Dynamic 3D Reconstruction Improvement via Intensity Video Guided 4D Fusion

Jie Zhang, Christos Maniatis, Luis Horna and Robert B. Fisher

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.

Index Terms

high-speed 3D video sensor, multi-frame 4D fusion, intensity tracking, dynamic object, noise reduction

I. 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.

Jie Zhang is with Beihang University, Beijing 100191, China

Jie Zhang, Luis Horna and Robert B. Fisher are with the School of Informatics, the University of Edinburgh, EH8 9AB, UK

Christos Maniatis was with the School of Mathematics, the University of Edinburgh, EH8 9AB, UK

Jie Zhang: zhangjie09@buaa.edu.cn; Robert B. Fisher: rbf@inf.ed.ac.uk

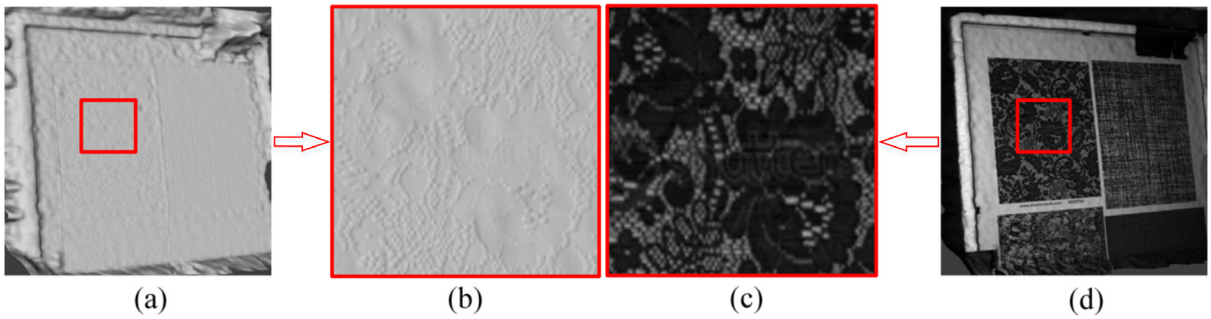

Fig. 1. Texture-related 3D noise on a static plane: (a) a 3D frame; (b) the region of interest of the 3D frame; (c) region of interest of the 3D frame with intensity texture; (d) the whole 3D frame with texture. The 3D noise in the 3D frame is closely related to the textures in the intensity image.

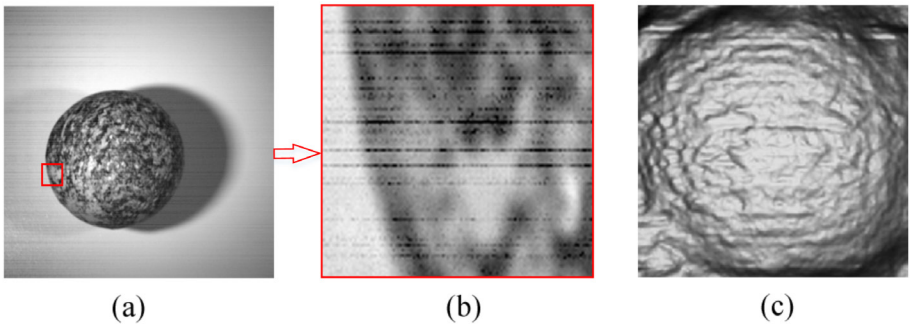

Fig. 2. Noise example: (a) an intensity frame of a falling sphere captured by a high-speed stereo video sensor; (b) invalid pixels in the intensity frames; (c) structural noise in a reconstructed 3D frame of the falling sphere.

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 IV); (2) a generic 2D intensity tracking guided multi-frame 4D fusion model that integrates spatial intra-frame filtering and temporal inter-frame fusion. (Section III). In Section V, we demonstrate the proposed method by denoising 3D sequences of stationary, dynamic and deformable objects from a 1000 fps 3D stereo video sensor.

II. Related Works

For 3D/depth noise reduction, 3D/depth noise characterization and models [6], [7], [8], [9], [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 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.

III. 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 $S^t = \{\mathbf{p}_i^t \in \mathcal{R}^3\}$ with pixel-wise registered intensity $I^t = \{a_i^t \in \mathcal{R}\}$ and depth images $D^t = \{d_i^t \in \mathcal{R}\}$, 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 IV.

IV. 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.

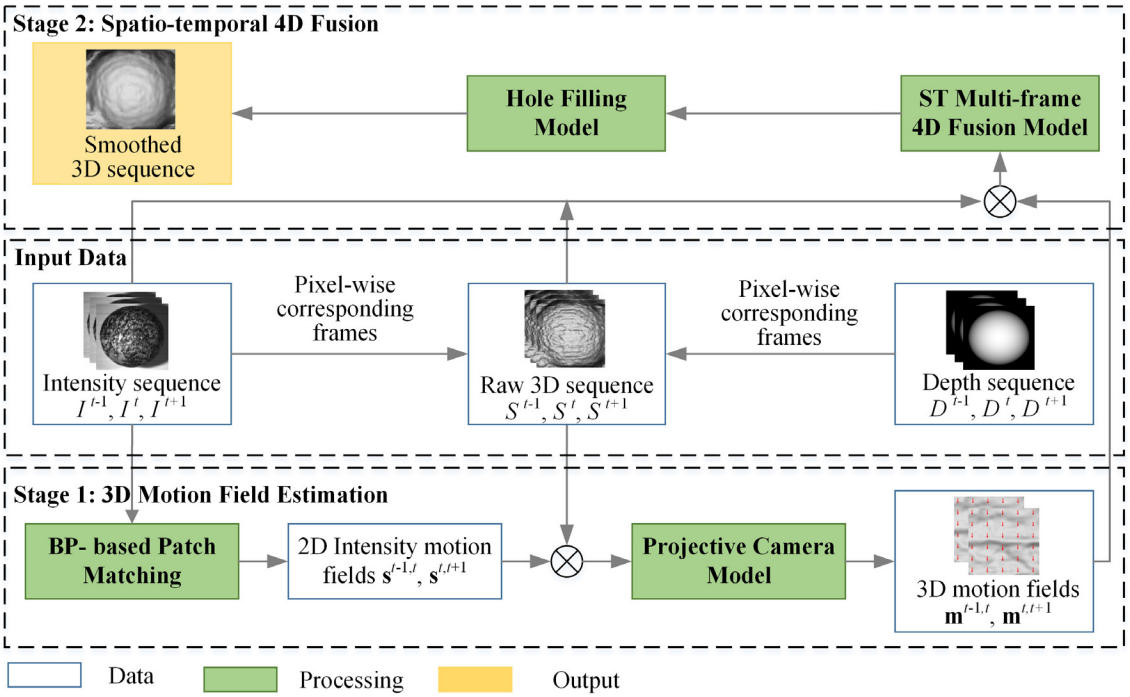

Fig. 3. The system framework (using 3 consecutive frames as an example).

A. *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 $\{\mathbf{s}^{t,t+1} \in \mathcal{R}^2\}$ 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:

$$\hat{\mathbf{s}}^{t,t+1} = \arg\min_{\mathbf{s}^{t,t+1}} \sum_i (\psi_1(\mathbf{s}_i^{t,t+1}) + \sum_{n \in N_I(i)} \psi_2(\mathbf{s}_i^{t,t+1}, \mathbf{s}_n^{t,t+1})) \tag{1}$$

In Eqn.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

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

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:

$$\mathbf{m}_i^{t,T} = \left[\begin{array}{c} f_x^{-1}(x_i^t - u_0)(d_i^T - d_i^t) + f_x^{-1} d_i^T s_{ix}^{t,T} \\ f_y^{-1}(y_i^t - v_0)(d_i^T - d_i^t) + f_y^{-1} d_i^T s_{iy}^{t,T} \\ (d_i^T - d_i^t) \end{array} \right] \tag{5}$$

By tracking from frame to frame, we can link the intensity image point a_i^t to its 3D position $\mathbf{p}_i^t$ in all frames.

B. Spatio-temporal Multi-frame 4D Fusion

Given n consecutive 3D frames linked by the $n-1$ continuous 3D motion fields, we seek to fuse them into one frame for piecewise spatio-temporal smoothness. Firstly, the outliers in each 3D motion field are removed by verifying pairwise forward and backward motion vectors, using a threshold constraint. Specifically, for a pair of 3D motion vectors $\mathbf{m}_i^{t,t+1}$ (or expressed as $\mathbf{m}^{t,t+1}(x_i, y_i, z_i)$) and $\mathbf{m}_i^{t+1,t}$ between a pair of corresponding points $\{\mathbf{p}_i^t, \mathbf{p}_i^{t+1}\}$ in the frame S^t and S^{t+1} respectively, the sum of the vectors should be smaller than a threshold ϑ as:

$$\left\| \mathbf{m}^{t,t+1}(x_i, y_i, z_i) + \mathbf{m}^{t+1,t}(x_i + \mathbf{m}_{ix}^{t,t+1}, y_i + \mathbf{m}_{iy}^{t,t+1}, z_i + \mathbf{m}_{iz}^{t,t+1}) \right\| < \vartheta \tag{6}$$

The 3D motion vectors that satisfy the threshold constraint are accepted as reasonable motion vectors.

The piecewise spatio-temporal 4D fusion performed on the consecutive 3D frames is expressed as

$$\hat{\mathbf{p}}_i^t = \frac{1}{\kappa_i} \sum_{T \in N_t(t)} \nu_i^{t,T} f(t,T) \left\{ \left[\frac{1}{\kappa_i^T} \sum_{n \in N(i)} d(\mathbf{p}_i^T, \mathbf{p}_n^T) g(I_i^T, I_n^T) \mathbf{p}_n^T \right] - \mathbf{m}_i^{t,T} \right\} \tag{7}$$

In the internal summation, $\mathbf{p}_n^T (n \in N(i))$ is a set of neighbors of the point i in a 3D frame S^T. $d(\mathbf{p}_i^T, \mathbf{p}_n^T) = e^{-\left\|\mathbf{p}_i^T - \mathbf{p}_n^T\right\|^2 / 2\delta_d^2}$ and $g(I_i^T, I_n^T) = e^{-\left|I_i^T - I_n^T\right|^2 / 2\delta_g^2}$ are Gaussian weights assigned according to the spatial distance and the intensity difference, where the parameter δ_d and δ_g are adjustable in experiments. The intensity-guided weights contribute to the spatial smoothness of the 3D frame, which reduces 3D noise but preserves some geometric structure information. This internal summation computes a bilaterally smoothed point in the frame S^T, which is then mapped back to frame S^t using the integrated motion vectors $\mathbf{m}_i^{t,T}$ (e.g. $\mathbf{m}_i^{t,t+2} = \mathbf{m}_i^{t,t+1} + \mathbf{m}_i^{t+1,t+2}$).

In the external summation, $N_t(t)$ is a set of neighboring frames S^T of the frame S^t. $\nu_i^{t,T}$ is a flag for the validity of the integrated 3D motion vector from frame S^t to S^T. $f(t,T) = e^{-(t-T)^2/2\delta_f^2}$ is a weight assigned according to the temporal distance. κ_i^T and κ_i are the cardinalities of the normalization factors for inter-frame fusion and intra-frame filtering respectively. Specifically, κ_i is the sum of the weights $\nu_i^{t,T} f(t,T)$, and κ_i^T is the sum of the weights $d(\mathbf{p}_i^T, \mathbf{p}_n^T) g(I_i^T, I_n^T)$. Overall, Eqn.7 gives a smoothed 3D point $\hat{\mathbf{p}}_i^t$ in the frame S^t. Both the spatial and temporal piecewise smoothness are guided by the 2D intensity information.

C. Hole Filling

Finally, a point $\mathbf{p}_i^t$ without spatial or temporal neighbors is filled with an interpolated point by using its spatial neighboring 3D points as

$$\hat{\mathbf{p}}_i^t = \begin{cases} \text{4D fusion using } Eqn.7 & if \text{ satisfying } Eqn.6 \\ \frac{1}{\kappa_i'} \sum\limits_{\mathbf{p}_n^t \in N_S(\mathbf{p}_i^t)} h(\mathbf{p}_i^t, \mathbf{p}_n^t) \left(\mathbf{p}_n^t + \left\langle \nabla_{\mathbf{p}_i^t \mathbf{p}_n^t}, \mathbf{p}_i^t - \mathbf{p}_n^t \right\rangle \right) & otherwise \end{cases} \tag{8}$$

where $N_S(\mathbf{p}_i^t)$ is a set of spatial neighbors of $\mathbf{p}_i^t$, n is the index of the neighbor, $h(\mathbf{p}_i^t, \mathbf{p}_n^t) = e^{-\left\|\mathbf{p}_i^t - \mathbf{p}_n^t\right\|^2/2\delta_h^2}$ is the Gaussian weight assigned according to the spatial distance. κ_i' is the cardinality of a normalization factor, which can be obtained by summing up the weights $h(\mathbf{p}_i^t, \mathbf{p}_n^t)$. $\nabla_{\mathbf{p}_i^t \mathbf{p}_n^t}$ is the 3D location gradient of the neighboring point $\mathbf{p}_n^t$. We can get the 3D gradient by computing the partial derivatives (along the direction x, y, z) of the 3D local shape fitted using all the spatial neighboring points $\{\mathbf{p}_n^t\}$. At last, $\langle\rangle$ is the inner (dot) product.

Theoretically, the intensity guidance avoids the impact of 3D noise on the 3D motion field estimation. Accurate temporal "stereo" correspondences lead inter-frame dense point fusion to reduce temporal fluctuations of the fused points, while without adding structural noise simultaneously. Additionally, the intra-frame filtering helps degrade local structural noise in the spatial domain. As a result, we can obtain a fused 3D sequence for every point i with less spatial noise and temporal fluctuations in the original 3D image sequence.

V. Results and Discussion

This section presents synthetic tests and real experiments to verify the effectiveness and the robustness of the proposed method. The synthetic tests focused on noise resistance and preserving shape correctness respectively. The real experiments investigated practical data improvement using a high frame rate 3D sensor (1000 fps).

A. Synthetic Noise Test

The synthetic measured object is a falling 3D ball with the radius of 140 mm, as shown in Fig.4. The synthetic 3D sequence contains 50 3D frames. The resolutions of the intensity image and depth image are 600×600 pixels and 600×600 points respectively. The sphere fell with the speed of 2 pixels/frame. The roughness of the 3D surface in one frame was measured by averaging (over the central area of the sphere) the local roughness Π_i of a 3D point $\mathbf{p}_i^t$ relative to its neighboring patch with the size of $n \times n$ points as

$$\Pi_i = \frac{1}{n^2} \sum_j^{n \times n} \frac{(\mathbf{p}_i^t - \mathbf{p}_j^t) \cdot \mathbf{n}_i}{|\mathbf{n}_i|} \tag{9}$$

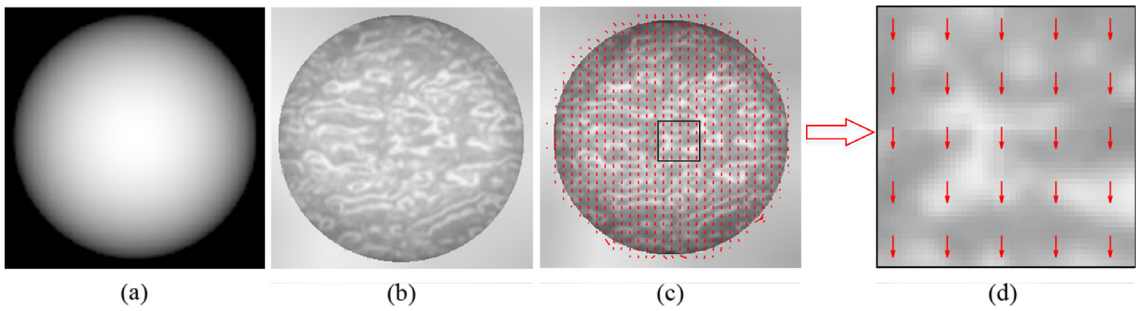

Fig. 4. Example frames of the synthetic falling ball sequence: (a) depth frame; (b) intensity frame; (c) 2D motion field of 2 neighboring frames; (d) 2D motion field of region of interest in (c).

where $\mathbf{p}_j^t$ is the neighboring point in the $n \times n$ window around the central point $\mathbf{p}_i^t$ (here we use $n = 5$), $\mathbf{n}_i$ is the normal vector of the fitted plane of the neighboring points. Note that this form of roughness measure does not have value zero when there is no noise, due to the curvature of the surface. We used this roughness measure to evaluate the performance because there is no ground truth for the real data experiments and we wanted to be able to compare the simulated and real results using the same measure.

We added Gaussian random noise with varying noise levels to the intensity and depth images, respectively, and then calculated the mean roughness of the reconstructed 3D sequence. The depth noise standard deviation varies from 0.1 mm to 0.5 mm (The range was chosen because it includes the standard deviation (0.15 mm) of the stereo sensor used in the real experiments). The intensity values are normalized to [0 1] and the intensity noise level varies from 2% to 10% of the highest intensity value. The results were compared with other existing methods including Ad-GF [9], Ad-BF [12], Guided filter [21], DNBF [18], TA [19], Ad-JBF [20], and ST-MF [22]. The mean roughness results (over all frames) w.r.t. different noise levels and algorithms are shown in Fig.5.

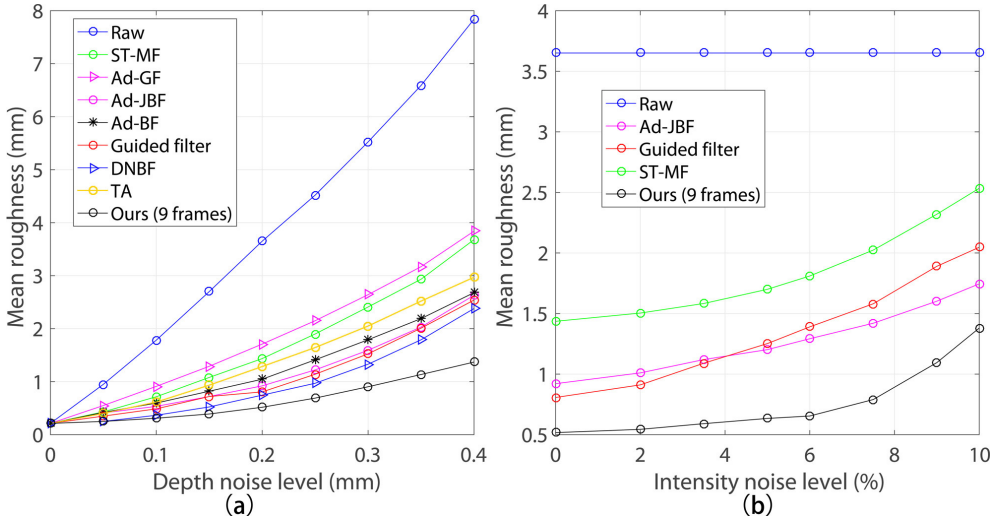

Fig. 5. Mean roughness vs. (a) Depth noise level (with intensity noise level of 3%); (b) Intensity noise level (with depth noise level of 0.2 mm).

The results in Fig.5a demonstrate that the performance of the proposed algorithm is superior to other algorithms especially at higher depth noise levels. Some intensity-joint or motion-joint algorithms (Ad-JBF, Guided filtering, DNBF) achieve better results on the synthetic noisy 3D ball than the single image based algorithms such as Ad-BF and Ad-GF. In Fig.5b, our algorithm has better performance over all the intensity noise levels, followed by the Guided filtering and Ad-JBF. Specifically, for our algorithm, the increase of the mean roughness at lower intensity noise levels is smaller than that at higher levels. This is because the 3D motion vectors are quantized to integral points and some wrong sub-pixel motion vectors are rejected at the stage of 3D motion field estimation, which increases the robustness of the intensity guided fusion method.

B. Roughness vs. Shape Correctness Test

Roughness and shape correctness are important coupled parameters for describing the quality of 3D reconstructed data. We seek to reduce the roughness of 3D data without losing the shape correctness by over-smoothing. Using the falling noisy synthetic sphere (with known ground truth), we investigated the balance between the reduction in roughness and in shape correctness of different algorithms, as the amount of smoothing is varied. The results are shown in Fig.6. The shape correctness is defined as

$$C = 1 - \frac{|r - \overline{r}|}{\overline{r}} \tag{10}$$

where r is the estimated radius of the sphere, computed by the MLESAC algorithm [25] over data from pixels 160 to 440 (as shown in Fig.6a); $\overline{r}$ is the ground truth radius.

Fig.6a illustrates the balance between the roughness and the shape correctness on the noisy ball from a side view. Our algorithm's smoothed depth values (black curve) have both lower roughness and better shape correctness than the raw values, while the DNBF smoothed depth values (red curve) has worse shape correctness when reaching the same roughness (Here we only show the performance of DNBF as an qualitative example, while the full quantitative comparable performance are shown in Fig.6b). That means the roughness improvement is achieved by sacrificing some shape correctness, which causes unexpected global deformations of the object.

When generating Fig.6b, for each algorithm, we varied the size of the smoothing neighborhood and the number of smoothing iterations to enable the algorithms to generate different roughnesses and to investigate the corresponding shape correctness simultaneously. Note that different parameter settings were used to generate (roughness, shape correctness) pairs, so the tendency curves are not functions. The initial depth noise level is 0.2 mm and the intensity noise level is 2%. The quantitative results are shown in Fig.6b. Overall, applying different noise reduction algorithms, the mean roughness decreases from the raw roughness (3.75 mm) in different degrees, with increasing shape correctness. However, after the best point, over-smoothing causes serious shape correctness loss with almost the same or even slightly decreasing roughness. Specifically, the curves show that our proposed algorithm achieves the best performance (nearest upper left corner), which demonstrates that it can denoise the 3D data while preserving the structural information better.

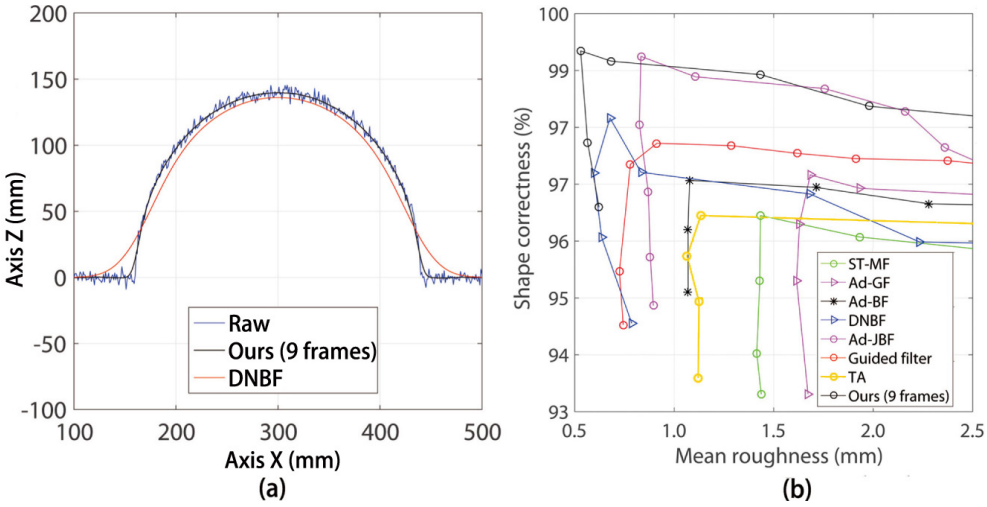

Fig. 6. (a) Qualitative illustration of the roughness and the shape correctness on the ball (Here we show the results from only the three most significant algorithms. The full quantitative comparison can be seen in the following sub-figure); (b) Roughness vs. shape correctness tendency curves. (The raw data has mean roughness of 3.75 mm and shape correctness of 88.34%, so the common starting point of those curves is on the lower right corner. Note that different parameter settings were used to generate (correctness, roughness) pairs, so the curves are not functions)

C. Results on High Frame Rate Sensors

We captured 3D sequences of four real 3D objects using a high-speed 3D stereo video sensor and investigated the performance of the proposed method. The high-speed 3D stereo video sensor is from Dimensional Imaging (DI4D) Ltd [26] and mainly consists of two intensity video cameras with the frame rate of 1000 fps. Pairwise images can be captured and then processed offline using a hierarchical dense area matching stereo algorithm proprietary to the DI4D Ltd, but derived from the research reported in [27].

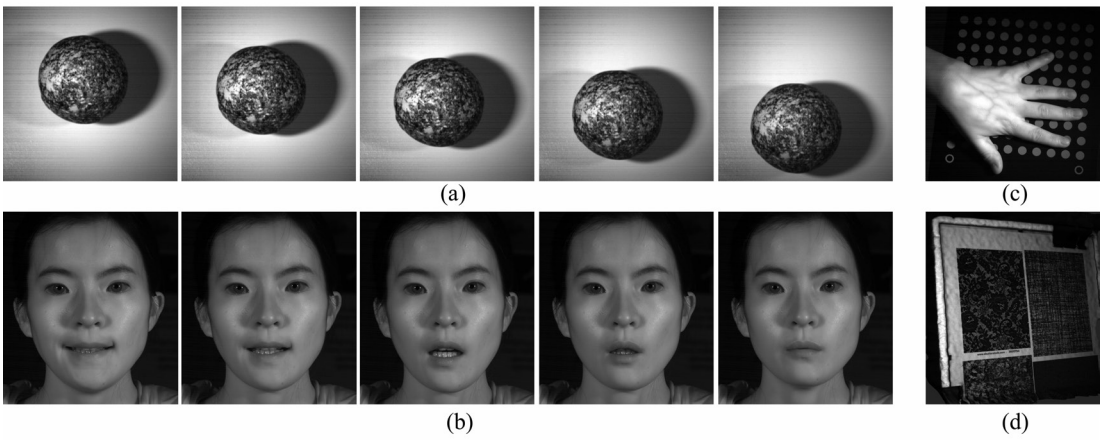

Fig. 7. Intensity image examples of 4 real measured objects: (a) falling rubber ball; (b) speaking human face; (c) static hand; (d) static texture plane.

The four real 3D objects are with different states and surface complexities, including a static plane, a static hand,

a falling rubber ball and a speaking human face (as shown in Fig.7). The measured stationary plane with textures is $\sim 120 \times 80$ mm. The radius of the ball is ~ 70 mm. The 3D sequence of the ball is time-varying since the ball deforms and rotates slightly during the falling. We applied the proposed method with varying numbers of fused frames to each measured object. For each number of fused frames, we calculated the mean roughness and standard deviation (std) of the 3D sequence. The results are shown in Fig.8.

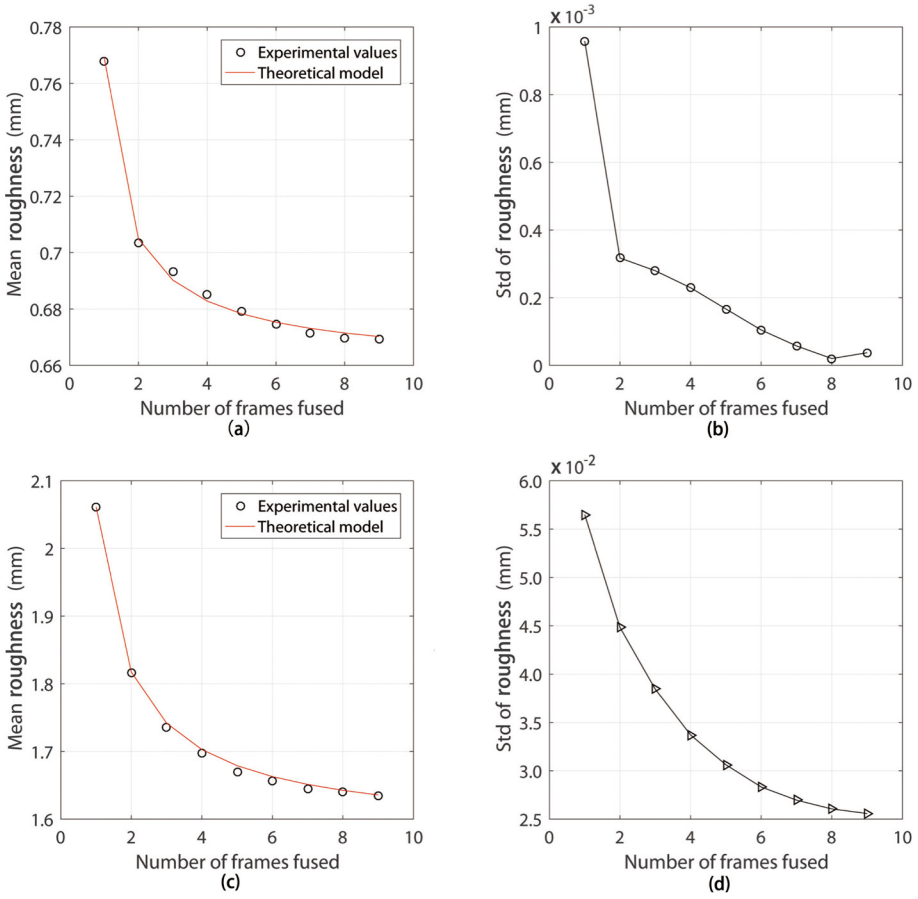

Fig. 8. Static Plane (first row): (a) mean roughness; (b) std of roughness vs. number of frames fused. Falling ball (second row): (c) mean roughness; (d) std of roughness vs. number of frames fused.

One can model the mean roughness presented in Fig.8 as $\sqrt{\delta_s^2 + (1/n)\delta_t^2}$, where δ_s is the std of the structural noise, δ_t is the std of the time-varying noise, and n is the number of frames fused. The red line in Fig.8a and Fig.8c show the above theoretical results fit the experimental results closely. It is obvious that both the mean and std of roughness decrease with the increasing number of frames fused. Compared with the static object, the std of roughness of the dynamic object falls more sharply, when the number of fused images varies from 2 to 9. This is because the number of fused frames mainly influences the temporal dynamic noise, while the dominant noise of the static object is regular structural noise. Overall, we can conclude that the proposed intensity-guided 4D fusion algorithm is more effective and suitable for boosting the 3D reconstruction of dynamic objects.

A qualitative example result of the proposed method when fusing 9 frames is shown in Fig.9, and we also show example intermediate results of the falling ball and the speaking face in Fig.10, including guiding intensity frames and their motion field with filtered holes (green motion vectors). Since the sequence uses a 1000 fps frame rate, the motion field is relatively tough to view, we choose two frames with a 10 frame gap. For quantitative comparison, results are shown in Table 1.

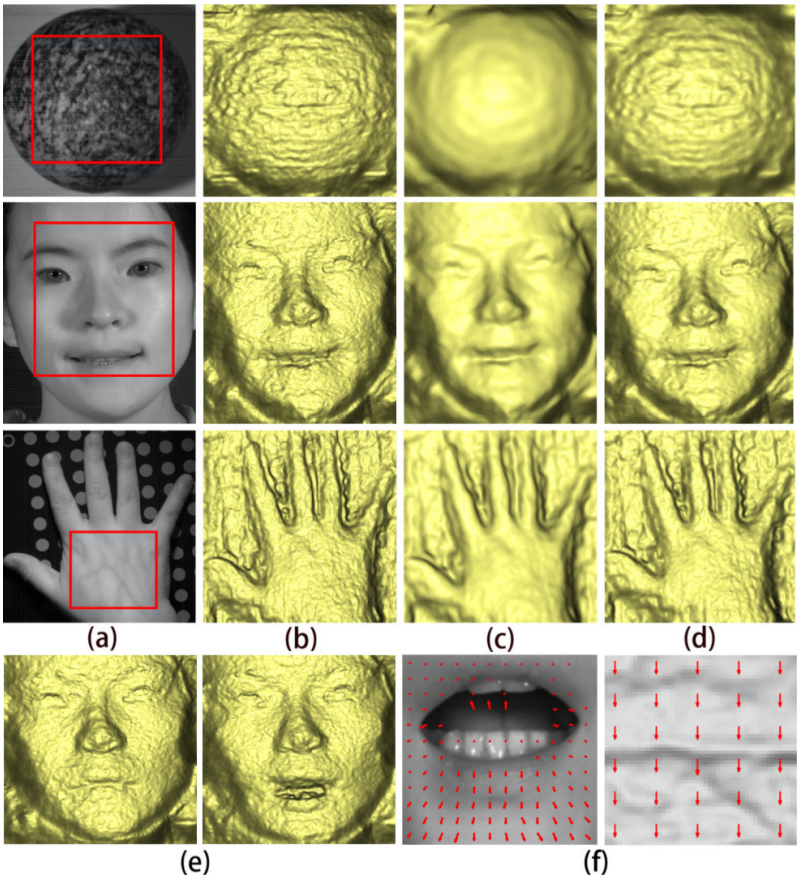

Fig. 9. From first to third row: falling ball, dynamic human face, static hand. (a) Intensity frame at time t with a ROI marked using a red box; (b) Raw registered 3D (cosine shaded) frame at time t; (c) Improved 3D frame by our algorithm; (d) Improved 3D image by Ad-JBF [20]; (e) Raw dynamic human face frames at time $t-100$ and $t+100$ respectively; (f) Motion field: the mouth region of the human face (left) and the center region of the falling ball (right).

From the qualitative results in Fig.9 we see that the 3D noise is obviously reduced by the proposed algorithm so that the surfaces of interest of the observed 3D objects are much smoother than those in the raw 3D images, especially for the falling ball. Correspondingly, the comparative results in Table 1 demonstrate that our method achieves the best performance with the lowest mean roughness (spatial noise) and the most stable roughness measure (std: temporal fluctuations).

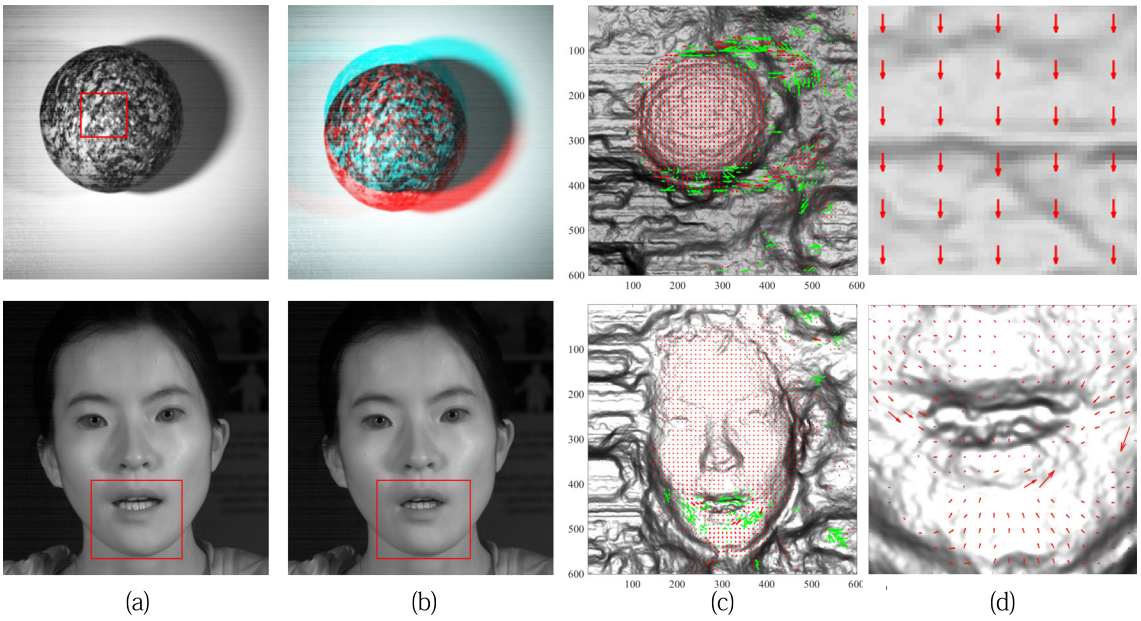

Fig. 10. Example intermediate results: (a) & (b) two guided intensity frames; (c) 2D motion field of the two consecutive frames, the red motion vectors are correct with motion consistence (in Eqn.6), while the green ones are wrong and are thus filtered out (causing holes); (d) the region of interest in the subfigure (c).

TABLE I
COMPARATIVE RESULTS OF 3D/DEPTH NOISE REDUCTION METHODS

	plane (mm)		hand (mm)		falling ball (mm)		dynamic face (mm)	
	mean	std $(\times 10^{-3})$	mean	std $(\times 10^{-2})$	mean	std $(\times 10^{-2})$	mean	std $(\times 10^{-2})$
Raw	0.62	1.91	1.34	1.81	1.48	7.92	1.10	6.76
Ad-GF [9]	0.39	1.12	0.89	1.41	1.11	3.95	0.77	6.08
Ad-BF [12]	0.26	0.82	0.64	1.21	0.89	3.84	0.59	5.45
Guided filter [21]	0.34	0.73	0.61	1.21	0.83	2.43	0.60	5.09
DNBF [18]	0.32	0.71	0.60	1.18	0.84	2.65	0.58	5.24
TA [19]	0.34	0.65	0.71	1.09	0.93	2.41	0.67	4.65
Ad-JBF [20]	0.27	0.61	0.64	1.21	0.89	3.81	0.59	5.45
ST-MF [22]	0.36	1.01	0.83	1.38	1.05	3.62	0.73	5.95
KinectFusion [13]	0.39	0.59	0.78	0.91	-	-	-	-
3D Deformable Scanning [28]	-	-	-	-	0.81	1.97	0.52	**2.67**
Ours (9 frames)	**0.22**	**0.31**	**0.55**	**0.83**	**0.71**	**1.14**	**0.40**	2.73

D. In Comparison to 6D Motion Field Based Fusion

In contrast to 4D fusion based on intensity motion fields for 3D/depth noise reduction, there is a group of algorithms that directly generate volumetric 6D motion fields $\{\mathbf{R}_i, \mathbf{T}_i\}$ using depth data from Kinect sensors and reconstruct improved 3D scenes via dense 3D/depth frame registration, such as KinectFusion [13], DynamicFusion [16], 3D Deformable Scanning [28], etc. In those works, the multi-view partial 2.5D scans from the Kinect sensors allow for large geometric and pose variations, while our algorithm works on consecutive frames from a fixed 1000

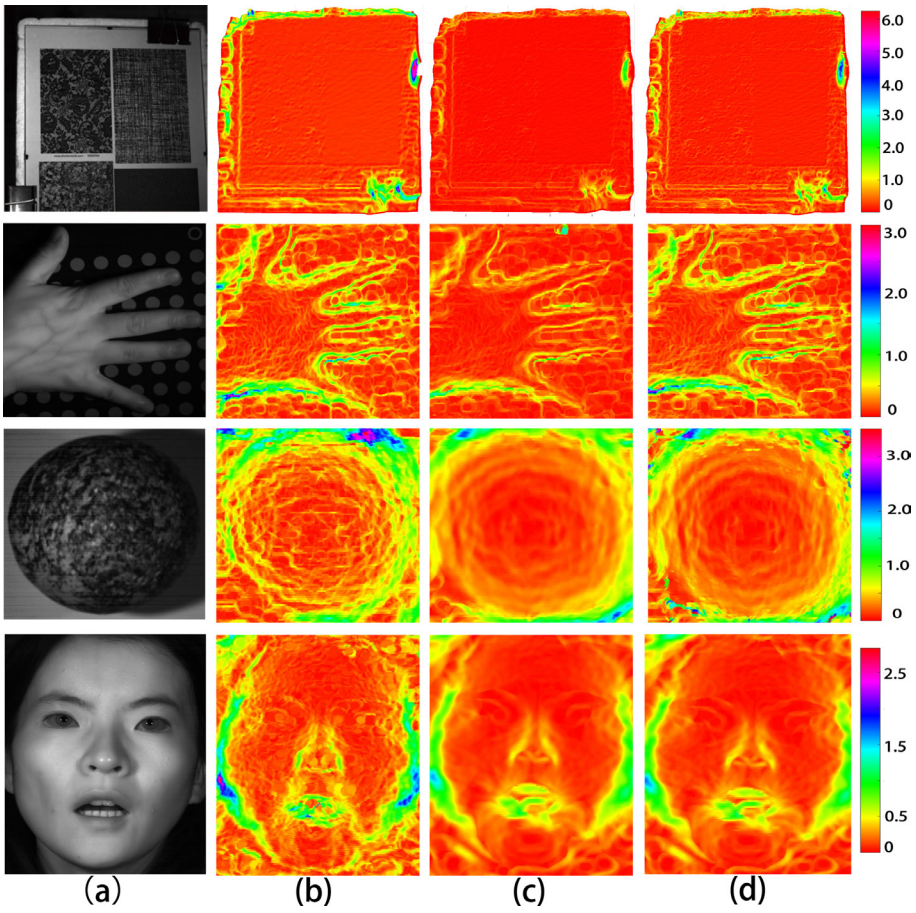

Fig. 11. From top to bottom row: static plane, static hand, falling ball, speaking human face. (a) Intensity frame at time t; (b) Roughness map of raw registered 3D frame at time t; (c) Roughness map of an improved 3D frame by our algorithm; (d) Roughness map of an improved 3D image by 3D motion field based algorithms [13], [28]. (Best viewed online in color).

fps 3D video sensor focusing on dense micro-deformation and fusion. Besides, the 3D noise from the 1000 fps video sensor is closely related to the textures of the observed 3D objects due to the uneven reflectance of the textures, as shown in Fig.1. Therefore, we directly use intensity information to generate intensity motion fields, guiding the spatio-temporal fusion.

We compared the performance of the proposed algorithm on the same four objects with the 6D motion field based fusion algorithms. For static objects including the static plane and the hand, a 6D transformation between a pair of consecutive 3D frames was generated using the rigid ICP algorithm, then all the registered 3D points were integrated into a volumetric representation for fusion. For the dynamic and deformable objects including the falling ball and the speaking human face, a dense 6D warp field between pairwise 3D frames was generated using the Embedded Deformable model (ED) based registration method [28]. Then, 9 consecutive frames were fused by leveraging the 8 dense flow fields between each pair of 3D frames. We calculated the roughnesses of the surface of each object and mapped them to the object as shown in Fig.11. The mean roughness and standard deviation of all 3D frames in a sequence were calculated, as listed in Table 1.

Overall, both the qualitative results in Fig.11 and comparable results in Table 1 show that our algorithm achieves better results on the datasets. The use of the 2D intensity frames increases the accuracy of dense correspondence and thus improves the spatio-temporal fusion for 3D noise reduction of high frame rate 3D video sensors, especially for the objects with less 3D shape characteristics, such as the plane, hand and ball. Also, our algorithm directly focuses on texture-related 3D noise (Fig.1), yielding a texture correspondence guided dense 3D motion field. It is more suitable for high frame rate 3D sequences of dynamic and deformable objects even with fewer 3D shape features.

VI. Conclusions

This paper presents a simple yet powerful pipeline for improving the 3D reconstruction of dynamic and deformable objects, using 2D intensity tracking guided multi-frame 4D fusion. Firstly, the continuous motion fields of a 3D sequence are estimated by leveraging the intensity motion fields that are obtained by dense tracking on a pixel-wise registered 2D intensity sequence. Then, using a spatial-temporal multi-frame 4D fusion model, consecutive 3D frame fusions are performed for improving the spatial smoothness and the temporal stability of the 3D sequence. The experimental results on stationary, dynamic and deforming objects verify that the proposed method achieves state-of-the-art performance with the lowest mean roughness over the reconstructed 3D surface in one frame and the best robustness over the whole 3D sequence. In the future, we would like to apply the proposed method as a part of dynamic 3D shape recognition (e.g. dynamic 3D human face and hand gesture recognition) to improve the accuracy and robustness of the 3D reconstruction and the recognition of highly dynamic and deformable objects.

Acknowledgments

This work is supported by the China Scholarship Council (No. 201606020087), National Council for Science and Technology (CONACyT) of Mexico.

References

[1] Y. Xiao, R. B. Fisher, and M. Oscar, "Performance characterization of a high-speed stereo vision sensor for acquisition of time-varying 3d shapes," *Machine Vision and Applications*, vol. 22, no. 3, pp. 535–549, 2011.

[2] S. Tabata, S. Noguchi, Y. Watanabe, and M. Ishikawa, "High-speed 3d sensing with three-view geometry using a segmented pattern," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 3900–3907.

[3] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[4] J. Wang and Z. Xu, "Stv-based video feature processing for action recognition," *Signal Processing*, vol. 93, no. 8, pp. 2151–2168, 2013.

[5] J. Xiang and R. Liang, "Motion recognition and synthesis based on 3d sparse representation," *Signal Processing*, vol. 110, pp. 82–93, 2015.

[6] T. Mallick, P. P. Das, and A. K. Majumdar, "Characterizations of noise in kinect depth images: A review," *IEEE Sensors journal*, vol. 14, no. 6, pp. 1731–1740, 2014.

[7] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[8] Y. Yu, Y. Song, Y. Zhang, and S. Wen, "A shadow repair approach for kinect depth maps," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 615–626.

[9] C. V. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3d reconstruction and tracking," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*. IEEE, 2012, pp. 524–530.

[10] J.-H. Park, Y.-D. Shin, J.-H. Bae, and M.-H. Baeg, "Spatial uncertainty model for visual features using a kinect sensor," *Sensors*, vol. 12, no. 7, pp. 8640–8662, 2012.

[11] M. A. Garduño-Ramón, I. R. Terol-Villalobos, R. A. Osornio-Rios, and L. A. Morales-Hernández, "A new method for inpainting of depth maps from time-of-flight sensors based on a modified closing by reconstruction algorithm," *J. Visual Communication and Image Representation*, vol. 47, pp. 36–47, 2017.

[12] L. Chen, H. Lin, and S. Li, "Depth image enhancement for kinect using region growing and bilateral filter," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3070–3073.

[13] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 559–568.

[14] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 192, 2016.

[15] D. Guo, C. Li, L. Wu, and J. Yang, "Improved marching tetrahedra algorithm based on hierarchical signed distance field and multi-scale depth map fusion for 3d reconstruction," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 491–501, 2017.

[16] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352.

[17] K. Essmaeel, L. Gallo, E. Damiani, G. De Pietro, and A. Dipandà, "Temporal denoising of kinect depth data," in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*. IEEE, 2012, pp. 47–52.

[18] J. Fu, S. Wang, Y. Lu, S. Li, and W. Zeng, "Kinect-like depth denoising," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 512–515.

[19] J. Wasza, S. Bauer, and J. Hornegger, "Real-time preprocessing for dense 3-d range imaging on the gpu: defect interpolation, bilateral temporal averaging and guided filtering," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1221–1227.

[20] M. Camplani, T. Mantecon, and L. Salgado, "Depth-color fusion strategy for 3-d scene modeling with kinect," *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1560–1571, 2013.

[21] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[22] S. Matyunin, D. Vatolin, Y. Berdnikov, and M. Smirnov, "Temporal filtering for depth maps generated by kinect depth camera," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*. IEEE, 2011, pp. 1–4.

[23] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, and A. El Saddik, "Evaluating and improving the depth accuracy of kinect for windows v2," *IEEE Sensors Journal*, vol. 15, no. 8, pp. 4275–4285, 2015.

[24] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz, "Pmbp: Patchmatch belief propagation for correspondence field estimation," *International Journal of Computer Vision*, vol. 110, no. 1, pp. 2–13, 2014.

[25] P. H. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[26] "Dimensional imaging (di4d)," http://www.di4d.com/, accessible on May. 2, 2018.

[27] C. W. Urquhart, J. P. Siebert, J. P. McDonald, and R. J. Fryer, "Active animate stereo vision." in *BMVC*, 1993, pp. 1–10.

[28] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3d scanning deformable objects with a single rgbd sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 493–501.