Full 3D Reconstruction of Non-Rigidly Deforming Objects

HASSAN AFZAL and DJAMILA AOUADA*, University of Luxembourg, Luxembourg
BRUNO MIRBACH, IEE S.A., Luxembourg
BJÖRN OTTERSTEN, University of Luxembourg, Luxembourg

In this work, we discuss enhanced full 360° 3D reconstruction of dynamic scenes containing non-rigidly deforming objects using data acquired from commodity depth or 3D cameras. Several approaches for enhanced and full 3D reconstruction of non-rigid objects have been proposed in the literature. These approaches suffer from several limitations due to requirement of a template, inability to tackle large local deformations and topology changes, inability to tackle highly noisy and low-resolution data, and inability to produce on-line results. We target on-line and template-free enhancement of the quality of noisy and low-resolution full 3D reconstructions of dynamic non-rigid objects. For this purpose, we propose a view-independent recursive and dynamic multi-frame 3D super-resolution scheme for noise removal and resolution enhancement of 3D measurements. The proposed scheme tracks the position and motion of each 3D point at every-time step by making use of the current acquisition and the result of the previous iteration. The affects of system blur due to per-point tracking are subsequently tackled by introducing a novel and efficient multi-level 3D bilateral total variation regularization. These characteristics enable the proposed scheme to handle large deformations and topology changes accurately. A thorough evaluation of the proposed scheme on both real and simulated data is carried out. The results show that the proposed scheme improves upon the performance of the state-of-art methods and is able to accurately enhance the quality of low-resolution and highly noisy 3D reconstructions while being robust to large local deformations.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; *3D imaging*; Active vision;

Additional Key Words and Phrases: 3D reconstruction, super-resolution, non-rigid registration, point-cloud enhancement, 3D bilateral total variation, 3D point tracking

ACM Reference Format:
Hassan Afzal, Djamila Aouada, Bruno Mirbach, and Björn Ottersten. 2018. Full 3D Reconstruction of Non-Rigidly Deforming Objects. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1 (January 2018), 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Acquiring high quality and complete 360° 3D reconstructions of dynamic scenes containing non-rigidly deforming objects is one of the fundamental goals of research in computer vision and robotics. Such reconstructions can be effective in solving various problems in the domains of security and surveillance, virtual reality and gaming, etc.

*The corresponding author

Authors' addresses: Hassan Afzal, hassan.afzal@uni.lu; Djamila Aouada, djamila.aouada@uni.lu, University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust (SnT), Luxembourg, L-1855, Luxembourg; Bruno Mirbach, IEE S.A., Advanced Engineering Department, Contern, L-5326, Luxembourg; Björn Ottersten, University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust (SnT), Luxembourg, L-1855, Luxembourg.

1551-6857/2018/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Reconstruction of 3D world has traditionally been achieved via photometric cameras by correlating same 3D points in 2D images across different views. Non-rigid objects have been reconstructed using both mono-view systems [27, 66], which provide partial coverage of the scene, and multi-view systems which provide complete coverage of the scene instantaneously [15, 20, 58]. Due to the limitations of photometric sensing systems most of these methods usually require expensive and highly constrained setups [59]. They either use pre-built templates of target objects [14, 56, 58], or build them as a first step [20, 26, 58, 66], as shape priors for e.g., non-rigid tracking, hole filling and shape completion etc.

Commodity depth cameras such as Microsoft Kinect v1 and v2 [41], Asus Xtion Pro Live [9] and PMD camboard nano [49], have opened further the possibilities of research in this domain by providing 2.5D information which can directly be converted into 3D point clouds. This, on the one hand, diminishes barriers on highly constrained setups but, on other hand, poses challenges due to noisy and, in some cases, low-resolution (LR) acquired measurements [49]. Hence, the goal of acquiring high quality and complete 3D reconstructions of non-rigidly deforming objects still remains unfulfilled.

Recently, researchers have tried to overcome these challenges by focusing on building complete and enhanced 3D reconstructions of non-rigidly deforming objects using commodity depth cameras. Apart from template based approaches [22, 65, 68, 70], there are mono-view methods which build complete and enhanced 3D reconstructions by incrementally fusing temporal information [18, 64, 67]. Such methods are view-dependent, and therefore cannot directly be used to provide full 3D reconstructions, instantaneously [3, 45]. View-independent temporal fusion based reconstruction methods such as VI-KinectDeform have also been proposed but they do not target LR measurements [1]. RecUP-SR, on the other hand, targets quality and resolution enhancement of 3D reconstructions using depth maps only, but is restricted to mono-view partial reconstructions and may not be robust to fast and abrupt 3D motions [33, 34].

Recently proposed multi-view systems based template-free methods which provide complete and enhanced reconstructions of non-rigidly deforming objects instantaneously are also not robust to LR depth measurements. Moreover they are either restricted to limited deformations [38, 57, 64, 67] or, are sensitive to certain topology changes [21, 52].

In this work we propose to tackle several challenges which lie in the way of achieving on-line, high quality, and complete 3D reconstructions of scenes, containing non-rigidly deforming objects with little constraints on their topology and motion. We base our work on measurements acquired via commodity depth cameras due to their low cost, flexibility and instantaneous capture of 3D information, but the challenges which arise due to LR and noisy measurements of these cameras need to be tackled as well. For this purpose we propose a view-independent recursive and dynamic multi-frame 3D super-resolution (SR) scheme. This scheme targets enhancement of resolution and quality of noisy LR 3D measurements, of non-rigidly deforming objects, acquired by commodity depth sensors.

The proposed approach is template-free and works directly on 3D points. This gives it flexibility to the types of objects being reconstructed, and the ability to capture their characteristics, i.e. position and motion in the 3D world more accurately. The affects of system blur are tackled via a novel and efficient multi-level 3D bilateral total variation (BTV) regularization. To our knowledge, the proposed algorithm is the first view-independent, recursive and dynamic multi-frame 3D SR method which targets complete 3D reconstructions of scenes/objects. The recursive nature of this algorithm allows it to produce enhanced high-resolution (HR) point clouds at each time-step by taking as input only the current noisy and LR measurement and the resulting noise-free HR point cloud obtained at the previous time-step. The pipeline of the proposed algorithm is shown in Fig. 1, and details follow.

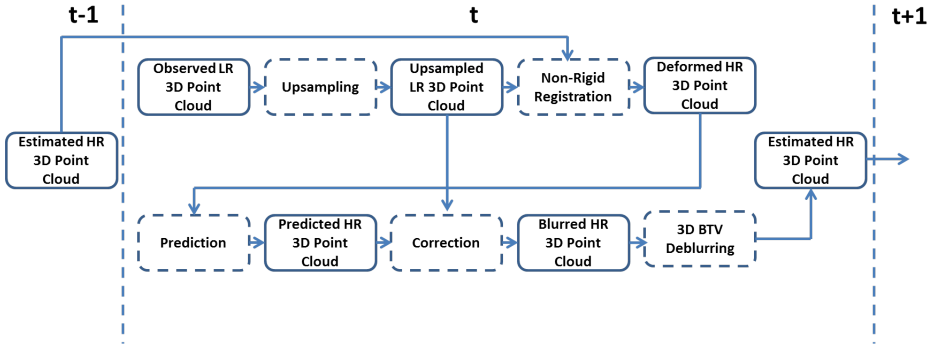


Fig. 1. Detailed pipeline of the proposed recursive dynamic multi-frame 3D super-resolution algorithm. For more details please see Section 4.

1.1 Contributions

We propose a novel recursive and dynamic multi-frame 3D super-resolution scheme for producing 3D videos containing enhanced and complete 3D reconstructions of non-rigidly deforming objects:

- The proposed scheme produces HR, enhanced and complete 3D reconstructions recursively by fusing the current acquisition, from a depth/3D camera system, and the result of previous iteration.
- It is a view-independent, and template-free, resolution and quality enhancement scheme based on per-point tracking in 3D space which allows it to be robust to changes in topology, and large motions.
- We have formulated this scheme to handle per-coordinate independent as-well-as depth camera specific noise in the acquired 3D points.
- A novel and efficient multi-level 3D Bilateral Total Variation (BTV) regularization is also proposed. It is used to handle system blur and correct per-point position and motion estimates, at every iteration.
- Detailed experimental, quantitative and qualitative, evaluations have been carried out using both simulated and real data. Results show that the proposed dynamic scheme out-performs the state-of-art filtering algorithms and produces accurate, smooth and feature-preserving 3D reconstructions.

1.2 Article Overview

This paper is organized as follows: We start by giving a brief over-view of state-of-art methods for acquiring complete and enhanced 3D reconstructions of non-rigidly deforming objects using depth cameras in Section 2, Section 3 formulates the problem of recursive dynamic multi-frame 3D SR. It is followed by details of the proposed algorithm which solves the given problem by tracking and filtering the position and motion of each 3D point, in LR upsampled measurements affected by per-coordinate independent as well as depth camera dependent noise, in Section 4. After per-point tracking a smoothing and deblurring step is required, at each time-step, for 3D point position and motion estimates. For this purpose a novel 3D BTV regularization is proposed. In Section 5, results of the proposed approach based on qualitative and quantitative experiments in comparison with state-of-art methods are presented which is followed by a conclusion in Section 6.

2 RELATED WORK

In this section we review the state-of-art related to enhancement of 3D dynamic videos containing non-rigidly deforming objects. Our focus will be on depth cameras based techniques which are used to acquire complete and noise-free 3D reconstructions of non-rigidly deforming objects.

Compared to photometric cameras, commodity 3D cameras based reconstruction approaches, although aided by 3D acquisitions, have to overcome problems related to noise and limited resolution. After the advent of commodity RGB-D or 3D cameras based enhanced 3D reconstruction techniques for rigid objects [11, 19, 46, 50], researchers have moved towards handling non-rigid deformations by proposing to construct complete and enhanced 3D models of mainly human subjects by fusing information from multiple views. This requires handling quasi-rigid motions between different views for which a global non-rigid registration is performed [38, 57], or a model-to-part registration based on deformation graph [53] or Shape Completion and Animation of People (SCAPE) model [5] is used to avoid error accumulation [63, 67]. The works of Cui et al. [17] and Shapiro et al. [24] are interesting in this regard as they try to tackle the limited-resolution of the data acquired from commodity 3D cameras as well. Before data fusion, a resolution enhancement step, called super-resolution (SR), is performed on data from individual views with the help of either high-resolution (HR) RGB images [17] or mono-view filtering under rigidity constraints [24, 46], to get enhanced HR 3D reconstructions.

To efficiently achieve enhanced 3D reconstructions of non-rigid objects, undergoing relatively large local motions, template based methods have been proposed in which a high quality template is built as a first step. Li et al. [38] and Zollhöfer et al. [70] propose to pre-build high quality complete templates of the target objects, which are then used to track non-rigid deformations before being fused with current measurements to produce enhanced 3D reconstructions. These methods are restricted to the class of objects which can stay static or undergo controlled rigid motions for a sufficient period of time for accurate template reconstruction.

On the other hand, methods based on different 3D non-rigid registration algorithms, using compact deformable parameterizations based on, e.g., Deformation Graphs [37, 53], Thin Plate Splines [10, 13], and skeleton extraction [58], consensus and matching under articulated motion assumptions [69], have been proposed [16, 40]. Ye et al., propose a performance capture method for complete human bodies based on skeleton fitting with three hand-held Kinect v1 cameras by making use of RGB information to aid in the registration process [65]. Li et al. [38] employ a visual hull prior, with pair-wise non-rigid scan registration based on deformation graphs [37] for hole-filling and shape completion based on relatively noise-free data.

Another class of template-free methods for complete reconstruction of 3D objects is based on spatio-temporal refinement and tracking of input data to build 4D models offline [43, 54]. Wand et al. use a topology-aware adaptive sub-space deformation technique to reduce the drift, together with as-rigid-as-possible and temporally coherent constraints on motion, to establish correspondences between acquisitions in 3D videos [60, 61]. The computed deformation field is used to construct a noise-free template from partial acquisitions. Sharf et al. relax the motion and spatial coherence constraints by using a bounded volume [52]. Their method suffers from flickering effects while still not being able to capture large deformations [38]. A recent work by Xu et al. [64] is interesting wherein a complete 3D model, and ultimately a 4D reconstruction, is iteratively built by fusing the non-rigidly deforming partial and low resolution observations and parameters of deformation subspace with the help of the Coherent Point Drift (CPD) algorithm [44]. CPD is a probabilistic non-rigid registration algorithm which is shown to handle arbitrary motions and arbitrary topologies accurately. The method of Xu et al. also has a tendency to suffer from drift due to large deformations.

Similar to Xu et al. [64], a recent body of work in this domain uses a recursive approach for temporal fusion and incremental construction of high quality 3D reference models without the need to build complete 4D reconstructions. In this vain, Dou and Fuchs, have proposed a recursive template-free scheme, using a multi-view system composed of ten Kinect v1 cameras, which tracks the motion of dynamic human subjects using deformation graphs [21]. After motion estimation, partial measurements and the reference frame are fused together using a directional distance function to produce enhanced 3D reconstructions [21, 22]. This method is restricted by the limitations of having open gesture topology for the reference frame. Moreover, the results lack quantitative analysis, and the technique has not been tested in setups with fewer cameras or with low-resolution acquisitions. DynamicFusion is a similar work which targets real-time enhancement and incremental surface completion of non-rigidly deforming objects using a mono-view system, but suffers from similar limitations as the work by Dou and Fuchs [21, 45]. VolumeDeform builds upon DynamicFusion by improving the non-rigid registration via computation of local deformations at a finer scale together with using sparse RGB features to reduce drift and improve loop closures [29].

To tackle the above mentioned challenges of recursive surface enhancement techniques, there are recent methods proposed by Afzal et al., namely KinectDeform [3] and VI-KinectDeform [1]. They are able to handle large local motions and do not require a reference model with a fixed topology. KinectDeform is a view-dependent method and hence can only produce partial reconstructions. VI-KinectDeform, on the other hand, is a view-independent moving least squares (MLS) and Kalman filter based, 3D video enhancement scheme which could directly be used in 3D multi-view systems. It has duly been tested for mono-view systems but has not been tested for and may not perform well on LR data [1].

To tackle LR and noisy non-rigidly deforming data we look into image-based SR techniques [6–8, 31–34]. It is important to mention the work of Al Ismaeil et al. in this regard which, though restricted to enhancement of mono-view dynamic depth videos, proposes to tackle the problem of LR sensing systems via a recursive dynamic multi-frame depth SR algorithm [33, 34]. This algorithm recursively estimates an HR and enhanced depth map at each time-step, by taking as input the current upsampled LR measurement and the result of previous time-step to track and correct the depth and radial displacement values of each 3D point, associated with a pixel, using a Kalman filter [35]. This method performs well on various non-rigid scenes but cannot be used for full 3D reconstructions. Moreover, due to range flow approximation this method can face difficulties to track fast and abrupt motions. Mac Aodha et al. have proposed a learning based SR approach known as *patch-based single image SR* (SISR) [6]. This approach targets resolution enhancement of LR image patches using a dictionary of noise-free synthetic HR patches. Bondi et al. [12], on the other hand, have proposed a 3D SR method which targets HR and noise-free 3D reconstruction of human faces for improved recognition. The 3D data, from LR mono-view dynamic depth videos, is accumulated via non-rigid registration based on CPD algorithm. The accumulated data is then used to estimate the 2D manifold of the face to get HR enhanced 3D reconstructions.

This overview of the state-of-art suggests that although several approaches for enhanced and complete 3D reconstructions of non-rigid objects, undergoing local motions, have been proposed, they suffer from several limitations. These limitations are due to the requirements for template generation, inability to tackle large deformations, inability to tackle highly noisy and low-resolution data, and inability to produce on-line results.

To tackle these limitations, we propose a template-free and recursive SR approach capable of handling highly noisy and low-resolution 3D data acquired via commodity depth/3D cameras. The pipeline of the proposed algorithm is shown in Figure 1. Following image-based SR approaches [32, 33], at every time-step, it upsamples the acquired measurement and uses it together with the result of previous time-step to track and correct the position and motion of each 3D point. It, therefore,

avoids error accumulation or drift caused by large deformations. Furthermore, regularization of positions and correction of motion is carried out, at each time-step, with the help of a novel 3D BTV regularization. Working directly with 3D points allows the proposed scheme to be view-independent. This enables the proposed scheme to produce high-quality full 3D reconstructions of dynamics scenes by making it generic to the number of cameras used in 3D acquisition systems. We validate the proposed approach via quantitative and qualitative analysis on simulated and real data.

3 PROBLEM FORMULATION

A 3D acquisition system captures a full 360° LR 3D video $\{\mathcal{L}_t\}$ of a scene containing non-rigidly deforming objects, with each unorganized point cloud represented as an ordered point-set $\mathcal{L}_t$, acquired at time t, and containing M 3D points, where $M \in \mathbb{N}^*$. The acquired points in $\mathcal{L}_t$ approximate the underlying surface of objects in the scene. The objective is to reconstruct an enhanced HR 3D video $\{\mathcal{H}_t\}$ where each point-set $\mathcal{H}_t = [\mathbf{p}_t^1, \cdots, \mathbf{p}_t^U]$. Each point $\mathbf{p}_t^i = (x_t^i, y_t^i, z_t^i)^\intercal$ where x_t^i, y_t^i and $z_t^i \in \mathbb{R}$, $\intercal$ is the transpose, and $i \in \{1, \cdots, U\}$. Also, $U = o \times M$, where $o \in \mathbb{N}^*$ is the factor by which the resolution of the input data is enhanced. It is also known as the SR factor.

Let us assume that each LR acquired point cloud $\mathcal{L}_t$ is related to the corresponding HR cloud $\mathcal{H}_t$ via the sensor model:

$$\mathcal{L}_t = r(\mathcal{H}_t) + \mathcal{W}_t, \tag{1}$$

where $r(.)$ is the measurement function which incorporates system blur and downsampling operators, and $\mathcal{W}_t$ represents additive white noise at time t and has same size as $\mathcal{L}_t$. We can perform dense upsampling on the acquired LR point clouds as a pre-processing step which eliminates the resolution difference between the measured data and the desired $\hat{\mathcal{H}}_t$ that we are to estimate, and helps in decreasing the registration error [31, 32]. Considering a dense upsampling operator $\uparrow$ which performs an increase or enhancement in resolution, with a factor o, (1) becomes:

$$\tilde{\mathcal{H}}_t = \mathcal{L}_t \uparrow = [r(\mathcal{H}_t)] \uparrow + \mathcal{W}_t \uparrow, \tag{2}$$

Moreover, each HR point cloud $\mathcal{H}_{t-1}$ undergoes a dynamic deformation at time t to give the HR point cloud $\mathcal{H}_t$ via:

$$\mathcal{H}_t = h_t(\mathcal{H}_{t-1}) + \mathcal{F}_t, \tag{3}$$

where $h_t(\cdot)$ is the local deformation function which deforms $\mathcal{H}_{t-1}$ to $\mathcal{H}_t$, and $\mathcal{F}_t$ is the innovation containing information about new and disappearing points [33, 34].

The objective of this paper is to devise an algorithm which recursively estimates $\mathcal{H}_t$, by taking into account the current upsampled input point cloud $\tilde{\mathcal{H}}_t$, the previous result $\hat{\mathcal{H}}_{t-1}$ and the estimated 3D non-rigid deformation relating them, such that:

$$\hat{\mathcal{H}}_t = \begin{cases} \tilde{\mathcal{H}}_t & \text{for } t = 0, \\ filt(\hat{\mathcal{H}}_{t-1}, \tilde{\mathcal{H}}_t) & t > 0, \end{cases} \tag{4}$$

where $filt(\cdot, \cdot)$ is a filtering function which mitigates the effects of cameras' measurement limitations which result in noisy measurements with limited resolution and system blur.

4 PROPOSED APPROACH

4.1 Overview

In this section, we present a solution to the problem formulated in Section 3 by proposing a view-independent recursive dynamic multi-frame 3D SR algorithm. Figure 1 gives an overview of this algorithm. After upsampling the acquired LR point cloud $\mathcal{L}_t$ to get $\tilde{\mathcal{H}}_t$, using (2), we estimate the non-rigid deformations which register the enhanced HR result of previous iteration $\hat{\mathcal{H}}_{t-1}$ with

$\tilde{\mathcal{H}}_t$. This registration is used to establish point-to-point correspondences between $\tilde{\mathcal{H}}_t$ and $\hat{\mathcal{H}}_{t-1}$, which allows to track and filter the position and motion of each point in $\tilde{\mathcal{H}}_t$. For this purpose, we use the CPD algorithm [44] which is a probabilistic method, wherein the matching of two point clouds is considered a probability density estimation problem [44]. The CPD algorithm non-rigidly registers $\hat{\mathcal{H}}_{t-1}$ to $\tilde{\mathcal{H}}_t$, which is followed by a nearest neighbor search for establishing point-to-point correspondences. For per-point refinement via tracking, in this work, we use a Kalman filter [35], which performs prediction and correction for each 3D point's motion and position using the point-to-point correspondence information. This results in a noise-free but blurred estimate of $\mathcal{H}_t$ [51]. We use a novel 3D BTV regularization to perform deblurring and produce a noise-free HR estimate $\hat{\mathcal{H}}_t$. After that a motion correction step using updated point positions in $\hat{\mathcal{H}}_t$ is also carried out. These steps are repeated for every measurement $\mathcal{L}_t$. This results in a recursive filtering process, as formulated in (4), which enhances the resolution and quality of $\mathcal{L}_t$ using the previous result.

In what follows, we describe the method for per-point tracking using the correspondence information provided by the non-rigid registration algorithm. After that we describe the proposed 3D BTV regularization based deblurring and correction method.

4.2 Per-point Refinement via Tracking

For simplification of notation, in what follows we remove the point indices i, i.e., $\mathbf{r}_t^i \equiv \mathbf{r}_t$, $\forall \mathbf{r}_t^i \in \mathbb{R}^3$. We assume that the non-rigid registration step, in Figure 1, establishes point-to-point correspondences between the points $\tilde{\mathbf{p}}_t$ and $\hat{\mathbf{p}}_{t-1}$. Now the measurement model for each point follows from (2) such that:

$$\tilde{\mathbf{p}}_t = \mathbf{p}_t + \mathbf{n}_t, \tag{5}$$

where $\mathbf{n}_t = (n_{(x,t)}, n_{(y,t)}, n_{(z,t)})^\intercal$ represents per coordinate independent Gaussian noise which affects each measured point $\tilde{\mathbf{p}}_t$ such that $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}_3, \mathbf{C})$ is a 3-dimensional noise vector where $\mathbf{0}_3$ is a 3D null vector, and $\mathbf{C} = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}$ is the covariance matrix. The per-point dynamic model follows from (3) such that:

$$\mathbf{p}_t = \mathbf{p}_{t-1} + \mathbf{w}_t, \tag{6}$$

where $\mathbf{w}_t$ is the noisy version of the innovation. We propose to treat each 3D point $\mathbf{p}_t$ in motion as an independent dynamic system decorrelated from other 3D points in the scene. The state $\mathbf{s}_t$ of this dynamic system is defined by the position $\mathbf{p}_t = (x_t, y_t, z_t)^\intercal$ and the velocity $\mathbf{v}_t = (v_{(x,t)}, v_{(y,t)}, v_{(z,t)})^\intercal$ of the corresponding 3D point such that $\mathbf{s}_t = (x_t, v_{(x,t)}, y_t, v_{(y,t)}, z_t, v_{(z,t)})^\intercal$. We propose to use the per point correspondence together with the measurement and dynamic models, and their corresponding measurement and motion uncertainties, to update and filter the system state using a Kalman filter [35].

Following from (5), the measurement model for state $\mathbf{s}_t$ is defined as:

$$\tilde{\mathbf{p}}_t = \mathbf{B}.\mathbf{s}_t + \mathbf{n}_t, \text{where } \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{7}$$

In this work we assume a constant velocity model, where the acceleration $\mathbf{a}_t$ of the point $\mathbf{p}_t$ is a random vector such that $\mathbf{a}_t \sim \mathcal{N}(\mathbf{0}_3, \mathbf{C}_\mathbf{a})$ where $\mathbf{C}_\mathbf{a} = \begin{pmatrix} \sigma_{ax}^2 & 0 & 0 \\ 0 & \sigma_{ay}^2 & 0 \\ 0 & 0 & \sigma_{az}^2 \end{pmatrix}$. Considering a time step Δt the dynamic model in (6) can be written as:

$$\mathbf{p}_t = \mathbf{p}_{t-1} + \mathbf{v}_{t-1}\Delta t + \frac{1}{2}\mathbf{a}_t \Delta t^2, \tag{8}$$

and the corresponding velocity is:

$$\mathbf{v}_t = \mathbf{v}_{t-1} + \mathbf{a}_t \Delta t, \tag{9}$$

which can, in turn, be written in the following matrix form:

$$\mathbf{s}_t = \mathbf{D}\mathbf{s}_t + \boldsymbol{\alpha}_t, \text{such that } \mathbf{D} = \begin{pmatrix} \mathbf{D}_x & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{D}_y & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{D}_z \end{pmatrix}, \tag{10}$$

where $\mathbf{D}_x = \mathbf{D}_y = \mathbf{D}_z = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}$. Moreover, $\boldsymbol{\alpha}_t$ represents the process error, such that $\boldsymbol{\alpha}_t \sim \mathcal{N}(\mathbf{0}_6, \mathbf{Q})$ where $\mathbf{0}_6$ is a 6 dimensional null vector and $\mathbf{Q} = \begin{pmatrix} \sigma^2_{a_x} A & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \sigma^2_{a_y} A & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \sigma^2_{a_z} A \end{pmatrix}$, where $\mathbf{A} = \Delta t^2 \begin{pmatrix} \Delta t^2/4 & \Delta t/2 \\ \Delta t/2 & 1 \end{pmatrix}$. Now using the standard Kalman equations, the prediction of the next state is given as:

$$\begin{cases} \hat{\mathbf{s}}_{t|t-1} = \mathbf{D}\mathbf{s}_{t-1|t-1}, \\ \hat{\mathbf{P}}_{t|t-1} = \mathbf{D}\mathbf{P}_{t-1|t-1}\mathbf{D}^\intercal + \mathbf{Q}, \end{cases} \tag{11}$$

where $\mathbf{P}_{t-1|t-1}$ is the covariance matrix corresponding to the previous state $\mathbf{s}_{t-1|t-1}$ and $\hat{\mathbf{P}}_{t|t-1}$ is the covariance matrix corresponding to the predicted state $\hat{\mathbf{s}}_{t|t-1}$. The error in the predicted state $\hat{\mathbf{s}}_{t|t-1}$ is corrected by comparing it with the observed measurement $\tilde{\mathbf{p}}_t$ based on the Kalman gain matrix $\mathbf{G}_{t|t}$ which is computed as follows:

$$\mathbf{G}_{t|t} = \hat{\mathbf{P}}_{t|t-1}\mathbf{B}^\intercal \left(\mathbf{B}\hat{\mathbf{P}}_{t|t-1}\mathbf{B}^\intercal + \mathbf{C}\right)^{-1}, \tag{12}$$

using this gain $\mathbf{G}_{t|t}$, the corrected state vector and covariance matrix are obtained via:

$$\begin{cases} \mathbf{s}_{t|t} = \hat{\mathbf{s}}_{t|t-1} + \mathbf{G}_{t|t}(\tilde{\mathbf{p}}_t - \mathbf{B}\hat{\mathbf{s}}_{t|t-1}), \\ \mathbf{P}_{t|t} = \hat{\mathbf{P}}_{t|t-1} - \mathbf{G}_{t|t}\mathbf{B}\hat{\mathbf{P}}_{t|t-1}. \end{cases} \tag{13}$$

This per-point filtering is performed for each $\tilde{\mathbf{p}}_t$ to obtain the filtered, but blurred, estimate of $\mathcal{H}_t$, i.e., $\hat{\mathcal{H}}^f_t$. Similarly, we get the filtered 3D velocity estimates for all points i.e., $\hat{\mathcal{V}}^f_t$, where $\hat{\mathcal{V}}^f_t$ contains U velocity vectors. The proposed method depends on an accurate non-rigid registration step to establish point-to-point correspondences for tracking and refinement. To handle incorrect correspondences we use a threshold distance parameter τ which allows for resetting the tracking of points [34]. For each correspondence between points $\tilde{\mathbf{p}}_t$ and $\hat{\mathbf{p}}_{t-1}$, if $\|\tilde{\mathbf{p}}_t - \hat{\mathbf{p}}_{t-1}\| > \tau$ then the state (and corresponding covariance) of $\hat{\mathbf{p}}_{t-1}$ is reset and tracking starts afresh. A filtered/accurate state for such a point is recovered after continuous tracking for a few frames. In the same vein, it is also interesting to discuss robustness of the proposed method in the face of changing or unstable camera views during a sequence. If there is no traumatic view change, the method should work fine as long as the points are being correctly registered/tracked via the non-rigid registration algorithm as discussed above. A traumatic view change, on the other hand, would mean loss of previous information in the worst case, as tracking for all, or most of, the, points would be lost.

It is to be noted that since the measurement noise and the process noise affect each coordinate of the 3D point independently, the per point Kalman filtering can be split into per coordinate Kalman filtering. This decreases the complexity of computation of the Kalman gain matrix $\mathbf{G}_{t|t}$ for each point. In the case of commodity depth cameras, the 3D point measurements suffer from depth dependant measurement noise instead of per-coordinate independent noise as discussed above [9, 41]. The details of depth dependant measurement noise model together with its affects on the noise covariance matrix $\mathbf{C}$ are discussed in the Appendix A.1.

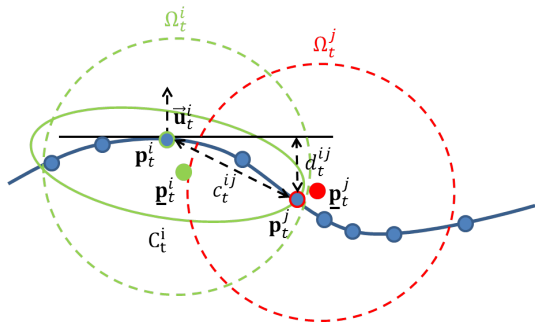


Fig. 2. Illustration of main components for per point gradient computation on a 2D surface. $\mathbf{p}_t^i$ is the query point and $\mathbf{p}_t^j$ lies in its neighborhood. Their corresponding neighborhoods are represented by Ω_t^i and Ω_t^j and the local point patches corresponding to these neighborhoods are classified by the mean and covariance of points in them i.e., $(\underline{\mathbf{p}}_t^i, \mathbf{C}_t^i)$ and $(\underline{\mathbf{p}}_t^j, \mathbf{C}_t^j)$, respectively. c_t^{ij} is the Euclidean distance between $\mathbf{p}_t^i$ and $\mathbf{p}_t^j$ and d_t^{ij} is the shortest distance of $\mathbf{p}_t^j$ to the plane, tangent at $\mathbf{p}_t^i$ to the local patch of $\mathbf{p}_t^i$, defined via the normal vector $\vec{\mathbf{u}}_t^i$.

4.3 Proposed 3D BTV Deblurring

Per-point refinement via tracking discussed in Section 4.2, although allows for view-independence but, does not explicitly cater for blurring in the measurement model in (2) [51]. Furthermore, blurring artifacts are introduced due to treating each point separately which affects the global smoothness property of point clouds [34]. This results in filtered but blurred estimates of 3D point positions in $\mathcal{H}_t$, i.e., $\hat{\mathcal{H}}_t^f$, together with the corresponding velocity estimates, i.e., $\hat{\mathcal{V}}_t^f$. Therefore after per-point tracking, at every time-step, it is necessary to carry out deblurring and regularization of position and motion estimates at hand to produce deblurred and globally smooth estimates [34]. We carry out the 3D BTV regularization of position estimates via the following minimization framework:

$$\hat{\mathcal{H}}_t = \arg\min_{\mathcal{H}_t} \mu|\nabla\mathcal{H}_t| + \frac{1}{2}\|\mathcal{H}_t - \hat{\mathcal{H}}_t^f\|_2^2, \tag{14}$$

which defines an L_2-optimization with an L_1-BTV regularization $|\nabla\mathcal{H}_t|$. $\nabla\mathcal{H}_t$ represents the discrete gradient of $\mathcal{H}_t$, $|.|$ denotes the L1-norm and μ is the regularization parameter. BTV regularization/denoising has been a topic of interest for researchers but most of the research has been restricted to organized color and depth images [28, 33, 36, 39, 42], where the neighborhoods are well defined and the gradient, based on intensity or depth values, is easy to compute e.g., via shift operators [34, 51]. In the current problem, $\hat{\mathcal{H}}_t^f$ is a set of unorganized 3D points without any connectivity or neighborhood information, therefore the extension of BTV regularization to 3D point clouds is not a straightforward problem. We are interested in finding a gradient operator ∇, which computes gradient per 3D point by taking into account the properties of the underlying surface in its local neighborhood. Therefore, we choose ∇ such that it exploits the properties of local point patches based on their unique locations, geometry and curvature, as illustrated in Fig. 2,

to formulate the 3D BTV regularization such that:

$$|\nabla\mathcal{H}_t| = \sum_{i,j} \|\nabla \mathbf{p}_t^{ij}\| = \sum_{\mathbf{p}_t^i, \mathbf{p}_t^j \in \Omega_t^i} \frac{w_{(t,c)}^{ij} w_{(t,d)}^{ij} \|((\mathbf{p}_t^i - \underline{\mathbf{p}}_t^i) - (\mathbf{p}_t^j - \underline{\mathbf{p}}_t^j))\|}{\omega_t^i}, \tag{15}$$

where Ω_t^i is the pre-computed neighborhood of $\mathbf{p}_t^i$ and $\omega_t^i = \sum_{\mathbf{p}_t^j \in \Omega_t^i} w_{(t,c)}^{ij} w_{(t,d)}^{ij}$. Each local patch corresponding to the neighborhood Ω_t^i of the query point $\mathbf{p}_t^i$ is characterized by the mean and covariance i.e., $(\underline{\mathbf{p}}_t^i, \mathbf{C}_t^i)$, of the points in it. Similarly the patch corresponding to $\mathbf{p}_t^j$ is characterized by $(\underline{\mathbf{p}}_t^j, \mathbf{C}_t^j)$. We assume equal distribution of points in $\hat{H}_t^f$, therefore we have $\mathbf{C}_t^i = \mathbf{C}_t^j$. The sizes of pre-computed niehgborhoods depend on the noise level in the data. The objective is to have a neighborhood large enough which can be used to compute the properties of the underlying patch as accurately as possible. On the other hand larger neighborhoods result in increase in computation complexity.

Now we localize $\mathbf{p}_t^i$ and $\mathbf{p}_t^j$ by subtracting from them the corresponding means and then find the difference between their local positions. This difference is then weighted by two factors $w_{(t,c)}^{ij}$ and $w_{(t,d)}^{ij}$, where $w_{(t,c)}^{ij}$ is defined as:

$$w_{(t,c)}^{ij} = \exp(-(c_t^{ij})^2/2\sigma_c^2), \tag{16}$$

where $c_t^{ij} = \|\mathbf{p}_t^i - \mathbf{p}_t^j\|$ is the Euclidean distance between $\mathbf{p}_t^i$ and $\mathbf{p}_t^j$, and σ_c^2 is a constant thresholding factor. The weight $w_{(t,c)}^{ij}$ serves to give more importance to points which lie closer to $\mathbf{p}_t^i$. On the other hand, $w_{(t,d)}^{ij}$ is defined as:

$$w_{(t,d)}^{ij} = \exp(-(d_t^{ij})^2/2\sigma_h^2), \tag{17}$$

where $d_t^{ij} = (\vec{\mathbf{u}}_t^i)^\intercal(\mathbf{p}_t^i - \mathbf{p}_t^j)$ is the shortest distance of $\mathbf{p}_t^j$ to the plane tangent, at $\mathbf{p}_t^i$, to the underlying surface sampled by the local patch of $\mathbf{p}_t^i$. The vector $\vec{\mathbf{u}}_t^i$ is the normal vector to the plane at $\mathbf{p}_t^i$, and σ_d^2 is a constant thresholding factor. $w_{(t,d)}^{ij}$ also serves to detect outliers and to preserve the edge information by taking into account the change in curvature in the local patch of $\mathbf{p}_t^i$.

The L_2-norm in (14) is convex and differentiable whereas the L_1-norm is convex and non-differentiable (non-smooth). Such type of problems cannot be solved by using simple gradient-decent methods [28]. Therefore we use the Foward-Backward Splitting (FBS) method (also known as proximal gradient solver), which relies on computing a *proximal* operator for the non-smooth part of the problem, which is implemented using Fast Adaptive Shrinkage/Thresholding Algorithm (FASTA) [28]. $|\nabla\mathcal{H}_t|$ is first reformulated to a simpler form which is differentiable, by defining a vector $\mathbf{r}_t^{ij} \in \mathbb{R}^3$ and using Cauchy-Swartz inequality to write [28]:

$$\max_{\|\mathbf{r}_t^{ij}\| \leq 1} \langle \mathbf{r}_t^{ij}, \nabla \mathbf{p}_t^{ij} \rangle = \|\nabla \mathbf{p}_t^{ij}\|, \tag{18}$$

where $\mathbf{r}_t^{ij}$ is assumed to be parallel to $\nabla \mathbf{p}_t^{ij}$, having a unit norm. Using this definition of $\|\nabla \mathbf{p}_t^{ij}\|$ in (14) and (15) respectively, solving (14) is equivalent to finding:

$$\max_{\|\mathbf{r}_t^{ij}\| \leq 1} \arg\min_{\mathcal{H}_t} \mu\langle \mathcal{R}_t, \nabla\mathcal{H}_t \rangle + \frac{1}{2}\|\mathcal{H}_t - \hat{\mathcal{H}}_t^f\|_2^2 \tag{19}$$

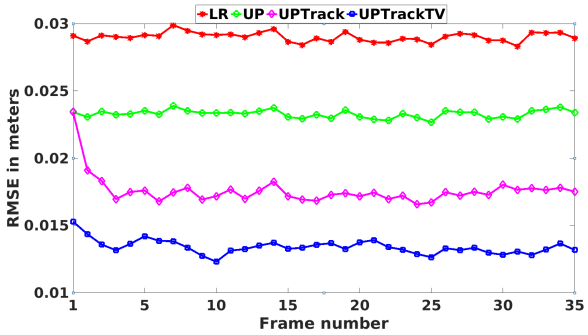


Fig. 3. Comparison of results of different steps of the proposed dynamic filtering pipeline, as shown in Figure 1, on 35 LR frames of the "Samba" dataset [58] with zero-mean Gaussian noise of standard deviation 3cm added to each coordinate of 3D points independently. The steps include dense upsampling (UP), UP with per-point tracking using a Kalman filter (UPTrack), and UP with per-point tracking and 3D BTV deblurring (UPTrackTV). Per-point tracking alone is not able to handle system blur, therefore the proposed method of per-point tracking together with 3D BTV deblurring produces the best results. The SR factor is $o = 4$.

where $\mathcal{R}_t = \{\mathbf{r}_t^{ij}\}$, and the inner minimization is now differentiable. The minimal value of $\mathcal{H}_t$ for a given value of $\mathcal{R}_t$ should satisfy $\mathcal{H}_t = \hat{\mathcal{H}}_t^f + \mu\nabla\cdot\mathcal{R}_t$ where $\nabla\cdot$ is the discrete divergence operator and can be computed by taking transpose of the gradient operator. We can reformulate (19) using the optimal value of $\mathcal{H}_t$ to get dual form of (14) such that:

$$\hat{\mathcal{R}}_t = \arg\min_{\|\mathcal{R}_t\|_\infty \leq 1} \frac{1}{2}\|\nabla\cdot\mathcal{R}_t - \frac{1}{\mu}\hat{\mathcal{H}}_t^f\|^2. \tag{20}$$

This problem is solved via the FBS method as explained in [28], and the final deblurred result at time t is obtained via:

$$\hat{\mathcal{H}}_t = \hat{\mathcal{H}}_t^f + \mu\nabla\cdot\hat{\mathcal{R}}_t. \tag{21}$$

In the case ∇ is linear it can be represented as a sparse matrix for which the corresponding discrete divergence operator can be computed by taking the transpose of this sparse matrix. This makes the solution of this problem very efficient. Therefore, for making ∇ linear we use the input $\hat{\mathcal{H}}_t^f$ to pre-compute the neighborhoods Ω_t^i and Ω_t^j, the weights $w_{(t,c)}^{ij}$ and $w_{(t,d)}^{ij}$ and the normals $\vec{\mathbf{u}}_t^i$, for all points. This method, although effective, is sensitive to parameters and can result in over-smoothing of the output. Therefore, similar to the work done in the image domain [34, 36, 39, 42], we propose to use iterative regularization with the minimization in Eq. (14) carried out multiple times, whereby in each iteration the regularization parameter μ is decreased in a dyadic way. This produces enhanced and feature preserving point clouds as shown in the results. In the next step we want to use the deblurred point cloud $\hat{\mathcal{H}}_t$ to correct the per point constant velocities estimates in $\hat{\mathcal{V}}_t^f$. For this purpose we use $\hat{\mathcal{H}}_t$ and the previous result $\hat{\mathcal{H}}_{t-1}$ to compute the per point corrected velocities estimate $\hat{\mathbf{v}}_t^i \in \mathbb{R}^3$ via:

$$\hat{\mathbf{v}}_t^i = (\hat{\mathbf{p}}_t^i - \hat{\mathbf{p}}_{t-1}^i)/\Delta t, \tag{22}$$

to get $\hat{\mathcal{V}}_t = \{\hat{\mathbf{v}}_t^i\}$. These corrected velocity estimates are then used to produce the per-point corrected state estimates which are then used in the next iteration.

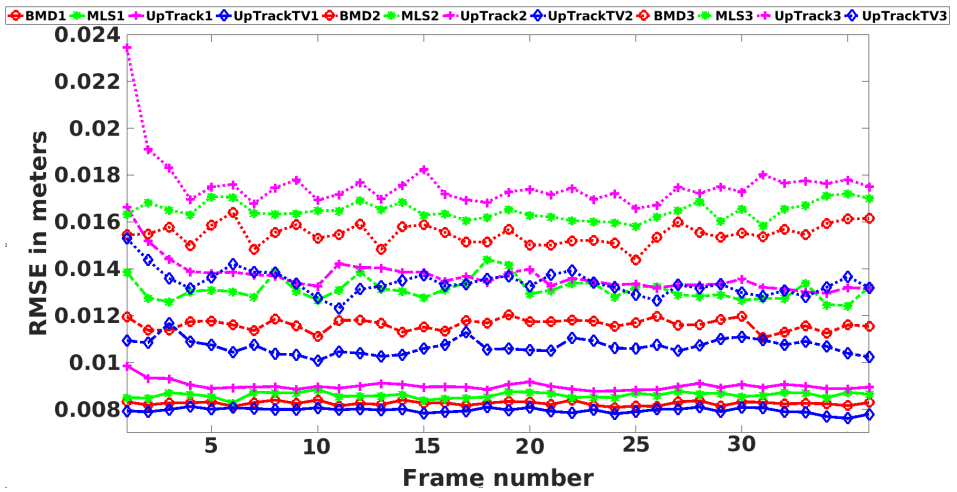


Fig. 4. Comparison of the proposed technique with the state-of-art methods for enhancement of 3D measurements, corresponding to non-rigid objects, affected by noise of varying magnitude. 35 LR frames of the "Samba" dataset [58], with zero-mean Gaussian noise of standard deviations $1cm$, $2cm$ and $3cm$ added to each coordinate of 3D points independently, are used respectively. The SR factor is $o = 4$. Two static filtering methods namely Bilateral Mesh Denoising (BMD) [25] and Moving Least Squares (MLS) [4] are compared with the proposed recursive and dynamic SR method with (UPTrackTV) and without (UPTrack) the 3D BTV deblurring. BMD1 is the result of BMD on data affected by Gaussian noise of standard deviation $1cm$, and so on. Results show that UPTrackTV provides the best performance, as compared to the other methods, across all noise levels with its comparative performance improvement increasing with increasing data noise. This is due to its ability to tackle noisy artifacts locally as well as globally, in contrast with other methods which are mainly local in nature and hence, are unable to tackle high magnitude of noise in the data.

5 EXPERIMENTS AND RESULTS

In this section we present the results of the quantitative and qualitative analysis of performance of the proposed recursive dynamic 3D SR method using both synthetic and real experimental data. The data is in the form of 3D videos and contains non-rigid objects undergoing local motions of various complexities. We start by analyzing the results of our experiments on synthetic data which includes evaluation of different steps of the proposed method and its comparison with the state-of-art methods. This is followed by an analysis of results of the proposed method using real data acquired by cameras in a multi-view system. We show the ability of the proposed 3D SR method to enhance LR and noisy 3D reconstructions of non-rigid objects undergoing local motions as well as significant topology changes.

5.1 Evaluation on Synthetic Data

In this section we analyze the performance of the proposed method, using synthetic data with available ground truth, both qualitatively and quantitatively. This performance analysis includes analyzing the affects of different steps of the proposed pipeline followed by a comparison with the state-of-art filtering methods under varying noise and SR levels.

We use the "Samba" dataset [58] which contains high quality meshes from which HR 3D point clouds are extracted. This HR data represents full 3D reconstructions of real scenes of a non-rigid human body, undergoing smooth and non-smooth local motions over time as shown in Figure. 8, which we call the ground truth (GT). We use 35 frames from this sequence for our experiments.

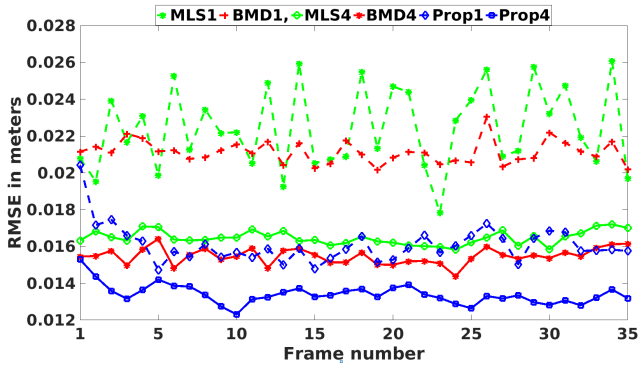


Fig. 5. Comparison of the proposed technique with the state-of-art methods for 3D point cloud enhancement for different SR factors. 35 LR frames (downsampled by a factor $o = 4$) of the "Samba" dataset [58], with zero-mean Gaussian noise of standard deviation 3cm added to each coordinate of 3D points independently, are used. The filtering is performed on the input data upsampled by a factor $o = 1$ and $o = 4$, respectively. Two static filtering methods namely BMD [25] and MLS [4] are compared with the proposed recursive and dynamic SR method. BMD1 is the result of BMD on input LR and noisy data upsampled by a factor $o = 1$, and so on. Although the proposed method has comparative performance at $o = 1$ with respect to the performance of the state-of-art methods at $o = 4$, it achieves best results at $o = 4$.

We start by analyzing the effects of different steps of the proposed SR pipeline as shown in Figure 1. For this purpose, the GT point clouds are first downsampled by a SR factor $o = 4$, then zero-mean Gaussian noise is added independently to each coordinate of 3D points, of the downsampled GT clouds, with standard deviations $\sigma_x = \sigma_y = \sigma_z = 3cm$. These LR noisy point clouds are given as input and SR results of upsampling based on mesh edge division using GT mesh information with o=4, upsampling and per-point tracking using a Kalman filter, and upsampling, per-point tracking together with multi–Level iterative 3D BTV deblurring, are obtained. Root Mean Squared Error (RMSE) for the result of each method is computed with respect to the HR GT data. Figure 3 shows the RMSE per frame for each of the steps mentioned before. Although per-point tracking using a Kalman filter recursively enhances the 3D point clouds and requires only 3-4 frames to converge, its performance is limited by its inability to handle system blur and its ability to introduce noisy artifacts. Adding a deblurring step based on 3D BTV regularization ($\sigma_c = 0.018m$, $\sigma_h = 0.0165m$) solves this problem and produces the best results.

In the next experiment, we perform a comparison of the state-of-art static 3D point cloud enhancement methods with the proposed dynamic SR scheme using the data affected by noise of varying magnitude. The GT point clouds are downsampled and upsampled by a factor $o = 4$ as explained above. Zero-mean Gaussian noise of standard deviations $\sigma_x = \sigma_y = \sigma_z = 1cm, 2cm$ and $3cm$, is added to the downsampled GT point clouds, respectively. In addition to the proposed method, we use static filtering schemes based on Bilateral Mesh Denoising (BMD) [25] and Moving Least Squares (MLS) [4] to enhance the upsampled point clouds. RMSE per frame for results of BMD, MLS, proposed method with per-point tracking only and proposed method with per-point tracking and 3D BTV deblurring, are plotted in Figure 4. Although the proposed method, with per-point tracking only, is able converge more quickly as the noise level decreases, its performance remains worse than the other methods due to introduction of blurring artifacts. The performance of BMD and MLS starts to get worse with the increase in noise magnitude due to their local nature and their inability to handle highly noisy artifacts. The proposed method with per-point tracking and 3D BTV blurring provides the best performance at all noise levels and can produce globally

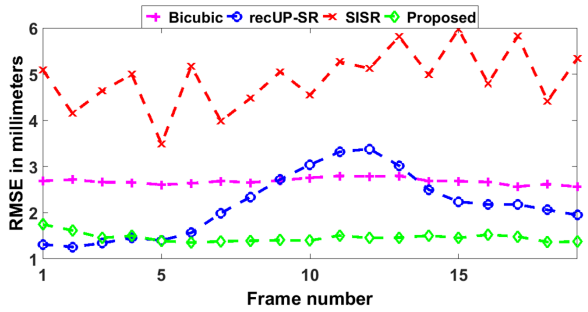


Fig. 6. Comparison of the proposed technique with the state-of-art SR methods, namely conventional bicubic interpolation, recUP-SR [34] and SISR [6], for enhancement of 3D/depth videos generated by simulating a mono-view depth system using the "Samba" dataset [58]. 19 LR depth frames with zero-mean Gaussian noise of standard deviation 3cm added to the depth measurements, are used [34]. The results show improved accuracy of the proposed method as compared to the other methods.

smooth and feature preserving point clouds even at high noise levels. The (σ_c, σ_h) values used for these experiments with Gaussian noise $\sigma_x = \sigma_y = \sigma_z = 1cm$, $2cm$ and $3cm$ are $(1.1cm, 0.65cm)$, $(1.5cm, 1.3cm)$ and $(1.6cm, 1.8cm)$, respectively.

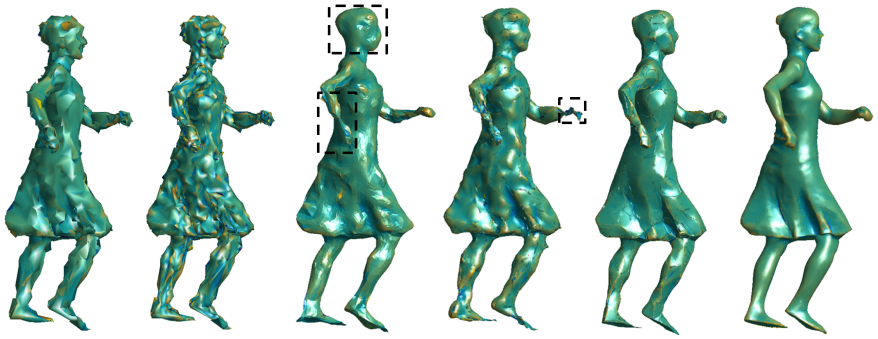
In Figure 7, we plot mesh reconstructions of an example frame (number 33) which are obtained as a result of; adding independent Gaussian noise to each coordinate of the downsampled GT data with standard deviation of $1cm$, dense upsampling of LR noisy data with $o = 4$ only, upsampling and BMD, upsampling and MLS, proposed pipeline with $o = 4$, together with HR ground truth meshes. The meshing of point clouds is carried out by using the mesh information available for GT. The results clearly show that the proposed technique produces enhanced, smoother and feature preserving reconstruction as compared to other methods. BMD and MLS fail to preserve smaller features such as hands, arm, nose, etc. To investigate further the quality of reconstructions obtained via the methods mentioned above we calculate the RMSE for different body parts for the reconstructed example Frame#33. Table 1 shows these results from which it is clear that even for separate body parts the conclusions drawn above hold.

Table 1. 3D RMSE in mm for different body parts, of Frame#33 of the "Samba" dataset [58], using different methods as shown in Figure 7.

	Arm	Leg	Torso	Full body
LR	11.31	11.61	11.03	11.48
UP	9.43	10.23	9.55	9.84
BMD	9.22	9.03	7.46	8.23
MLS	10.07	8.83	7.75	8.69
Proposed	**8.05**	**7.55**	**7.26**	**7.83**

Figure. 8 shows plots of 3D mesh reconstruction of 5 frames (#1, 3, 12, 21, 30), from the sequence, obtained as a result of the proposed method. It shows that the proposed method is able to recursively enhance the noisy input measurements while successfully tackling non-rigid smooth and non-smooth local motions.

In the next experiment, we perform a comparison of the state-of-art static 3D point cloud enhancement methods, i.e., BMD and MLS, with the proposed dynamic SR scheme for different SR



(a) LR (b) UP (c) BMD (d) MLS (e) Proposed (f) GT

Fig. 7. 3D mesh plots of a super-resolved resultant Frame#33 from the "Samba" dataset [58] after: b) dense upsampling (UP), c) Bilateral Mesh Denoising (BMD), d) Moving Least Squares (MLS) and e) Proposed recursive and dynamic SR scheme. a) is the 3D plot of LR noisy data, and e) is the GT HR mesh respectively. Proposed technique produces smooth, enhanced and feature preserving reconstruction as compared to the rest. The SR factor is $o = 4$. Display color-scale is based on mean surface curvature.

factors. This means that GT point clouds are first downsampled by a SR factor $o = 4$, then zero-mean Gaussian noise is added independently to each coordinate of 3D points, of the downsampled GT clouds, with standard deviations $\sigma_x = \sigma_y = \sigma_z = 3cm$. Filtering is carried on this data with upsampling factors of $o = 1$ and $o = 4$, respectively. RMSE per frame is plotted in Figure 5. Results show that proposed method clearly outperforms both BMD and MLS when used on same data. The results also show that even at upsampling factor $o = 1$ the proposed dynamic scheme gives comparative performance with respect to both BMD and MLS used on upsampled noisy data with $o = 4$. This is outperformed by applying the proposed dynamic filtering scheme at $o = 4$. The reason for this is that at $o = 1$, the method recursively denoises the noisy input. On the other hand, at $o = 4$, the method applies the full recursive dynamic super-resolution pipeline which together with denoising, enhances the quality of data by preserving useful features.

Lastly, we perform a comparison of the proposed dynamic 3D SR method with the state-of-art SR methods which include the conventional bicubic interpolation, the dynamic depth SR method proposed by Al Ismaeil et al. [33, 34], called recUP-SR, and the learning based SR method called SISR proposed by Mac Aodha et al. [6]. We again make use of the “Samba" dataset [58], and simulate a depth camera, placed at a distance of approx. 2 meters, in V-Rep [23] to generate a mono-view synthetic depth sequence [34]. This GT depth sequence is downsampled by a factor $o = 4$, and zero mean Gaussian noise of variance $\sigma_z = 3cm$ is added to the depth measurements. This LR noisy depth sequence is given as input to the state-of-art methods, and is converted to a 3D sequence via the known camera parameters and given as input to the proposed method. To compare the super-resolved (by a factor $o = 4$) results of all methods, the resulting depth sequences from state-of-art methods and the GT depth sequence are converted to 3D sequences as explained before. After that per frame RMSE for the result of each method with respect to the 3D GT is computed. The results are reported in the Figure 6. The results show the robustness and improved accuracy of the proposed method as compared to the state-of-art methods. SISR produces HR 3D reconstructions but its accuracy suffers due to its patch-based nature which prevents it from preserving finer details. Moreover, both SISR and bicubic interpolation suffer due to not taking into account the temporal information. recUP-SR tries to overcome this weakness by proposing a dynamic and recursive

(a) Frame#1 (b) Frame#3 (c) Frame#12 (d) Frame#21 (e) Frame#30

Fig. 8. 3D mesh plot of different frames of super-resolved resultant full 3D point clouds (Frames#1, 3, 12, 21, 30) from the "Samba" dataset [58]. Display color-scale is based on mean surface curvature.

SR scheme, but it is restricted in tracking large and abrupt motions due to working with depth images and range flow motion approximation. In contrast, the results show improved accuracy of the proposed method due to its robustness to handle non-smooth/abrupt motions, undergone by the non-rigid object, which results from its ability to accurately track the positions of points in 3D space. The (σ_c, σ_h) values for deblurring used in these experiments are $(1.125cm, 1cm)$.

5.2 Evaluation on Real Data

In this section we analyze the performance of the proposed method using real data acquired via multi-view systems composed of photometric and commodity depth cameras, respectively. In addition to showcasing the ability of the proposed method to enhance the quality of LR and noisy data to produce smooth and feature preserving full 3D reconstructions of non-rigid objects, this experimental analysis also demonstrates the capabilities of the proposed method to produce accurate and enhanced 3D reconstructions of objects with changing topologies.

In the first experiment we use full 3D point-clouds extracted from meshes of the “adult child ball" sequence from Inria’s "4D-Repository" [30]. This dataset is acquired via a fully calibrated multi-view system based on photometric RGB cameras. The sequence, used here, has two characteristics; the resolution of data is quite low (approx. 10000 points per scene) resulting in non-smooth surfaces, and it contains an object, i.e. a ball, with changing topology as shown in Figure 9. Due to these characteristics this type of dataset is very challenging for the class of methods to which belong the works by Dou and Fuchs [21, 22], and DynamicFusion [45] etc. These methods do not explicitly target LR data and are very sensitive to objects with changing topologies due to their design of always fusing the current measurement with the first frame which is considered to be the reference. The proposed method, on the other hand, explicitly targets LR 3D data and produces HR, smooth and feature preserving 3D reconstructions as shown in Figure 9. Moreover, it works by recursively fusing the current measurement and the result of the previous iteration/time-step and hence, can accurately reconstruct objects, in this case a ball, with changing topologies. The (σ_c, σ_h) values for deblurring used in these experiments are $(0.5cm, 0.3cm)$.

In the next experiment we use point clouds from the full 3D video of the “jumping in place" action performed by a human subject from the Berkeley Multimodal Human Action Database (MHAD) [47]. This dataset is acquired via a fully calibrated multi-view system composed of two Kinect v1 cameras placed at opposite corners of the acquisition space. As explained in Section A.1,

the depth acquisition system of cameras, built on structured-light principle, such as Kinect v1 suffers from depth dependent measurement noise. The distance of Kinect cameras from the subjects in MHAD's multi-view setup is approximately $3.5 - 4$ meters. This results in highly noisy 3D measurements with non-smooth surfaces and diminished features as shown in Figure 10. Figure 10 also shows the point clouds which are received as the output of the proposed algorithm. The input data is upsampled by a factor $o = 1.5$. Moreover, to tackle the depth dependent measurement noise specific to Kinect v1 cameras the measurement model presented in Section A.1 is used during the per-point tracking step. The resolution enhancement together with per-point tracking and 3D BTV deblurring results in point clouds which are relatively noise-free, have smoother surfaces with less holes/gaps and better preserved features/details. The (σ_c, σ_h) values for deblurring used in these experiments are $(1.25cm, 0.75cm)$.

Lastly, it is to be noted that we have used different upsampling operators to test the proposed approach. These include the mesh sub-division operator for datasets with available mesh information, e.g., full 3D "Samba" dataset [58] (Section 5.1 and Figure 3). We have also used the bi-linear interpolation operator for datasets based on depth cameras, e.g., "Berkley MHAD" dataset [47] (Figure 10). The proposed approach has been shown to be robust to both types of upsampling operators.

5.3 Performance & Implementation Details

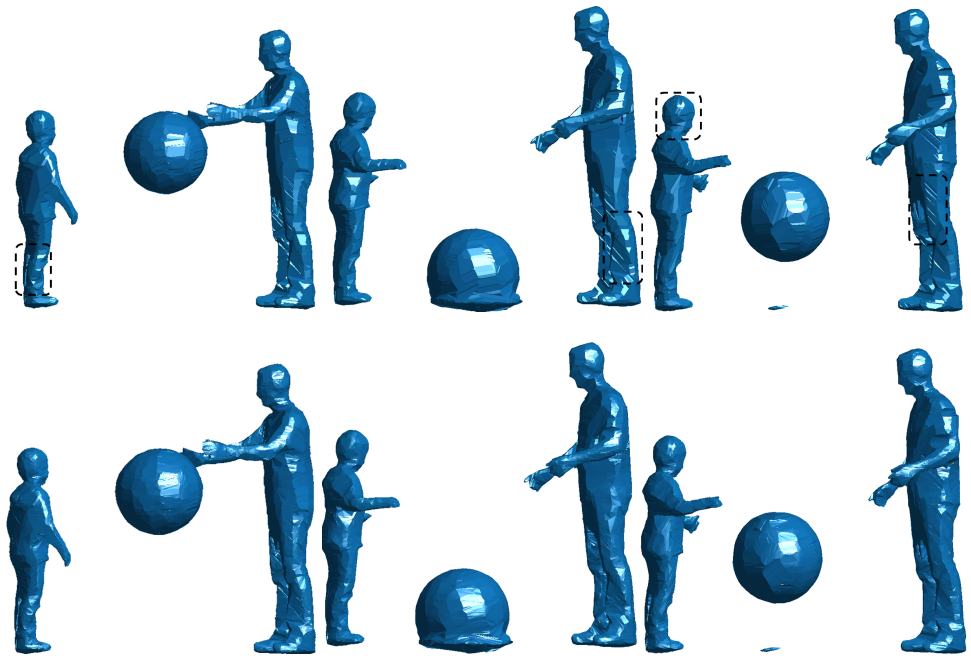
In this section we report the run-time performance of the proposed scheme on the datasets we evaluated in the previous sections. The implementation of the proposed scheme together with the experimental evaluation has been carried out in Ubuntu 14.04 on a desktop system with Intel Xeon 3.4 GHz (8 cores) processor and 8 GB of RAM.

The proposed scheme can be divided into three main steps, first is the non-rigid registration based on CPD algorithm, second is per-point tracking and refinement based on Kalman filter, and third is deblurring based on 3D BTV. For the non-rigid registration, we have used the standard CPD implementation provided by the authors [44]. Similarly, for 3D BTV deblurring we use the FASTA implementation of the FBS method provided by the authors [28]. The per-point refinement via Kalman filter has been implemented in C++.

Table 2. Computation-times (sec) for different stages of the proposed scheme for each frame of the "Samba" dataset [58], as shown in Figure 7, and the "Berkley MHAD" dataset [47], as shown in Figure 10.

	Points/frame	Registration	Tracking	Deblurring
Samba	10,000	120 sec	3.5 sec	40 sec
Berkley MHAD	54,000	1500 sec	16 sec	350 sec

We take the full 3D sequence from the "Samba" dataset [58], as shown in Figure 7, and the "Berkley MHAD" dataset [47], as shown in Figure 10 as examples. Each frame of "Samba" dataset contains approx. $10,000$ 3D points whereas each frame of the "Berkley MHAD" dataset contains approx. $54,000$ 3D points. The computation-times for the 3 stages for processing one frame of these datasets are given in Table 2. This table shows that the maximum time is required by the CPD based non-rigid registration algorithm followed by the 3D BTV deblurring and per-point tracking respectively. It is to be noted that the tracking implementation is not optimized for tackling each point independently and in parallel fashion. Therefore, we believe that we can achieve real-time performance for tracking with better implementation. Lastly, around 80% of the computation time during deblurring is used in computation of the matrices related to the gradient operator. This operation can also be highly optimized via parallelization e.g., with the help of a GPU processor.



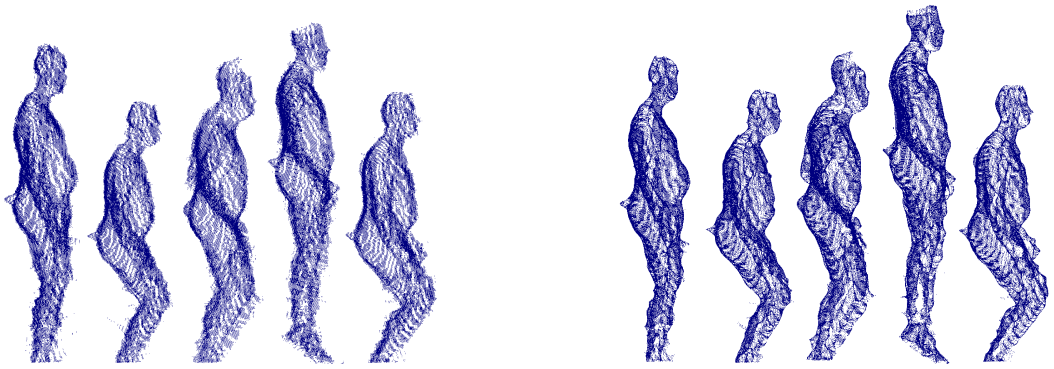
(a) Frame #5 (b) Frame #12 (c) Frame #15

Fig. 9. 3D mesh plots of three LR frames (#5, #12 and #15) from Inria's "4D-Repository" [30], i.e. the top row, and the corresponding super-resolved (using SR factor $o = 4$) results of the proposed algorithm, i.e. the bottom row. The input data has low-resolution which results in non-smooth surfaces, thick edges and loss of details. The results show super-resolved, smooth, and feature preserving 3D reconstructions of non-rigid objects. They also show ability of the proposed method to produce enhanced reconstructions of objects with changing topologies e.g., the ball in the above plots. Display color-scale is based on mean surface curvature.

6 CONCLUSION & FUTURE WORK

In this paper we have presented a framework for acquiring high quality and full 360° 3D reconstructions of dynamic scenes containing non-rigid objects undergoing large local motions/deformations. We target noisy and LR data acquired from commodity 3D cameras in mono-view or multi-view systems. This framework is based on a view-independent recursive and dynamic multi-frame 3D SR algorithm which is capable of filtering out the noise as well as enhancing the resolution of the raw measurements obtained from multi-view systems. The proposed algorithm tracks and filters the position and motion of every 3D point recursively hence making use of complete 3D characteristics of the input data. It is able to handle generic 3D as well as structured-light sensing based depth specific noise in 3D measurements. Moreover, it uses a 3D BTV regularization for deblurring and smoothing of the point clouds after per point tracking. Quantitative and qualitative evaluation of the proposed framework shows its better performance as compared to state-of-art methods for producing noise-free and smooth full 3D reconstructions.

As future work, we would be interested in incorporating the proposed 3D SR scheme to improve the accuracy of applications based on extracting meaningful information from 3D measurements. A facial recognition system is an example of such an application which could benefit from the high quality 3D reconstructions obtained from the proposed scheme in un-constrained environments [7,



(a) LR (b) Prop.

Fig. 10. Plots of LR 3D point-clouds of five frames (#6, #10, #18, #22 and #35) from the "Berkeley MHAD" Kinect dataset [47], on the left, and the corresponding super-resolved (using SR factor $o = 1.5$) results of the proposed algorithm on the right. The input data suffers from high magnitude of noisy artifacts, in the form of non-smooth surface and jagged edges, due to large distance of the human subject from the cameras. The results show super-resolved, smooth, and feature preserving full 3D reconstructions of the human subject.

11]. We would also like to investigate the use of texture information which is available via RGB sensors in commodity RGB-D cameras [9, 41]. This information can be used to produce high quality textured 3D reconstructions [21]. Furthermore, its utility in improving the performance of the proposed 3D BTV framework can also be investigated [62]. Lastly, we would also like to investigate the use of a curvature operator instead of a gradient operator in the proposed 3D BTV framework [48].

A APPENDIX

A.1 Depth Dependent Measurement Noise

The measurement model in (5) assumes per coordinate independent Gaussian noise affecting each 3D point $\mathbf{p}_t$. In reality the 3D points are computed from depth images acquired via commodity 3D cameras built on structured-light or time-of-flight principles [9, 41, 49]. The acquired per-point depth measurement, i.e., $\tilde{\mathbf{q}}_t = (\tilde{u}_t, \tilde{v}_t, \tilde{z}_t)^\intercal$ is defined by the approximated pixel position $(\tilde{u}_t, \tilde{v}_t)$, in the depth image, and the measured depth value $\tilde{z}_t$ such that $\tilde{\mathbf{q}}_t = \mathbf{q}_t + \dot{\mathbf{n}}_t$, where $\dot{\mathbf{n}}_t = (n_{(u,t)}, n_{(v,t)}, n_{(z,t)})^\intercal$ represents noise in the measured pixel position and depth value. Let us consider a structured-light depth camera [9], for which the depth measurement $\tilde{z}_t$ suffers due to noise $n_{(d,t)}$ in disparity d, which is the distance (in pixels) between locations of a point in observed and projected pattern, via the relation $n_{(z,t)} = -\frac{z_t^2}{f.b} n_{(d,t)}$, where f is camera's horizontal focal length, b the baseline distance between the camera and the projector, and $n_{(d,t)}$ is the noise in the corresponding disparity measurement $\tilde{d}_t$ [2, 55]. The main factor affecting both the pixel and disparity measurements is the noise due to quantization [2], therefore we can assume it to be drawn from independent Gaussian distributions such that $n_{(u,t)} \sim \mathcal{N}(0, \sigma_u^2)$, $n_{(v,t)} \sim \mathcal{N}(0, \sigma_v^2)$ and $n_{(d,t)} \sim \mathcal{N}(0, \sigma_d^2)$. This allows us to model the noise in depth measurement i.e., $n_{(z,t)} \sim \mathcal{N}(0, \sigma_{(z,t)}^2)$ where $\sigma_{(z,t)}^2 = (-\frac{z_t^2}{f.b})^2 \sigma_d^2$.

To convert the depth measurement $\tilde{\mathbf{q}}_t$ to the corresponding 3D position $\tilde{\mathbf{p}}_t$, the intrinsic matrix $\mathbf{K} = \begin{pmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix}$, where (f_u, f_v) represent the focal lengths (where $f = f_u$), and (c_u, c_v) represent center of camera's imager such that:

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{Z}}_t \mathbf{K}^{-1} \tilde{\mathbf{q}}_t = \tilde{\mathbf{Z}}_t \mathbf{K}^{-1} (\mathbf{q}_t + \dot{\mathbf{n}}_t), \tag{23}$$

where $\tilde{\mathbf{Z}}_t = \begin{pmatrix} \tilde{z}_t & 0 & 0 \\ 0 & \tilde{z}_t & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\tilde{z} = z + n_{(z,t)}$. Therefore the measurement model for each 3D point can now be defined as:

$$\tilde{\mathbf{p}}_t = \mathbf{p}_t + \mathbf{n'}_t, \tag{24}$$

where:

$$\mathbf{n'}_t = \left(\frac{z_t n_{(u,t)} + (u_t - c_u) n_{(z,t)} + n_{(z,t)} n_{(u,t)}}{f_u}, \frac{z_t n_{(v,t)} + (v_t - c_v) n_{(z,t)} + n_{(z,t)} n_{(v,t)}}{f_v}, n_{(z,t)} \right)^\mathsf{T}. \tag{25}$$

Here $\mathbf{n'}_t \sim \mathcal{N}(\mathbf{0}_3, \mathbf{C}'_t)$ where the entries of covariance matrix $\mathbf{C}'_t$ are defined as:

$$\begin{cases} cov(n_{(x,t)}, n_{(x,t)}) = \left(\frac{z_t^2 \sigma_u^2 + (u_t - c_u)^2 \sigma_{(z,t)}^2 + \sigma_u^2 \sigma_{(z,t)}^2}{f_u^2} \right), cov(n_{(x,t)}, n_{(y,t)}) = \frac{(u_t - c_u)(v_t - c_v)}{f_u f_v} \sigma_{(z,t)}^2, \\ cov(n_{(y,t)}, n_{(y,t)}) = \left(\frac{z_t^2 \sigma_v^2 + (v_t - c_v)^2 \sigma_{(z,t)}^2 + \sigma_v^2 \sigma_{(z,t)}^2}{f_v^2} \right), cov(n_{(z,t)}, n_{(z,t)}) = \sigma_{(z,t)}^2, \\ cov(n_{(x,t)}, n_{(z,t)}) = \frac{(u_t - c_u)}{f_u} \sigma_{(z,t)}^2, cov(n_{(y,t)}, n_{(z,t)}) = \frac{(v_t - c_v)}{f_v} \sigma_{(z,t)}^2, \end{cases} \tag{26}$$

where $cov(.,.)$ computes the covariance between two random variables. This covariance matrix, specific to each point, can therefore be replaced in (12) when dealing with data acquired from depth cameras. To compute this covariance matrix, the noise-free pixel and depth values are required, but are not available in practice. Therefore, we propose to use the measured pixel and depth values instead, which are the closest approximation of the noise-free values we can get. Using the $\mathbf{C}'_t$ increases complexity of the proposed approach as now we have to deploy a Kalman filter per point, instead of per coordinate which was the case previously, but it captures the noise characteristics of depth cameras more accurately.

ACKNOWLEDGMENTS

The work was supported by the National Research Fund, Luxembourg under the CORE project: C11/BM/1204105/FAVE/Ottersten.

REFERENCES

[1] Hassan Afzal, Djamila Aouada, François Destelle, Bruno Mirbach, and Björn Ottersten. 2015. View-Independent Enhanced 3D Reconstruction of Non-rigidly Deforming Objects. In *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II*. Springer International Publishing, 712–724. https://doi.org/10.1007/978-3-319-23117-4_61

[2] H. Afzal, D. Aouada, D. Fofi, B. Mirbach, and B. Ottersten. 2014. RGB-D Multi-view System Calibration for Full 3D Scene Reconstruction. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. 2459–2464. https://doi.org/10.1109/ICPR.2014.425

[3] H. Afzal, K. Al Ismaeil, D. Aouada, F. Destelle, B. Mirbach, and B. Ottersten. 2014. KinectDeform: Enhanced 3D Reconstruction of Non-Rigidly Deforming Objects. In *The 3DV Workshop on Dynamic Shape Measurement and Analysis*.

[4] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and Claudio T. Silva. 2003. Computing and rendering point set surfaces. *Visualization and Computer Graphics, IEEE Transactions on* 9, 1 (Jan 2003), 3–15. https://doi.org/10.1109/TVCG.2003.1175093

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph.* 24, 3 (July 2005), 408–416. https://doi.org/10.1145/1073204.1073207

[6] Oisin Mac Aodha, Neill D. F. Campbell, Arun Nair, and Gabriel J. Brostow. 2012. Patch Based Synthesis for Single Depth Image Super-resolution. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III (ECCV'12)*. Springer-Verlag, Berlin, Heidelberg, 71–84. https://doi.org/10.1007/978-3-642-33712-3_6

[7] Djamila Aouada, Kassem Al Ismaeil, Kedija Kedir Idris, and Björn E. Ottersten. 2014. Surface UP-SR for an improved face recognition using low resolution depth cameras. In *11th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2014, Seoul, South Korea, August 26-29, 2014*. 107–112. https://doi.org/10.1109/AVSS.2014.6918652

[8] Djamila Aouada, Kassem Al Ismaeil, and Björn E. Ottersten. 2015. Patch-based Statistical Performance Analysis of Upsampling for Precise Super-Resolution. In *VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 1, Berlin, Germany, 11-14 March, 2015*. 186–193. https://doi.org/10.5220/0005316001860193

[9] Asus. 2010. Xtion PRO LIVE. (2010). https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/

[10] Ilya Baran and Jovan Popović. 2007. Automatic Rigging and Animation of 3D Characters. *ACM Trans. Graph.* 26, 3, Article 72 (July 2007). https://doi.org/10.1145/1276377.1276467

[11] Stefano Berretti, Pietro Pala, and Alberto Del Bimbo. 2014. Face Recognition by Super-Resolved 3D Models From Consumer Depth Cameras. *IEEE Trans. Information Forensics and Security* 9, 9 (2014), 1436–1449. https://doi.org/10.1109/TIFS.2014.2337258

[12] Enrico Bondi, Pietro Pala, Stefano Berretti, and Alberto Del Bimbo. 2016. Reconstructing High-Resolution Face Models From Kinect Depth Sequences. *IEEE Trans. Information Forensics and Security* 11, 12 (2016), 2843–2853.

[13] F. L. Bookstein. 1989. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 6 (June 1989), 567–585. https://doi.org/10.1109/34.24792

[14] Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Probabilistic Deformable Surface Tracking From Multiple Videos. In *ECCV 2010 - 11th European Conference on Computer Vision*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.), Vol. 6314. Springer, Heraklion, Greece, 326–339. https://doi.org/10.1007/978-3-642-15561-1_24

[15] Rodrigo L. Carceroni and Kiriakos N. Kutulakos. 2002. Multi-View Scene Capture by Surfel Sampling: From Video Streams to Non-Rigid 3D Motion, Shape and Reflectance. *International Journal of Computer Vision* 49, 2-3 (2002), 175–214. https://doi.org/10.1023/A:1020145606604

[16] Will Chang and Matthias Zwicker. 2009. Range Scan Registration Using Reduced Deformable Models. *Comput. Graph. Forum* 28, 2 (2009), 447–456. http://dblp.uni-trier.de/db/journals/cgf/cgf28.html#ChangZ09

[17] Yan Cui, Will Chang, Tobias Nöll, and Didier Stricker. 2013. KinectAvatar: Fully Automatic Body Capture Using a Single Kinect. In *Proceedings of the 11th International Conference on Computer Vision - Volume 2 (ACCV'12)*. Springer-Verlag, Berlin, Heidelberg, 133–147. https://doi.org/10.1007/978-3-642-37484-5_12

[18] Y. Cui, S. Schuon, S. Thrun, D. Stricker, and C. Theobalt. 2013. Algorithms for 3D Shape Scanning with a Depth Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 5 (May 2013), 1039–1050. https://doi.org/10.1109/TPAMI.2012.190

[19] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. 2017. BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. *ACM Transactions on Graphics 2017 (TOG)* (2017).

[20] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance Capture from Sparse Multi-view Video. *ACM Trans. Graph.* 27, 3, Article 98 (Aug. 2008), 10 pages. https://doi.org/10.1145/1360612.1360697

[21] Mingsong Dou and H. Fuchs. 2014. Temporally enhanced 3D capture of room-sized dynamic scenes with commodity depth cameras. In *Virtual Reality (VR), 2014 iEEE*. 39–44. https://doi.org/10.1109/VR.2014.6802048

[22] Mingsong Dou, Henry Fuchs, and Jan Michael Frahm. 2013. *Scanning and tracking dynamic objects with commodity depth cameras*. 99–106. https://doi.org/10.1109/ISMAR.2013.6671769

[23] M. Freese E. Rohmer, S. P. N. Singh. 2013. V-REP: a Versatile and Scalable Robot Simulation Framework. In *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*.

[24] Andrew Feng, Ari Shapiro, Wang Ruizhe, Mark Bolas, Gerard Medioni, and Evan Suma. 2014. Rapid Avatar Capture and Simulation Using Commodity Depth Sensors. In *ACM SIGGRAPH 2014 Talks (SIGGRAPH '14)*. ACM, New York, NY, USA, Article 16, 1 pages. https://doi.org/10.1145/2614106.2614182

[25] Shachar Fleishman, Iddo Drori, and Daniel Cohen-Or. 2003. Bilateral Mesh Denoising. In *ACM SIGGRAPH 2003 Papers (SIGGRAPH '03)*. ACM, New York, NY, USA, 950–953.

[26] Yasutaka Furukawa and Jean Ponce. 2010. Dense 3D Motion Capture from Synchronized Video Streams. In *Image and Geometry Processing for 3-D Cinematography*. 193–211. https://doi.org/10.1007/978-3-642-12392-4_9

[27] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. 2013. Dense Variational Reconstruction of Non-rigid Surfaces from Monocular Video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. 1272–1279. https://doi.org/10.1109/CVPR.2013.168

[28] Tom Goldstein, Christoph Studer, and Richard G. Baraniuk. 2014. A Field Guide to Forward-Backward Splitting with a FASTA Implementation. *CoRR* abs/1411.3406 (2014). http://arxiv.org/abs/1411.3406

[29] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. *VolumeDeform: Real-Time Volumetric Non-rigid Reconstruction.* Springer International Publishing, Cham, 362–379. https://doi.org/10.1007/978-3-319-46484-8_22

[30] Inria. 2009. 4D-Repository. http://4drepository.inrialpes.fr/pages/home. (2009).

[31] K. A. Ismaeil, D. Aouada, B. Mirbach, and B. Ottersten. 2013. Dynamic super resolution of depth sequences with non-rigid motions. In *2013 IEEE International Conference on Image Processing*. 660–664. https://doi.org/10.1109/ICIP.2013.6738136

[32] Kassem Al Ismaeil, Djamila Aouada, Bruno Mirbach, and Björn Ottersten. 2016. Enhancement of dynamic depth scenes by upsampling for precise super-resolution (UP-SR). *Computer Vision and Image Understanding* (2016). https://doi.org/10.1016/j.cviu.2016.04.006

[33] K. Al Ismaeil, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten. 2015. Real-time non-rigid multi-frame depth video super-resolution. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 8–16. https://doi.org/10.1109/CVPRW.2015.7301389

[34] K. Al Ismaeil, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten. 2016. Real-Time Enhancement of Dynamic Depth Videos with Non-Rigid Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).

[35] Rudolph Emil Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering* 82, Series D (1960), 35–45.

[36] A. Kheradmand and P. Milanfar. 2014. A General Framework for Regularized, Similarity-Based Image Restoration. *IEEE Transactions on Image Processing* 23, 12 (Dec 2014), 5136–5151. https://doi.org/10.1109/TIP.2014.2362059

[37] Hao Li, Robert W. Sumner, and Mark Pauly. 2008. Global Correspondence Optimization for Non-Rigid Registration of Depth Scans. *Computer Graphics Forum (Proc. SGP'08)* 27, 5 (July 2008).

[38] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T. Barron, and Gleb Gusev. 2013. 3D Self-Portraits. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2013)* 32, 6 (November 2013).

[39] Wenshu Li, Chao Zhao, Qiegen Liu, Qingjiang Shi, and Shen Xu. 2012. A parameter-adaptive iterative regularization model for image denoising. *EURASIP Journal on Advances in Signal Processing* 2012, 1 (2012), 1–10. https://doi.org/10.1186/1687-6180-2012-222

[40] C. Malleson, M. Klaudiny, A. Hilton, and J. Y. Guillemaut. 2013. Single-View RGBD-Based Reconstruction of Dynamic Human Geometry. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. 307–314. https://doi.org/10.1109/ICCVW.2013.48

[41] Microsoft. 2014. Kinect. (2014). https://dev.windows.com/en-us/kinect//

[42] P. Milanfar. 2013. A Tour of Modern Image Filtering: New Insights and Methods, Both Practical and Theoretical. *IEEE Signal Processing Magazine* 30, 1 (Jan 2013), 106–128. https://doi.org/10.1109/MSP.2011.2179329

[43] Niloy J. Mitra, Simon Flory, Maks Ovsjanikov, Natasha Gelfand, Leonidas Guibas, and Helmut Pottmann. 2007. Dynamic Geometry Registration. In *Symposium on Geometry Processing*. 173–182.

[44] A. Myronenko and X. Song. 2010. Point Set Registration: Coherent Point Drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 12 (Dec 2010), 2262–2275. https://doi.org/10.1109/TPAMI.2010.46

[45] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[46] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time Dense Surface Mapping and Tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR '11)*. IEEE Computer Society, Washington, DC, USA, 127–136. https://doi.org/10.1109/ISMAR.2011.6092378

[47] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. 2013. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. 53–60. https://doi.org/10.1109/WACV.2013.6474999

[48] C. Olsson and Y. Boykov. 2012. Curvature-based regularization for surface approximation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1576–1583. https://doi.org/10.1109/CVPR.2012.6247849

[49] PMDTechnologies. 2012. Camboard Nano. (2012). http://www.pmdtec.com

[50] Henry Roth and Marsette Vona. 2012. Moving Volume KinectFusion. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 112.1–112.11. https://doi.org/10.5244/C.26.112

[51] Sebastian Schuon, Christian Theobalt, James Davis, and Sebastian Thrun. 2009. LidarBoost: Depth Superresolution for ToF 3D Shape Scanning. *In Proc. of IEEE CVPR 2009* (2009).

[52] Andrei Sharf, Dan A. Alcantara, Thomas Lewiner, Chen Greif, Alla Sheffer, Nina Amenta, and Daniel Cohen-Or. 2008. Space-time Surface Reconstruction Using Incompressible Flow. *ACM Trans. Graph.* 27, 5, Article 110 (Dec. 2008), 10 pages. https://doi.org/10.1145/1409060.1409063

[53] Robert W. Sumner, Johannes Schmid, and Mark Pauly. 2007. Embedded Deformation for Shape Manipulation. *ACM Trans. Graph.* 26, 3, Article 80 (July 2007). https://doi.org/10.1145/1276377.1276478

[54] Jochen Sussmuth, Marco Winter, and Gunther Greiner. 2008. Reconstructing Animated Meshes from Time-varying Point Clouds. In *Proceedings of the Symposium on Geometry Processing (SGP '08)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 1469–1476. http://dl.acm.org/citation.cfm?id=1731309.1731332

[55] Richard Szeliski. 2010. *Computer Vision: Algorithms and Applications* (1st ed.). Springer-Verlag New York, Inc., New York, NY, USA.

[56] Christian Theobalt, Naveed Ahmed, Hendrik Lensch, Marcus Magnor, and Hans-Peter Seidel. 2007. Seeing People in Different Light-Joint Shape, Motion, and Reflectance Capture. *IEEE Transactions on Visualization and Computer Graphics* 13, 4 (2007), 663–674. https://doi.org/10.1109/TVCG.2007.1006

[57] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. 2012. Scanning 3D Full Human Bodies Using Kinects. *Visualization and Computer Graphics, IEEE Transactions on* 18, 4 (April 2012), 643–650. https://doi.org/10.1109/TVCG.2012.56

[58] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated Mesh Animation from Multi-view Silhouettes. *ACM Trans. Graph.* 27, 3, Article 97 (Aug. 2008), 9 pages. https://doi.org/10.1145/1360612.1360696

[59] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic Shape Capture using Multi-View Photometric Stereo. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 28, 5 (Dec. 2009).

[60] Michael Wand, Bart Adams, Maksim Ovsjanikov, Alexander Berner, Martin Bokeloh, Philipp Jenke, Leonidas Guibas, Hans-Peter Seidel, and Andreas Schilling. 2009. Efficient Reconstruction of Nonrigid Shape and Motion from Real-time 3D Scanner Data. *ACM Trans. Graph.* 28, 2, Article 15 (May 2009), 15 pages. https://doi.org/10.1145/1516522.1516526

[61] Michael Wand, Philipp Jenke, Qixing Huang, Martin Bokeloh, Leonidas Guibas, and Andreas Schilling. 2007. Reconstruction of Deforming Geometry from Time-varying Point Clouds. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing (SGP '07)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 49–58. http://dl.acm.org/citation.cfm?id=1281991.1281998

[62] Oliver Wasenmüller, Gabriele Bleser, and Didier Stricker. 2015. Combined Bilateral Filter for Enhanced Real-time Upsampling of Depth Images. In *VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 1, Berlin, Germany, 11-14 March, 2015*. 5–12. https://doi.org/10.5220/0005234800050012

[63] A. Weiss, D. Hirshberg, and M. J. Black. 2011. Home 3D body scans from noisy image and range data. In *2011 International Conference on Computer Vision*. 1951–1958. https://doi.org/10.1109/ICCV.2011.6126465

[64] Weipeng Xu, Mathieu Salzmann, Yongtian Wang, and Yue Liu. 2015. Deformable 3D Fusion: From Partial Dynamic 3D Observations to Complete 4D Models. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, Washington, DC, USA, 2183–2191. https://doi.org/10.1109/ICCV.2015.252

[65] Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. 2012. *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Performance Capture of Interacting Characters with Handheld Kinects, 828–841. https://doi.org/10.1007/978-3-642-33709-3_59

[66] Rui Yu, Chris Russell, Neill D. F. Campbell, and Lourdes Agapito. 2015. Direct, Dense, and Deformable: Template-Based Non-rigid 3D Reconstruction from RGB Video. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 918–926. https://doi.org/10.1109/ICCV.2015.111

[67] Ming Zeng, Jiaxiang Zheng, Xuan Cheng, and Xinguo Liu. 2013. Templateless Quasi-rigid Shape Modeling with Implicit Loop-Closure. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[68] Q. Zhang, B. Fu, M. Ye, and R. Yang. 2014. Quality Dynamic Human Body Modeling Using a Single Low-Cost Depth Camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 676–683. https://doi.org/10.1109/CVPR.2014.92

[69] Qian Zheng, Andrei Sharf, Andrea Tagliasacchi, Baoquan Chen, Hao Zhang, Alla Sheffer, and Daniel Cohen-Or. 2010. Consensus Skeleton for Non-Rigid Space-Time Registration. *Computer Graphcis Forum (Special Issue of Eurographics)* 29, 2 (2010), 635–644.

[70] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. 2014. Real-time Non-rigid Reconstruction using an RGB-D Camera. *ACM Transactions on Graphics (TOG)* 33, 4 (2014).