Real-Time Enhancement of Dynamic Depth Videos with Non-Rigid Deformations

Kassem Al Ismaeil, *Student Member, IEEE*, Djamila Aouada, *Member, IEEE*,
Thomas Solignac, Bruno Mirbach, and Björn Ottersten, *Fellow, IEEE*

Abstract—We propose a novel approach for enhancing depth videos containing non-rigidly deforming objects. Depth sensors are capable of capturing depth maps in real-time but suffer from high noise levels and low spatial resolutions. While solutions for reconstructing 3D details in static scenes, or scenes with rigid global motions have been recently proposed, handling unconstrained non-rigid deformations in relative complex scenes remains a challenge. Our solution consists in a recursive dynamic multi-frame super-resolution algorithm where the relative local 3D motions between consecutive frames are directly accounted for. We rely on the assumption that these 3D motions can be decoupled into lateral motions and radial displacements. This allows to perform a simple local per-pixel tracking where both depth measurements and deformations are dynamically optimized. The geometric smoothness is subsequently added using a multi-level L_1 minimization with a bilateral total variation regularization. The performance of this method is thoroughly evaluated on both real and synthetic data. As compared to alternative approaches, the results show a clear improvement in reconstruction accuracy and in robustness to noise, to relative large non-rigid deformations, and to topological changes. Moreover, the proposed approach, implemented on a CPU, is shown to be computationally efficient and working in real-time.

Index Terms—Depth enhancement, super-resolution, non-rigid deformations, registration, Kalman filtering, bilateral total variation

1 Introduction

Sensing using 3D technologies, structured light cameras or time-of-flight (ToF) cameras, has seen a revolution in the past years where sensors such as the Microsoft Kinect version 1 and 2 are today part of accessible consumer electronics [1]. The ability of these sensors in directly capturing depth videos in real-time has opened tremendous possibilities for applications in gaming, robotics, surveillance, health care, etc. These sensors, unfortunately, have major shortcomings due to their high noise contamination, including missing and jagged measurements, and their low spatial resolutions. This makes it impossible to capture detailed 3D features indispensable for many 3D computer vision algorithms. The face data in Fig. 1a is an example of such challenging raw depth measurements. Running a traditional face recognition algorithm on this type of data would result in a very low recognition rate [2], [3], [4].

Some solutions have been proposed in the literature for recovering these details but mostly in the context of static 3D scene scanning, with *LidarBoost* [5] and its extension [6] and *KinectFusion* [7] being the most known methods. The current major challenge is when the object or objects in the scene are subject to non-rigid deformations. Indeed, *LidarBoost* and *KinectFusion* are rigid depth fusion approaches, and they immediately fail in providing any reasonable result on non-rigidly deforming scenes, the focus of this paper.

In [8], [9], [10], we proposed the *UP-SR* algorithm, which stands for *Upsampling for Precise Super-Resolution*, as the first dynamic multi-frame depth video super-resolution (SR) algorithm that can enhance depth videos containing non-rigidly deforming scenes without any prior assumption on the number of moving objects they contain or on the topology of these objects. These advantages were possible thanks to a direct processing on depth maps without using connectivity information inherent to meshing as used in subsequent methods, namely, *KinectDeform* [11] and *DynamicFusion* [12]. The *UP-SR* algorithm is, however, limited to lateral motions as it only computes 2D dense optical flow but does not account for the full motion in 3D, known as scene flow, or the 2.5D motion, known as range flow. It consequently fails in the case of radial deformations. Moreover, it is not practical because of a heavy cumulative motion estimation process applied to a number of frames buffered in the memory.

This paper presents a solution that improves over the *UP-SR* algorithm by keeping its advantages and solving its two limitations–not considering 3D motions and using an inefficient cumulative motion estimation. The proposed solution is based on the assumption that the 3D motion of a point can be approximated by decoupling the radial component from the lateral ones. This approximation allows the handling of non-rigid deformations while reducing the computational complexity associated with an explicit full 3D motion estimation at each point. Moreover, a recursive depth multi-frame SR is formulated by replacing *UP-SR*'s cumulative motion estimation with a point-tracking

- K. Al Ismaeil, D. Aouada, and B. Ottersten are with Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, Esch-sur-Alzette L-2721, Luxembourg.
E-mail: {kassem.alismaeil, djamila.aouada, bjorn.ottersten}@uni.lu.
- T. Solignac and B. Mirbach are with the Advanced Engineering Department, IEE S.A, Contern L-5326, Luxembourg.
E-mail: {thomas.solignac, bruno.mirbach}@iee.lu.

Manuscript received 6 Aug. 2015; revised 21 Aug. 2016; accepted 30 Sept. 2016. Date of publication 26 Oct. 2016; date of current version 12 Sept. 2017.
Recommended for acceptance by T. Brox.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2622698

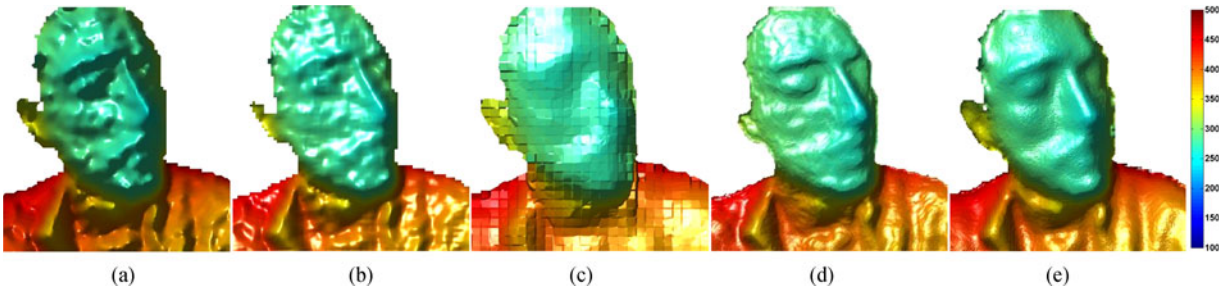

Fig. 1. Results of different SR methods for a scale factor of $r = 4$ applied to a low resolution dynamic depth video captured with a ToF camera at a rate of 50 frames per ms. (a) Raw frame, (b) Bicubic interpolation, (c) *SISR* [31], (d) *UP-SR* [8], (e) Proposed *recUP-SR*. Units are in mm.

operation locally at each pixel. Similarly to earlier approaches for a recursive SR [13], [14], [15], [16], we use a Kalman filter for tracking except that we treat each pixel separately as opposed to considering the full image. As a result, the proposed solution efficiently runs multiple Kalman filters in parallel on local depth values and on their radial displacements. A subsequent processing is required in order to recover the smoothness property of a depth map and correct the artifacts caused by this per-pixel filtering. To that end, we propose a multi-level version of the L_1 minimization with a bilateral total variation (BTV) regularization originally given in [17]. The proposed algorithm leads to a new approach for estimating range flow by pixel tracking with a Kalman filter. It is important to note that while this flow contributes in handling non-rigid deformations in 3D, the effectiveness of the proposed SR algorithm comes from how this flow is employed in the *UP-SR* reconstruction framework. Indeed, merely applying the estimated range flow in another depth SR method does not give satisfactory results. An overview of the proposed algorithm, named *recUP-SR*, is given in Fig. 2. A first visual illustration of the performance of the proposed algorithm on a low quality depth video of a highly non-rigidly deforming face is given in Fig. 1e. In summary, the contribution of this paper is a new mutli-frame depth SR algorithm which has the following properties: 1) Accuracy in depth video reconstruction. 2) Robustness to non-rigid deformations and to noise. 3) Robustness to topological changes. 4) Independence of the number of moving objects in the scene. 5) Real-time implementation on a CPU. This paper is an extended version of [18] with additional theoretical details and clarifications explaining the transition from the earlier *UP-SR* algorithm to the *recUP-SR* algorithm proposed herein, and an extended explanation of the proposed multi-level iterative deblurring. A significantly extended experimental part is also provided containing original analyses based on new results.

The remainder of the paper is organized as follows: Section 2 gives an overview of the different categories of depth enhancement approaches and a review of range flow algorithms. Section 3 formulates the problem of depth video SR and describes the necessary background on *UP-SR*. The proposed recursive depth video SR algorithm is presented in Section 4. Experimental evaluations are presented and discussed in Section 5. Finally, the conclusion is given in Section 6.

2 Related Work

Three categories of approaches for the enhancement of depth videos containing non-rigid deformations may be distinguished as detailed below.

2.1 Multi-Modal Fusion Approaches

This category of methods is based on the assumption that there is a correspondence between the edges of a depth map and the edges of another modality of a better quality, often chosen to be a 2D intensity image of the same scene [19], [20], [21], [22], [23], [24], [26]. Such an image is considered to be a guidance image whose properties, in terms of structure, signal to noise ratio (SNR) and resolution, can be

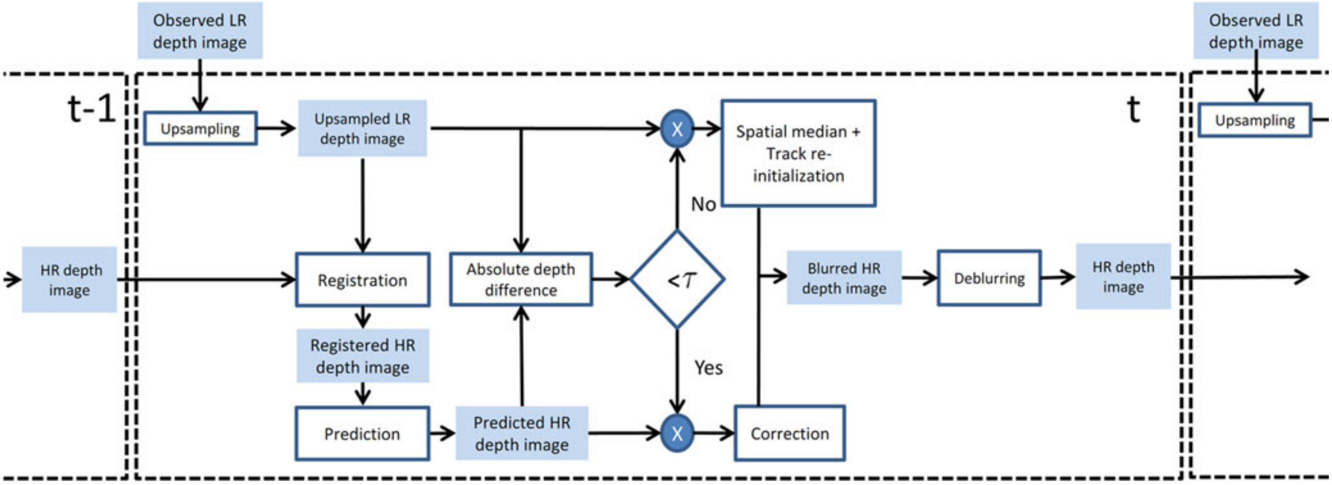

Fig. 2. Flow chart of *recUP-SR*: A new multi-frame depth super-resolution algorithm for dynamic depth videos of non-rigidly deforming objects.

transferred to its corresponding depth map in order to obtain an enhanced version. Diebel and Thrun proposed the first work in this category [19]. Their approach consists in a multiresolutional Markov Random Field (MRF) defined to integrate both modalities, depth and 2D intensity, and used to estimate a depth map of the same resolution as the resolution of the intensity image. In 2007, Kopf et al. [20] and Yang et al. [21] proposed the concept of joint bilateral upsampling (JBU), which is an adapted version of the bilateral filter [27] for fusing a low resolution (LR) depth map with a high resolution (HR) 2D image. In [22], [26], the JBU multi-modal filter was extended to considering not only a 2D image as a guidance image but also the depth map itself. These weighting-based filters may be viewed as implicit guided regularization approaches. In an explicit guided regularization approach, a regularization function that translates the prior properties of a guidance image is added to a data fidelity term to form a cost function to be optimized as in [28], [30]. These more sophisticated approaches are as of today the most effective ones in the category of multi-modal fusion. Another variant in this category are methods using passive stereo imaging in combination with low quality active depth sensors [23]. However, besides the fact that all approaches in this category require an additional camera, they are also highly dependent on the assumption of correspondence between 2D intensity and depth images. This requires a perfect calibration and synchronization of a multi-camera system and a perfect data mapping.

2.2 Learning-Based Approaches

The first work in this category is by Mac Aodha et al. known as *patch-based single image SR* (SISR) [31]. This method, as implied by its name, follows a patch-based approach where an LR patch is reconstructed from a large dictionary of synthetic noise-free HR depth patches. The boundaries of overlapping patches follow a special treatment to keep smoothness properties. An important contribution of [31] is the large database of HR depth maps used to create the dictionary. Hornacek et al. proposed in [32] an adaptation of [33] to depth maps. The idea there is to run a search for similar patches only using the repetitive information already available in the depth map to be enhanced. To compute the similarity between depth patches, the *PatchMatch* algorithm by Barnes et al. [34] was redefined in 3D under invariance to 3D rigid motions. In [35], a patchwork assembly algorithm for depth single image SR was proposed merging the concept of using a training database [31] and the concept of self-similarity [32]. The problem was formulated as an MRF optimization which could optionally add information from an HR 2D intensity image following the same principle of multi-modal fusion approaches. Another work at the frontier of the two categories, learning-based and multi-modal fusion approaches, was proposed earlier by Li et al. in [36]. The authors proposed to learn a mapping function between HR 2D patches and their corresponding LR depth patches using a training database. A sparse coding algorithm was then used for the final enhancement. The performance of the methods in this category rely on the quality of the training database and/or on the repetitive structure of the input depth map. In general, these conditions are not always available or verified.

2.3 Dynamic Multi-Frame Approaches

Using multiple frames to recover depth details has been successful in the case of static scenes or scenes with global rigid motion [5], [6], [7]. Since these methods and their immediate derivatives, the real challenge that the research community has been facing is extending the multi-frame depth enhancement concept to scenes with non-rigid deformations. There have been few attempts to handle single object scanning under relative small non-rigidities by replacing a global rigid registration with a non-rigid alignment [37], [38], [39]. These techniques, however, cannot handle large deformations, and are not very practical for real-time applications. Real-time non-rigid reconstruction approaches have been achieved with the help of a template which is first acquired then used for tracking of non-rigidities with a good flexibility [40], [41]. Recently, we have proposed *KinectDeform* [11], the first non-rigid version of *KinectFusion*. It does not require any template, and similarly to *KinectFusion*, provides an enhanced smoother reconstruction over time with the addition of handling non-rigid deformations in the scene. *KinectDeform* has been successfully tested on an Asus Xtion Pro Live camera [42], equivalent to Microsoft Kinect structured light version 1. It cannot, however, perform well on lower resolution, noisier ToF cameras such as the PMD camboard nano [43]. Indeed, its registration module requires denser raw acquisitions. *DynamicFusion* is another recent non-rigid version of *KinectFusion*. Thanks to a GPU implementation, it has been tested on a Kinect camera in real-time. However, its reconstruction accuracy has not been evaluated, and it has only been validated visually. Moreover, it builds on the assumption of having only one moving object in the scene. In addition, its reported limitations are its sensitivity to complex scenes and scenes with changes in topology. Also, similarly to *KinectDeform*, one may suspect *DynamicFusion* not to be able to perform well on a lower resolution noisier ToF camera.

The *UP-SR* algorithm falls under this category of dynamic multi-frame approaches. Indeed, it exploits the deformations collected over time in an inverse SR reconstruction framework. *UP-SR* was shown in [8] and [10] to have a higher reconstruction accuracy as compared to representative methods from the first and second categories of depth enhancement methods. Its major drawbacks, however, are its limitation to lateral motions and its computational complexity. Using range flow in place of optical flow is a natural first step towards reconstructing non-lateral, i.e., radial, local deformations. In what follows, we review related literature on range flow estimation.

2.4 Range Flow Estimation

In order to handle 3D non-rigid deformations, it is important to consider the full 3D motion per pixel or the range flow. The range flow constraint has been first proposed in [45], and later appeared in other references [46], [47], [48], [49]. This constraint is usually used in a variational framework to estimate the range flow. However, estimating a dense range flow, i.e., a three dimensional vector for each point is still computationally complex and not achievable in real-time, at least, not with a sub-pixel accuracy [50]. Using RGB-D depth cameras has allowed a multi-modal approach for range flow estimation by defining a global energy functional combining the range flow and the 2D optical flow constraints. In addition to the challenge of reducing the computational complexity, these algorithms have to handle erroneous measurements of depth sensors such as flying pixels, and missing and invalid values. A common approach is to define a smoothness condition to be used as a regularization term in the global optimization [29], [49], [51]. In [52] and [53], a probabilistic approach is followed by using a particle filter. This concept is the closest to our range flow estimation where information is recursively propagated to the next frame. Recently, the *aTGV-SF* algorithm has been proposed where the flow is directly calculated in 3D by back-projection, resulting in a joint estimation of the lateral and radial motions [29]. To our knoweledge, this is reported to be currently the best performing range flow algorithm in terms of accuracy and runtime. We will use it as a reference in our evaluation.

As explained in Section 1, the current paper improves *UP-SR*. In Section 3, the necessary background on *UP-SR* is given after formalizing the problem of multi-frame SR.

3 Background and Problem Formulation

The following notation will be adopted: matrices are denoted by boldface uppercase letters $\mathbf{A}$, and vectors or column images are denoted by boldface lowercase letters $\mathbf{a}$. Scalars are denoted by italic letters, a, A. $\hat{\mathbf{a}}$: estimate of $\mathbf{a}$. $\tilde{\mathbf{a}}$: measured $\mathbf{a}$. $\mathbf{a}^{\mathtt{i}}$: element $\mathtt{i}$ of $\mathbf{a}$. $\mathbf{a}_t$: $\mathbf{a}$ at time t. $\overline{\mathbf{a}}_{t_1}^{t_0}$: registered $\mathbf{a}$ from t_1 to t_0. $\mathbf{a}\uparrow$: upsampled $\mathbf{a}$.

3.1 Multi-Frame Super-Resolution

Let us consider an LR video $\{\mathbf{g}_t\}$ acquired with a depth sensor. The captured scene is assumed to be dynamically and non-rigidly deforming without any assumption on the number of moving objects. Each LR observation $\mathbf{g}_t$ is represented by a column vector of length m corresponding to the lexicographic[1] ordering of frame pixels. The objective of depth SR is to reconstruct an HR depth video $\{\mathbf{f}_t\}$ using $\{\mathbf{g}_t\}$, where each frame $\mathbf{f}_t$ is of length n with $n = r^2 \times m$ such that $r \in \mathbb{N}^*$ is the SR scale factor. In the classical multi-frame depth SR problem, in order to reconstruct a given frame $\mathbf{f}_{t_0} \in \{\mathbf{f}_t\}$, also known as the reference frame, the N preceding observed LR frames are used.

An LR observation $\mathbf{g}_t$ is related to the reference frame through the following data model

$$\mathbf{g}_t = \mathbf{DHM}_{t_0}^{t}\mathbf{f}_{t_0} + \mathbf{n}_t, \qquad t_0 \geq t, \tag{1}$$

where $\mathbf{D}$ is a known constant downsampling matrix of dimension $(m \times n)$. The system blur is represented by the time and space invariant matrix $\mathbf{H}$. The $(n \times n)$ matrices $\mathbf{M}_{t_0}^{t}$ correspond to the motion between $\mathbf{f}_{t_0}$ and $\mathbf{g}_t$ before downsampling. The vector $\mathbf{n}_t$ is an additive white noise at time instant t. Without loss of generality, both $\mathbf{H}$ and $\mathbf{M}_{t_0}^{t}$ are assumed to be block circulant commutative matrices. As a result, the estimation of $\mathbf{f}_{t_0}$ may be decomposed into two steps; estimation of a blurred HR image $\mathbf{z}_{t_0} = \mathbf{Hf}_{t_0}$, followed by a deblurring step to recover $\hat{\mathbf{f}}_{t_0}$.

The above framework has been first proposed in the case of static 2D scenes in [17] and for static depth scenes in [44]. In [8] and in [10] it has been extended to dynamic depth scenes defining the *UP-SR* algorithm.

3.2 Upsampling for Precise Super-Resolution (*UP-SR*)

The *UP-SR* algorithm starts by a dense upsampling of all the LR observations. This is shown to ensure a more accurate registration of frames. The resulting r^2-times upsampled image is defined as $\mathbf{g}_t\uparrow = \mathbf{U}\cdot\mathbf{g}_t$, where $\mathbf{U}$ is the $(n \times m)$ dense upsampling matrix. It is chosen to be the transpose of the downsampling matrix $\mathbf{D}$. As a result, the product $\mathbf{UD} = \mathbf{A}$ gives a block circulant matrix $\mathbf{A}$ that defines a new blurring matrix $\mathbf{B} = \mathbf{AH}$. Therefore, the estimation of $\mathbf{f}_{t_0}$ goes through the estimation of its new blurred version $\mathbf{z}_{t_0} = \mathbf{Bf}_{t_0}$.

In order to estimate $\mathbf{z}_{t_0}$, the N frames preceding $\mathbf{g}_{t_0}$ are required. Every two consecutive frames are related by the following dynamic model

$$\mathbf{g}_{t+1}\uparrow = \mathbf{M}_t^{t+1}\mathbf{g}_t\uparrow + \boldsymbol{\delta}_{t+1}, \tag{2}$$

where $\mathbf{M}_t^{t+1}$ is the motion between them. The vector $\boldsymbol{\delta}_{t+1}$ is referred to as the *innovation* image. It contains novel measurements appearing, or disappearing due to occlusions or large motions. Note that, in the *UP-SR* framework, the *innovation* is assumed to be negligible, and the matrix $\mathbf{M}_t^{t+1}$ is assumed to be an invertible permutation. It can be estimated using classical dense 2D optical flow. The elements of the estimated $\hat{\mathbf{M}}_t^{t+1}$ matrix are 1's and 0's, basically indicating the address of the source pixels in $\mathbf{g}_t\uparrow$ and the address of the destination pixel in $\mathbf{g}_{t+1}\uparrow$. This information is equivalent to finding for each pixel position $\mathrm{p}_t^{\mathtt{i}} = (x_t^{\mathtt{i}}, y_t^{\mathtt{i}})$, $\mathtt{i} = 1, \ldots, n$, the horizontal and vertical displacements in pixels $u_t^{\mathtt{i}}$ and $v_t^{\mathtt{i}}$, respectively. In the continuous case, these displacements correspond to the lateral motions $(u_t^{\mathtt{i}}, v_t^{\mathtt{i}})$ where $u_t^{\mathtt{i}} = \frac{dx_t^{\mathtt{i}}}{dt}$ and $v_t^{\mathtt{i}} = \frac{dy_t^{\mathtt{i}}}{dt}$.

The small motions between consecutive frames are then cumulated, and a frame $\mathbf{g}_t\uparrow$ is registered to the reference frame $\mathbf{g}_{t_0}\uparrow$ with the following cumulative motion compensation approach

$$\overline{\mathbf{g}}_t^{t_0}\uparrow = \hat{\mathbf{M}}_t^{t_0}\mathbf{g}_t\uparrow = \underbrace{\hat{\mathbf{M}}_{t_0-1}^{t_0}\cdots\hat{\mathbf{M}}_t^{t+1}}_{(t_0-t)\ \text{times}}\cdot\mathbf{g}_t\uparrow. \tag{3}$$

As a result, the original data model in (1) is simplified to define the following *UP-SR* data model

$$\overline{\mathbf{g}}_t^{t_0}\uparrow = \mathbf{z}_{t_0} + \boldsymbol{\nu}_t, \qquad t_0 \geq t, \tag{4}$$

where $\boldsymbol{\nu}_t = \hat{\mathbf{M}}_t^{t_0}\mathbf{U}\cdot\mathbf{n}_t$ is an additive noise vector of length n. It is assumed to be independent and identically distributed. Using an L_1-norm, the blurred estimate $\hat{\mathbf{z}}_{t_0}$ is found by pixel-wise temporal median filtering of the upsampled registered LR observations $\{\overline{\mathbf{g}}_t^{t_0}\uparrow\}$. As a second and final step, follows an image deblurring to estimate $\hat{\mathbf{f}}_{t_0}$ from $\hat{\mathbf{z}}_{t_0}$.

The only considered motions in the *UP-SR* algorithm are lateral ones using 2D dense optical flow. Radial

[1] Lexicographic ordering: Concatenation of the columns of the image.

displacements in the depth direction, often encountered in depth sequences, are therefore not handled. In order to address this problem, we propose to consider range flow in the *UP-SR* framework.

3.3 Range Flow Approximation

A time-varying depth surface $\mathcal{Z}$ may be viewed as a mapping of a pixel position $\mathrm{p}_t^{\mathtt{i}} = (x_t^{\mathtt{i}}, y_t^{\mathtt{i}})$ on the sensor image plane, at a time instant t, such that $\mathrm{p}_t^{\mathtt{i}} \mapsto \mathcal{Z}(x_t^{\mathtt{i}}, y_t^{\mathtt{i}})$. The value $\mathcal{Z}(x_t^{\mathtt{i}}, y_t^{\mathtt{i}})$ corresponds to the ith element of the depth image $\mathbf{z}_t$ written in lexicographic vector form, that we will denote in what follows as $\mathbf{z}_t^{\mathtt{i}}$. The deformation of the surface $\mathcal{Z}$ from $(t-1)$ to t takes the point $\mathrm{p}_{t-1}^{\mathtt{i}}$ to a new position $\mathrm{p}_t^{\mathtt{i}}$. It may be expressed through the derivative of $\mathcal{Z}$ following the direction of the 3D displacement resulting in a range flow $(u_t^{\mathtt{i}}, v_t^{\mathtt{i}}, w_t^{\mathtt{i}})$ where the radial displacement in the depth direction $w_t^{\mathtt{i}} = \frac{d\mathbf{z}_t^{\mathtt{i}}}{dt}$ is added as the third component to the lateral displacement. In this work, we propose to decouple the estimation of the lateral motions $(u_t^{\mathtt{i}}, v_t^{\mathtt{i}})$ from the estimation of the radial displacement $w_t^{\mathtt{i}}$. Indeed, depth cameras provide an intensity image $\mathbf{a}_t$ (2D image in the case of RGB-D sensors, or an amplitude image in the case of ToF sensors), which makes it possible to estimate the lateral motions directly using the 2D optical flow constraint between two consecutive intensity images $\mathbf{a}_{t-1}$ and $\mathbf{a}_t$. This decoupling approach enables to reduce the complexity, but also to introduce a probabilistic framework that allows to recursively estimate $w_t^{\mathtt{i}}$ and the corrected depth value at the same point. Once $(u_t^{\mathtt{i}}, v_t^{\mathtt{i}})$ is estimated, we proceed with estimating $w_t^{\mathtt{i}}$ under a probabilistic framework where we account for radial motion uncertainties.

4 Proposed Approach

The proposed depth video enhancement approach is based on an extension of the *UP-SR* algorithm. As our goal is a real-time processing, the major difference resides in replacing the cumulation of N frames in *UP-SR* for processing a reference frame at time t_0, by a recursive processing that only considers two consecutive frames at $(t-1)$ and t where the current frame is to be enhanced each time. The measurement model for each current frame may be defined by setting $t_0 = t$ in (4), resulting in

$$\tilde{\mathbf{z}}_t^{\mathtt{i}} := [\mathbf{g}_t \uparrow]^{\mathtt{i}} = \mathbf{z}_t^{\mathtt{i}} + [\mathbf{n}_t \uparrow]^{\mathtt{i}} \qquad \forall t, \tag{5}$$

where $[\mathbf{n}_t \uparrow]^{\mathtt{i}}$ is assumed to be zero mean Gaussian with the variance σ_n^2, i.e., $[\mathbf{n}_t \uparrow]^{\mathtt{i}} \sim \mathcal{N}(0, \sigma_n^2)$. The problem at hand is then to estimate $\mathbf{z}_t^{\mathtt{i}}$ given a noisy measurement $\tilde{\mathbf{z}}_t^{\mathtt{i}}$ and an enhanced noise-free depth value $\mathbf{z}_{t-1}^{\mathtt{i}}$ estimated at the preceding iteration. The time-deforming depth scene is viewed as a dynamic system where the state of each pixel is defined by its depth value and radial displacement. These states are estimated dynamically over time using a Kalman filter. The *UP-SR* dynamic model in (2) is directly used to characterize the dynamic system and introduce the uncertainties of depth measurements and radial deformations in one probabilistic framework. The proposed recursive approach, that we refer to as *recUP-SR*, is summarized in the flow chart of Fig. 2. The main steps are described in what follows.

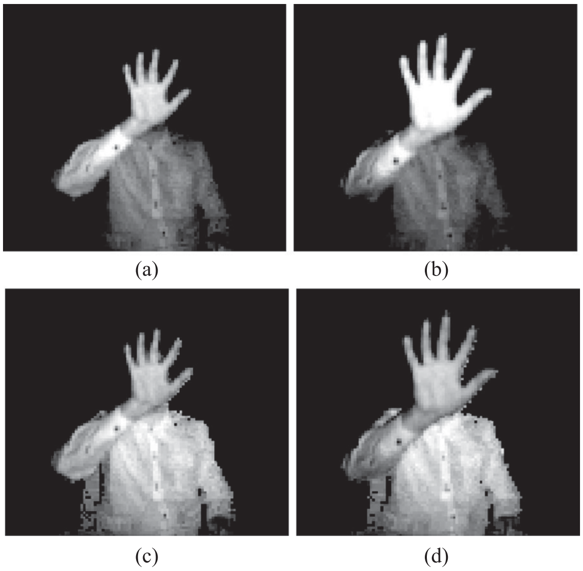

Fig. 3. Correcting amplitude images using a standardization step [54]. (a) and (b) show the original LR amplitude images for a dynamic scene containing a hand moving towards the camera where the intensity (amplitude) values differ significantly depending on the object distance from the camera. The corrected amplitude images for the same scene are presented in (c) and (d), where the intensity consistency is preserved.

4.1 Lateral Registration

In order to be able to carry a per-pixel processing, essential for handling non-rigid deformations, one needs to properly align these pixels between consecutive frames. This is achieved by registration through 2D dense optical flow that estimates the lateral motion between the intensity images $\mathbf{a}_{t-1}$ and $\mathbf{a}_t$. In the case of RGB-D cameras, these images are provided directly. Mapping and synchronization have to be ensured, though, as in [49] and [51]. In the case of ToF cameras, the provided intensity images, known as amplitude images, can not be used directly. Their intensity values differ significantly depending on the camera integration time and on the distance of the scene from the camera; hence, not verifying the optical flow assumption of brightness consistency. Thus, in order to guarantee an accurate registration, it is necessary to apply a standardization step similar to the one proposed in [54] prior to motion estimation, see Fig. 3. If intensity images are not available, for example when using synthetic data, the 2D optical flow can be directly estimated using LR raw depth images, but after a denoising step (e.g., using a bilateral filter). We note that this denoising should only be used in the preprocessing step. The original raw depth data is the one to be mapped in the registration step. In all cases, as for *UP-SR*, we register the upsampled versions of the LR images after upscaling the motion vectors estimated from the LR images. We define the registered depth image from $(t-1)$ to t as $\bar{\mathbf{z}}_{t-1}^{t}$. Consequently, the radial displacement $w_t^{\mathtt{i}}$ may be initialized by the temporal difference between depth measurements, i.e.,

$$w_t^{\mathtt{i}} \approx \tilde{\mathbf{z}}_t^{\mathtt{i}} - [\bar{\mathbf{z}}_{t-1}^{t}]^{\mathtt{i}}. \tag{6}$$

This first approximation of $w_t^{\mathtt{i}}$ is an initial value that requires further refinement directly accounting for the system noise. We propose to do that using a per-pixel tracking with a Kalman filter as detailed in Section 4.2.

4.2 Refinement by Per-Pixel Tracking

According to the definition of image pixel registration, we have $\mathbf{z}_{t-1}^{\mathtt{i}} := [\bar{\mathbf{z}}_{t-1}^{t}]^{\mathtt{i}}$. The dynamic model follows from (2) as

$$\mathbf{z}_t^{\mathtt{i}} = \mathbf{z}_{t-1}^{\mathtt{i}} + \boldsymbol{\mu}_t^{\mathtt{i}} \qquad \forall t, \tag{7}$$

where $\boldsymbol{\mu}_t$ is a noisy version of the innovation $\boldsymbol{\delta}_t$. Whereas in *UP-SR* this innovation is neglected, in *recUP-SR* it is assimilated to the uncertainty considered in the dynamic model. In this work, we assume a constant velocity model with an acceleration $\gamma_t^{\mathtt{i}}$ following a Gaussian distribution $\gamma_t^{\mathtt{i}} \sim \mathcal{N}(0, \sigma_a^2)$. As a result, the noisy innovation maybe expressed as

$$\boldsymbol{\mu}_t^{\mathtt{i}} = w_{t-1}^{\mathtt{i}} \Delta t + \frac{1}{2} \gamma_t^{\mathtt{i}} \Delta t^2. \tag{8}$$

The dynamic model in (7) can then be rewritten as:

$$\begin{cases} \mathbf{z}_t^{\mathtt{i}} = \mathbf{z}_{t-1}^{\mathtt{i}} + w_{t-1}^{\mathtt{i}} \Delta t + \frac{1}{2} \gamma_t^{\mathtt{i}} \Delta t^2 \\ w_t^{\mathtt{i}} = w_{t-1}^{\mathtt{i}} + \gamma_t^{\mathtt{i}} \Delta t \end{cases}. \tag{9}$$

Considering the following state vector

$$\mathbf{s}_t^{\mathtt{i}} = \begin{pmatrix} \mathbf{z}_t^{\mathtt{i}} \\ w_t^{\mathtt{i}} \end{pmatrix}, \tag{10}$$

where both the depth measurement and the radial displacement are to be filtered, (9) becomes

$$\mathbf{s}_t^{\mathtt{i}} = \mathbf{K} \mathbf{s}_{t-1}^{\mathtt{i}} + \boldsymbol{\gamma}_t^{\mathtt{i}}, \tag{11}$$

with $\mathbf{K} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}$, and $\boldsymbol{\gamma}_t^{\mathtt{i}} = \gamma_t^{\mathtt{i}} \begin{pmatrix} \frac{1}{2} \Delta t^2 \\ \Delta t \end{pmatrix}$ is the process noise which is white Gaussian with the covariance

$$\mathbf{Q} = \sigma_a^2 \Delta t^2 \begin{pmatrix} \Delta t^2/4 & \Delta t/2 \\ \Delta t/2 & 1 \end{pmatrix}. \tag{12}$$

Using standard Kalman equations, the prediction is achieved as

$$\begin{cases} \hat{\mathbf{s}}_{t|t-1}^{\mathtt{i}} = \mathbf{K} \mathbf{s}_{t-1|t-1}^{\mathtt{i}}, \\ \hat{\mathbf{P}}_{t|t-1}^{\mathtt{i}} = \mathbf{K} \mathbf{P}_{t-1|t-1}^{\mathtt{i}} \mathbf{K}^T + \mathbf{Q}, \end{cases} \tag{13}$$

where $\mathbf{P}_t^{\mathtt{i}}$ is the error covariance matrix. The error in the prediction of $\hat{\mathbf{s}}_{t|t-1}^{\mathtt{i}}$ is corrected using the observed measurement $\tilde{\mathbf{z}}_t^{\mathtt{i}}$. This error is considered as the difference between the prediction and the observation, and weighted using the Kalman gain matrix $\mathbf{G}_{t|t}^{\mathtt{i}}$ which is calculated as follows:

$$\mathbf{G}_{t|t}^{\mathtt{i}} = \hat{\mathbf{P}}_{t|t-1}^{\mathtt{i}} \mathbf{b}^T \left(\mathbf{b} \hat{\mathbf{P}}_{t|t-1}^{\mathtt{i}} \mathbf{b}^T + \sigma_n^2 \right)^{-1}, \tag{14}$$

such that the observation vector is $\mathbf{b} = (1, 0)^T$. The corrected state vector $\mathbf{s}_{t|t}^{\mathtt{i}} = \begin{pmatrix} \mathbf{z}_{t|t}^{\mathtt{i}} \\ w_{t|t}^{\mathtt{i}} \end{pmatrix}$ and corrected error covariance matrix $\mathbf{P}_{t|t}^{\mathtt{i}}$ are computed as follows:

$$\begin{cases} \mathbf{s}_{t|t}^{\mathtt{i}} = \hat{\mathbf{s}}_{t|t-1}^{\mathtt{i}} + \mathbf{G}_{t|t}^{\mathtt{i}} \left(\tilde{\mathbf{z}}_t^{\mathtt{i}} - \mathbf{b} \hat{\mathbf{s}}_{t|t-1}^{\mathtt{i}} \right), \\ \mathbf{P}_{t|t}^{\mathtt{i}} = \hat{\mathbf{P}}_{t|t-1}^{\mathtt{i}} - \mathbf{G}_{t|t}^{\mathtt{i}} \mathbf{b} \hat{\mathbf{P}}_{t|t-1}^{\mathtt{i}}. \end{cases} \tag{15}$$

This per-pixel filtering is extended to all the depth frame resulting in n Kalman filters run in parallel. Each filter tracks the trajectory of one pixel. At this level, pixel trajectories are assumed to be independent. The advantage of the processing per pixel is to reduce all the required matrix inversions to simple scalar inversions. The burden of traditional Kalman filter-based SR as in [13] will consequently be avoided. Moreover, for a recursive multi-frame SR algorithm, instead of using a video sequence of length N to recover one frame, we use the preceding recovered frame $\hat{\mathbf{f}}_{t-1}$ to estimate $\mathbf{f}_t$ from the current upsampled observation $\mathbf{g}_t \uparrow$. Furthermore, in order to separate background from foreground depth pixels, and tackle the problem of flying pixels, especially around edges, we define a condition for track re-initialization. This condition is based on a fixed threshold τ such that

$$\begin{cases} \text{Continue the track} & \text{if } |\tilde{\mathbf{z}}_t^{\mathtt{i}} - \hat{\mathbf{z}}_{t|t-1}^{\mathtt{i}}| < \tau; \\ \text{New track \& spatial median} & \text{if } |\tilde{\mathbf{z}}_t^{\mathtt{i}} - \hat{\mathbf{z}}_{t|t-1}^{\mathtt{i}}| \geqslant \tau. \end{cases}$$

The choice of the threshold value τ is related to the type of the used depth sensor and the level of the sensor-specific noise. This step is very important for the overall performance of *recUP-SR*. Without it, the temporal filtering would continue even if the registered pixels from time $(t-1)$ to t are not matching (e. g. a background pixel wrongly registered to a foreground pixel) and hence the outcome of filtering will be completely wrong for the pixel under consideration at time t. This is an ad hoc solution that provides satisfactory results for the considered examples.

The assumption of independent trajectories leads to blurring artifacts, and requires a corrective step to bring back the correlation between neighbouring pixels from the original depth surface $\mathcal{Z}$. To that end, we use an L_1 minimization where we propose a multi-level iterative BTV regularization as detailed in Section 4.3.

4.3 Multi-Level Iterative Bilateral TV Deblurring

Similarly to the *UP-SR* algorithm, $\mathbf{f}_t$ is estimated in two steps; first, finding a blurred version $\hat{\mathbf{z}}_t$, which is the result of Section 4.2. Then the deblurring takes place to recover $\hat{\mathbf{f}}_t$ from $\hat{\mathbf{z}}_t$. To that end, we apply the following deblurring framework

$$\hat{\mathbf{f}}_t = \underset{\mathbf{f}_t}{\operatorname{argmin}} \left(\|\mathbf{B}\mathbf{f}_t - \hat{\mathbf{z}}_t\|_1 + \lambda \Gamma(\mathbf{f}_t) \right), \tag{16}$$

where λ is a regularization parameter that controls the amount of regularization needed to recover the original blur and noise-free frame. The matrix $\mathbf{B}$ is the blur matrix introduced in Section 3.2. We choose to use a BTV regularizer [17] in order to enforce the properties of bilateral filtering on the final solution [27], [55], [56]. It is a filter that has been shown to perform well on depth data [20], [21], [22]. Indeed, it is a filter that smoothes an image while preserving its sharp edges based on pixel similarities in both the spatial and in the intensity domains. The BTV regularizer is defined as

$$\Gamma(\mathbf{f}_t) = \sum_{p=-P}^{p=P} \sum_{q=0}^{q=P} \alpha^{|p|+|q|} \parallel \mathbf{f}_t - \mathbf{X}^p \mathbf{Y}^q \mathbf{f}_t \parallel_1 . \tag{17}$$

TABLE 1
3D RMSE in mm for the Super-Resolved Dancing Girl Sequence Using Different SR Methods

	$\sigma = 25$ mm				$\sigma = 50$ mm			
	Arm	Torso	Leg	Full body	Arm	Torso	Leg	Full body
Bicubic	10.5	7.5	8.9	8.8	25.2	14.9	13.1	16.5
SISR [31]	**9.0**	5.6	8.4	6.6	14.1	6.9	9.6	9.7
UP-SR [8], [10]	22.2	15.6	9.3	15.9	29.7	17.4	12.8	23.5
Proposed *recUP-SR*	9.6	**3.6**	**7.5**	**6.3**	**9.9**	**4.8**	**8.1**	**9.5**

The SR scale factor is $r = 4$.

The matrices $\mathbf{X}^p$ and $\mathbf{Y}^q$ are shifting operators which shift $\mathbf{f}_t$ by p and q pixels in the horizontal and vertical directions, respectively. The parameters P defines the size of the exponential kernel and the scalar $\alpha \in]0, 1]$ is its base which controls the speed of decay. Minimizing the cost function in (16) has shown to give good results in *UP-SR* [9], [10]; however, unless all the parameters are perfectly chosen, which is a challenge in itself, the final result can end up being a denoised and deblurred version of $\mathbf{f}_t$, which is also over-smoothed. This issue has been addressed by iterative regularization in the case of denoising [57], [58], [59], [60], and in the more general case of deblurring [61]. In the same spirit, we use an iterative regularization where we propose to focus on the choice of the regularization parameter λ. Specifically, our deblurring method consists in running the minimization (16) multiple times where the regularization strength is progressively reduced in a dyadic way. We define, thus, a multi-level iterative deblurring with a BTV regularization such that the solution at level l is

$$\hat{\mathbf{f}}_t^{(l)} = \underset{\mathbf{f}_t^{(l)}}{\operatorname{argmin}} \left(\|\mathbf{B}\mathbf{f}_t^{(l)} - \mathbf{f}_t^{(l-1)}\|_1 + \frac{\lambda}{2^l}\Gamma(\mathbf{f}_t^{(l)}) \right), \text{ with } \mathbf{f}_t^{(0)} = \hat{\mathbf{z}}_t. \quad (18)$$

Combined with a steepest descent numerical solver, the proposed solution is described by the following pseudocode:

```
for $l = 1, \dots, L$
for $k = 1, \dots, K$
$\hat{\mathbf{f}}_t^{(l,k)} = \hat{\mathbf{f}}_t^{(l,k-1)} - \beta\Big\{\mathbf{B}^T \operatorname{sign}\Big(\mathbf{B}\hat{\mathbf{f}}_t^{(l,k-1)} - \mathbf{z}_t\Big) + \frac{\lambda}{2^l}\sum_{p=-P}^{p=P}$
$\sum_{q=0}^{q=P} \alpha^{|p|+|q|}(\mathbf{I} - \mathbf{Y}^{-q}\mathbf{X}^{-p}) \operatorname{sign}\Big(\hat{\mathbf{f}}_t^{(l,k-1)} - \mathbf{X}^p\mathbf{Y}^q\hat{\mathbf{f}}_t^{(l,k-1)}\Big)\Big\}$
end for
$\mathbf{z}_t \longleftarrow \hat{\mathbf{f}}_t^{(l,K)}$
end for
```

The parameter β is an empirically chosen scalar which represents the step size in the direction of the gradient, $\mathbf{I}$ is the identity matrix, and $\operatorname{sign}(\cdot)$ is the sign function. The parameter L is the number of levels considered, and K is the number of iterations for one level. We note that the correct formulation of the problem at the beginning of Section 4 is to use the final deblurred depth value $\mathbf{f}_{t-1}^{\hat{\mathrm{i}}}$ obtained as a solution of (18) instead of $\mathbf{z}_{t-1}^{\hat{\mathrm{i}}}$.

5 Experimental Results

We evaluate the performance of *recUP-SR* against state-of-the-art methods and evaluate the impact of each step in the algorithm. Finally, we give additional examples illustrating the features of *recUP-SR* on real data from simple scenes with one moving object to more complex cluttered scenes containing multiple moving objects with non-rigid deformations. Depth videos of dynamic scenes with non-rigid deformations were captured with a ToF camera, PMD camboard nano [43] or the 3D MLI [65].

5.1 Comparison with State-of-Art Methods

In order to compare our algorithm with state-of-art methods, we use a scene with a highly non-rigidly moving object. We use the publicly available "Samba" [62] data. This data corresponds to a real sequence of a dancing lady scene in full 3D. This sequence contains both non-rigid radial motions and self-occlusions, represented by arms and leg movements, respectively. We use the publicly available toolbox V-REP [63] to create from the "Samba" data a synthetic depth sequence with fully known ground truth. We choose to fix a depth camera at a distance of 2 meters from the 3D scene. Its resolution is $1{,}024^2$ pixels. The camera is used to capture the depth sequence. Then we downsample the obtained depth sequence with $r = 4$ and further degrade it with additive Gaussian noise with standard deviation σ varying from 0 to 50 mm. The created LR noisy depth sequence is then super-resolved using state-of-art methods: the conventional bicubic interpolation, *UP-SR* [8], *SISR* [31], and the proposed *recUP-SR*. Table 1 reports the 3D reconstruction error of each method at different noise levels. Then, we compare the accuracy of the reconstructed 3D super-resolved scene with state-of-art results. The comparison is done by back-projecting the reconstructed HR depth images to the 3D world using the camera matrix and calculating the 3D Root Mean Squared Error (RMSE) of each back-projected 3D point cloud with respect to the ground truth 3D point cloud. The comparison is done at two levels: (i) Different parts of the reconstructed 3D body, namely, arm, torso, and leg, and (ii) full reconstructed 3D body. As expected, by applying the conventional bicubic interpolation method directly on depth images, a large error is obtained. This error is mainly due to the flying pixels around object boundaries. Thus, we run another round of experiments using a modified bicubic interpolation, where we remove all flying pixels by defining a fixed threshold. Yet, the 3D reconstruction error is still relatively high across all noise levels, see Table 1. This is due to the fact that bicubic interpolation does not profit from the temporal information provided by the sequence. We observe in Table 1 that the proposed method provides, most of the time, better results as compared to state-of-art algorithms. In order to visually evaluate the performance of the proposed *recUP-*

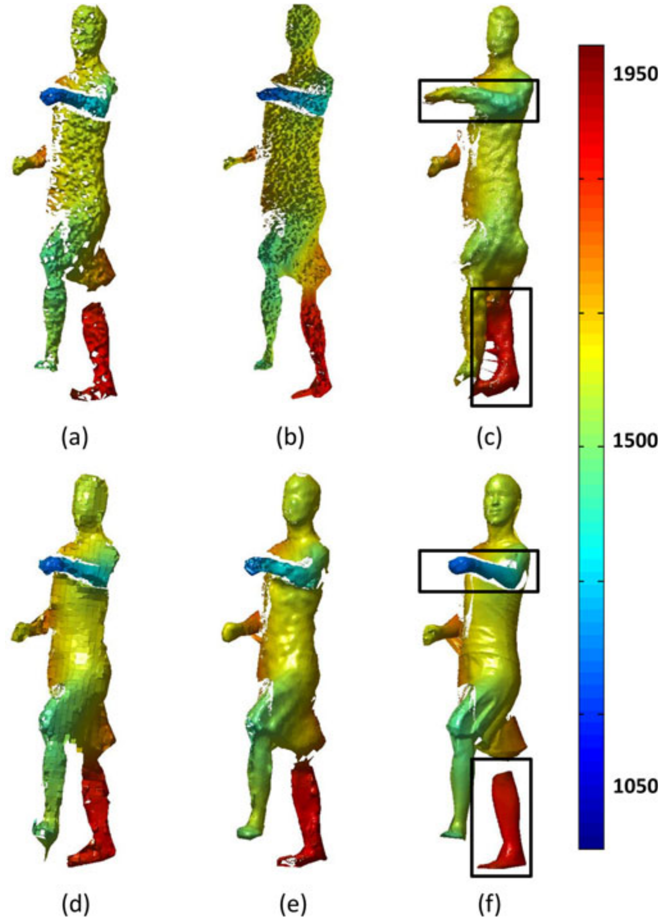

Fig. 4. 3D Plotting of one super-resolved depth frame with $r = 4$ using: (b) bicubic interpolation, (c) *UP-SR* [8], (d) *SISR* [31], (e) proposed *recUP-SR* with $L = 3, K = 7, \lambda = 2.5$. (a) LR noisy depth frame. (f) 3D ground truth. Color bar in mm.

SR algorithm, we plot the super-resolved results for one frame of the dancing girl sequence in 3D. We show the results for the sequence corrupted with $\sigma = 30$ mm. We note that *recUP-SR* outperforms state-of-art methods by keeping the fine details (e.g., the details of the face). Note that the *UP-SR* algorithm fails in the presence of radial movements and self-occlusions, see black boxes in Fig. 4c. In contrast, the *SISR* algorithm can handle these cases, but cannot keep the fine details due to its patch-based nature, see Fig. 4d. In addition, a heavy training phase is required for *SISR*.

We next compare *recUP-SR* with multi-modal fusion approaches. To that end, we use the same moving chairs data presented in [8] and in [10]. The used setup to capture this data was an LR ToF camera, the 3D MLI of resolution (56×61) [65], mounted in the ceiling at a height of 2.5 m, and coupled with an HR 2D camera, the Dragonfly2 CCD camera of resolution (648×488) from Point Grey. Both cameras looking at the scene of two persons sitting on chairs sliding in two different directions. In multi-modal fusion approaches, the HR 2D image is a guidance image used to enhance the resolution of the LR depth image. We consider three representative algorithms; the *Pixel Weighted Average Strategy* (*PWAS*) filter [25], and its improved version called *Unified Multi-Lateral* (*UML*) filter [26] which are two fusion filters based on the concept of bilateral filtering, and the method of umpsampling using *anisotropic Total Generalized Variation* (*aTGV*) which is based on a global optimization. The results for one frame are given in Fig. 5. One can see that *PWAS* and *UML* suffer from texture copying from the 2D image while *aTGV* gives large errors on boundaries. The result of *recUP-SR* is cleaner with smooth clear edges and no texture copied from 2D. It is important to note that *recUP-SR* falls under the category of multi-frame SR (Section 2). As such, the HR 2D images were not used; instead a sequence of 30 frames was used to obtain the reported result. The lateral flow was computed from amplitude images directly captured by the 3D MLI camera.

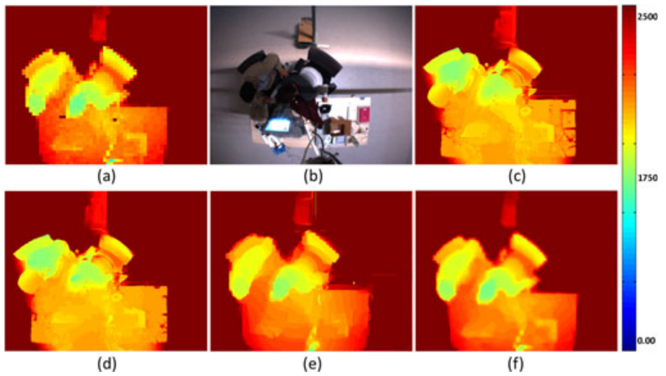

Fig. 5. Comparison with multi-modal fusion methods. (a) Raw LR depth image. (b) HR 2D image. (c) *PWAS* [25], (d) *UML* [26], (e) *aTGV* [28], (f) Proposed *recUP-SR* and corresponding full video is available through this link. Color bar in mm.

5.2 Evaluation of Different Steps

We evaluate the performance of the *recUP-SR* algorithm at different levels. First, we show how it is efficient in filtering both the depth value as well as the radial displacement and hence the corresponding velocity. Then we evaluate the range flow estimation, and finally, we show the importance of deblurring.

5.2.1 Filtering of Depth and Radial Displacement

We start with a simple and fully controlled scene containing one 3D object moving radially with respect to the camera. The considered object is a synthetic hand. A sequence of 20 depth frames is captured at a radial distance of 5 cm between each two successive frames, and with $\Delta t = 0.1$ s. The generated sequence is downsampled with a scale factor of $r = 2$, and $r = 4$, and further degraded with additive Gaussian noise with a standard deviation σ varying from 10 to 80 mm. We then super-resolve the obtained LR noisy depth sequences by applying the proposed algorithm with a scale factor of $r = 1$, $r = 2$, and $r = 4$. Obtained results show that by increasing the scale factor r, a higher 3D error is introduced as seen in Fig. 6. In the simple case where $r = 1$, the SR problem is merely a denoising one, and hence there is no blur due to upsampling. In contrast, by increasing the SR factor r, more blurring effects occur leading to a higher 3D error. Furthermore, in order to evaluate the quality of the filtered depth data and the filtered velocity, we randomly choose one pixel $\mathrm{p}_t^{\mathrm{i}}$ from each super-resolved sequence with $r = 1$, $r = 2$, and $r = 4$, and a fixed noise level for σ=50 mm. For each one of these pixels, we track the corresponding enhanced depth value $\mathbf{f}_t^{\mathrm{i}}$ and the

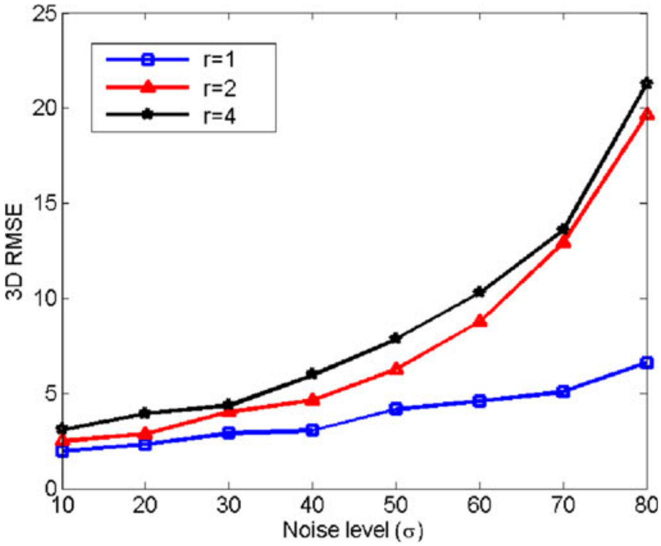

Fig. 6. 3D RMSE in mm of the super-resolved hand sequence in Section 5.2.1 using *recUP-SR*.

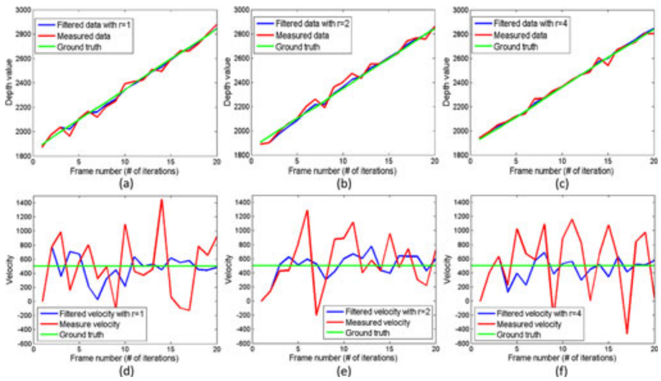

Fig. 7. Tracking results for different depth values randomly chosen from the super-resolved hand sequence in Section 5.2.1 with different SR scale factors $r = 1, r = 2,$ and $r = 4$, are plotted in (a), (b), and (c), respectively. The corresponding filtered velocities are shown in (d), (e), and (f), respectively.

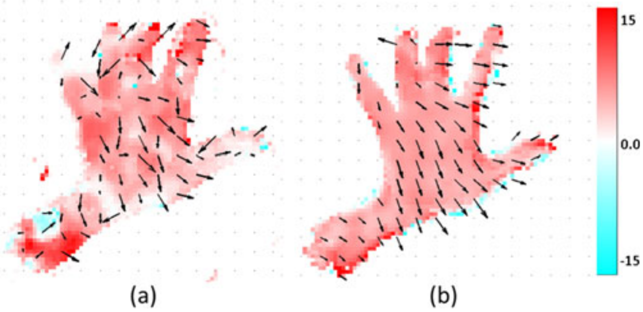

Fig. 8. Estimated range flow on the hand deforming sequence of Section 5.2.2: (a) *aTGV-SF* [29], (b) proposed *recUP-SR*. Arrows represent the lateral motion. The color represents the radial motion in mm. Full video available through this link.

corresponding enhanced velocity $\frac{\Delta \mathbf{f}_t^i}{\Delta t}$ through the super-resolved sequence. In Figs. 7a, 7b, and 7c, we can see how the depth values are filtered (blue lines) as compared to the noisy depth measurements (red lines) for all scale factors. Similar behaviour is observed for the corresponding filtered velocities.

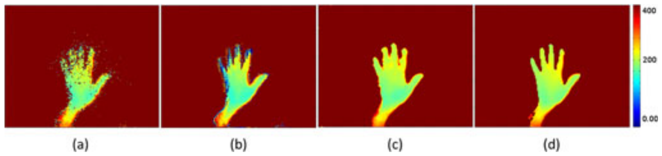

Fig. 9. Results of SR on the hand deforming sequence of Section 5.2.2: (a) *aTGV-SF (flow+SR)* [29] (b) *recUP-SR (flow)* + *aTGV-SF (SR)* (c) *aTGV-SF (flow)* + *recUP-SR (SR)*, (d) proposed *recUP-SR (flow+SR)*.

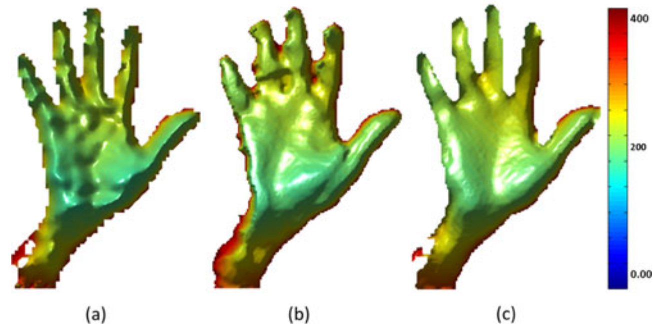

Fig. 10. Results of SR on the hand deforming sequence of Section 5.2.2: (a) Raw LR, (b) *aTGV-SF (flow)+recUP-SR (SR)*, (c) proposed *recUP-SR (flow+SR)*.

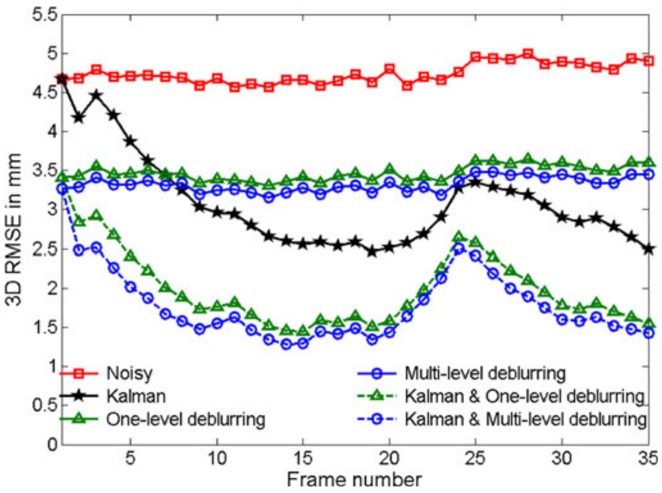

Fig. 11. Effects of applying different steps separately and combined on a sequence of 35 LR noisy depth frames with $\sigma = 10$ mm. The combination of the Kalman filter with the spatial multi-level deblurring provides the best performance in reducing the 3D RMSE.

5.2.2 Estimated Range Flow

One of the main contributions of this paper is the estimation of range flow by point tracking with a Kalman filter. We visually evaluate the accuracy of this flow on a known sequence of a hand non-rigidly deforming and captured with the PMD CamBoard Nano camera. We qualitatively compare this result with the state-of-art range flow algorithm *aTGV-SF* [29]. The result for one frame is given in Fig. 8. One can see that the proposed approach provides a smoother and more homogeneous flow, which is more accurate. Moreover, this flow is computed at a frame rate of 19 frames per second on a CPU and without parallelization considering $r = 2$, which is faster than *aTGV-SF*'s reported runtime of 1 frame per second using an optimized code. We have used the publically available *aTGV-SF* Matlab code which took around 200 seconds for one frame. We have kept the same parameters used with the

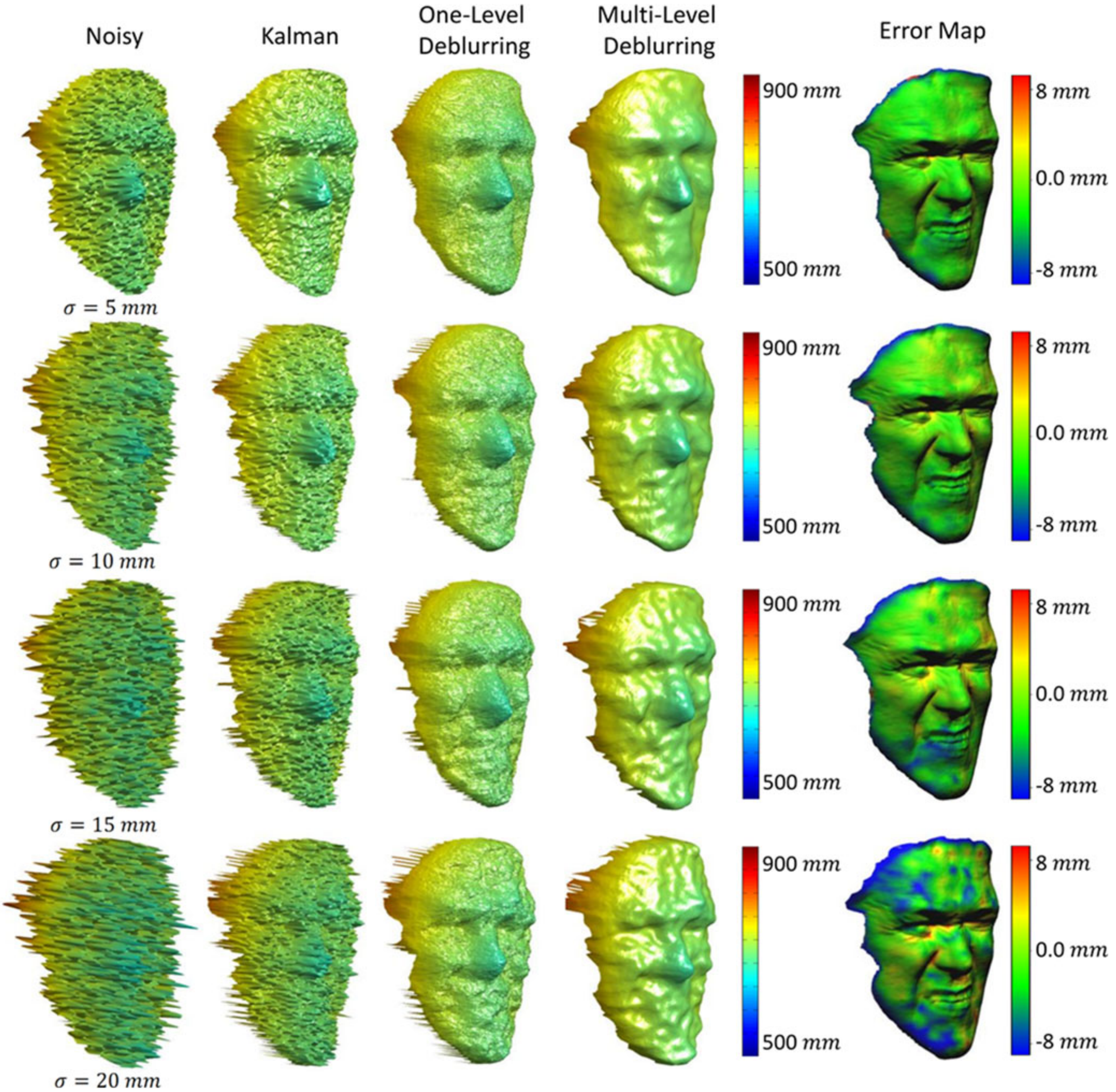

Fig. 12. 3D plotting of (starting from left column): 1) LR noisy depth frames, 2) super-resolved depth frames with $r = 4$ using Kalman filter, 3) super-resolved depth frame with $r = 4$ using the proposed method with one-level deblurring step with $L = 1, K = 25$ 4) super-resolved depth frame with $r = 4$ using the proposed method with the proposed multi-level deblurring step with $L = 5, K = 25$, 5) error map of comparing the obtained results in forth column with the 3D ground truth.

PMD CamBoard Nano camera as in [29]. The *aTGV-SF* flow was also used to super-resolve a depth scene containing a non-rigidly deforming object. The proposed SR consists in median filtering all registered frames using the estimated flow after back-projecting them to the 3D world. We give the result of *aTGV-SF (flow+SR)* as well as our result *recUP-SR (flow+SR)* in Figs. 9a and 9d, respectively. This result shows that the proposed algorithm outperforms *aTGV-SF* at the level of the flow and also at the SR level. Fig. 9b and 9c give the result of using the flow of *recUP-SR (flow)* in the SR of *aTGV-SF (SR)*, and vice versa. This shows that while the flow estimated through *recUP-SR* is good and outperforms state-of-art methods, it is not sufficient to directly use it in a multi-frame SR algorithm. Using the *aTGV-SF (flow)* in the proposed SR algorithm, however, provides a more acceptable result although still inferior to the proposed *recUP-SR (flow+SR)*. This is confirmed by the corresponding 3D rendering given in Fig. 10, and can be explained by the fact that the proposed *recUP-SR* algorithm is a simultaneous filtering of the flow and also the depth values. This ensures an effective reconstruction of non-rigid depth scenes.

5.2.3 Deblurring

In order to better understand the contribution of deblurring, we consider the “Facecap” data [64] which is a simple scene of a real 3D face sequence with non-rigid deformations. We use a similar setup to the one used with the “Samba” dataset by fixing a camera at a distance

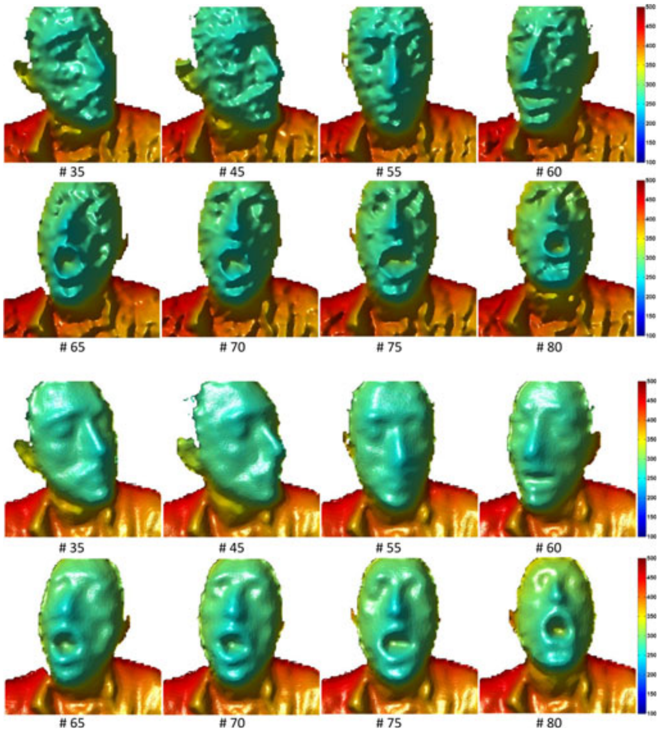

Fig. 13. Results of applying the proposed algorithm on a real sequence captured by a LR ToF camera (120×160 pixels) of a non-rigidly moving face. First and second rows contain a 3D plotting of selected LR captured frames. Third and fourth rows contain the 3D plotting of the super-resolved depth frames with $r = 4$. Distance units on the coloured bar are in mm. Full video available https://dropit.uni.lu/invitations?share=b6ed393cddde9693250b&dl=0

of 0.7 m from the 3D face. We create a new synthetic depth sequence of the moving face.

Then, we downsample the obtained depth sequence with $r = 4$ and further degrade it with additive Gaussian noise with standard deviation σ varying from 0 to 20 mm. The obtained LR noisy depth sequence is then super-resolved with $r = 4$ using: 1) Kalman filter, 2) spatial deblurring, and 3) the proposed *recUP-SR* algorithm. In the deblurring process, two different techniques are considered, one-level deblurring and the proposed multi-level deblurring. The accuracy of the reconstructed 3D face sequences is measured by calculating the 3D RMSE.

In Fig. 11, we report the obtained results for the super-resolved LR noisy depth sequence with $\sigma = 10$ mm. We see how the Kalman filter attenuates the noise gradually and hence decreasing the 3D RMSE for an increased number of frames (black solid line). We notice that, in the presence of a non-smooth motions, the constant velocity filtering model needs few number of iterations (frames) before converging which affects the reconstruction quality of the super-resolved depth frame. For example, due to the up and down non-smooth and fast motions of the eye brows between frame number 20 to 25, the per-pixel temporal filtering is not converged yet, and hence the 3D error increases for a few number of frames before decreasing again Fig. 11 (Black solid line). By considering the deblurring step alone without engaging in the per-pixel temporal filtering process, we can see that the 3D RMSE is almost constant throughout the sequence as shown in Fig. 11 (solid blue and green lines). This can be explained by the fact that there is no engagement of temporal information. Instead only a spatial filtering is applied at each frame independently of each other. Finally, by looking at the obtained results in Fig. 11, we find that the best performance is achieved by combining the spatial and the temporal filters (blue and green dashed lines), with an advantage of using the proposed multi-level deblurring approach over the one-level conventional deblurring approach. Note that an intensive search is applied to find the best deblurring parameters which lead to the smallest 3D RMSE error.

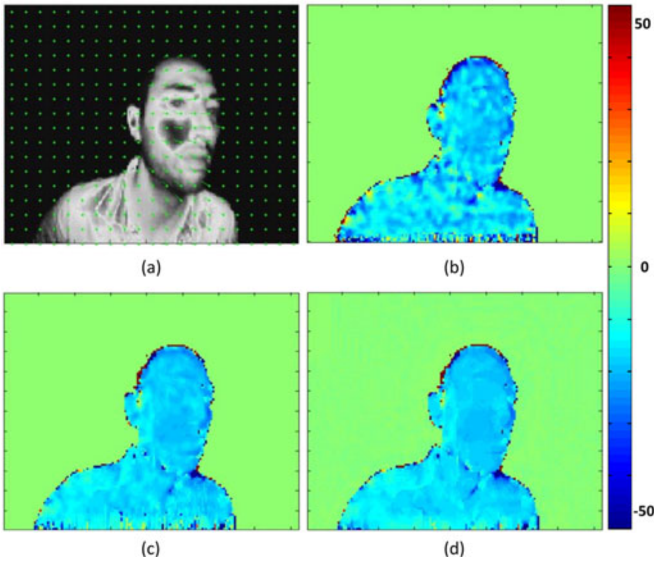

Fig. 14. Radial depth displacement filtering. (a) 2D optical flow calculated from the normalized LR amplitude images. (b) Raw noisy depth radial depth displacement. (c) Filtered radial depth displacement using Kalman filter. (d) Filtered radial depth displacement using the proposed method. Color bar in mm.

In Fig. 12, we show the physical effects of the previously discussed cases by plotting the corresponding 3D super-resolved results of the last HR depth frame in the sequence. Starting from the first column, we show the LR noisy faces for different noise levels. The filtered results using a per-pixel Kalman filtering are shown in the second column where we see how the noise has been attenuated. The results of the proposed algorithm using the one-level deblurring step, with $L = 1$ and $K = 25$, and the multi-level deblurring step, with $L = 5$ and $K = 5$, are plotted in the third and fourth columns, respectively. By visually comparing the obtained results, we find that the proposed algorithm with the multi-level deblurring process provides the best results and hence confirms the quantitative evaluation of Fig. 11 (blue dashed line) where it provides the lowest 3D RMSE.

5.3 Additional Examples

We run the *recUP-SR* on different LR real depth sequences captured with the PMD CamBoard Nano of resolution (120×160) [43]. First, we start with a simple scene with one non-rigidly moving face. Then, we show the robustness of *recUP-SR* to topology changes by testing it on a more complex and cluttered scene containing multiple moving objects. The algorithm's runtime on all these sequences for an SR scale factor of $r = 1$ is 38 frames per second, $r = 2$ is 20 frames per second, and $r = 3$ is 9 frames per second. This is using a 2.2 GHz i7 processor with 4 Gigabyte RAM and an unoptimized code.

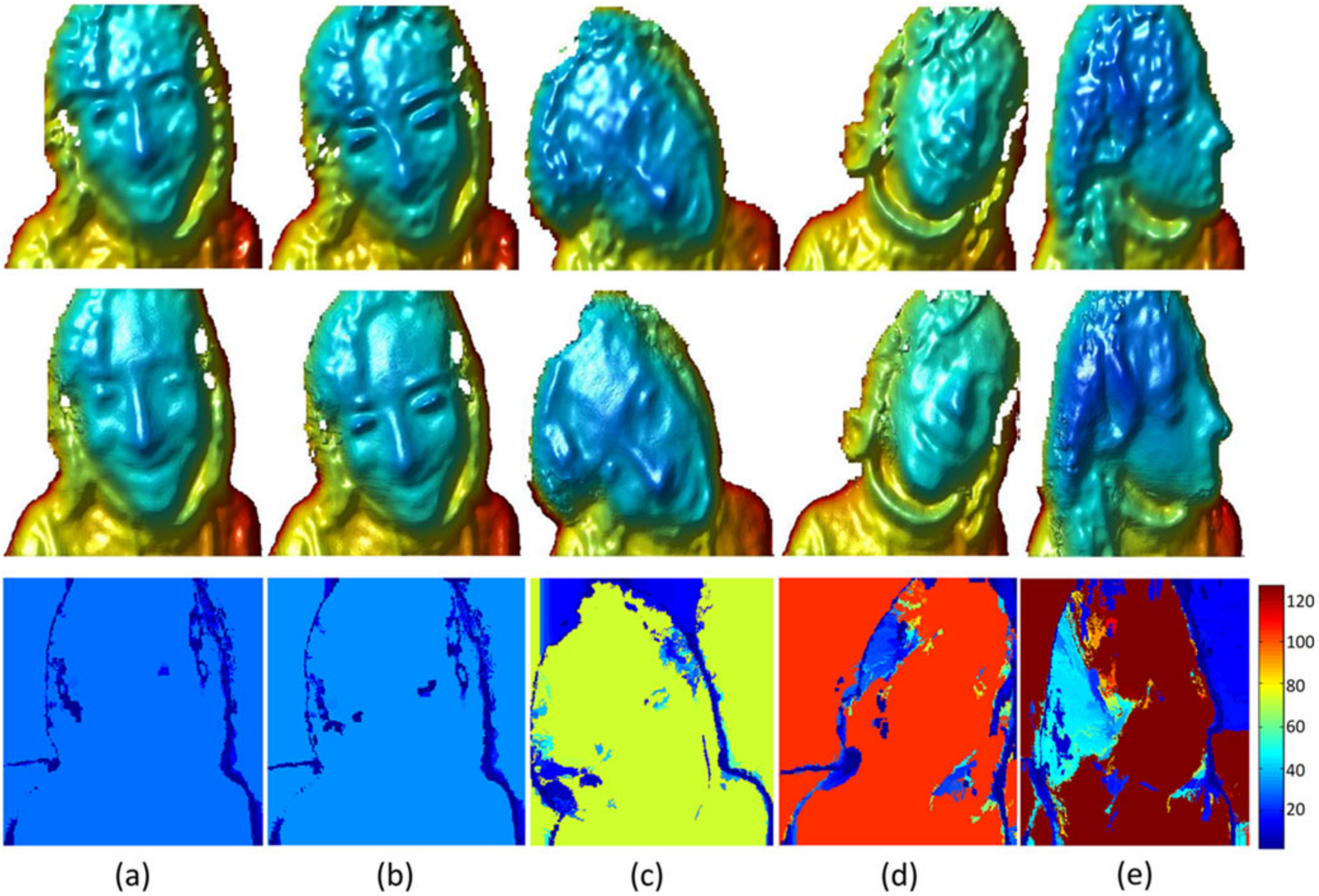

Fig. 15. Results of applying the proposed algorithm on a real sequence captured by a LR ToF camera (120×160 pixels) of a non-rigidly moving face. First and second rows contain a 3D plotting of selected LR captured frames and the 3D plotting of the super-resolved depth frames with $r = 6$, respectively. Third row shows the tracking life for each pixel through the sequence. Units of the coloured bar represents the tracking life (iterations).

5.3.1 One Non-Rigidly Moving Object

We test the proposed algorithm on a real LR depth sequence of a non-rigidly deforming face with large motions and local non-rigid deformations. We super-resolve this sequence using the proposed algorithm with an SR scale factor of $r = 4$. Obtained results are given in 3D in Fig. 13. They visually show the effectiveness of the proposed algorithm in reducing noise, and further increasing the resolution of the reconstructed 3D face under large non-rigid deformations. Full video of results is available through this link. To visually appreciate these results as compared to state-of-art methods, we tested the bicubic, *UP-SR*, and *SISR* methods on the same LR real depth sequence. Obtained results show the superiority of the *recUP-SR* as compared to other methods, see Fig. 1. We show in Fig. 14b how the raw radial depth displacement is noisy and ranges from -50 to -10 mm while in fact the real displacement of the face in this frame has to be smooth and homogeneous. By applying the proposed algorithm, the noisy displacement is refined to match the real homogeneous displacement of an approximate value of -20 mm, see Fig. 14d.

We run another experiment on a second real sequence composed of 120 depth frames of a face moving with long hair causing strong self-occlusions. The goal of this experiment is to show how the tracking process is reinitialized in the self-occlusion case for all pixels representing the self-occluded area. We super-resolve the acquired sequence with a scale factor of $r = 6$. Obtained results are shown in Fig. 15. It is interesting to see in the third row how the tracking life for each pixel is evolving through the time with stronger occlusions causing shorter tracking, hence illustrating the impact of the threshoulding parameter τ on the performance of the proposed algorithm. For example, all pixels with the dark red colors in Fig. 15d have been appeared through the full sequence and no self-occlusion happened and hence the track continues. In contrast, for most of the boundary pixels the tracking process has been reinitialized (blue dark) and thus a spatial median filter is applied for these pixels.

5.3.2 Cluttered Scene

Finally, we tested *recUP-SR* on a cluttered scene of moving hands transferring a ball from one hand to another. This scene is quite complex where it contains multiple objects moving with non-rigid deformations, and self-occlusions with one hand passing over the second one. Moreover, the scene contains a challenging case of topology changes represented by hands touching each other and then separating. We note that a strong temporal filtering leads to a longer time for convergence in the case of self-occlusions or non-smooth motions. Similarly, a strong spatial filtering leads to undesired over-smoothing effects and hence removing the fine details from the final reconstructed HR depth sequence. Thus, in order to handle such a scene, a trade-off between

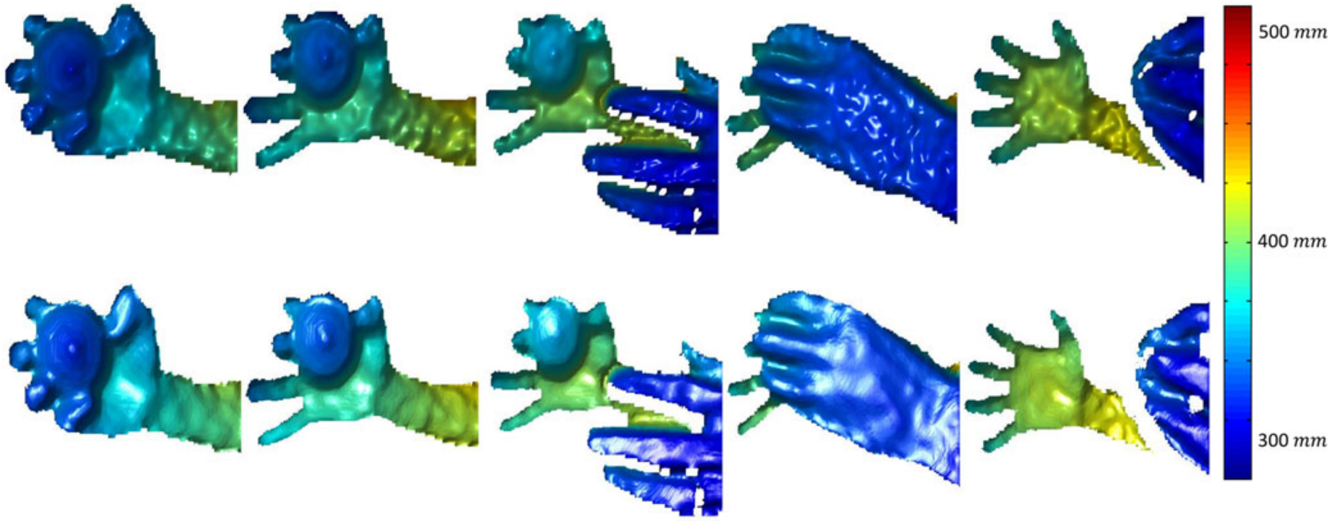

Fig. 16. Results of applying the proposed algorithm on a real sequence captured by a LR ToF camera (120×160) of a cluttered scene. First row contains a 3D plotting of selected LR captured frames. Second row contains a 3D plotting of the corresponding super-resolved depth frames with $r = 3$. Full video available through this link.

the temporal and spatial filtering has to be achieved. Obtained results in Fig. 16 show the robustness of the proposed algorithm in handling this kind of scenes. Full video of results is available through this link.

6 Discussion and Conclusions

We have proposed a new algorithm to enhance the quality of low resolution noisy depth videos acquired with cost-effective depth sensors. This algorithm improves upon the *UP-SR* algorithm [8], [10] which was able to handle non-rigid deformations but limited to lateral motions. The newly proposed algorithm, *recUP-SR*, is designed to handle non-rigid deformations in 3D thanks to a per-pixel filtering that directly accounts for radial displacements in addition to lateral ones. This algorithm is formulated in a dynamic recursive way that allowed a computationally efficient real-time implementation on CPU. Moreover, as compared to state-of-the-art methods, the processing on depth maps while approximating local 3D motions has allowed to maintain a good robustness against topological changes and independence of the number of moving objects in the scene. This property is a clear advantage over most recent methods that explicitly compute a flow in 3D and apply a processing on meshed point clouds [11], [12]. In order to keep smoothness properties without losing details, each filtered depth frame is further refined using a multi-level iterative bilateral total variation regularization after filtering and before proceeding to the next frame in the sequence. This post-processing is shown experimentally to give the best final results in terms of 3D error without having access to full information about the scene and the sensor. Supported by the experimental results on both synthetic and real data, we believe that *recUP-SR* opens new possibilities for computer vision applications using cost-effective depth sensors in dynamic scenarios with non-rigid motions. Further developments in the case of strong self-occlusions are still required. As shown, the proposed *recUP-SR* algorithm needs a number of depth measurements before converging, which is not suitable for some applications. Moreover, a realistic depth sensor noise model is more complex than the Gaussian model considered in this work. An extended work would be to adapt the proposed solution to a more elaborate noise model on depth measurements.

Acknowledgments

This work was supported by the National Research Fund, Luxembourg, under C11/BM/1204105/FAVE/Ottersten. The "Facecap" data were provided courtesy of the research group Graphics, Vision & Video of the Max-Planck-Institute for Informatics. The authors would like to thank David Ferstl et al. for making the codes for [28] and [29] publically available, and Konstantinos Papadopoulos and Michel Antunes for their help with the experiments. They also thank the Associate Editor and the anonymous reviewers for their suggestions and comments.

References

[1] [Online]. Available: https://www.microsoft.com/en-us/kinectforwindows/, Accessed on: Nov. 04, 2016.

[2] S. Berretti, A. Del Bimbo, and P. Pala, "Superfaces: A superresolution model for 3D faces," in *Proc. 5th Workshop Non-Rigid Shape Anal. Deformable Image Alignment*, 2012, pp. 73–82.

[3] D. Aouada, K. Al Ismaeil, K. K. Idris, and B. Ottersten, "Surface UP-SR for an improved face recognition using low resolution depth cameras," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, 2014, pp. 107–112.

[4] S. Berretti, P. Pala, and A. Del Bimbo, "Face recognition by super-resolved 3D models from consumer depth cameras," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 9, pp. 1436–1449, Sep. 2014.

[5] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "LidarBoost: Depth superresolution for ToF 3D shape scanning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 343–350.

[6] Y. Cui, S. Schuon, S. Thrun, D. Stricker, and C. Theobalt, "Algorithms for 3D shape scanning with a depth camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1039–1050, May 2013.

[7] S. Izadi, et al., "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 559–568.

[8] K. Al Ismaeil, D. Aouada, B. Mirbach, and B. Ottersten, "Dynamic super resolution of depth sequences with non-rigid motions," in *Proc. 20th IEEE Int. Conf. Image Process.*, 2013, pp. 660–664.

[9] D. Aouada, K. Al Ismaeil, and B. Ottersten, "Patch-based statistical performance analysis of upsampling for precise super-resolution," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, 2015, pp. 186–193.

[10] K. Al Ismaeil, D. Aouada, B. Mirbach, and B. Ottersten, "Enhancement of dynamic depth scenes by upsamplig for precise super-resolution (UP-SR)," *Comput. Vis. Image Understanding*, vol. 147, pp. 38–49, 2016.

[11] H. Afzal, K. Al Ismaeil, D. Aouada, F. Destelle, B. Mirbach, and B. Ottersten, "KinectDeform: Enhanced 3D reconstruction of non-rigidly deforming objects," in *Proc. 3DV Workshop Dynamic Shape Meas. Anal.*, 2014, vol. 2, pp. 7–13.

[12] R. Newcombe, D. Fox, and S. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 343–352.

[13] M. Elad and A. Feuer, "Super-resolution restoration of an image sequence: Adaptive filtering approach," *IEEE Trans. Image Process.*, vol. 8, no. 3, pp. 387–395, Mar. 1999.

[14] B. C. Newland, A. D. Gray, and D. Gibbins, "Modified Kalman filtering for image super-resolution: Experimental convergence results," in *Proc. 9th Int. Conf. Signal Image Process.*, 2007, pp. 58–63.

[15] S. Farsiu, M. Elad, and P. Milanfar, "Video-to-video dynamic super-resolution for grayscale and color sequences," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 232–232, 2006.

[16] J. Tian and K.-K. Ma, "A new state-space approach for super-resolution image sequence reconstruction," in *Proc. IEEE Int. Conf. Image Process.*, 2005, vol. 1, pp. 881–885.

[17] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[18] K. Al Ismaeil, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten, "Real-time non-rigid multi-frame depth video super-resolution," in *Proc. IEEE Comput. Vis. Pattern Recognit. Workshop*, 2015, pp. 8–16.

[19] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. Conf. Neural Inf. Process. Syst.*, 2005, pp. 291–298.

[20] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, 2007, Art. no. 96.

[21] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[22] F. Garcia, D. Aouada, B. Mirbach, T. Solignac, and B. Ottersten, "A new multi-lateral filter for real-time depth enhancement," in *Proc. 8th IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, 2011, pp. 42–47.

[23] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1400–1414, Jul. 2011.

[24] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. M. Bruckstein, "RGBD-fusion: Real-time high precision depth recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5407–5416.

[25] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *Proc. IEEE Conf. Image Process.*, 2010, pp. 2805–2808.

[26] F. Garcia, D. Aouada, B. Mirbach, T. Solignac, and B. Ottersten, "Unified multi-lateral filter for real-time depth map enhancement," *Image Vis. Comput.*, vol. 41, pp. 26–41, 2015.

[27] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 836–846.

[28] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 993–1000.

[29] D. Ferstl, C. Reinbacher, G. Riegler, M. Rüther, and H. Bischof, "aTGV-SF: Dense variational scene flow through projective warping and higher order regularization," in *Proc. 2nd Int. Conf. 3D Vis.*, 2014, vol. 1, pp. 285–292.

[30] B. Ham, M. Cho, and J. Ponce, "Robust image filtering using joint static and dynamic guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4823–4831.

[31] O. Mac Aodha, N. Campbell, A. Nair, and G. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 71–84.

[32] M. Hornacek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1123–1130.

[33] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 349–356.

[34] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, pp. 24:1–24:11, 2009.

[35] J. Li, Z. Lu, G. Zeng, R. Gan, and H. Zha, "Similarity-aware patchwork assembly for depth image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3374–3381.

[36] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2012, pp. 152–157.

[37] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3D full human bodies using Kinects," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 4, pp. 643–650, Apr. 2012.

[38] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, "3D self-portraits," *ACM Trans. Graph.*, vol. 32, pp. 187:1–187:9, 2013.

[39] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3D scanning deformable objects with a single RGBD sensor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 493–501.

[40] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives," *ACM Trans. Graph.*, vol. 32, pp. 42:1–42:10, 2013.

[41] M. Zollhofer, et al., "Real-time non-rigid reconstruction using an RGB-D camera," *ACM Trans. Graph.*, vol. 33, 2014, Art. no. 156.

[42] [Online]. Available: http://www.primesense.com/, Accessed on: 2014.

[43] PMD Technologies. Siegen, Germany. Camboard Nano. [Online]. Available: http://www.pmdtec.com, Accessed on: Nov. 04, 2016.

[44] K. Al Ismaeil, D. Aouada, B. Mirbach, and B. Ottersten, "Depth super-resolution by enhanced shift and add," in *Proc. 15th Int. Conf. Comput. Anal. Images Patterns*, 2013, pp. 100–107.

[45] M. Yamamoto, P. Boulanger, J. A. Beraldin, and M. Rioux, "Direct estimation of range flow on deformable shape from a video rate range camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 1, pp. 82–89, Jan. 1993.

[46] H. Spies, B. Jahne, and J. Barron, "Regularised range flow," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 785–799.

[47] H. Spies, B. Jahne, and J. L. Barron, "Range flow estimation," *Comput. Vis. Image Understanding*, vol. 85, pp. 209–231, 2002.

[48] H. Spies and J. L. Barron, "Evaluating the range flow motion constraint," in *Proc. 16th Int. Conf. Pattern Recognit.*, 2002, vol. 3, pp. 517–520.

[49] J.-M. Gottfried, J. Fehr, and C. S. Garbe, "Computing range flow from multi-modal kinect data," in *Proc. 7th Int. Symp. Advances Visual Comput.*, 2011, pp. 758–767.

[50] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers, "Stereoscopic scene flow computation for 3D motion understanding," *Int. J. Comput. Vis.*, vol. 95, pp. 29–51, 2011.

[51] E. Herbst, X. Ren, and D. Fox, "RGB-D flow: Dense 3-D motion estimation using color and depth," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 2276–2282.

[52] S. Hadfield and R. Bowden, "Kinecting the dots: Particle based scene flow from depth sensors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2290–2295.

[53] S. Hadfield and R. Bowden, "Scene particles: Unregularized particle-based scene flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 564–576, Mar. 2014.

[54] M. Sturmer, J. Penne, and J. Hornegger, "Standardization of intensity-values acquired by time-of-flight-cameras," in *Proc. IEEE Workshop Comput. Vis. Pattern Recognit.*, 2008, pp. 1–6.

[55] K. Al Ismaeil, D. Aouada, B. Mirbach, and B. Ottersten, "Bilateral filter evaluation based on exponential kernels," in *Proc. 20th IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 258–261.

[56] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1141–1151, Oct. 2002.

[57] W. Li, C. Zhao, Q. Liu, Q. Shi, and S. Xu, "A parameter-adaptive iterative regularization model for image denoising," *EURASIP J. Advances Signal Process.*, vol. 2012, 2012, Art. no. 222.

[58] M. R. Charest, M. Elad, and P. Milanfar, "A general iterative regularization framework for image denoising," in *Proc. 40th Annu. Conf. Inf. Sci. Syst.*, 2006, pp. 452–457.

[59] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multi-Scale Model Simulation*, vol. 4, pp. 460–489, 2005.

[60] P. Milanfar, "A tour of modern image filtering: New insights and methods, both practical and theoretical," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 106–128, Jan. 2013.

[61] A. Kheradmand and P. Milanfar, "A general framework for regularized, similarity-based image restoration," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5136–5151, Dec. 2014.

[62] [Online]. Available: http://people.csail.mit.edu/drdaniel/mesh_animation/, Accessed on: Nov. 04, 2016.

[63] [Online]. Available: http://www.k-team.com/mobile-robotics-products/v-rep, Accessed on: Nov. 04, 2016.

[64] L. Valgaerts, C. Wu, A. Bruhn, H. P. Seidel, and C. Theobalt, "Lightweight binocular facial performance capture under uncontrolled lighting," *ACM Trans. Graph.*, vol. 31, pp. 187:1–187:11, 2012.

[65] [Online]. Available: http://www.iee.lu/home-page, Accessed on: Nov. 04, 2016.

Kassem Al Ismaeil received the BSc degree in electronic engineering and the MSc degree in computer science both from the University of Aleppo, Aleppo, Syria, in 2006 and 2008, respectively. He received the Erasmus Mundus MSc degree in computer vision and robotics from the University of Burgundy, Le Creusot, France, in 2011, and the PhD degree in computer science from Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg, in 2015. He received the Best Paper Award at the IEEE 2nd CVPR Workshop on Multi-Sensor Fusion and Dynamic Scene Understanding (MSF'15). His research interests include 3D computer vision, image and video processing, with a focus on depth super-resolution, 3D reconstruction, and structure from motion. He is a student member of the IEEE.

Djamila Aouada (S'05-M'09) received the state engineering degree in electronics from Ecole Nationale Polytechnique (ENP), Algiers, Algeria, in 2005, and the PhD degree in electrical engineering from North Carolina State University (NCSU), Raleigh, North Carolina, in 2009. She is a research scientist with Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg. She has been leading the computer vision activities with SnT since 2009. She has worked as a consultant for multiple renowned laboratories (Los Alamos National Laboratory, Alcatel Lucent Bell Labs., and Mitsubishi Electric Research Labs.). Her research interests span the areas of signal and image processing, computer vision, pattern recognition and data modelling. She is the co-author of two IEEE Best Paper Awards. She is a member of the IEEE.

Thomas Solignac received the engineering degree in telecommunications from the Institut Supérieur d'Électronique du Nord, Lille, France, in 1995, and the DEA degree in electronics from the Institut d'Électronique et de Micro-électronique du Nord, Valenciennes, France, in 1995. He worked as development engineer and project leader in industry automation with main focus on fully automated optical inspection systems for the wood industry. Since 2005, he has been a senior computer vision engineer with IEE S.A., Luxembourg working in the research and developent of innovative 2D- and 3D-camera based machine vision applications with main focus on real-time object detection and tracking. The application fields are automotive safety systems, advanced driver assistant systems, as well as advanced building security and automation.

Bruno Mirbach received the PhD degree in theoretical physics from the University of Kaiserslautern, Kaiserslautern, Germany, in 1996. He was a postdoctoral researcher with Center for Nonlinear and Complex Systems, Como, Italy, Max-Planck-Institute of the Physics of Complex Systems, Dresden, Germany, and the University of Ulm, Germany. His research interests include focussing on non-linear dynamics and quantum chaos. After that he joined automotive industry, working on the research and development of intelligent optical systems for safety applications. He is currently senior algorithm group leader with IEE S.A., Luxembourg, responsible for the development of computer vision algorithms for advanced driver assistant systems, as well as for systems in advanced building security and automation. His current research interests span the areas of 3D-vision, 2D/3D sensor fusion, image processing and machine learning, with the main focus on real-time object detection and tracking.

Björn Ottersten (S'87-M'89-SM'99-F'04) received the MS degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986. In 1989, he received the PhD degree in electrical engineering from Stanford University, Stanford, California. He has held research positions in the Department of Electrical Engineering, Linköping University, Information Systems Laboratory, Stanford University, Katholieke Universiteit Leuven, Leuven, and University of Luxembourg. During 96/97, he was director of research with ArrayComm Inc, a start-up, San Jose, California, based on Ottersten's patented technology. He has co-authored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, and 2013 and three IEEE conference papers receiving Best Paper Awards. In 1991 he was appointed a professor of signal processing with the Royal Institute of Technology (KTH), Stockholm. From 1992 to 2004 he was head of the department for signals, sensors, and systems with KTH and from 2004 to 2008 he was dean of the School of Electrical Engineering, KTH. Currently, he is a director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg. His research interests include security and trust, reliable wireless communications, and statistical signal processing. He is a fellow of the IEEE.