Enhancement of Dynamic Depth Scenes by Upsampling for Precise Super-Resolution (UP-SR)

Kassem Al Ismaeil[a], Djamila Aouada[a], Bruno Mirbach[b], Björn Ottersten[a]

[a] *Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg.*
[b] *Advanced Engineering Department, IEE S.A., Luxembourg*

Abstract

Multi-frame super-resolution is the process of recovering a high resolution image or video from a set of captured low resolution images. Super-resolution approaches have been largely explored in 2-D imaging. However, their extension to depth videos is not straightforward due to the textureless nature of depth data, and to their high frequency contents coupled with fast motion artifacts. Recently, few attempts have been introduced where only the super-resolution of static depth scenes has been addressed. In this work, we propose to enhance the resolution of dynamic depth videos with non-rigidly moving objects. The proposed approach is based on a new data model that uses densely upsampled, and cumulatively registered versions of the observed low resolution depth frames. We show the impact of upsampling in increasing the sub-pixel accuracy and reducing the rounding error of the motion vectors. Furthermore, with the proposed cumulative motion estimation, a high registration accuracy is achieved between non-successive upsampled frames with relative large motions. A statistical performance analysis is derived in terms of mean square error explaining the effect of the number of observed frames and the effect of the super-resolution factor at a given noise level. We evaluate the accuracy of the proposed algorithm theoretically and experimentally as function of the SR factor, and the level of contaminations with noise. Experimental results on both real and synthetic data show the effectiveness of the proposed algorithm on dynamic depth videos as compared to state-of-art methods.

Keywords: Super-resolution, moving objects, non-rigid motion, depth video, upsampling, cumulative motion.

1. Introduction

Interactive computer vision applications using depth data have literally exploded in recent years thanks to the development of new depth sensors that are currently accessible to everyone. Most of these applications deal with dynamic scenes containing one or multiple moving objects. Depth sensors, such as time-of-flight (ToF) cameras are, however, still limited by their high contamination with noise and their low pixel resolutions. Moreover, such cameras can be highly sensitive to fast motions leading to motion artifacts; hence, affecting the reliability of depth measurements [1]. Some examples of such cameras are the MLI by IEE S.A. [2] of resolution (56×61) pixels, and the PMD CamBoard nano [3] of resolution (120×165) pixels.

Most of the works proposed to enhance the resolution and quality of depth images have been based on fusion with a high resolution (HR) image acquired with a second camera, e.g., a 2-D camera [4, 5], a stereo camera [6], or both 2-D and stereo cameras [7]. These multi-modality methods suffer from drawbacks such as undesired texture copying, and blurring artifacts. In addition, the performance of these systems depends on parameter tuning, and may encounter difficulties related to data mapping and synchronization.

The multi-frame super-resolution (MFSR) framework offers an alternative solution where an HR image is to be recovered from a set or a sequence of low resolution (LR) images captured with the same camera [8]. The observed LR images are subject to deviations from the reference image due to relative motion and to aliasing errors caused by the acquisition system. MFSR can be formulated as an inverse problem where the deviations on LR frames are explored to estimate the reference HR image. Super-resolution (SR) techniques have been largely explored in 2-D imaging. However, their extension to depth data is not straightforward as presented in [9, 10, 11] where only the SR of a static object has been addressed. The difficulty of applying SR to depth videos is further illustrated in the context of single image SR (SISR) in [12] where a dedicated preprocessing followed by a heavy training were proposed. Indeed, depth data is characterized by its textureless nature with high frequency contents. Moreover, fast motions and surface reflectivity of objects in the scene create invalid pixels and the so-called flying pixels [1]; thus, making most existent 2-D SR algorithms fail when directly applied on dynamic depth videos.

In this paper, we propose an MFSR algorithm for dynamic depth scenes. The proposed solution can handle scenes containing one or more moving objects even non-rigidly without

Email addresses: kassem.alismaeil@uni.lu (Kassem Al Ismaeil), djamila.aouada@uni.lu (Djamila Aouada), bruno.mirbach@iee.lu (Bruno Mirbach), bjorn.ottersten@uni.lu (Björn Ottersten)

[1]This work was supported by the National Research Fund, Luxembourg, under the CORE project C11/BM/1204105/FAVE/Ottersten.

prior assumptions on their shape, and without training. Our algorithm referred to as *Upsampling for Precise Super Resolution* (UP-SR) builds on our work in [13, 14, 15]. We herein give a unified framework and provide additional details and proofs, and a more extensive experimental part, where we evaluate the accuracy of the proposed algorithm theoretically and experimentally as function of the SR factor, and the level of contaminations with noise.

UP-SR is based on a new data model that uses densely upsampled, and cumulatively registered versions of the observed LR frames. It is these two key components, together, that constitute the working principle of UP-SR as detailed below:

1) *Upsampling*: Most SR algorithms are directly related to a registration based on a too coarse pixel correspondence as compared to the scale of details in the scene. This leads to failure in handling local deformations of moving objects. It is therefore necessary to call upon a very accurate sub-pixel correspondence. In what follows, we argue that this accuracy is significantly increased after upsampling the observed sequence as supported by [16]. Moreover, we prove that the upsampling process reduces the errors caused by rounding the motion vectors.

2) *Cumulative motion estimation*: In order to achieve a high registration accuracy between non-successive upsampled frames with relative large motions, we propose a new cumulative motion estimation process. The proposed method is based on using the temporal information provided by the intermediate frames between the reference frame and the frame under consideration.

The remainder of this paper is organized as follows: Section 2 reviews state-of-the-art SR techniques in the context of their extension to depth data, and to dynamic scenes. Section 3 introduces the problem formulation for the classical MFSR. We prove the improvements in accuracy and robustness due to estimating motion from densely upsampled depth images in Section 4. The proposed data model is presented in Section 5 along with the proposed cumulative motion estimation, leading to the UP-SR algorithm for dynamic depth scenes with moving objects. Then, a statistical performance analysis is given in Section 6. Section 7 reports a thorough experimental evaluation of the UP-SR approach and its comparison with state-of-the-art methods. Discussions and conclusion are provided in Section 8.

2. Related Work

MFSR is the process of recovering an HR image from a set of captured LR frames. It is based on using the deviation between these frames and a reference frame as provided by relative motion, where the ratio between HR and LR defines the SR factor. Depending on the type of motion, two categories of scenes may be distinguished, and accordingly two categories of SR algorithms; SR for static scenes and SR for dynamic scenes. In the static case, the motion is global where frames could be seen as slightly different perspectives of the same scene. The scene is said to be dynamic if there is at least one moving object with non-rigid deformations; thus, the estimation of a local motion becomes necessary. In order to understand the challenges related to applying SR to depth data, we review state-of-art approaches for both static and dynamic scenes.

2.1. SR for Static Scenes

The SR estimation is solved numerically using iterative methods starting from an initial image. This image may be obtained by interpolation [17], which is not suitable in the case of textureless depth data, as interpolating depth data would induce erroneous values and flying pixels that are difficult to attenuate. Another approach is known as *Shift & Add* (S&A) [18, 19] which includes a filling procedure based on the global relative motion of the considered LR images. Schuon et al. have applied in [9] the S&A method of [19] to depth images acquired with a ToF camera. In [10], the same authors proposed to replace the regularization term in [19] by a new term tailored for depth data, specifically, ToF data, leading to a new depth-dedicated SR method referred to as *LidarBoost*. The aim of LidarBoost is to preserve areas with a smooth geometry by using a regularization term that is a function of spatial gradients approximated with finite differences. The original LidarBoost uses an L_2-norm of weighted depth gradients. In order to better accommodate the needs of detailed 3-D object scanning, Cui et al. proposed a new version of LidarBoost where the regularization term is set to be an anisotropic non-linear function of gradients [11]. In both cases, however, the initial HR is obtained by means of averaging, which is not appropriate for sensing cluttered scenes. This adaptation of SR to static depth data is quite promising but remains restricted to static scanning where the method assumes a perfectly controlled setup with a turning table-like procedure implying a large motion diversity by construction, but not handling non-rigid motions.

2.2. SR for Dynamic Scenes

Dynamic scenes are challenging scenarios for MFSR as they require the local motion of moving objects to be computed accurately. They, hence, may face the problem of self-occlusions especially in the case of non-rigidly moving objects. This difficulty arises in depth videos, but also for 2-D sequences [21, 22, 23, 24]. Most of the methods in the literature are limited due to strong assumptions on the shape and number of moving objects. For this reason, the enhancement of the resolution of dynamic depth scenes has been so far mostly based on fusion with higher resolution 2-D data that has to be simultaneously captured [4, 5]; thus, requiring a perfect alignment, synchronization, and mapping of the 2-D and depth images, and assuming the correspondence of edges on the two modalities. These methods may be computationally efficient, but unfortunately they frequently suffer from artifacts caused by the heuristic nature of the enforced statistical model, mainly copying the intensity texture of 2-D images to depth images.

In this paper, we propose a new MFSR algorithm for dynamic depth scenes with moving objects. Our algorithm is largely independent of surface texture and does not suffer from the texture copying problem since it only deals with LR depth frames as inputs without fusion with any other type of sensors.

In what follows, in Section 3, we formulate the problem of dynamic MFSR.

3. Problem Formulation

The aim of dynamic MFSR algorithms is to estimate a sequence of HR images $\{\mathbf{x}_{t_0}\}$ of size $(\sqrt{n} \times \sqrt{n})$ from observed LR sequences. The dynamic SR problem can be simplified by reconstructing one HR image at a time, $\mathbf{x}_{t_0}$, for $t_0 \in \mathbb{N}$ using an LR sequence $\{\mathbf{y}_t\}_{t_0-N+1}^{t_0}$ of length N, where each LR image $\mathbf{y}_t$ is of size $(\sqrt{m} \times \sqrt{m})$ pixels, with $\sqrt{n} = r \cdot \sqrt{m}$, where r is the SR factor, such that $r \geq 1$. Note that for the sake of simplicity, and without loss of generality, we assume squared images. Every image $\mathbf{y}_t$ may be viewed as an LR noisy and deformed realization of $\mathbf{x}_{t_0}$ at the acquisition time t, with $t \leq t_0$. Rearranging all images in lexicographic order, i.e., column vectors of lengths n for $\mathbf{x}_t$, and m for $\mathbf{y}_t$, we consider the following data model:

$$\mathbf{y}_t = \mathbf{D}\mathbf{H}\mathbf{M}_{t_0}^t \mathbf{x}_{t_0} + \mathbf{n}_t, \quad t \leq t_0, \tag{1}$$

where $\mathbf{D}$ is a matrix of dimension $(m \times n)$ that represents the downsampling operator, and which we assume to be known and constant over time. The system blur is represented by the time and space invariant matrix $\mathbf{H}$. The vector $\mathbf{n}_t$ is an additive Laplacian noise at time t, as justified in [18, 19]. The matrices $\mathbf{M}_{t_0}^t$ are $(n \times n)$ matrices corresponding to the geometric motion between the considered HR image $\mathbf{x}_{t_0}$ and the observed LR image $\mathbf{y}_t$ prior to its downsampling.
Based on the data model in (1), and using an L_1-norm between the observations and the model, the Maximum Likelihood (ML) estimate of $\mathbf{x}_{t_0}$ is obtained as follows:

$$\hat{\mathbf{x}}_{t_0} = \arg\min_{\mathbf{x}_{t_0}} \sum_{t=t_0-N+1}^{t_0} \|\mathbf{D}\mathbf{H}\mathbf{M}_{t_0}^t \mathbf{x}_{t_0} - \mathbf{y}_t\|_1. \tag{2}$$

Using the same approach as in [19, 27], we consider that $\mathbf{H}$ and $\mathbf{M}_{t_0}^t$ are block circulant matrices. Therefore: $\mathbf{H}\mathbf{M}_{t_0}^t = \mathbf{M}_{t_0}^t\mathbf{H}$. The minimization in (2) can then be decomposed into two steps; initialization by estimating the blurred HR image $\mathbf{z}_{t_0} = \mathbf{H}\mathbf{x}_{t_0}$, followed by a deblurring step to recover $\hat{\mathbf{x}}_{t_0}$. In what follows, we assume that $\mathbf{y}_t$ is simply the noisy and decimated version of $\mathbf{z}_t$ without any geometric warp. We may thus write $\mathbf{M}_t^t = \mathbf{I}_n, \forall t$, $\mathbf{I}_n$ being the identity matrix of size $(n \times n)$, hence, $\mathbf{M}_{t_0}^t \mathbf{z}_{t_0} = \mathbf{z}_t = \mathbf{H}\mathbf{x}_t$. This operation can be assimilated to registering $\mathbf{z}_{t_0}$ to $\mathbf{z}_t$. We draw attention to the fact that in the case of static MFSR, instead of a sequence, a set of observed LR images is considered, i.e., there is no order between frames. Such an order becomes crucial in dynamic SR because the estimation of motion, based on the optical flow paradigm, happens between consecutive frames only. An accurate dynamic SR estimation is consequently highly dependent on the accuracy of estimating the registration matrices between consecutive frames $\mathbf{M}_t^{t-1}$, as well as the motion between non-consecutive frames $\mathbf{M}_{t_0}^t$ with $t < t_0 - 1$.
In Section 4, we discuss the higher accuracy of estimating consecutive motion matrices $\mathbf{M}_t^{t-1}$ using upsampled images, and leading to an enhanced pyramidal motion estimation. In Section 5, we present our strategy for a cumulative estimation of the non-consecutive motion matrices $\mathbf{M}_{t_0}^t$, leading to the final proposed UP-SR algorithm.

4. Enhanced Pyramidal Motion

In the UP-SR approach, a highly accurate motion estimation with a $\pm\frac{1}{2}$ sub-pixel accuracy at the HR level is desired. This corresponds to a sub-pixel accuracy of $\pm\frac{1}{2r}$ at the LR level. To reach this objective, two ways may be considered: 1) tuning the parameters of the chosen optical flow algorithm until the desired accuracy is reached, then multiplying the LR motion vectors by the SR factor r; 2) upsampling the LR frames prior to estimating motion. The main disadvantage of the former solution is that full knowledge of the used optical flow algorithm and its parameters is needed. In addition, modifying the parameters in order to increase the accuracy requires increasing the number of iterations in the optical flow related optimization process. On the other hand, the latter solution could be seen as a more systematic option. The choice between these two solutions is totally based on the targeted application. Either ways, the registration has to be done at the upsampled level in order to attenuate the rounding error of motion vectors.

In this work, we propose to follow the second option, and to upsample the observed LR images even before registering them. We further detail the advantages of this approach in the context of pyramidal motion estimation (PyrME) [25, 26]. Indeed, PyrME is the principle followed by most optical flow algorithms used in the SR framework. PyrME uses the pyramidal strategy to increase sub-pixel accuracy and robustness to large motions as compared to estimating motions directly from observed frames. In what follows, we describe PyrME as it is currently used. Then, we present how we further improve its performance in the context of the SR problem. Let $\mathsf{w}_t = (u_t, v_t)$ be the motion vector between a frame $\mathbf{y}_t$ and the reference frame $\mathbf{y}_{t_0}$ at a given target point p. This motion vector is estimated by minimizing the following error:

$$\xi(\mathsf{w}_t) = \sum_{\mathsf{q}=\mathsf{p}-\mu}^{\mathsf{p}+\mu} \|\mathbf{y}_{t_0}(\mathsf{q}) - \mathbf{y}_t(\mathsf{q} + \mathsf{w}_t)\|_2^2. \tag{3}$$

This error is calculated within an integration disc of radius μ, which corresponds to the largest motion that can be detected within this framework. The center of this disc is represented by the target pixel position p. A small value of μ increases the sub-pixel motion accuracy while a large value is preferable in order to increase robustness to large motions. PyrME was proposed as a trade-off solution for these conflicting characteristics. The main idea is to follow a coarse to fine strategy that progressively downsamples the images $\mathbf{y}_t$ and $\mathbf{y}_{t_0}$ starting from the bottom of the pyramid. These images are downsampled by a factor 2^ℓ in the dyadic case, where ℓ indicates the pyramidal level, $\ell = 0, \cdots, L$. Considering two consecutive levels ℓ and $\ell - 1$, the downsampling process may be defined as follows:

$$\mathbf{y}_t^\ell(\mathsf{p}) = \mathbf{y}_t^{\ell-1}(2\mathsf{p}) \quad s.t. \quad \mathbf{y}_t^0 = \mathbf{y}_t, \quad \forall t. \tag{4}$$

In fact, the number of the pyramidal levels L is directly related to the considered minimum size of the downsampled image at the highest level of the pyramid. Let us define this minimum size as $(d \times d)$ pixels. Then, we may define the maximal number of pyramidal levels as:

$$\frac{\sqrt{m}}{2^L} = d \quad \Rightarrow \quad L = \log_2\left(\sqrt{m}\right) - \log_2(d). \tag{5}$$

Starting from the top of the pyramid, the motion is first estimated from the images of lowest resolution, i.e. at the highest level $\ell = L$, before progressively going back down to the images of highest resolution, i.e., at the initial level $\ell = 0$. At each level ℓ, the motion w_t^ℓ between the two images $\mathbf{y}_t^\ell$ and $\mathbf{y}_{t_0}^\ell$ consists of an initial estimate ω_t^ℓ and a residual motion ϕ_t^ℓ. The initial estimate ω_t^ℓ is obtained from the preceding level $(\ell + 1)$ such that $\omega_t^\ell = 2 \cdot \mathsf{w}_t^{(\ell+1)}$, and initially set to zero at the level $\ell = L$. The two images $\mathbf{y}_t^\ell$ and $\mathbf{y}_{t_0}^\ell$ are then pre-registered using the initial motion vector. This pre-registration step reduces the process of finding the optimal motion w_t^ℓ to finding the optimal residual motion. The estimation of the optimal residual motion is then defined by the following minimization:

$$\phi_t^\ell = \underset{\nu}{\operatorname{argmin}} \sum_{\mathsf{q}=\mathsf{p}-\mu}^{\mathsf{p}+\mu} \|\mathbf{y}_{t_0}^\ell(\mathsf{q}) - \mathbf{y}_t^\ell(\mathsf{q} + \omega_t^\ell + \nu)\|_2^2. \tag{6}$$

The optimal motion at level ℓ is then defined as $\mathsf{w}_t^\ell = \omega_t^\ell + \phi_t^\ell$. In order to have a high sub-pixel resolution accuracy, a small neighbourhood disc of radius μ is considered in the refinement operation defined in (6). By repeating the operation in (6) for all the levels of the pyramid, the finest motion vector is obtained at $\ell = 0$ defining w_t as:

$$\mathsf{w}_t := \mathsf{w}_t^0 = \omega_t^0 + \phi_t^0. \tag{7}$$

We may also express this motion using the refined residuals at all levels as follows:

$$\mathsf{w}_t = \sum_{\ell=0}^{L} 2^\ell \phi_t^\ell. \tag{8}$$

The maximal pixel motion vector that can be detected at any level ℓ is restricted by the initial motion vector from the preceding level and the radius of the neighbourhood disc μ in (6). By considering all the refined residuals as in (8), the maximal overall pixel motion that can be detected at the level $\ell = 0$ by PyrME is within a maximum radius of:

$$\mu_{\max} = \mathcal{G}(L) \times \mu \quad \text{with} \quad \mathcal{G}(L) = 2^{(L+1)} - 1. \tag{9}$$

From (9), we see that the maximal motion is controlled by the gain $\mathcal{G}(L)$ and the radius of the neighbourhood disc μ. The gain $\mathcal{G}(L)$ is a function of the height L of the pyramid. By considering a small μ while increasing the number of pyramidal levels, PyrME may estimate large motions up to $\mu_{\max}$; hence, verifying the robustness property in addition to the accuracy one.

In the context of the SR problem, our target is to increase the resolution of the LR images up to the resolution of the final HR images with size $(\sqrt{n} \times \sqrt{n})$ pixels. By increasing the resolution, we thus increase the number of pyramidal levels. This gives us a natural way to further improve the performance of PyrME by upsampling the LR frames up to the SR factor r prior to any motion estimation. This upsampling step directly impacts the two properties of PyrME :

1) *Robustness:*

The upsampling step leads to changing the size of the pyramid base and hence changing the starting point in PyrME. These changes result, in turn, to an increased pyramidal height $L \uparrow^r$ by $\log_2(r)$ which results in a new and higher gain $\mathcal{G}(L \uparrow^r)$:

$$\mathcal{G}(L \uparrow^r) = r \cdot \mathcal{G}(L) + (r - 1), \qquad \text{with} \;\; r > 1. \tag{10}$$

The result in (10) shows that, in the SR context, the robustness to large motions for PyrME, may further be enhanced with a new larger gain $\mathcal{G}(L \uparrow^r)$.

2) *Accuracy:*

By increasing the resolution with a factor r, the initial motion vector at the new level can be estimated from w_t^0 in (7) as $\omega_t^{-\log_2(r)} = r \cdot \mathsf{w}_t^0$. Hence, the optimal refined final motion can be further defined as:

$$\begin{aligned} \mathsf{w}_t := \mathsf{w}_t^{-\log_2(r)} &= \omega_t^{-\log_2(r)} + \phi_t^{-\log_2(r)} \\ &= r \cdot (\omega_t^0 + \phi_t^0) + \phi_t^{-\log_2(r)}. \end{aligned} \tag{11}$$

By back projecting the newly refined motion in (11) to the original resolution at the level $\ell = 0$, we have:

$$\mathsf{w}_t^0 = \omega_t^0 + \phi_t^0 + \frac{\phi_t^{-\log_2(r)}}{r}. \tag{12}$$

Comparing (7) and (12), we find an increase in accuracy of $\delta\mathsf{w}_t(r) = \frac{\phi_t^{-\log_2(r)}}{r}$. This confirms the result in [16] which shows that higher image resolutions help in increasing the accuracy of motion estimation. We note that the advantage of upsampling for PyrME saturates when a certain accuracy increase is reached, i.e., $\lim_{r\to\infty} \delta\mathsf{w}_t(r) = 0$. For the example in Section 7.1, we observed a saturation at $r = 2^3$, as illustrated in Table 2.

5. Novel Reduced SR Data Model

Following the result in Section 4, we use the enhanced PyrME and follow an upsampling strategy as a starting point for a new improved SR algorithm. We thus introduce the concept of Upsampling for Precise Super Resolution (UP-SR). As shown in Section 4, upsampling the observed LR images $\mathbf{y}_t$ prior to any operation should lead to a more accurate and robust motion estimation, which enhances the registration of frames. We define the resulting r-times upsampled image as $\mathbf{y}_t \uparrow = \mathbf{U} \cdot \mathbf{y}_t$, where $\mathbf{U}$ is an $(n \times m)$ upsampling matrix.

5.1. Dense Upsampling

Due to the specific properties of depth data, classical interpolation-based methods, such as bicubic interpolation, cannot be used as they lead to flying pixels and to blurring

effects especially for boundary pixels. Thus, the upsampling $\mathbf{U}$ has to be dense, which is also known as nearest neighbour upsampling. For our problem, it is defined by the following matrix:

$$\mathbf{U} = \begin{bmatrix} \mathbf{Q} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q} \end{bmatrix}, \tag{13}$$

where $\mathbf{0}$ is a zero matrix, and $\mathbf{Q}$ represents the blocks of $\mathbf{U}$ of size ($\sqrt{n}r \times \sqrt{m}$). The dense upsampling implies that

$$\mathbf{Q} = \left[\underbrace{\mathbf{P}^T, \cdots, \mathbf{P}^T}_{r \text{ times}} \right]^T, \tag{14}$$

where T denotes the matrix transpose, and $\mathbf{P}$ is a matrix of size ($\sqrt{n} \times \sqrt{m}$) such that:

$$\mathbf{P} = \begin{bmatrix} \mathbb{1}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbb{1}_r & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbb{1}_r \end{bmatrix} \quad \text{with} \quad \mathbb{1}_r = [\underbrace{1, \cdots, 1}_{r \text{ times}}]^T. \tag{15}$$

We assume in what follows that the upsampling matrix $\mathbf{U}$ is the transpose of the downsampling matrix $\mathbf{D}$. Their product $\mathbf{UD} = \mathbf{A}$ gives another block circulant matrix $\mathbf{A}$ that defines a new blurring matrix $\mathbf{B} = \mathbf{AH}$. The matrix $\mathbf{A}$ is actually a block diagonal matrix with the square matrix $\mathbf{QQ}^T$ repeated $\sqrt{m}$ times on its diagonal. Considering that $\mathbf{B}$ and $\mathbf{M}_{t_0}^t$ are block circulant matrices, we have $\mathbf{BM}_{t_0}^t = \mathbf{M}_{t_0}^t\mathbf{B}$. As a result, the initialization described in Section 3 gets modified where a new blurred HR image $\mathbf{z}_{t_0} = \mathbf{Bx}_{t_0}$ is to be estimated first.

5.2. Cumulative Motion Estimation

Most of optical flow approaches, including the proposed enhanced PyrME, work under the assumption of small motions. Thus, by considering the frames which are far from the reference frame at t_0, high registration errors are introduced as compared to the errors introduced by frames that are closer to t_0. Further frames are therefore considered as outliers. To tackle this problem, we propose a new registration method. This method is based on a cumulative motion estimation where we use the temporal information provided by intermediary frames between the reference frame and the frame under consideration. Each two consecutive upsampled frames $\mathbf{y}_t \uparrow$ and $\mathbf{y}_{t+1} \uparrow$ in the sequence are related as follows:

$$\mathbf{y}_{t+1} \uparrow = \mathbf{M}_t^{t+1}\mathbf{y}_t \uparrow + \mathbf{v}_{t+1}, \tag{16}$$

where $\mathbf{v}_{t+1}$ represents the innovation which is assumed to be negligible. We apply the enhanced PyrME strategy described in Section 4 to estimate the local motion $\mathbf{M}_t^{t+1}$ for all the pixel positions p. By so doing we obtain a dense optical flow.

$$\hat{\mathbf{M}}_t^{t+1} = \arg\min_{\mathbf{M}} \Psi\left(\mathbf{y}_{t+1} \uparrow, \mathbf{y}_t \uparrow, \mathbf{M}\right), \tag{17}$$

where Ψ is a dense optical flow-related cost function, in the simplest case based on local mean squared errors as in (3). The motion from $\mathbf{y}_t \uparrow$ to $\mathbf{y}_{t+1} \uparrow$ is computed in a similar way; thus, leading to the registration of $\mathbf{y}_t \uparrow$ to $\mathbf{y}_{t+1} \uparrow$ as follows:

$$\overline{\mathbf{y}}_t^{t+1} \uparrow = \hat{\mathbf{M}}_t^{t+1}\mathbf{y}_t \uparrow . \tag{18}$$

The main target is to define $\overline{\mathbf{y}}_t^{t_0} \uparrow$, which represents the registered version of $\mathbf{y}_t \uparrow$ to the reference $\mathbf{y}_{t_0} \uparrow$ by using all the registered upsampled images $\overline{\mathbf{y}}_t^{t+1} \uparrow$, as defined in (18), for $t < t_0$, see Figure 1. This approach is similar to the concept proposed in [28], with an additional improvement where we further reduce the cumulated motion error by recomputing $\hat{\mathbf{M}}_t^{t+1}$ using the already registered frame $\overline{\mathbf{y}}_{t-1}^{t} \uparrow$ as follows:

$$\hat{\mathbf{M}}_t^{t+1} = \arg\min_{\mathbf{M}} \Psi\left(\mathbf{y}_{t+1} \uparrow, \overline{\mathbf{y}}_{t-1}^{t} \uparrow, \mathbf{M}\right). \tag{19}$$

We prove by induction (see Appendix A) the following registration equation for non-consecutive frames:

$$\overline{\mathbf{y}}_t^{t_0} \uparrow = \hat{\mathbf{M}}_t^{t_0}\mathbf{y}_t \uparrow = \underbrace{\hat{\mathbf{M}}_{t_0-1}^{t_0} \cdots \hat{\mathbf{M}}_t^{t+1}}_{(t_0 - t) \text{ times}} \cdot \mathbf{y}_t \uparrow, \tag{20}$$

where

$$\hat{\mathbf{M}}_t^{t_0} = \hat{\mathbf{M}}_{t_0-1}^{t_0} \cdots \hat{\mathbf{M}}_t^{t+1}. \tag{21}$$

Note that due to the high noise level in depth raw data, we apply a preprocessing step with a bilateral filter before motion estimation. The bilateral filter is only used in the preprocessing step while the original depth data is mapped in the registration step and further used in the fusion process.

5.3. Proposed UP-SR Algorithm

The classical data model for a dynamic scene is given in (1). The additive noise $\mathbf{n}_t$ follows a white multivariate Laplace distribution as it has been shown to better fit the SR problem as compared to a Gaussian noise model [18, 19]. This distribution is defined as follows:

$$p(\mathbf{n}_t) = \prod_{i=1}^{m} \frac{\sqrt{2}}{2\sigma} \exp\left(-\frac{\sqrt{2}|\mathbf{n}_t(i)|}{\sigma} \right), \tag{22}$$

where $\frac{\sigma}{\sqrt{2}}$ is a positive Laplace scale factor leading to the diagonal covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{I}_m$, with $\mathbf{I}_m$ being the identity matrix of size $(m \times m)$.
Considering the reference frame $\mathbf{x}_{t_0}$, and by left multiplying (1) by $\mathbf{U}$, we find:

$$\mathbf{y}_t \uparrow = \mathbf{M}_{t_0}^t\mathbf{Bx}_{t_0} + \mathbf{Un}_t, \qquad t < t_0. \tag{23}$$

In addition, similarly to [29], for analytical convenience, we assume that all pixels in $\mathbf{y}_t \uparrow$ originate from pixels in $\mathbf{x}_{t_0}$ in a one to one mapping. Therefore, each row in $\mathbf{M}_{t_0}^t$ contains 1 for each position corresponding to the address of the source pixel in $\mathbf{x}_{t_0}$. This bijective property implies that the matrix $\mathbf{M}_{t_0}^t$ is an invertible permutation, $[\hat{\mathbf{M}}_{t_0}^t]^{-1} = \hat{\mathbf{M}}_t^{t_0}$. Following the result in Section 4, and using the cumulative motion proposed in Section 5.2, the motion matrix $\hat{\mathbf{M}}_{t_0}^t$ is obtained from upsampled LR

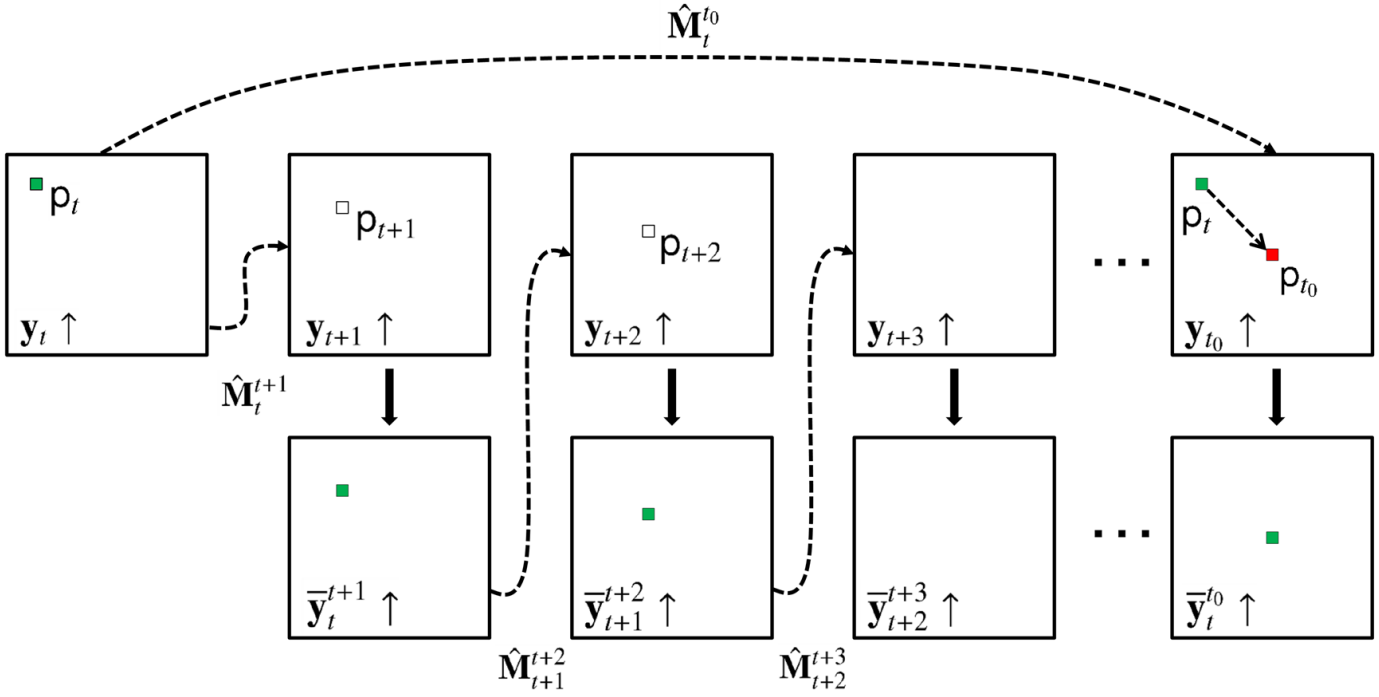

Figure 1: UP-SR Cumulative Motion Estimation: All intermediate registered upsampled depth frames are used to register the pixel p_t in frame $\mathbf{y}_t \uparrow$ to its corresponding pixel at the position p_{t_0} from the reference frame $\mathbf{y}_{t_0} \uparrow$ where $\mathbf{y}_t \uparrow$ and $\mathbf{y}_{t_0} \uparrow$ are non-consecutive upsampled frames.

frames $\mathbf{y}_t \uparrow$, $t = t_0 - N + 1, \cdots, t_0$, as in (21). Thus, the corresponding registrations to the reference $\mathbf{y}_{t_0} \uparrow$ are performed as

$$\mathbf{y}_t \uparrow = \hat{\mathbf{M}}_{t_0}^{t} \overline{\mathbf{y}}_t^{t_0} \uparrow . \tag{24}$$

Given (24), and by left multiplying (23) by $[\hat{\mathbf{M}}_{t_0}^{t}]^{-1}$, we find

$$\overline{\mathbf{y}}_t^{t_0} \uparrow = \mathbf{B}\mathbf{x}_{t_0} + \boldsymbol{\nu}_t, \qquad t < t_0. \tag{25}$$

This finally leads to a new simplified SR data model which is analogous to a classical image denoising problem from multiple observations, specifically

$$\overline{\mathbf{y}}_t^{t_0} \uparrow = \mathbf{z}_{t_0} + \boldsymbol{\nu}_t, \qquad t < t_0, \tag{26}$$

where $\boldsymbol{\nu}_t = \hat{\mathbf{M}}_t^{t_0}\mathbf{U} \cdot \mathbf{n}_t$ is an additive Laplacian noise vector of length n with mean zero and covariance $\tilde{\boldsymbol{\Sigma}} = \hat{\mathbf{M}}_t^{t_0}\mathbf{U}\boldsymbol{\Sigma}\mathbf{D}\hat{\mathbf{M}}_{t_0}^{t}$. Given the data model in (26), the two steps of initialization and deblurring are described below.

Step 1: *Initialization*

The log-likelihood function associated with (26) becomes

$$\begin{aligned} \ln p(\overline{\mathbf{y}}_{t_0-N+1}^{t_0} \uparrow, \cdots, \overline{\mathbf{y}}_{t_0}^{t_0} \uparrow \mid \mathbf{z}_{t_0}) &= \\ &= \ln\left(\prod_{t=t_0-N+1}^{t_0} \frac{\sqrt{2}}{2\sigma} \exp\left(-\frac{\sqrt{2}\|\overline{\mathbf{y}}_t^{t_0} \uparrow - \mathbf{z}_{t_0}\|_1}{\sigma} \right) \right) \\ &= -N \ln \frac{\sigma}{\sqrt{2}} - \frac{\sqrt{2}}{\sigma} \sum_{t=t_0-N+1}^{t_0} \|\mathbf{z}_{t_0} - \overline{\mathbf{y}}_t^{t_0} \uparrow\|_1. \end{aligned} \tag{27}$$

Maximizing (27) with respect to $\mathbf{z}_{t_0}$, we obtain

$$\hat{\mathbf{z}}_{t_0} = \arg\min_{\mathbf{z}_{t_0}} \sum_{t=t_0-N+1}^{t_0} \|\mathbf{z}_{t_0} - \overline{\mathbf{y}}_t^{t_0} \uparrow\|_1 \;\Rightarrow\; \hat{\mathbf{z}}_{t_0} = \mathrm{med}_t\{\overline{\mathbf{y}}_t^{t_0} \uparrow\}_{t=t_0-N+1}^{t_0}. \tag{28}$$

In fact, the equations in (28) represents a temporal pixel-wise median filter med_t, which constitutes the fusion step in the UP-SR algorithm. Taking the median filter as a temporal filter solves the problem of invalid pixels caused by depth sensors [1], and guarantees that no flying pixels are generated, such erroneous pixels are caused, in classical SR methods [10, 11], by averaging background and foreground pixels.

Step 2: *Deblurring*

In this work, we adopt Maximum A Posteriori (MAP) estimation using the robust bilateral total variation (BTV) as a regularization term as defined in [19]. This choice is motivated by the fact that the properties of a bilateral filter, namely, noise reduction while preserving edges, is now established as an appropriate method for depth data processing [12, 32, 33]. The BTV regularization is defined as follows:

$$\Gamma_{BTV}(\mathbf{x}_{t_0}) = \sum_{i=-l}^{i=l} \sum_{j=-l}^{j=l} \alpha^{|i|+|j|} \parallel \mathbf{x}_{t_0} - \mathbf{S}_x^i \mathbf{S}_y^j \mathbf{x}_{t_0} \parallel_1 . \tag{29}$$

The matrices $\mathbf{S}_x^i$ and $\mathbf{S}_y^j$ are shifting matrices that shift $\mathbf{x}_{t_0}$ by i, and j pixels in the horizontal and vertical directions, respectively. The scalar $\alpha \in]0, 1]$ is the base of the exponential kernel

UP-SR: Upsampling for Precise Super-Resolution

for t_0,
1. Choose the reference frame $\mathbf{y}_{t_0}$.

for t, *s.t.*, $t_0 - N + 1 \leq t \leq t_0$,
do
2. Compute $\mathbf{y}_t \uparrow$ using (13).
3. Estimate the registration matrices $\hat{\mathbf{M}}_t^{t_0}$ using (21).
4. Compute $\overline{\mathbf{y}}_t^{t_0} \uparrow$ using (20).

end do
end for
5. Find $\hat{\mathbf{z}}_{t_0}$ by applying a temporal median estimator (28).
6. Estimate $\hat{\mathbf{x}}_{t_0}$ by deblurring using (30).

end for

Table 1: Proposed UP-SR Algorithm

which controls the speed of decay [20].
The final solution is:

$$\hat{\mathbf{x}}_{t_0} = \underset{\mathbf{x}_{t_0}}{\operatorname{argmin}}\Big(\|\mathbf{B}\mathbf{x}_{t_0} - \mathbf{z}_{t_0}\|_1 + \lambda\Gamma_{BTV}(\mathbf{x}_{t_0})\Big), \tag{30}$$

where λ is the regularization parameter. The UP-SR algorithm is summarized in Table 1.

Because of the complexity of dynamic scenes with moving objects, the choice of the order of the reference frame $\mathbf{y}_{t_0}$ with respect to the frames used to super-resolve it plays a major role. Since we use a temporal median filter in fusing the registered depth frames, taking $\mathbf{y}_{t_0}$ to be in the middle is a natural choice to estimate the corresponding HR depth image $\mathbf{x}_{t_0}$.

6. Statistical Performance Analysis

In this section we derive the performance of the UP-SR algorithm in terms of mean square error (MSE) for a fixed noise level. This derivation helps in better understanding the effect of the number of frames N and the effect of the SR factor r on the performance of the UP-SR algorithm. In [34, 35], there have been some attempts to derive the asymptotic limits of SR. However, these attempts do not take into account the bias of an SR estimator, which is always part of an image reconstruction process [36]. Moreover, a Gaussian noise model is usually assumed while UP-SR exploits an additive Laplacian noise model [18]. Taking into account the considered problem, we propose to adapt the affine bias model of [37] based on a representation with patches, which leads to an approximation of the UP-SR bias. This bias is related to two main factors, namely, the error due to gradient-based motion estimation [36], and to the SR factor r. Few assumptions are introduced for simplicity of analysis but we will show that they hold in the experimental evaluation, both quantitatively and qualitatively.
Thanks to the new data model proposed in (26), we look into the performance of the median estimator $\hat{\mathbf{z}}_{t_0}$ as defined in (28) in terms of MSE. Let us define $\mathrm{tr}(\cdot)$ and $\mathrm{cov}(\cdot)$ to be the trace and the covariance functions, respectively. Then, the MSE may be decomposed into two parts; the $\mathrm{bias}(\cdot)$, and the variance $\mathrm{var}(\cdot)$, defined for a given vector $\mathbf{z}$ as $\mathrm{var}(\mathbf{z}) = \mathrm{tr}\,(\mathrm{cov}(\mathbf{z}))$. By considering a known ground truth $\mathbf{x}_{t_0}$, we may then express the MSE as follows:

$$\mathrm{MSE}\,(\hat{\mathbf{z}}_{t_0}, \mathbf{x}_{t_0}) = \mathrm{var}(\hat{\mathbf{z}}_{t_0}) + \|\mathrm{bias}\,(\hat{\mathbf{z}}_{t_0})\|^2. \tag{31}$$

6.1. Bias Computation

Chatterjee and Milanfar have proposed in [37] an affine bias model for image denoising. The processing is done on patches, thus making the model in [37] local. We have shown in Section 5 how the SR problem can be formulated as a denoising problem (26). We may therefore apply the model in [37] after some modifications to fit the estimation in (28).
We decompose the ground truth image $\mathbf{x}_{t_0}$ into n patches $\{\mathbf{q}_{t_0}(i), i = 1, \cdots, n\}$ where each patch $\mathbf{q}_{t_0}(i)$ is of size $(r \times r)$ pixels and centered at the pixel $\mathbf{x}_{t_0}(i)$. Similarly, the frames $\overline{\mathbf{y}}_t^{t_0} \uparrow$ are decomposed into n overlapping patches $\{\mathbf{p}_t(i), i = 1, \cdots, n\}$. In fact, the estimation in (28) corresponds to the process of locally selecting the element with the highest ranking among the N patches at the same position $\{\mathbf{p}_t(i), t = t_0 - N + 1, \cdots, t_0\}$. Let $\mathbb{E}(\cdot)$ be the expectation operator, and $\mathbf{I}_r$ the identity matrix of size $(r \times r)$. By considering two frames at different times t and t', we may calculate the local bias per patch as explained in [15] as follows:

$$\mathrm{bias}\,(\hat{\mathbf{q}}_{t_0}(i)) = \mathbf{S}_i\mathbf{q}_{t_0}(i) + \mathbf{u}_i, \tag{32}$$

with

$$\mathbf{S}_i = \left(\mathbb{E}\left(\mathbf{W}_{t_0}^{t'}(i)\right) - \mathbf{I}_r\right)\mathbf{q}_{t_0}(i),$$

and

$$\mathbf{u}_i = \mathbb{E}\left(\mathbf{W}_{t_0}^{t'}(i)\boldsymbol{\eta}_{t_0}(i) + \mathbf{w}_{t_0}^{t'}(i)\right),$$

where $\mathbf{W}_{t_0}^{t'}(i)$ and $\mathbf{w}_{t_0}^{t'}(i)$ are the sub-block of $\hat{\mathbf{M}}_{t_0}^{t'}$ centered at position i, and the local innovation directly related to cumulated innovations defined in (16), respectively. The vector $\boldsymbol{\eta}_{t_0}(i)$ represents the patch measurement error due to noise and to blur. The final bias is then defined as:

$$\|\mathrm{bias}\,(\hat{\mathbf{z}}_{t_0})\|^2 = \sum_{i=1}^{n} \|\mathrm{bias}\,(\hat{\mathbf{q}}_{t_0}(i))\|^2. \tag{33}$$

In the simple case where the average motion per patch and its innovation $\mathbf{w}_{t_0}^{t'}(i)$ are close or equal to zero, the per-patch bias term becomes $\mathbb{E}\,(\boldsymbol{\eta}_t(i))$. This bias is in fact due to the effects of the per-patch blur and to noise. The statistical properties of the noise are the same as those of $\boldsymbol{\nu}_t$. The blur effect is due to the $(r^2 - 1)$ pixels per patch generated by the upsampling step. Assuming that they induce a fixed mean error ρ, the total bias may be simplified as follows:

$$\|\mathrm{bias}\,(\hat{\mathbf{z}}_{t_0})\|^2 = \sum_{i=1}^{n} \|\mathbb{E}\,(\boldsymbol{\eta}_t(i))\|^2 = n \cdot (r^2 - 1)\rho^2. \tag{34}$$

We can see in (34) that, for $r = 1$, the estimation becomes unbiased. This is due to the fact that there is no blur caused by the

upsampling process. Generally, the bias term is data dependent because of $\mathbf{q}_{t_0}(i)$ in (32). It also depends on the SR factor r, and the local motions and noise. From (34), we conclude that the bias is proportional to the squared SR factor r^2 and to the image size n.

6.2. Variance Computation

Assuming that the noise $\boldsymbol{\nu}_t$ follows an i.i.d. n-multivariate Laplace distribution, we may write: $\text{var}(\hat{\mathbf{z}}_{t_0}) = \text{tr}\left(\text{cov}(\hat{\mathbf{z}}_{t_0})\right) = n \cdot \text{var}\left(\hat{\mathbf{z}}_{t_0}(i)\right)$, $i = 1, \cdots, n$. Therefore, we may define the variance as [38]

$$\text{var}\left(\hat{\mathbf{z}}_{t_0}(i)\right) = 2\sigma^2 f(N), \quad i = 1, \cdots, n, \tag{35}$$

where for N even,

$$f(N) = \frac{4N!}{\left(\left(\frac{N-1}{2}\right)!\right)^2}\left(\frac{1}{2}\right)^{\frac{N+1}{2}} \sum_{k=0}^{\frac{N-1}{2}} \frac{\binom{\frac{N-1}{2}}{k}\left(-\frac{1}{2}\right)^k}{(N+1+2k)^3}, \tag{36}$$

and for N odd,

$$f(N) = \frac{N!}{\left(\frac{N}{2}\right)!\left(\frac{N}{2}-1\right)!}\left(\frac{1}{2}\right)^{\frac{N}{2}}\Big(\frac{1}{N^3}\left(\frac{1}{2}\right)^{\frac{N}{2}} + \sum_{k=0}^{\frac{N}{2}-1}\binom{\frac{N-1}{2}}{k}\left(-\frac{1}{2}\right)^k \frac{7N^2+8N(k+1)+4(k+1)^2}{N^2(N+2k+2)^3}\Big). \tag{37}$$

Our model assumes that the effect of overlapping patches is expressed in the bias term. Thus, the variance is independent of r, which corresponds to the simple denoising operation where no SR is involved and $r = 1$. It is proportional to the noise variance σ^2 and to the number of measurements N. The Cramèr Rao bound corresponding to the variance in (35) is equal to $\frac{\sigma^2}{2N}$. Thus, for a very long sequence, where N tends to ∞, the variance $\text{var}(\hat{\mathbf{z}}_{t_0})$ tends to 0.

7. Experimental Results

In order to evaluate the performance of the UP-SR algorithm, we start by separately looking at the impact of the two key components, upsampling and cumulative motion estimation, designed to handle the motion of freely moving and deforming objects in depth LR videos. Then, we provide a quantitative evaluation comparing with state-of-the-art approaches by testing on synthetic data with ground truth. We give qualitative examples using the same synthetic data in addition to real data acquired in a laboratory environment. Finally, for different SR factors and varying noise levels, we compare the obtained results to the theoretical analysis given in Section 6.

7.1. Upsampling and Motion Estimation

To demonstrate the effect of the upsampling step on the motion estimation process, we conduct the following experiment. We consider the "Art" depth image from the Middlebury dataset [39]. We shift it with one pixel in both x and y directions at the resolution $r = 1$. As a result, the corresponding motion vector at a given scale $r = R$ is $\mathsf{w}^{L\uparrow^R} = (R, R)$ pixels, which represents the ground truth motion. In this experiment, we take $R = 8$. Next, we estimate motion vectors for different SR factors, i.e., r varying from 1 to R. These vectors are further upscaled with the factor $\frac{R}{r}$ in order to be compared with the motion ground truth $\mathsf{w}^{L\uparrow^R}$. The error of the estimated motion is calculated as follows: $\epsilon_r = \|\frac{R}{r} \cdot \mathsf{w}^{L\uparrow^r} - \mathsf{w}^{L\uparrow^R}\|_2$. The obtained results are shown in Table 2. They clearly support our claim where the error decreases by a factor of $\frac{1}{r}$ by increasing the SR factor r. We can see that estimating motion from upsampled images with the factor $r = R$ is more accurate than upscaling the estimated motion from the lowest level with $r = 1$.

	r=1	r=2	r=4	r=6	r=8
ϵ_r (pixels)	0.51	0.25	0.13	0.08	0.06
Gain in accuracy (%)	0%	50%	75%	84%	88%

Table 2: Errors ϵ_r between estimated motions upscaled with a factor of $(\frac{R}{r})$ with $r = 1, ..., R$, and estimated motions from upsampled frames with a resolution factor $R = 8$.

7.2. Cumulative Registration

To illustrate the effectiveness of the cumulative registration proposed in Section 5.2, we consider a challenging case of four persons moving with a large motion in different directions. The used setup is an LR ToF camera, the 3-D MLI [2], mounted in the ceiling and looking at the scene from the top. One of the LR frames is shown in Figure 2 (a). We apply the UP-SR algorithm on this sequence using three different registration techniques, namely, non-cumulative registration, cumulative registration using the upscaled motion vectors estimated from LR frames, and the proposed cumulative registration using the estimated motion from upsampled LR frames. The corresponding results are shown in Figure 2 (b), (c), and (d), respectively. They show the superiority of the third technique over the first two techniques, which confirms the advantage of using the proposed cumultative motion estimation.

7.3. Qualitative Comparison

We use the "Samba" dataset available in [40], which provides a real sequence of a 3-D dynamic scene with HR ground truth, Figure 3 (e). We downsample a sub-sequence of 9 LR frames with a scale factor $r = 4$. The obtained LR sequence is of resolution (256×147) pixels. This sequence is degraded with additive Laplacian noise with σ varying from 0 to 100 mm. The created LR noisy depth sequence is then super-resolved. In order to visually evaluate the performance of UP-SR, we plot in 3-D the super-resolved results of the "Samba"-generated sequence for the noise level of $\sigma = 30$ mm. As expected, the UP-SR algorithm provides a better result by keeping the fine details as compared to the bicubic interpolation and to the patch-based SISR methods. By zooming on the face part and plotting the 3-D error map, it is clear that UP-SR gives the closest result as compared to the ground truth, see Figure 3 for more details.

Using the same setup of the LR ToF camera mounted in the ceiling at a 2.5m height, we captured an LR depth video of two persons sitting on chairs sliding in two different directions. A

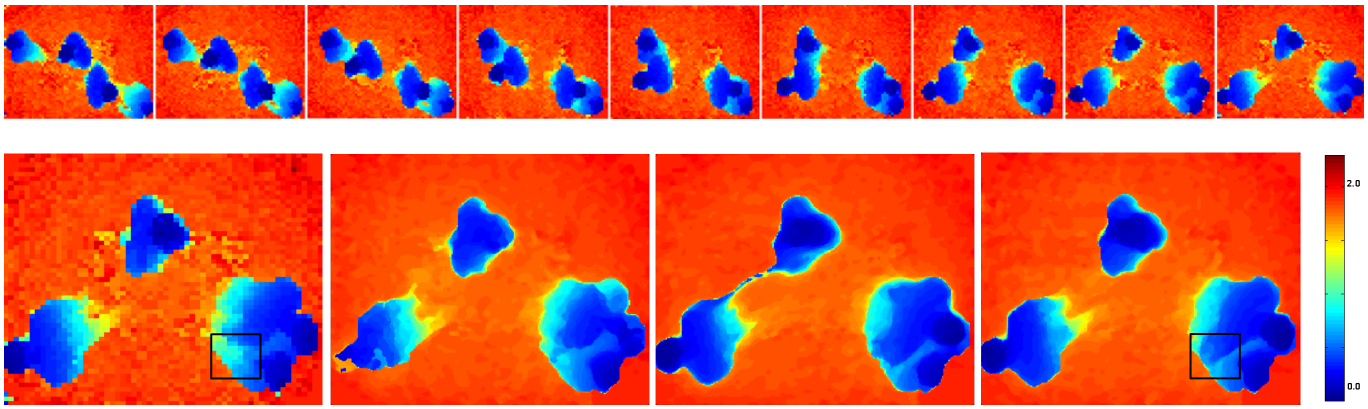

Figure 2: UP-SR results with $r = 4$ using different registration techniques of a dynamic scene with four persons moving in different directions. The sequence consists of 9 LR (56×61) depth images. (a) Last frame in the LR sequence. (b) UP-SR without cumulative motion. (c) UP-SR with cumulative motion upscaled from LR frames. (d) UP-SR with the proposed cumulative motion from upsampled frames. The largest measured depth in this scene is 2.5 m.

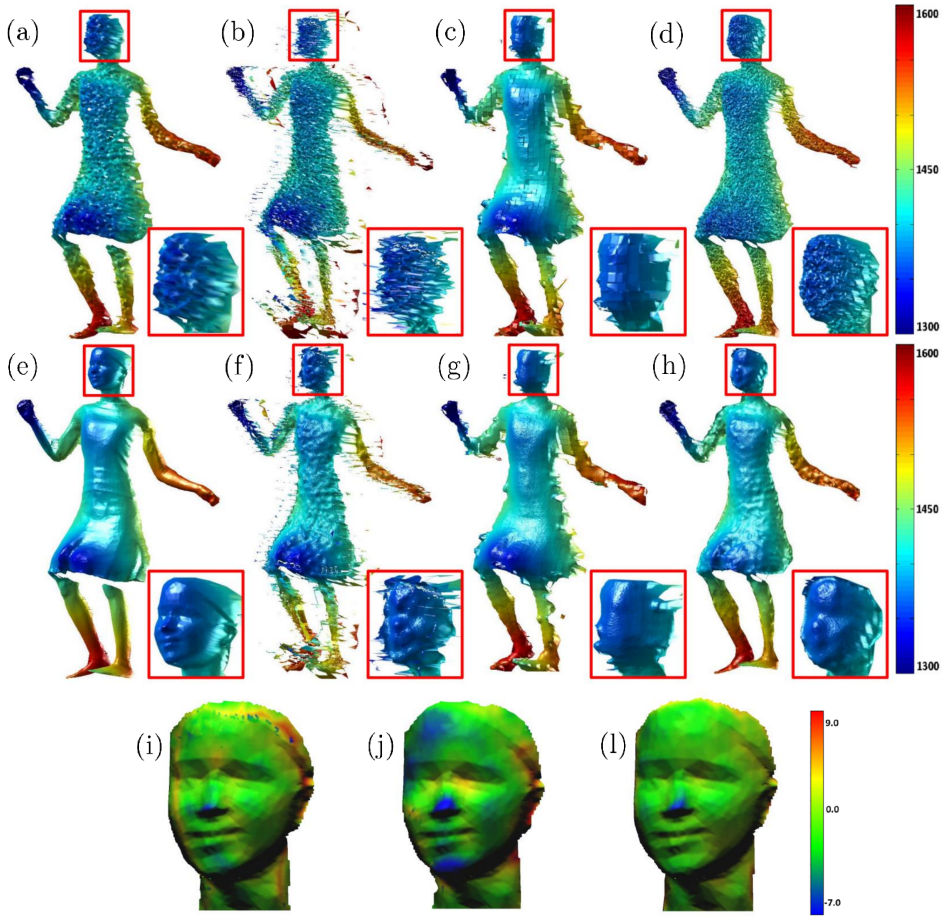

Figure 3: 3-D results of different SR methods applied on the "Samba" sequence [40]. (a) LR noisy input. (b) Bicubic interpolation. (c) Patch-based SISR [12]. (d) UP-SR, initial estimate. (e) Ground truth. (f) Deblurred bicubic. (g) Deblurred patch-based SISR. (h) Deblurred UP-SR. Third row represents the 3-D error maps for: (i) Bicubic. (j) Patch-based SISR. (l) Proposed UP-SR. We can see that the obtained error using the the proposed UP-SR (l) is quite small as compared to other methods where the bicubic interpolation leads to noisy depth measurements in addition to the flying pixels represented by the yellow and orange collors in the 3D error map in (i). The obtained results using the patch-based SISR is quite smooth and lead to removing fine details, and hence, resulting in large 3-D reconstruction errors, see blue patches in (j). The depth is measured in mm.

sequence of 9 LR depth images, of size (56×61) pixels, was super-resolved with an SR factor $r = 5$ using bicubic interpolation, 2-D/depth fusion [5], dynamic S&A [41], patch-based SISR [12], and the proposed UP-SR. Visual results for one

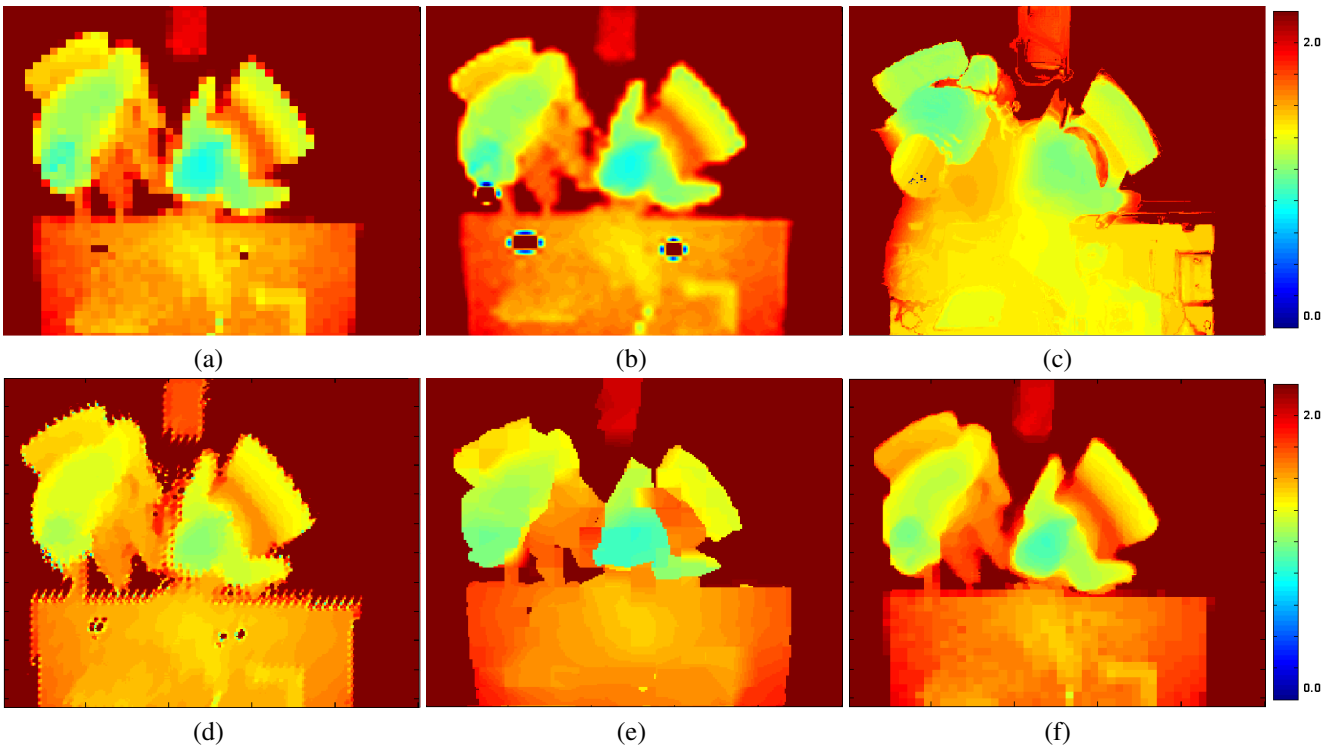

Figure 4: Moving chairs sequence: comparison of the results for different SR methods with SR factor of $r = 5$: (a) Last frame of 9 LR (56×61) depth images. (b) Bicubic interpolation of the last depth frame in the sequence. (c) 2-D/depth fusion [5]. (d) Dynamic S&A [41]. (e) SISR S&A [12]. (f) Proposed UP-SR.

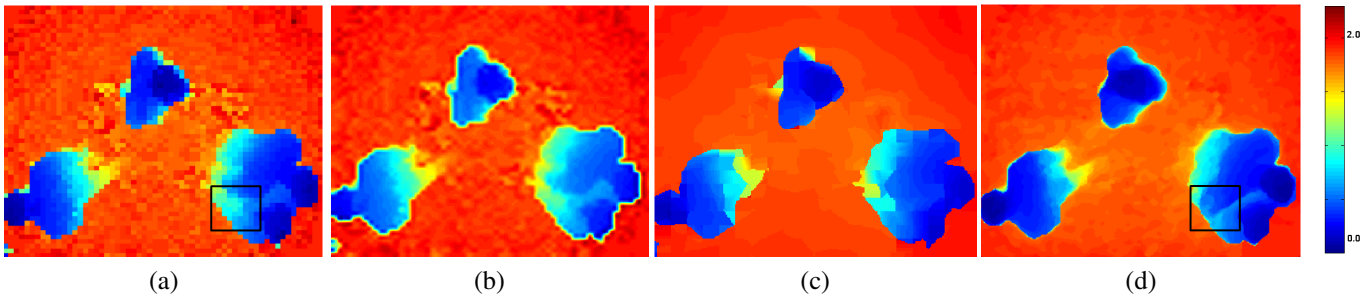

Figure 5: Comparison of the results for different SR methods with SR factor of $r = 4$. These methods are applied on a dynamic sequence of four persons with fast motion in different directions. (a) Last frame of LR (56×61) depth images. (b) Bicubic interpolation of the last depth frame in the sequence. (c) SISR [12]. (d) Proposed UP-SR.

frame are given in Figure 4 (b), (c), (d), (e), and (f), respectively. Obtained results show that bicubic interpolation and dynamic S&A fail on depth data mainly on boundary pixels, while the result of the 2-D/depth fusion suffers from strong 2-D texture copying on the final super-resolved depth frame as shown in Figure 4 (c). We can see the results of SISR in Figure 4 (e), where the inaccuracies are also observed especially on objects' boundaries. We show in Figure 4 (f) the result of the UP-SR algorithm where we obtained clear sharp edges in addition to an efficient removal of noisy pixel values. This is mostly due to the proposed sub-pixel motion estimation combined with an accurate cumulative registration leading to a successful temporal fusion of the sequence. Similar results are observed in Figure 5 by testing the different methods on the challenging case of the sequence of four moving persons.

7.4. Quantitative Comparison

We provide a quantitative evaluation of the proposed UP-SR algorithm as compared to two methods, namely, the conventional bicubic interpolation and the patch-based single image SR (SISR) given in [12]. We start with the "Samba" dataset, where the previously created LR noisy depth sequences are super-resolved using these methods and the proposed method. We compare the obtained results at two levels, initial and deblurred using the deblurring step proposed in Section 5. For the deblurring step we use an exhaustive search to find the best optimization parameters corresponding to the smallest 3-D reconstruction error. The quantitative results are reported in Figure 6. As expected, by applying the conventional bicubic interpolation method directly on depth images, a large error in the reconstructed HR depth image is obtained. This error is mainly due to flying pixels around object's boundaries, Figure 3 (b).

Thus, for a fair comparison we run another round of experiments using a modified bicubic interpolation, where we remove all flying pixels by defining a fixed threshold. Yet, the 3-D reconstruction error remains relatively high. This is due to the fact that bicubic interpolation does not profit from the temporal information provided by the sequence. Only in the case of one moving object and a very low noise level (less than 10 mm) the modified bicubic interpolation may be considered as shown by the red solid line in Figure 6. The performances of SISR, original and deblurred, are given in green lines, solid, and dashed, respectively. SISR can be seen to be robust to noise as its performance is stable even for high noise levels. The addition of the deblurring step of UP-SR improves the MSE of the original SISR algorithm. The result of the proposed UP-SR algorithm is shown with a blue dashed line. Its MSE is the lowest among all the tested methods, and is also shown to be robust across all noise levels. This result can be explained by the fact that SISR is a patch-based method where no temporal information is used in recovering the fine details even after applying a deblurring step. In contrast, the good quality of the UP-SR results is obtained thanks to the temporal fusion using the pixel-wise median filtering after a cumulative registration. This fusion plays a major role in attenuating the temporal noise and represents an appropriate process to deal with the problem of flying pixels. Moreover, the spatial deblurring step leads to further adding a smoothing effect while keeping sharp edges, hence, recovering fine details.

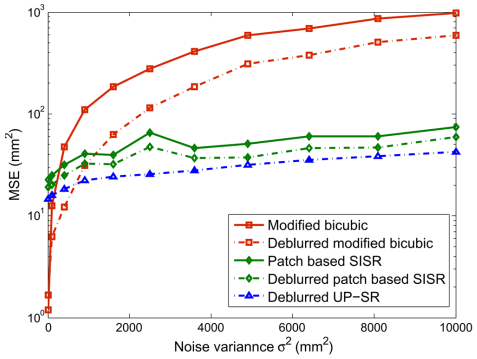

Figure 6: MSE at different noise levels for different SR methods applied to an LR depth sequence created from the "Samba" dynamic data [40], with $r = 4$ and $N = 9$.

7.5. Evaluation for Varying SR Factors

In order to evaluate the performance of the proposed UP-SR algorithm for different SR factors and varying noise levels, as compared to the statistical performance analysis of Section 6, we setup the following experiment. We use the publicly available toolbox V-REP [42] to create synthetic data with fully known ground truth of a laterally moving person with less complex motions as compared to the "Samba" dataset. Three depth cameras with the same field of view are fixed at the same position. These cameras are of different resolutions, namely, 512^2, 256^2, and 128^2 pixels. They are used to capture three sequences of the moving person. These sequences are further degraded with additive Laplacian noise with a standard deviation σ varying from 0 *mm* to 60 *mm*. Each sequence is super-resolved using UP-SR by considering 9 successive frames. The corresponding MSE performance of the first fusion step and the second deblurring step of UP-SR are reported in Figure 7 in solid and dashed lines, respectively. In the simple case where $r = 1$, the SR resolution problem is merely a denoising one where the ground truth is estimated from 9 noisy measurements. In other words, the objective is not to increase resolution, and hence there is no blur due to upsampling. Since consecutive motions between frames are small, they led to an approximately unbiased median estimation, which validates (34). Indeed, as seen in Figure 7,

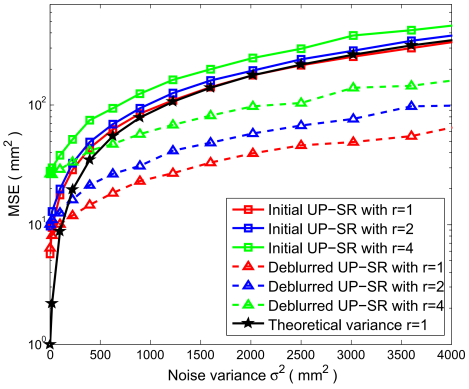

Figure 7: MSE versus noise variance for the V-REP simulated dynamic scene.

starting from $\sigma \approx 30$ *mm*, the solid red line overlaps with the dashed-dotted black line which corresponds to the theoretical variance obtained using (37). A non-zero bias is found for $r = 2$ and $r = 4$ where the corresponding blue and green solid lines are above the theoretical variance. This suggests a correlation between motion and upsampling blur as expressed by the vector $\mathbf{u}_i$ in (32). We note an increased bias for a larger SR factor r. This is justified by a larger blur effect due to the dense upsampling and to local motions. Finally, the dashed lines in Figure 7 confirm the performance enhancement after applying the optimization in (30); thus, ensuring an effective deblurring.

These quantitative results can be appreciated visually in Figure 8 where the noise level is fixed at $\sigma = 35$ *mm*. First, second and third columns correspond respectively to $r = 1$, $r = 2$, and $r = 4$ where (a), (b) and (c) are the noisy LR observations; (g), (h), and (i) are the result of UP-SR. The corresponding error maps as compared with the ground truth are given in (j), (k), and (l). The effective resolution enhancement, with a factor of 4, and the denoising power of UP-SR for a moving object on depth data is seen in 3-D in Figure 8 (i). The average root MSE in 3-D as shown in Figure 8 (l) is about 9 mm.

8. Discussion and Conclusions

A new multi-frame super-resolution algorithm for dynamic depth scenes has been proposed. It has been shown to be effec-

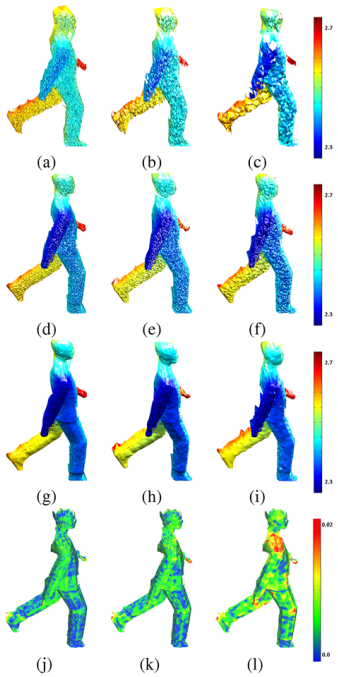

Figure 8: UP-SR qualitative results for the V-REP simulated moving person with different SR factors r. First, second and third columns correspond respectively to $r = 1$, $r = 2$, and $r = 4$ where (a), (b) and (c) are the noisy LR observations; (d), (e), and (f) are the result of the initialization step of UP-SR; (g), (h), and (i) are the result of the deblurring step of UP-SR. The corresponding error maps as compared to the ground truth are given in (j), (k), and (l)

tive in enhancing the resolution of dynamic scenes with one or multiple non-rigidly moving objects. The proposed algorithm relies on two main components; first, an enhanced motion estimation based on a prior upsampling of the observed low resolution depth frames up to the super-resolution factor. Second, it uses a cumulative motion estimation accurately relating non-consecutive frames in the considered depth sequence, even for relatively large motions. In addition, the multi-frame super-resolution problem has been reformulated defining a simplified data model which is analogous to a classical image denoising problem with additive Laplacian noise, and using multiple observations. This has led to a median initial estimate, further refined by a deblurring operation using a bilateral total variation as the regularization term. For a thorough understanding of the impact of the different parameters, namely, number of observed frames N and the super-resolution factor r, a statistical model for the proposed approach in terms of MSE has been derived. One important conclusion is that the blur effect is due to both upsampling, motion and occlusions. Extensive evaluations using synthetic and real data have been carried out, showing the consistent good performance of the proposed approach in full correspondence with the derived theoretical statistical model. We note, nevertheless, interesting limitations in the case, for example, of intersecting or touching objects, as can be seen within the bounding boxes in Figure 2 (d) and Figure 5 (d). This is due to the textureless nature of depth images which may cause two objects to be allocated to the same depth value, and hence makes them wrongly appear as one object. In the future, we will consider a full 3-D motion for a more accurate registration that should solve such ambiguous cases. Furthermore, we plan to investigate recursive approaches for a real time dynamic depth super-resolution. The results of this work are very novel as compared to state-of-art multi-frame super-resolution techniques applied to depth data. They are expected to have a significant impact in increasing the deployment of cost-effective low resolution depth cameras in many applications, such as, robotics, gaming, and security.

Appendix A. Proof of the Cumulative Motion Estimation

We prove by induction the following $\zeta(n)$ statement:

$$\begin{cases} \mathbf{M}^{t_0}_{t_0-n}\mathbf{y}_{t_0-n}\uparrow &= \overline{\mathbf{y}}^{t_0}_{t_0-n}\uparrow, \\ \text{s.t. } \mathbf{M}^{t_0}_{t_0-n} &= \mathbf{M}^{t_0}_{t_0-1}\mathbf{M}^{t_0-1}_{t_0-2}...\mathbf{M}^{t_0-n+1}_{t_0-n} \end{cases} \cdots \zeta(n).$$

Proof. Let us consider that $\zeta(n-1)$ is true, i.e.

$$\begin{cases} \mathbf{M}^{t_0}_{t_0-(n-1)}\mathbf{y}_{t_0-(n-1)}\uparrow &= \overline{\mathbf{y}}^{t_0}_{t_0-(n-1)}\uparrow, \\ \text{s.t. } \mathbf{M}^{t_0}_{t_0-(n-1)} &= \mathbf{M}^{t_0}_{t_0-1}\mathbf{M}^{t_0-1}_{t_0-2}...\mathbf{M}^{t_0-(n-1)+1}_{t_0-(n-1)} \end{cases} \quad (A.1)$$

From (A.1) we have:

$$\mathbf{M}^{t_0}_{t_0-(n-1)}\mathbf{M}^{t_0-(n-1)}_{t_0-n} = \mathbf{M}^{t_0}_{t_0-n} \quad (A.2)$$

Base case: When $n = 1$ we have

$$\mathbf{M}^{t_0}_{t_0}\mathbf{y}_{t_0}\uparrow = \overline{\mathbf{y}}^{t_0}_{t_0}\uparrow, \quad (A.3)$$

and

$$\mathbf{M}^{t_0}_{t_0}\mathbf{M}^{t_0-1}_{t_0} = \mathbf{M}^{t_0-1}_{t_0}. \quad (A.4)$$

Both (A.3) and (A.4) are verified because $\mathbf{M}^{t_0}_{t_0} = I_n$. Then,
Induction step: We need to show that $\zeta(n-1) \Rightarrow \zeta(n)$.
Given two consecutive frames: $\mathbf{y}_{t_0-n}$ and $\mathbf{y}_{t_0-(n-1)}$, we have:

$$\mathbf{M}^{t_0-(n-1)}_{t_0-n}\mathbf{y}_{t_0-n}\uparrow = \overline{\mathbf{y}}^{t_0-(n-1)}_{t_0-n}\uparrow, \quad (A.5)$$

where

$$\hat{\mathbf{M}}^{t_0-(n-1)}_{t_0-n} = \arg\min_{\mathbf{M}} \Psi\left(\mathbf{y}_{t_0-(n-1)}\uparrow, \mathbf{y}_{t_0-n}\uparrow, \mathbf{M}\right). \quad (A.6)$$

Multiplying (A.5) by $\mathbf{M}^{t_0}_{t_0-(n-1)}$ we find

$$\mathbf{M}^{t_0}_{t_0-(n-1)}\mathbf{M}^{t_0-(n-1)}_{t_0-n}\mathbf{y}_{t_0-n}\uparrow = \mathbf{M}^{t_0}_{t_0-(n-1)}\overline{\mathbf{y}}^{t_0-(n-1)}_{t_0-n}\uparrow. \quad (A.7)$$

From (A.2) and (A.7) we have

$$\mathbf{M}^{t_0}_{t_0-n}\mathbf{y}_{t_0-n}\uparrow = \overline{\mathbf{y}}^{t_0}_{t_0-n}\uparrow.$$

□

[1] M. Lindner, A. Kolb, "Compensation of motion artifacts for timeof-flight cameras," in Proc. Dynamic 3D Vision Workshop, vol. 5742, Jena, Sep. 2009, pp. 1627.
[2] 3D MLI, 2015, http://www.iee.lu/home-page.
[3] pmd CamBoard nano, 2015, http://www.pmdtec.com/products_ services/reference_ design.php.

[4] Q. Yang, R. Yang, J. Davis, D. Nister, "Spatial-Depth Super Resolution for Range Images," IEEE Int. Conf. Computer Vision and Pattern Recognition, vol., pp. 1-8 2007.

[5] F. Garcia, D. Aouada, B. Mirbach, T. Solignac, B. Ottersten, "Real-time Hybrid ToF Multi-Camera Rig Fusion System for Depth Map Enhancement," IEEE Int. Conf. Computer Vision and Pattern Recognition Workshops, vol., pp. 1-8, 2011.

[6] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," IEEE Int. Conf. Computer Vision and Pattern Recognition, vol., pp. 1-8, 2008.

[7] Q. Yang, K. Tan, B. Culbertson, J. Apostolopoulos, "Fusion of Active and Passive Sensors for Fast 3D Capture," IEEE Int. Workshops Multimedia Signal Processing, vol., pp. 69-74, 2010.

[8] "Super-Resolution Imaging", by Peyman Milanfar, in CRC Press, 2010.

[9] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," IEEE Computer Vision and Pattern Recognition Workshops, vol., pp. 1-7, 2008.

[10] S. Schuon, C. Theobalt, J. Davis, S. Thrun, "LidarBoost: Depth superresolution for ToF 3D shape scanning," IEEE Int. Conf. Computer Vision and Pattern Recognition., vol., pp.343-350, 2009.

[11] Y. Cui, S. Schuon, S. Thrun, D. Stricker, C. Theobalt, "Algorithms for 3D Shape Scanning with a Depth Camera," IEEE Trans. on Pattern Analysis and Machine Intelligence., vol.35, pp. 1039-1050, 2013.

[12] O. M. Aodha, N. Campbell, A. Nair, G. Brostow, "Patch Based Synthesis for Single Depth Image Super-Resolution," European Conf. on Computer Vision., vol. Part III, pp. 71-84, 2012.

[13] K. Al Ismaeil, D. Aouada, B. Mirbach, B. Ottersten, "Dynamic Super-Resolution of Depth Sequences with Non-Rigid Motions," IEEE Int. Conf. Image processing., vol., pp. 660-664, 2013.

[14] K. Al Ismaeil, D. Aouada, B. Mirbach, B. Ottersten, "Multi-Frame Super-Resolution by Enhanced Shift & Add", IEEE Int. Symposium on Image and Signal Processing and Analysis., vol., pp. 171-176, 2013.

[15] D. Aouada, K. Al Ismaeil, B. Ottersten, "Patch-based Statistical Performance Analysis of Upsampling for Precise SuperResolution," Int. Conf. on Computer Vision Theory and Applications, 2015.

[16] L. Xu, J. Jia, S. B. Kang, "Improving sub-pixel correspondence through upsampling," Journal on Computer Vision and Image Understanding., vol. 116, pp. 250-261, 2012.

[17] S. D. Babacan, R. Molina, and A.K. Katsaggelos. "Variational Bayesian Super Resolution". IEEE Trans. Image Process., vol. 20, pp. 984-999, 2011.

[18] S. Farsiu, D. Robinson, M. Elad, P. Milanfar, "Robust Shift and Add Approach to Super-Resolution," Int. Symposium on Optical Science and Technology., vol. 5203, pp.121-130, 2003.

[19] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Fast and Robust Multi-Frame Super-Resolution," IEEE Trans. Image Process., vol. 13, pp. 1327-1344, 2004.

[20] K. Al Ismaeil, D. Aouada, B. Mirbach, B. Ottersten, "Bilateral Filter Evaluation Based on Exponential Kernels," Int. Conf. on Pattern Recognition, pp. 258-261, 2012.

[21] R. Hardie, T. Tuinstra, K. Barnard, J. Bognar, and E. Armstrong, "High resolution image reconstruction from digital video with global and non-global scene motion." IEEE Int. Conf. Image Process., vol. 1, pp. 153-156, 1997.

[22] S. Farsiu, M. Elad, adn P. Milanfar, "Video-to-Video Dynamic Superresolution for Grayscale and Color Sequences," Journal on Advances in Signal Processing., 2006.

[23] A. W. M. van Eekeren, K. Schutte, J. Dijk, Dirk-Jan de Lange, and L. J. van Vliet, "Super-Resolution on Moving Objects and Background," IEEE Int. Conf. Image Process., vol., pp. 2709-2712, 2006.

[24] A. W. M. van Eekeren, K. Schutte, and L. J. van Vliet, "Multiframe Super-Resolution Reconstruction of Small Moving Objects," IEEE Trans. on Image Process., vol. 19, pp. 2901-2912, 2010.

[25] J. Y. Bouguet, "Pyramidal implementation of the Lukas Kanade feature tracker. Description of the algorithm". http://robots.stanford.edu/cs223b04/algo_tracking.

[26] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. " Hierarchical Model-Based Motion Estimation," European Conf. on Computer Vision., vol. 588, pp. 237-252, 1992.

[27] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Superresolution without Explicit Subpixel Motion Estimation," IEEE Trans. on Image Process., vol. 18, pp. 1958-1975, 2009.

[28] F. Garcia, D. Aouada, B. Mirbach, T. Solignac, B. Ottersten, "Spatio-Temporal ToF Data Enhancement by Fusion," IEEE Int. Conf. Image Process., vol., pp. 981-984, 2012.

[29] M. Elad and A. Feuer, "Super-Resolution reconstruction of Continuous Image Sequence," IEEE Trans. on Pattern Analysis and Machine Intelligence., Vol. 21, No. 9, pp. 817-834, 1999.

[30] D. Chan, H. Buisman, C. Theobalt, S. Thrun, "A Noise-Aware Filter for Real-Time Depth Upsampling", Workshop on Multi-camera and Multi-modal Sensor Fusion, 2008.

[31] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust Super-Resolution," IEEE Int. Conf. on Computer Vision and Pattern Recognition., vol. 1, pp. 645-650, 2001.

[32] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Andrew Fitzgibbon, "KinectFusion: Real-Time Dense Surface Mapping and Tracking", IEEE ISMAR, October 2011.

[33] K. Al Ismaeil, D. Aouada, B. Mirbach, B. Ottersten, "Depth Super-Resolution by Enhanced Shift and Add," 15th International Conference in Computer Analysis of Images and Patterns, 2013.

[34] A. Rajagopalan and P. Kiran, "Motion-free superresolution and the role of relative blur," J. Opt. Soc. Amer., vol. 20, pp. 20222032, 2003.

[35] D. Robinson, and P. Milanfar, "Statistical Performance Analysis of Super-resolution," IEEE Trans. on Image Processing., Vol. 15, pp. 1413-1428, 2006.

[36] Robinson and Milanfar, "Bias-minimizing filters for motion estimation," IEEE Asilomar Conf. on Signals, Systems and Computers., 2003.

[37] P. Chatterjee and P. Milanfar, "Bias modeling for image denoising," IEEE Asilomar Conf. on Signals, Systems and Computers., vol., pp.856-859, 2009.

[38] N.C. Beaulieu and S. Jiang, "ML estimation of signal amplitude in Laplace noise," IEEE Conf. Global Telecommunications., vol., pp. 1-5, 2010.

[39] http://vision.middlebury.edu/stereo/data/

[40] http://people.csail.mit.edu/drdaniel/mesh_animation/

[41] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and Challenges in Super-Resolution," Int. Journal of Imaging Systems and Technology, vol. 14, pp. 47-57, 2004.

[42] http://www.k-team.com/mobile-robotics-products/v-rep