Depth Super-Resolution by Enhanced Shift and Add*

Kassem Al Ismaeil[1], Djamila Aouada[1], Bruno Mirbach[2], and Björn Ottersten[1]

[1] Interdisciplinary Centre for Security, Reliability and Trust
Universtity of Luxembourg
{kassem.alismaeil,djamila.aouada,bjorn.ottersten}@uni.lu
[2] Advanced Engineering Department, IEE S.A.
bruno.mirbach@iee.lu

Abstract. We use multi-frame super-resolution, specifically, *Shift & Add*, to increase the resolution of depth data. In order to be able to deploy such a framework in practice, without requiring a very high number of observed low resolution frames, we improve the initial estimation of the high resolution frame. To that end, we propose a new data model that leads to a median estimation from densely upsampled low resolution frames. We show that this new formulation solves the problem of undefined pixels and further allows to improve the performance of pyramidal motion estimation in the context of super-resolution without additional computational cost. As a consequence, it increases the motion diversity within a small number of observed frames, making the enhancement of depth data more practical. Quantitative experiments run on the Middlebury dataset show that our method outperforms state-of-the-art techniques in terms of accuracy and robustness to the number of frames and to the noise level.

Keywords: Time-of-flight depth data, Super-resolution, Dense upsampling, Pyramidal optical flow, Motion diversity.

1 Introduction

The usage of depth data captured by time-of-flight (ToF) cameras is often limited because of its low resolution (LR). Most of the work proposed to enhance the resolution of this data has been based on fusion with high resolution (HR) images acquired with a second camera, e.g., 2D camera [1,2], stereo camera [3], or both 2D and stereo cameras [4]. These multi-modality methods provide solutions with undesired texture copying artifacts in addition to being highly dependent on parameter tuning. Moreover, using an additional camera requires dealing with data mapping and synchronization issues.

The super-resolution (SR) framework offers an alternative solution where an HR image is to be recovered from a set of LR images captured with the same

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

R. Wilson et al. (Eds.): CAIP 2013, Part II, LNCS 8048, pp. 100–107, 2013.

camera. The key idea is to explore the deviation between these LR images and a reference frame. SR techniques have been largely explored in the 2D case. The extension of these algorithms to depth data is not straightforward as presented in [5] where a dedicated preprocessing has been proposed to achieve depth SR from a single image, hence calling upon a heavy training. Earlier, the classical *Shift & Add* (*S&A*) was applied on depth data [6] in a multi-frame setup. While this work showed that SR may be used successfully on depth data without any training, it is still not a practical solution as it requires a large number of frames to ensure sufficient depth discontinuities between frames. An extended version has been proposed by the same authors in [7] by defining a new cost function dedicated to depth data. Both approaches in [6] and [7] do not solve the limitation on the number of frames inherent to classical *S&A*; thus, they remain unpractical solutions. In what follows, we show that this limitation goes back to the initialization step even before reaching to the iterative optimization step. Indeed, estimating the initial HR frame relies on LR frames only. In the case where the motion diversity within these frames is not sufficient, the initial estimate ends up with undefined pixels that affect the result of any iterative optimization. Therefore, we propose to estimate the initial HR frame from upsampled LR frames. By doing so, we ensure that no undefined pixels are present; moreover, we prove that, in the SR context, a more accurate motion estimation using pyramidal optical flow may be achieved resulting in an increased motion diversity within a smaller number of frames. In contrast to [13], this work is dedicated to depth data, where the upsampling has to be a dense one.

The remainder of the paper is organized as follows: Section 2 gives the classical SR formulation and a description of *S&A*. In Section 3, a new data model is provided leading to our proposed algorithm. Experimental results on the Middlebury dataset and on real ToF data are given in Section 4.

2 Motivation and Background

Let $\mathbf{X}$ be an HR depth image of size $(m \times n)$ and $\mathbf{Y}_k$, $k = 0, ..., (N-1)$, N observed LR images, where each LR image is of size $(\acute{m} \times \acute{n})$ pixels, such that $n = r \cdot \acute{n}$ and $m = r \cdot \acute{m}$, where r is the SR factor. Every frame $\mathbf{Y}_k$ may be viewed as a LR noisy and deformed realization of $\mathbf{X}$ caused by the ToF imaging system at the k^{th} acquisition. Considering $\mathbf{Y}_k$'s and $\mathbf{X}$'s respective lexicographic vector forms $\mathbf{y}_k$ and $\mathbf{x}$, the SR data model may be defined as follows:

$$\mathbf{y}_k = \mathbf{DHW}_k\mathbf{x} + \mathbf{n}_k, \qquad k = 0, ..., (N-1), \tag{1}$$

where $\mathbf{W}_k$ is an $(mn \times mn)$ matrix corresponding to the geometric motion between $\mathbf{x}$ and $\mathbf{y}_k$. In this framework, this motion is assumed to be global translational; hence, $\mathbf{W}_k$ represents a global shifting operator by u_k in x direction, and by v_k in y direction. The point spread function (PSF) of the ToF camera is modelled by the $(mn \times mn)$ space and time invariant blurring matrix $\mathbf{H}$. The matrix $\mathbf{D}$ of dimension $(\acute{m}\acute{n} \times mn)$ represents the downsampling operator, and

the vector $\mathbf{n}_k$ is the additive noise at k. Using the same approach as in [8], we consider that $\mathbf{H}$ and $\mathbf{W}_k$ are block circulant matrices. Therefore:

$$\mathbf{H}\mathbf{W}_k = \mathbf{W}_k\mathbf{H}. \tag{2}$$

Estimating $\mathbf{x}$ may thus be decomposed into two main steps: estimation of a blurred HR image $\mathbf{z}_0 = \mathbf{H}\mathbf{x}_0$, where $\mathbf{x}_0$ is an initial guess for $\mathbf{x}$, followed by a deblurring step by an iterative optimization. The classical *S&A* approach [8] defines $\mathbf{z}_0$ by first setting its corresponding full HR image grid $\mathbf{Z}_0$ to zeros, i.e., $\mathbf{Z}_0 = \mathbf{0}_{m\times n}$. Then, all LR images $\mathbf{Y}_k$ are used to update the pixel values in $\mathbf{Z}_0$. To that end, given a reference LR image $\mathbf{Y}_0$ chosen as the closest one to the target HR image $\mathbf{X}$, the global translational motions $\mathbf{w}_k = (u_k, v_k)$ between each image $\mathbf{Y}_k$ and $\mathbf{Y}_0$ are computed for $k = 1, \cdots, (N-1)$. These motions are used to register all LR images $\mathbf{Y}_k$ with respect to the reference image $\mathbf{Y}_0$. The resulting registered images $\overline{\mathbf{Y}}_k$ are simply defined at each pixel position $\mathbf{p} = (x, y)$ as follows:

$$\overline{\mathbf{Y}}_k(\mathbf{p}) = \mathbf{Y}_k(\mathbf{p} + \mathbf{w}_k). \tag{3}$$

These images are then grouped into M sets based on their relative motions $\mathbf{w}_k$. Note that to avoid aliasing problems, the range of this motion is forced to be within the SR factor r by a simple modulo function, i.e., $u_k = u_k \mathrm{mod}(r)$ and $v_k = v_k \mathrm{mod}(r)$. The frames in one set are fused by median filtering resulting in one LR image $\overline{\overline{\mathbf{Y}}}_i$ per motion $\mathbf{w}_i$, with $1 \leq i \leq M \leq N$. Each frame is then used to update the pixels of $\mathbf{Z}_0$ as follows:

$$\mathbf{Z}_0\,(r \cdot \mathbf{p} + \mathbf{w}_i) = \overline{\overline{\mathbf{Y}}}_i(\mathbf{p}). \tag{4}$$

This operation is known as zero filling in the *S&A* approach. We note that for a successful filling, there should be enough motion diversity in the considered LR frames. Indeed, in order to further update the zero pixels in $\mathbf{Z}_0$, an additional $(r \times r)$ median filtering is applied. Given that the median filter's breakdown point is $\frac{1}{2}$, a meaningful filling that does not leave pixels undefined should be achieved if the following condition is satisfied:

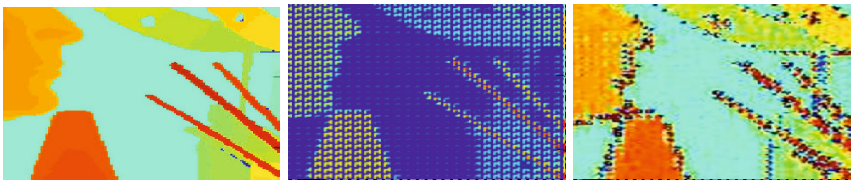
$$\mathrm{round}\left(\frac{r^2}{2}\right) \leq M. \tag{5}$$

As a second step, the estimation of $\mathbf{x}$ follows a maximum likelihood approach that, by assuming $\mathbf{n}_k$ as a Laplacian white noise, leads to the following minimization:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\mathrm{argmin}}\Big(\|\mathbf{H}\mathbf{x} - \mathbf{z}_0\|_1 + \lambda\Gamma(\mathbf{x})\Big), \tag{6}$$

where $\Gamma(\mathbf{x})$ is a regularization term added to compensate undetermined cases by enforcing prior information about $\mathbf{x}$, and λ being the regularization parameter.

Starting with an accurate initial guess $\mathbf{z}_0$ has a strong impact on the final solution of (6). We show the effect of undefined pixels in $\mathbf{Z}_0$ caused by classical *S&A* in Fig. 1(b). A similar phenomenon is observed using interpolation-based



(a) Original HR (b) *S&A* (c) *VBSR*

Fig. 1. Undefined pixels using state-of-the-art SR methods (Red colors are the closest objects and green colors are the furthest ones.)

initialization such as variational Bayesian SR (*VBSR*) [9] as seen in Fig. 1(c), suggesting that interpolation is not a sufficient solution to remove undefined pixels. Moreover, it creates additional artifacts on depth data such as jagged values on edges. It is common to face this serious problem of undefined pixels in practice. It is dealt with by restricting the SR factor to low values, e.g., $r = 2$, and by taking a relatively large number of frames, e.g., $N > 30$, thus indirectly attempting to satisfy inequality (5), which, in turn, limits the practical usage of SR algorithms on depth data. In what follows, our aim is to increase motion diversity M to give more freedom in the choice of r without having to increase N. We propose to tackle the aforementioned problem by a new non-zero initialization of $\mathbf{Z}_0$ as detailed in Section 3.

3 Proposed Algorithm

Estimating the motions $\mathbf{w}_k$ with high sub-pixel accuracy is crucial in capturing the full diversity in motion as contained in the observed LR depth frames; hence, important in increasing M. Indeed, for two frames $\mathbf{Y}_i$ and $\mathbf{Y}_j$ with respective relative motions $\mathbf{w}_i$ and $\mathbf{w}_j$, such that $\|\mathbf{w}_i - \mathbf{w}_j\|_2 = \epsilon$; if the motion estimation approach has an accuracy that is smaller than ϵ, the two frames will be wrongly fused and labeled under the same motion. Classical *S&A* uses pyramidal motion estimation (*PyrME*) [10,11]. This method represents state-of-art in motion estimation increasing both accuracy and robustness by a gain $\mathcal{G}(L) = 2^{(L+1)} - 1$, where L is the number of pyramidal levels [10]. In the case of SR, we note that the target resolution at which we want to land is the HR $(m \times n)$. This gives us a natural way to further improve the performance of *PyrME*. We thus propose to start by upsampling the LR depth frames up to the SR factor r prior to any motion estimation such that $\mathbf{y}_k \uparrow = \mathbf{U} \cdot \mathbf{y}_k$, where $\mathbf{U}$ is a dense upsampling matrix of size $(mn \times \acute{m}\acute{n})$. By doing so, we increase the size of the basis of the pyramid by a factor r. Changing the starting point in *PyrME* leads to an increased pyramid height $L \uparrow$ by $\log_2(r)$ which results in a new increased gain $\mathcal{G}(L \uparrow) = r \cdot \mathcal{G}(L) + (r - 1)$. This demonstrates that, in the SR context, the performance of *PyrME*, in terms of accuracy and robustness, may further be enhanced. Therefore, to estimate $\mathbf{w}_k$ more accurately, we now work with $\mathbf{Y}_k \uparrow$, $k = 0, \cdots, (N - 1)$, the N upsampled LR frames corresponding to the vectors

$\mathbf{y}_k \uparrow$. Performing the registration process as in (3) on the upsampled images $\mathbf{Y}_k \uparrow$ gives: $\overline{\mathbf{Y}}_k \uparrow (\mathbf{p}) = \mathbf{Y}_k \uparrow (\mathbf{p} + \mathbf{w}_k) \quad \Leftrightarrow \quad \mathbf{y}_k \uparrow = \mathbf{W}_k \overline{\mathbf{y}}_k \uparrow$.

It is easy to note that the new corresponding matrices $\mathbf{W}_k$ still verify (2). Furthermore, we choose $\mathbf{U}$ to be the transpose of $\mathbf{D}$, such that $\mathbf{UD} = \mathbf{A}$, where $\mathbf{A}$ is a block circulant matrix that defines a new blurring $\mathbf{B} = \mathbf{AH}$. Therefore, we redefine $\mathbf{z}_0$ as $\mathbf{z}_0 = \mathbf{Bx}_0$, and by left multiplying (1) by $\mathbf{U}$ we find:

$$\mathbf{y}_k \uparrow = \mathbf{W}_k \mathbf{Bx} + \mathbf{Un}_k, \qquad k = 0, ..., (N-1). \tag{7}$$

In addition, similarly to [12], for analytical convenience, we assume that all pixels in $\mathbf{y}_k \uparrow$ originate from pixels in $\mathbf{x}$ in a one to one mapping. Therefore, each row in $\mathbf{W}_k$ contains 1 for each position corresponding to the address of the source pixel in $\mathbf{x}$. This bijective property implies that the matrix $\mathbf{W}_k$ is an invertible permutation. By left multiplying (7) by $\mathbf{W}_k^{-1}$, we define the following new data model from upsampled registered observed LR frames:

$$\overline{\mathbf{y}}_k \uparrow = \mathbf{Bx} + \boldsymbol{\nu}_k, \qquad k = 0, ..., (N-1), \tag{8}$$

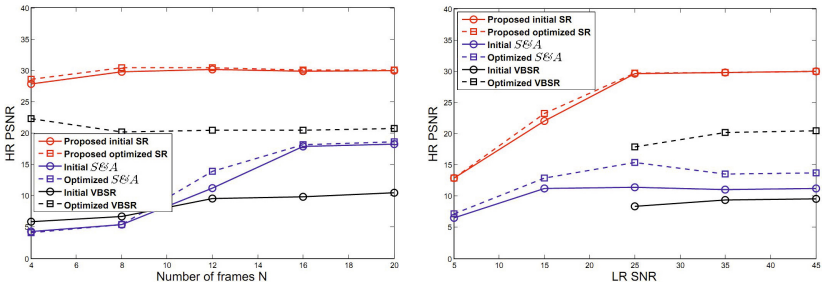
where $\boldsymbol{\nu}_k$ is an upsampled additive white Laplacian noise at k, leading to the following estimation of the initial guess:

$$\hat{\mathbf{z}}_0 = \arg\min_{\mathbf{z}_0} \sum_{k=0}^{(N-1)} \| \mathbf{z}_0 - \overline{\mathbf{y}}_k \uparrow \|_1. \tag{9}$$

The non-zero initialization in (9) releases the condition in (5), thus solving the problem of undefined pixels. In order not to fall under the same artifacts as those present with interpolation-based SR approaches, e.g., *VBSR* (Fig. 1(c)), it is necessary to perform the filling operation from registered and clustered LR images as in (4). Indeed, the values from LR frames remain more reliable sources of information than the ones due to upsampling. They are further processed by a (3×3) median filtering to smooth out noisy depth pixels. We point out that the higher accuracy in the estimation of $\mathbf{w}_k$ leads to a higher discrimination between motions, resulting in a higher diversity M and a better update of the pixel values in $\mathbf{z}_0$ as compared to the case of classical *S&A*. In our algorithm, it is more accurate to refer to this operation as initialization update rather than filling. After the new initialization and update step described above, a last deblurring step is performed to recover $\hat{\mathbf{x}}$ from $\hat{\mathbf{z}}_0$ using (6).

4 Experimental Results

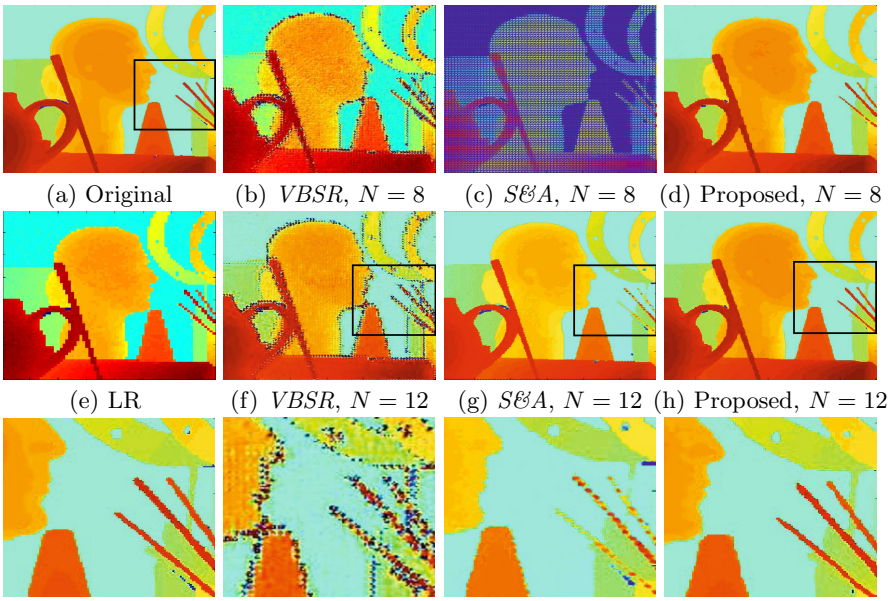
To evaluate the performance of the proposed algorithm, we tested its robustness on synthetic and real depth images against two parameters: number of considered LR images N, and image contamination with noise measured by signal to noise ratio (SNR). Each time we compare with two state-of-the-art SR methods, namely, classical *S&A* [8], and *VBSR* [9]. First, we ran Monte-Carlo simulations on synthetic sequences of a static scene subjected to a randomly



(a) Different N values (b) Different input SNR levels

Fig. 2. Mean PSNR values for different SR methods applied to a (75×75) LR sequence of a static depth scene with $r = 4$

generated global frame motion. These sequences were created by downsampling the HR image "ART" from the Middlebury dataset [10] with a factor $r = 4$, and PSF of $\sigma = 0.4$, and further degrading them by additive white Gaussian noise (AWGN). For a fixed noise level corresponding to SNR = 45dB, and 100 different realizations, Fig. 2(a) shows the average PSNR for N progressively increasing from 4 to 20 frames. It is clear that the proposed method outperforms both *S&A* and *VBSR* across different numbers of LR frames. This difference is even more noticeable for very low values of N, which illustrates the practicality of the proposed method. Next, we ran another round of experiments to evaluate the performance across different noise levels. In this experiment, a sequence of 12 (75×75) LR depth images was used. It was generated in the same way as in the previous experiment, and further degraded by AWGN with SNR of 5, 15, 25, 35 and 45dB. Fig. 2(b) shows that the proposed method is consistently more robust to noise. Furthermore, the textureless property of depth images combined with dense upsampling boost the performance of the proposed initial HR frame estimation, even for a very high noise level, e.g., SNR = 5dB, leading to comparable results before and after optimization with (6) as shown, respectively, with the dashed and continuous red lines in Fig. 2(a) and Fig. 2(b). This result suggests that the non-zero initialization may be considered as a standalone approach in the case of depth data as it does not deviate much from the assumptions related to the data model in (8). We give, in Fig. 3, an example of an HR estimated image of "ART" using 8 and 12 LR images in the first and second rows, respectively. Due to the condition (5), it is not surprising to see the artifacts caused by undefined pixels where the number of images is not sufficient to cover the motion range. Moreover, as seen in Fig. 3(d),(h), it is clear that our method provides the best visually enhanced HR depth images with sharper edges as compared to the results of *S&A* and *VBSR*. Finally, we tested the proposed algorithm on two real short depth sequences. The first sequence contains 8 LR depth images acquired using an IEE MLI ToF camera of resolution (56×61) pixels. The second sequence contains 5 LR frames acquired using a PMD CamBoard nano of resolution (120×165) pixels. Considering an SR factor of 4, the final results are given in Fig. 4, clearly showing that for these practical cases with a small N,

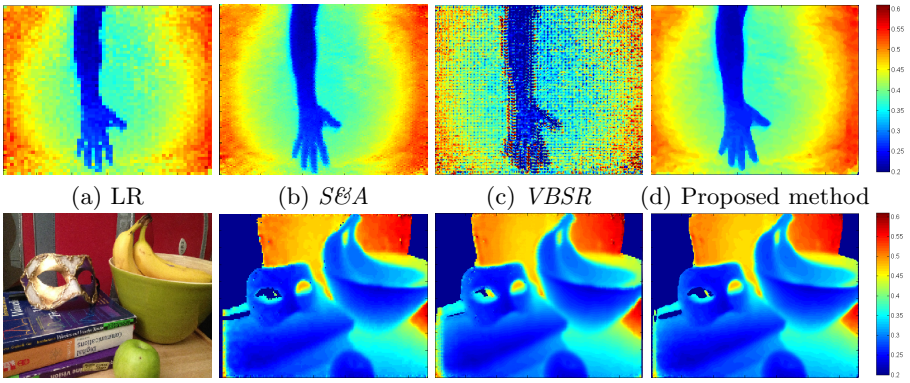


(a) Original (b) *VBSR*, $N = 8$ (c) *S&A*, $N = 8$ (d) Proposed, $N = 8$

(e) LR (f) *VBSR*, $N = 12$ (g) *S&A*, $N = 12$ (h) Proposed, $N = 12$

(i) (j) (k) (l)

Fig. 3. Results of different SR methods on a static ToF depth scene with different frame numbers ($N = 8, N = 12$) and SR factor of $r = 4$.)



(a) LR (b) *S&A* (c) *VBSR* (d) Proposed method

(e) 2D Image (f) LR (g) *S&A* (h) Proposed method

Fig. 4. Results of different SR methods on real LR ToF short sequences

the proposed method nicely super-resolves the LR frames by preserving edges and details while *S&A* and *VBSR* fail due to undefined pixels. Note that for the sake of practical deployment, to avoid any additional computal cost in the proposed method, the motion estimation from upsampled LR frames may be approximated by upscaling the corresponding LR motion vectors.

5 Conclusion

We proposed a practical SR solution for LR depth data acquired by ToF cameras. Our algorithm is based on a new SR data model that uses the upsampled and registered versions of the observed LR images. The benefits of this new formulation are twofold: It leads to a non-zero initialization of the estimate of the HR depth frame which solves the problem of undefined pixels inherent to classical SR techniques. Furthermore, it increases the accuracy and robustness of pyramidal motion estimation, which contributes in increasing the motion diversity within the observed frames. Both results help to reach good SR performances even in the challenging case of a relatively small number of LR frames, hence making the proposed algorithm usable in practice. While this method may be applied to 2D data, it is specifically designed to improve depth data thanks to a dense upsampling. Moreover, the textureless nature of depth data allows to use the proposed initialization step as a standalone algorithm where additional optimization is only required in the presence of high level non Laplacian noise.

References

1. Garcia, F., Aouada, D., Mirbach, B., Solignac, T., Ottersten, B.: Real-time Hybrid ToF Multi-Camera Rig Fusion System for Depth Map Enhancement. In: IEEE CVPRW 2011 (2011)
2. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-Depth Super Resolution for Range Images. In: IEEE CVPR 2007 (2007)
3. Zhu, J., Wang, L., Yang, R., Davis, J.: Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps. In: IEEE CVPR 2008 (2008)
4. Yang, Q., Tan, K., Culbertson, B., Apostolopoulos, J.: Fusion of Active and Passive Sensors for fast 3D Capture. In: MMSP 2010 (2010)
5. Mac Aodha, O., Campbell, N.D.F., Nair, A., Brostow, G.J.: Patch Based Synthesis for Single Depth Image Super-Resolution. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 71–84. Springer, Heidelberg (2012)
6. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: High-Quality Scanning Using Time-of-Flight Depth Superresolution. In: IEEE CVPRW 2008 (2008)
7. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: LidarBoost: Depth Superresolution for ToF 3D Shape Scanning. In: IEEE CVPR (2009)
8. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Fast and Robust Multi-Frame Super-Resolution. In: IEEE TIP 2003 (2003)
9. Babacan, S.D., Molina, R., Katsaggelos, A.K.: Variational Bayesian Super Resolution. In: IEEE TIP 2011 (2011)
10. Bouguet, J.Y.: Pyramidal Implementation of the Lukas Kanade Feature Tracker. Description of the Algorithm, `http://robots.stanford.edu/cs223b04/algo_tracking`
11. Bergen, J.R., Anandan, P., Hanna, K.J., Hingorani, R.: Hierarchical Model-Based Motion Estimation. In: Sandini, G. (ed.) ECCV 1992. LNCS, vol. 588, pp. 237–252. Springer, Heidelberg (1992)
12. Elad, M., Feuer, A.: Super-Resolution Reconstruction of Continuous Image Sequence. In: IEEE PAMI 1999 (1999)
13. Al Ismaeil, K., Aouada, D., Mirbach, B., Ottersten, B.: Multi-Frame Super-Resolution by Enhanced Shift & Add. In: IEEE ISPA 2013 (2013)