ICASSP 2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

Raw Data Processing for Practical Time-of-Flight Super-resolution

Miguel Heredia Conde

Center for Sensor Systems (ZESS), University of Siegen, Paul-Bonatz-Straße 9-11, 57076 Siegen, Germany. E-mail: heredia@zess.uni-siegen.de

Introduction

Time-of-Flight Imaging

- **Time-of-Flight** (ToF) imaging is an active imaging technology that aims to estimate distance from the delay experienced by an intensity-modulated optical signal.
- To this end, both modulated illumination and an array of pixels with demodulating capabilities are needed.
- **Continuous wave** (CW) operation: light is modulated and demodulated according to periodic (e. g., sinusoidal) functions of the same frequency.
- Let f_0 denote the base frequency, then the i^{th} frequency can be defined as $f_i = (i+1)f_0$, $\forall 0 \leq i \leq n_{\text{freq}}-1$. Similarly, for each f_i, n_{phase} phase steps are considered, where $\theta_j^\Gamma = \theta_0^\Gamma + 2\pi j/n_{\text{phase}}$, $\forall 0 \leq j \leq n_{\text{phase}}-1$. Then, in the sinusoidal case, the **raw measurements** for the ToF pixel channel $\Gamma \in \{\text{A}, \text{B}\}$ follow:

$$m_\Gamma[i,j] = \frac{\alpha_0 A_{\text{illu}} A_{p_\Gamma}}{2}\left(1+\cos\left(2\pi f_i\,(t-t_0)+\theta_j^\Gamma\right)\right) \tag{1}$$

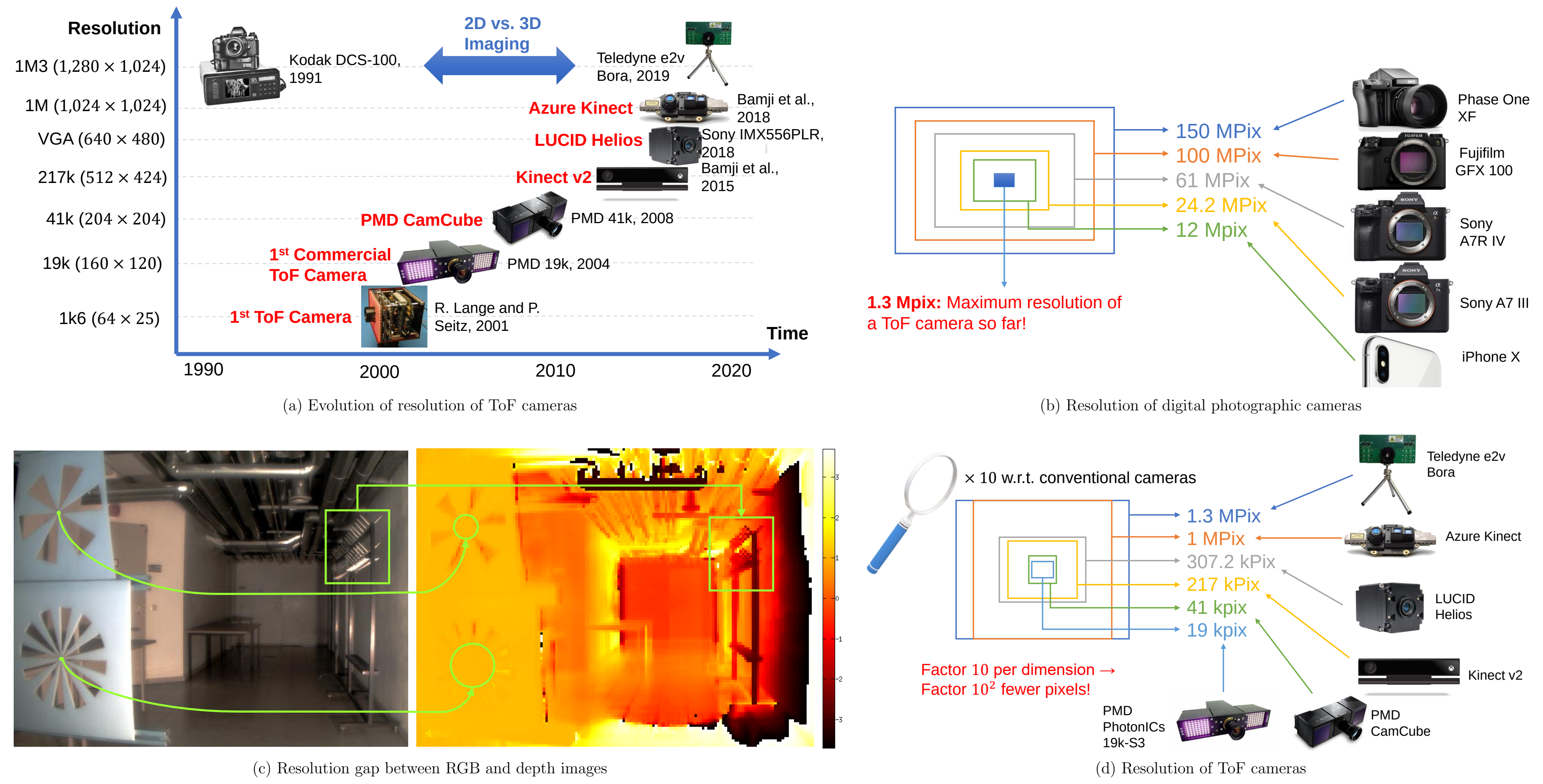
- Summing the measurements of both pixel channels obtained from a raw image acquisition yields an intensity image that is ideally independent from f_{mod} and θ:

$$I[i,j] = m_{\text{A}}[i,j] + m_{\text{B}}[i,j] = I,\ \forall i,j,\ \text{if}\ \mathcal{T} \coloneqq \{\text{A},\text{B}\} \tag{2}$$

- Two major limitations of ToF cameras w.r.t. lidar are:
  - **Power consumption:** ToF cameras need to densely illuminate the observed scene, as opposed to laser scanners, which perform punctual measurements.
  - **Motion artifacts:** ToF cameras require several raw image acquisitions to generate one depth image, each of them in the ms range, while lidar measurements take few tens of μs.
- The overarching idea of this work is that temporal oversampling leads to spatial super-resolution (SR).
- Existing multi-frame SR approaches exploit *inter*-frame motion, but neglect *intra*-frame motion!
- Multiple raw images from consecutive frames will be combined to obtain SR in raw image domain.
- From the super-resolved raw data, a high-resolution (HR) depth image is obtained using a *four phases algorithm*.

The ToF Resolution Gap

- The resolution of ToF cameras is one order of magnitude lower than conventional cameras. → $\times 10^{-2}$ fewer pixels.
- This means a delay of three decades w.r.t. digital photographic cameras (Fig. 1a) in terms of resolution.



(a) Evolution of resolution of ToF cameras

(b) Resolution of digital photographic cameras

(c) Resolution gap between RGB and depth images

(d) Resolution of ToF cameras

Figure 1: The resolution gap in ToF imaging. (a): A schematic representation of the evolution of the resolution of ToF sensors in the last two decades. (b): Relosution of commercial 3:2 digital photographic cameras, as compared to the highest-resolution ToF camera. (c): Illustration of the resolution gap between RGB (left) and depth (right) data from a ZESS MultiCam. (d): Resolution of ToF cammeras. The infography in (d) is scaled by a factor 10 w.r.t. that in (b).

Methodology

Super-resolution in Raw Image Domain

- **Basis:** Our SR framework is based on the image formation model in [1], consisting on *motion*, *blur*, and *downsampling*.
- Let the tuple $(x,y) \in \mathbb{R}^2$ denote the 2D spatial image domain and let the tuple $(u,v) \in \mathbb{N}^2$ denote the indices of the ToF pixels in the array, $0 \leq u \leq n_{\text{col}}-1$, $0 \leq v \leq n_{\text{row}}-1$, being $n_{\text{col}} \times n_{\text{row}}$ the array size. Equations (1) and (2) can be rewritten for the $(u,v)^{\text{th}}$ pixel of the k^{th} frame, yielding $m_\Gamma[i,j,k,u,v]$ and $I[i,j,k,u,v] = I[u,v]$.
- Then, in the noiseless case and neglecting atmospheric blur, the measurements obtained according to the model in [1] would be:

$$m_\Gamma[i,j,k,u,v] = \mathcal{D}\left(B(x,y) *^2 \mathcal{M}_{i,j,k}\left(m_\Gamma[i,j,k,x,y]\right)\right), \tag{3}$$

where $*^2$ denotes 2D-convolution, $B(x,y)$ is the blur kernel modeling the effective PSF of the ToF camera, $\mathcal{M}_{i,j,k}$ is a 2D-motion operator, and $\mathcal{D}$ the downsampling operator.

- The *fast and robust* SR methodology proposed in [1] consists on the following two consecutive steps:
  1. Non-iterative **data fusion**
  2. Iterative **deblurring**
- Let $\underline{Z}_\Gamma[i,j] \coloneqq \boldsymbol{B}\underline{m}_\Gamma^{\text{HR}}[i,j]$ be the blurred version of the HR image we aim to estimate in step 1, then we seek:

$$\hat{\underline{Z}}_\Gamma[i,j] = \arg\min_{\underline{Z}} \sum_{k=0}^{n_{\text{frame}}-1} \|\boldsymbol{DM}_{i,j,k}\underline{Z} - \underline{m}_\Gamma[i,j,k]\|_p^p, \tag{4}$$

which admits a closed-form solution for $p=1$ and $p=2$, namely, the *median* or the *mean* value of the *registered* images.

- For attaining accurate *registration*, the method in [2] is applied to the intensity images exploiting their ideal invariability across acquisitions, $I[i,j,k,u,v] = I[u,v]$. For a set of frequencies, $f_{\text{H}}, f_{\text{V}} \in \Omega^2$ (low-pass region), the unknown displacements, $\Delta x_{i,j,k}$ and $\Delta y_{i,j,k}$ are estimated via least squares from a set of equations of the shape:

$$2\pi \begin{bmatrix} f_{\text{H}} & f_{\text{V}} \end{bmatrix} \begin{bmatrix} \Delta x_{i,j,k} \\ \Delta y_{i,j,k} \end{bmatrix} = \arg\left(\frac{\mathcal{F}_{f_{\text{H}},f_{\text{V}}}\boldsymbol{I}[i,j,k]}{\mathcal{F}_{f_{\text{H}},f_{\text{V}}}\boldsymbol{I}[i_{\text{ref}},j_{\text{ref}},k_{\text{ref}}]}\right), \tag{5}$$

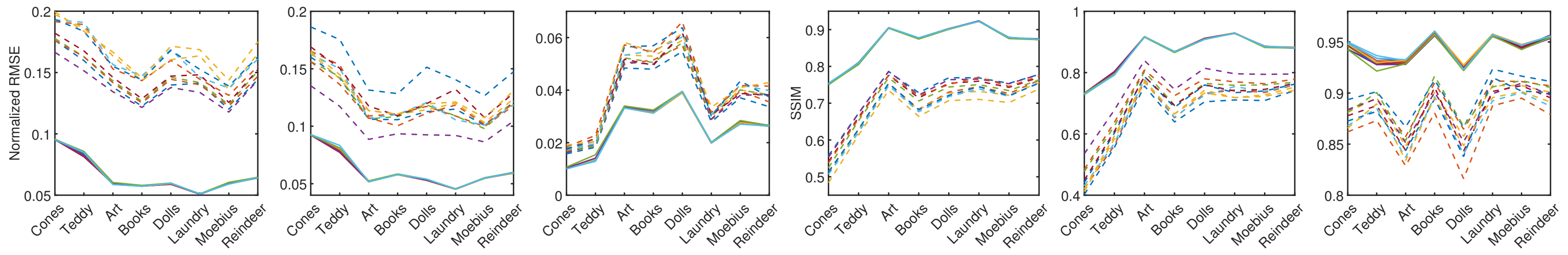
- Step 2 is then solved using the sparsity-aware *collaborative filtering* extension of BM3D introduced in [3].

Experimental Results and Conclusions

Evaluation with Synthetic Data

- **Datasets:** Middlebury stereo datasets from 2003 and 2005, providing RGB (unused, showed in Fig. 2) and disparity images for 8 complex scenes in total.
- For each scene, 15 frames of HR synthetic ToF raw data (2 pixel channels, 4 phases, 1 frequency) are generated according to (1).
- Raw images of both pixel channels are randomly 2D-shifted, up to ±5 pixels per spatial dimension in HR domain. 10 independent realizations per scene.
- The shifted HR raw data is blurred and downsampled by a factor 2 to generate the LR raw data.
- The proposed SR pipeline is applied to LR raw data (SR factor 2). HR depth images are obtained from the super-resolved raw data via the four phases algorithm.
- Fig. 3 plots the obtained RMSE and SSIM of the raw, amplitude, and depth super-resolved images and compares it to the RMSE and SSIM obtained applying bicubic interpolation. Fig. 4 compares the super-resolved depth maps to ground truth.

Figure 2: RGB images of the scenes from the Middlebury datasets of 2003 (first two images) and 2005 (rest).



(a) Raw Data RMSE

(b) Amplitude Image RMSE

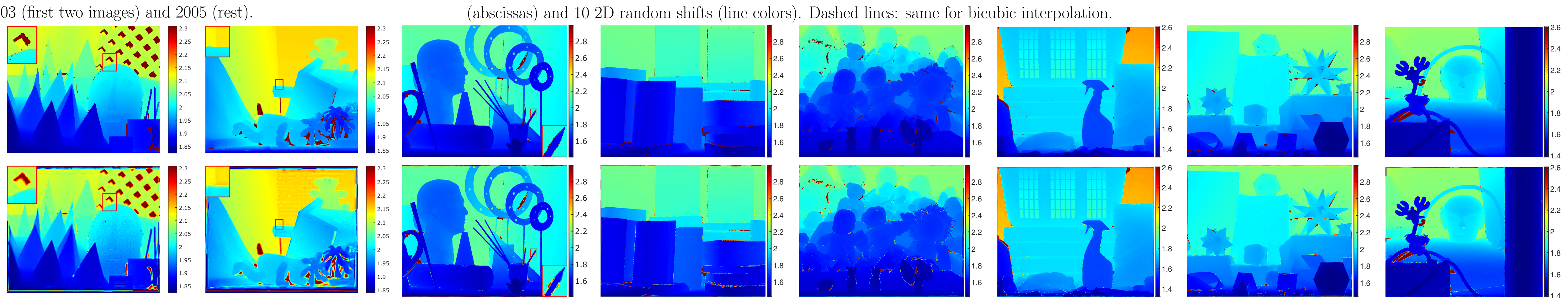
(c) Depth Image RMSE

(d) Raw Data SSIM

(e) Amplitude Image SSIM

(f) Depth Image SSIM

Figure 3: Solid lines: RMSE and SSIM of the super-resolved raw images and the resulting HR amplitude and depth images w.r.t. ground truth, for all 8 data sets (abscissas) and 10 2D random shifts (line colors). Dashed lines: same for bicubic interpolation.



(a) Cones

(b) Teddy

(c) Art

(d) Books

(e) Dolls

(f) Laundry

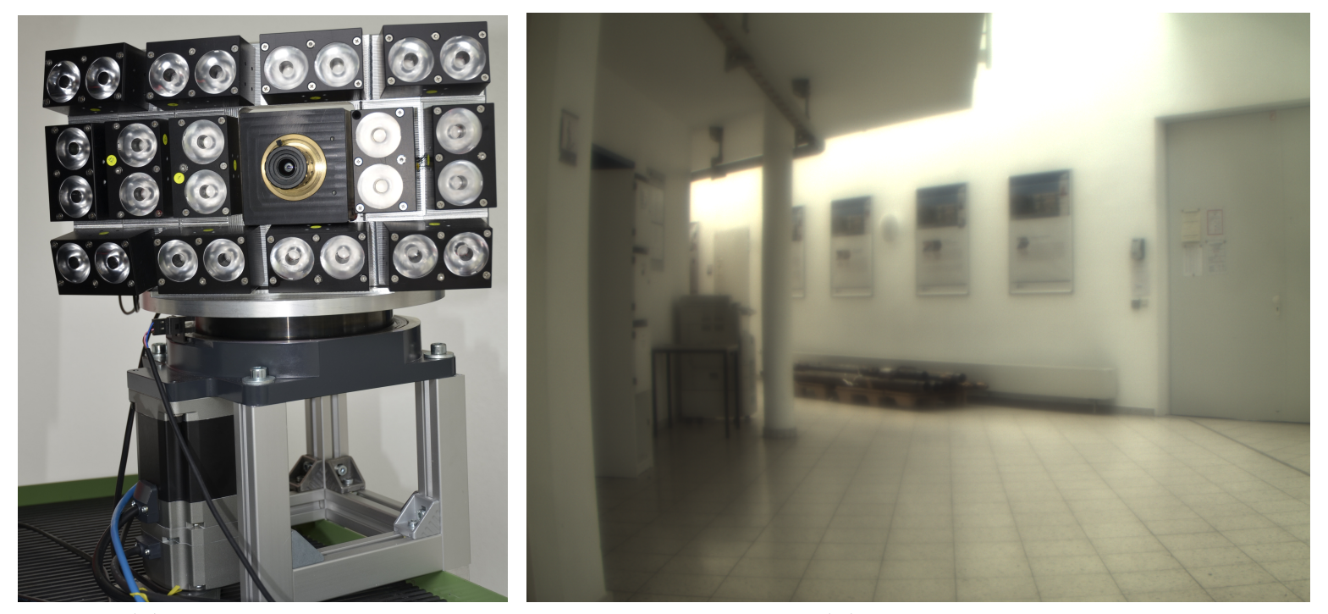
(g) Moebius

(h) Reindeer

Figure 4: Ground truth HR depth images obtained from the disparity maps of the Middlebury datasets of 2003 and 2005 (middle row), and corresponding depth images obtained from super-resolved raw data (bottom row).

Evaluation with Real Data

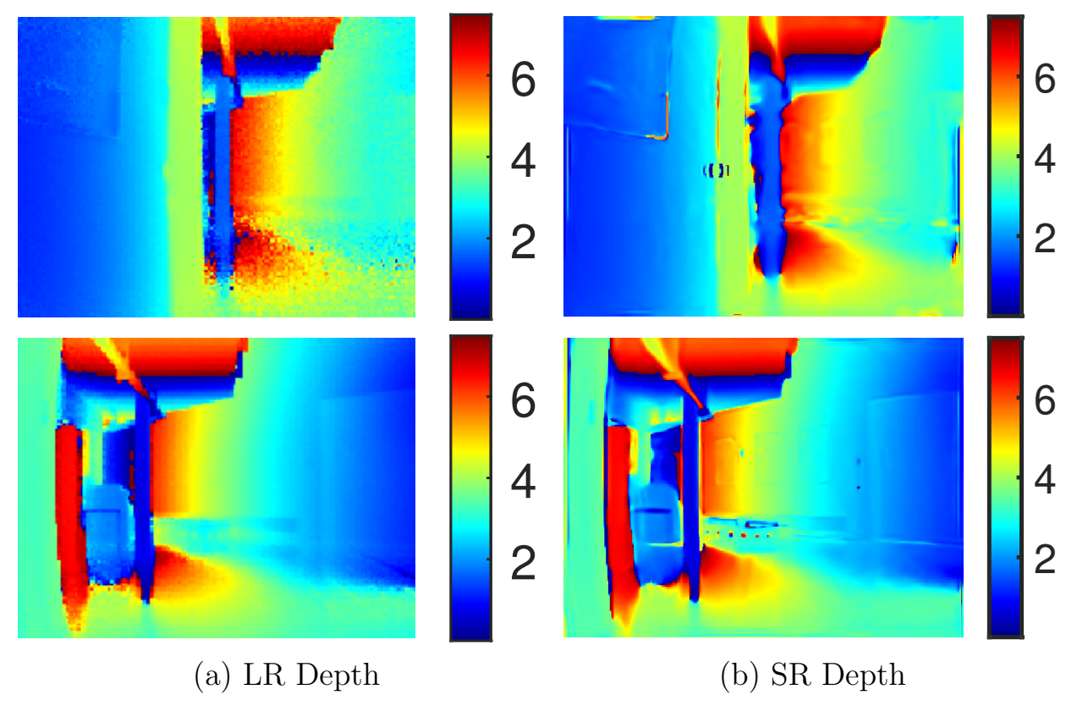
- **Hardware:** ZESS MultiCam with medium-range NIR (850 nm) LED illumination system mounted on a rotary table (Fig. 5a).
- **Modalities:** RGB (Aptina 3 Mpix sensor) + depth (PMD 19k-S3 chip, 160 × 120 pixels), single lens, fully registrable images.
- **Scene:** Hall of the ZESS building (Fig. 5b). **Datasets:** Two datasets were acquired using two different horizontal inter-frame displacements: 1.34 pix and 6.43×10^{-2} pix.
- For each dataset, 15 consecutive frames of ToF raw data (2 pixel channels, 4 phases, 1 frequency) were acquired. Depth SR and registration results are given in Fig. 6 and Fig. 7.



(a) ZESS MultiCam

(b) Test Scene

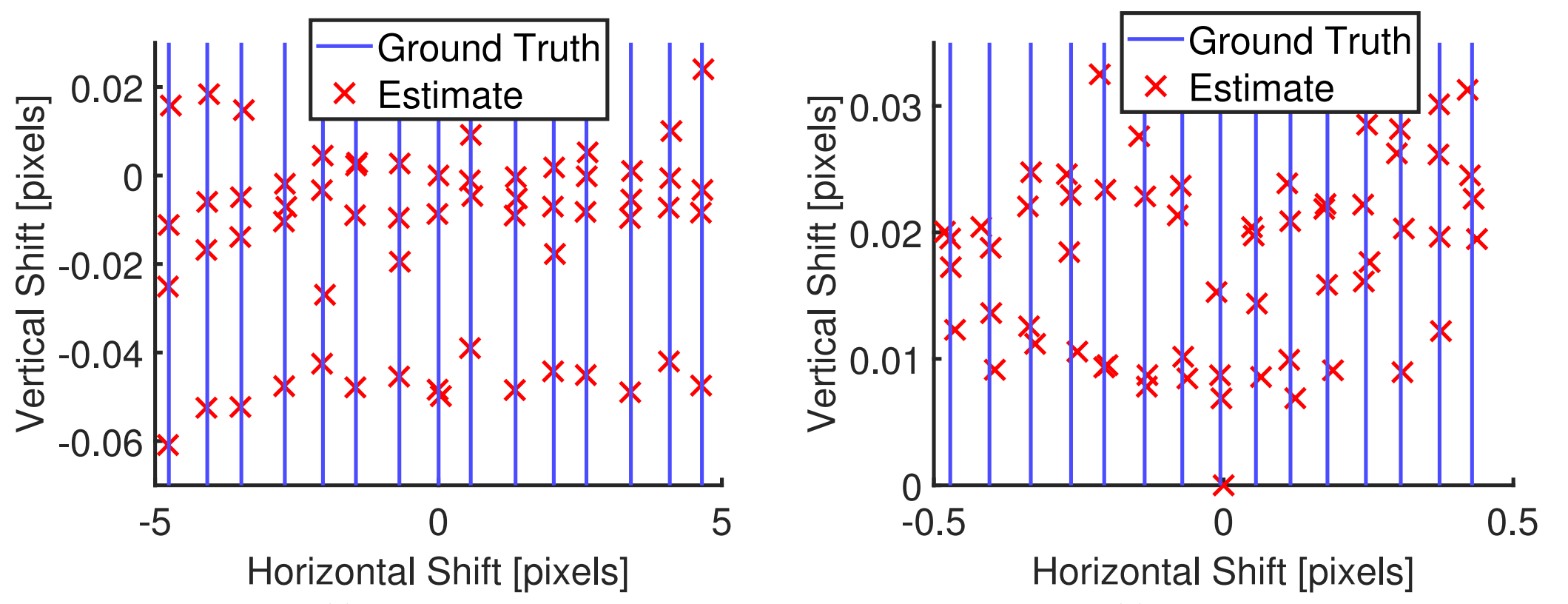
Figure 5: A ZESS MultiCam (a) was used to acquire real (LR) raw data. The observed scene was the hall of the ZESS biulding, shown in (b), with a depth range exceeding 16.5 m.



(a) LR Depth

(b) SR Depth

Figure 6: Single-frame LR depth images (a) vs. reconstructed HR depth images from LR real raw data (b). All scales are in meters. Fine detail and sharp borders become visible in the images in (b).



(a) Dataset 1

(b) Dataset 2

Figure 7: Estimated 2D displacements for the two real datasets considered. The ordinates in (b) witness registration accuracy in the order of 10^{-3} pixels.

Conclusions

- **ToF** cameras can retrieve 3D geometry, but its resolution is an order of magnitude lower than conventional 2D cameras, compromising data fusion.
- Existing multi-frame SR methods ignore *intra*-frame motion and operate directly on depth images.
- We have presented a SR framework that works on ToF raw data and accounts for both *inter*- and *intra*-frame motion.
- The proposed modular framework not only allows for seamless integration and benchmarking of new methods for the separable tasks of *raw data fusion* and *deblurring*, but enables detecting which task may constitute a bottleneck for the overall performance.
- Experiments on both synthetic and real ToF raw data from challenging scenes demonstrated the good performance of the proposed approach.

References

[1] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004. [Online]. Available: http://people.duke.edu/ sf59/srfinal.pdf

[2] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 2006: 071459, pp. 1–14, 2006. [Online]. Available: http://infoscience.epfl.ch/record/33804

[3] Y. Mäkinen, L. Azzari, and A. Foi, "Collaborative filtering of correlated noise: Exact transform-domain variance for improved shrinkage and patch matching," *IEEE Transactions on Image Processing*, vol. 29, pp. 8339–8354, 2020. [Online]. Available: http://www.cs.tut.fi/ foi/papers/Ymir-Collaborative_Filtering_of_Correlated_Noise-TIP.pdf

More information on related projects and publications of the author: