TRANSIENT DICTIONARY LEARNING FOR COMPRESSED TIME-OF-FLIGHT IMAGING

Miguel Heredia Conde

Center for Sensor Systems (ZESS), University of Siegen, Paul-Bonatz-Str. 9-11, 57076 Siegen, Germany

ABSTRACT

Time-of-Flight imaging aims to retrieve the 3D geometry of a scene from the delay that a modulated light waveform experiences when interacting with the former. Multi-path interference, arising from translucent objects or concave geometries, poses a challenge when the problem is to be solved from few measurements. In this work, we step aside from mainstream deep learning methods to invert the problem and propose exploiting underlying sparsity in an appropriate basis, in combination with *compressive sampling* schemes. More specifically, we show that the temporal response functions of real-life scenes are of bounded complexity and can be sparsely represented in a learned dictionary. A variety of sparse dictionary learning techniques are considered to find appropriate time-domain bases. Uniform frequency-domain sampling is compared to random sampling schemes and sparse rulers. Simulations acknowledge the superiority of non-uniform sampling and confirm that both transient profiles and millimeter-accurate depth images can be successfully reconstructed from few measurements.

***Index Terms*—** Time-of-Flight, depth imaging, transient imaging, dictionary learning, non-uniform sampling

1. INTRODUCTION

Time-of-Flight (ToF) cameras are active imaging sensors able to retrieve the 3D geometry of the environment. To this end, the scene is flood-illuminated with modulated light and the backscattered light is collected by an array of pixels with *demodulating* capabilities, referred to as ToF pixels. The conjunction of a time-domain modulated probing signal and an array of detectors with time-domain demodulation capacity yields temporal resolution. Different from conventional imaging sensors, which are blind to the transient response of the scene, measurements from a ToF camera effectively sample the latter. Exploiting the crucial fact that the distance or *depth* of an object is encoded as a time-delay in the backscattered photons, ToF cameras are able to retrieve depth images of the scene.

However, accurately retrieving the depth from the raw ToF measurements may become challenging in real-world conditions, as the light reaching a ToF pixel might be the superposition of the direct retroreflection from the target plus light returning along the same direction as a result of more complex light transport effects. This situation is commonly denominated as Multipath Interference (MPI) and may include both diffuse components and reflective components. A large number of works have aimed to compensate the effect of MPI on the depth estimate and even retrieve the depth corresponding to multiple return paths [1, 2, 3, 4]. Approaches aiming for compensation leveraging light transport models based on the scene geometry [1, 3] are too slow for real-time processing. Techniques relying on parametric spectral estimation methods [2, 4] are specially attractive, for being both accurate and outstandingly fast. However, these methods rely on the hypothesis that the time-domain *scene response function*, $h_{\mathrm{sce}}(t)$, also known as *transient* response, is sparse, that is, it can be modelled as a sum of scaled and shifted Dirac delta functions. As a matter of fact, this situation does not hold in general. For instance, concavity of objects and scene geometries often yields dense transient profiles due to multiple interreflections [5].

Waiving the sparsity constraint on h_{sce} has two negative consequences: first, it disables the use of parametric inversion and brings the need for substitutive regularization constraints and, second, it increases the number of measurements required to retrieve h_{sce}. Methods attempting to solve MPI for a general h_{sce} for an array of ToF pixels enter into the realm of *transient imaging*. In fact, arrays of ToF pixels have already been exploited to this end [6, 7, 5]. Measurements at different frequencies are obtained from ToF pixels operating in Continuous Wave (CW) mode, such as those based on the Photonic Mixer Device (PMD) [8]. The overarching idea is that CW-ToF pixels are able to deliver approximate samples of h_{sce} in frequency domain. Alternatively, binary codes with good autocorrelation properties, i. e., wide frequency spread, can be used as modulation/demodulation functions instead of sinusoids, in order to retrieve h_{sce} [9].

Recently, deep learning methods have been proposed to cope with the effect of MPI [10, 11, 12, 13, 14]. As the problem formulation merges, in its most general version, with transient imaging, recent works on MPI estimation and compensation have resorted to deep learning models to reconstruct a discrete version of the transient function, h_{sce}, and simultaneously deliver an MPI-free depth estimate [15, 16]. A common denominator to all of them is the underlying hypothesis that the transient images can be accurately reconstructed from compact representations in a *low-dimensional latent space*. This justifies the multiple architectures with U-shaped and encoder-decoder structures. A question that remains unanswered if whether one can find a representation basis to efficiently encode the transient profiles. Analytical expressions for the shape of h_{sce} have been discussed in the literature [17, 16], but without proof of optimality in a given sense.

In this work, inspired by recent approaches aiming to use transient profiles as a proxy for MPI-free depth estimation [15], we propose learning optimal frames that allow for sparse representations of such functions. We capitalize on the fact that an underlying sparse representation opens the door for a *compressive sensing* (CS) [18, 19] formulation, in which the sparse coefficients may be retrieved from a reduced number of measurements. Furthermore, and adopting a CS perspective on the reconstruction problem, we revisit the sampling scheme to be implemented in the frequency domain. Most related work requiring a set of measurements at different frequencies either adopt a uniform sampling grid or select the frequencies to be coincident with those implemented in commercial CW-ToF cameras, without further analysis. However, [20] has shown that non-uniform frequency sampling may enable a reduction of the number of measurements required to solve the MPI problem, while retaining accuracy. Consequently, we will evaluate both uniform and non-uniform frequency sampling schemes in a CS setting.

2. RELATED WORK

As briefly commented in section 1, recent work on MPI compensation has stepped away from both explicitly modelling light transport or aiming for closed-form solutions derived from simple models of h_{sce}. Instead, deep learning has been the tool of choice for solving this inverse problem. In the following we comment representative works of this trend that serve as a basis for this work.

In [10], a neural network (NN) is proposed that accepts a depth image affected by MPI as input and delivers an MPI-corrected depth image as output. The neural network model is, initially, a convolutional autoencoder, which is subject to a later process of training as supervised decoder, fixing the encoder side and adding skip connections at multiple scales, for the task of MPI compensation. Instead of correcting a depth image affected by MPI, in [11], a NN is proposed to predict a corrected depth image from dual-frequency raw ToF data. The network covers three functions in an aggregated fashion: denoising, phase unwrapping, and MPI compensation. The architecture combines a symmetrically skip-connected encoder-decoder generator network with a patchGAN discriminator network. Similarly to [10], skip connections are contemplated at multiple scales. Similarly to [11], the two-stage NN architecture introduced in [12] also operates on multi-frequency raw ToF measurements and corrects them for posterior depth estimation. The first stage is an encoder-decoder architecture aiming to compensate for motion artifacts. The second stage consists of a kernel-predicting network that aims to jointly compensate the effect of MPI and shot noise. In [13], the input data are depth and amplitude images obtained at three frequencies. The actual data plugged to the network are ratios between amplitudes and differences between unwrapped depths at different frequencies The network learns the MPI correction in depth domain. A convolutional NN (CNN) architecture is employed. In [14] work, in parallel to a coarse-to-fine CNN trained on synthetic data in a supervised manner, a discriminator CNN is trained in parallel in a GAN fashion to perform unsupervised pixel-level domain adaptation to real-world data.

A more ambitious alternative than correcting the effect of MPI is to train a NN to reconstruct transient images from ToF measurements, that is, a discrete version of h_{sce} for each pixel. Then, single- and multiple-path depth estimation boils down to appropriate peak detection in the transient profiles. This is the approach presented in [15]. More specifically, the NN proposed in [15] performs frequency interpolation and extrapolation, followed by Inverse Fourier Transform (IFT) and filtering to reconstruct the transient profiles. The NN architecture is a U-net with skip connections. In [16], a three-stage NN model is proposed. The first stage aims to cancel the effect of zero-mean temporal noise sources, such as shot noise. The second stage estimates the direct component, which encodes the depth, and the third one reconstructs the transient profiles.

The aforementioned works implicitly rely on the fact that the transient profiles can be effectively represented in a low-dimensional space [15]. Parametric models have been suggested to model h_{sce}. For instance, exponentially-modified Gaussian functions were used to model the transient responses of highly-scattering media in [17]. In [16], Rayleigh and Weibull distributions were found to bear resemblance to common shapes of h_{sce}. The authors chose the Weibull distribution to encode the global component of h_{sce}, being the direct component implicitly represented as a Dirac delta function. However, the subspace or union of subspaces that best represent the transient profiles remain largely unexplored. Proposed analytical models have shown good performance, but have not been proven optimal in terms of representation efficiency and accuracy. In this work we suppose an underlying linear subspace (or union thereof) and propose a data-driven optimization of this representation. The resulting frames allow for a sparse representation of h_{sce} and pave the way for a CS formulation [18].

3. METHODOLOGY

In this section we present how we model the sensing process, seek an efficient representation of the transient functions, and finally formulate the problem inversion from a CS perspective.

3.1. Sensing Model

Without loss of generality and for coherence with prior work, we focus on correlation-based ToF cameras operated in CW mode, such as those based on PMD technology [8]. Regardless of the operation mode, omitting the pixel indexing for notation simplicity, measurements obtained from a given ToF pixel can be modeled as samples of the convolution

$$y\left(\tau\right) = h_{\mathrm{cam}} * h_{\mathrm{sce}}(\tau), \tag{1}$$

where $h_{\mathrm{cam}}(t)$ is the instrument response function of the ToF camera in time domain. Differently from conventional cameras, where $h_{\mathrm{cam}}(t)$ is fixed, ToF cameras allow certain programmability of $h_{\mathrm{cam}}(t)$, which yields time resolution. More specifically, in CW-ToF it is customary to adopt the sinusoidal model

$$h_{\mathrm{cam},k}(t;\theta) = A\cos\left(2\pi f_k t - \theta\right), \tag{2}$$

where A is the amplitude, f_k, $1 \leq k \leq K$ denotes the modulation frequency and θ is a configurable phase shift. It is this configurable parameter that allows for sampling over time shifts in (1). Clearly, by acquiring measurements using two values of θ with $\pi/2$ separation, a complex Fourier coefficient, $\mathcal{F}_{f_k}\left(h_{\mathrm{sce}}\right)$, at the selected frequency f_k can be obtained. Thus, a set of measurements at appropriately chosen frequencies may provide an accurate representation of h_{sce}. Provided that h_{sce} is a real function, we restrict the model to the real numbers and assume that the in-phase and quadrature components are extracted for $\theta = 0$ and $\theta = \pi/2$, respectively. This is coherent with PMD-based ToF modules, which use multiples of $\pi/2$, and with the dataset in [15]. Discretizing the $h_{\mathrm{sce}}(t)$ for some fine time step, t_{step}, yields a vector $\vec{h}_{\mathrm{sce}} \in \mathbb{R}^n$, with $T = nt_{\mathrm{step}}$ the considered time domain. This also allows, in combination with (1) and (2), for the following sensing model:

$$\begin{aligned} y_k^{\Re} &= \vec{\phi}_k^{\Re\top}\vec{h}_{\mathrm{sce}}, \quad \text{with } \vec{\phi}_k^{\Re\top}[i] = A\cos\left(2\pi f_k i t_{\mathrm{step}}\right) \\ y_k^{\Im} &= \vec{\phi}_k^{\Im\top}\vec{h}_{\mathrm{sce}}, \quad \text{with } \vec{\phi}_k^{\Im\top}[i] = A\sin\left(2\pi f_k i t_{\mathrm{step}}\right) \end{aligned} \tag{3}$$

Now, stacking together the measurements, $y_k^{\Re}$, $y_k^{\Im}$ obtained for all K frequencies yields the following aggregated sensing model:

$$\vec{y} = \boldsymbol{\Phi}\vec{h}_{\mathrm{sce}} \quad \text{with} \quad \boldsymbol{\Phi} \coloneqq \begin{bmatrix} \boldsymbol{\Phi}^{\Re} \\ \boldsymbol{\Phi}^{\Im} \end{bmatrix}, \quad \begin{aligned} \boldsymbol{\Phi}^{\Re} &\coloneqq \left[\vec{\phi}_k^{\Re\top}\right]_{k=1}^{K} \\ \boldsymbol{\Phi}^{\Im} &\coloneqq \left[\vec{\phi}_k^{\Im\top}\right]_{k=1}^{K} \end{aligned} \tag{4}$$

The simultaneous requirements of high depth resolution and long range translate into a large n, whereas K is limited by frame rate requirements. Thus, $2K \ll n$ and the linear system in (4) is heavily underdetermined.

3.2. Learning an Optimal Representation

As mentioned in previous sections, the complexity of the transient responses is limited and, therefore, they are expected to admit an efficient representation in an appropriate subspace or union of subspaces. In this work we consider that there exists a collection of representation vectors, $\vec{\psi}_i \in \mathbb{R}^n, \leq i \leq N$, which constitutes a frame or *dictionary* where the signal $\vec{h}_{\text{sce}}$ admits an s-sparse representation with $s \ll n$. For generality, we consider the case of an overcomplete dictionary, i. e., $N > n$. Grouping the dictionary atoms together in a matrix, $\boldsymbol{\Psi}$, the linear model in (4) can now be written in terms of the unknown sparse vector of coefficients $\vec{x} \in \mathbb{R}^N$:

$$\vec{y} = \boldsymbol{\Phi} \underbrace{\boldsymbol{\Psi}\vec{x}}_{\vec{h}_{\text{sce}}} = \boldsymbol{A}\vec{x}, \quad \text{with} \quad \boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}, \quad \boldsymbol{\Psi} \coloneqq \left[\vec{\psi}_i\right]_{i=1}^{K} \tag{5}$$

Let the matrix $\boldsymbol{H}_{\text{sce}} = \left[\vec{h}_{\text{sce},i}\right]_{1\leq i\leq M}$ denote a representative set of M transient profiles, where $M > N$. Finding an optimal dictionary for $\boldsymbol{H}_{\text{sce}}$ means solving the following sparse dictionary learning problem:

$$\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{\Psi},\boldsymbol{X}} \|\boldsymbol{H}_{\text{sce}} - \boldsymbol{\Psi}\boldsymbol{X}\|_F^2, \quad \text{s.t.: } \|\vec{x}_i\|_0 \leq s_{\max}, \ \forall i, \tag{6}$$

where $\boldsymbol{X} = [\vec{x}_i]_{1\leq i\leq M}$, $\|\cdot\|_F$ denotes Frobenius norm, and $s_{\max}$ is an hypothesized upper bound for s. We select the following five alternatives to solve (6): the Method of Optimal Directions (MOD) [21], K-SVD [22], the approximate K-SVD in [23], the Online Dictionary Learning (ODL) approach in [24], and the Reweighted Least Squares Dictionary Learning Algorithm (RLS-DLA) from [25]. For all methods, the same tight frame is used as seed and Orthogonal Matching Pursuit (OMP) as sparse approximation method for speed.

3.3. Compressive Sensing Formulation

Once a dictionary $\hat{\boldsymbol{\Psi}}$ has been learned by solving (6), the sparsity constraint on $\vec{x}$ can be leveraged to uniquely retrieve it from $\vec{y}$. The inverse problem to solve becomes a linearly-constrained ℓ_0 minimization:

$$\hat{\vec{x}} = \arg\min_{\vec{x}} \|\vec{x}\|_0 \quad \text{subject to} \quad \vec{y} = \boldsymbol{A}\vec{x} \tag{7}$$

Let Σ_k denote the set of all k-sparse vectors in $\mathbb{R}^N$, then uniqueness of the solution can only be ensured if $\mathcal{N}(\boldsymbol{A}) \cap \Sigma_{2s} = \emptyset$, where $\mathcal{N}(\cdot)$ denotes nullspace. Thus, given $\vec{y} \in \mathbb{R}^m$, the condition $\text{spark}\,(\boldsymbol{A}) > 2s$ is sufficient for uniqueness, where $\text{spark}\,(\boldsymbol{A}) \coloneqq \min_{\vec{x}} \|\vec{x}\|_0$ s.t. $\boldsymbol{A}\vec{x} = 0$. Provided that $\boldsymbol{A} \in \mathbb{R}^{m\times N}$ with $m \ll N$, the minimum number of measurements required to recover $\vec{x}$ is $m \geq 2s$, which means $K \geq s$ frequencies. In other words, K scales linearly with s and not with n. It remains to define how these K frequencies are selected. Inspired by [20], apart from uniform sampling (US), we consider three non-uniform sampling (NUS) alternatives: random sampling, both subject to a grid and gridless, and sparse rulers. We consider *optimal* rulers when they exist for a given K and *perfect* rulers otherwise. For coherence with the dictionary learning stage, we use OMP to solve (7). From the obtained $\hat{\vec{x}}$ and $\hat{\boldsymbol{\Psi}}$, an estimate of the transient profile can be retrieved as $\hat{\vec{h}}_{\text{sce}} = \hat{\boldsymbol{\Psi}}\hat{\vec{x}}$. Finally, an MPI-immune depth estimate is obtained from $\hat{\vec{h}}_{\text{sce}}$ via peak detection, which in its simplest form reads:

$$\hat{d} = \frac{c}{2\,(\hat{i}\,t_{\text{step}})}, \quad \hat{i} = \arg\max_i \hat{\vec{h}}_{\text{sce}}[i] \quad \text{s.t.} \quad \hat{\vec{h}}_{\text{sce}}[i] > \epsilon, \tag{8}$$

where c is the speed of light and ϵ is an amplitude threshold to discard negligible signals. In practice, multiple peaks may be detected to deal with reflective MPI and flying pixels at depth edges [15].

4. EXPERIMENTS AND RESULTS

In this section we evaluate the methodology in section 3. For constructing $\boldsymbol{H}_{\text{sce}} \in \mathbb{R}^{n\times M}$, we make use of the set of 25 transient images of different indoor scenes from the iToF2dToF dataset [15]. Their length is $n = 2000$, with $t_{\text{step}} = 33.333\,\text{ps}$, spanning $T = 66.666\,\text{ns}$. This accounts for light paths of up to 20 m length, yielding an unambiguous depth range of 10 m. Each transient image contains 120×160 pixels, yielding 1.92×10^4 transient profiles per scene. For notation simplicity, we use the position in which they appear in the dataset as ID for the scenes. Similarly to [15], we consider only 21 scenes for training, reserving four (IDs 3, 10, 12, 14) for posterior validation purposes. From the $> 4 \times 10^5$ available profiles for training, we randomly select $M = 10^5$ and use them to train a dictionary ($\Psi \in \mathbb{R}^{n\times N}$ in (6)) with $N = 8000$ atoms using the five methods referred in section 3.2. This yields an overcompleteness factor of 4, coherently with prior work [24]. Pixels with negligible signals are excluded. OMP is set to enforce $s_{\max} = 16$ during training. Table 1 provides the normalized sparse reconstruction RMSE and execution times for the five learning methods considered. The intended paralellization potential of ODL was exploited by running 32 parallel threads in independent cores, thus the inferior execution time. The second fastest alternative is the approximate version of K-SVD, which scores an RMSE only marginally superior than K-SVD. The latter provides the best performance in terms of RMSE.

4.1. Sparse Representation of Transient Profiles

The relatively low RMSEs shown in Table 1 witness that the training data could be sparsely represented with $s \leq s_{\max} = 16$ nonzero coefficients in the learnt dictionaries. In this section we aim to provide statistical characterizations of both the accuracy of the sparse representations and the error that they produce when used to calculate the depth estimate. Additionally, we study the dependency of the error metrics with $s_{\max}$. The statistical characterization considers four non-overlapping error percentiles: 0-75%, 75-85%, 85-95%, and 95-99%. The accuracy of the sparsely reconstructed $\hat{\vec{h}}_{\text{sce}} = \hat{\Psi}\vec{x}_s$, where $\vec{x}_s \in \Sigma_s$ is the s-sparse approximation of $\vec{h}_{\text{sce}}$ in $\hat{\Psi}$, is evaluated in terms of RMSE with respect to the ground truth (GT) $\vec{h}_{\text{sce}}$. For computing depth errors, GT depth images are generated by applying multiple-peak detection, followed by heuristic selection, to the original profiles in all transient images. This avoids false depth errors arising from border effects present in the GT depth images contained in the dataset. For coherence with prior work, the Mean Absolute Error (MAE) is adopted to evaluate depth accuracy. RMSEs for the sets of $\hat{\vec{h}}_{\text{sce}}$ and MAEs for the depth images are provided in Tables 2 and 3, respectively, for the worst-performing scene in the validation set (IDs 12). Despite K-SVD yielded minimal training RMSE (cf. Table 1), MOD seems to generalize best, consistently yielding minimal RMSE (bold) in Table 2 and minimal depth MAE (bold) in Table 3 for all percentiles. The last row of Table 3 compares to the best depth MAEs in Table II of [15].

Despite all dictionaries were trained enforcing $s \leq s_{\max} = 16$, it is of interest to study how the representation accuracy and the resulting depth error vary with the s used to obtain $\vec{x}_s \in \Sigma_s$. The corresponding results, for the two best-performing dictionaries and $s \in [8, 24]$, are presented in Fig. 1, both in terms of RMSE of the

Method	MOD	K-SVD	Approx. K-SVD	ODL	RLS-DLA
RMSE [a.u.]	1.098×10^{-2}	1.016×10^{-2}	1.405×10^{-2}	2.280×10^{-2}	5.866×10^{-2}
Time [s]	8.457×10^{3}	2.376×10^{4}	1.498×10^{4}	6.543×10^{3}	2.428×10^{5}

Table 1: Normalized sparse reconstruction errors and training times obtained for each of the dictionary learning methods considered.

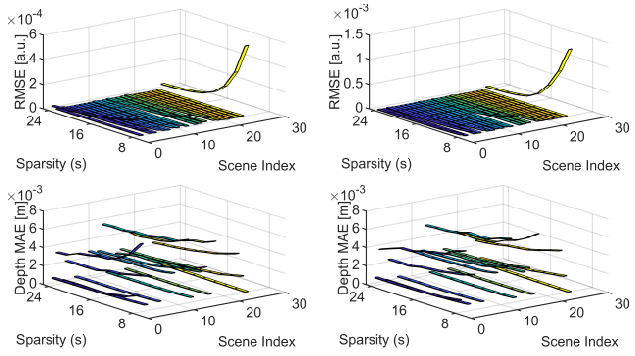
Dict. Learn. Method	Norm. RMSE per Percentile $\times 10^{-5}$ [a.u.] 0-75%	75-85%	85-95%	95-99%
MOD	**0.0301**	**0.1914**	**0.2486**	**0.6386**
K-SVD	0.0663	0.2018	0.3634	0.8074
App. K-SVD	0.0943	0.2891	0.5366	1.215
ODL	0.0697	0.2263	0.4628	1.251
RLS-DLA	0.3868	1.236	2.511	5.901

Table 2: Percentile normalized transient RMSE obtained from sparse representations in each learnt dictionary, for the worst-performing validation scene (ID 12).

Dict. Learn. Method	Depth MAE per Percentile [mm] 0-75%	75-85%	85-95%	95-99%
MOD	**0**	**3.248**	**7.041**	**24.38**
K-SVD	**0**	4.571	7.270	24.97
App. K-SVD	0.1240	5.000	8.162	26.29
ODL	0.3212	5.000	10.22	31.07
RLS-DLA	1.077	5.341	12.27	40.47
Best of [15]	7.19	20.40	32.17	71.56

Table 3: Percentile depth MAE obtained from sparse representations in each learnt dictionary, for the worst-performing validation scene (ID 12). The last row compares to the best depth MAEs in [15].

transient images (top row) and depth MAE (bottom row). The reconstruction of the transient profiles is accurate for all scenes, with only scene 25 (heavy diffuse MPI) showing comparatively lower performance, whereas the quality of the depth reconstruction shows larger variability among scenes. Errors are of only few mm and increase for lower s, but no relevant improvement is observed for $s > 16$.

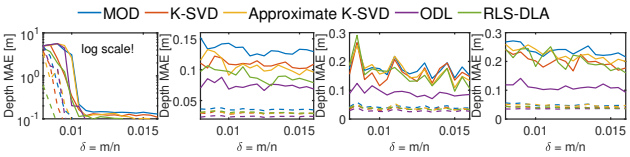


(a) MOD (b) K-SVD

Fig. 1: Evolution of the normalized RMSE of the transient profiles (top row) and of the MAE of the depth images (bottom row) with respect to the sparsity, s, for each of the 25 scenes.

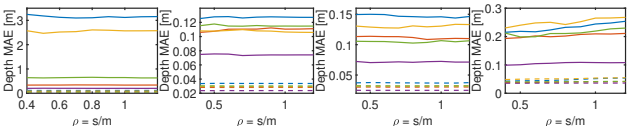
4.2. Compressive Sensing Reconstruction

In this section we evaluate the performance in the CS scenario, in which only a low-dimensional vector of measurements, $\vec{y}$, is available. The K f_k frequencies are selected both according to US and NUS schemes in the range (0-480]MHz. The reference US grid contains $K_{\max} = 32$ sampling points with 15 MHz step. On-grid random sampling is restricted to this grid. All results are given for the worst-performing scene (ID 12). Fig. 2 shows the evolution of depth MAE versus $\delta = m/n$ for the last two percentiles (dashing) and for all dictionary types (coloring). Coherently with [20], we observe successful reconstruction over a wide range of δ for NUS alternatives, while US fails for $\delta \lessapprox 0.01$. For the 85-95% percentile, the MAE is steadily below 4 cm for random sampling on grid. Fig. 3 shows the evolution of depth MAE versus $\rho = s/m$ for $m = 20$ ($\delta = 0.01$). All NUS alternatives yield successful depth retrieval, while US fails in the 85-95% percentile due to insufficient m.



(a) Uniform (b) Random w. grid (c) Random (d) Sparse ruler

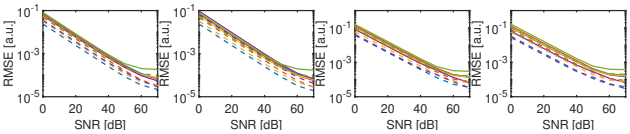
Fig. 2: Evolution of the MAE of the depth images with respect to $\delta = m/n$ for the scene 12. Solid lines are for the 95-99% percentile and dashed lines for the 85-95% percentile.



(a) Uniform (b) Random w. grid (c) Random (d) Sparse ruler

Fig. 3: Evolution of the MAE of the depth images with respect to $\rho = s/m$ for the scene 12. Solid lines are for the 95-99% percentile and dashed lines for the 85-95%. Color coding as in Fig. 2.

Fig. 4 shows the evolution of the RMSE of the reconstructed transient profiles with respect to measurement SNR for $m = 20$ in logarithmic scale. Beyond 50 dB the performance is limited by the representation error rather than by noise and the curves flatten out.



(a) Uniform (b) Random w. grid (c) Random (d) Sparse ruler

Fig. 4: Evolution of the RMSE of the transient images with respect to SNR for the scene 12. Solid lines are for the 95-99% percentile and dashed lines for the 85-95%. Color coding as in Fig. 2.

5. CONCLUSIONS

In this work we have dealt with the problem of robust depth estimation from MPI-corrupted CW-ToF measurements. The underlying low-complexity of the scene response functions has been leveraged making use of learnt sparse representations. The latter have been used to enable a CS formulation, in which the s sparse coefficients are retrieved from $m \sim \mathcal{O}(s)$ frequency measurements. Extensive simulations confirm the validity of known sparse dictionary learning techniques for finding an efficient representation. Despite K-SVD yields the lowest representation error during training, ODL shows superior performance in reconstructing transient profiles and retrieving depth. The experiments also show that NUS-CW-ToF enables transient imaging and depth retrieval from fewer measurements than US-CW-ToF. The learnt representations do not limit transient profile reconstruction below 50 dB measurement SNR. Future work includes leveraging shift-invariant transient dictionary learning.

6. REFERENCES

[1] Stefan Fuchs, "Multipath interference compensation in time-of-flight camera images," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3583–3586.

[2] Ahmed Kirmani, Arrigo Benedetti, and Philip A. Chou, "Spumic: Simultaneous phase unwrapping and multipath interference cancellation in time-of-flight cameras using spectral methods," in *IEEE Intl. Cong. on Mult. and Expo (ICME)*, July 2013, pp. 1–6.

[3] David Jiménez, Daniel Pizarro, Manuel Mazo, and Sira Palazuelos, "Modeling and correction of multipath interference in time of flight cameras," *Image and Vision Computing*, vol. 32, no. 1, pp. 1–13, 2014.

[4] Ayush Bhandari, Andrew M. Wallace, and Ramesh Raskar, "Super-resolved time-of-flight sensing via FRI sampling theory," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc.*, Mar. 2016, pp. 4009–4013.

[5] Matthew O'Toole, Felix Heide, Lei Xiao, Matthias B. Hullin, Wolfgang Heidrich, and Kiriakos N. Kutulakos, "Temporal frequency probing for 5D transient analysis of global light transport," *ACM Trans. Graph.*, vol. 33, no. 4, jul 2014.

[6] Felix Heide, Matthias B. Hullin, James Gregson, and Wolfgang Heidrich, "Low-budget transient imaging using photonic mixer devices," *ACM Trans. Graph. (Proc. SIGGRAPH 2013)*, vol. 32, no. 4, pp. 45:1–45:10, 2013.

[7] Jingyu Lin, Yebin Liu, Matthias B. Hullin, and Qionghai Dai, "Fourier analysis on transient imaging with a multifrequency time-of-flight camera," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 3230–3237.

[8] Rudolf Schwarte, Zhanping Xu, Horst-Guenther Heinol, Joachim Olk, Ruediger Klein, Bernd Buxbaum, Helmut Fischer, and Juergen Schulte, "New electro-optical mixing and correlating sensor: facilities and applications of the photonic mixer device (PMD)," in *Proc. SPIE*, 1997, vol. 3100, pp. 245–253.

[9] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar, "Coded time of flight cameras: Sparse deconvolution to address multipath interference and recover time profiles," *ACM Trans. Graph.*, vol. 32, no. 6, Nov. 2013.

[10] Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrian Jarabo, Min H. Kim, Xin Tong, and Diego Gutierrez, "DeepToF: Off-the-shelf real-time correction of multipath interference in Time-of-Flight imaging," *ACM Trans. Graph.*, vol. 36, no. 6, nov 2017.

[11] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich, "Deep end-to-end Time-of-Flight imaging," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6383–6392.

[12] Qi Guo, Iuri Frosio, Orazio Gallo, Todd Zickler, and Jan Kautz, "Tackling 3D ToF artifacts through learning and the FLAT dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[13] Gianluca Agresti and Pietro Zanuttigh, "Deep learning for multi-path error removal in ToF sensors," in *Computer Vision – ECCV 2018 Workshops*, Laura Leal-Taixé and Stefan Roth, Eds., Cham, 2019, pp. 410–426, Springer International Publishing.

[14] Gianluca Agresti, Henrik Schaefer, Piergiorgio Sartor, and Pietro Zanuttigh, "Unsupervised domain adaptation for ToF data denoising with adversarial learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5579–5586.

[15] Felipe Gutierrez-Barragan, Huaijin Chen, Mohit Gupta, Andreas Velten, and Jinwei Gu, "iToF2dToF: A robust and flexible representation for data-driven time-of-flight imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1205–1214, 2021.

[16] Adriano Simonetto, Gianluca Agresti, Pietro Zanuttigh, and Henrik Schäfer, "Lightweight deep learning architecture for MPI correction and transient reconstruction," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 721–732, 2022.

[17] Felix Heide, Lei Xiao, Andreas Kolb, Matthias B. Hullin, and Wolfgang Heidrich, "Imaging in scattering media using correlation image sensors and sparse convolutional coding," *Opt. Express*, vol. 22, no. 21, pp. 26338–26350, Oct 2014.

[18] Emmanuel J. Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[19] David L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[20] Miguel Heredia Conde, Ayush Bhandari, and Otmar Loffeld, "Nonuniform sampling of echoes of light," in *13th International conference on Sampling Theory and Applications (SampTA)*, 2019, pp. 1–4.

[21] Kjersti Engan, Sven O. Aase, and John Hakon Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, 1999, vol. 5, pp. 2443–2446.

[22] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[23] Ron Rubinstein, Michael Zibulevsky, and Michael Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," Tech. Rep., Technion - Computer Science Department, 2018.

[24] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, p. 689–696, Association for Computing Machinery.

[25] Karl Skretting and Kjersti Engan, "Recursive least squares dictionary learning algorithm," *Trans. Sig. Proc.*, vol. 58, no. 4, pp. 2121–2130, apr 2010.