Recurrent Neural Networks for Compressive Video Reconstruction

Antonio Lorente Mur, Françoise Peyrin, Nicolas Ducros

RECURRENT NEURAL NETWORKS FOR COMPRESSIVE VIDEO RECONSTRUCTION

Antonio Lorente Mur, Françoise Peyrin, Nicolas Ducros

Univ Lyon, INSA-Lyon, UCB Lyon 1,CNRS, Inserm, CREATIS UMR 5220, U1206, Lyon, France

ABSTRACT

Single-pixel imaging allows low cost cameras to be built for imaging modalities where a conventional camera would either be too expensive or too cumbersome. This is very attractive for biomedical imaging applications based on hyperspectral measurements, such as image-guided surgery, which requires the full spectrum of fluorescence.

A single-pixel camera essentially measures the inner product of the scene and a set of patterns. An inverse problem has to be solved to recover the original image from the raw measurement. The challenge in single-pixel imaging is to reconstruct the video sequence in real time from under-sampled data.

Previous approaches have focused on the reconstruction of each frame independently, which fails to exploit the natural temporal redundancy in a video sequence. In this study, we propose a fast deep-learning reconstructor that exploits the spatio-temporal features in a video. In particular, we consider convolutional gated recurrent units that have low memory requirements. Our simulation shows than the proposed recurrent network improves the reconstruction quality compared to static approaches that reconstruct the video frames independently.

Index Terms— Computational optics, video reconstruction, single-pixel camera, recurrent neural networks, image-guided surgery.

1. INTRODUCTION

Traditional imaging approaches based on arrays of sensors are ill-suited to imaging modalities where each detector is either too expensive or too cumbersome to be arranged as an array. Single-pixel imaging is a computational imaging strategy that relies on only a single point detector. Therefore, unlike conventional cameras, single-pixel cameras (SPCs) are suited to imaging at wavebands where silicon-based detectors are blind [1] or to hyperspectral imaging [2] where each pixel is a spectrometer. This makes SPCs very attractive for biomedical imaging applications based on hyperspectral measurements. In particular, we are interested in fluorescence-guided neurosurgery where exploitation of the full spectrum of fluorescence allows tumours to be detected that would have otherwise gone undetected [3].

A SPC measures the inner product of the image of a scene with a set of user-defined patterns [4], through a spatial light modulator and a set of lenses. As the measurements are performed sequentially, it is usually necessary to keep the number of patterns small for real-time applications. Therefore, the reconstruction problem in single-pixel imaging is typically under-determined, with more unknowns than measurements.

Traditional reconstruction strategies for single-pixel imaging can be categorized into two groups. The ℓ_2-regularized approaches [5], and the ℓ_1- (or total variation-) regularized algorithms [4, 6]. On the one hand, ℓ_2 approaches are fast, but they lead to reduced image quality. On the other hand, while ℓ_1 regularization leads to improved image quality, the resulting iterative algorithms are too slow to be implemented in real time. Recently, deep neural networks have been used successfully in medical image reconstruction problems, such as computed tomography [7, 8] and magnetic resonance imaging [9]. In [10], they proposed an auto-encoder network for single-pixel image reconstruction. Although this represents a useful step towards real-time imaging, the network in [10] reconstructs every frame independently, and cannot exploit the temporal redundancy of video sequences.

In the present study, we propose a fast deep-learning reconstructor that exploits the spatio-temporal features in a video. In particular, we consider a recurrent neural network (RNN) that is suited to handling image sequences through its internal state that memorizes previous inputs. Among recurrent neural networks, the long short-term memory cells are probably the most popular deep-learning variant [11]. Here, we consider gated recurrent units (GRUs), which have been shown to have similar performance to long short-term memory cells [12], although they have less memory requirements.

This paper is organized as follows. In section 2, we introduce the mathematical framework of single-pixel imaging alongside the classic single-pixel reconstruction approaches. In Section 3, we present our proposed RNN for solving the single-pixel video problem. Section 4 describes our numerical experiments, and our findings are reported and discussed in Section 5.

This work was supported by the French National Research Agency (ANR), under Grant ANR-17-CE19-0003 (ARMONI Project). It was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the programme "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the ANR.

2. RELATED WORK

2.1. Single-pixel acquisition

We consider here a video sequence $(\mathbf{f}_t)_{t\in\mathbb{N}} \in \mathbb{R}^{N\times 1}$, where $\mathbf{f}_t$ is the t-th frame in the sequence. A single-pixel camera provides access to the measurement sequence $(\mathbf{m}_t)_{t\in\mathbb{N}} \in \mathbb{R}^{K\times 1}$, which is given by [4]

$$\mathbf{m}_t = \mathbf{P}_t \mathbf{f}_t \Delta t, \quad \forall t, \tag{1}$$

where $(\mathbf{P}_t)_{t\in\mathbb{N}} \in \mathbb{R}^{K\times N}$ is the sequence of patterns that are uploaded to the spatial light modulator, and Δt is the integration time for each single pattern. At each time frame, $\mathbf{P}_t = (\mathbf{p}_{t,1}, \ldots, \mathbf{p}_{t,K})^\top \in \mathbb{R}_+^{K\times N}$ is a matrix that contains a sequence of K patterns. Patterns are typically chosen on an orthogonal basis (e.g., Hadamard, Fourier, wavelets [5]). For simplicity, the same patterns are usually chosen for measuring each frame; i.e., $\mathbf{P}_t = \mathbf{P}, \forall t$.

We assume that $\mathbf{f}_t$ is constant for a time period of $K\Delta t$, which corresponds to the acquisition of each measurement frame $\mathbf{m}_t$.

2.2. Single-pixel image reconstruction

2.2.1. *Static reconstruction*

Static reconstruction recovers $\mathbf{f}_t^* \approx \mathbf{f}_t$ by designing an inverse mapping Φ that relies solely on the current measurements $\mathbf{m}_t$; i.e.,

$$\mathbf{f}_t^* = \Phi(\mathbf{m}_t), \quad \forall t, \tag{2}$$

Traditional static approaches solve a sequence of optimization problems of the form

$$\mathbf{f}_t^* \in \operatorname{argmin} \mathcal{R}(\mathbf{f}_t) \quad \text{s.t.} \quad \mathbf{P}_t \mathbf{f}_t \Delta t = \mathbf{m}_t \tag{3}$$

where $\mathcal{R}$ is typically the ℓ_2 norm [5]), the ℓ_1 norm [4], or the total variation norm [13].

In [10], they proposed the use of an auto-encoder that processes each measurement frame independently

$$\mathbf{f}_t^* = \Phi_{\boldsymbol{\theta}^*}(\mathbf{m}_t), \quad \forall t, \tag{4}$$

where $\boldsymbol{\theta}^*$ represents the weights of the networks that are computed during a training phase. Although the training phase is time consuming, evaluation of (4) is fast. However, this approach fails to exploit the spatio-temporal redundancy within video sequences, as the same network $\Phi_{\boldsymbol{\theta}^*}$ is used for all of the time frames and it has no feedback mechanism.

2.2.2. *Dynamic reconstruction*

Dynamic reconstructions exploit temporal features by designing inverse mapping that takes into account the measurements of previous frames $(\mathbf{m}_{t'})_{0\le t'\le t}$ for the reconstruction of the current frame $\mathbf{f}_t$:

$$\mathbf{f}_t^* = \Phi(\mathbf{m}_t, \ldots, \mathbf{m}_0) \tag{5}$$

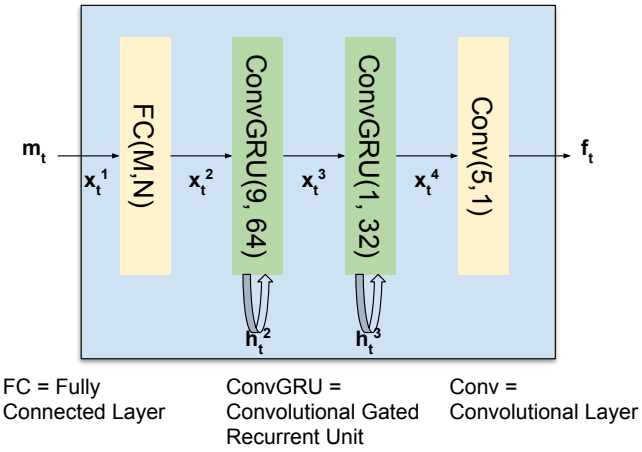

Fig. 1: Proposed recurrent neural network for single-pixel video reconstruction. The design is inspired by [10], but the convolutional layers are replaced by convolutional gated recurrent units (ConvGRUs) that maintain the long-term temporal dependency. ConvGRU(K_s, F_m) designates a ConvGRU cell with convolutional kernels of size $K_s \times K_s$ and F_m output feature maps.

In particular, sparsity promoting solutions that rely on minimizing a problem of the form

$$\mathbf{f}_t^* \in \operatorname{argmin} \mathcal{R}(\mathbf{f}_t, \ldots, \mathbf{f}_0) \quad \text{s.t.} \quad \mathbf{P}_{t'} \mathbf{f}_{t'} \Delta t' = \mathbf{m}_{t'}, \quad 0 \le t' \le t \tag{6}$$

have been proposed by different authors [6]. Despite their elegance, such approaches require iterative schemes that lead to reconstruction times (∼min) that are too long for real-time applications.

3. PROPOSED RECURRENT NETWORK FOR DYNAMIC RECONSTRUCTION

We propose to reconstruct a video sequence using a RNN that makes use of a hidden memory state. The current frame $\mathbf{f}_t^*$ is estimated from the current measurement vector $\mathbf{m}_t$ and the previous hidden state $\mathbf{h}_{t-1}$:

$$(\mathbf{f}_t^*, \mathbf{h}_t) = \Psi_{\boldsymbol{\theta}^*}(\mathbf{m}_t, \mathbf{h}_{t-1}) \tag{7}$$

where Ψ represents the RNN and $\boldsymbol{\theta}^*$ are the parameters of the RNN obtained after training. Note that the hidden state is also updated so as to maintain long-term dependency.

We consider the four-layer network depicted in Fig. 1. The first layer is a fully connected layer that projects the measurement frame $\mathbf{m}^t$ to the image domain through the statistical completion presented in [14]. The next two layers are two ConvGRUs, followed by a (regular) convolutional layer.

The output of a ConvGRU given an input $\mathbf{x}_t$ is the same

as its memory state $\mathbf{h}_t$ and can be computed as follows [15]:

$$\begin{cases} \mathbf{z}_t = \sigma(\mathbf{W}_z * \mathbf{x}_t + \mathbf{U}_z * \mathbf{h}_{t-1} + \mathbf{b}_z) & \text{(8a)} \\ \mathbf{r}_t = \sigma(\mathbf{W}_r * \mathbf{x}_t + \mathbf{U}_r * \mathbf{h}_{t-1} + \mathbf{b}_r) & \text{(8b)} \\ \tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h * \mathbf{x}_t + \mathbf{U}_h * (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) & \text{(8c)} \\ \mathbf{h}_t = (\mathbf{1} - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t & \text{(8d)} \end{cases}$$

where σ is the sigmod function, $\odot$ is the Hadamard product of two vectors, $*$ is a convolution operator, and $(\mathbf{W}_r, \mathbf{W}_z, \mathbf{W}_h, \mathbf{U}_r, \mathbf{U}_z, \mathbf{U}_h, \mathbf{b}_r, \mathbf{b}_z, \mathbf{b}_h)$ are the parameters of the unit. The update gate $\mathbf{z}_t^l$ defines how much information from the memory state we want to keep, and the reset gate $\mathbf{r}_t^l$ decides if we want to forget our previous memory state.

Given a training set $\{(\mathbf{f}^q_{\{1,\dots,T\}}, \mathbf{m}^q_{\{1,\dots,T\}})\}_{1\leq q\leq Q}$, where $\mathbf{m}_t^q = \mathbf{P}\mathbf{f}_t^q \Delta$ according to (1), we trained our network using the following loss function:

$$\boldsymbol{\theta}^* \in \arg\min \sum_{q=1}^{Q} \sum_{t=1}^{T} \frac{\|\mathbf{f}_t^q - \Psi_{\boldsymbol{\theta}}(\mathbf{m}_t^q, \mathbf{h}_{t-1}^q)\|_2^2}{2QT} + \lambda \|\boldsymbol{\theta}\|_2^2 \quad (9)$$

where λ is the weight decay parameter.

4. NUMERICAL SIMULATIONS

In our experiments, we chose $M = 333$ Hadamard patterns of size $N = 64 \times 64$. The size of the convolutional kernels and number of feature maps of our RNN were chosen to mimic those in [10].

4.1. Training dataset

We use the UCF-101 [16] dataset to train and test our network. The UCF-101 dataset is an action recognition dataset consisting of 13 320 videos from 101 action categories. Each video has a different number of frames and a different resolution. Therefore we down-sample all of the frames to 64×64.

4.2. Training procedure

We train our RNN using Pytorch [17]. For training, we consider the ADAM optimiser for 150 epochs. The step size is initialized to 10^{-3} and divided by 5 every 40 epochs. At each epoch, we randomly extract 10 consecutive frames from each video. The weight decay regularization parameter λ is set to 10^{-6}. The number of learned parameters is 1033601. Note that the fully connected layer is computed beforehand [14], so that our network does not need to learn it.

5. RESULTS AND DISCUSSION

We compare our dynamic reconstruction method with three static approaches. We consider the pseudo inverse solution, the total variation solution, and a deep-learning approach. More precisely, the pseudo inverse and the total variation solutions correspond to choosing $\mathcal{R}$ as the ℓ_2-norm and as the total variation in (3), respectively. The deep-learning approach is the network proposed in [10]. As reported in Table 1, our proposed dynamic reconstruction method yields better average peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) results than the other two static reconstruction methods.

Table 1: Average peak signal-to-noise ratio (PSNR) and average structural similarity (SSIM) over the UCF-101 test dataset for the three different methods

Method	PSNR	SSIM
Pseudo-inverse	22.05	0.9278
Static network [10]	23.63	0.9479
Proposed dynamic network	**23.92**	**0.9508**

We also compare our proposed method using a real fluorescence-guided neurosurgery video. From the original video, we consider the fluorescence signal from protoporphyrin IX within the tumours. We consider 20 frames that we reshape to 64×64 pixels. Fig. 2 shows the fluorescence ground-truth and reconstruction for two frames. Although our network is trained on an action recognition dataset, it provides better reconstruction results than the other three methods, both visually and it terms of PSNR and SSIM.

6. CONCLUSION AND PERSPECTIVES

We propose a RNN to solve the single-pixel video problem. We use ConvGRUs that enable the spatio-temporal redundancy within natural videos to be exploited. Compared to strategies based on static reconstructions, the reconstruction error is decreased and the reconstruction quality is visually improved. In future, we will evaluate this method on experimental data. Although this study focuses on single-pixel videos, our method can be adapted for similar problems that incompletely sample the Fourier domain, such as magnetic resonance imaging.

References

[1] Jaewook Shin et al., "Single-pixel imaging using compressed sensing and wavelength-dependent scattering," *Opt. Lett.*, vol. 41, no. 5, pp. 886–889, Mar 2016.

[2] Florian Rousset, Nicolas Ducros, Françoise Peyrin, Gianluca Valentini, Cosimo D'Andrea, and Andrea Farina, "Time-resolved multispectral imaging based on an adaptive single-pixel camera," *Opt. Express*, vol. 26, no. 8, pp. 10550–10558, Apr 2018.

[3] Pablo A Valdés et al., "Quantitative fluorescence using 5-aminolevulinic acid-induced protoporphyrin ix biomarker as a

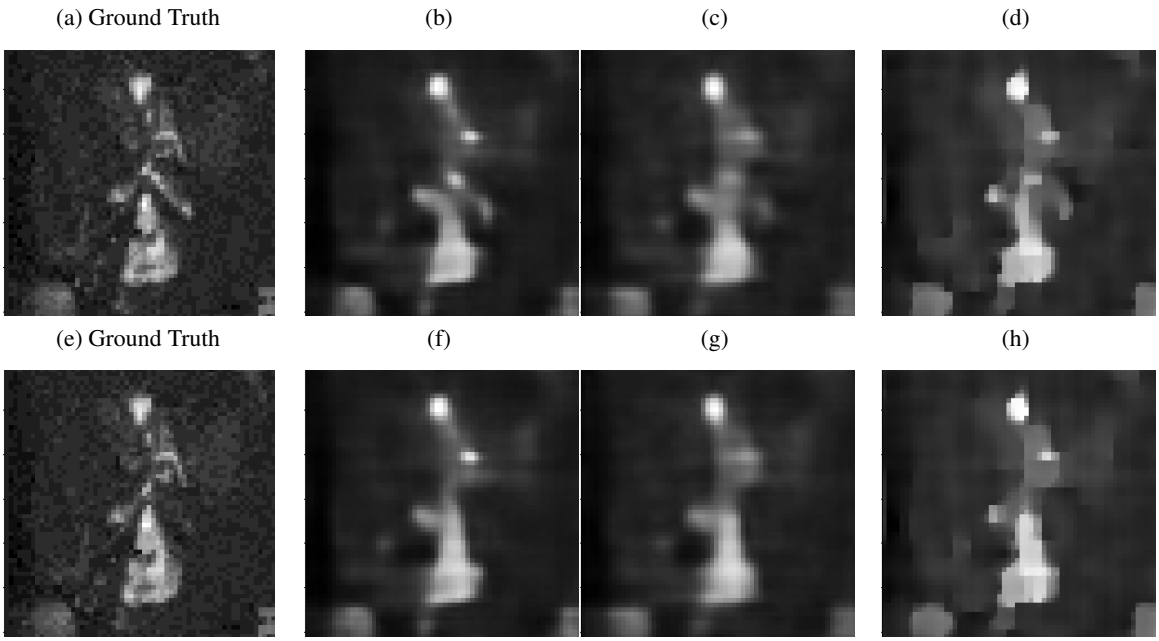

Fig. 2: Reconstruction through different methods of a frame of a fluorecence-guided neurosurgery video. (a) Ground truth frame 10. (b) Proposed recurrent network; peak signal-to-noise ratio (PSNR) = 24.50, structural similarity (SSIM) = 0.86. (c) Static network proposed in [10]; PSNR = 24.06; SSIM = 0.84; (d) Total Variation [13]; PSNR = 24.35, SSIM = 0.85. (e) Ground truth frame 10. (f) Proposed recurrent network; peak signal-to-noise ratio (PSNR) = 24.64; SSIM = 0.87. (g) Static network proposed in [10]; PSNR = 24.35; SSIM = 0.86; (h) Total Variation [13]; PSNR = 24.16, SSIM = 0.85.

surgical adjunct in low-grade glioma surgery," *Journal of neurosurgery*, vol. 123, no. 3, pp. 771–780, 2015.

[4] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, Ting Sun, K.F. Kelly, and R.G. Baraniuk, "Single-pixel imaging via compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 83–91, March 2008.

[5] Florian Rousset, Nicolas Ducros, Andrea Farina, Gianluca Valentini, Cosimo D'Andrea, and Françoise Peyrin, "Adaptive Basis Scan by Wavelet Prediction for Single-pixel Imaging," *IEEE Transactions on Computational Imaging*, 2016.

[6] R. G. Baraniuk, T. Goldstein, A. C. Sankaranarayanan, C. Studer, A. Veeraraghavan, and M. B. Wakin, "Compressive video sensing: Algorithms, architectures, and applications," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 52–66, Jan 2017.

[7] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, "Cnn-based projected gradient descent for consistent ct image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1440–1453, June 2018.

[8] Il Yong Chun, Xuehang Zheng, Yong Long, and Jeffrey A. Fessler, "Bcd-net for low-dose ct reconstruction: Acceleration, convergence, and generalization," 2019.

[9] Kerstin Hammernik et al., "Learning a variational network for reconstruction of accelerated mri data," *Magnetic resonance in medicine*, vol. 79 6, pp. 3055–3071, 2017.

[10] Catherine Higham, Roderick Murray-Smith, Miles Padgett, and Matthew. Edgar, "Deep learning for real-time single-pixel video," *Scientific Reports.*, , no. 8, Feb 2018.

[11] Sepp Hochreiter and Jrgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[12] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[13] Chengbo Li, *An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing*, Ph.D. thesis, Rice University, 2010.

[14] N. Ducros et al., "A completion network for compressed acquisition," in *IEEE 17th International Symposium on Biomedical Imaging (submitted)*, October 2020.

[15] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville, "Delving deeper into convolutional networks for learning video representations," *arXiv preprint arXiv:1511.06432*, 2015.

[16] Khurram Soomro et al., "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, p. 2012.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.