Joint optimization for compressive video sensing and reconstruction under hardware constraints

Michitaka Yoshida[1], Akihiko Torii[2], Masatoshi Okutomi[2], Kenta Endo[3], Yukinobu Sugiyama[3], Rin-Ichiro Taniguchi[1] Hajime Nagahara[4]
1 Kyushu University
2 Tokyo Institute of Technology
3 Hamamatsu Photonics K. K.
4 Osaka University

Abstract Compressive video sensing is the process of encoding multiple sub-frames into a single frame with controlled sensor exposures and reconstructing the sub-frames from the single compressed frame. It is known that spatially and temporally random exposures provide the most balanced compression in terms of signal recovery. However, sensors that achieve a fully random exposure on each pixel cannot be easily realized in practice because the circuit of the sensor becomes complicated and incompatible with the sensitivity and resolution. Therefore, it is necessary to design an exposure pattern by considering the constraints enforced by hardware. In this paper, we propose a method of jointly optimizing the exposure patterns of compressive sensing and the reconstruction framework under hardware constraints. By conducting a simulation and actual experiments, we demonstrated that the proposed framework can reconstruct multiple sub-frame images with higher quality.

Keywords: Compressive sensing, video reconstruction, deep neural network

1. Introduction

Recording a high-frame video with high spatial resolution has various uses in practical and scientific applications. Such video sensing can be achieved by using a high-speed camera, but such special sensor is expensive and low resolution and low sensitivity.

A feasible approach consists of capturing video by using compressive sensing techniques [1-4]. In contrast to the standard images captured with a global shutter, where all pixels are exposed concurrently, a compressive video sensor samples temporal information and compresses it into a single image, while randomly changing the exposure pattern for each pixel. This non-continuous exposition enables the recovery of high-quality video. Formally, compressive video sensing is expressed as follows:

$$y = \emptyset x$$

where x is the high-frame video to be compressed, $\emptyset$ is the measurement matrix (exposure patterns), and y is the compressed single image. The following tasks are included in compressive video sensing: reconstruct a high-frame video $\bar{x}$ from a single image y by using pattern $\emptyset$; optimize the pattern that enables high-quality video reconstruction (Figure 1).

In signal recovery theory, the best exposure pattern is random sampling from a uniform distribution. However, this is not an optimal pattern in terms of practical image sensing, because a practical scene does not always maintain the sparsity assumed in compressive sensing theory. However, implementing such completely random exposures with a practical CMOS sensor is not realistic, owing to hardware limitations. Therefore, it is necessary to optimize the exposure patterns by recognizing the hardware constraints of actual sensors.

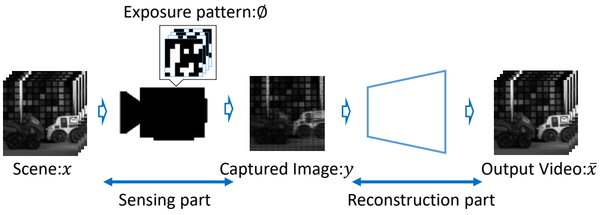

Fig. 1. Compressive video sensing.

In this paper, we propose a new pipeline to optimize both the exposure pattern and reconstruction decoder of compressive video sensing by using a deep neural network (DNN) framework. The proposed method is a general framework for optimizing the exposure patterns with and without hardware constraints.

2. Hardware constraints of exposure controls

There exist hardware constraints that prevent the generation of completely random exposure patterns, which are a theoretical requirement of compressive video sensing. In this section, we detail the hardware constraints resulting from sensor architecture.

Hitomi et al. [2] assume the CMOS sensor that has a row and column address decoder and can be read out pixel-wise. However, it does not have a per-pixel buffer which is only a single bump exposure in a frame. In this paper, we termed this sensor as the single bump exposure (SBE) sensor.

Sonoda et al. [1] used the prototype CMOS sensor with additional reset and transfer signal lines to control the exposure time. These signal lines are shared by the pixels in the columns and rows. Therefore, the exposure patterns depended spatially on the rows or columns of the neighboring pixels. This image sensor has an additional transistor and exposure control signal line, and can perform multi-bump exposure. In this paper, we termed this sensor as the row-column wise exposure (RCE) sensor.

Figure 2 shows the architecture of SBE (left side) and RCE (right side) sensor.

3. Joint optimization for sensing and reconstruction under hardware constraints

In this section, we describe the proposed optimization method of jointly optimizing the exposure pattern of compressive video sensing, and performing reconstruction by using a DNN. The proposed DNN consists of two main parts. The first part is the sensing layer (encoding) that optimizes the exposure pattern (binary weight) under the constraint imposed by the hardware structure, as described in Section 2. The second part is the reconstruction layer that recovers the multiple sub-frames from a single captured image, which was compressed by using the

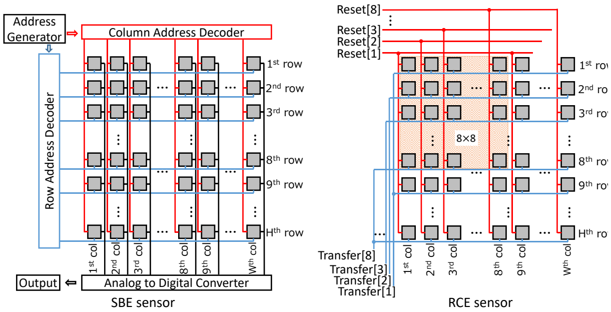

Fig. 2. Architecture of SBE and RCE image sensors.

optimized exposure pattern. The overall framework is shown in Figure 3. Training was carried out in the following steps:

1. At the time of forward propagation, the binary weight is used for the sensing layer, while the reconstruction layer uses the continuous weights.
2. The gradients are computed by backward propagation.
3. The continuous weights of sensing and reconstruction layers are updated according to the computed gradients.
4. The binary weights of the sensing layer are updated with the continuous weights of the sensing layer.

3.1 Compressive sensing layer

We sought an exposure pattern that would be capable of reconstructing video frames with high quality when trained along with the reconstruction (decoding) layer. More importantly, the compressive sensing layer had to be capable of handling the exposure pattern constraints imposed by actual hardware architectures. Because implementing nested spatial pattern constraints (Section 2) in the DNN layer was not trivial, we used a binary pattern (weight) chosen from the precomputed binary weights at forward propagation in the training. The binary weight was relaxed to a continuous value [5] to make the network differentiable by backward computation. Next, the weight was binarized for the next forward computation by choosing the most similar patterns in the precomputed binary weights. The similarity between the continuous-value weight and the precomputed binary pattern was computed by the normalized dot product.

3.2 Reconstruction layer

The reconstruction layer decodes high-frame videos from a single image compressed by using the learned exposure pattern, as was described in the previous section. This decoding expands the single image to multiple sub-frames by non-linear mapping, which can be modeled and learned by a multi-layer perceptron (MLP). As illustrated in Figure 3, the MLP consisted of four hidden layers and each layer was truncated by rectified linear unit (ReLU). The network was trained by minimizing the errors between the training videos and the reconstructed videos. We used the mean squared error (MSE) as the loss function because it was directly related with the peak signal-to-noise ratio (PSNR).

4. Experiments

We carried out simulation experiments to evaluate our method. The network size was determined based on the size of the patch volume to be reconstructed. We used the controllable exposure sensor [1]. Therefore, the volume size of $W_p \times H_p \times T$ was set to $8 \times 8 \times 16$ in the experiments. We assumed two different types of hardware constraints for the SBE and RCE sensors. The details of the SBE and RCE sensor constraints are described in Section 2. We generated a captured image simulated for the SBE and RCE sensors and input the simulated images to the reconstruction network to recover the video. Figure 4 shows that, in our result, the edges of the letter mark were reconstructed sharper than in the result of the handcrafted exposure pattern. The reconstruction qualities in the optimized exposure pattern were always better than those of the handcrafted exposure pattern. Table 1 shows the average PSNRs of the handcrafted and optimized results for the SBE and RCE sensors. Owing to the pattern's joint optimization and the reconstruction layers, the proposed method always outperformed the original handcrafted patterns.

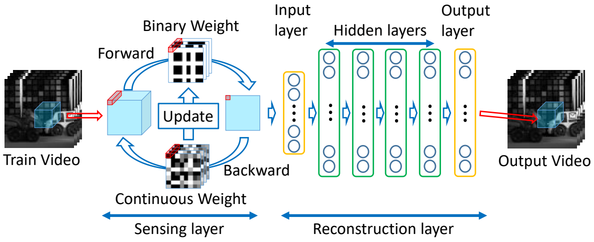

Fig. 3. Network Structure

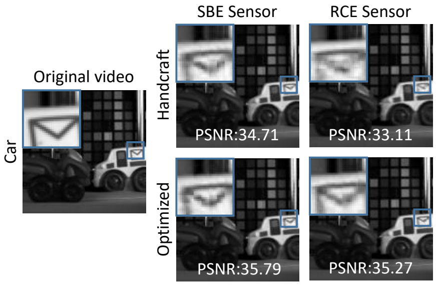

Fig. 4. Reconstruction results of 3rd sub-frame

Table 1. Average PSNR of video reconstruction

	Handcraft	Optimized	Handcraft	Optimized
	SBE［2］	SBE	RCE［1］	RCE
PSNR	29.41	30.05	28.51	29.45

4. Conclusion

In this paper, we first argued that real sensor architectures for developing controllable exposure have various hardware constraints that make non-practical the implementation of compressive video sensing based on completely random exposure patterns.

References

[1] T. Sonoda, H. Nagahara, K. Endo, Y. Sugiyama, R.Taniguchi, "High-speed imaging using CMOS image sensor with quasi pixel-wise exposure", International Conference on Computational Photography (ICCP), 1-11, 2016.

[2] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, S. K. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary", International Conference on Computer Vision (ICCV), 287-294, 2011.

[3] M. Iliadis, L. Spinoulas, A. K. Katsaggelos, "Deep fully-connected networks for video compressive sensing." Digital Signal Processing 72 ,9-18, 2018.M. Iliadis, L. Spinoulas, A. K. Katsaggelos, "Deep Binary Mask: Learning a Binary Mask for Video Compressive Sensing.", arXiv preprint arXiv:1607.03343 2016.

[4] M.Courbariaux, I.Hubara, D.Soudry, R.El-Yaniv, Y.Bengio, "Binarized neural networks: Training neural networks with weights and activations constrained to+ 1 or － 1." arXiv preprint arXiv:1602.02830, 2016.