INP5-3
Invited

Deep Sensing

- Joint Optimization of Hardware and Software for Imaging -

Hajine Nagahara[1]

nagahara@ids.osaka-u.ac.jp
[1]Institute for Datability Science, Osaka University
2-8, Yamadaoka, Suita, Osaka, 565-0871, Japan
Keyword: Computational Photography, Deep Learning, Compressive Sensing

ABSTRACT

Deep neural network (DNN) is a powerful tool for solving computer vision tasks such as object recognition, scene understanding and image reconstruction etc. However, DNN have been used for only digital domain in the imaging pipeline such as the feature extractor and classifier models after image is captured and digitized as shown in blur part of figure 1. In this research, we propose a new framework, called "deep sensing" as shown in figure 1. The proposed framework also models the analog layer to neural network model, and jointly optimize the parameters in optics and sensor designs of a camera as well as reconstruction and classification models by same training strategy.

1 Introduction

Deep neural network (DNN) is a powerful tool for solving computer vision tasks such as object recognition, scene understanding and image reconstruction etc. It realizes to drastically improve the accuracy of recognition and reconstruction to the classical methods, since feature extractor and classifier models are designed by a training based on the target data. However, DNN have been used for only digital domain in the imaging pipeline such as the feature extractor and classifier models after image is captured and digitized as shown in blur part of figure 1. On the other hand, optics and sensor in analog layer still have been designed by hand based on theoretical or empirical analysis. It is not always grantee that the designs and hardware setting parameters are optimal to the applications and target tasks. In this research, we propose a new framework, called "deep sensing" as shown in figure 1. The proposed framework also models the analog layer to neural network model, and jointly optimize the parameters in optics and sensor designs of a camera as well as reconstruction and classification models by same training strategy. In this talk, we introduce the concept of deep sensing and show our work; compressive light field sensing [1,4], compressive video sensing [2], and action recognition by a coded image [3] as examples of the applications. The details are in the corresponding papers. Similar approaches which called deep optics or neural sensing are also getting popular in the research area and they applied to variety of the applications; hyper spectrum imaging [5], lens design for depth estimation [6], high-dynamic range [7], etc.

2 Compressive video sensing[2]

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. We propose a method of jointly optimizing the exposure patterns of compressive sensing and the reconstruction framework under hardware constraints as shown in figure 2. By conducting a simulation and actual experiments, we demonstrated that the proposed framework could reconstruct multiple sub-frame images with higher quality as shown in figure 3.

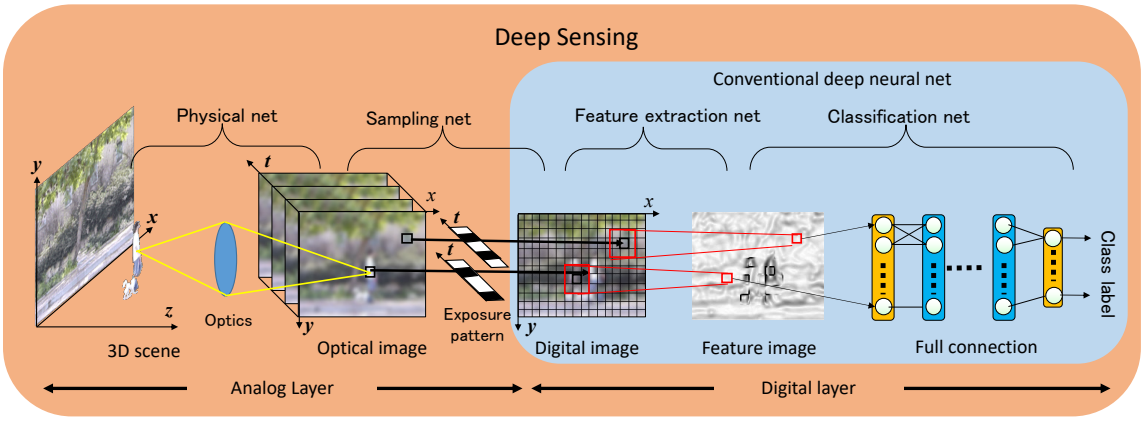

Figure 1: Conceptual figure of Deep sensing

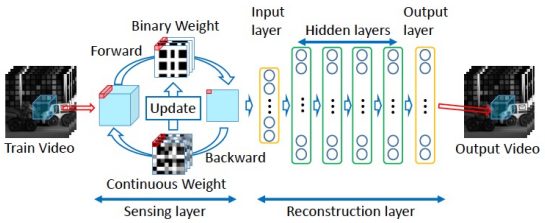

Figure 2: Compressive video sensing

Original video(PSNR) Handcraft(29.77) Optimized(31.39)

Figure 3: Reconstruction results

3 Action recognition from a single coded image[3]

Cameras are prevalent in society at the present time, for example, surveillance cameras, and smartphones equipped with cameras and smart speakers. There is an increasing demand to analyze human actions from these cameras to detect unusual behavior or within a man-machine interface for Internet of Things (IoT) devices. For a camera, there is a trade-off between spatial resolution and frame rate. It is difficult to capture or send a high resolution and high frame rate because of the limitation of the read-out of the sensor and band limit of the network between a client camera and processing server. Low resolution or low frame rate video is not suitable input for action recognition because the low resolution loses object detail, and the low frame rate loses motion detail. A feasible approach to solve this problem is compressive video sensing. Compressive video sensing uses random coded exposure and reconstructs higher than read out of sensor frame rate video from a single coded image. We believe that it is possible to recognize action in a scene from a single coded image because the image contains multiple temporal information for reconstructing a video. We propose using a coded exposure image for action recognition. We use a deep learning framework as shown in figure 4 to optimize the coded exposure pattern in addition to learning the classification model simultaneously. We demonstrated that the proposed method could recognize human actions from only a single coded image. We also compared it with competitive inputs, such as low-resolution video with a high frame rate and high-resolution video with a single frame and demonstrated some advantages of the proposed method as shown in Table 1.

4 Conclusions

This talk proposes deep sensing which is jointly optimized sensor design as well as decoder/classifier in deep learning framework. We introduce examples of the task specific sensing and optimization. Please refer the details to the original papers in the references.

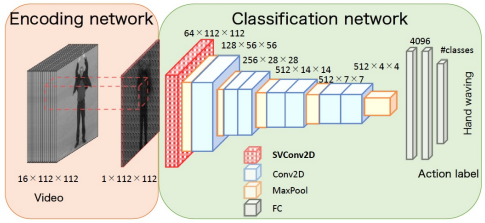

Figure 4: Model for single image action recognition

Table 1: Experimental comparisons

			Simulation			Real		
Input		Model	Top1	Top3	Top5	Top1	Top3	Top5
Video (upper bound)		C3D	47.1	69.4	76.9	71.0	88.0	88.0
Single image	Coded exposure (Proposed)	SVC2D	41.6	58.9	67.2	72.0	84.0	88.0
	Long exposure	C2D	13.8	30.4	39.4	20.0	40.0	52.0
	Short exposure	C2D	14.6	32.5	40.5	21.0	47.0	60.0

References

[1] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, “Learning to capture light fields through a coded aperture camera,” European Conference on Computer Vision, 2018, pp. 418–434.

[2] M. Yoshida, A. Torii, M. Okutomi, K. Endo, Y. Sugiyama, R. Taniguchi, and H. Nagahara, “Joint optimization for compressive video sensing and reconstruction under hardware constraints,” European Conference on Computer Vision, 2018, pp. 634–649.

[3] T. Okawara, M. Yoshida, H. Nagahara, and Y. Yagi, “Action Recognition from a Single Coded Image,” in Proceedings of IEEE International Conference on Computational Photography, 2020.

[4] K. Sakai, K. Takahashi, T. Fujii, H. Nagahara: "Acquiring Dynamic Light Fields through Coded Aperture Camera", European Conference on Computer Vision 2020.

[5] S. Nie, L. Gu, Y. Zheng, A. Lam, N. Ono, and I. Sato, “Deeply learned filter response functions for hyperspectral reconstruction,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4767–4776.

[6] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghav, “Phasecam3d—learning phase masks for passive single view depth estimation,” in Proceedings of IEEE International Conference on Computational Photography, 2019, pp. 1–12.

[7] C. A. Metzler, H. Ikoma, Y. Peng, G. Wetzstein, “Deep Optics for Single-shot High-dynamic-range Imaging,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1375–1385.