Deep sensing
-Jointly optimize imaging and processing-

Hajime Nagahara[1]
1 Institute for Datability Science, Osaka University
2-8, Yamadaoka, Suita, Osaka, 565-0871, Japan
E-mail:nagahara@ids.osaka-u.ac.jp

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

Keywords: Deep neural network, Computational photography, Compressive sensing

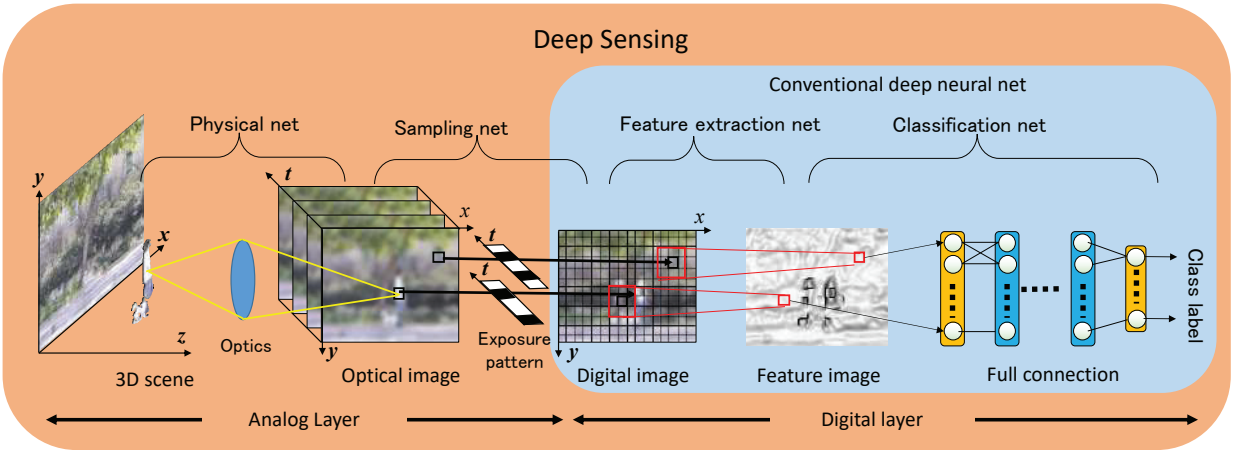


Figure 1: Conceptual figure of Deep sensing

1. Overview

Deep neural network (DNN) is a powerful tool for solving image processing and computer vision tasks such as image and video reconstructions, object recognition, and scene understanding, etc. It realizes to drastically improve the quality of reconstruction and the accuracy of recognition to the classical methods since feature extractor and classifier models are designed by training based on the target data. However, DNN have been used for only the digital domain in the imaging pipeline, such as the feature extractor and classifier models after an image is captured and digitized, as shown in the blue part of figure 1. On the other hand, optics and sensors in the 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 the neural network model and jointly optimizes the parameters in optics and sensor designs of a camera as well as reconstruction and classification models by the same training strategy. In this talk, I introduce the concept of deep sensing and show our work; compressive light field sensing [1,4,5], compressive video sensing [2], action recognition by a coded image [3], and privacy-preserving imaging. The details are in the corresponding papers.

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] R. Mizuno, K. Takahashi, M. Yoshida, C. Tsutake, T. Fujii, H. Nagahara, "Acquiring a Dynamic Light Field through a Single-Shot Coded Image", IEEE Conference on Computer Vision and Pattern Recognition, June, 2022.

[6] S. Kumawat, T. Okawara, M. Yoshida, H. Nagahara, Yasushi Yagi, "Action Recognition From a Single Coded Imaging", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-14, Aug., 2022.

[7] S. Kumawat, H. Nagahara, "Privacy-Preserving Action Recognition via Motion Difference Quantization", European Conference on Computer Vision, Oct, 2022.

Deep sensing

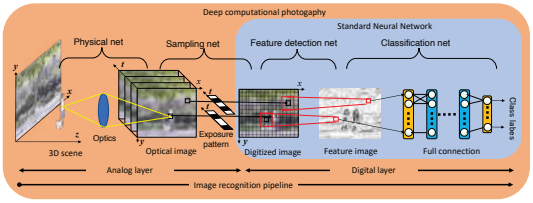
- Jointly optimizing sensing and processing -

Hajime Nagahara

Institute for Datability Science, Osaka University

1

Concept of Deep sensing



H30-R1科研挑戦的研究(萌芽)

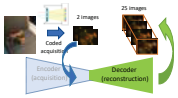
Proposing a general framework to design a camera hardware by DNN optimization

2

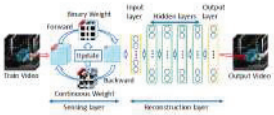
Our research examples

Learning to Capture Light Fields through A Coded Aperture Camera [Inagaki+ ECCV2018]

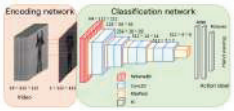
Acquiring a Dynamic Light Field through a Single-Shot Coded Image [Mizuno+ CVPR2022]



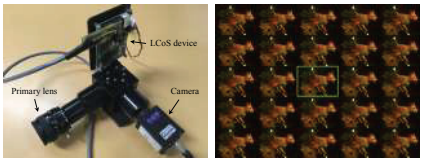
Joint optimization for Compressive video sensing [Yoshida+ ECCV2018]



Action Recognition from a Single Coded Image [Okawara+ ICCP2020]



3



Compressive LF sensing

Presented at:

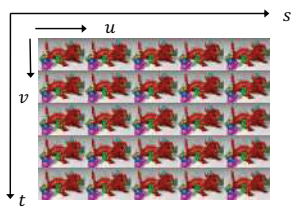
Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, H. Nagahara:
"Learning to Capture Light Fields through A Coded Aperture Camera",
European Conference on Computer Vision (ECCV) 2018.

4

What is light field (LF)?

Dense set of multi-viewpoint images

Cameras

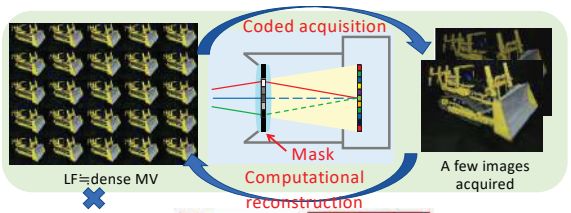


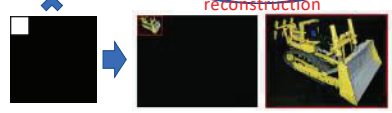
(s,t) Viewpoint

(u,v) Pixel position

5

Naïve acquisition





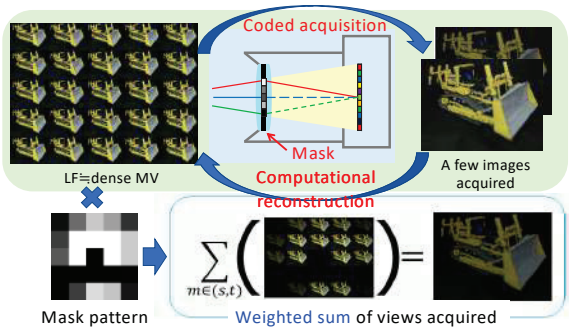
Mask pattern

View by view acquisition

- 25 acquisitions
- Noisy images

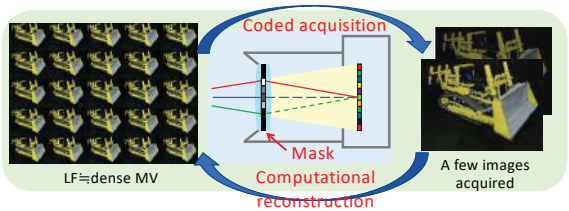
6

Coded Acquisition



7

Problem statement



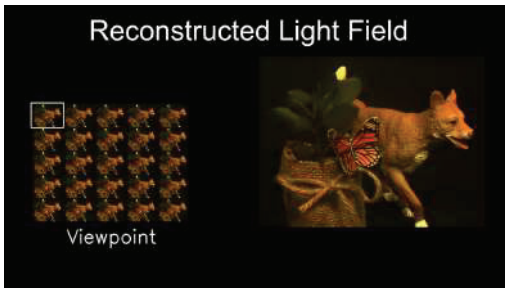
Maximize the quality of reconstructed LF
s.t. limited number of acquisitions

Find best mask pattern and reconstruction method

8

Our achievement

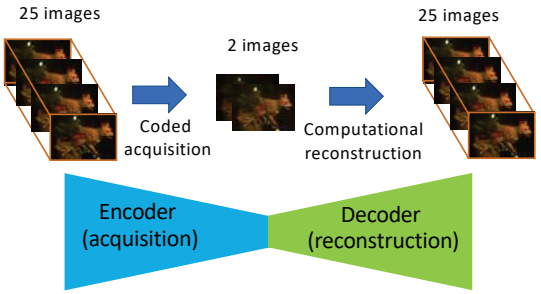
- Reconstruction of 5x5 or 8x8 LF images from only 2-4 acquisitions



9

How did we do that?

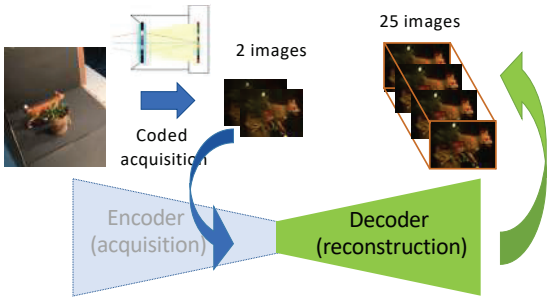
- Modeling as an auto-encoder



10

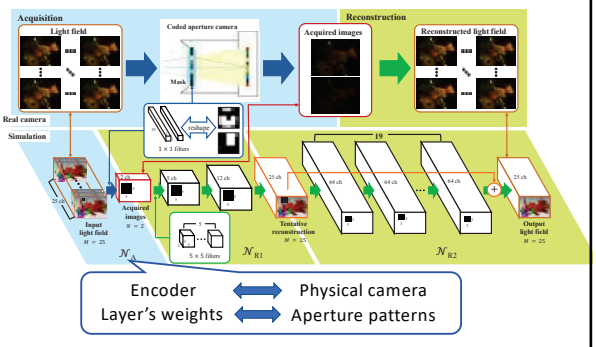
How did we do that?

- Replace the encoder with a coded aperture camera



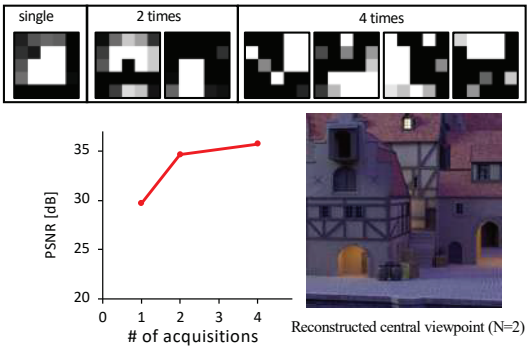
11

Network architecture



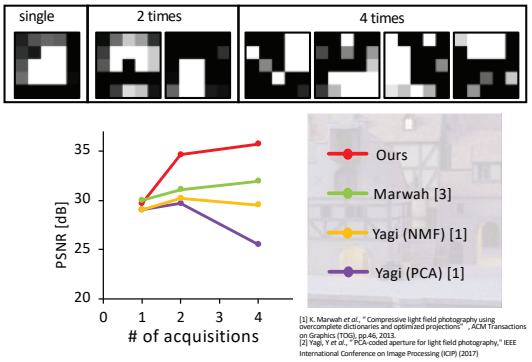
12

Learned aperture patterns (5x5)



13

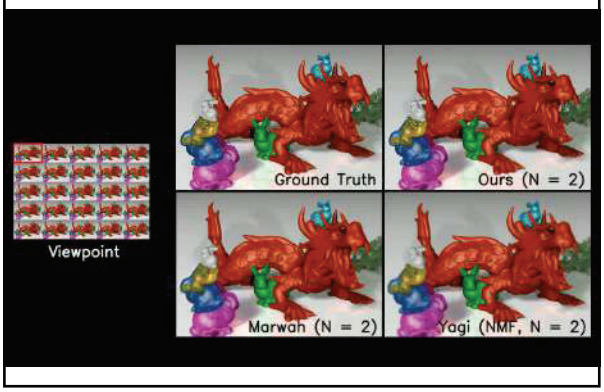
Comparison (coded acquisition)



[1] K. Marwah *et al.*, " Compressive light field photography using overcomplete dictionaries and optimized projections" , ACM Transactions on Graphics (TOG), pp.46, 2013.
[2] Yagi, Y *et al.*, " PCA-coded aperture for light field photography," IEEE International Conference on Image Processing (ICIP) (2017)

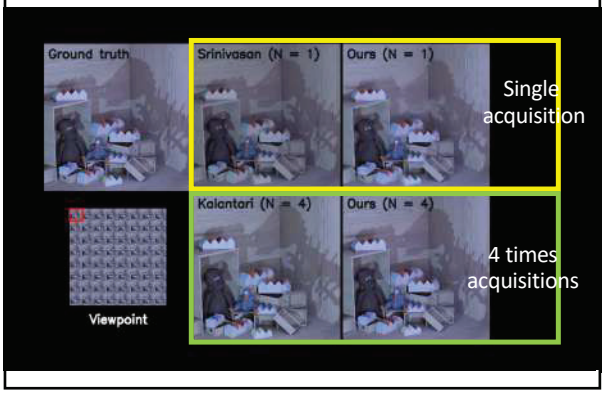
14

Comparison (2 coded acquisitions)



15

Comparison (against view synthesis)



16

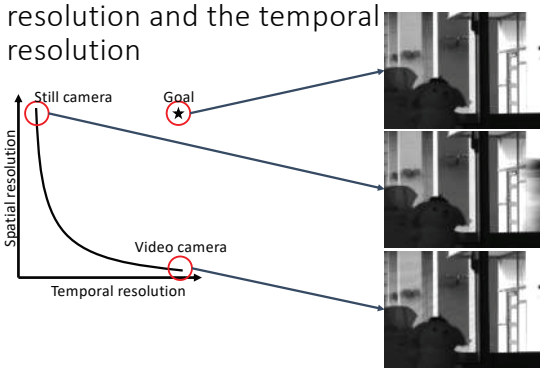
Joint optimization for Compressive video sensing

Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi,
Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi,
Hajime Nagahara

ECCV2018

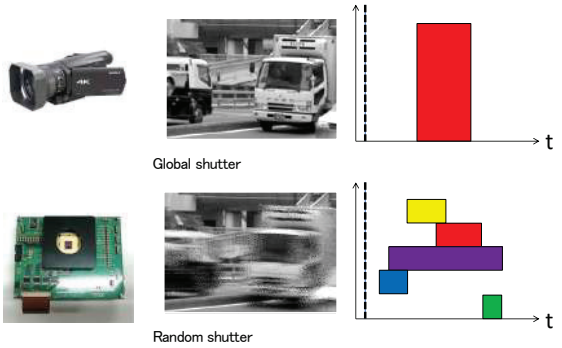
17

A trade-off between the spatial resolution and the temporal resolution



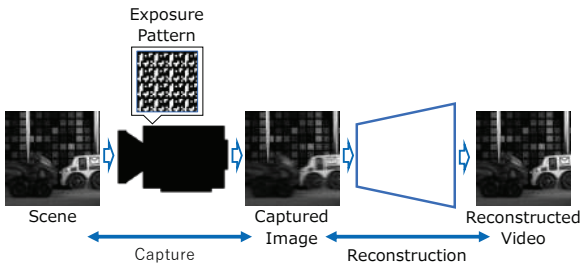
18

Random shutter



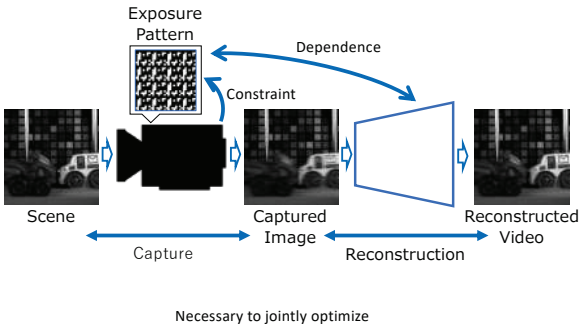
19

Compressive video sensing



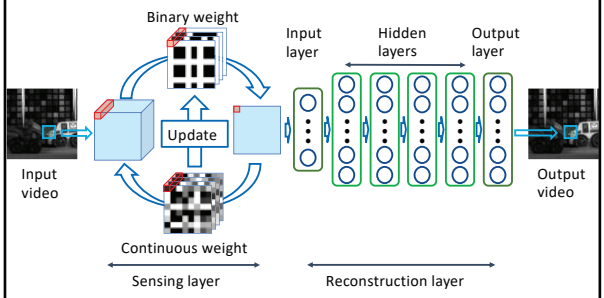
20

Compressive video sensing



21

Joint optimization for compressive video sensing

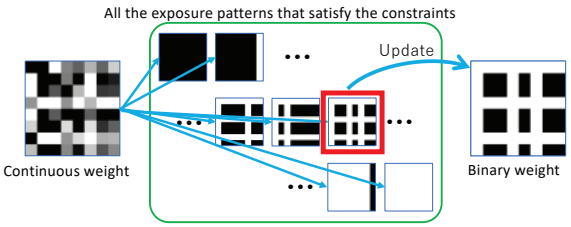


22

Update the binary weights with continuous weights

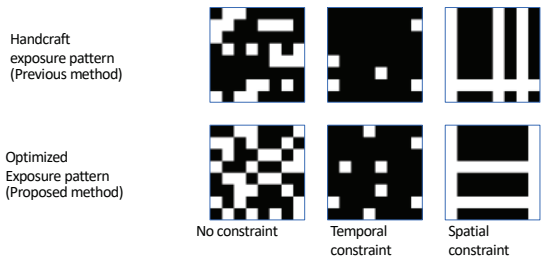
- Step to compute the possible patterns

1. Generate the possible sets of signal
2. Simulate the exposure pattern



23

Optimized exposure pattern (8×8)



24

Simulation experiments

- No constraint
  - Sarhangnejad et al.

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

14 videos Average	Handcraft	Optimize
PSNR(dB)	29.17	29.99

25

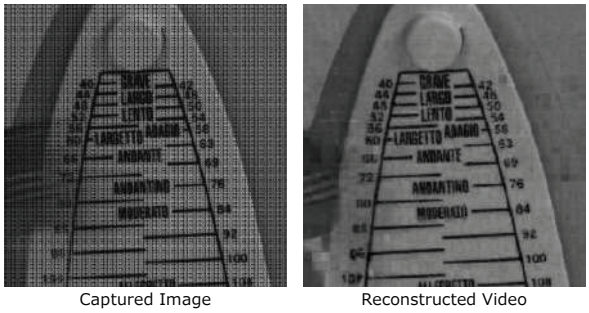
Real experiments

- Spatial constraint on exposure pattern
- Capturing coded images with 15 FPS
- Set 16 exposure patterns per frame

The reconstructed video was equivalent to 240 FPS

26

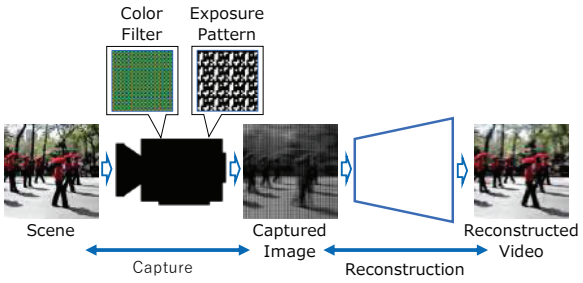
Real experiments



Captured Image Reconstructed Video

27

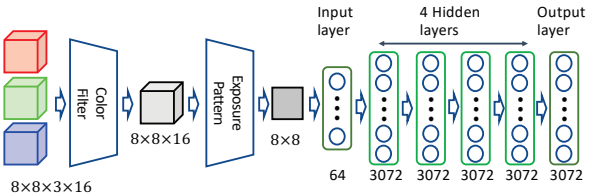
Compressive color video sensing



28

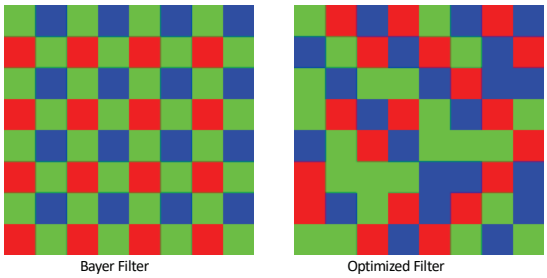
Network structure to optimize the color filter, exposure pattern and reconstruction

- Input : RGB Video (8×8×3×16)
- Output : RGB Video (8×8×3×16)



29

Color filter (8×8)



Bayer Filter

R:16 G:32 B:16

Optimized Filter (100 epoch)

R:19 G:23 B:22

30

Exposure pattern (Different color filter)

Monochrome (500epoch) | Bayer filter (100epoch) | Optimized filter (100epoch)

31

Reconstruction results

Original Video

Color Filter	Bayer	Bayer	Optimize
Exposure Pattern	Fixed	Optimize	Optimize
Epoch	100	100	100
PSNR	24.18	23.92	24.34

Average 25 Videos	Only decoder	Exposure pattern	Filter + Pattern
PSNR	26.56	26.43	26.76

32

Acquiring a Dynamic Light Field through a Single-Shot Coded Image

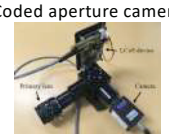
Ryoya Mizuno, Keita Takahashi, Michitaka Yoshida,
Chihiro Tsutake, Toshiaki Fujii, Hajime Nagahara

CVPR2022

33

Combination of aperture and pixel codings

Coded aperture camera



Coding viewpoint dimension
to reconstruct Light field

+

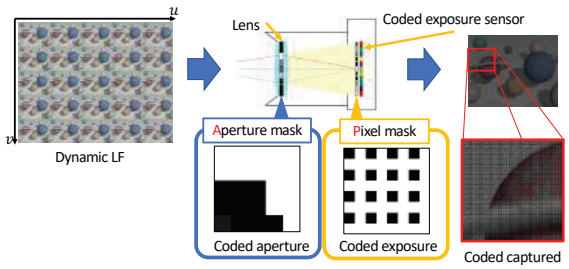
Coded exposure camera

Coding space-time dimension
to reconstruct video

= 5D Light field video can be captured and reconstructed

34

Coded aperture + Coded exposure

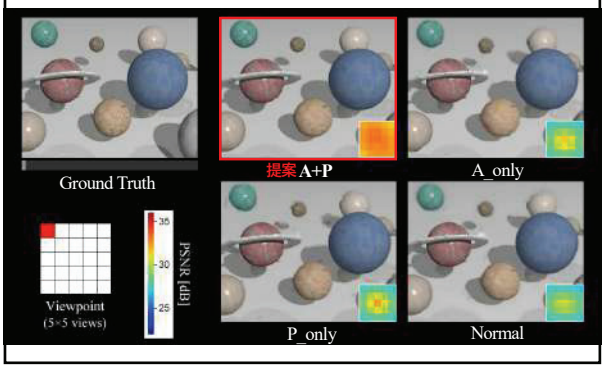


Synchronized 4 aperture and 4 exposure pattens

Compressing 25 views x 4 frames into a single coded image

35

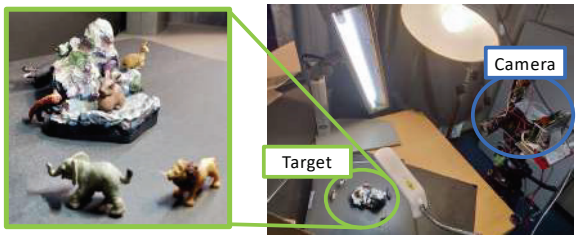
Simulation experiment (wakusei)



36

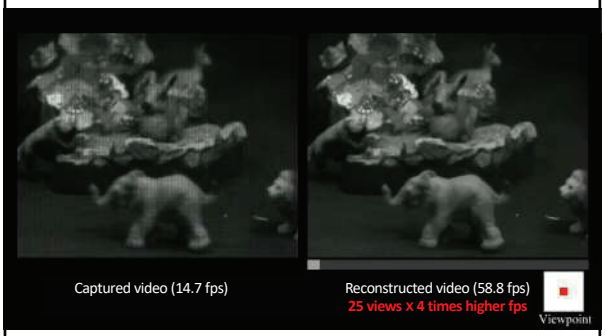
Real experiment using prototype camera

The target objects moved by turntable.
The camera captured coded images with 14.7 fps.



37

Coded and Reconstructed videos



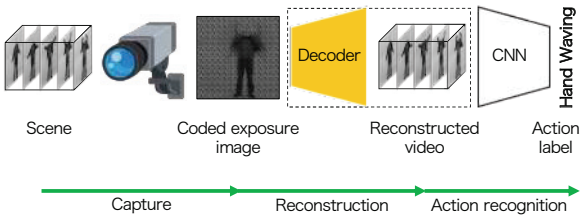
38

Action recognition from a single coded image

Tasashi Okawara, Michitaka Yoshida,
Hajime Nagahara, Yasushi Yagi
ICCP2020

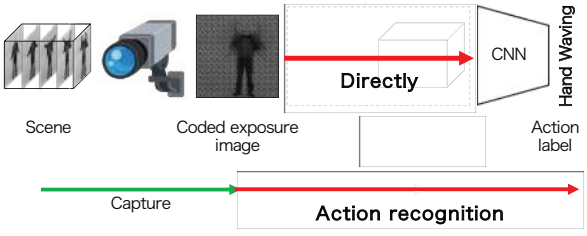
39

Apply compressive video sensing to video analysis



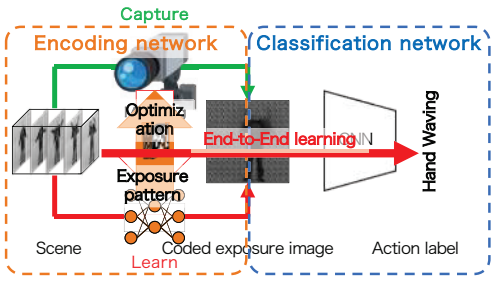
40

[Proposed] Action recognition from a single coded image



41

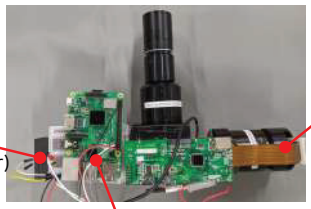
[Proposed] Joint optimization of encoding and classification networks



42

Prototype camera for real experiments

Coded exposure camera

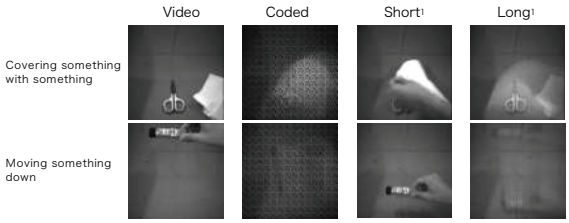


43

Dataset in real environment

- Chose 25 actions from the Something-Something V2 dataset
- Captured 100 coded images and 100 videos with 4 objects per action

Captured image and videos



1: generate from a video

44

Results

			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

45

Privacy Preserving Action Recognition via Motion Difference Quantization

Sudhakar Kumawat and Hajime Nagahara

ECCV2022

46

Privacy-preserving action recognition

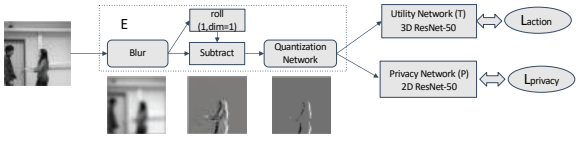
Some methods for privacy-preserving action recognition.

Goal: To develop an efficient encoder for the camera system that allows important features for action recognition while protecting actor(s) visual privacy.

47

Blur Difference Quantization

- Finding good balance of action recognition and image quality by adversarial learning



Given a frame v_t, we define a video as $V = \{v_t | i = 1, 2, .., t\}$ where t is the number of frames.

Blur module: $B_{v_t} = G_\sigma v_t$ where $G_\sigma = \frac{1}{2\pi\sigma^2} exp[-\frac{x^2+y^2}{2\sigma^2}]$

Difference module: $D(B_{v_t}, B_{v_1}] = B_{v_t} - B_{v_1}$

Quantization module: $Q(D(B_{v_t}, B_{v_1})) = \sum_{n=1}^{N-1} \sigma(H(D(B_{v_t}, B_{v_1}] - b_i))$, where $N = 16, 1 \cdot$ Temperature, $b_i = [0.5, 1.5, \ldots, N-1.5]$

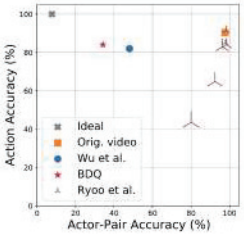
BDQ Training: Repeat following two steps iteratively until convergence.

1. Fix P, train E and T using loss function $\mathcal{L}(V, \theta_E, \theta_T) = \mathcal{XE}(T(E(V)), L_{action}) - \alpha\mathcal{E}(P(E(V)))$
2. Fix E and T, train P using loss function $\mathcal{L}(V, \theta_P) = \mathcal{XE}(P(E(V)), L_{privacy})$

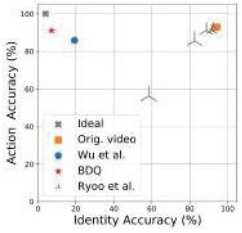
48

Evaluation on

SBU dataset (Actions-8, Privacy-13)



KTH dataset (Actions-6, Privacy-25)



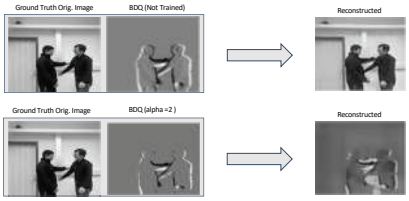
Wu et al. PAMI2020 used a UNet like encoder-decoder for video degradation.
Ryoo et al. AAAI2017 used downsampling for video degradation.

Method	Params.	Size	FLOPs
Wu et al	1.3M	3.8Mb	166.4G
BDQ	16	3.4Kb	120.4M

49

Reconstruction Attack

Visualization of reconstruction by a 3D UNet model (attacker) trained with input as BDQ output videos and labels as original videos.

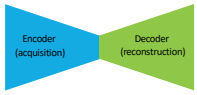


50

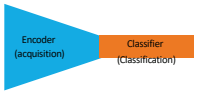
Insights of Deep Sensing

- Optimize the encoder to simulate the imaging devise.
- Finding the good 2D coded image as a bottleneck.

- Reconstruction
  - Important cues for reconstruction.
  - Omitting redundant information which decoder can be interpolated.



- Classification
  - Distinguishable features btw classes.
  - Filter out the redundant information.
  - Rich information is not always better: Cost, data rate, and privacy.



51

Conclusion

- Proposing deep sensing
  - Parameters for optical designs can be jointly optimized with decoder by same deep learning frame work.
- Demonstrating two example cases:
  - Compressive light field acquisition (4D→2D)
  - Compressive video sensing (3D -> 2D)
  - Compressive LF video sensing (5D-> 2D)
  - Action recognition directly from coded image (3D -> 174 classes)
  - Balancing the accuracy and privacy.

52