EXTRACTING KEY FRAMES FROM FIRST-PERSON VIDEOS IN THE COMMON SPACE OF MULTIPLE SENSORS

Yujie Li[1], Atsunori Kanemura[1,2], Hideki Asoh[1], Taiki Miyanishi[2], Motoaki Kawanabe[2]

[1]National Institute of Advanced Industrial Science and Technology (AIST) [2]Advanced Telecommunications Research Institute International (ATR)

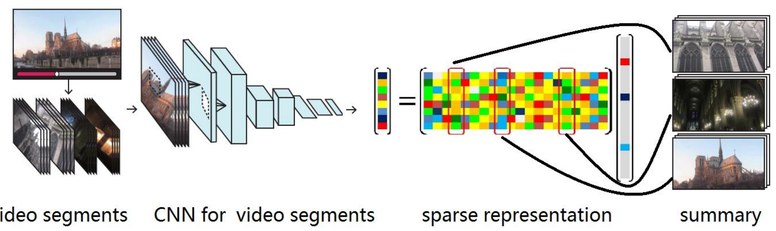
INTRODUCTION

Selecting authentic scenes about activities of daily living (ADL) is useful to support our memory of everyday life. Key-frame extraction for first-person vision (FPV) videos is a core technology to realize such memory assistant.

FPVs
√ Shift background
√ With much noise
√ Only parts of subjects

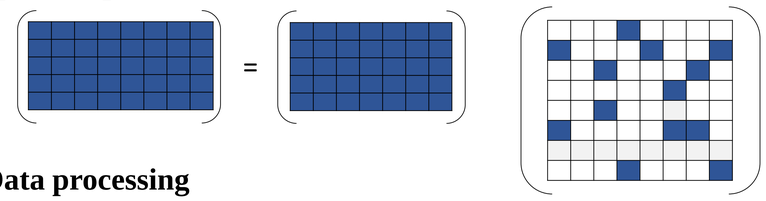
Normal videos
√ Stable background
√ With less noise

➢The overview of the video feature selection.



PROPOSED METHOD

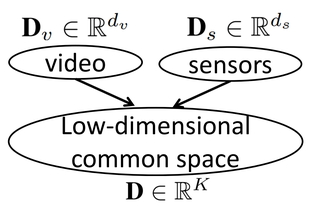
➢Sparse representation: Y=DH.



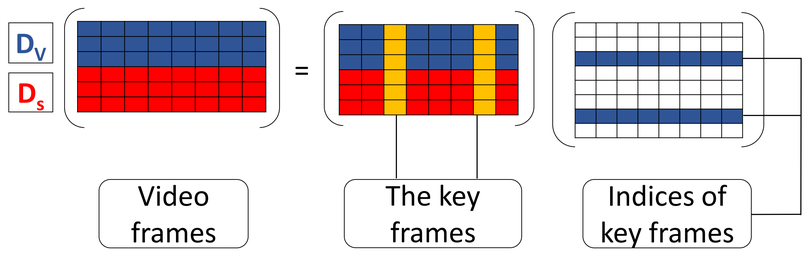
➢Data processing

In our situation, we use (Probabilistic Canonical Correlation Analysis) PCCA to embed the multi-sensor integration data (video and sensors) to a common space.

$$\min\left(\frac{1}{2}\sum_{n_v=1}^{N_v}\sum_{n_s=1}^{N_s} w_{ij}^{n_v n_s}\|\mathbf{A}^v\mathbf{x}_i^v-\mathbf{A}^s\mathbf{x}_j^s\|^2 + \frac{1}{2}\sum_{n_s=1}^{N_s}\sum_{n_v=1}^{N_v} w_{ij}^{n_s n_v}\|\mathbf{A}^s\mathbf{x}_i^s-\mathbf{A}^v\mathbf{x}_j^v\|^2\right)$$



➢Sparse model based key frame extraction.



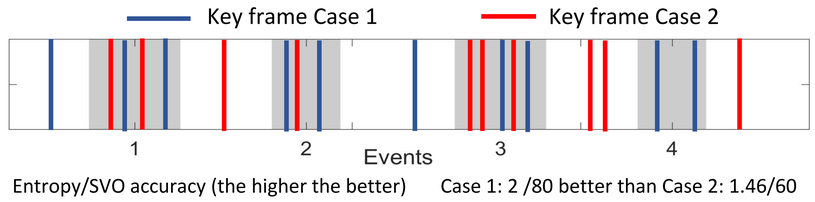
$$\min_{\mathbf{H}}\|\mathbf{Y}-\mathbf{Y}\mathbf{H}\|_2^2+\alpha\|\mathbf{H}\|_{1,2} \quad \text{s.t. } \mathbf{1}^T\mathbf{H}=\mathbf{1}^T, \text{where } \|\mathbf{H}\|_{1,2}=\sum_{i=1}^{N}\|\mathbf{h}_i\|_2$$

EXPERIMENTS AND DISCUSSION

➢Evaluation

We use accuracy (A) and entropy (S) as metrics for the extraction experiments.

$$A=\frac{N_{\text{Correct}}}{N_{\text{Whole}}} \qquad S=-\sum_{i=1} p_i \log_2 p_i$$



Entropy/SVO accuracy (the higher the better) Case 1: 2 /80 better than Case 2: 1.46/60

➢Experimental results.

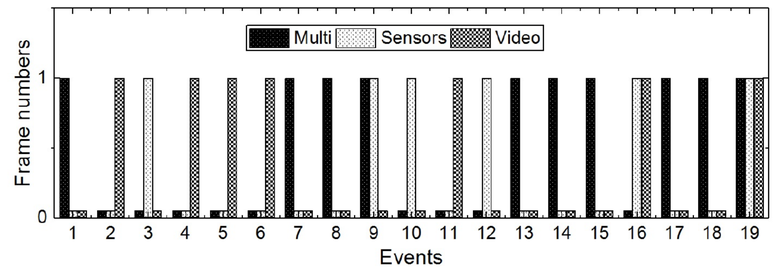
Table 1. The accuracy various kinds of information

Session	α	Video	Sensors	Multi
1	$4\sqrt{2}$	0.15	0.13	**0.23**
2	4	0.21	0.16	**0.23**
3	$8\sqrt{2}$	**0.24**	0.19	0.22
4	$8\sqrt{2}$	**0.19**	0.07	0.15
5	$4\sqrt{2}$	0.15	0.13	**0.24**
6	$4\sqrt{2}$	0.13	0.15	**0.20**
7	$8\sqrt{2}$	**0.22**	0.08	**0.22**
8	$8\sqrt{2}$	0.13	0.13	**0.15**
9	8	0.13	0.18	**0.20**
10	$2\sqrt{2}$	0.10	0.10	**0.14**

Table 2. The entropy various kinds of information

Session	α	Video	Sensors	Multi
1	$4\sqrt{2}$	2.81	2.59	**3.46**
2	4	2.81	2.81	**3.00**
3	$8\sqrt{2}$	**3.59**	3.32	3.32
4	$8\sqrt{2}$	**3.32**	2.00	3.00
5	$4\sqrt{2}$	3.32	3.17	**3.46**
6	$4\sqrt{2}$	2.32	2.59	**3.00**
7	$8\sqrt{2}$	**3.46**	2.00	3.32
8	$8\sqrt{2}$	2.59	**2.81**	**2.81**
9	8	2.81	3.17	**3.32**
10	$2\sqrt{2}$	2.00	2.00	**2.32**

➢The number of representatives found by our method for each of the events in the video.



CONCLUSION

➢We proposed a novel framework for key frame extraction of FPVs by sparse modeling representation selection from multisensory integration.

➢We use video features from DNN instead of raw frames.

➢Experimental results show that our proposed approach achieves modest improvements over a pure video information and the accuracy and entropy results predict the efficient of the proposed algorithm.

➢Moving forward, we plan to improve our method by using other nonvideo information such as audio and eWatch information.

This study was supported in part by NEDO, JST CREST JPMJCR15E2, JST SICORP, and JSPS KAKENHI 26730130 and 15K12112.