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), Tsukuba, Japan
{yujie-li, atsu-kan, h.asoh}@aist.go.jp
[2]Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan
{atsu-kan, miyanishi, kawanabe}@atr.jp

ABSTRACT

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. However, most existing key-frame extraction methods have mainly focused on stable scenes not related to ADL and only used visual signals of the image sequence even though the activities usually associate with our visual experience. To deal with dynamically changing scenes of FPV about daily activities, integrating motion and visual signals are essential. In this paper, we present a novel key-frame extraction method for ADL, which integrates multi-modal sensor signals to temper noise and detect salient activities. Our proposed method projects motion and visual features to a shared space by a probabilistic canonical correlation analysis and selects key frames there. The experimental results using ADL datasets collected in a house suggest that our key-frame extraction technique running in the shared space improves the precision of extracted key frames and the coverage of the entire video.

Index Terms— Video summarization, sparse estimation, multi-sensors, convolution neural network (CNN), probabilistic canonical correlation analysis (PCCA).

1. INTRODUCTION

With the omnipresence of smartphones and other devices for capturing and storing daily activities, more and more multiple information data (video, audio, pulse, etc.) is simultaneously obtained. For instance, almost one third of the people online use YouTube to upload or review videos [1]. This increasing popularity of Internet videos has accelerated the demand for efficient video retrieval [2]. To efficiently find and access the key information becomes a challenge. Nowadays video summarization has become the key tool for efficient browsing, access and manipulation of large video collections.

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

Much progress has been made in developing a variety of ways to summarize a single video, by exploring different design criteria in an unsupervised manner, or developing supervised algorithms. However summarizing videos still remains as a challenging problem for the video from with multiply sensors information. In this paper, we focus on the task of summarizing videos from the camera along with the information from non-video sources, such as audio and sensors, which is an insignificant task of computer visual.

Key frames are a subset of all still frames extracted from different video shots [3], and can be theoretically defined as the most informative and representative frames that reflect most of the visual contents in one video [4–7]. Some novel and improved key frame extraction methods are presented recently. X. Liu et al. [8] proposed a method based on Maximum a Posteriori (MAP) to estimate the positions of key frames. N. Ejaz et al. [3] proposed an aggregation mechanism to combine the visual features extracted from the correlation of RGB color channels, the color histogram and the moments of inertia to extract key frames [4]. In 2003, E. Elhamifar et al. proposed Sparse Modeling Representation Selection (SMRS) [9], which is an efficient method for video summarization and classification [10]. However, the direct application of SMRS produced null-information frames because of the instability and noise in FPVs, and our simple technique extends SMRS for better key frame extraction from FPVs.

Our task is to deal with multi-sensor data, one of the methods for associating data from multiple modalities is to embed these data instances into a common real-valued vector space [11–15]. Canonical Correlation Analysis (CCA) [16] is a standard method of multivariate analysis in statistics for achieving embedding two sets of data vectors into a common space by linear projections. PCCA [17] is probabilistic interpretation of CCA.

1.1. Contributions

Most existing methods for key frame extraction is only using the video information, and cannot efficiently used for first per-

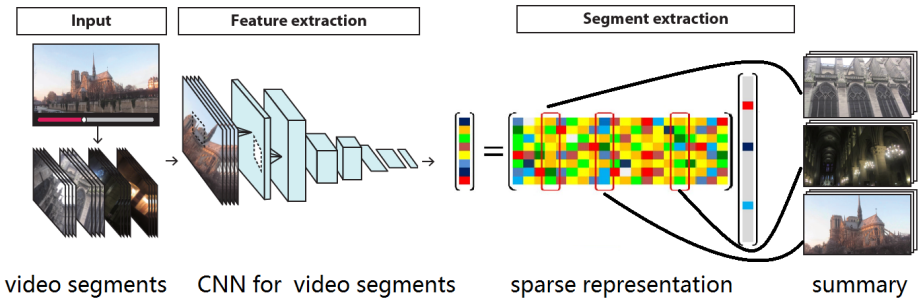

Fig. 1. The overview of the video feature selection.

son video, which are obtained by a wearable camera. The direct application of SMRS, which is efficiently used to process general videos, produced null-information frames because of the instability and noise in FPVs. Our idea is to use multi-sensor integration other than FPVs to reliably find key frames.

For reducing the computational complexity and enhancing the effectiveness, we use video features instead of raw images. We use a sliding window method to obtain the video features, which consisted of successive image features in a window. For feature extraction, we use a pre-trained model of VGG [18], which can produce discriminative visual features. Then we embed video features data and sensors data into a common space using PCCA.

2. MODEL AND FORMULATION

2.1. Video Feature Selection

Figure 1 shows an overview of our approach for video summarization. We first extract uniform length video segments from the input video in a temporal sliding window manner and compute their deep semantic features using a trained DNN. Inspired by the paper from DeMenthon [19], we represent the input video as a sequence of deep features in the semantic space, each of which corresponds to a video segment. This representation can encode the semantic transition of the video and thus can be useful for various tasks including video retrieval, video description generation, etc. Some clusters can be observed, each of which are expected to contain semantically similar video frames. Based on this assumption, we can pick out a subset of video frames by optimizing an objective function of the sparse representation. In this paper, we use video features, which are obtained by DNN, as video information data instead of raw images. For feature extraction, we used a pre-trained model of VGG [18], which can produce discriminative visual features.

2.2. Selection with Multi-sensors Integration

In our situation, we use PCCA to embed the multi-sensor integration data (video and sensors) to a common space (Fig. 2). Let $\mathbf{X}^{\mathrm{v}}$ as video data and $\mathbf{X}^{\mathrm{s}}$ as sensors data. Linear projections $\mathbf{A}^{v}$, $\mathbf{A}^{\mathrm{s}}$ from video domain and sensors domain to common space with the same dimension, which can be obtained by the following objective function

$$\min\Bigg(\frac{1}{2}\sum_{n_{\mathrm{v}}=1}^{N_{\mathrm{v}}}\sum_{n_{\mathrm{s}}=1}^{N_{\mathrm{s}}} w_{ij}^{n_{\mathrm{v}}n_{\mathrm{s}}}\|\mathbf{A}^{\mathrm{v}}\mathbf{x}_i^{\mathrm{v}}-\mathbf{A}^{\mathrm{s}}\mathbf{x}_j^{\mathrm{s}}\|^2 + \frac{1}{2}\sum_{n_{\mathrm{s}}=1}^{N_{\mathrm{s}}}\sum_{n_{\mathrm{v}}=1}^{N_{\mathrm{v}}} w_{ij}^{n_{\mathrm{s}}n_{\mathrm{v}}}\|\mathbf{A}^{\mathrm{s}}\mathbf{x}_i^{\mathrm{s}}-\mathbf{A}^{\mathrm{v}}\mathbf{x}_j^{\mathrm{v}}\|^2\Bigg) \quad (1)$$

We can estimate optimal projection matrices $\mathbf{A}^{\mathrm{v}}$ and $\mathbf{A}^{\mathrm{s}}$ by simply solving an eigenvalue problem. Once the projection matrices $\mathbf{A}^{\mathrm{v}}, \mathbf{A}^{\mathrm{s}}$ have been learned, we can utilize these matrices for projecting data vectors from video domain and sensors domain into the K-dimensional common space where the relevant pairs of information get close.

2.3. Sparse representation for frames selection

Signal processing applications are typically concerned with only a specific subset (or family) of signals which forms the informative content. With the development of mathematics, linear representation methods have been well studied and received considerable attention [20,21]. Recently, the most representative methodology of the linear representation methods is Sparse representation [22–25]. The basic of sparse representation is to represent a signal with a linear combination of atoms from a dictionary. Namely, the signal can be represented as a linear combination of a few dictionary atoms.

We assume there is a set of informative matrix $\mathbf{Y}$, which can also be the multi-sensor data of $\mathbf{X}^{\mathrm{v}}$ and $\mathbf{X}^{\mathrm{s}}$. It can be approximate by a linear combinations of very few atoms $\mathbf{w}$ of the dictionary $\mathbf{W}$. Here "few" means "sparse" in sparse representation. In other words, the signal matrix $\mathbf{Y}$ can be represented the multiplication of the dictionary matrix $\mathbf{W}$ and a corresponding sparse coefficient matrix $\mathbf{H}$ as $\mathbf{Y} = \mathbf{WH}$, where $\mathbf{H}$, which is sparse, contains the representation coefficients of signals matrix $\mathbf{Y}$ with respect to the dictionary $\mathbf{W}$.

The sparsest representation with the fewest number of nonzero coefficients is currently one of the most appealing

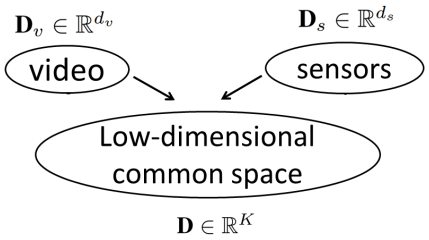

Fig. 2. Graphical model for canonical correlation analysis.

representations. The sparsest representation is the solutions of the following optimization problems,

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{Y} - \mathbf{W}\mathbf{H}\|^2 + \lambda \|\mathbf{H}\|_0 \tag{2}$$

where $\|\cdot\|_0$ denotes l_0-norm as sparsity constraint, namely $\|\mathbf{h}\|_0$, which counts the nonzero entries of a vector h. λ is a regularization parameter which can be adjusted for controlling the tradeoff between the approximation error and the sparsity of the coefficient matrix $\mathbf{H}$.

Naturally, one can find the sparsest representation by solving either of the above problem. Considering the fact that the l_0-norm optimization problem is generally NP-hard that is quite difficult to solve. Thus, approximate solutions are considered instead, and several efficient algorithms have been proposed. The simplest ones are Matching Pursuit (MP) and orthogonal matching pursuit (OMP). However, these methods cannot provide good enough estimation and are normally cannot suitable for high-dimensional problems. The relaxation method uses l_1-norm instead of l_0-norm can achieve large-scale optimization problem [26]. The relaxation objective function with l_1-norm can be expressed as follows,

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{Y} - \mathbf{W}\mathbf{H}\|^2 + \lambda \|\mathbf{H}\|_1 \tag{3}$$

l_1-norm regularized problem which has increasing popularity for the reason that the problem is convex quadratic optimization so it can be efficiently solved using Basis Pursuit (BP). Recently, gradient method is an very effective approach to solve l_1-norm optimization. Gradient methods are also called first-order methods or iterative soft-thresholding methods, which need more iterations especially if the sparsity of the solution is not enough or the initialization is not ideal.

We aim to select the key frames from the video, Each frame from the video can be expressed as a linear combination of a few common key frames selected from the entire frames in the video. We employ sparse representation to select key frames from video via convex optimization to minimize the reconstruction error. For this propose, the sparse representation formulation can be rewritten as follows,

$$\min_{\mathbf{H}} \|\mathbf{Y} - \mathbf{Y}\mathbf{H}\|^2 \quad \text{s.t. } \|\mathbf{H}\|_{0q} < s \tag{4}$$

where the mixed l_{0q}-norm is defined as

$$\|\mathbf{H}\|_{0q} = \sum_{n=1}^{N} I(\|\mathbf{h}_n\|_q > 0), \tag{5}$$

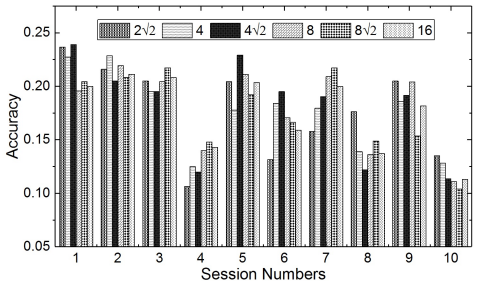

Fig. 3. The accuracy of training database various β using the multi-information 2: video and audio.

where $\mathbf{h}_n$ is the n-th row of the matrix $\mathbf{H}$ and I denotes the indicator function. Namely, $\|\mathbf{H}\|_{0q}$ counts the nonzero rows of the sparse representation $\mathbf{H}$. The indices of the nonzero rows of $\mathbf{H}$ correspond to the indices of the columns of $\mathbf{Y}$ which are chosen as the data representation. We want the selection of the representation to be invariant with respect to the global translation of the data, thus enforce the affine constraint $\mathbf{1}^{\mathrm{T}}\mathbf{H} = \mathbf{1}^{\mathrm{T}}$. Since l_0-norm is an NP-hard problem, we introduce l_1 relaxation of this optimization, which can be rewritten as,

$$\min_{\mathbf{H}} \|\mathbf{Y} - \mathbf{Y}\mathbf{H}\|^2 + \alpha \|\mathbf{H}\|_{1q} \quad \text{s.t. } \mathbf{1}^{\mathrm{T}}\mathbf{H} = \mathbf{1}^{\mathrm{T}}. \tag{6}$$

where

$$\|\mathbf{H}\|_{1,2} = \sum_{n=1}^{N} \|\mathbf{h}_n\|_2. \tag{7}$$

The solution of the above optimization program indicates the representatives as the nonzero rows of $\mathbf{H}$. The index of the nonzero rows of $\mathbf{H}$ refer to as the key frames of the video.

3. EXPERIMENTS AND DISCUSSION

We use a dataset which are collected the daily activities of eight subjects (not the researchers) in a house. Subjects wore wearable motion sensors, which contain three-axis accelerometers and gyroscopes, and a wearable camera. The subjects performed 20 daily activities at different places based on written instructions on a worksheet without direct supervision from the experimenters. For example, subjects "drink coffee" in the living room and "wash dishes" in the kitchen. For each subject, there are several sessions containing different activities of the subject. In total, the recorded sensor signals consist of 17 hours videos and motion data about 20 ADLs performed by eight subjects. The proposed method extracts key-frames from these FPV videos using both video and motion features [27].

3.1. Metrics

We use accuracy (A) to measure the efficient of our proposed method for key frame extraction from FPVs with Multi-

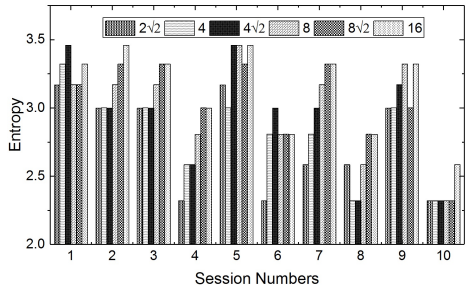

Fig. 4. The entropy value of training database various β using the multi-information 2: video and audio.

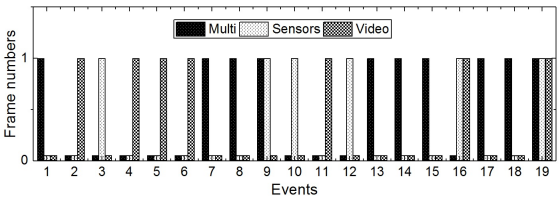

Fig. 5. The number of each events from Session 1.

Sensor Integration, which can be expressed as following,

$$A = \frac{N_{\text{Correct}}}{N_{\text{Whole}}}, \tag{8}$$

where N_{Correct} means the number of selected key frames which is correct chose regarding to the label, N_{Whole} means the number of key frames selected by the methods.

To evaluate the information content of events in the video, we introduce entropy as a metric for the extraction experiments, which can be expressed as follows.

$$S = -\sum_{i=1} p_i \log_2 p_i, \tag{9}$$

where p_i is the probability of each event M selected by our proposed method. The measure should be maximal if all the outcomes are equally likely. Thus, the higher entropy value is the key frames selection results are better.

3.2. Experimental Results

To investigate the effect of changing the regularization parameter α in the quality of obtaining representations, we run our proposed algorithm with $\alpha = 2, 2\sqrt{2}, 4, 4\sqrt{2}, 8, 8\sqrt{2}, 16$ to investigate the effect of various α. Fig. 3 shows the accuracy (A) results of selected frames for different values of α. Fig. 4 shows entropy results for different values of α. Then we choose the parameter α which has the highest value of accuracy and entropy as the optimal α. With the optimal α, we obtain key frames selection results with with different data.We first demonstrate the results only with video information. Next, we obtain key frames selection results only with the information from sensors, which contains accelerometers and gyroscopes. At last, we use the multi-information: the video information along with the sensors information. The results are shown in Table 1 and Table 2, from which we can see the key frames using multiple information can obtain higher value of entropy and accuracy.

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**

Then we calculate the number of representatives found by our method for each of the events in the videos using different information data. We take session 1 for example. Fig. 5 shows the results, from which we can see the key frames using multiple information can represent more events compared with pure video information and pure sensors information.

4. CONCLUSION

We proposed a novel framework for key frame extraction of FPVs by sparse modeling representation selection from multi-sensor integration. We use video features from DNN instead of raw frames. The index of the key frames is then produced, which proves to be more elegant and informative in representing the key frames of a FPV. 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 non-video information such as audio and eWatch information.

5. REFERENCES

[1] "Statistics – YouTube," 2016, https://www.youtube.com/yt/press/statistics.html.

[2] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," in *Asian Conf. Comput. Vis. (ACCV)*, 2016.

[3] N. Ejaz, T. B. Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *J. Vis. Commun. Image Represent.*, vol. 23, no. 7, pp. 1031–1040, 2012.

[4] Q. Xu, Y. Liu, X. Li, Z. Yang, J. Wang, Sbert M., and R. Scopigno, "Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based jensen divergence," *Inf. Sci.*, vol. 278, pp. 736–756, 2014.

[5] L. Pan, X. J. Wu, and X. Shu, "Key frame extraction based on sub-shot segmentation and entropy computing," in *Chin. Conf. Pattern Recognit. (CCPR)*, 2009.

[6] A. Kanemura, J. Yuan, and Y. Kawahara, "Finding structured dictionary representation by network-flow optimization," in *Workshop Data Discret. Segment. Knowl. Discov. (DDS)*, 2013.

[7] H. Wang, Y. Kawahara, C. Weng, and J. Yuan, "Representative selection with structured sparsity," *Pattern Recognit.*, vol. 63, pp. 268–278, 2017.

[8] X. Liu, M. L Song, L. M. Zhang, and S. L. Wang, "Joint shot boundary detection and key frame extraction," in *Int. Conf. Pattern Recognit. (ICPR)*, 2012.

[9] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012.

[10] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe, "Key frame extraction from first-person video with multi-sensor integration," in *IEEE Int. Conf. Multimedia and Expo (ICME)*, 2017.

[11] J. Sang, J. Liu, and C. Xu, "Exploiting user information for image tag refinement," in *ACM Int. Conf. Multimedia (ACMMM)*, 2011, pp. 1129–1132.

[12] C. Xu, D. Tao, and C. Xu, "A survey on multiview learning," *arXiv:1304.5634*, 2013.

[13] A. Lazaridou, E. Bruni, and M. Baroni, "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world," in *Annu. Meet. Assoc. Computat. Linguist. (ACL)*, 2014.

[14] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multiview embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[15] E. Bruni, N. K. Tran, and N. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res.*, vol. 49, no. 1, pp. 1–47, 2014.

[16] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[17] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep., Department of Statistics, University of California, Berkeley, 2005.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent. (ICLR)*, 2014.

[19] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *ACM Int. Conf. Multimedia (ACMMM)*, 1998, pp. 211–218.

[20] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[21] M. Huang, Wei Yang, Jun Jiang, Yao Wu, Yu Zhang, Wufan Chen, and Qianjin Feng, "Brain extraction based on locally linear representation-based classification," *NeuroImage*, vol. 92, pp. 322–339, 2014.

[22] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[23] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.

[24] K. Kreutz-Delgado, J. F. Murray, and B. D. Rao, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, pp. 349396, 2003.

[25] Z. Li, S. Ding, and Y. Li, "A fast algorithm for learning overcomplete dictionary for sparse representation based on proximal operators," *Neural Comput.*, vol. 27, no. 9, pp. 1951–198, 2015.

[26] D. L. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, 2003.

[27] T. Miyanishi, J. Hirayama, Q. Kong, T. Maekawa, H. Moriya, and T. Suyama, "Egocentric video search via physical interactions," in *AAAI Conf. Artif. Intell. (AAAI)*, 2016.