Egocentric Video Search via Physical Interactions

Taiki Miyanishi[†], **Jun-ichiro Hirayama**[†], **Quan Kong**[†‡],
Takuya Maekawa[†‡], **Hiroki Moriya**[†], **and Takayuki Suyama**[†]
[†]ATR Brain Information Communication Research Laboratory Group, Kyoto, Japan
[‡]Graduate School of Information Science and Technology, Osaka University, Osaka, Japan
{miyanishi, hirayama, kong, t.maekawa, moriyah, suyama}@atr.jp

Abstract

Retrieving past egocentric videos about personal daily life is important to support and augment human memory. Most previous retrieval approaches have ignored the crucial feature of human-physical world interactions, which is greatly related to our memory and experience of daily activities. In this paper, we propose a gesture-based egocentric video retrieval framework, which retrieves past visual experience using body gestures as non-verbal queries. We use a probabilistic framework based on a canonical correlation analysis that models physical interactions through a latent space and uses them for egocentric video retrieval and re-ranking search results. By incorporating physical interactions into the retrieval models, we address the problems resulting from the variability of human motions. We evaluate our proposed method on motion and egocentric video datasets about daily activities in household settings and demonstrate that our egocentric video retrieval framework robustly improves retrieval performance when retrieving past videos from personal and even other persons' video archives.

Introduction

Recent developments in wearable computing (Mann 1997) allow us to digitally capture everything that we have ever seen as well as our daily actions (Bell 2001). In particular, recorded egocentric images and videos of daily activities from wearable cameras are important to assist memory recollection for both memory-impaired and unimpaired persons (Berry et al. 2007; Hodges et al. 2006; Kalnikaite et al. 2010; Sellen et al. 2007). Since egocentric images and videos about daily activities are long and unstructured, the ability to retrieve past egocentric images and videos could support and augment human memory. The current egocentric image and video retrieval methods use manually and automatically labeled texts (Gemmell, Bell, and Lueder 2006; Hori and Aizawa 2003; Nakayama, Harada, and Kuniyoshi 2009) or images as user queries (Chandrasekhar et al. 2014). However, these approaches let users who need memory support describe what they forgot in their own words as user queries or prepare image queries similar to the past visual experience of what users want to remember. Furthermore, most previous methods have ignored the valuable feature of human-physical world interactions, which usually associate our daily activities and visual experience. For example, if you drink a cup of coffee, you might first look at your coffee cup. Therefore, the question about incorporating physical interactions into egocentric video retrieval is still open.

To address this question, in this paper, we propose a gesture-based egocentric video search framework that uses gesture motions as user queries. The underlying idea is that most of our experience in daily life is not explicitly verbalized but is associated with bodily behaviors. Body gestures associated with activities will thus be a natural modality for people to search for daily-life episodes. Moreover, body gestures can sometimes be easier to recall than query words for memory- or language-impaired persons. In fact, procedural memory, which involves the memory of how to perform actions, is reportedly unaffected by age when contrasted against semantic and episodic memories, which involve the memories of naming and autobiographical events (Hodges, Salmon, and Butters 1990; Light et al. 1992). By using the proposed framework, we can retrieve our past visual experience just by describing it with gesture motions. For example, if you want to remember something you ate last evening, you could retrieve and remember the past visual experience with an eating gesture. To achieve the gesture-based retrieval system, we introduce multimodal retrieval and re-ranking models.

Our proposed retrieval system works as in Fig. 1. First, the system simultaneously records motion and egocentric video about the daily activities of users on their personal digital archives using wearable devices. Second, the system learns the recorded motion and video matching using probabilistic canonical correlation analysis (PCCA), which models human-physical world interactions for indexing paired motions and videos. Third, we retrieve a motion-video pair for any given query using the similarity between motions in the learned latent space. Moreover, we re-rank retrieved videos in latent space to robustly improve retrieval performance. We assumed that by incorporating physical interactions into the retrieval models, we could solve the problems resulting from the variety of human motions related to daily activities, which leads to decreasing search performance.

Our experiments using motion and egocentric videos about daily activities collected in household settings reveal two main findings. (i) The proposed gesture-based egocentric

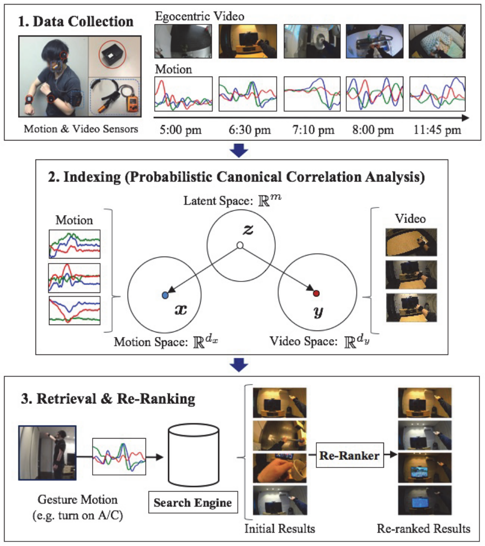

Figure 1: Our system retrieves videos from the egocentric video archive through a video search engine in response to a given gesture motion as a query.

video retrieval framework can retrieve past egocentric videos from personal and even another person's videos in response to gesture motions. (ii) Our re-ranking method modeling physical interactions robustly improves egocentric video retrieval performance.

The remainder of the paper is organized as follows. First, we provide some background in egocentric video research. Next, we present our egocentric video search framework and the re-ranking model. Finally, we present an experimental evaluation on motion and egocentric video datasets collected in house-like settings followed by our conclusions.

Related Work

Recently, there has been significant research interest in egocentric video such as activity recognition, summarization, and retrieval since the hardware development of a wearable camera enables capturing everyday activities of human life. To remember past human activities, egocentric activity recognition is useful. The existing egocentric vision-based activity recognition methods use segmented hand region (Fathi, Farhadi, and Rehg 2011), gaze location (Fathi, Farhadi, and Rehg 2011), and detected object (Ramanan 2012) as features of human activities. These approaches mainly classify egocentric videos using supervised techniques that require manually labeled training data, so that predictable activities are limited. In contrast, our egocentric video retrieval uses unsupervised techniques for retrieving videos in order to address various information needs. Egocentric summarization enables us to quickly browse the long-past visual experience. Current works use visual features focusing on important regions including people and objects (Ghosh 2012; Lee and Grauman 2015) and their story (Lu and Grauman 2013). Even though summarized videos can be a first step for retrieving important objects and activities, an information retrieval framework is essential in finding useful information from a vast amount of egocentric videos about daily life.

One recent egocentric retrieval study uses image annotation techniques to label images for text-based retrieval (Nakayama, Harada, and Kuniyoshi 2009). However, text-based retrieval methods only let users describe the past in concrete language. Another direction is using images or videos as user queries (Imura et al. 2011; Chandrasekhar et al. 2014), which involves a content-based image retrieval method (Smeulders et al. 2000). However, content-based image retrieval requires images or videos including objects and specific locations about what we want to remember as queries. These limitations defeat the purpose of supporting and augmenting human memory. Moreover, these egocentric retrieval methods ignore the important features of human-physical world interactions. In contrast, our proposed method can retrieve past videos anywhere and does not require texts, images or videos since we use gesture motion as a query. Moreover, the main topic of this paper is investigating the effectiveness of physical interaction for retrieving and re-ranking egocentric videos for improving retrieval performance.

Egocentric Video Search Framework

In this section, we present a framework of an egocentric video search system. Our retrieval system outputs an egocentric video in response to a given gesture motion based on a retrieval model constructed in advance. To construct the model, first we collect motions and videos of daily activities using motion sensors and a wearable camera. Then, we segment the motion and video pairs into a series of n observations $\mathcal{M} := \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1, 2, \ldots, n\}$, where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$ are the observed feature vectors of motion and video. We call $\mathcal{M}$ external memory for augmenting human memory.

Our goal is to retrieve optimal past video $\boldsymbol{y}^*$ from external memory $\mathcal{M}$ in response to given motion query $\boldsymbol{x}'$. The output $\boldsymbol{y}^*$ is obtained by maximizing a reproducibility function $R : \boldsymbol{x}' \mapsto (\boldsymbol{x}, \boldsymbol{y})$, i.e.,

$$f((\boldsymbol{x}^*, \boldsymbol{y}^*)) = \underset{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{M}}{\operatorname{argmax}}\, R(\boldsymbol{x}', (\boldsymbol{x}, \boldsymbol{y})). \tag{1}$$

The reproducibility function R, depending on the specific retrieval model we use, measures how current query $\boldsymbol{x}'$ reproduces each instance in the external memory $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{M}$.

Below, we design a reproducibility function using a probabilistic IR framework based on the probabilistic canonical correlation analysis (PCCA) (Bach and Jordan 2005). First, we introduce PCCA and then present the corresponding reproducibility function by extending PCCA into the IR context. Although the induced final formula is almost the same as CCD (Nakayama, Harada, and Kuniyoshi 2010), we naturally extend it to the re-ranking model that re-ranks motion-video pairs in latent space.

Many multi-modal retrieval frameworks that retrieve images using text queries (Jeon, Lavrenko, and Manmatha 2003; Guillaumin et al. 2009; Xu et al. 2015) have been proposed. However, we use the PCCA-based retrieval model, which easily combines motion and video features and can be easily extended to the re-ranking model.

Probabilistic CCA

PCCA has been widely used in information-matching tasks. We use PCCA to learn the shared representations of motion and video features, assuming that the shared linear representation approximately models the essential interaction between humans and the physical world. Furthermore, being a formal probabilistic model, PCCA gives a natural probabilistic IR framework for egocentric video retrieval.

Figure 1 (middle) shows our PCCA where paired motion and video features $(\boldsymbol{x}, \boldsymbol{y})$ are given by a linear model with shared latent variables $\boldsymbol{z} \in \mathbb{R}^m$ ($m \le \min\{d_x, d_y\}$): $\boldsymbol{x}=\mathbf{B}_x\boldsymbol{z}+\boldsymbol{\epsilon}_x, \boldsymbol{y}=\mathbf{B}_y\boldsymbol{z}+\boldsymbol{\epsilon}_y$. Both the latent and error variables are assumed to be Gaussian, such that $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m), \boldsymbol{\epsilon}_x \sim N(0, \boldsymbol{\Psi}_x), \boldsymbol{\epsilon}_y \sim N(0, \boldsymbol{\Psi}_y)$, where $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ are covariance matrices. Then, the posterior distribution of $\boldsymbol{z}$ given motion feature $\boldsymbol{x}$ and that given video feature $\boldsymbol{y}$ are obtained respectively by

$$\boldsymbol{z} \mid \boldsymbol{x} \sim \mathcal{N}(\mathbf{W}_x\boldsymbol{x}, \mathbf{V}_x), \quad \boldsymbol{z} \mid \boldsymbol{y} \sim \mathcal{N}(\mathbf{W}_y\boldsymbol{y}, \mathbf{V}_y), \qquad (2)$$

where $\mathbf{V}_x := (\mathbf{I}+\mathbf{B}_x^\top\boldsymbol{\Psi}_x^{-1}\mathbf{B}_x)^{-1}$, $\mathbf{V}_y := (\mathbf{I}+\mathbf{B}_y^\top\boldsymbol{\Psi}_y^{-1}\mathbf{B}_y)^{-1}$, $\mathbf{W}_x := (\mathbf{I} + \mathbf{B}_x^\top\boldsymbol{\Psi}_x^{-1}\mathbf{B}_x)^{-1}\mathbf{B}_x^\top\boldsymbol{\Psi}_x^{-1}$, and $\mathbf{W}_y := (\mathbf{I} + \mathbf{B}_y^\top\boldsymbol{\Psi}_y^{-1}\mathbf{B}_y)^{-1}\mathbf{B}_y^\top\boldsymbol{\Psi}_y^{-1}$. We estimate the parameters $\mathbf{B}_x$ and $\mathbf{B}_y$ by maximum-likelihood estimation that selects the best model and parameters to explain the simultaneously recorded pairs of motion and video features. The maximum likelihood estimates of $\mathbf{B}_x$, $\mathbf{B}_y$ are given (Bach and Jordan 2005) by $\mathbf{B}_x = \mathbf{C}_{xx}\mathbf{U}_x\mathbf{M}_x$, $\mathbf{B}_y = \mathbf{C}_{yy}\mathbf{U}_y\mathbf{M}_y$, where $\mathbf{C}_{xx} \in \mathbb{R}^{d_x \times d_x}$ and $\mathbf{C}_{yy} \in \mathbb{R}^{d_y \times d_y}$ are the sample covariance matrixes in the motion and video features. $\mathbf{M}_x, \mathbf{M}_y \in \mathbb{R}^{m \times m}$ are arbitrary matrices satisfying $\mathbf{M}_x\mathbf{M}_y^\top = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the first m canonical correlations. The columns in $\mathbf{U}_x$ and $\mathbf{U}_y$ are the corresponding canonical vectors obtained by conventional CCA (Hardoon, Szedmak, and Shawe-Taylor 2004) for motion and video features, respectively. We set them using diagonal matrices $\mathbf{M}_x = \mathbf{M}_y = \Lambda^{1/2}$, which equally weight the motion and video features.

Probabilistic CCA for Retrieval

To derive the reproducibility function R, we formulate the IR problem from a probabilistic point of view based on the PCCA model. According to Eq. (2), the latent vector $\boldsymbol{z}$ can be estimated as the posterior mean $\boldsymbol{z} = \mathbf{W}_x\boldsymbol{x}$ for any query $\boldsymbol{x}$. Now suppose that we have already obtained latent vectors $\boldsymbol{z}_i$ corresponding to all pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in the external memory $\mathcal{M}$, and assume virtually that any new latent query $\boldsymbol{z}' = \mathbf{W}_x\boldsymbol{x}'$ is stochastically generated from one of the $|\mathcal{M}|$ latent vectors $\boldsymbol{z}_i$ in the following manner: (i) first pick one instance $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in the external memory according to the probability of π_i, (ii) then, add a stochastic noise $\boldsymbol{r}$ to the corresponding latent vector $\boldsymbol{z}_i = \mathbf{W}_x\boldsymbol{x}_i$, so that we finally obtain $\boldsymbol{z}' = \boldsymbol{z}_i + \boldsymbol{r}$.

The IR problem can then be formulated as estimating $\boldsymbol{z}_i$, from which the new $\boldsymbol{z}'$ was generated. If $\boldsymbol{r}$ has the probability density function given by f, a reasonable approach is to maximize the posterior probability $p(\boldsymbol{z}_i \mid \boldsymbol{z}') = \pi_i f(\boldsymbol{z}' - \boldsymbol{z}_i)/\sum_i \pi_i f(\boldsymbol{z}' - \boldsymbol{z}_i)$ with respect to $i = 1, 2, \ldots, |\mathcal{M}|$. Here, we set the prior probability π_i as uniform, although more informative priors can also be available. Thus, we propose a generic form of reproducibility function R_f for any specific choice of f:

$$R_f(\boldsymbol{x}', (\boldsymbol{x}, \boldsymbol{y})) = \frac{f(\mathbf{W}_x\boldsymbol{x}' - \mathbf{W}_x\boldsymbol{x})}{\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} f(\mathbf{W}_x\boldsymbol{x}' - \mathbf{W}_x\boldsymbol{x})}. \qquad (3)$$

Note that with the uniform prior π_i, the problem is equivalent to minimizing $-\log f(\mathbf{W}_x\boldsymbol{x}' - \mathbf{W}_x\boldsymbol{x})$, which is typically done in practice.

We also design a specific f according to the PCCA model. We define it as f_{pcca}. The model implies that if given the query $\boldsymbol{x}_i$ without observing $\boldsymbol{y}_i$, the noise vector $\boldsymbol{r} = \boldsymbol{z}' - \boldsymbol{z}_i$ should follow $\mathcal{N}(\mathbf{0}, \mathbf{V}_x)$; alternatively, if given the target $\boldsymbol{y}_i$ without observing $\boldsymbol{x}_i$, it should follow $\mathcal{N}(\mathbf{0}, \mathbf{V}_y)$. Here, we combine the two different views in a simple manner:

$$f_{pcca}(\boldsymbol{r}) = \mathcal{N}(\boldsymbol{r} \mid \mathbf{0}, \mathbf{V}_x + \mathbf{V}_y), \qquad (4)$$

implying with Eq. (3) that the corresponding R is given by

$$\begin{aligned} R(\boldsymbol{x}', (\boldsymbol{x}, \boldsymbol{y})) &\propto f_{pcca}(\mathbf{W}_x\boldsymbol{x}' - \mathbf{W}_x\boldsymbol{x}) \qquad (5) \\ &\propto \exp\{-(\boldsymbol{z}' - \boldsymbol{z})^\top(\mathbf{V}_x + \mathbf{V}_y)^{-1}(\boldsymbol{z}' - \boldsymbol{z})\}, \end{aligned}$$

which can be seen as a similarity measure between $\boldsymbol{x}'$ and $\boldsymbol{x}$ in the latent space with the metric integrating the posterior uncertainties in two different views. Note that the last formula is almost the same as the CCD model (Nakayama, Harada, and Kuniyoshi 2010), which uses the KL-divergence between $p(\boldsymbol{z}_i|\boldsymbol{x}')$ and $p(\boldsymbol{z}_i|\boldsymbol{x})$. Naturally, we extend this formula into the egocentric video re-ranking model using the probability density function f_{pcca}.

Probabilistic CCA for Re-Ranking

To further improve retrieval performance, we extend the probabilistic CCA model into the video re-ranking framework. Although many re-ranking models have been proposed in the IR community, we used a pseudo-relevance feedback method based on a kernel density estimation (KDE) (Efron et al. 2014) since KDE can naturally incorporate multimodal features in the learned latent space. The past work used weighted kernel densities for re-ranking search target $\boldsymbol{x}$ according to

$$p(\boldsymbol{x}) = \frac{1}{m}\sum_{i=0}^{m} \omega_i f(\boldsymbol{x} - \boldsymbol{x}_i), \qquad (6)$$

where m is the number of pseudo-relevance documents used for re-ranking, and ω_i is the weight of a kernel $f(\cdot)$. We use *rank-based weights* $\omega_i = \lambda e^{-\lambda r_i}$, which weight search results with exponential decay regarding their ranks, where $\lambda = \frac{1}{\bar{r}}$, which is the maximum likelihood estimate of $\bar{r}$, is the mean of the ranks $1, 2, \ldots, m$. The *rank-based weights* are the most successful weights for re-ranking, and Gaussian kernel was used for calculating distance between retrieved search targets in the past work (Efron et al. 2014). However, this existing model only considers a single modality (i.e. motion in our case). We assumed the fusion of motions and videos leads to robustness for re-ranking results since visual feature are stable and temper the ambiguity of motions. To leverage

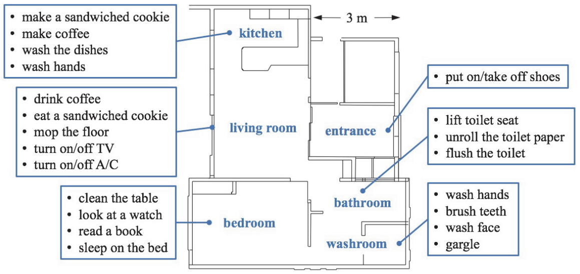

Figure 2: Room layout of experimental environment and 20 daily activities for a with-object task.

combined multi-modal features (i.e. motion and video), we use the posterior probability of $p(\boldsymbol{z}_i|\boldsymbol{z}) \propto f_{pcca}(\boldsymbol{z}_i - \boldsymbol{z})$ in the latent space produced by Probabilistic CCA following Eq. (4). Thus, we have a new KDE:

$$p(\boldsymbol{x}) = \frac{1}{m}\sum_{i=0}^{m} \omega_i f_{pcca}(\mathbf{W}_x\boldsymbol{x} - \mathbf{W}_x\boldsymbol{x}_i). \qquad (7)$$

This simple formula re-ranks search results in the latent space considering multi-modalities modeling physical interactions for re-ranking as well as the PCCA-based initial retrieval Eq. (5).

Temporal Diversification

Finally, we introduce a search result diversification technique. In our study, the proposed framework retrieved three seconds of video shifted by 0.1 seconds over all the videos. As a result, the search results are redundant. For example, the retrieval methods may retrieve a video about "drink coffee" at 10:24 and 17.5 seconds and another short video about "drink coffee" at 10:24 and 17.6 seconds. To avoid this problem, we diversified the search results using a simple temporal diversification technique. First, we retrieved video using the retrieval methods and obtained their search scores, which are the values of the reproducibility function in our case. Then, we segmented the retrieved video sets into a series of 30-second segments. We selected a high-scoring video from each segment and sorted them in decreasing order of their search scores. If duplicated videos exist in a single activity label due to temporally equal segregation, we also removed the low-scoring video. Then, we diversified search results of initial retrieval and re-ranked diversified results using our proposed re-ranking methods.

Evaluation

This section reports on the empirical evaluation for the proposed gesture-based egocentric video retrieval framework in terms of the retrieval effectiveness of the search results. We demonstrate that our framework can retrieve past visual experience and robustly improve retrieval performance by using the proposed re-ranking method that models physical interactions.

Data Sets

We built a dataset by collecting the daily activities of eight subjects (not the researchers) in a house. Even though the most natural data would be acquired from the normal daily lives of the subjects, collecting sufficient samples of such data in their own individual homes/apartments is too difficult. In this study, we used a semi-naturalistic collection protocol (Bao and Intille 2004) to collect more variable behavior data than in a laboratory setting.

Procedure Eight subjects whose ages ranged from 21 to 26 (mean = 23.13, SD = 1.69) wore wearable motion sensors, LP-WS1101, which contain three-axis accelerometers and gyroscopes, and a wearable camera, Panasonic HX-A100 (1280 × 720 pixels, 29.97 fps). They performed the 20 written activities at different places based on written instructions on a worksheet without direct supervision from the experimenters. The subjects performed relatively free activities in specified places related to them. For example, subjects "turn on/off TV" in the living room and "open/close the refrigerator" in the kitchen. We randomly shuffled the order of places where subjects did the daily activities in each session. We call this experiment a *with-object* task. Figure 2 shows the room layout of the experimental environments and lists the 20 daily activities at each place performed by the subjects in each session of the with-object task. A single session averaged 10.86 minutes (SD = 1.14) among the subjects. Sessions were repeated 12 times (including two initial practice sessions); they were allowed short breaks. No researcher supervised the subjects while they collected data under the semi-naturalistic collection protocol. We used the motion and video data from the 3rd to 12th sessions of the with-object task as the search target.

After the with-object task, to collect gesture motions for retrieving past activities we asked the subjects to remember and repeat 20 activities that they did in the with-object task experiments as gesture motions used for queries. This second experiment is called a *without-object* task. We explained the without-object task and gave more written instructions that listed the activities of the without-object task that consisted of 20 activities per session. Its activities were slightly different from the with-object task to complete each activity during specified times. For example, we added "pour hot water" and "stir a cup of coffee" instead of "make coffee" and removed "sleep on the bed." Subjects then repeated the 20 activities, this time without objects and in a new environment. All sessions were repeated six times including one practice session. Note that the 1st session in the without-object task was a practice session, so we removed it and used gesture motions from the 2nd to 6th sessions as queries.

Relevance Judgments Our goal is to return a relevant ranked list of egocentric videos about daily life using the gesture-based retrieval framework. To make daily activity datasets for retrieval evaluation, we specified 20 activities and their start and end points to the collected sensor dataset. Two annotators labeled the 20 activities listed on the worksheet of the without-object task and the sensor data collected in both the with/without-object tasks by watching 17 hours of egocentric video captured by cameras attached to the eight subjects. We obtained 1,982 labels for the with-object tasks and 799 for the without-object tasks. In terms of relevance judgments, each retrieved motion and video pair was judged with these labels. We defined a relevant pair as one that temporally overlaps more than 50% (overlapping more than 1.5 seconds) of the corresponding activity label.

Feature Extraction

In this section, we explain the extraction of the motion and video features. These features were used for the input of our proposed egocentric video retrieval system.

Motion Feature For the motion feature extraction, first we down-sampled the acceleration and gyro signals from 50 to 25 Hz to denote 25 samples a second. Then, we used a moving average with four overlapping samples to smooth the signals. To obtain temporal feature of motions, we also applied a short time Fourier transform (STFT) to the smoothed signals with a sliding window. The window width was set to 75 samples and was shifted by one sample. Then, we down-sampled the transformed signals from 25 to 10 Hz to align the sampling rate to the video features. Note that each motion feature sample had 684 dimensions. Finally, we standardized the features with the mean and the variance.

Video Feature We used a sliding window method to obtain the video features, which consisted of successive image features in a window. To extract the image features, we used Caffe,[1] which is a well-known deep learning framework, and prepared several pre-trained models. For feature extraction, we used a pre-trained model of VGG (Simonyan and Zisserman 2014), which can produce discriminative visual features. We used the activations in the second to last fully connected layer as image features. As a result, we obtained 4,096 dimensions per frame extracted from the egocentric video. Then, we applied principal component analysis (PCA) to the extracted image features and reduced the dimensions from 4,096 to 250. Then, to align the sampling rate to the motion features, we down-sampled all of the transformed image features from 29.97 to 10 Hz by a rolling mean of the time intervals. We combined the image features by three seconds by shifting one sample. Each video feature sample had 7,500 dimensions. Finally, we standardized the video features with the mean and the variance.

Baselines

Our approach first conducts initial retrieval according to the value of the reproducibility function using the probability function of Eq. (5) and re-ranks search results by the KDE of Eq. (7) after temporal diversification. We denote the proposed initial retrieval and re-ranking methods as MR + PCCA and KDE (PCCA), respectively. Note that even though MR + PCCA is almost the same as past work (Nakayama, Harada, and Kuniyoshi 2010), the effectiveness of a PCCA-based method for gesture-based egocentric video retrieval is still unclear.

To evaluate our retrieval method, we also prepared several baseline methods. The first baseline retrieves and re-ranks motion and video pairs using the following Gaussian kernel in the original motion space instead of $f_{pcca}(\cdot)$ used in MR + PCCA and KDE (PCCA).

$$f_{gauss}(\boldsymbol{x}_1 - \boldsymbol{x}_2) \propto \exp\{-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\}.$$

We denote these initial retrieval and re-ranking methods as MR and KDE, respectively. Note that KDE is the same as the reranking method (Efron et al. 2014) when the kernel bandwidth is set to one. By comparing these with PCCA and KDE (PCCA), we can quantify the benefit of combining the motion and video features and retrieve motion and videos pairs in the learned latent space. The second baseline uses the following probability density, which uses the latent space learned by a standard canonical correlation analysis (CCA) instead of $f_{pcca}(\cdot)$ used in MR + PCCA and KDE (PCCA).

$$f_{cca}(\dot{\mathbf{W}}_x\boldsymbol{x}_1 - \dot{\mathbf{W}}_x\boldsymbol{x}_2) \propto \exp\{-\|\dot{\mathbf{W}}_x\boldsymbol{x}_1 - \dot{\mathbf{W}}_x\boldsymbol{x}_2\|\},$$

where $\dot{\mathbf{W}}_x$ is a transformed matrix learned by CCA in the motion view. By comparing this with PCCA, we can quantify the benefit of considering the uncertainty for modeling physical interactions. We denote this initial retrieval and re-ranking as MR + CCA and KDE (CCA), respectively.

For CCA and PCCA, we tuned the parameter as the number of dimensions d in the latent space learned by CCA and PCCA. For KDE , KDE (CCA) KDE (PCCA) , we tuned the feedback motion and video pairs m. We used leave-one-subject-out cross-validation to tune these parameters among candidates $d = \{50, 100, 150, 200, 250, 300\}$ and $m = \{1, 2, 4, 8, 16, 32\}$, which are optimized for the best performance of the average precision on the validation data of seven subjects (without involving a target subject), and tested it with the target subject dataset.

Experimental Results

In the experiments, what we want to evaluate in this study is the retrieval effectiveness of the ranked list when retrieving and re-ranking past videos from personal and another person's videos. Thus, we evaluated the search results in two conditions: the inner-subject and cross-subject conditions. The inner-subject condition assumed that users retrieved past egocentric video from their personal video archives. Under this condition, we used the motion queries and videos from each subject. The cross-subject condition assumed that the users retrieved egocentric video from the video archives of others. Under this condition, we retrieved the videos of a single subject with multiple subject queries and averaged their evaluation results. We averaged the evaluation results over the queries of each activity by each subject, and so we evaluated the retrieval methods using 160 samples (20 activities × 8 subjects) for motion queries under both conditions.

To evaluate the retrieval effectiveness, we used average precision (AP), which is the mean of the precision scores obtained after each retrieved relevant video. We discuss statistical significance of results using a two-tailed paired t-test with $p < 0.05$ on 160 samples using Bonferroni correction with the number of subjects for multiple testing. To compare our proposed methods to baselines, we use this evaluation framework under both inner/cross-subject conditions throughout this paper.

[1] http://caffe.berkeleyvision.org/

Table 1: Performance comparison of initial retrieval when using the proposed methods and baselines under the inner-subject and cross-subject conditions. The best performing run is indicated in bold and statistically significant differences are marked using the symbols in the top-right corner of each method name.

Method	Inner-Subject	Cross-Subject
MR$^{\heartsuit}$	0.2699	0.1477
MR + CCA$^{\spadesuit}$	0.3253$^{\heartsuit}$	0.1856$^{\heartsuit}$
MR + PCCA	**0.3557**$^{\heartsuit\spadesuit}$	**0.2087**$^{\heartsuit\spadesuit}$

Table 2: Performance comparison of the proposed methods and baselines under the inner-subject and cross-subject conditions when re-ranking results of MR+PCCA. The best performing run is indicated in bold and statistically significant differences are marked using the symbols in the top-right corner of each method name.

Method	Inner-Subject	Cross-Subject
MR + PCCA $^{\diamondsuit}$	0.3557	0.2087
+ KDE$^{\heartsuit}$	0.3554	0.2076
+ KDE (CCA)$^{\spadesuit}$	0.3697$^{\heartsuit}$	0.2187$^{\diamondsuit\heartsuit}$
+ KDE (PCCA)	**0.3822**$^{\diamondsuit\heartsuit\spadesuit}$	**0.2238**$^{\diamondsuit\heartsuit\spadesuit}$

Re-ranking Performance Table 2 shows the retrieval performance under both inner/cross-subject conditions when using our proposed re-ranking methods. We used initial retrieval as MR + PCCA for all methods. The results show that the proposed MR + PCCA + KDE (PCCA) significantly outperformed all baselines with statistical significance. The single modality re-ranking approach MR + PCCA + KDE decreases retrieval performance compared to the initial retrieval. Moreover, from Fig. 3, our PCCA-based retrieval and re-ranking consistently improved retrieval performance over all subjects compared to KDE and KDE (CCA). The results suggest that the PCCA-based re-ranking model boosts retrieval performance when re-ranking personal and another person's past egocentric videos. Figure 4 shows the top five retrieved videos with MR, MR + PCCA, and our method MR + PCCA + KDE (PCCA) using gesture motions "clean the table" and "make a sandwiched cookie". For both cases, MR + PCCA using PCCA-based retrieval improved results of MR. Furthermore, MR + PCCA + KDE (PCCA) successfully re-ranked relevant videos at the top since some relevant video already exists at initial retrievals MR + PCCA.

Initial Retrieval Performance Table 1 shows the initial retrieval performance under both inner/cross-subject conditions. Multimodal methods MR + CCA and MR + PCCA remarkably outperformed single modality method MR with statistical significance under both inner/cross-subject conditions, which suggests that modeling physical interactions is important for improving gesture-based egocentric video retrieval performance when retrieving videos from personal and another person's video archives. MR + PCCA also outperformed the CCA-based approach MR + CCA with statistical significance indicating the effectiveness of Probabilistic CCA when modeling physical interactions for the gesture-based egocentric video retrieval.

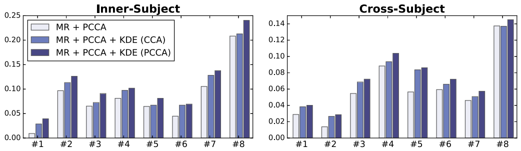

Figure 3: Improvement of AP about each subject from the initial retrieval MR under both inner-subject (left) and cross-subject (right) conditions. The x-axis shows subject ID. The y-axis shows the AP improvements.

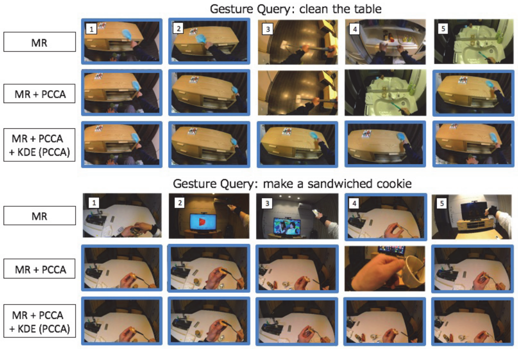

Figure 4: Example of search results retrieved by the methods MR, MR + PCCA, and MR + PCCA + KDE (PCCA) using gesture motions: *"clean the table"* (upper) and *"make a sandwiched cookie"* (lower). Retrieved videos are ordered by search score from left to right. Blue boxes are relevant videos.

Robustness We showed that KDE (PCCA) significantly improves retrieval performance against the initial search results MR + PCCA and other baselines on averaged evaluation measures. In this section, we demonstrated the robustness of our proposed method, which is defined as the number of queries improved/degraded as the results of applying these methods (Metzler and Croft 2007). A highly robust re-ranking technique will significantly improve many queries and only minimally degrade very few. Figure 5 shows the retrieval performance (i.e. AP values) of each query when using KDE, KDE (CCA), and KDE (PCCA), which suggests KDE (PCCA) is a robust re-ranking method compared to the baselines. For example, KDE (PCCA) improved search results for 17 out of 20 queries in both inner/cross-subject conditions over the initial retrieval MR + PCCA, whereas KDE and KDE (CCA) improved results over the initial retrieval MR + PCCA for nine and 11 queries in the inner-subject condition and for eight and 15 queries in the cross-subject condition, respectively. Moreover, KDE (PCCA) improved results for 16 queries over KDE (CCA) in both conditions. These results indicate that our proposed PCCA-based multimodal re-ranking method robustly improves gesture-based egocentric video retrieval by modeling physical interactions.

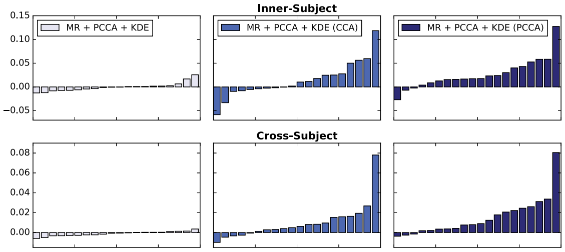

Figure 5: Improvement of AP about each query from the initial retrieval MR + PCCA under both inner-subject (upper) and cross-subject (lower) conditions. The x-axis shows queries ordered by AP improvements. The y-axis shows the AP improvements.

Conclusion

In this paper, we proposed a novel framework for egocentric video search using gesture motions as queries. Our proposed framework used the probabilistic IR model, which fuses motion and video features by introducing probabilistic canonical correlation analysis. We induced the similarity between a given motion and the past motion and video pairs in canonical space and naturally extended the retrieval model into the re-ranking method. The experimental results show that our proposed gesture-based egocentric video search framework can retrieve past egocentric video using gesture motions and robustly improve retrieval performance when re-ranking search results in the latent space learned by PCCA.

Acknowledgments

This work was supported by CREST and SICORP from JST, JSPS KAKENHI 15H02759, and ImPACT from CSTI.

References

Bach, F. R., and Jordan, M. I. 2005. A probabilistic interpretation of canonical correlation analysis. Technical Report 668, Department of Statistics, University of California, Berkeley.

Bao, L., and Intille, S. S. 2004. Activity recognition from user-annotated acceleration data. In *Pervasive*, 1–17.

Bell, G. 2001. A personal digital store. *Communications of the ACM* 44(1):86–91.

Berry, E.; Kapur, N.; Williams, L.; Hodges, S.; Watson, P.; Smyth, G.; Srinivasan, J.; Smith, R.; Wilson, B.; and Wood, K. 2007. The use of a wearable camera, sensecam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: a preliminary report. *Neuropsychological Rehabilitation* 17(4-5):582–601.

Chandrasekhar, V.; Tan, C.; Min, W.; Liyuan, L.; Xiaoli, L.; and Hwee, L. J. 2014. Incremental graph clustering for efficient retrieval from streaming egocentric video data. In *ICPR*, 2631–2636.

Efron, M.; Lin, J.; He, J.; and de Vries, A. 2014. Temporal feedback for tweet search with non-parametric density estimation. In *SIGIR*, 33–42.

Fathi, A.; Farhadi, A.; and Rehg, J. M. 2011. Understanding egocentric activities. In *ICCV*, 407–414.

Gemmell, J.; Bell, G.; and Lueder, R. 2006. Mylifebits: a personal database for everything. *Communications of the ACM* 49(1):88–95.

Ghosh, J. 2012. Discovering important people and objects for egocentric video summarization. In *CVPR*, 1346–1353.

Guillaumin, M.; Mensink, T.; Verbeek, J.; and Schmid, C. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 309–316.

Hardoon, D.; Szedmak, S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: an overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London.

Hodges, S.; Williams, L.; Berry, E.; Izadi, S.; Srinivasan, J.; Butler, A.; Smyth, G.; Kapur, N.; and Wood, K. 2006. Sensecam: a retrospective memory aid. In *UbiComp*.

Hodges, J. R.; Salmon, D. P.; and Butters, N. 1990. Differential impairment of semantic and episodic memory in alzheimer's and huntington's diseases: a controlled prospective study. *Journal of Neurology, Neurosurgery & Psychiatry* 53(12):1089–1095.

Hori, T., and Aizawa, K. 2003. Context-based video retrieval system for the life-log applications. In *SIGMM international workshop on multimedia information retrieval*, 31–38.

Imura, J.; Fujisawa, T.; Harada, T.; and Kuniyoshi, Y. 2011. Efficient multi-modal retrieval in conceptual space. In *ACM-MM*, 1085–1088. ACM.

Jeon, J.; Lavrenko, V.; and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, 119–126.

Kalnikaite, V.; Sellen, A.; Whittaker, S.; and Kirk, D. 2010. Now let me see where i was: Understanding how lifelogs mediate memory. In *CHI*, 2045–2054.

Lee, Y. J., and Grauman, K. 2015. Predicting important objects for egocentric video summarization. *IJCV* 114(1):38–55.

Light, L. L.; LaVoie, D.; Valencia-Laver, D.; Albertson Owens, S. A.; and Mead, G. 1992. Direct and indirect measures of memory for modality in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(6):1284.

Lu, Z., and Grauman, K. 2013. Story-driven summarization for egocentric video. In *CVPR*, 2714–2721.

Mann, S. 1997. Wearable computing: a first step toward personal imaging. *Computer* 30(2):25–32.

Metzler, D., and Croft, W. B. 2007. Latent concept expansion using Markov random fields. In *SIGIR*, 311–318.

Nakayama, H.; Harada, T.; and Kuniyoshi, Y. 2009. AI goggles: real-time description and retrieval in the real world with online learning. In *CRV*, 184–191.

Nakayama, H.; Harada, T.; and Kuniyoshi, Y. 2010. Evaluation of dimensionality reduction methods for image auto-annotation. In *BMVC*, 1–12.

Ramanan, D. 2012. Detecting activities of daily living in first-person camera views. In *CVPR*, 2847–2854.

Sellen, A. J.; Fogg, A.; Aitken, M.; Hodges, S.; Rother, C.; and Wood, K. 2007. Do life-logging technologies support memory for the past?: an experimental study using sensecam. In *CHI*, 81–90.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Smeulders, A. W. M.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(12):1349–1380.

Xu, R.; Xiong, C.; Chen, W.; and Corso, J. J. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*.