EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines

Keita Higuch[1], **Ryo Yonetani**[1], **and Yoichi Sato**[1]

[1]The University of Tokyo

ABSTRACT

This work presents EgoScanning, a video fast-forwarding interface that helps users find important events from lengthy first-person videos continuously recorded with wearable cameras. This interface features an elastic timeline that adaptively changes playback speeds and emphasizes egocentric cues specific to first-person videos, such as hand manipulations, moving, and conversations with people, on the basis of computer-vision techniques. The interface also allows users to input which of such cues are relevant to events of their interest. Through our user study, we confirm that users can find events of interest quickly from first-person videos thanks to the following benefits of using the EgoScanning interface: 1) adaptive changing of playback speeds allows users to watch fast-forwarded videos more easily; 2) emphasized parts of videos can act as candidates of events actually significant to users; and 3) users are able to select relevant egocentric cues depending on events of their interest.

ACM Classification Keywords

H.5.2. Graphical User Interfaces

Author Keywords

First-person videos; Content-aware video fast-forwarding

INTRODUCTION

Recent advancements in camera technologies have led to a variety of portable camera devices that can be worn on the head of a person, such as Google Glass and GoPro Hero. Videos recorded with such wearable head-mounted cameras are called *first-person videos* and are used as a continuous record of everyday activities from our own perspective. First-person videos are commonly used to share our experiences in various activities such as sports and sightseeing with other people. In the HCI community, first-person videos are also used for applications such as video surveillance [35], remote collaboration [12, 13, 21], and life-logging [5, 15, 16].

Despite the widespread use of first-person videos, browsing techniques for such videos have not yet been well established. One main difficulty with watching first-person videos arises when people try to find events related to their interest from long, continuously recorded first-person videos, as important events can be distributed sparsely in such videos. Moreover, first-person videos tend to have severe camera shake due to frequent and rapid head motion. These make it difficult for users to scan first-person videos quickly to locate frames of interesting events.

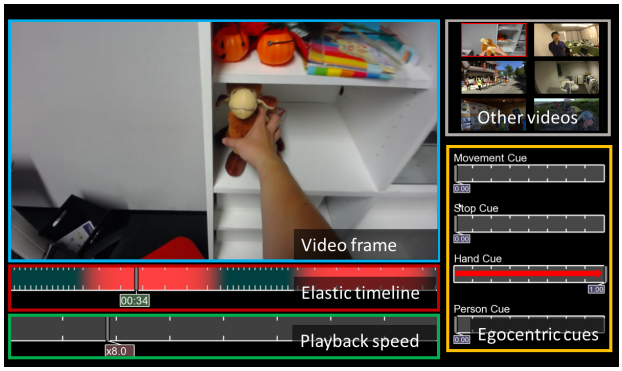

Figure 1. The proposed interface: The elastic timeline emphasizes salient parts of the video timeline on the basis of egocentric cues. The interface also allows users to input which of such cues are relevant to events of their interest. The red arrow at the bottom right indicates a user's input. In this case, the interface emphasizes hand-related events such as picking up an object.

In this work, we propose a novel interface design called **EgoScanning** for assisting users to browse first-person videos quickly to find target events of interest. This interface features an elastic timeline that adaptively changes playback speeds and emphasizes egocentric cues specific to first-person videos, such as hand manipulations, walking/standing still, and conversing with people (Figure 1). Users can easily select different egocentric cues that are relevant to specific events of their interest. For example, when a user prioritizes hand manipulations, parts of a video that contain hand-related events such as picking up a tool from a shelf are highlighted in the timeline and played at slower speed, while the remaining parts of the video can be fast-forwarded.

The main contribution of this work is twofold: 1) we introduce a video browsing interface named EgoScanning that allows users to control playback speeds easily on the basis of different egocentric cues, and 2) we confirm through a user study the effectiveness of the proposed interface for helping users to locate events of interest in long first-person videos. Concretely, the following benefits of the proposed interface were found from statistical and qualitative evidences.

- With our elastic timeline, users can scan fast-forwarded first-person videos more easily by adaptively changing playback speeds on the basis of egocentric cues.
- Parts of a video emphasized on the elastic timeline can act as candidates that contain important events. The user can then examine those parts easily by moving the slider of the elastic timeline to the emphasized parts, rather than watching the entire video. This helps significantly reduce the task completion times in search tasks.
- Users can freely select different egocentric cues that are relevant to events of their interest out of a set of fundamental egocentric cues. In fact, we observed that the egocentric cues utilized in our user study were widely different depending on the types of events in the search tasks.

RELATED WORK

Browsing Support for Conventional Videos

Various techniques to support users in browsing conventional videos recorded with fixed or hand-held cameras have been well studied. One example is a video stabilization technique for videos taken by a shaky hand-held camera [30, 31, 32]. Fast-forwarding techniques such as [17, 2] are also useful to help users watch videos in reduced time. Several researchers developed a content-aware fast-forwarding technique that changes playback speeds dynamically depending on the importance given to each video frame, enabled by using key clips [43], audio [27], a skimming model [6], and the viewing histories of other people [24]. Several novel forms of video visualizations have also been studied: spatio-temporal volume [37]; positional information [42]; and video synopsis [46, 44, 45]. Direct manipulation techniques have enabled users to manipulate object positions in video frames to seek specific video timelines [8, 38, 18, 19]. Video lens allowed users to interactively explore large collections of baseball videos and related metadata [36].

Unlike these studies, our focus is on providing an efficient way to browse first-person videos that have several unique properties such as recorder-centric scene representations and continuous recordings with significant camera motion. This requires us to develop a novel assistance for browsing videos effectively.

Browsing Support for First-Person Videos

In contrast to conventional videos, research efforts for developing browsing support techniques for first-person videos have been limited. The existing techniques can be categorized into two types: fast-forwarding techniques [26, 41] and video summarization techniques [28, 50, 33, 3]. The fast-forwarding techniques can reduce camera shake of first-person videos played at high speed based on 3D geometrical analysis of captured scenes [26] or by carefully sub-sampling frames [41]. However, such fast-forwarding techniques may not necessarily be helpful in finding target events of interest because both important and unimportant parts of a video are played equally at high speed. In contrast, video summarization techniques try to extract significant video shots automatically, where the significance is evaluated on the basis of various factors such as the presence of people [28], gaze [50], storyline [33], or joint attention [3]. Yao *et al.*also proposed a method for highlight detection for video summarization by using pairwise deep ranking of first-person videos[51]. Although summarization can help users comprehend video contents in a shorter amount of time, there is a risk of some important events being wrongly omitted in the summarized video.

BROWSING FIRST PERSON VIDEOS

Assume a scenario where users watch first-person videos with a standard laptop or a computer monitor. Note that such users are not necessarily the same people who put on a wearable camera to record such videos. One typical example of such a scenario is life-logging. Users can replay a day of themselves from their own points-of-view and relive memorable moments in life. Another example is learning professional skills. With first-person videos of sports or cooking scenes recorded by professionals, one can easily see where to pay attention or how to use the hands to perform a certain task.

As stated in the introduction, the main difficulty in such scenarios is to find events of interest quickly from a video. First-person videos of a day are often full of mundane scenes like a daily commute. Even if videos are intentionally taken by professionals, not all of them have significant parts that showcase their skills to viewers. As a result, parts of videos that are potentially significant are distributed sparsely in lengthy videos and are not apparent to users.

To overcome this difficulty, we propose analyzing various *egocentric cues* that can be extracted automatically from first-person videos. Many computer vision techniques tailored to first-person videos are now available for a variety of tasks including hand activity recognition [9, 11, 4, 34, 14, 29], activity segmentation [25, 39, 48, 40], and interaction recognition [10, 52, 53, 22]. This work takes an approach of content-aware video fast-forwarding that changes playback speeds dynamically depending on the importance of each video frame to quickly seek events of interest from first-person videos. The importance of each frame can be automatically extracted by computer-vision techniques. This approach allows users to selectively browse extracted frames with slower speeds than other frames.

However, these computer-vision techniques sometimes cause false alarms and missed detection (*i.e.*, false positives and negatives) under conditions such as extreme lighting environments and crowded scenes. As reported in a previous work, such detection errors adversely affect users' compliance with and reliance on these systems [7]. In particular, missed detection may result in the overlooking of important events in the case of video summarization approaches like [28, 50, 33, 3]. An advantage of our approach is that users can still have access to the entire video because the undetected parts are just fast-forwarded, not eliminated. We also aim to reduce missed detection of important events by using relaxing thresholds. As discussed in the following section, the techniques can allow us to have access to various cues related to important events in everyday life.

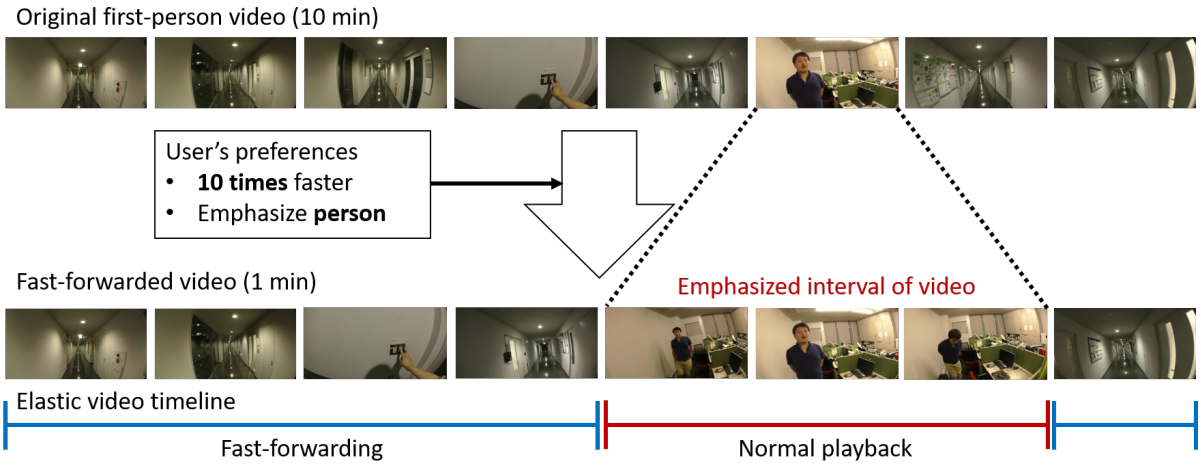

Figure 2. An example of the elastic timeline: The elastic timeline generates fast-forwarded videos from original first-person videos. In this scenario, users input their preferences of playback speed (*10 times faster*) and egocentric cues (*emphasize person*). The elastic timeline then highlights the corresponding parts of a video that contain the camera wearer's interaction with people. The highlighted parts are played at normal speed, while the remaining parts are fast-forwarded.

THE EGOSCANNING INTERFACE

In this section, we first outline the proposed EgoScanning interface. Essentially, this interface fast-forwards first-person videos while keeping some parts of the videos played at slower speed if they are potentially significant to users. Figure 2 shows an example of how videos are played with help from the EgoScanning interface. The three main components are described below.

Extraction of Egocentric Cues

First, it is automatically determined if certain egocentric cues are observed in each frame of a video. These cues should be general and applicable to first-person videos taken in various scenarios. In addition, cue extraction methods should be robust and fast enough to process long videos or multiple videos in a reasonable amount of time. We therefore use the following cues:

Movement & stop cues that reflect the movement of camera wearers, such as walking and standing still. For example, these cues can let us know when camera wearers change places of activities or when they stop to find something interesting. Since wearable cameras are typically mounted on the head, such cues can be extracted easily by analyzing camera motion [41].

Hand cue indicating when camera wearers use their hands in the front. This cue is particularly crucial when reviewing a variety of scenarios that require fine hand manipulations (*e.g.*, cooking or assembling). Accurate hand detection is one of the major tasks in computer vision with first-person videos and robust methods have been proposed [29].

Person cue that aims at detecting moments when camera wearers interact with another person. This cue can be useful when videos capture social scenes such as parties, meetings, and collaborative work. Various off-the-shelf human detection methods (*e.g.*, [47]) are available for the cue extraction.

User Inputs

Unlike typical video fast-forwarding or summarization techniques, the proposed interface accepts user interactions to take into account how much each egocentric cue is significant in their ongoing tasks. For example, users who try to learn how to use a skillet from professional videos can give more weight to the "hand" cue, or those who want to recall who they met yesterday with their own videos can emphasize the "person" cue. Users are also allowed to adjust how fast videos are played (*e.g.*, 10 times faster), as with a standard fast-forwarding interface. These user inputs change the playback speed of videos adaptively and give feedback to users in the elastic timeline that follows.

Elastic Timeline

On the basis of the combination of cue extraction results and user inputs, the timeline of a video is "elasticized" to emphasize parts that are potentially significant to users. Intuitively, if a certain part of a video contains the egocentric cues to which a user gives more weight, this part will be played at original speed. Otherwise, the part is fast-forwarded to the specified speed (see also Figure 2). In addition, intervals of videos where extracted cues are selected by users are highlighted in red in the timeline, as shown in Figure 1.

A sketch of the algorithm to determine playback speeds is as follows. Let $c_{i,t} \in [0,1]$ be the extraction result (1 if extracted and 0 otherwise) of the i-th egocentric cue at frame t. User inputs are defined by w_i in $[0,1]$ as a relative weight on the i-th cue and $s \in [1,2,\dots]$ as the desired playback speed (*i.e.*, s-times faster). With these notations, the importance of frame t is first given by $I(t) = \sum_i w_i c_{i,t} + \alpha$, where $\alpha \in \mathbb{R}_+$ is a positive constant. Then, we fix the time to display frame t to $S(t) = \frac{T}{s} \frac{I(t)}{\sum_t I(t)}$, where T is the total number of frames. If some frames are played slower than the original speed (*i.e.*, $S(t) > 1$), we further modify display times of those frames to

Figure 3. Extracting egocentric cues from first-person videos. (a) Sparse optical flows for movement and stop cues; (b) Pixel-level hand detection for a hand cue; (c) Object detection for a person cue.

$S(t) = 1$ and those of the other frames accordingly so that the total playback time remains $\frac{T}{s}$.

IMPLEMENTATIONS

This section introduces several key implementations to reproduce the proposed EgoScanning interface. As stated earlier, Figure 1 depicts the prototype interface that we used in the user studies.

Extracting Egocentric Cues from First-Person Videos

To extract egocentric cues, we adopt off-the-shelf computer vision techniques that can be implemented easily. Detection thresholds are set for the hand and person detectors to reduce missed detection of corresponding events at the cost of increased false alarms. Detection thresholds are set for the hand and person detectors to reduce missed detection of corresponding events at the cost of increased false alarms.

Movement & stop cues

To detect camera motion (*i.e.*, wearer's head motion), we implement the optical-flow-based motion detector proposed in [39]. In this method, optical flows are computed at several fixed points (Figure 3 (a)) and smoothed over time. Some features extracted from the flows (*e.g.*, flow amplitude) are then trained with a binary classifier such as a support vector machine to classify if a wearer was moving or not at every frame. Frames that are classified into the 'moving' class finally obtain 1 for the movement cue and 0 for the stop cue. Otherwise, the values 0 and 1 are assigned to the movement and stop cues, respectively.

Hand cue

We use the hand detector pre-trained with a first-person video dataset [29]. Based on color and texture features, this detector judges if each pixel in the frames belongs to hands or not, and thus is able to extract regions of hands accurately (Figure 3(b)). In our interface, the hand cue obtains 1 if hand regions occupy more than 5 % of a video frame size, and otherwise gets 0.

Person cue

We detect regions where people appear in first-person videos by a state-of-the-art object detector named faster-RCNN [47]. With this detector, objects are detected with a rectangular bounding box and classified into one of several classes (Figure 3 (c)). The value 1 is assigned to the person cue if frames contain bounding boxes of the people class that are larger than 20 % of a video frame size. The cue obtains 0 otherwise. To reduce detection time, the system applies the detector to every six frames of a video. The skipped frames have the same values as the most recent frame.

Preprocessing and postprocessing

While original videos are taken at the HD resolution, we resize them into VGA or QVGA sizes to complete the aforementioned processes in a reasonable time. In order to cope with spontaneous detection failures, for each cue we apply a median filter with the kernel size of 31 over time. We also use a Gaussian filter with the kernel size of 45 to smooth detection results over time (and as a result, each cue is represented by a real value).

User Interface

Our prototype interface (shown in Figure 1) has several control sliders that allow users to specify importance weights to each egocentric cue (bottom-left) and preferred playback speeds (bottom). Inputs through these sliders are reflected immediately to the playback speed of the video at the top left as well as the elastic timeline (second line from the bottom); intervals that obtain more importance weights are highlighted in red and played at original speed. Users are allowed to jump to arbitrary frames via the elastic timeline. Links to other videos are also implemented at the top-right.

USER STUDY

We conducted a series of user studies to investigate how the proposed EgoScanning interface helped users to scan first-person videos. The main hypothesis posed in this work is: *"with the help from the combination of egocentric cues and user inputs, the proposed EgoScanning interface can allow participants to find events related to their interests quickly."* To validate this hypothesis qualitatively and quantitatively, participants were asked to perform three realistic tasks of finding some pre-defined events from videos.

Conditions of Experiments

We recruited 16 participants (Female: 3) who were graduate students or postdoc in the computer science and engineering fields. All of the participants had experience using video player interfaces (*e.g.*, YouTube).

We compared the EgoScanning interface with a simple baseline interface that just allowed users to fast-forward videos at arbitrary uniform speed. With the EgoScanning, participants were able to input their preferences on egocentric cues and playback speeds. In contrast, we disabled the control sliders of egocentric cues in the baseline interface. Note that

other widgets (*e.g.*, the control slider of the playback speed) remained available, as in the proposed interface. By default, the playback speed was set to 8 times faster in both of the interfaces.

Tasks

We designed three tasks to find some events from first-person videos. To make our user study realistic, videos taken in a variety of scenarios were used: visual surveillance (Task 1), navigation (Task 2), and cooking (Task 3). We recorded a new first-person video dataset for Task 1 while existing datasets were used in the remaining tasks (the Navigation dataset [54] for Task 2 and the GeorgiaTech Egocentric Activities dataset [11] for Task 3).

We divided each dataset into two video subsets, A and B, and used one with the EgoScanning and the other with the baseline. For each dataset, we preliminarily defined several events that participants were asked to find. Table 1 presents more details on the datasets.

Task 1: Finding events from a long video

In Task 1, participants tried to find certain events from first-person videos during a visual surveillance task. In each video, camera wearers walked inside a large building for 30 minutes and took the following actions: pushing light switches (a number of times), feeding paper into a printer (4 times), and reporting to a colleague (twice), as shown in Figure 4. We asked participants to find the latter two events. This task is designed to be a practical task of reviewing the work histories of others in the context of work training.

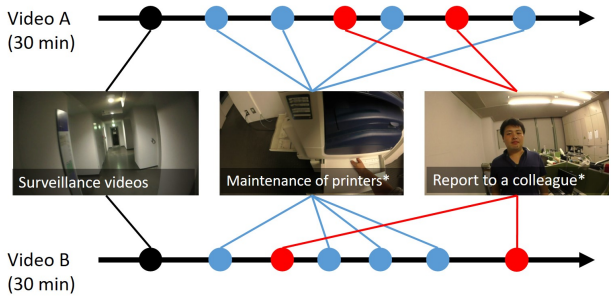

Figure 4. Videos used in Task 1: Participants were asked to search each video for the actions (feeding paper into a printer and reporting to a colleague.)

Task 2: Finding attractive objects from multiple videos

In Task 2, participants were asked to find some attractive objects from multiple first-person videos recorded in several indoor/outdoor environments [54]. Specifically, camera wearers walked around the environment to look at attractive objects such as map signs, vending machines, and the entrance of a store for a short time period. We chose 16 videos from the original 44 videos for our study. Specific appearances of objects used in the experiments are shown in Figure 5. Task 2 can be interpreted as a scenario of making a navigation plan using first-person videos.

	Task 1		Task 2		Task 3	
Dataset	Our dataset		[54]		[11]	
Test sets	A	B	A	B	A	B
Videos	1	1	8	8	8	8
Total times	30:00	30:00	23:30	24:03	26:40	27:45
Fps	30	30	30	30	15	15
Target time						
1st event	8:00	4:57	3:39	2:49	4:15	10:19
2nd event	14:13	8:30	9:03	10:08	13:41	15:44
3rd event	15:00	11:20	22:23	16:21	16:38	22:30
4th event	17:10	13:57	–	–	–	–
5th event	21:45	20:28	–	–	–	–
6th event	27:55	24:20	–	–	–	–
Computation time						
M&S cues	41:33	39:36	46:40	47:44	23:32	24:00
Hand cue	29:45	30:26	25:55	25:48	20:07	20:34
Person cue	56:21	55:56	46:08	46:09	24:56	25:57
Detection accuracy						
M&S cues	0.85	0.93	0.86	0.85	–	–
Precision	0.84	0.94	0.86	0.85	–	–
Recall	0.99	0.99	0.98	0.98	–	–
Hand cues	0.90	0.88	–	–	0.62	0.68
Precision	0.54	0.33	–	–	0.62	0.68
Recall	0.83	0.93	–	–	0.95	0.95
Person cues	0.99	0.99	0.98	0.92	–	–
Precision	0.68	0.69	0.26	0.47	–	–
Recall	0.96	0.94	1.00	0.89	–	–

Table 1. Three datasets used in our user study: The first row gives basic information on the three datasets. The second row shows computation time of egocentric cues in the preprocessing. The third row indicates time of target events in the test sets of each task. We measured the computation time of each egocentric cue on a desktop computer with an Intel i7-4790K CPU, NVIDIA TITAN X GPU, and 32Gb RAM. The movement, stop, and hand cues were extracted in a single processor. To process the person cue, we used GPU acceleration. The fourth row shows detection accuracy of egocentric cues (Recall and Precision). We mostly did not observe hand-related events in Task 2 or movement-related and person-related events in Task 3.

Task 3: Finding actions from multiple videos

Task 3 required participants to find important hand actions from multiple first-person videos of a cooking scene. We used the GeorgiaTech Egocentric Activities (GTEA) dataset [11] that contains 28 videos of cooking activities recorded by four camera wearers. We chose to use 16 videos for the experiments. The action of *pouring a drink* was the target to be found. Figure 6 shows some examples of target actions. Although the presence of hands in videos is a salient cue for finding hand actions, this task is probably the hardest among the three even for the proposed interface because hands are visible almost all the time during the recordings. With this task, we aim to see how the proposed interface works with such a difficult situation.

Procedure

At the beginning of the session, participants filled out a pre-study questionnaire on their prior experience of using video interfaces. We then gave them brief explanations of the user

Figure 5. Target objects in Task 2.

Figure 6. Examples of target actions.

study and tasks. The order of experiment conditions and test sets was randomized to maintain a counterbalance (*i.e.*, this study needed 16 participants). Before using each interface, the participants performed one or two example tasks to understand how the interfaces worked. Sessions proceeded as follows: with one interface, participants first conducted all three tasks once. Then they filled out a questionnaire about the interface being used. After that, participants used the other interface and conducted the three tasks again with different video subsets. The questionnaire was then filled out again. Finally, we conducted an interview session with the participants. Each experiment took about half an hour in total.

Evaluation Measures

Task completion time

To determine the effectiveness of the EgoScanning, we measured the task completion time for each task. We expect that this time will decrease if participants find events quickly. We therefore compared the two interfaces quantitatively by using the Wilcoxon signed-rank test and confidence interval of the difference of means to reveal significance effects.

Average scanning speed

We then calculated the average scanning speed for each task and for all tasks. This metric shows how fast each interface allows users to use the playback speed. To calculate the metric, we used task completion times and final event times of each test set, as the participants reviewed videos of each test set from beginning to end in order. We thus simply calculated the reviewed times of the video per second by $\frac{FinalEventTime}{TaskCompletionTime}$ as scanning speeds. Larger values indicated faster scanning speeds. We also compared the two interfaces by using the Wilcoxon signed-rank test and confidence interval of the difference of means to reveal significance effects.

Utilization ratio of egocentric cues

We computed the utilization ratios of egocentric cues for each task to see which cues were useful. We assumed that the utilization ratio can be diverse depending on the task. To calculate utilization ratios, we used manipulation logs of the proposed interface. We judged whether participants used an egocentric cue if the corresponding user input value exceeds 0.1. We then computed temporal averages of each egocentric cue as the utilization ratios for all tasks.

Questionnaire

After each condition , the participants answered questions on a seven-point scale (strongly disagree = 1, neutral = 4, strongly agree = 7). We also used the Wilcoxon signed-rank test to determine whether there were significant differences in the participants' experience. Specifically, we asked the participants the following five questions: Q1) Could you use the interface easily? Q2) Could you complete tasks easily? Q3) Did you understand the contents of the videos? Q4) Did you feel any visually induced motion sickness? and Q5) Did you enjoy using the interface?

After completing the tasks in both conditions, the participants answered an extra five questions about the functions of the proposed interface on a seven-point scale. The extra questions were: EQ1) Was the elastic timeline useful? EQ2) Were *the Movement & Stop cues* useful? EQ3) Was *the Hand cue* useful? EQ4) Was *the Person cue* useful? and EQ5) Were the multiple egocentric cues necessary?

Observation and interview

We observed the participants to see how they found the events of search targets using each interface. After each user completed all tasks, we interviewed them for about 5 min. For the most part, our interviews were semi-structured so as to focus on the experience. First, we asked the participants about the proposed interface, namely, "*Did you use the functions of the elastic timeline for finding events? And if so, how?*"[1] We further inquired as to their usages of the proposed interface based on observations. Finally, we asked the participants, "*Can you recommend any new functions to improve the proposed interface?*" and "*How did you find the events using the baseline interface?*"

RESULTS

Task Completion Time

The upper half of Table 2 shows the results of task completion time. Overall, the proposed interface achieved faster results than the baseline interface. The Wilcoxon signed-rank test revealed a significant difference between the proposed and baseline interface in both the individual tasks and the tasks as a whole ($p = 0.02, p = 0.01, p = 0.05$, and $p = 0.01$, respectively). A 95 % confidence interval of the difference of means revealed that the proposed interface could significantly reduce task completion times in all tasks. These results support our hypothesis that the proposed interface enables faster task completion time.

[1]In this paper, italic fonts in double quotations denote translated speech from other languages.

			Wilcoxon: Baseline/Proposed		95% confidence interval of difference of means		99% confidence interval of difference of means	
	Baseline: Average (Std)	*Proposed: Average (Std)*	*p*	*Z*	*Lower bound*	*Upper bound*	*Lower bound*	*Upper bound*
Task Completion Time								
Task 1	173.65 (62.50)	**126.00** (25.38)	**.02***	-2.39	-81.80	**-13.48**	-92.60	**-2.68**
Task 2	182.23 (52.65)	**116.20** (38.55)	**.01***	-3.08	-88.54	**-43.52**	-95.66	**-36.40**
Task 3	144.29 (55.95)	**108.75** (52.34)	**.05***	-2.18	-70.41	**-0.67**	-81.44	10.36
Total	500.17 (131.33)	**350.95** (91.19)	**.01***	-3.26	-203.36	**-95.08**	-220.49	**-77.96**
Average Scanning Speed								
Task 1	10.05 (3.18)	**13.26** (3.05)	**.02***	-2.32	**0.80**	5.62	**0.03**	6.39
Task 2	7.15 (2.19)	**11.02** (3.44)	**.01***	-2.70	**1.78**	5.96	**1.11**	6.62
Task 3	9.31 (4.52)	**12.29** (3.40)	**.01***	-2.70	**0.44**	5.51	-0.37	6.32
Total	8.84 (2.87)	**12.20** (2.22)	**.01***	-2.95	**1.79**	4.91	**1.30**	5.40

Table 2. Results of task completion time and average scanning speed: This table contains averages and standard deviations in baseline and proposed interfaces for the Wilcoxon signed-rank test ("*" indicates the significance). Results of 95 % and 99 % confidence intervals of difference of means are also shown.

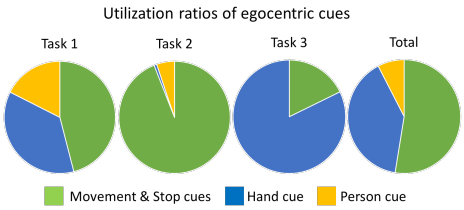

Figure 7. Utilization ratios of egocentric cues.

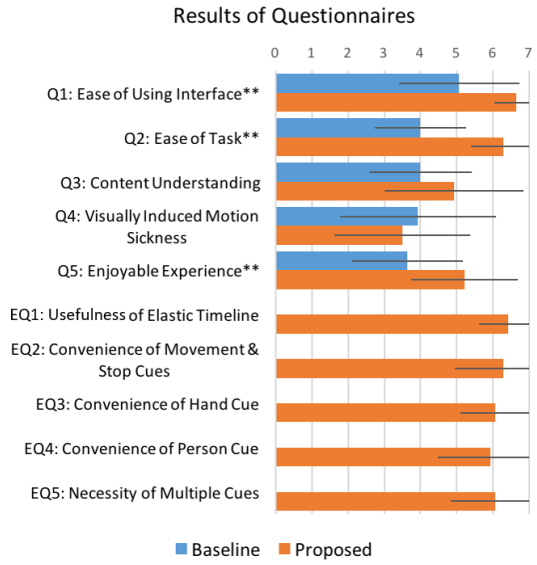

Figure 8. Results of questionnaire.

Average Scanning Speed

The lower half of Table 2 shows the average scanning speeds in each task and total averages. Not surprisingly, the results are similar to the task completion times. By conducting the Wilcoxon signed-rank test, we observed a significant increase of average scanning speed ($p = 0.02, p = 0.01, p = 0.01$, and $p = 0.01$, respectively). A 95 % confidence interval of the difference of means also indicates that the proposed interface can help users increase the scanning speeds in all tasks. As for the total average, we confirmed a 38 % increase of the scanning speed with the proposed interface.

Utilization ratio of egocentric cues

Figure 7 shows the utilization ratios of egocentric cues in each task. The result indicates there is a large variance in the utilization ratio across tasks. This implies that the participants tried to select relevant cues to events of search targets. In Task 1, nearly half of the participants combined the Hand and Person cues to find query events, while the other participants used only the Stop cue. In Tasks 2 and 3, the participants mainly used the Stop and Hand cues, respectively.

Questionnaire

Figure 8 shows the questionnaire results. The Wilcoxon signed-rank test revealed significant differences in Q1) ease of using interface, Q2) ease of task, and Q5) enjoyable experience ($p = 0.01$). Interestingly, 63% of the participants (10/16) did not report any visually induced motion sickness during the three tasks in both of the interfaces.

As for the extra questions, we received positive feedback (*i.e.*, 5, 6, or 7 scales) from most participants. The positive feedback ratios of EQ1–5 were 100% (16/16), 88% (14/16), 88%, 88%, and 88%, respectively.

Feedback and Observation

Here, we discuss the qualitative results elicited from observation and feedback of the participants. Overall, it seems that the participants used the proposed interface effectively, as indicated by comments such as "*I really like this interface because I could see important scenes when I used very high speed playback*," "*I didn't need to see unnecessary information. I could find targets when playback speeds became slower*," and "*This function (the elastic timeline) was very useful for watching videos. I could reduce the time to watch the videos*."

(a) Task 1 + Person cue (b) Task 2 + Stop cue (c) Task 3 + Hand cue

Figure 9. Typical results of elastic timeline usage in the user study.

Most of the participants preferred the highlighted parts of the elastic timeline. Figure 9 (a)(b) shows typical results of the elastic timeline that emphasize the candidates of query events in Tasks 1 and 2. Almost half the participants moved the slider of the video timeline to emphasized regions, which enabled them to quickly review candidates of target events. User comments included "*I could understand where important scenes were located*," "*I felt that I could watch videos more easily because I only needed to check the highlighted areas*," and "*I manually moved the slider to the highlighted parts. I felt finding targets was very easy*."

Using the proposed interface, participants could quickly select egocentric cues depending on the tasks. Selected cues were mostly relevant to the events of search targets. User comments included, "*Manipulating the interface was very easy. I could quickly grasp how to use it," "I could predict what cues should be selected. This interface immediately emphasized the video timeline*," and "*I really like the Hand and Stop cues. It was really useful to find the camera wearers' actions*."

Surprisingly, in Task 3, the proposed interface could reduce the task completion time by emphasizing the Hand cue, even though all videos used in this task frequently showed the camera wearers' hands and contained false alarms of hand detection (Figure 9 (c)). We received positive feedback from some participants, with comments including "*When I used the Hand cue, many parts of the video were emphasized. But, for me, it is still useful to specify the candidates of target actions*" and "*I could still believe that the interface made playback speeds slower in important scenes.*" In contrast, we also received some negative feedback, such as "*I was confused in Task 3, because lots of parts in the timeline were highlighted*."

Many participants requested new features of the proposed interface in order to find objects in the first-person videos for Task 3, as indicated by comments such as "*In Task 3, I focused on objects that were manipulated by hands. I wish there was a function that enabled me to search objects from videos*" and "*I believe objects-based retrieval is really important for this interface. But, I have no idea how that could be done*."

Using the baseline interface, the participants set a slower playback speed than the proposed interface, or manually sought the slider of the video timeline to find events of search targets: "*I set 4–6 times faster playback speeds to view the videos. I felt it was difficult to find events at a faster speed*" and "*I didn't think I would be able to find events with simple fast-forwarding playback, so I always moved the slider manually*."

DESIGN IMPLICATION

Overall, we found statistical and qualitative evidence that the proposed interface was beneficial to the participants in carrying out event-finding tasks from first-person videos. We now summarize three important benefits provided by the proposed interface.

Faster speed playback of first-person videos with egocentric elastic timeline

With the elastic timeline, the participants could watch a fast-forwarded first-person video with reduced effort, as the playback speed of the video was changed adaptively depending on egocentric cues selected by the participants. Video frames corresponding to the selected cues were played more slowly than other parts of the video. This allowed the participants to examine those frames more carefully to find target events. We also observed increased scanning speeds with statistical significance in all of the tasks. We thus argue that the adaptive control of playback speed enables users to watch fast-forwarded first-person videos more easily.

Emphasized frames as candidates of important events

Emphasized parts of videos can act as candidates of events that are actually significant to users. On the basis of selected egocentric cues, the proposed interface extends and emphasizes parts of the video timelines. The participants could thus predict where target events were located in the videos. Most of the participants believed that emphasized regions included important events. Some participants further moved the slider of the video timeline to emphasized regions. As a result, we observed significant reduction of task completion time in all tasks. We argue that users quickly grasp candidates of events from the emphasized timeline to achieve faster event findings.

Providing a set of important egocentric cues

The participants were able to select relevant egocentric cues depending on the query events of their interest. The proposed interface provided a set of egocentric cues that can be robustly extracted by computer vision techniques despite significant variance in first-person videos. We observed in the user study that the participants quickly selected egocentric cues that were appropriate for finding particular targets, and that selected egocentric cues varied significantly among the three tasks. This suggests that it is important to provide a set of important fundamental cues in order to assist users to find events related to their interest more efficiently.

LIMITATION

We used three computer vision techniques to analyze first-person videos in the preprocessing. Extracting egocentric cues worked well with the three first-person video datasets used in the user study that contained various scenes such as outdoor, office, and large hallway environments. However, the computer vision techniques could still fail in some first-person videos recorded in complicated scenes. In particular, the current techniques cannot find *Hand* and *Person* cues correctly in crowded scenes under severe lighting conditions. In such cases, it is not easy to choose appropriate detection thresholds automatically. One possible way to alleviate the problem is to

allow users to adjust thresholds as they perform a task using the proposed interface.

Furthermore, it was reported by the participants that object-level cues to locate objects of certain types would have been helpful in addition to the egocentric cues. Other egocentric cues such as staring at an object, standing up/sitting down, and looking around can be useful for carrying out various tasks, so allowing users to use additional cues in the proposed interface would be beneficial. On the other hand, having too many egocentric cues may make the proposed interface overly complicated and thus more difficult for users to choose and adjust the cues needed to do a certain task. We therefore plan to further investigate which sets of egocentric cues are robustly extracted and useful for the elastic timeline in many situations.

CONCLUSION AND FUTURE WORK

We introduced a video fast-forwarding interface and investigated how this interface helped users scan first-person videos efficiently to find events related to their interest. The results of user studies with three realistic tasks showed that users with the proposed interface were able to reduce task completion times and scan videos faster on average. We also confirmed three significant effects of using the interface. We believe that our interface can be applied to various tasks that use first-person videos, including but not limited to life-logging [1, 16, 49, 23] and collaborative systems [13, 20].

While we tested a limited number of egocentric cues, a variety of computer vision techniques can be applied to use other egocentric cues. For example, action recognition from first-person videos [34] could help us to find events of a specific fine-grained action (*e.g.*, the action of "pouring milk into a bowl"). One interesting direction for future work is to introduce many types of egocentric cues and manage them intelligently so that they can be accessed by users easily. This will allow us to address more challenging scenarios where finer-grained egocentric cues and more efficient methods of user input will be necessary, such as helping users learn skills from videos taken by professionals or helping them find important moments from videos recorded over the period of a week.

ACKNOWLEDGMENTS

This research was supported by CREST, JST. We thank Yuki Koyama, Hiroshi Kera and Rie Kamikubo for their support.

REFERENCES

1. Kiyoharu Aizawa, Datchakorn Tancharoen, Shinya Kawasaki, and Toshihiko Yamasaki. 2004. Efficient Retrieval of Life Log Based on Context and Content. In *In Proc. ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*. DOI: http://dx.doi.org/10.1145/1026653.1026656
2. Abir Al-Hajri, Matthew Fong, Gregor Miller, and Sidney Fels. 2014. Fast Forward with Your VCR: Visualizing Single-video Viewing Statistics for Navigation and Sharing. In *In Proc. Graphics Interface (GI '14)*. http://dl.acm.org/citation.cfm?id=2619648.2619669
3. Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic Editing of Footage from Multiple Social Cameras. *ACM Transaction on Graphics* 33, 4 (2014). DOI: http://dx.doi.org/10.1145/2601097.2601198
4. Minjie Cai, Kris M Kitani, and Yoichi Sato. 2015. A scalable approach for understanding the visual structures of hand grasps. In *In Proc. IEEE International Conference on Robotics and Automation (ICRA '15)*. 1360–1366. DOI: http://dx.doi.org/10.1109/ICRA.2015.7139367
5. Yi Chen and Gareth JF Jones. 2010. Augmenting human memory using personal lifelogs. In *In Proc. Augmented Human International Conference (AH '10)*. 24. DOI: http://dx.doi.org/10.1145/1785455.1785479
6. Kai-Yin Cheng, Sheng-Jie Luo, Bing-Yu Chen, and Hao-Hua Chu. 2009. SmartPlayer: User-centric Video Fast-forwarding. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '09)*. DOI: http://dx.doi.org/10.1145/1518701.1518823
7. Stephen R Dixon, Christopher D Wickens, and Jason S McCarley. 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 4 (2007).
8. Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video Browsing by Direct Manipulation. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '08)*. DOI: http://dx.doi.org/10.1145/1357054.1357096
9. Alireza Fathi, Ali Farhadi, and James M. Rehg. 2011. Understanding Egocentric Activities. In *In Proc. International Conference on Computer Vision (ICCV '11)*. DOI:http://dx.doi.org/10.1109/ICCV.2011.6126269
10. Alireza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *In Proc. IEEE Conference On Computer Vision and Pattern Recognition*. IEEE, 1226–1233.
11. Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *In Proc. IEEE Conference On Computer Vision and Pattern Recognition (CVPR '11)*. DOI: http://dx.doi.org/10.1109/CVPR.2011.5995444
12. Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. 2003. Effects of Head-mounted and Scene-oriented Video Systems on Remote Collaboration on Physical Tasks. In *In Proc. ACM CHI conference on Human factors in computing systems (CHI '03)*.
13. Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016. Can Eye Help You?: Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In *In Proc. ACM CHI conference on Human factors in computing systems (CHI '16)*. 5180–5190. DOI: http://dx.doi.org/10.1145/2858036.2858438

14. De-An Huang, Minghuang Ma, Wei-Chiu Ma, and Kris M Kitani. 2015. How do we use our hands? discovering a diverse set of common grasps. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298666

15. Yoshio Ishiguro, Adiyan Mujibiya, Takashi Miyaki, and Jun Rekimoto. 2010. Aided Eyes: Eye Activity Sensing for Daily Life. In *In Proc. Augmented Human International Conference (AH '10)*. 25. DOI: http://dx.doi.org/10.1145/1785455.1785480

16. Yoshio Ishiguro and Jun Rekimoto. 2012. GazeCloud: A Thumbnail Extraction Method Using Gaze Log Data for Video Life-Log. In *In Proc. International Symposium on Wearable Computers*. 72–75. DOI: http://dx.doi.org/10.1109/ISWC.2012.32

17. Neel Joshi, Wolf Kienzle, Mike Toelle, Matt Uyttendaele, and Michael F. Cohen. 2015. Real-time Hyperlapse Creation via Optimal Frame Selection. *ACM Transaction on Graphics* 34, 4 (2015). DOI: http://dx.doi.org/10.1145/2766954

18. Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers. 2008. DRAGON: A Direct Manipulation Interface for Frame-accurate In-scene Video Navigation. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '08)*. DOI: http://dx.doi.org/10.1145/1357054.1357097

19. Thorsten Karrer, Moritz Wittenhagen, and Jan Borchers. 2012. DragLocks: Handling Temporal Ambiguities in Direct Manipulation Video Navigation. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '12)*. DOI: http://dx.doi.org/10.1145/2207676.2207764

20. Shunichi Kasahara, Mitsuhito Ando, Kiyoshi Suganuma, and Jun Rekimoto. 2016. Parallel Eyes: Exploring Human Capability and Behaviors with Paralleled First Person View Sharing. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '16)*. 12. DOI:http://dx.doi.org/10.1145/2858036.2858495

21. Shunichi Kasahara and Jun Rekimoto. 2014. Jackin: Integrating first-person view with out-of-body vision generation for human-human augmentation. In *In Proc. Augmented Human International Conference (AH '14)*. DOI:http://dx.doi.org/10.1145/2582051.2582097

22. Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. 2016. Discovering Objects of Joint Attention via First-Person Sensing. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '16)*. DOI: http://dx.doi.org/10.1109/CVPRW.2016.52

23. Ig-Jae Kim, Sang Chul Ahn, Heedong Ko, and Hyoung-Gon Kim. 2008. Automatic Lifelog media annotation based on heterogeneous sensor fusion. In *In Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI '08)*.

24. Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven Interaction Techniques for Improving Navigation of Educational Videos. In *In Proc. ACM Symposium on User Interface Software and Technology (UIST '14)*. DOI: http://dx.doi.org/10.1145/2642918.2647389

25. Kris. M. Kitani, Takanori Okabe, Yoichi Sato, and Akihiro Sugimoto. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *In Proc. IEEE Conference On Computer Vision and Pattern Recognition (CVPR '11)*. 3241–3248. DOI: http://dx.doi.org/10.1109/CVPR.2011.5995406

26. Johannes Kopf, Michael F. Cohen, and Richard Szeliski. 2014. First-person Hyper-lapse Videos. *ACM Transaction on Graphics* 33, 4 (2014). DOI: http://dx.doi.org/10.1145/2601097.2601195

27. Kazutaka Kurihara. 2012. CinemaGazer: A System for Watching Videos at Very High Speed. In *In Proc. International Working Conference on Advanced Visual Interfaces (AVI '12)*. DOI: http://dx.doi.org/10.1145/2254556.2254579

28. Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*. DOI: http://dx.doi.org/10.1109/CVPR.2012.6247820

29. Cheng Li and Kris M Kitani. 2013. Pixel-level hand detection in ego-centric videos. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*. DOI: http://dx.doi.org/10.1109/CVPR.2013.458

30. Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. 2009. Content-preserving Warps for 3D Video Stabilization. *ACM Transaction on Graphics* 28, 3 (2009). DOI:http://dx.doi.org/10.1145/1531326.1531350

31. Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. 2011. Subspace Video Stabilization. *ACM Transaction on Graphics* 30, 1 (2011). DOI: http://dx.doi.org/10.1145/1899404.1899408

32. Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. 2014. Steadyflow: Spatially smooth optical flow for video stabilization. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. 4209–4216. DOI:http://dx.doi.org/10.1109/CVPR.2014.536

33. Zheng Lu and Kristen Grauman. 2013. Story-Driven Summarization for Egocentric Video. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*. 2714–2721. DOI: http://dx.doi.org/10.1109/CVPR.2013.350

34. Minghuang Ma, Haoqi Fan, and Kris M Kitani. 2016. Going Deeper into First-Person Activity Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). DOI: http://dx.doi.org/10.1109/CVPR.2016.209

35. S. Mann. 1997. Wearable computing: a first step toward personal imaging. *IEEE Computer* 30, 2 (1997). DOI: http://dx.doi.org/10.1109/2.566147

36. Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video Lens: Rapid Playback and Exploration of Large Video Collections and Associated Metadata. In *In Proc. ACM Symposium on User Interface Software and Technology (UIST '14)*. 10. DOI: http://dx.doi.org/10.1145/2642918.2647366

37. Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2012. Video Summagator: An Interface for Video Summarization and Navigation. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '12)*. DOI: http://dx.doi.org/10.1145/2207676.2207767

38. Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2013. Direct Manipulation Video Navigation in 3D. In *In Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI '13)*. DOI: http://dx.doi.org/10.1145/2470654.2466150

39. Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal Segmentation of Egocentric Videos. In *In Proc. IEEE Conference On Computer Vision and Pattern Recognition (CVPR '14)*. DOI: http://dx.doi.org/10.1109/CVPR.2014.325

40. Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. 2016. Compact CNN for Indexing Egocentric Videos. In *In Proc. IEEE Winter Conference on Applications of Computer Vision (WACV '16)*.

41. Yair Poleg, Tavi Halperin, Chetan Arora, and Shmuel Peleg. 2015. EgoSampling: Fast-forward and stereo for egocentric videos. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. DOI:http://dx.doi.org/10.1109/CVPR.2015.7299109

42. Suporn Pongnumkul, Jue Wang, and Michael Cohen. 2008. Creating Map-based Storyboards for Browsing Tour Videos. In *In Proc. ACM Symposium on User Interface Software and Technology (UIST '08)*. DOI: http://dx.doi.org/10.1145/1449715.1449720

43. Suporn Pongnumkul, Jue Wang, Gonzalo Ramos, and Michael Cohen. 2010. Content-aware Dynamic Timeline for Video Browsing. In *In Proc. ACM Symposium on User Interface Software and Technology (UIST '10)*. DOI: http://dx.doi.org/10.1145/1866029.1866053

44. Yael Pritch, Alex Rav-Acha, Avital Gutman, and Shmuel Peleg. 2007. Webcam Synopsis: Peeking Around the World. In *In Proc. IEEE International Conference on Computer Vision (ICCV' 07)*. DOI: http://dx.doi.org/10.1109/ICCV.2007.4408934

45. Yael Pritch, Alex Rav-Acha, and Shmuel Peleg. 2008. Nonchronological Video Synopsis and Indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 11 (2008). DOI: http://dx.doi.org/10.1109/TPAMI.2008.29

46. Alex Rav-Acha, Yael Pritch, and Shmuel Peleg. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR' 06)*.

47. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *In Proc. Annual Conference on Neural Information Processing Systems (NIPS '15)*.

48. Julian Steil and Andreas Bulling. 2015. Discovery of Everyday Human Activities From Long-Term Visual Behaviour Using Topic Models. In *In Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*. 75–85. DOI: http://dx.doi.org/10.1145/2750858.2807520

49. Zhe Wang, Matthew D. Hoffman, Perry R. Cook, and Kai Li. 2006. VFerret: Content-based Similarity Search Tool for Continuous Archived Video. In *In Proc. ACM Workshop on Continuous Archival and Retrival of Personal Experences (CARPE '06)*. DOI: http://dx.doi.org/10.1145/1178657.1178663

50. Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M. Rehg, and Vikas Singh. 2015. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. DOI:http://dx.doi.org/10.1109/CVPR.2015.7298836

51. Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. (2016). DOI: http://dx.doi.org/10.1109/CVPR.2016.112

52. Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. 2015. Ego-surfing first person videos. In *In Proc. IEEE Conference On Computer Vision and Pattern Recognition (CVPR '15)*. DOI: http://dx.doi.org/10.1109/CVPR.2015.7299183

53. Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. 2016a. Recognizing Micro-Actions and Reactions From Paired Egocentric Videos. In *In Proc. IEEE Conference On Computer Vision and Pattern Recognition (CVPR '16)*. DOI:http://dx.doi.org/10.1109/CVPR.2016.288

54. Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. 2016b. Visual Motif Discovery via First-Person Vision. In *In Proc. European Conference on Computer Vision (ECCV '16)*.