Dynamic Object Scanning: Object-Based Elastic Timeline for Quickly Browsing First-Person Videos

Seita Kayukawa
Waseda University
Tokyo, Japan
k940805k@ruri.waseda.jp

Keita Higuchi
Ryo Yonetani
Institute of Industrial Science,
The University of Tokyo
Tokyo, Japan
khiguchi@iis.u-tokyo.ac.jp
yonetani@iis.u-tokyo.ac.jp

Masanori Nakamura
Waseda University
Tokyo, Japan
m-nakamu@ruri.waseda.jp

Yoichi Sato
Institute of Industrial Science,
The University of Tokyo
Tokyo, Japan
ysato@iis.u-tokyo.ac.jp

Shigeo Morishima
Waseda Research Institute for
Science and Engineering
Tokyo, Japan
shigeo@waseda.jp

Abstract

This work presents the Dynamic Object Scanning (DO-Scanning), a novel interface that helps users browse long and untrimmed first-person videos quickly. The proposed interface offers users a small set of object cues generated automatically tailored to the context of a given video. Users choose which cue to highlight, and the interface in turn fast-forwards the video adaptively while keeping scenes with highlighted cues played at original speed. Our experimental results have revealed that the DO-Scanning arranged an efficient and compact set of cues, and this set of cues is useful for browsing a diverse set of first-person videos.

Author Keywords

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

ACM Classification Keywords

H.5.2. [Graphical User Interfaces]

Introduction

We envision a future where people are equipped with wearable cameras, such as Google Glass and GoPro Hero, habitually to record visual experience of everyday life. Such a continuous use of wearable cameras will produce a very large and diverse collection of long and untrimmed first-person points-of-view videos. These

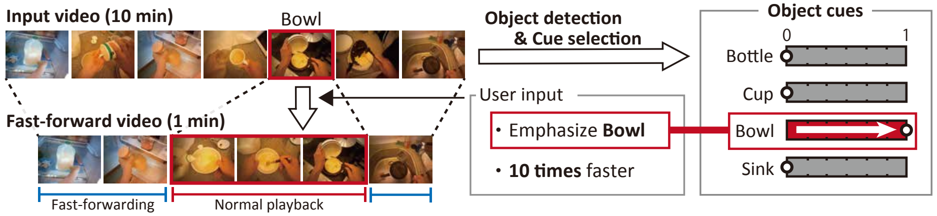


Figure 1: Dynamic Object Scanning (DO-Scanning)

videos contain a variety of moments such as daily conversations with colleagues, cooking at home, or even more special events like traveling to another country. Our goal in this work is to develop a novel user interface that assists people to quickly browse first-person videos of such diverse visual experiences.

A typical solution to browse long and untrimmed first-person videos is video summarization techniques that extract significant scenes automatically from the videos under certain criteria (*e.g.*, [3, 8], and Google Clips[1]). However, video summaries generated automatically do not always reflect specific interests that users may have and can unexpectedly omit parts of videos that they seek.

Another solution is to adaptively fast-forward videos while playing significant scenes at a lower speed [7]. In particular, we are interested in *elastic timeline* [2], which allows users to input their preferences interactively. Based on this interaction, the elastic timeline fast-forwards videos adaptively according to the specified significance. For example, if users set high significance to ‘people,’ the timeline fast-forwards videos while playing all scenes with people at the original speed.

Despite its conceptual novelty, practical applications of the elastic timeline are still limited. While [2] allows users to highlight a set of cues specific to first-person videos, such as hand manipulations, walking/standing still and conversations, these cues have been fixed for any given video. As a result, the choice of these cues does not necessarily reflect the underlying semantic context of videos, which significantly limits the variety of videos that can enjoy the benefit of elastic timeline. For example, consider a scenario where users browse first-person videos of cooking. Since such videos would typically capture recorder’s hands nearly in every time like shown in Figure 1, the hand cue would never help users to browse the video. To work around a diverse set of first-person videos, the interface requires a more sophisticated choice of semantic cues to describe a variety of scenes in detail.

In this work, we develop a novel interface based on the elastic timeline which we code-named *Dynamic Object Scanning (DO-Scanning)*. As illustrated in Figure 1, the DO-Scanning offers a set of *object cues*, categories of objects detected automatically in a given video and arranged dynamically to describe the context of the video. These object cues allow users to enhance various types of scenes. For instance, if users set high significance to the ’bowl’ cue, the interface will allow the users to access all scenes with bowls (the frame highlighted in red in Figure 1) at the original speed.

As the backbone of DO-Scanning, we present an algorithm to generate a compact and efficient set of object cues from a diverse set of object categories found in videos. Our algorithm selects a set of cues in a greedy manner while excluding useless object categories such as irrelevant categories hardly observed in the videos and

[1] store.google.com/product/google_clips

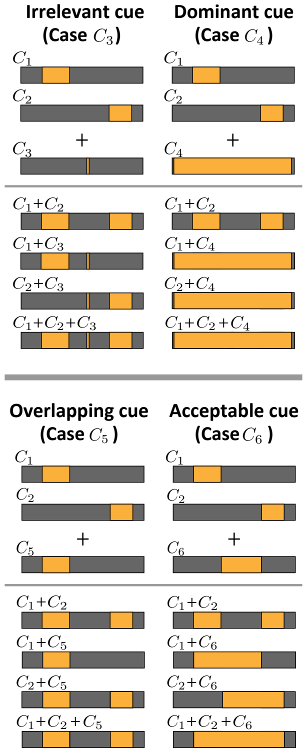


Figure 2: Greedy selection of object categories for constructing an efficient and compact set of cues. Each rectangle represents a sequence of frames. Orange rectangles describe frames which will be emphasized by setting high significance to individual object categories from C_1 to C_6 or their combinations like $C_1 + C_2$.

temporally dominant categories observed uniformly in videos and cannot be used for adaptive fast-forwarding.

Dynamic Object Scanning

In this work, we will particularly focus on how to generate a set of cues to adaptively fast-forward a diverse collection of first-person video. Below we first discuss some requirements for cue selection, and then present a concrete algorithm to discover a relevant set of cues automatically.

Dynamic and Semantic cue

As a cue that reflects a semantic context of each given video, we focus on *object cues*, the presence of certain object categories in videos. Just as object detection plays an important role for tasks of video summarization [3] and scene recognition [4] in computer vision, we expect the object cues to assist users not only to infer the overall semantic context of videos (*e.g.*, types of activities and places) but also to easily access specific scenes with certain contexts (*e.g.*, some scenes when recorders looked at particular objects). More specifically, we first run a pre-trained YOLOv2 [6] to detect 80 object categories.

Compact and Efficient Sets of Cues

While object cues help users to access to a semantic context of videos, displaying a large number of object categories without considering each relevance to the videos would be redundant and unhelpful for users. To cope with this problem, we propose a greedy algorithm that can select a compact and efficient set of cues.

We show how our algorithm works with an example in Figure 2. Suppose that object categories C_1 and C_2 are given as a part of the final set of cues, and we try to add a new cue from C_3, C_4, C_5, and C_6. Instances of C_3 do not appear frequently and would be *irrelevant* to the overall semantic context of a given video. On the other hand, instances of C_4 appear in nearly every frame. While this object certainly describes the context of the input video, this cue will fast-forward videos *uniformly* and be redundant for adaptive fast-forwarding. While instances of C_5 are observed in a moderate part of videos, they are significantly overlapped by those of C_1. This is another redundant case where users will obtain highly similar fast-forwarding patterns by setting high significance to either C_1 or C_5. Finally, C_6 does not violate any of the problems shown above. Fast-forwarding patterns obtained by selecting C_1, C_2, and C_6 are all dissimilar and not redundant.

Algorithm Details

We formalize the aforementioned cue selection criteria as follows. Let $C_{\text{all}} = \{C_1, \ldots, C_N\}$ be a set of all object categories obtained via object detection and $C \subset C_{\text{all}}$ be a set of the categories already selected as a cue. Our algorithm is based on the following objective function defined over a set of categories:

$$F(C) \quad = \quad A(C) - B(C). \qquad (1)$$

$A(C)$ is the *overall coverage* term that indicates the number of frames where at least one of the categories in C is observed. On the other hand, $B(C)$ is the *the individual coverage* term describing the number of frames with the object category observed most frequently. In each greedy step, we select $c \in C_{\text{all}} \setminus C$ that maximizes $F(C \cup \{c\})$. With this maximization, $A(C)$ helps to avoid an irrelevant category like C_3 and a temporally-overlapping category like C_5 in the previous example. On the other hand, $B(C)$ acts as a constraint to prevent each step from selecting categories that are temporally dominant like C_4 in our example. The initial cues (C_1 and C_2 in the example) are selected by exhaustively searching a set of two categories for the ones that maximize the function $F(C)$.

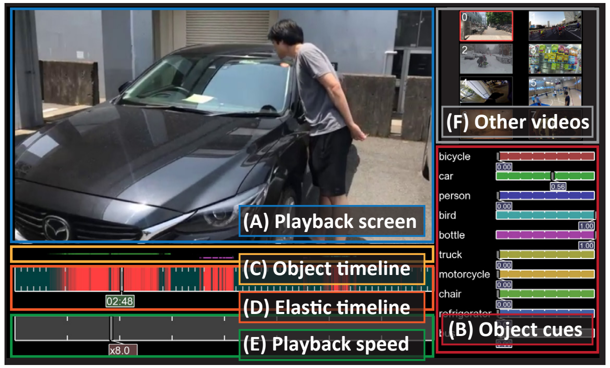


Figure 3: DO-Scanning interface. While inheriting the general layout of EgoScanning on (A) playback screen, (D) elastic timeline, (E) playback speed bar, and (F) links to other objects, we present as a new functionality, (B) a set of object cues to emphasize a part of videos and (C) object timeline indicating where specified objects are located.

In DO-Scanning, the number of object cues is fixed to 10. For processing 10 minute video recorded at 30 fps, the object detector required about 15 minutes on a GPU (NVIDIA Titan X), while the cue selection process can be done in only 1.8 seconds on the CPU.

DO-Scanning Interface

Figure 3 shows the layout of DO-Scanning. Videos are played in area (A), cues are arranged in area (B), links to other videos are listed in area (F), the playback speed is specified with bar (E), and the elastic timeline is shown in (D). Moreover, we introduced the object timeline which indicates where specified objects are located in area (C).

Experiment

As a preliminary study to validate the effectiveness of DO-Scanning, we first observed what objects cues were selected by the DO-Scanning for first-person videos of various scenes. More specifically, we applied our cue selection algorithm to 5 diverse scenes: strolling in the street, playing in an amusement park [1], cooking at home [5], strolling in the park, playing volleyball, some of which were available as public datasets for computer vision research or the others were uploaded to YouTube.

Then, we conducted a user study to compare the proposed DO-Scanning interface with EgoScanning [2]. Sixteen university students (Female: 4) served as a participant. We asked them to watch fast-forwarded videos shown above while manipulating object (DO-Scanning) or egocentric (EgoScanning) cues provided by each interface. Finally, we conducted an interview session for about 10 minutes to receive qualitative feedback.

Results

Cue Selection

Figure 4 shows cue selection results for two videos: A) Strolling in the street and B) Playing in an amusement park. Figure 4 also depicts some examples of video frames, objects that are selected (top five ones) or omitted (the bottom ones) by our algorithm, and timelines that represent a sequence of frames. In addition, results for the other video are shown in Figure 5: some examples of video frames, a part of most frequently-observed object categories obtained via object detection [6] as well as a part of object cues that our algorithm selected greedily from a set of detected objects.

Qualitative Feedback

After using both of DO-Scanning and EgoScanning in the user study, many participants reported positive feedback on DO-Scanning. Some qualitative feedback shared by multiple participants are summarized below: **A1 (10 participants out of 16):** *"It was easy to infer what types of*

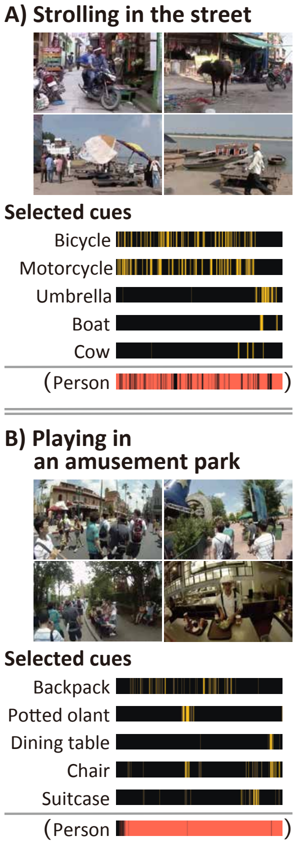


Figure 4: Some examples of video frames, objects that are selected (yellow) or omitted (red) by our algorithm and timelines that represent a sequence of frames.

scenes were emphasized by each object category"; **A2 (8):** *"DO-Scanning was user-friendly because the temporal range emphasized by cues was limited"*. On the other hand, EgoScanning interface got negative feedback: **A3 (13):** *"cues in EgoScanning could not be used effectively as they often emphasized most parts of videos and did not fast-forward videos adaptively"*.

Participants provided mixed feedback on the number of cues: **A4 (5):** *"As the DO-Scanning provided 10 cues, I was able to choose as many cues"*; **A5 (2):** *"Too many and adaptively changing object cues made it difficult to find appropriate cues to find target scenes"*. Finally, negative opinions about DO-Scanning were mainly about video E) playing volleyball: **A6 (5):** *"As target scenes in volleyball videos were not characterized by objects but human motion, it was not a good idea to focus on object categories. I would want the "hand" cue to watch the video"*.

Discussions

As shown in Figure 4 and 5, we confirmed that the object cues were certainly arranged dynamically as different object categories were selected for each video. For instance, while 'person' was detected among all videos, they were not selected as a cue in the video (A): strolling in the street and video (B): playing in an amusement park. This is because pedestrians were detected in nearly every frame in those videos (the red rectangles in Figure 4). Our algorithm can prevent such object categories from being a part of temporally-dominant cues and instead pick objects observed in a moderate part of videos.

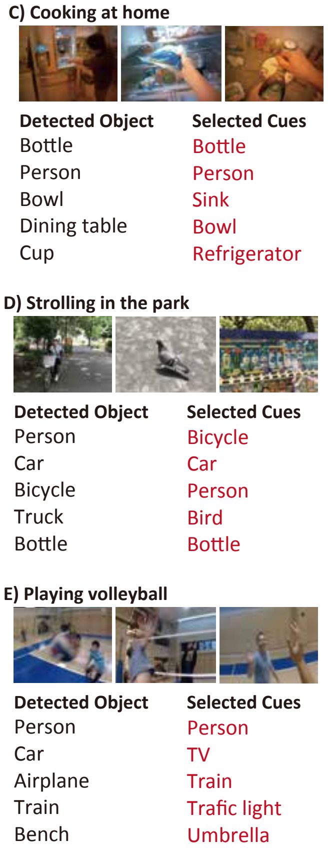
Qualitative feedback mentioned that object cues from our algorithm assisted users to watch the video (**A1** and **A2**). Particularly, participants acknowledged object cues since they often emphasized only a limited part of videos (**A2**), and EgoScanning obtained negative feedback on this point (**A3**). We found that this type of cues were particularly helpful for users to browse long and untrimmed videos. We also found that the number of cues is a relatively sensitive parameter that affects user experience on the use of DO-Scanning (**A4** and **A5**). This number of cues is currently fixed and has to be specified in advance. One interesting extension of the DO-Scanning is to adaptively choose not only object categories to add but also the maximum number of categories that can be accepted based on user preferences.

On the other hand, our experiments suggest a potential limitation of the current form of DO-Scanning. Object cues were not useful when target scenes did not particularly involve objects but were characterized mostly by camera wearers' motion (**A6**). Such scenes were, for instance, blocking a ball and setting a ball in volleyball videos. Although we implemented object cues and human behavior cues separately in DO-Scanning and EgoScanning, these two types of cues can be used together.

Conclusions

We presented DO-Scanning, an interactive video player based on the elastic timeline that adaptively fast-forwards videos based on automated content analysis and user inputs. As the key technical contribution, the DO-Scanning generates a set of object cues tailored to the context of a given video.

We believe that our approach based on dynamically-arranged object cues has made the concept of elastic timeline much more applicable to videos taken under a variety of scenarios. Along this direction of development, one promising extension for future work is to generate a set of cues in the same manner but from a large variety of



C) Cooking at home

Detected Object	Selected Cues
Bottle	Bottle
Person	Person
Bowl	Sink
Dining table	Bowl
Cup	Refrigerator

D) Strolling in the park

Detected Object	Selected Cues
Person	Bicycle
Car	Car
Bicycle	Person
Truck	Bird
Bottle	Bottle

E) Playing volleyball

Detected Object	Selected Cues
Person	Person
Car	TV
Airplane	Train
Train	Trafic light
Bench	Umbrella

Figure 5: Some examples of video frames, some most frequently-observed objects, and a part of object cues selected by our algorithm.

content analysis results, including not only object detection used in DO-Scanning. Such an extension will further help the elastic timeline work on a variety of user needs to watch first-person videos, such as finding scenes with specific persons, specific places, and specific activities, which we visually experience in our everyday life.

Acknowledgements

This work was supported by JST ACCEL (grant JPMJAC1602) and JST CREST (grant JPMJCR14E1).

REFERENCES

1. Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. 2012. Social Interactions: A First-Person Perspective. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Washington, DC, USA, 1226–1233. DOI: **http://dx.doi.org/10.1109/CVPR.2012.6247805**

2. Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 6536–6546. DOI: **http://dx.doi.org/10.1109/CVPR.2013.350**

3. Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering Important People and Objects for Egocentric Video Summarization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Washington, DC, USA, 1346–1353. DOI: **http://dx.doi.org/10.1109/CVPR.2012.6247820**

4. Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. 2012. Objects As Attributes for Scene Classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer-Verlag, Berlin, Heidelberg, 57–69. DOI: **http://dx.doi.org/10.1007/978-3-642-35749-7_5**

5. Yin Li, Alireza Fathi, and James M. Rehg. 2013. Learning to Predict Gaze in Egocentric Video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Washington, DC, USA, 3216–3223. DOI: **http://dx.doi.org/10.1109/ICCV.2013.399**

6. Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Washington, DC, USA, 7263–7271.

7. Michel Melo Silva, Washington Luis Souza Ramos, Joao Pedro Klock Ferreira, Mario Fernando Montenegro Campos, and Erickson Rangel Nascimento. 2017. Towards Semantic Fast-Forward and Stabilized Egocentric Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer International Publishing, Cham, 557–571. DOI: **http://dx.doi.org/10.1007/978-3-319-46604-0_40**

8. Jia Xu, Lopamudra Mukherjee, Yin Lo, Jamieson Warner, James M Rehg, and Vikas Singh. 2015. Gaze-Enabled Egocentric Video Summarization via Constrained Submodular Maximization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Washington, DC, USA, 2235–2244. DOI: **http://dx.doi.org/10.1109/CVPR.2015.7298836**