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

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

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

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

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

CHI'18 Extended Abstracts, April 21–26, 2018, Montréal, QC, Canada.
ACM ISBN 978-1-4503-5621-3/18/04.
http://dx.doi.org/10.1145/3170427.3186501

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 adaptively fast-forwards the video while keeping scenes with highlighted cues played at original speed. Our experimental results have revealed that the DO-Scanning has an efficient and compact set of cues arranged dynamically 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

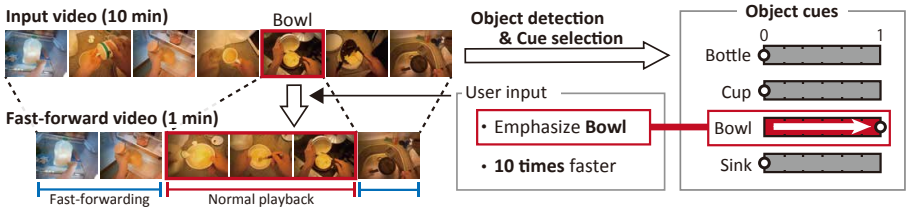

Figure 1: Dynamic Object Scanning (DO-Scanning)

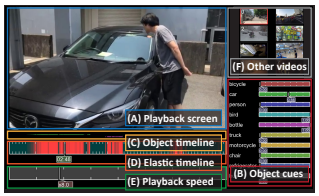

Figure 2: DO-Scanning interface. While inheriting the general layout of EgoScanning on (A) playback screen, (D) elastic timeline, (E) playback speed bar, and (F) links to other objects, we present as a new functionality, (B) a set of object cues to emphasize a part of videos and (C) object timeline indicating where specified objects are located.

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 browse first-person videos of such diverse visual experiences quickly.

Various techniques to support browsing long and untrimmed first-person videos have been studied [1, 3]. In particular, we are interested in *elastic timeline* [3], which allows users to input their preferences interactively. Based on the inputs, the elastic timeline adaptively fast-forwards videos while playing significant scenes at a lower speed.

Despite its conceptual novelty, practical applications of the elastic timeline are still limited. While [3] allows users to operate 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 get 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 the *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 generates a set of cues in a greedy manner while excluding useless object categories such as irrelevant categories hardly observed in the videos and temporally dominant categories observed uniformly in videos and cannot be used for adaptive fast-forwarding.

Dynamic Object Scanning

Figure 2 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).

Dynamic and Semantic cue

In order to generate a good set of cues for adaptive fast-forwarding a diverse collection of first-person videos, first,

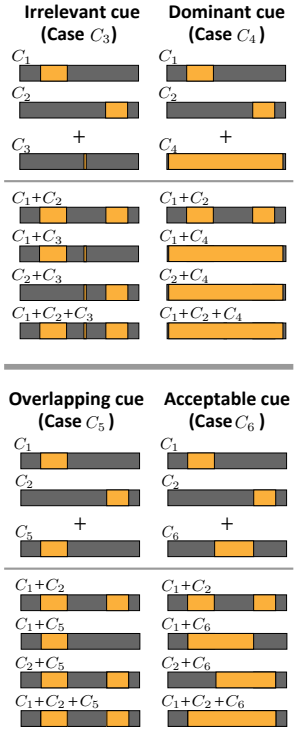

Figure 3: Greedy selection of object categories for constructing an efficient and compact set of cues. Each rectangle represents a sequence of frames. Orange rectangles describe frames which will be emphasized by setting high significance to individual object categories from C_1 to C_6 or their combinations like C_1+C_2.

we chose cues that describe the semantic context of each given video dynamically. Particularly, we propose to use *object cues*, the presence of certain object categories in videos. In this work, we first run the YOLOv2 object detector [5] trained on several different object databases including MS COCO [4] to detect 80 object categories.

Compact and Efficient Sets of Cues

While object cues get users access to the semantic context of a given video, it is not obvious how to arrange them dynamically to use handily in the elastic timeline. To generate a compact and efficient set of cues, we propose a greedy algorithm. We show how our algorithm works with an example in Figure 3. 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.

Irrelevant cue: Instances of C_3 do not appear frequently and would be irrelevant to the overall semantic context of a given video.

Dominant cue: 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*.

Overlapping cue: 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.

Acceptable cue: 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. So C_6 is chosen as a new cue in the end. This way, our algorithm grows a set of object cues in a greedy fashion.

Algorithm Details

More formally, 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:

$$F(C) \quad = \quad A(C) - B(C). \tag{1}$$

$A(C)$ 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 $F(C)$.

Experiment

As a preliminary study, we extracted objects cues from a dataset of first-person videos recorded in various scenes. We collected the dataset under 3 diverse scenarios: strolling in the street, playing in an amusement park [2], shopping at a store, some of which were available as a part of public datasets for computer vision research or the others were uploaded to YouTube.

Results and Discussions

Figure 4 shows cue selection results for three videos: A) Strolling in the street, B) Playing in an amusement park and C) Shopping at a store. Figure 4 also depicts some

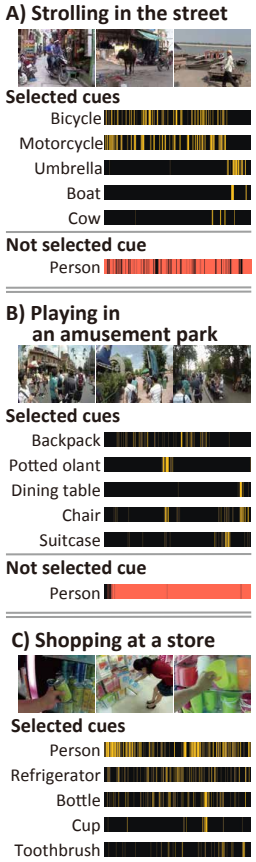

Figure 4: some examples of video frames, objects that are selected or omitted by our algorithm and timelines that represent a sequence of frames.

examples of video frames, objects that are selected or omitted by our algorithm and timelines that represent a sequence of frames. Top five objects are selected by our algorithm. Moreover, about video A) and B), we add an object which is detected in many frames but not adopted as a cue. Yellow or red rectangle describe frames where each object is detected.

As shown in Figure 4, 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 object cues and pick objects that are observed in a moderate part of videos.

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 confirmed that our algorithm successfully omits many cues that just appear most parts of videos like 'person' cues in videos of the strolling in the street.

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 content analysis results. 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. Connor Dickie, Roel Vertegaal, Jeffrey S. Shell, Changuk Sohn, Daniel Cheng, and Omar Aoudeh. 2004. Eye Contact Sensing Glasses for Attention-Sensitive Wearable Video Blogging. In *Extended Abstracts of CHI'04*. ACM, 769–770. DOI: **http://dx.doi.org/10.1145/985921.985927**
2. Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. 2012. Social Interactions: A First-Person Perspective. In *Proc. CVPR '12*. 1226–1233. DOI: **http://dx.doi.org/10.1109/CVPR.2012.6247805**
3. Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proc CHI '17*. ACM, 6536–6546. DOI: **http://dx.doi.org/10.1109/CVPR.2013.350**
4. Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. ECCV '14*. 740–755. DOI: **http://dx.doi.org/10.1007/978-3-319-10602-1_48**
5. Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proc. CVPR '17*. 7263–7271.