Ego-Surfing First Person Videos

Ryo Yonetani[1], Kris M. Kitani[2], Yoichi Sato[1],
[1]The University of Tokyo. [2]Carnegie Mellon University.

We envision a future time when wearable cameras (e.g., small cameras in glasses or pinned on a shirt collar) are worn by the masses and record first-person point-of-view (POV) videos of everyday life. While these cameras can enable new assistive technologies and novel research challenges, they also raise serious privacy concerns. For example, first-person videos passively recorded by wearable cameras will necessarily include anyone who comes into the view of a camera – with or without consent. Motivated by these benefits and risks, we argue that one important technology to develop is the ability to automatically search first-person video repositories for the videos of a single user. Much like ego-surfing enables us to perform an Internet search with our own name, we believe that self-search in first-person videos can empower users to monitor and manage their own personal data.

In this paper, we develop a video-based self-search technique tailored to first-person videos. Since the appearance of people in these types of videos often comes under heavy occlusions, motion blur and extreme pose changes (see Fig. 1), a robust approach beyond what can be accomplished by face recognition alone is required. To this end, we propose to use motion as our primary feature. The key insight is that first-person videos of target users (target videos) can act as a unique identifier to enable a target-specific search on the first-person videos recorded by other observers (observer videos). Specifically, we utilize a global motion pattern in target videos induced when target people shake their heads. The same shake pattern can be observed where they appear in observer videos (observers see the targets shaking their heads), but in the form of a local motion pattern. Therefore, this correlation between global and local motions is expected to serve as a salient cue for target instances in the observer videos.

Our proposed algorithm first generates candidate regions of target instances in observer videos and extracts a local motion pattern from the regions. With a global motion pattern extracted from a target video, we then estimate 'targetness' for each region based on the correlation between those two patterns. Instead of utilizing people detection for candidates [3], we introduce a more general supervoxel representation [5]. Supervoxels as the candidates can extend correlation-based self-search to accept first-person videos with significant egomotion where people detection is not applicable.

The main technical challenge of this work arises when localizing target instances based on the correlation evaluated at each supervoxel. Namely, an image region corresponding to a target person is not necessarily found as one supervoxel, but likely to be under- or over-segmented. For example, supervoxels under-segmenting a target region merge with a part of backgrounds. To solve this problem, our approach generates a hierarchy of supervoxels to seek preferable segmentation of target regions. While evaluating the correlation for each supervoxel, we also learn a discriminative model to consider generic targetness of supervoxels and avoid incorrect segmentation of target regions potentially involved among the hierarchy. We frame an overall procedure as a binary-class Bayesian inference problem. That is, we aim to refine the likelihood derived from correlation-based targetness by the prior targetness learned in the discriminative model.

Specific measures of targetness are given as follows. Correlation-based targetness is computed on the subspace spanned by global motion patterns in target videos. Since the global motions are usually consistent with target head motions, projecting local motions onto this subspace effectively eliminates many irrelevant (*i.e.,* non-target head) local motions such as hand gestures of observers. Prior targetness is learned from many pairs of observer video clips and corresponding masks annotating target regions. Features are designed to consider traits of generic targetness and incorrect segmentation such as the size and motion sparsity of supervoxels. While manual annotations of target people are required here, this prior is independent of specific people and backgrounds. Therefore, learning of the prior needs to be carried out only once, and is not necessary for each target video.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

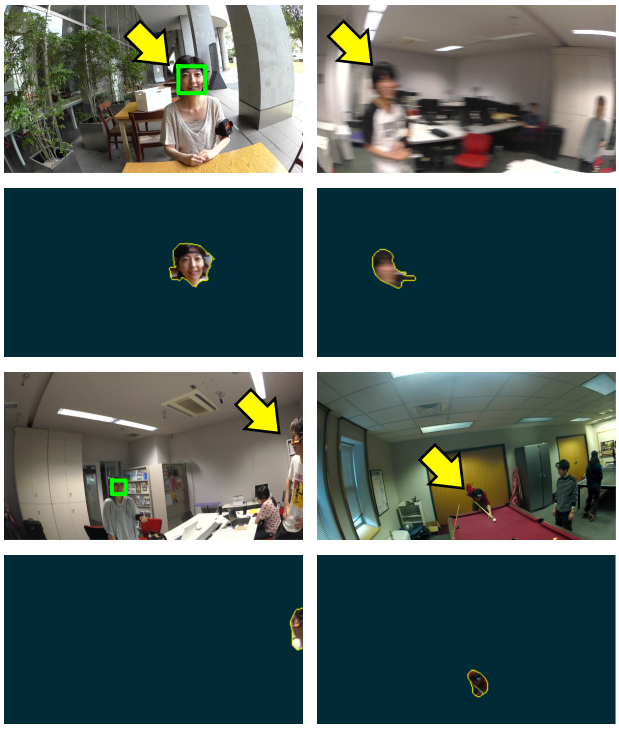

Figure 1: Robust self-search results. Self instances (**yellow arrows**) are detected (**unmasked regions**) despite heavy occlusion (**bottom-left**), motion blur (**top-right**) and extreme pose (**bottom-right**), where face recognition fails (**green rectangles**).

In our experiments, we built a new first-person video dataset to evaluate how our approach performs on self-search in details. The dataset is available at http://www.hci.iis.u-tokyo.ac.jp/datasets/. We also evaluated the effectiveness of proposed approach on the CMU-group first-person video dataset used in [2]. Our approach successfully improved search performance (area under ROC curve) over several well-known face detectors and recognizers such as [1, 4], on average by 12 percentage points on our dataset and 24 percentage points on CMU dataset. In the paper, we also suggested several proof-of-concept application scenarios that require self search as an important pre-processing, such as privacy filtering of target instances, automated video collection based on targetness and social group discovery via clustering with targetness-based affinities.

Acknowledgment

This research was supported by CREST, JST and Kayamori Foundation of Informational Science Advancement.

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28(12):2037–2041, 2006.

[2] H. S. Park, E. Jain, and Y. Sheikh. Predicting Primary Gaze Behavior using Social Saliency Fields. In *ICCV*, pages 3503 – 3510, 2013.

[3] Y. Poleg, C. Arora, and S. Peleg. Head Motion Signatures from Egocentric Videos. In *ACCV*, pages 1–15, 2014.

[4] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 57(2): 137–154, 2004.

[5] C. Xu, C. Xiong, and J. Corso. Streaming Hierarchical Video Segmentation. In *ECCV*, pages 1–14, 2012.