# Unconstrained ego-centric videos with eye-tracking data

Keng-Teck Ma, Rosary Lim, Peilun Dai, Liyuan Li and Joo-Hwee Lim
Institute for Infocomm Research, A*STAR, Singapore
{makt, rosary-lim, daip, lyli, joohwee}@i2r.a-star.edu.sg

## Abstract

*We present the first eye-tracking dataset for unconstrained ego-centric videos. The dataset captures over 6 hours of subjects performing common daily activities. These activities are manually annotated as socializing, walking, object manipulating, transiting and observing.*

## 1. Introduction

Computer-based scene understanding systems process image sequence frame by frame, and pixel by pixel within each frame, aiming to aggregate pixels into coherent regions (e.g. segmentation) for meaningful interpretation (e.g. object recognition). Is this exhaustive approach a good way for solving ill-posed visual perception and cognition problems? Human visual systems are driven by visual attention whereby eye movements facilitate the selection of relevant areas in scene image to focus (by fovea) and process, while keeping a broad picture with summary statistics in peripheral visual field. Why can't we develop a saccade-based visual information processing approach which is both more natural and efficient? Do we have proper dataset and evaluation metric to study and benchmark this type of research?

Although bottom-up saliency-based attention helps to anchor visual fixation and has been an active area of research for many years, more often than not, task-based top-down visual attention and contextual priming play more important role in directing our visual computational resources to accomplish our activities [4, 6, 10, 11].

We aim to facilitate the saccade-based visual information processing approach of scene understanding by creating unconstrained ego-centric videos with eye-tracking dataset for such research purposes. The video and eye-tracking data is recorded while participants are engaged in daily activities (*e.g.* socializing, commuting and object manipulations) in unconstrained settings. The unconstrained settings are similar to *life-logging* in ego-centric video research.

This setup is different from existing ego-centric eye-tracking video datasets in controlled environments [1, 5, 12]. To the best of our knowledge, this is the first unconstrained ego-centric video dataset with eye-tracking information.

## 2. Unconstrained dataset

Six participants, 4 males and 2 females, were recruited from a population of graduate students and office workers. Their ages range from 23 to 31 years old. They have normal eye-sight or are wearing contact lens.

We used the SMI Eye-Tracking Glasses (ETG) version 1. It records the video in 24 frames per second and the gazes are sampled at $30Hz$. The resolution of the front facing camera is 1280x960. The field-of-view is 60 degrees visual angle. The eye-tracker was calibrated with the 3-point calibration for each recording session.

Participants were instructed to wear the mobile eye-trackers whenever it was convenient for them. There was no instruction on the type of activities they should participate, except that they should avoid sports due to the risks of damaging the equipment; and avoid driving due to limited field-of-view. They were further instructed to record at least 10 minutes of data for each session. Fixations were extracted from the gaze samples with the vendor's software (BeGaze).

## 3. Annotations

The dataset was annotated by 3 volunteers for the following overall information: Time of Day, Place (*e.g.* home, office, subway), Indoor/Outdoor and a short description.

For each manually selected video segment, the annotators also labeled a short description; and assigned it one or more of the activities: Social, Walk, Object, Transit, Observe.

Social refers to socializing activities such as talking, listening, meetings *etc*. Walk refers to self-locomotive activities such as walking and running *etc*. Object refers to activities in which hands are used to manipulate objects such as packing, holding, assembling *etc*. Transit refers to activities on moving platform such as elevators, escalators *etc*. Observe refers to passive viewing such as scenery viewing.

# 4. Qualitative Analysis

We performed qualitative analysis on our datasets to better understand the various factors and cues which influence fixations (visual attention).

## 4.1. Task-centric

In daily activities, fixations are determined by the current task [4, 10, 11]. During social interactions, speakers' faces capture attention. Subjects will orient their gazes and heads to face the speakers.

For object manipulation, objects of interests are fixated upon [2, 4, 5]. However, besides object manipulation activities, the hands are rarely visible within the field of view [4]. Indirectly, this highlights the fact that vision, as a scarce resource, will be uncoupled from an action once another sensory modality, *e.g*. proprioception, takes over.

In walking/running, the fixations are divided between direction of travel and ground plane [4].

In other tasks such as transiting and observing, bottom-up saliency often plays a major role.

## 4.2. Embodiment

Embodiment refers to the fact that unlike screen viewing, the participants are physically present in the visual environment. There are several implications as a result:

(a) Attention is shifting to avoid hazards and collisions, especially during walking. This is clearly not present in third person videos (TPV) [8].

(b) Gazes are directed towards a speaker's face in social interactions [7] while averted from strangers' gazes. Gaze, in this case, is deployed both as an attention mechanism and a communication tool. A mutual gaze indicates willingness to communicate and vice versus.

(c) Oculomotor structure pre-determines the distributions of fixations. It is easier to move our eyes to look down; and to move our head to look up. For sustained attention, it is tiring to maintain our gaze in off-center positions [3].

## 4.3. Multi-tasking

Subjects are frequently engaged in multiple tasks, e.g. walking and talking; object manipulation and walking etc. The subjects alternate their gaze between the two (or more) targets for several cycles [9].

Attention models [2, 5, 12] derived from single tasks (*e.g*. walking, food preparation) in controlled environments are unable to represent the richness and prevalence of multi-tasking in our daily activities. This is the distinctive difference from other prior datasets.

# 5. Conclusion and Acknowledgment

We presented and described an unconstrained egocentric videos with eye-tracking dataset. This dataset reveals several interesting characteristics which will facilitate development of saccade-based visual information processing approach for scene understanding.

We would like to thank Chengyao Shen, Gabriel Kreiman, Qi Zhao and Tomaso Poggio for their advice. This work is supported by the Reverse Engineering Visual Intelligence for cognitiVe Enhancement (REVIVE) programme funded by the Joint Council Office (JCO) of A*STAR.

# References

[1] H. Boujut, J. Benois-Pineau, and R. Megret. Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 436–445. Springer, 2012.

[2] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012*, pages 314–327. Springer, 2012.

[3] D. Guitton, R. Kearney, W. N, and B. Peterson. Visual vestibular and voluntary contributions to human head stabilization. *Experimental Brain Research*, 64:59–69, 1986.

[4] M. F. Land. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, 25(3):296–324, 2006.

[5] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3216–3223. IEEE, 2013.

[6] K.-T. Ma, T. Sim, and M. Kankanhalli. VIP: A unifying framework for computational eye-gaze research. In *4th International Workshop on Human Behavior Understanding*. Springer, 2013.

[7] A. Rahman, D. Pellerin, and D. Houzet. Influence of number, location and size of faces on gaze in video. *Journal of Eye Movement Research*, 7(2):1–11, 2014.

[8] C. Tan, H. Goh, V. Chandrasekhar, L. Li, and J.-H. Lim. Understanding the nature of first-person videos: Characterization and classification using low-level features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[9] B. Tatler, I. Gilchrist, and M. Land. Visual memory for objects in natural scenes: from fixations to object files. *The Quarterly Journal of Experimental Psychology*, 58:931–960, 2005.

[10] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5, 2011.

[11] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[12] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Advances in Image and Video Technology*, pages 277–288. Springer, 2012.