Egocentric Multi-Modal Dataset with Visual and Physiological Signals

Katsuyuki Nakamura[1,2] Alexandre Alahi[1] Serena Yeung[1] Li Fei-Fei[1]
[1]Stanford University [2]Hitachi, Ltd.
katsuyuki.nakamura.xv@hitachi.com, alahi@stanford.edu, {syyeung,feifeili}@cs.stanford.edu

Abstract

We introduce a dataset of egocentric video augmented with physiological heart rate and acceleration signals. To the best of our knowledge, our work is the first to leverage visual information with physiological signals. We believe this dataset will have great value by enabling new work directions in egocentric video understanding, ranging from activity detection to video summarization.

1. Introduction

The use of wearable sensors, such as heart rate monitors and accelerometers, is widespread as a way to track daily activities. To detect these activities as accurately as possible, these sensors are used in many combinations and configurations and are attached to body parts ranging from the chest to the wrist and foot. More recently, egocentric first-person cameras [2] have gained popularity as a new modality of wearable sensor. Activity recognition from visual data is a widely studied problem in computer vision, and a number of works have investigated this task in the domain of egocentric video [4, 3, 5, 7] and in combination with other wearable sensors [8, 6].

However, the degree to which computer vision algorithms alone can be effective for activity understanding has not been established. We would also like to understand how much relevant information still requires other sensor modalities and how these other signals should be best fused with vision. Pioneering works [8, 6] showed that systems of inertial measurement units and audio and accelerometer sensors, respectively, could improve activity segmentation compared to using only egocentric cameras in indoor settings. However, we are not aware of an existing dataset that supports research in egocentric video augmented with physiological sensors.

We therefore introduce the ECM^2 dataset, which comprises 30 hours of egocentric video augmented with heart rate and acceleration data (Figure 1). We believe that ECM^2 is a comprehensive, realistic, and challenging dataset for egocentric multi-modal activity understanding and that it can also be of interest for additional applications including video summarization.

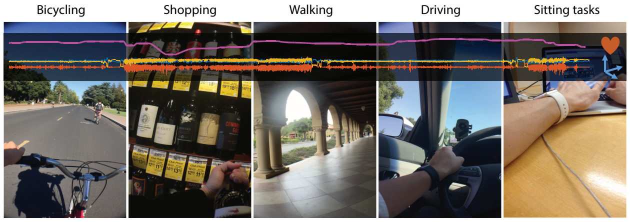

Figure 1. **ECM^2 dataset**, consisting of egocentric video enhanced with heart rate and acceleration signals.

2. ECM^2 Dataset

Data Collection Data was collected using a sensing system comprising a mobile phone and a wrist-worn heart rate sensor. The mobile phone was placed in the chest pocket of the shirt of each subject to collect egocentric video and accelerometer data, and the wrist sensor provided the corresponding heart rate data. Eight subjects wore sensing systems, recording 100 videos for a total of 30 hours. The lengths of the individual videos varied from 5 minutes to 2 hours. Subjects were only instructed to perform daily life activities without constraint on how, where, or in what environments. Data collection was therefore performed under natural daily conditions.

The mobile phone collected egocentric video data at 1280×720 resolution and 30 fps, as well as triaxial acceleration data at 30Hz. The mobile phone was equipped with a wide-angle lens, so that the horizontal field of view was enlarged from 60 degrees to about 82 degrees. The wrist-worn heart rate sensor was used to capture the heart rate data every 5 seconds (0.2 Hz). The phone and heart rate monitor were time-synchronized through Bluetooth, and all data was stored in the phone's memory. Piecewise cubic polynomial interpolation was used to fill in any gaps in heart rate data. Finally, data was aligned to the millisecond level at 30 Hz.

Activity and MET Annotations We defined the activity classes in ECM^2 as subsets of a compendium of physical activities [1]. Each of these activity classes is associated with a metabolic equivalent of task (MET) measure [1].

Activity	MET	Activity	MET
Bicycling Uphill	14.0	Shopping	2.3
Bicycling	7.5	Strolling	2.0
Circuit Training	4.3	Talking-standing	1.8
Stair Climbing	4.0	Talking-sitting	1.5
Playing with Children	3.5	Sitting Tasks	1.5
Touring	3.5	Meetings	1.5
Walking	3.5	Eating	1.5
Cooking	3.3	Riding	1.3
Presenting*	3.0	Reading	1.3
Driving	2.5	Background*	N/A

Table 1. **Definitions of activity classes and their MET values** [1]. * indicates additional class.

MET is a physiological measure defined relative to the resting metabolic rate, and expresses the energy cost of physical activities. For instance, sitting quietly is considered to be 1.0 MET, and bicycling is considered to be 7.5 MET ($\mathrm{kcal} \cdot \mathrm{kg}^{-1} \cdot \mathrm{h}^{-1}$). MET is a simple and practical measure and widely used to quantify activities. We selected 18 activity classes from the compendium [1] and added two additional classes, *Presenting* and *Background* (see Table 1).

3. Dataset Statistics

Among 20 activity classes, *walking* appears in the highest number of videos; 42 of the 100 videos contain some walking. *presenting* appears in the lowest number of videos. The average number of distinct activity types per video is 3.35, and there are often multiple occurrences of the same activity in a video. The durations range from several seconds to half-an hour. The longest duration is *presenting*, which occurred for 2,740 seconds (45.6 minutes) in a single video. The shortest duration is an instance of *talking-sitting*, which occurred for 5.8 seconds. This range presents challenging scenarios for activity detection.

Figure 2(a) shows a scatter plot of median heart rate vs. acceleration variance per class. This plot quantifies the correlation between heart rate and acceleration variance, which has a correlation coefficient of $r = 0.76$. Figure 2(b) visualizes the MET distribution for heart rate vs. acceleration variance and shows that heart rate is a strong indicator of MET.

4. Conclusion

In this work, we introduced ECM^2, which is a dataset of egocentric video augmented with heart rate and acceleration signals. We believe this dataset will enable many new work directions on various aspects of egocentric video understanding, ranging from activity detection to video summarization. We will make the ECM^2 dataset and software for data collection publicly available.

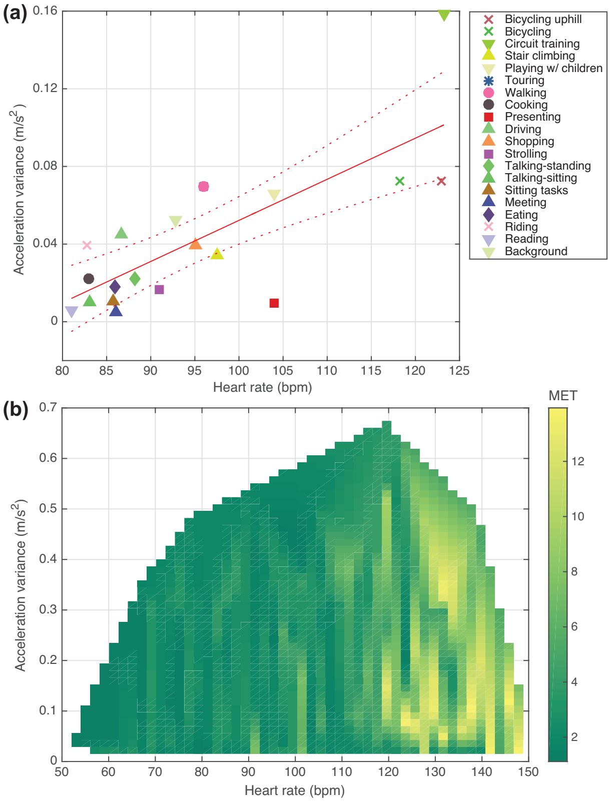

Figure 2. **Statistics of ECM^2 dataset**. (a) Scatter plot of heart rate vs. acceleration variance, and (b) MET distribution for heart rate vs. acceleration variance.

References

[1] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Bassett, C. Tudor-Locke, J. L. Greer, J. Vezina, M. C. Whitt-Glover, and A. S. Leon. 2011 Compendium of Physical Activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, 43(8):1575–1581, 2011. 1, 2

[2] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The Evolution of First Person Vision Methods: A Survey. *IEEE Trans. Circuits and Systems for Video Technology*, 25(5):744–760, 2015. 1

[3] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social Interactions : A First-Person Perspective. In *CVPR*, pages 16–21, 2012. 1

[4] P. Hamed and D. Ramanan. Detecting Activities of Daily Living in First-person Camera Views. In *CVPR*, 2012. 1

[5] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, pages 3241–3248, 2011. 1

[6] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome. Object-based activity recognition with heterogeneous sensors on wrist. In *Pervasive*, pages 246–264, 2010. 1

[7] Y. Poleg, C. Arora, and S. Peleg. Temporal Segmentation of Egocentric Videos Yair Poleg. In *CVPR*, 2014. 1

[8] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshop on Egocentric Vision*, 2009. 1