

# Detecting Engagement in Egocentric Video

Yu-Chuan Su and Kristen Grauman

University of Texas at Austin

Our supplementary materials consists of:

- Video examples comparing our method to the baselines
- Details on the annotation interface used for our data collection (B)
- Details on the precision/recall evaluation metric (C)
- Details on the start point detection experiment (D)
- Breakdown analysis of per-scenario performance (E)
- Analysis for the required amount of training data (F)

## A Example Intervals

Please refer to our project webpage for example video, our method’s predictions, the ground truth and multiple baselines.

## B Annotation Interface

In this section, we show the interface and instructions for engagement annotation on Amazon Mechanical Turk. Please refer to our project webpage for the interface in action.

### B.1 Task Description

George is wearing a camera on his head. The camera captures video constantly as George goes about his daily life. Because the camera is on his head, when George moves his head to look around, the camera moves too. Basically, it captures the world just as George sees it.

Your job is to watch a video excerpt from George’s camera that lasts 1-2 minutes, and determine when **something in the environment has captured George’s attention**. You will first watch the entire video. Then you will go back and use a slider to navigate through the video frames and mark the intervals (start and end points) where he is paying close attention to something. **Note, the video may have more than one interval where George is paying close attention to something.**

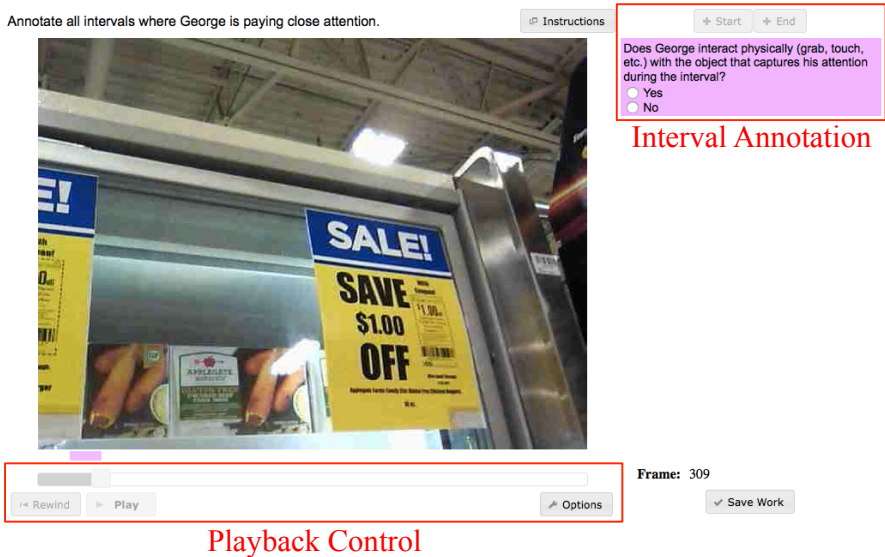


Fig. 1. Screen shot of annotation interface.

## Definition of Attention

The following instructions will describe what we mean by “capturing George’s attention” in more detail: Humans’ cognitive process has different levels of attention to the surrounding environment. For example, people pay very little attention to their surroundings when they are walking on a route they are familiar with, but the attention level will rise significantly if there are unusual events (such as a car accident) or something attracts their curiosity (such as a new advertisement on the wall), or if they want to inspect something more closely (such as a product on the shelf when shopping). You are asked to identify these “high attention intervals” in the video.

**In particular, we ask you to identify intervals where George’s attention is focussed on an object or a specific location in the scene.** During these intervals, George is attracted by an object and tries to have a better view/understanding about it intentionally. In general, George may:

- Have a closer look at the object
- Inspect the object from different views
- Stare at the object

In some situations, George may even interact physically with the object capturing his attention to gather more information. For example, he may grab the object to have a closer view of it, or he may turn the object to inspect it from different views. To identify these situations, we also ask you to annotate **whether George touched the object** capturing his attention during the interval.

The following video shows examples of attention interval: *please refer to our webpage*.

## Important Notes

- You should watch the entire video (3 minutes) first before doing any annotation. This will give you the context of the activity to know when George is paying close attention.
- A video may contain **multiple or no** intervals where George’s attention is captured. You should label each one separately. The intervals are mutually exclusive and should not overlap.
- Each interval where George’s attention is captured may vary in length. Some could be a couple seconds long, others could be closer to a minute long. The minimum length of each interval is 15 frames (1 second).
- You may need to scroll back and forth in the video using our slider interface to determine exactly when the attention starts and stops. Mark the interval as tightly as possible.
- After labeling where an attention interval starts and ends, you will mark whether George has physical contact (grab, touch, etc.) with the object during the interval or is just looking at it.
- You will also mark your confidence in terms of how strongly George’s attention was captured in that interval (Obvious, Fairly clear, Subtle).

## B.2 Interface Introduction

The following introduction will give you tips on how to best use the tool. Please watch the below video (and/or read the below section) for instructions: *please refer to our webpage*.

### Getting Started

- Press the **Play** button to play the video.
- After the video finished, press the **Rewind** button and start annotation.



- Play the video, **Pause** the video when you reach the frame at the beginning of high attention interval.
- Click the **Start** button to mark the “Start” of the interval.



Attention - Start from: 118



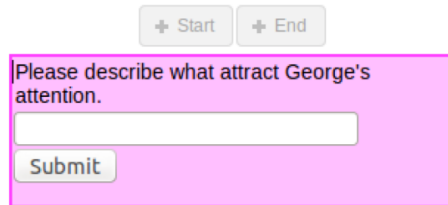
- On the right, directly below the Start button, you will find a colorful box showing the frame number corresponding to the ‘Start’ of the interval.
- Similarly, click the **End** button to mark the “End” of the interval.
- After you mark the end of the interval, you will be asked whether George contact (grabbing, touching, etc.) the object that captures his attention.
- Next, you will be asked about how obvious is the attention interval. Specify whether the interval is **Obvious**, **Fairly clear**, **Subtle**.



How obvious is the attention interval?

☐ Obvious  
☒ **Fairly clear**  
☐ Subtle

- Finally, you will be asked to describe what attracts George’s attention. Type in what attracts George’s attention (object, scene, event, etc.) and **Submit** the interval.



Please describe what attract George's attention.

- When you are ready to submit your work, rewind the video and watch it through one more time. Do the “Start” and “End” you specified cover the complete high attention interval? After you have checked your work, press the **Submit HIT** button. We will pay you as soon as possible.
- Do **not** reload or close the page before redirected to next hit. This may cause submission failure.

## How We Accept Your Work

We will hand review your work and we will only accept high quality work. Your annotations are not compared against other workers.

## Keyboard Shortcuts

These keyboard shortcuts are available for your convenience:

- **t** toggles play/pause on the video
- **r** rewinds the video to the start

- **d** jump the video forward a bit
- **f** jump the video backward a bit
- **v** step the video forward a tiny bit
- **c** step the video backward a tiny bit

## C Evaluation Metric

In this section, we describe our evaluation metric in detail. Let  $G$  denote a set of ground truth intervals for engagement. The set of intervals is consistent if none of the intervals within the set overlap with others, denoted by  $|g_1 \cap g_2| = 0$ ,  $\forall g_1, g_2 \in G$ .  $g_1 \cap g_2$  denotes the frames that are in both interval  $g_1$  and  $g_2$ . Also, let  $P$  denote a set of predicted intervals that is consistent.

We consider a predicted interval  $p$  to be covered by a ground truth interval  $g$  if  $\frac{1}{2}|p \cap g| > |p|$ , denoted by  $p \subset g$ . Given the ground truth intervals  $G$  and predictions  $P$ , we define the interval precision as follows:

$$Precision = \frac{|\{\exists g \in G \text{ s.t. } p \subset g \mid \forall p \in P\}|}{|P|}.$$

Similarly, we consider a ground truth interval  $g$  to be covered by a predicted interval  $p$  if  $\frac{1}{2}|p \cap g| > |g|$ , and we compute the interval recall as

$$Recall = \frac{|\{\exists p \in P \text{ s.t. } g \subset p \mid \forall g \in G\}|}{|G|}.$$

Note the recall monotonically increases as we prolong the length of each prediction  $p$  in  $P$ . Roughly speaking, a predicted interval  $p$  is considered correct if more than 50% of the prediction overlaps with some ground truth interval, and a ground truth interval is considered predicted if more than 50% of the interval is covered by some prediction.

## D Start point correctness

Figure 5 in the main paper evaluates our method and the baselines for the “start point” detection task. To compare the start point accuracy of different methods, we plot the  $F_1$  score as a function of error tolerance window (in seconds) allowed between the predicted and the nearest ground truth start point. Our method outperforms all other methods under all error tolerances. This is evidence that our method has promise for both the online and offline setting, though we think there remains interesting future work to best account for streaming data.

The Motion Magnitude baseline is our nearest competitor for this setting. This indicates that an abrupt decline in motion is predictive for the transition between engagement and non-engagement (e.g., as a person slows to examine something). However, it remains weaker than our method, and, as we see in the other results in the main paper, it cannot predict the continuation and subsequent drop of engagement level.

	Mall Market Museum		
CNN	0.640	0.691	0.624
Ours – frame	0.643	0.713	0.682
Ours – interval	0.665	0.739	0.693

**Table 1.** Frame  $F_1$ -score for each scenario in cross-scenario setting on UT EE.

## E Breakdown per-scenario result

In this section, we show the breakdown per-scenario result of the Cross-Scenario setting in Table 3 of the main paper. This analysis provides further insight about when our motion-based method outperforms the baseline appearance based method. The frame  $F_1$  scores are shown in Table 1.

While our method performs better in all scenarios, the breakdown reveals some properties of the scenarios. For example, results on Market are the best, suggesting the set of actions people perform in their daily life (Market) is smaller than that in more rare situations (Mall/Museum), and there are more actions that people perform only in Mall/Museum scenarios than those people only perform in Market. Our margins over CNN are largest on Museum, suggesting our motion based method better generalizes from daily life to rare/unique scenes.

## F Influence of training data size

The following results show that our method can achieve reasonable performance with modest amount of training data: when training one model for each recorder from their own data (i.e., only 10% of the training data), the F1-score drops by 5%; in the Cross-Recorder setting, our method still outperforms the best baseline (CNN) using only 1 training video ( $\sim 30$  minutes). Nevertheless, increasing the number of training data always help.