Supplementary Material: KrishnaCam: Using a Longitudinal, Single-Person, Egocentric Dataset for Scene Understanding Tasks

Krishna Kumar Singh[1,3] Kayvon Fatahalian[1] Alexei A. Efros[2]

[1]Carnegie Mellon University [2]UC Berkeley [3]UC Davis

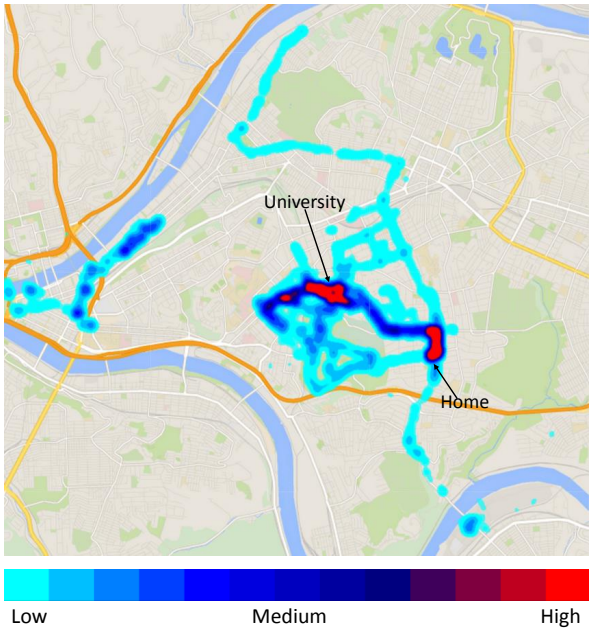


Figure 2. Recording was conducted throughout the city. Locations visited rarely are shown in light blue. Red regions are the most frequently visited, and correspond to areas surrounding the student's home and university.

1. An Outdoor Egocentric Video Dataset

In total, the dataset contains 70.2 hours of 720p, 30 fps video (7.6 million total frames) making it significantly larger than prior single-individual egocentric datasets recently studied in computer vision [1, 3]. As shown in Figure 1-A, data was collected regularly over the span of nine months (164 of the 250 days in this span), with the largest amount of recording on a single day being 149 minutes. Most recording took place between the hours of 11am and 9pm (Figure 1-B), and thus the dataset contains video recorded at dusk and at night. The most active day for recording was Sunday (Figure 1-C) since the student had more time to be outdoors collecting video.

Figure 2 plots the location and density of recordings in the student's home city, where most recording took place (92% of dataset). While recording has occurred in neighborhoods throughout the city (the displayed map spans about 30 square kilometers), as expected most recording took place near the student's home and university (the red color indicates a larger number of distinct recordings at a location). The most common sequence in the dataset is walking from the student's home to his local bus stop, which occurs 95 times.

1.1. First Occurrence of Common Scenes

It is interesting to consider when common scenes are observed for the first time in the data, since the distribution of these first encounters provides intuition about the long-tail of life situations. We identify these situations by finding frames whose similarity to the top-5 neighbors drawn from *prior* recordings is low (no scene like it has been seen before), but whose similarly to the top-5 neighbors drawn from the *entire* dataset is high (the image is common in the context of all recording). Figure 3 shows a selection of frames that meet this criteria for various time ranges during recording. (For these frames, the average cosine distance to the five most similar frames in the entire dataset is greater than that to the five most similar prior frames by at least 0.2.) The first five hours of recording contain the first instances of many common scenes such as walking around campus and to and from work. The next 15 hours capture the first instances of waiting at a bus stop, eating outdoors, and less traveled paths. Although the locations in the bottom row of Figure 3 were often visited during early recordings, they are observed with snow for the first time after 20 hours of recording.

1.2. Distribution of People

Using Dollar's pedestrian detector [4, 2], we determined that 17% of frames in the dataset contain at least one person (all bars, Figure 4-left). Thus it is interesting to consider when the student sees them. The fraction of frames containing at least one person is highest around noon (when the

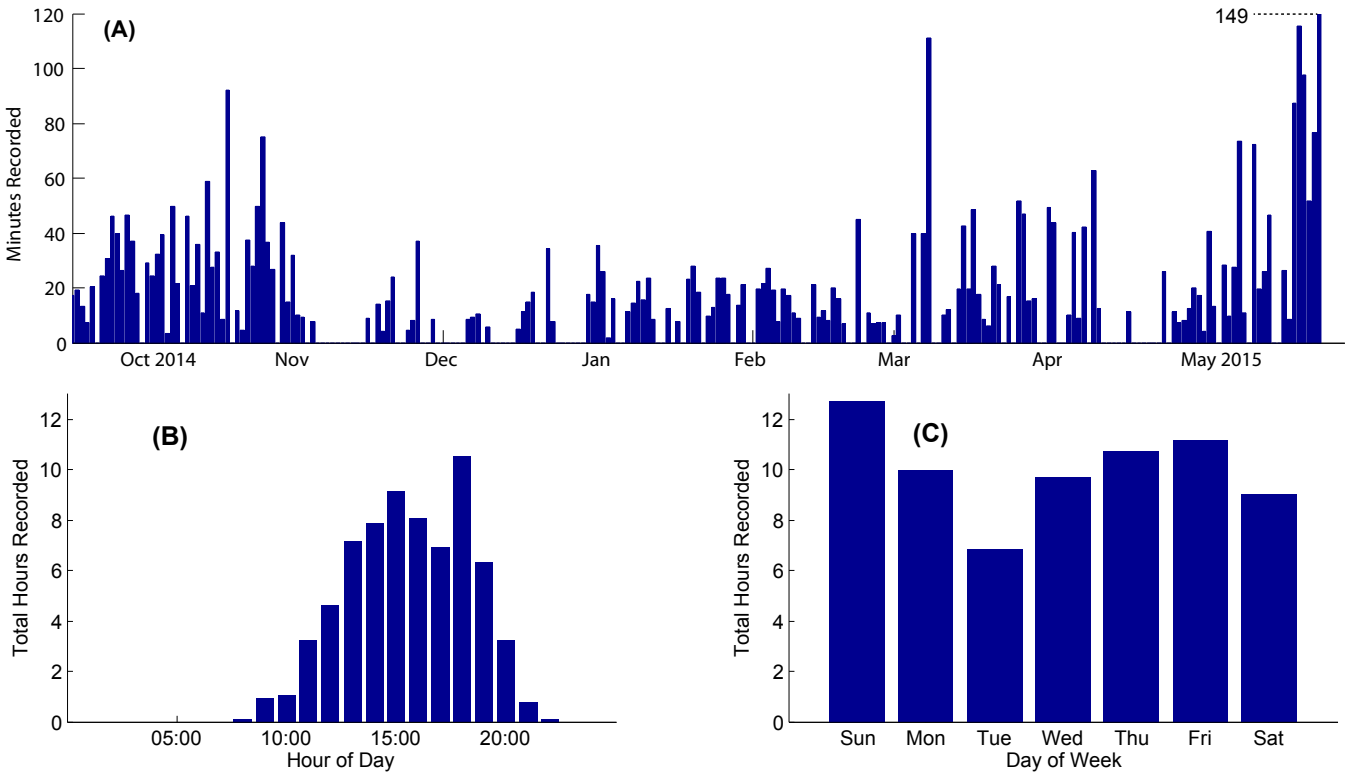


Figure 1. Breakdown of the duration of recording by day, by time of day, and day of week.

Figure 3. First occurrences of commonly observed scenes.

student is walking to lunch, Figure 4-center), and in general higher on weekdays than weekends (campus and the surrounding eateries are more lively, Figure 4-right).

2. Comparison with GPS

While it may seem reasonable to use GPS position for motion prediction, measurements from present consumer GPS technology are too spatially and temporally coarse to make high-quality fine-scale, short-time horizon motion predictions. Figure 5-top shows an example trajectory recorded using GPS. Even though the student moved along the sidewalk in this situation, the GPS trajectory indicates incorrect motion which deviates from the actual movement.

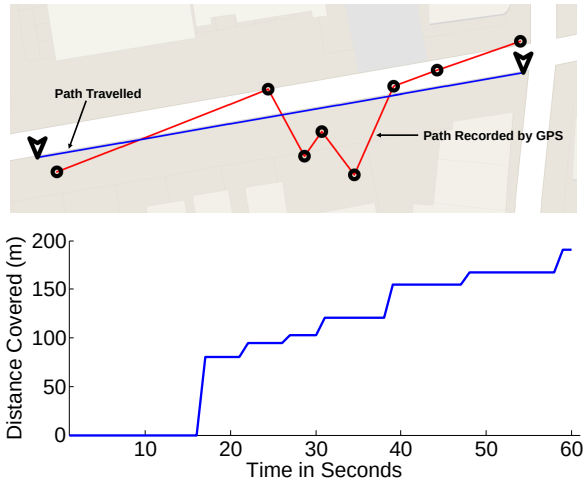


Figure 5. Even though the student was walking continuously on the sidewalk, GPS measurements were too spatially (top) and temporally coarse (bottom) to be useful for motion prediction.

Further, GPS coordinates updated only eight times in the span of 60 seconds (Figure 5-top) in this example. We find that even when the camera had previously visited at the same location as the test frame, when compared to predictions using visual similarly provided by MIT Places-Hybrid network, GPS-position-based similarity yielded trajectory prediction error 26.3% greater (on average for the 20,000

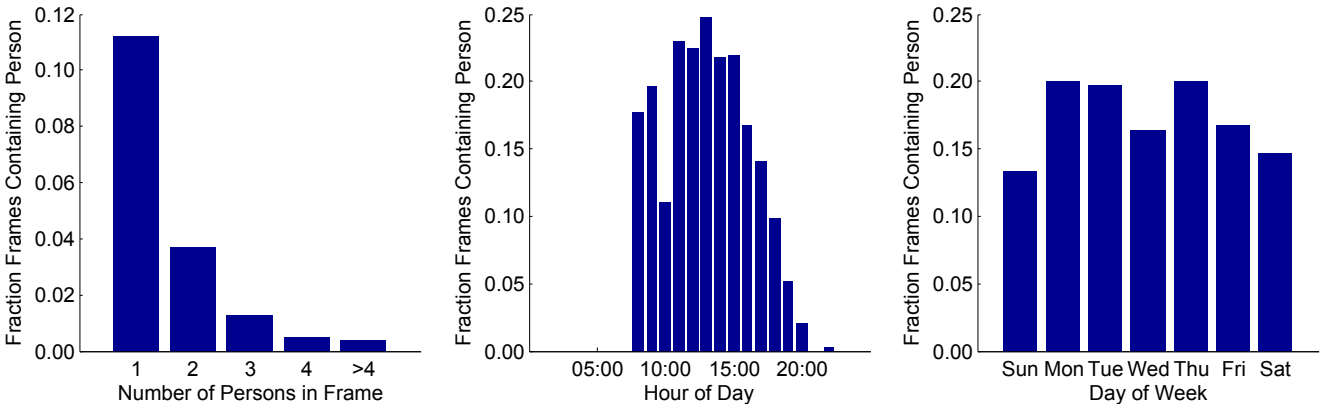
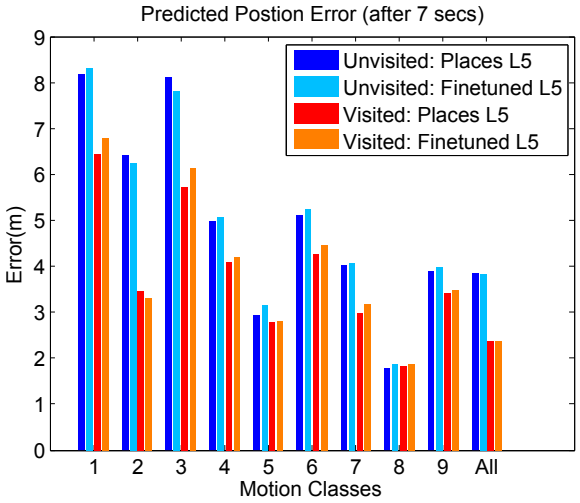


Figure 4. The fraction of frames containing people, by number of people (left), time of day (center), and day of week (right).



	Unvisited	Visited	Overall
MIT Places-Hybrid L5	3.85	2.38	3.11
Fine Tuned L5	3.83	2.36	3.10

Figure 6. Similar trajectory prediction error is obtained from neighbors computed using MIT Places-Hybrid layer-5 descriptors and layer-5 descriptors from our network finetuned for motion class prediction.

visited location frames randomly chosen between 38 and 52 hours of the dataset and error is distance in meters between predicted and ground-truth position seven seconds into the future). Although it is likely possible to interpolate and filter the GPS signal to get better results, techniques based on geographic position will not generalize to new locations and will never be able to predict camera movement that is based on transient properties of a scene. For example, GPS based predictions fail in situations like stopping to talk to friends or to avoid a moving car (visual-similarity can successfully makes such predictions).

3. Comparison with Fine-tuning

Using layer-5 response for the MIT Places-Hybrid network and our fine-tuned network, we evaluate the quality of predictions for 40,000 frames: 20,000 unvisited location frames and 20,000 visited location frames randomly chosen between 38 and 52 hours of the dataset. We assess prediction quality by evaluating the distance (in meters) between the predicted position and the measured T_i seven seconds into the future.

Figure 6 plots prediction error for both similarity metrics. (For clarity we report overall error for the test set, and also categorize results by the motion classes.) Overall, we fail to observe notable benefit from class-based fine-tuning. Closer inspection reveals that that fine-tuning improves prediction accuracy in stationary situations, at the cost of decreasing performance slightly for the other motion classes.

Inspection of the nearest neighbor results generated by the MIT Places-Hybrid network and the fine tuned network does reveal significant qualitative differences in the character of the resulting neighbors. As shown in Figure 7-(A-B), the fine tuned network emphasizes objects in the foreground in its definition of similarity. (These visual elements, such as hands, people, and food are indicative of a need to stop.) We also find the fine-tuned features also emphasize general features of sidewalks and roads (C,D,F).

Figure 7 also provides evidence of why, even though fine tuning yields compelling visual neighbors, those neighbors don't translate into quantitatively better predictions. For example in (E) while all neighbors provided by the fine-tuned network contain images of stairs to the left, these neighbors are images of different stairwells, and the student's motion behavior differs in these locations (the place network produces exact matches in this scenario, and the student's motion is consistent in this location). In case (F), while the fine-tuned network produces arguably better visual matches capturing the shape of the sidewalk ahead, the student departed from sidewalk.

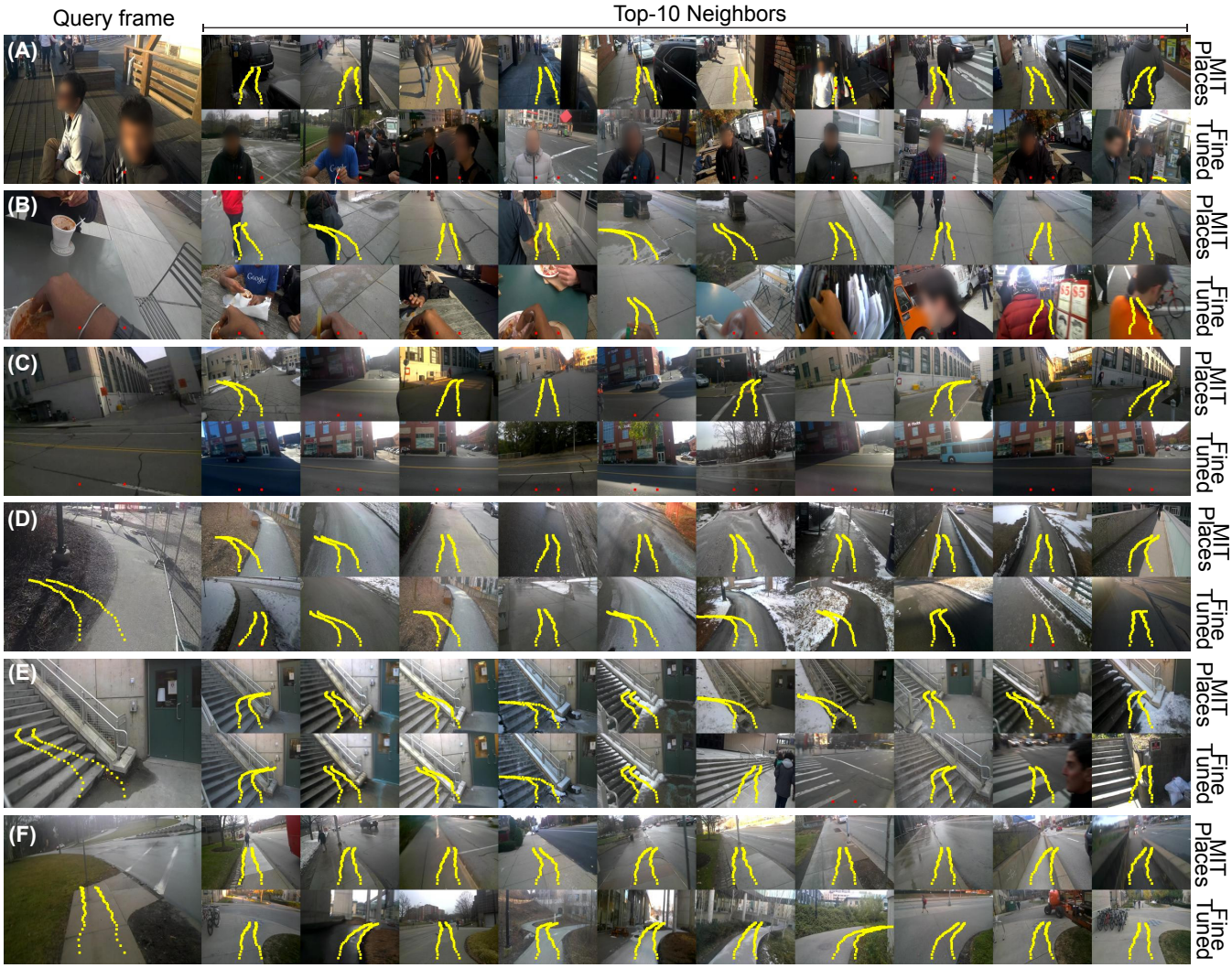


Figure 7. (A) Fine-tuned nearest neighbors matching people. (B) Fine-tuned nearest neighbors matching eating scenarios with hand in focus (original MIT Places-Hybrid network prioritizes elements from the road) (C) Fine-tuned nearest neighbors picks up horizontal road strips which indicates observer is waiting. (D) Fine-tuned neighbors capture left turn of the query frame. (E) Top 10 neighbors generated by the placed network are exact match with query image (same location), fine-tuned neighbors matches different staircases (F) Fine-tuned neighbors capture the right turn pattern of the query image, much the student does not follow the sidewalk.

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3297–3304, Washington, DC, USA, 2011. IEEE Computer Society.

[2] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`.

[3] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1346 –1353, june 2012.

[4] J. H. H. Woonhyun Nam, Piotr Dollár. Local decorrelation for improved pedestrian detection. In *NIPS*, 2014.

[5] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.