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

1. Introduction

We record, and analyze, and present to the community, KrishnaCam, a large (7.6 million frames, 70 hours) egocentric video stream along with GPS position, acceleration and body orientation data spanning nine months of the life of a computer vision graduate student. We explore and exploit the inherent redundancies in this rich visual data stream to answer simple scene understanding questions such as: How much novel visual information does the student see each day? Given a single egocentric photograph of a scene, can we predict where the student might walk next? We find that given our large video database, simple, nearest-neighbor methods are surprisingly adept baselines for these tasks, even in scenes and scenarios where the camera wearer has never been before. For example, we demonstrate the ability to predict the near-future trajectory of the student in broad set of outdoor situations that includes following sidewalks, stopping to wait for a bus, taking a daily path to work, and the lack of movement while eating food.

2. The KrishnaCam Dataset

Over a period of nine months (September 2014 to May 2015) we collected egocentric video of the daily outdoor activities of a single graduate student. Whenever possible, the student attempted to continuously record video outdoors (technical, legal, and social constraints limited the scope recording that could be performed). The dataset, acquired using Google Glass, consists of 460 unique video recordings, each ranging in length from a few minutes to about a half hour of video. Recording took place over a wide geographic area (including many different neighborhoods of the student's home city and trips out of the city), contains visual diversity due to seasonal change (snow in winter months), and day-and-night recording. The videos capture the student's movement and interactions with others in a diverse set of residential, campus, and urban areas, as well as in multiple city parks. The student's GPS position, acceleration, and orientation was also captured using a smart phone in the student's pocket, and subsequently synced with the video data. Given this collection methodology, the dataset's non-visual sensor readings describe the configuration of the student's body, not the orientation or acceleration of the head-mounted camera.

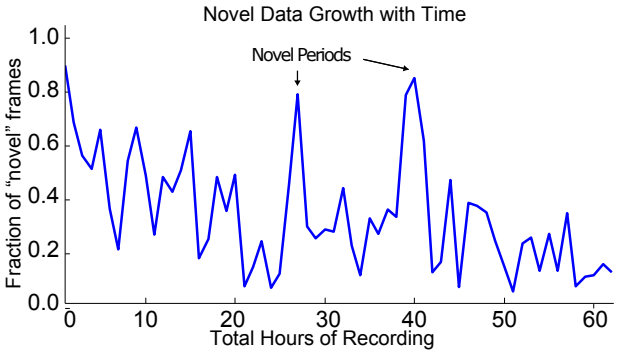
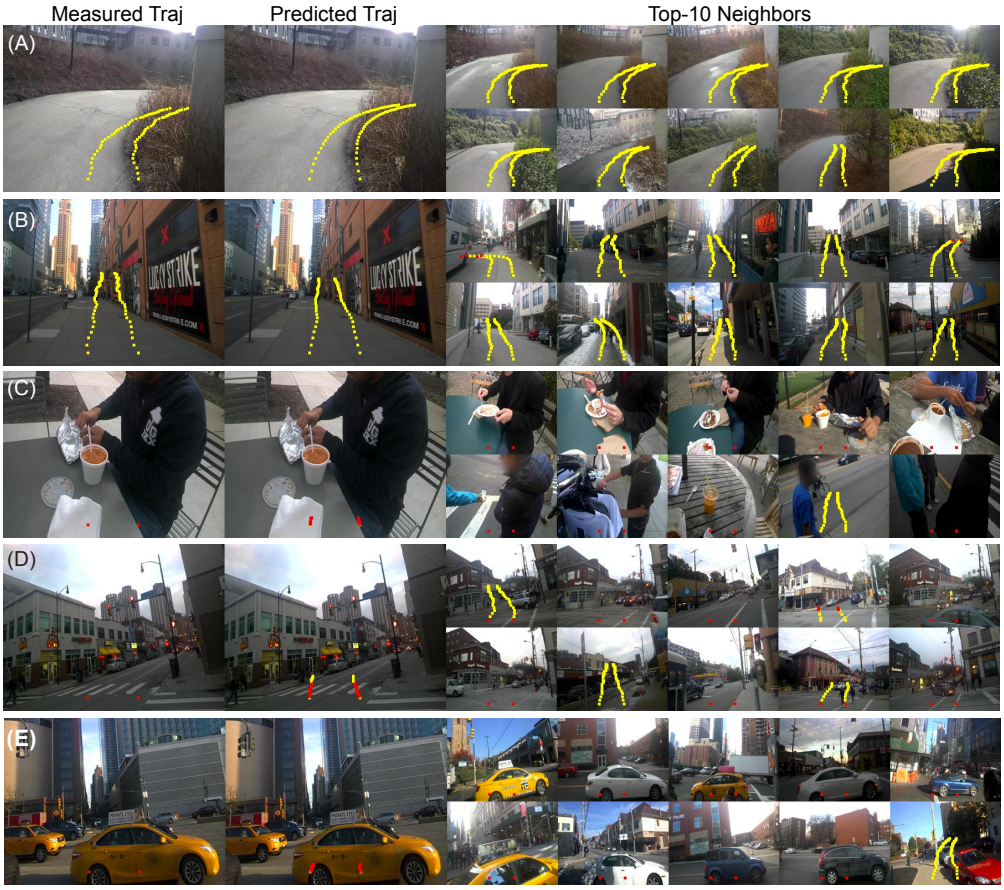


Figure 1. Due to the redundency in daily life, the rate novel frames are observed decreases with time. Days recording in new locations are easily identified as spikes in the graph near 26 and 40 hours

In total, the dataset contains 70.2 hours of 720p, 30 fps video (7.6 million total frames) making it larger than prior single-individual egocentric datasets recently studied in computer vision [1, 2].

3. Novel Visual Data Growth

Hypothesizing that the life of a computer vision graduate student is highly redundant, we attempted to quantify the amount of novel visual data observed by the camera each day. Specifically, for each frame, we identify its top-5 nearest neighbor frames from prior recordings. We use cosine similarity between layer-5 outputs (after pooling) of the MIT Places-Hybrid network [3] as a distance metric for nearest neighbor computations. We label a frame as novel if the average similarity of its top-5 nearest neighbors is below a threshold, or if five valid neighbors do not exist given the selection constraints. Given this definition, Figure 1 plots the fraction of novel frames observed in each hour of the first 60 hours of recording. As to be expected, at the start of recording a large fraction of frames are novel, but this



Prediction of general behaviors that hold across different events and/or locations: (A-B) following a sidewalk (in both frequently visited and novel locations) (C) remaining stationary while eating food, (D-E) stopping at new intersections or when there is traffic.

Figure 2. Examples of successful trajectory predictions.

fraction drops as more data is recorded. The two peaks of the graph (steep rises in the amount of novel visual data) correspond to days the student spent outside his home city.

4. Predicting Trajectories

We attempted to use the motion-annotated video dataset (KrishnaCam) to address the simple scene-understanding question: *given a single image, can we predict where the student would walk next in the scene*. We estimate the student's motion from accelerometer and orientation sensor readings taken from a smart phone in the student's pocket. For trajectory prediction, we lean on the rich visual history contained in our database and pursue a nearest-neighbor-based approach. Given each new frame, we estimate the camera wearer's future trajectory as the average of the trajectories of its top-10 nearest neighbors.

Figure 2 shows that nearest neighbor-based prediction approach yields surprisingly accurate predictions across a variety of scenes. In (A-B) the system is able to predict common navigation behaviors such as the camera wearer following the path and sidewalk, (C) remaining stationary when eating, (D) stopping at an intersection, and not walking into the middle of traffic (E). In rows B-E of Figure 2, a large fraction of the nearest neighbor set comes from locations different from the query, resulting in the successful transfer of motion information to new situations and environments. (This transfer would not be possible using GPS!)

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[2] 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.

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