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

Abstract

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.

1. Introduction

"A baby has brains, but it doesn't know much. Experience is the only thing that brings knowledge, and the longer you are on earth the more experience you are sure to get."
—L. Frank Baum, The Wonderful Wizard of Oz

With video cameras rapidly becoming tinier, cheaper, and more power efficient, computers will soon have the ability to observe an increasingly large fraction of life's events. Whether it be life-logging videos, household webcams, dashboard cams, security cams, or even daytime TV, these emerging (always-on) data streams provide an endless collection of examples that document how people and objects function in the world. Of course, this data presents the challenge of finding new ways to extract value from these visual data sources.

In this paper we make two contributions. First, we have created our own streaming data source—we have recorded a large egocentric video stream (KrishnaCam) that spans nine months of the life of a computer vision graduate student. Today, the dataset consists of 7.6 million frames, but the camera will continue to be on and this dataset will grow every day (at least until the students says "I've had enough"). Second, we explore and exploit the inherent redundancies in this unique visual data stream to provide answers to simple scene understanding questions such as: *given a single egocentric photograph of a scene, can we predict where the student might walk next in the scene? How much novel visual information does the student see each day? Can we mine the dataset to establish scene changes over time, or yield unique insights into the student's environment?*

Our decision to analyze an egocentric video stream presents the challenge of interpreting data from a moving camera and from a diverse set of everyday situations (as opposed to a stationary camera in a room). However, by constraining collection to a single individual, as opposed to recording a shorter span from many individuals (or using unstructured image or video collections from the internet), we limit the events we observe to not be too diverse—a single individual's daily life experiences are only so broad. From these redundancies across a large database of observations, valuable patterns begin to emerge. For example, using simple, nearest neighbor methods we observe (and can predict), that like most humans, the student generally stops at intersections and walks straight in sidewalks, but we can be surprised by a harmless jaywalk. While we record many predictable mornings of taking the same walk to campus, we also record trips to parks and lunches with friends. As more data continues to be collected, we expect our ability to predict to continue to improve.

With new data, perhaps involving new life situations, arriving daily from our continuous video stream, we are confronted with the challenge that it is simply intractable to involve humans in the labeling of data. Thus our analyses focus on tasks that are well-suited for always-on streams. For example, properties such as visual uniqueness can be explored directly from the data itself. Also, camera motion can be reliably estimated from auxiliary sensors, affording the ability to use a large corpus of examples to make (and

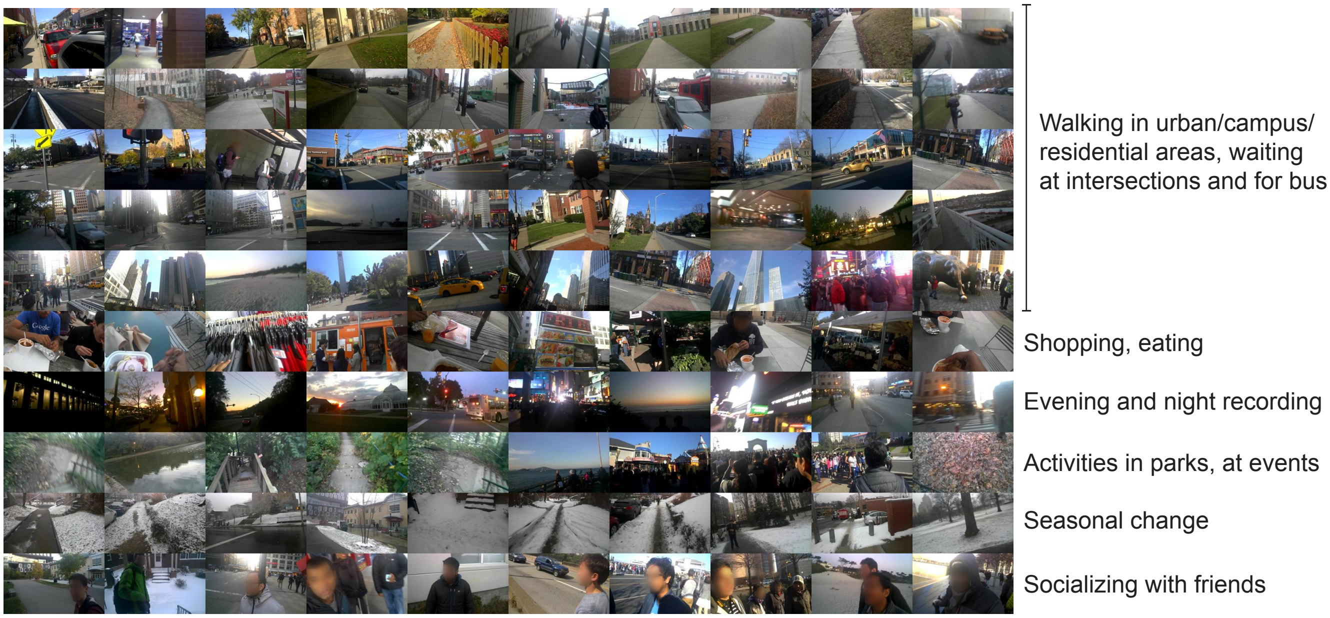

Walking in urban/campus/ residential areas, waiting at intersections and for bus
Shopping, eating
Evening and night recording
Activities in parks, at events
Seasonal change
Socializing with friends

Figure 1. Over nine months we acquired a 70 hour (7.6 million frame) egocentric video dataset capturing daily outdoor activities of a computer vision graduate student. The dataset spans a wide variety of environments and life experiences.

then automatically validate) new motion predictions.

2. Prior Work

Although early explorations of egocentric image capture such as the MyLifeBits [8] system from Microsoft Research and the U.K.'s "Memories for Life" grand challenge initiative [1] were longitudinal studies of data from a single individual, nearly all recent egocentric datasets [20, 16] have been collected by multiple individuals for very short durations and span. Aghazadeh et al. [2] collected data from a single individual for an entire month, but recorded for a only few minutes daily and along the same walking route each day. Perhaps the closest dataset to our recordings is the 6-month egocentric dataset recently produced by Castro et al. [3] (and used to learn a model to predict everyday activities). However, this dataset consists of images, not video, taken at a fixed interval (from once a minute to once per 5 minutes). Our dataset consists of video collected for several months and in a diverse set of environments.

As it has become easier to acquire egocentric video en masse, the uniqueness of egocentric content, and the difficulty of its analysis, has made it an increasingly popular target of recent study. Researchers have explored activity recognition [5, 6, 12, 20], object recognition [21, 7], summarization [16, 18, 10], and pose estimation [22] on egocentric videos. Aghazadeh et al. [2] leverage redundancy in videos to identify novel events. Lee et al. [16] attempted to remove the "boring" parts of egocentric videos by predicting important objects and events. Given that egocentric video cameras are fundamentally mobile devices, we view camera movement prediction as a challenging new task for researchers in the area to consider.

Our work in Section 5 takes a purely data-driven approach to the task of temporal prediction. Similar to prior methods [17, 27] we make no assumptions about the visual environment, require no semantic labeling of the scene, and leverage simple nearest-neighbor search of large visual databases to find examples that are likely to predict future behavior. While [17, 27] also sought to transfer object motion across different scenes in unstructured video collections, we benefit from a database that is substantially larger and also heavily biased toward the experiences of a single individual. Kaneva et al. [11] also adopt a data-driven approach to prediction using a Street View image dataset that has similar characteristics to our outdoor egocentric videos, but their work aimed to predict what might be observed if the camera undertook a specified motion, not make predictions about the camera motion itself.

Modeling and predicting the motion of agents in a scene is a primary focus of the field of trajectory-based activity analysis [19, 13, 14]. Much of this work, as well as recent work on unsupervised visual prediction [24], assumes a fixed camera viewpoint (e.g. a security camera or webcam recording a single scene), not a mobile egocentric camera that encounters a diverse array of environments and life situations. Further, most prior trajectory analysis efforts seek to predict the behavior of agents *in the scene*, whereas in our case, the agent is the videographer himself.

3. 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 dataset captures many repetitive daily activities as walking from the student's residence to and from campus, walking with colleagues to local restaurants, and waiting in line for and eating meals at outdoor fast-food stands with friends. (The most common sequence in the dataset is walking from the student's home to his local bus stop, which occurs 95 times.) However, due to the extended nature of the recording we also have captured a number of rare events, such as a trip an amusement park and sledding on a snow day. 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.

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 [2, 16]. Due to its large size, our experiments operate on a 5 fps sampling of the source videos.

4. 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 [28] as a distance metric for nearest neighbor computations since this network was trained on scene categories that bare similarity to many scenes present in our dataset. (The MIT Places-Hybrid network is the Krizkevsky et al. network topology [15] trained on 1183 categories: 205 scene categories from the MIT Places database [28] and 978 object categories from Imagenet [23].) We also experimented with similarity based on layer-7 outputs, but found the layer-5 results to yield more intuitive neighbors. To ensure diversity in the resulting neighbor list, all nearest neighbor frames are constrained to be separated by at least 10 minutes.

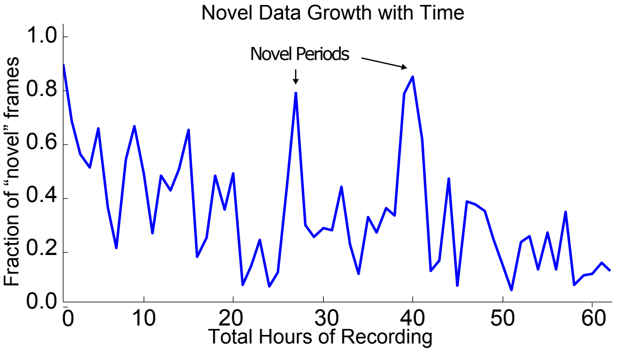


Figure 2. Due to the redundency in daily life, the rate novel frames are observed decreases with time. Days recording in new locations are easily identified as spikes in the graph near 26 and 40 hours

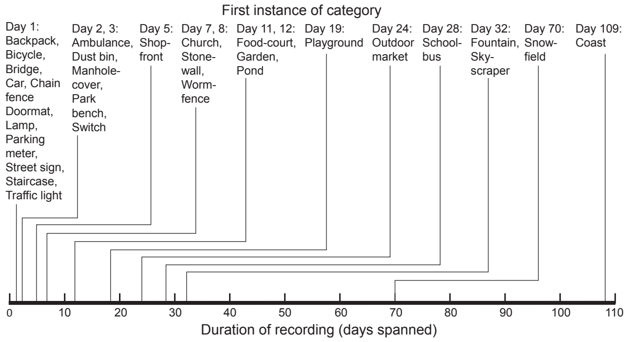


Figure 3. First occurrences of semantic classes in the dataset (as evaluated by the MIT Places-Hybrid CNN).

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. (We empirically determined 0.55 to be a reasonable threshold for novelty in our data). Given this definition, Figure 2 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 fraction drops as more data is recorded. For example, after only a few days of recording, many frames observed on the commonly traveled path between the student's home and campus are no longer novel. The two peaks of the graph (steep rises in the amount of novel visual information) correspond to days the student spent outside his home city.

In addition to identifying visually redundant frames we also identified the first occurrences of MIT Places-Hybrid CNN classes in the dataset (Figure 3). (To make the figure we chronologically sorted top classifications and manually removed false positives.) Notice that significant amounts of recording are necessary before the first instances of classes

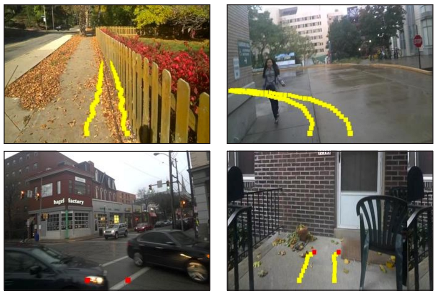


Figure 4. Example motion trajectories obtained from accelerometer and orientation sensor readings. Red dots indicate moments when the camera wearer is stationary.

like "snow field" (due to the arrival of winter), "fountain" and "skyscraper" (the first walk downtown), and "coast" (due to travel).

5. Student Motion Prediction

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 chose to focus on the task of predicting future camera motion from a single image (rather than attempt predictions based on recent video history or motion clues), because single-view scene understanding is a task that is performed by humans quite well, but remains very difficult for computers. The task is well-established in the literature (e.g. [11, 17, 27]) and is an important to address because it forces a higher-level of semantic understanding than temporal-based methods. For example, MPEG-style low-level processing can perform short-term prediction quite well without understanding anything about the scene—we wish to avoid this bias by providing only a single image at test time.

5.1. Estimating Motion Trajectories

We found GPS position measurements by commercial smart phones to be too spatially and temporally coarse to be viable motion measurements for short time scales. In urban environments, consecutive consumer GPS position readings can differ by tens of meters (significant noise compared to pedestrian velocities) yielding poor motion predictions. (See supplementary material for a quantitative comparison with the results in Section 5.3.) Instead, we estimate the student's motion from accelerometer and orientation sensor readings taken from a smart phone in the student's pocket. We trained a multi-class-SVM classifier on two-second windows of accelerometer readings to classify the student's velocity v_i at frame f_i as stationary (0 m/s), slow (0.375 m/s), regular (1.0 m/s), or fast (1.375 m/s). Given this velocity estimate and measured changes in body orientation between frames, we annotate each f_i with a 2D trajectory $T_i = [x_0, y_0, x_1, y_1 \ldots x_{N-1}, y_{N-1}]$ of the student for the next M seconds. Trajectories are sampled at 5 fps, so $|T_i| = 10M$. The motion trajectories encode position relative to the student's position and measured orientation θ_i at f_i. Therefore, for each T_i, the coordinate of the j^{th} point in the trajectory is:

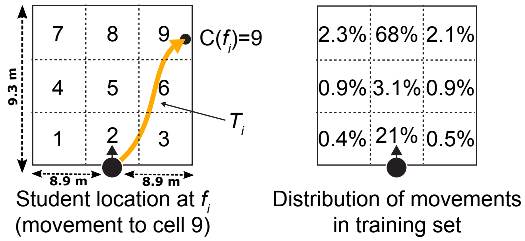


Figure 5. We classify motion by the grid cell the camera wearer ends up in after seven seconds. Real-life motion is highly biased towards walking straight (turns are rare.)

Motion Class Prediction Accuracy (Unweighted)

	Unvisited (%)	Visited (%)	Overall (%)
Fine Tuned	58.4	81.2	73.4
NN	54.9	81.4	72.2
Chance	43.2	51.3	48.5

Table 1. Motion class prediction accuracy from the fine-tuned MIT Places-Hybrid network is better than the chance and nearest-neighbor baselines. Accuracy for test images taken at never-before visited and previously visited locations is reported separately.

$$\begin{aligned} x_0 &= 0 \\ y_0 &= 0 \\ x_j &= x_{j-1} + v_{i+j} \times cos(\theta_{i+j} - \theta_i) \\ y_j &= y_{j-1} + v_{i+j} \times sin(\theta_{i+j} - \theta_i) \end{aligned}$$

We find this simple approach is sufficient to automatically generate "ground truth" trajectory information for the entire video dataset (no human labeling). Figure 4 shows several examples of these 7-second trajectories, rendered with perspective from the point-of-view of the egocentric camera. (Stationary moments of the trajectory are red.) For example, the bottom-left image depicts a future where the student stands still for 7-seconds on the street corner. In the bottom-right image the student briefly continues forward and then stops at the door.

5.2. Predicting Motion Classes

As an initial prediction experiment, we sought to predict simple, discrete descriptions of student movement. As visualized in Figure 5-left, we form discrete motion classes by partitioning the ground plane in front of the student into a 3×3 grid, and label each frame f_i with $C(f_i)$, the grid cell containing the last point in T_i (i.e., where the student will be in seven seconds). The grid is scaled to encompass the maximum 7-second displacement of the camera-wearer in the dataset. (We clamp backward vertical motion to zero.) To learn a motion predictor $C(f_i)$, we modify the

Class	Unvisited (%)	Visited (%)	Overall (%)
1 (Hard-Left)	1.2	11.2	7.9
2 (Stop)	26.2	56.2	41.8
3 (Hard-Right)	4.8	29.3	22.9
4 (Med-Left)	9.6	27.7	23.6
5 (Med-Straight)	25.6	27.3	26.6
6 (Med-Right)	7.1	22.6	18.9
7 (Soft-Left)	20.4	48.4	40.3
8 (Straight)	35.4	57.4	51.3
9 (Soft-Right)	16.8	38.9	31.6
Overall	16.4	35.4	29.4

Table 2. Per-class motion prediction accuracy resulting from weighted network training.

final softmax layer of the MIT Places-Hybrid Network [28] to predict nine motion classes rather than the original 1183 classes and then fine-tune the network (using Caffe [9]) on a training set of motion class labels from the first 38 hours (681,565 frames, September 18 to March 2) of the dataset.

When evaluated on a test set of 252,209 frames (collected between 38 and 52 hours of the dataset, March 5 to April 11), the classifier achieves 73% prediction accuracy. Its performance exceeds chance (guessing a motion category based on the distribution of training class labels yields 48.5% accuracy) as well as that of a nearest-neighbor baseline that uses layer-5 responses of the MIT Places-Hybrid network to compute neighbors of frame f_i, then computes $C(f_i)$ as the grid cell most often traveled to after these neighbors. The test set contains recordings in locations that the student has previously visited often (e.g., his walk to work), as well as images of never-before-visited locations (approximately 34% of the test set is recorded in completely novel locations). Table 1 reports classification accuracy separately for both types of testing frames.

Reflecting the real-life behavior of the student, the dataset is heavily towards instances of walking straight. (Figure 5-right shows that 68% of training frames involve a walking straight scenario: $C(f_i) = 8$, and less than 1% involve hard left or right turns.) The classifier described above reflects this skew, rarely predicting less common motion events. To improve prediction of infrequent (and arguably more interesting) motions, we also trained a classifier to maximize performance on a *single-frame* classification task (i.e., to predict motion category from a single frame, when all motion categories are equally likely).

Rather than supply equal numbers of training examples for each motion class (which would dictate using only a small fraction of the training data), we use all training frames, but for each training frame, scale the gradient used for back-propagation by the size of the frame's motion category. This modification reweights loss so all categories may influence training equally despite having unequal numbers of exemplars. Per-class prediction accuracy is given in Table 2 (chance is now 11.1%). The classifier's full confusion matrix is provided in the supplementary material.

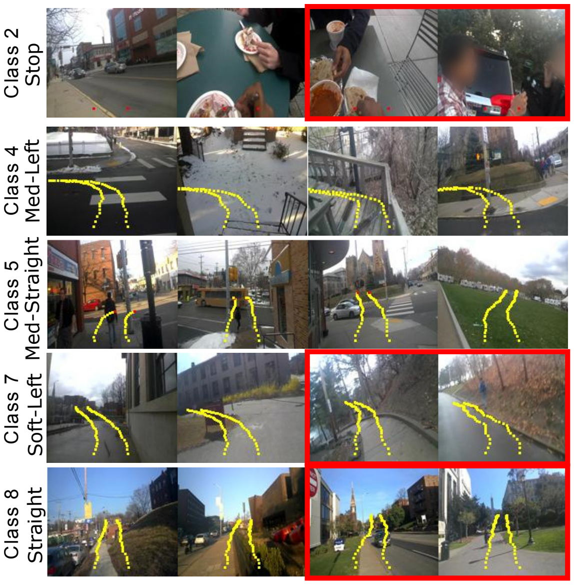


Figure 6. Most confident predictions for selected motion classes reveal visual clues that suggest upcoming motion. Red boxes indicate images from locations not visited in the training data.

As shown from the examples of top predictions in Figure 6, we observe the classifier is often able to predict future motions, even if images are from locations the student had never visited before (red squares). Visual characteristics such as roads or sidewalks suggest forward movement (Class 8), intersections indicate the possibility of stopping (Class 5), and food and people in the foreground are usually observed when stationary (Class 2).

5.3. Predicting Trajectories

Class-based prediction provides only a coarse description of future motion. We are also interested to predict richer movements, such as a gradual turn to the right, or a swerving path around an upcoming obstacle. Specifically, we seek to predict the immediate future trajectory T_i of the camera wearer in a scene depicted in a single input image.

We lean on the rich visual history contained in our database and pursue a nearest-neighbor-based approach to trajectory prediction. Given each new frame, we estimate the camera wearer's future trajectory as the average of the trajectories of its top-10 nearest neighbors. As with the prior class prediction experiments, we search the first 38 hours of recording (681,565 frames after temporal subsampling) during nearest neighbor search. Since our image database consists of frames from long video recordings (not unstructured image collections), to ensure neighbors come from a diverse set of prior experiences, we require that all returned nearest neighbor frames are separated by at least ten minutes in time. Without this condition, top nearest neighbors are often consecutive frames from a single video in the database. As described in Section 4 we use cosine

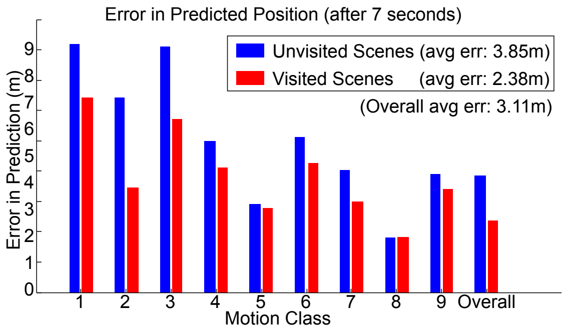


Figure 7. Error in predicted position (in meters). Results are seperated by actual motion class and by whether test image is from a location present in the training set.

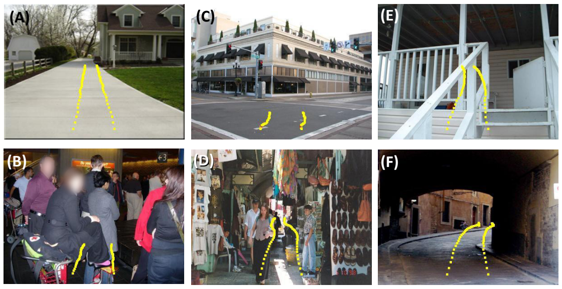


Figure 9. Motion trajectory predictions for SUN database images.

similarity of MIT Places-Hybrid layer-5 outputs [28] as a metric for nearest neighbor computations. We also experimented with similarity based on layer-7 outputs, but found the layer-5 results to yield more intuitive neighbors.

5.3.1 Quantitative Evaluation

We evaluate the quality of trajectory predictions for 40,000 frames (20,000 unvisited location frames, 20,000 previously-visited location frames) randomly chosen between 38 and 52 hours of the dataset. We assess prediction error as the distance (in meters) between the predicted position and the measured position seven seconds into the future. Figure 7 plots the error in these predictions for images. (To examine prediction error in different contexts, we separate results into previously visited and unvisited test frame groups and by the motion classes used in Section 5.2). Prediction error is greater for test images of previously unvisited places. It is also higher in the presence of turns.

5.3.2 Qualitative Evaluation

Figure 8 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, G of Figure 8, 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!)

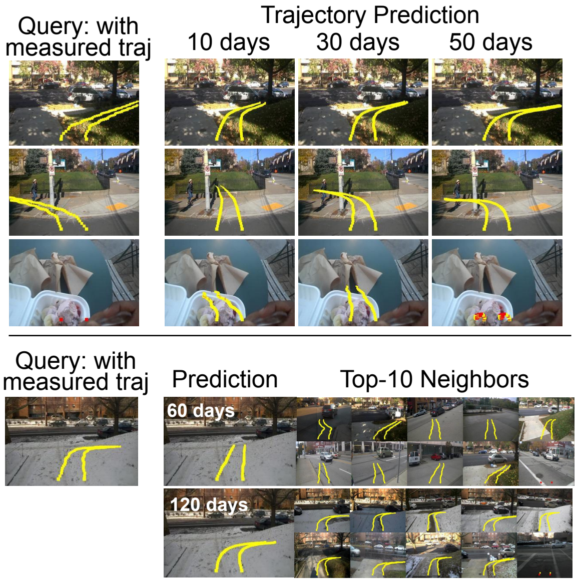
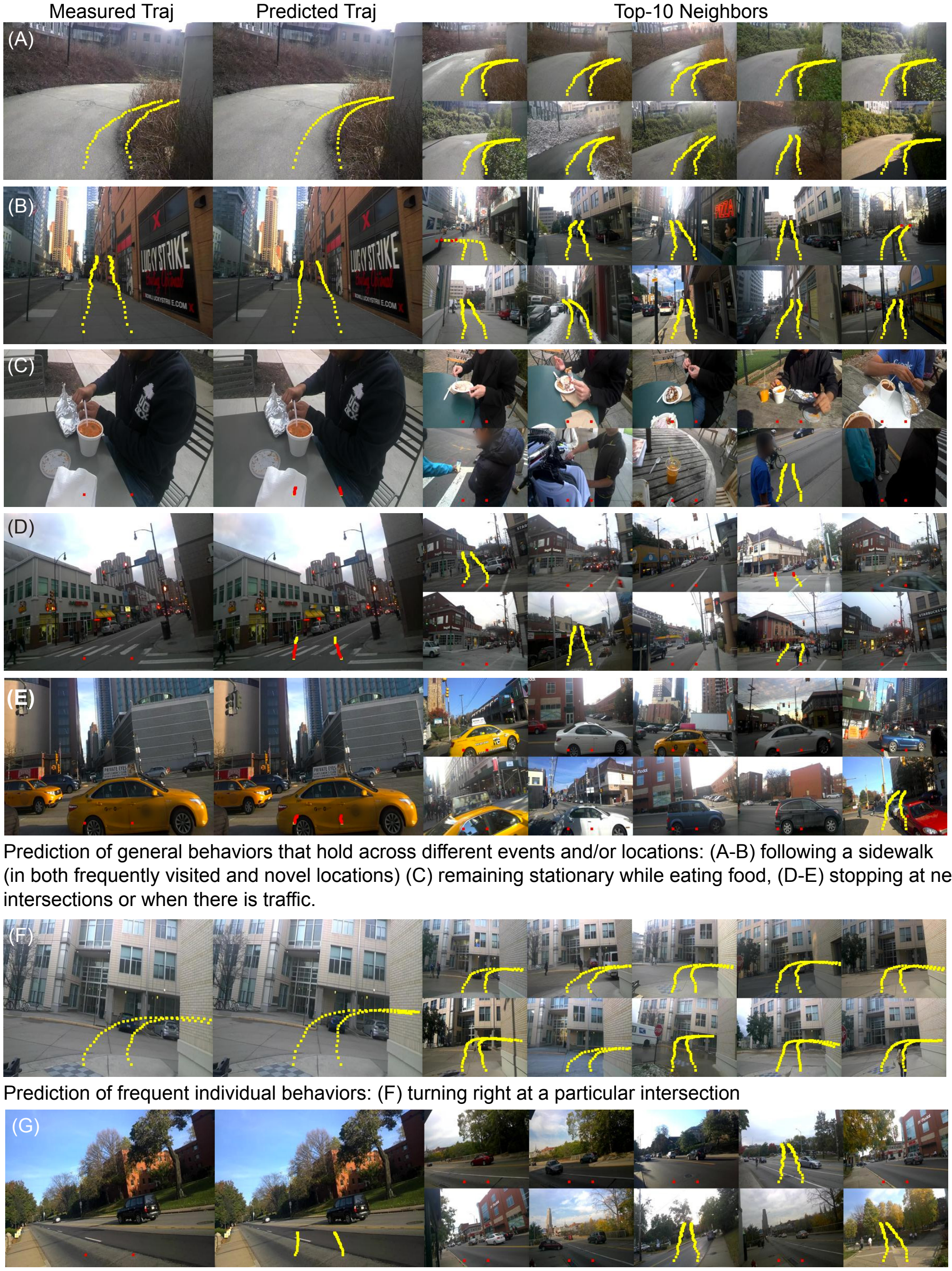


Figure 10. Top: Longer recording is needed to adequately sample rare events. Bottom: Not until four months of recording had occurred did snowing days begin to appear in the dataset, making prediction robust to seasonal change.

In addition, our dataset is sufficiently large to capture common patterns and redundancies in the student's daily life (he is a grad student, after all!), resulting in accurate predictions that are highly specific to the individual. For example, even though intersections offer the possibility for travel in several directions, the student almost always moves in the same direction at many commonly visited locations (F). Finally, some of the system's correct predictions were quite unexpected. For example, in (G) the system correctly predicts a future of standing still at a bus stop. Interestingly, the nearest neighbors that produced this prediction come from a similar views of the road at bus stops in a different part of town. Both situations feature a similarly angled view of the road from the sidewalk (likely the student looking back for a bus).

Figure 9 emphasizes the single-image prediction aspect of our chosen task by illustrating camera wearer motion predictions for images from the SUN dataset. [26]. As before, for each query image, we find its top-10 nearest neighbors in our training data and predict motion by averaging the trajectories of these neighbors. While ground truth motion is not available to validate predictions (the camera wearer was never present in these scenes), the predicted trajectories depict plausible motions.



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.

Prediction of frequent individual behaviors: (F) turning right at a particular intersection

Unexpected prediction: (G) staring at road at angle is indicative of waiting at bus stop. (Note: nearest neighbors are from bus stops in a different part of town.)

Figure 8. Examples of successful trajectory predictions.

5.3.3 Value of Extended Recording

To gain a better understanding of the relationship between dataset size and prediction accuracy, (i.e., "how much data is actually necessary?"), we repeated the reweighted motion class training from Section 5.2 with smaller training sets containing only 25% (first 10 hours) with 50% (first 19 hours) of data. The prediction accuracy of the resulting classifiers fell from 29.4% to 14.5% and 20.8%, respectively.

Figure 10 illustrates the value of extended video recording. While only a 10-day span of recording is sufficient to confidently predict the right turn taken every day outside the student's home (top row), longer amounts of record-

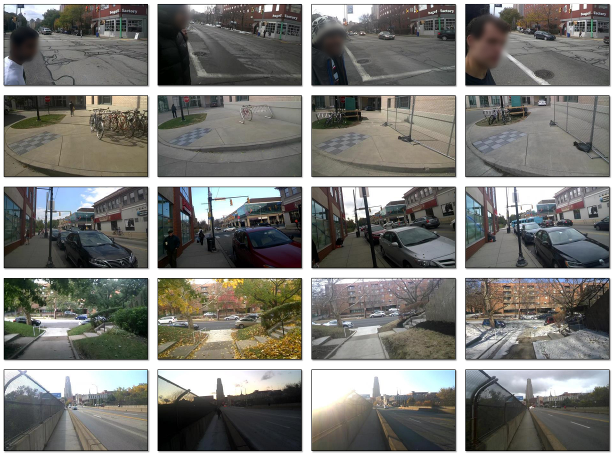


Figure 11. Although the egocentric camera is not stationary, long-term recording captures changes in a scene over time. From top to bottom: changes in companion, movement of a bicycle stand, changes in parked cars, season, and lighting.

ing are needed to predict a left turn at frequently traveled intersection (30 day span), or sitting while eating (50 day span). Figure 10-bottom shows that four months of recording were necessary for the same right turn outside the student's home to be accurately predicted on snowy days. (Prediction becomes robust to seasonal changes only after a sufficient number of snowy days exist in the training set.)

6. Virtual Webcams

Even though the egocentric camera is not stationary, many scenes appear frequently over the nine months of recording. We visualize this repetition by constructing "virtual webcams" that reveal interesting aspects of the time evolution of these scenes. Webcams are constructed by finding nearest neighbors for dataset frames using the same method as discussed in Section 4, and then manually selecting a subset of top neighbors that are well distributed in time. Each row of Figure 11 shows a selection of frames from one such webcam. The webcams reveal the collection of friends the student walks to lunch with (row 1: different people at the same intersection), change in physical structures like the movement of a bike rack due to on-campus construction (row 2), or cars parked on the same block on different days (row 3). These webcams also depict seasonal change (row 4) and the diversity of lighting conditions throughout the day (row 5).

7. Detecting Popular Places

Our egocentric video stream also faciliates analyzes that shed light on aspects of the student's environment. For example, what are the most popular locations that the student visits? Using Dollar's pedestrian detector [25, 4], we determined that 17% of frames in the dataset contain at least one person. By correlating these pedestrian detections with GPS measurements, it is possible to use the dataset to identify popular locations. For example, the heatmaps in Figure 12 plot geographic locations where more than four persons are observed on average for all images captured near that location. The figure identifies locations on campus where students often congregate (e.g., outside the student center, at an intersection between campus and the university's largest dormitory) as well as local establishments (e.g., movie theater and popular restaurants) where lines often build up.

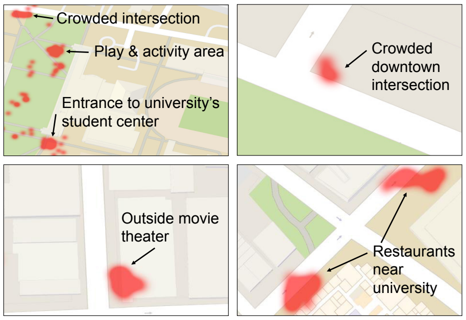

Crowded intersection
Play & activity area
Entrance to university's student center
Crowded downtown intersection
Outside movie theater
Restaurants near university

Figure 12. Red regions indicate locations where (on average) more than four people are present in images. These locations are university hangouts areas, blocks with popular restaurants and movie theaters, and busy intersections.

8. Discussion

In this work we collected a large-scale, motion annotated, egocentric video stream documenting the daily life of a single graduate student. We demonstrate that the unique size and longitudinal characteristics of the KrishnaCam dataset enable new opportunities to explore novel scene understanding tasks, such as egocentric camera motion prediction, and that the dataset enables new analyses that shed light on the nature of an individual's daily visual environment (novel data estimation, virtual webcams, popular place detection).

We observe that given enough data, nearest neighbor methods employing deep feature similarity metrics can be surprisingly effective at these tasks. We hope the dataset and these baseline results inspire future work improving upon these techniques and trying new tasks such as motion prediction based on recent video history, not only a single image (using the temporal aspect of video). We plan to continue our recording effort each day, making the data available to the community. As more data is collected, we anticipate our ability to attempt more sophisticated analyses and make more accurate predictions will continue to improve.

9. Acknowledgments

Support for this research was provided by the National Science Foundation (IIS-1422767), the Intel Corporation's Egocentric Video ISRA, and by Google.

References

[1] Memories for life. http://www.memoriesforlife.org, 2013.

[2] 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.

[3] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. Predicting daily activities from egocentric images using deep learning. *ISWC*, 2015.

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

[5] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. 2011.

[6] A. Fathi and J. M. Rehg. Modeling actions through state changes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2579–2586. IEEE, 2013.

[7] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*, pages 3281–3288. IEEE, 2011.

[8] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: A personal database for everything. *Commun. ACM*, 49(1):88–95, Jan. 2006.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[10] N. Jojic, A. Perina, and V. Murino. Structural epitome: a way to summarize one's visual experience. In *Neural Information Processing Systems*, pages 1027–1035, 2010.

[11] B. Kaneva, J. Sivic, A. Torralba, S. Avidan, and W. T. Freeman. Matching and predicting street level images. In *In Workshop for Vision on Cognitive Tasks, ECCV*, 2010.

[12] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3241–3248. IEEE, 2011.

[13] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *Computer Vision - ECCV 2012*, volume 7575 of *Lecture Notes in Computer Science*, pages 201–214. Springer Berlin Heidelberg, 2012.

[14] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pages 618–633, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[16] 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.

[17] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):978–994, May 2011.

[18] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2714–2721, June 2013.

[19] B. Morris and M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(8):1114–1127, Aug 2008.

[20] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012.

[21] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3137–3144, June 2010.

[22] G. Rogez, J. S. S. III, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4325–4333, 2015.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[24] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3302–3309, June 2014.

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

[26] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3485–3492, June 2010.

[27] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *Proceedings of the 11th European Conference on Computer Vision: Part II*, ECCV'10, pages 707–720, Berlin, Heidelberg, 2010. Springer-Verlag.

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