Learning to Recognize Daily Actions using Gaze

Alireza Fathi, Yin Li, and James M. Rehg

College of Computing
Georgia Institute of Technology

Abstract. We present a probabilistic generative model for simultaneously recognizing daily actions and predicting gaze locations in videos recorded from an egocentric camera. We focus on activities requiring eye-hand coordination and model the spatio-temporal relationship between the gaze point, the scene objects, and the action label. Our model captures the fact that the distribution of both visual features and object occurrences in the vicinity of the gaze point is correlated with the verb-object pair describing the action. It explicitly incorporates known properties of gaze behavior from the psychology literature, such as the temporal delay between fixation and manipulation events. We present an inference method that can predict the best sequence of gaze locations and the associated action label from an input sequence of images. We demonstrate improvements in action recognition rates and gaze prediction accuracy relative to state-of-the-art methods, on two new datasets that contain egocentric videos of daily activities and gaze.

1 Introduction

Ever since the pioneering experiments of Yarbus [27], it is well known that human attention and gaze are directed in a top-down task-dependent and goal-oriented manner. This is summarized in the following quote from [27]: "Eye movement reflects the human thought processes; so the observer's thought may be followed to some extent from records of eye movement." Hayhoe and Ballard [10] note that the point of fixation in the scene may not be the location which is the most visually salient, but rather will correspond to the best location given the spatio-temporal demands of the task. However, in computer vision, research on visual attention has been primarily based on bottom-up approaches [11]. Research on attention based on top-down components such as scene content, actions and objects has been very limited [28,8,2].

A basic challenge in the top-down study of gaze is that there is not always a direct relationship between actions and fixations. For example, a person can easily carry an object in her hand and put it on the table without looking at it. To address this issue, in this paper, we focus on object-manipulation tasks that require hand-eye coordination. These are actions that are hard to accomplish without using both hands and eyes in coordination. For example, when pouring a liquid into a bottle, subjects initially fixate on the mouth of the bottle, and then switch to monitoring the level of liquid in the container once they are past the

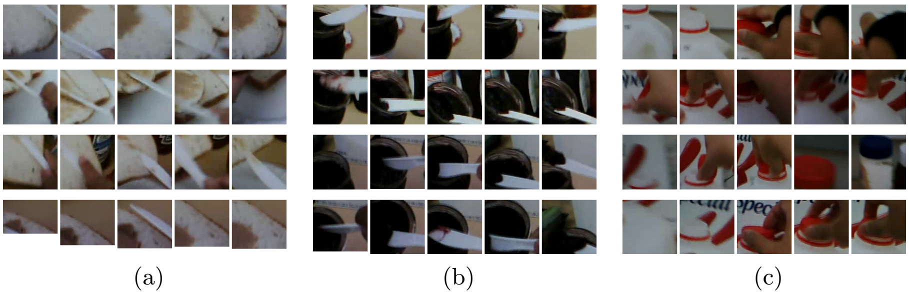


Fig. 1: Humans often attend to the location that contains the spatio-temporal information of the task. While this might not be true in some cases such as covert gaze, but in general the region around the gaze location provides significant information about the action. In the figures above, in each row we show a sequence of bounding boxes extracted around the gaze point from a particular instance of the action. For each of the action types, we show four rows of boxes, each selected from one instance of the action. The actions are (a) spread peanut-butter on bread using knife, (b) scoop jam using knife and (c) close milk.

half-way mark. In their classic study [14], Land and Hayhoe demonstrated that during object manipulation tasks a substantial percentage of fixations (around 80%) fall upon the task-relevant objects.

As an illustration of the close association between gaze and activities of daily living, Fig 1 contains small windows of pixels which have been extracted from around the gaze location. Columns correspond to frames, sampled at every two seconds. Rows correspond to different instances of a particular action. We observe that the appearance of these small windows is very consistent among instances of the same action performed by different individuals. Moreover, window contents vary significantly between actions. This observation illustrates the close relationship between eye movement, action and objects in such tasks.

Previous investigations of eye movement have largely been based on studies of static scene viewing, using gaze tracking technology affixed to a monitor screen. However, in order to study gaze in the context of object manipulation tasks, a mobile system that captures human gaze in real-life setting is required. Recently, wearable gaze tracking systems, such as [3], Tobii[1] and SMI[2], have become available. These systems combine an outward-facing camera, which captures an ego-centric or first-person view of the scene, with inward-facing gaze sensing cameras that estimate the line of sight into the scene. Calibration of the multi-camera system makes it possible to continuously measure the point of gaze within the scene in front of the user. These systems create new opportunities to exploit gaze measurements in the context of real-world tasks and naturalistic settings. In this paper, we address the question of how such gaze measurements could be useful for activity recognition in egocentric video.

[1] http://www.tobii.com/
[2] http://www.eyetracking-glasses.com/

This paper addresses the following questions:

- How consistent are the fixation patterns of different individuals performing the same action?
- Does knowing the fixation location in images of a sequence help to better recognize actions?
- Can we develop a method that can learn where to look and how to recognize actions given egocentric video with gaze measurements?

We show that action and gaze behavior are highly coordinated in daily object manipulation tasks. We show that knowing gaze location significantly improves action recognition results, and knowing the action enables more accurate prediction of gaze location. We use these observations and findings in order to learn from humans where to look for and how to recognize the daily actions in egocentric videos.

2 Previous Work

We divide the previous work into three groups: (1) daily activity recognition, (2) wearable sensors, and (3) gaze.

Daily activity recognition: Recognizing daily human activities is central to a number of different areas such as human-computer interaction, humanoid robots and elder care. The recognition of human conduct of daily object-manipulation tasks has attracted considerable attention [14,26,9,5], yet it is far from being solved. In contrast to traditional action recognition, which focuses on whole body movements, object context plays an important role in recognizing daily actions [26]. Mann et al. [17] derive force dynamic relations between objects to understand their interactions. Wu et al. [26] use RFID-tagged objects to bootstrap an appearance-based object classifier and perform activity recognition using temporal patterns of object use. Gupta et al. [9] follow a Bayesian approach using a likelihood model based on hand trajectories to analyze human-object interactions. All of these methods use static cameras mounted in the environment. However, to capture daily activities of a person, even if the office and the home are densely instrumented with cameras, the system needs to go through the non-trivial challenge of focusing on hands and objects and coping with occlusions. In contrast to these methods, in this paper we recognize daily actions from first-person point of view.

First-Person Vision: The idea of using wearable cameras is not new [22], however, recently there has been a growing interest in using them in the computer vision community, motivated by the advances in hardware technology [23,5,13,21,28,7,6,15,19]. Spriggs et al. [23] classify daily activities using a head-mounted camera and accelerometers. Pirsiavash and Ramanan [19] reocgnize activities of daily living by learning active object detectors. Yi and Ballard [28] use a wearable eye-tracking system and wearable sensors on the hands to detect the grasped and gazed object for recognizing daily actions. In contrast to [28], we develop a method that can perform action recognition both with and without observed gaze during the testing phase. In addition, we introduce a simple generative model that captures the relationship between action and gaze. Our previous method [5] for recognizing daily actions in an egocentric setting is the closest work to this paper. In that work, we use motion cues to segment hands and foreground objects and then extract features from the foreground region to recognize actions. However, our previous method fails when the object is not moving, for example when spreading peanut-butter on a slice of bread which is resting on a plate. We show that our new method presented in this paper achieves significantly better performance in comparison to [5].

Gaze: Gaze allocation models are usually derived from static picture viewing studies. This has led to methods for computation of image salience [11] which uses low-level image features such as color contrast or motion to provide a good explanation of how humans orient their attention. However, these models fail for many aspects of picture viewing [27] and natural task performance. Einhauser et al. [4] and Borji et al. [2] observe that object-level information can better predict fixation locations than low-level saliency models. Torralba et al. [24] uses global scene context features to predict the image regions fixated by humans performing natural search tasks. Judd et al. [12] show that incorporating top-down image semantics such as faces and cars improves saliency estimation in images. In this paper, we show that we can significantly enhance daily action recognition given gaze and further we show that knowing the first-person action as a prior can significantly improve gaze allocation in images. Further, we introduce a method for simultaneously inferring gaze and first-person action in egocentric videos of daily activities.

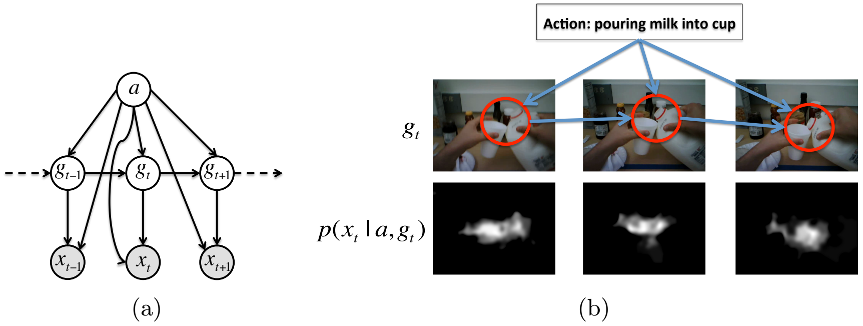
3 Method

Our algorithm estimates the action and the most likely sequence of gaze locations in an image sequence by leveraging the fact that human gaze is often focused at locations where the task is being performed. Usually the immediate surroundings of the gaze point contain most of the informative features, and other parts of image contain less relevant information.

3.1 Model

We use a generative model to describe the relationship between the egocentric action and the gaze location in each frame of an image sequence, as depicted in Fig 2(a). In this model, an action a can be inferred from the local image features x_t that are observed in the vicinity of the sequence of fixation points g_t. We have visually illustrated the concept of our model in Fig 2(b).

In our model, we have two conditional probabilities: likelihoods $p(x_t|a, g_t)$ and transitions $p(g_t|g_{t-1}, a)$. We model the probability of transition from a gaze location g_{t-1} in frame $t-1$ to gaze location g_t in frame t of an action a with a Gaussian on the distance of the two points in image coordinates. We learn a separate Gaussian model for each action.



(a) (b)

Fig. 2: In (a), we show the model for predicting the gaze location in images and action. We have visualized our model in context of a few frames in (b). The likelihood map of $p(x_t|g_t, a)$ is shown for action a set to "pour milk into cup". The brighter the pixels in images shown for $p(x_t|g_t, a)$ are, the higher the likelihood.

$$p(g_t|g_{t-1}, a) = \frac{1}{\sigma_a\sqrt{2\pi}} exp(\frac{-(\parallel g_t - g_{t-1} \parallel - \mu_a)^2}{2\sigma_a^2})$$

The mean μ_a and the variance σ_a^2 of the Gaussian models are learned separately for each action a from training data. In the following we describe our features x_t, and in Sec 3.2 we describe the procedure for computing $p(x_t|g_t, a)$. Our method uses the image content in the neighborhood of the gaze location to infer the action.

Based on our observations and experiments, we use three sets of features for each pixel location in an image: (1) features representing the set of objects around that point, (2) appearance features, and (3) features capturing if the image location belongs to an object that will be manipulated by the hands in the near future.

Object-based Features: Objects play an important role in discriminating daily actions. In an action such as "spreading peanut-butter on the bread using knife", usually it is possible to see parts of peanut, knife and bread in a local neighborhood of the gaze point. It is very uncommon to find the same pattern in an area of an image from another action. To build our object-based features, for each pixel in the image, we concatenate the maximum scores of different object classifiers in its local neighborhood to build a feature vector. We describe the details of our object detectors in Sec 5.1.

Appearance Features: Captures the appearance of the gaze location. This feature is used to determine the fixated part of the object. For example the appearance of a milk jar at its handle is different from its appearance at its mouth. In different actions, different parts of an object will be fixated. We compute the histogram of color and texture in a circular area around each pixel and use that as appearance feature.

Future Manipulation Features: This feature is based on the well known fact in the psychology literature that the gaze is usually ahead of the hands in the hand-eye coordinate system [14,18]. Eyes usually lead to another task before the hands, in order to provide additional input for planning further movements. Land and Hayhoe [14] observed that the average lead time for the tea-making task was 0.56 s and for sandwich-making was 0.9 s. As a result, hand activity in a few frames ahead provides a strong cue for predicting the gaze location in the current frame. In order to build a feature that captures whether an object is manipulated by hands in the future, we first use the method in [21] to segment each frame of the video into foreground and background regions. The foreground regions contain the hands and the manipulated objects. To verify if a pixel in frame f belongs to foreground in t frames later in video, we transfer the computed foreground mask of frame $f+t$ to frame f using the chain of optical flow vectors between adjacent frames. An example is shown in Fig 3. We do this for multiple values of t, and build a $0-1$ feature vector for each pixel location that describes if it is part of the foreground in t frames later or not.

3.2 Inference

For each action we learn a SVM classifier that fires on the pixels that are more likely to correspond to the gaze location for that particular action, given the described set of features. To train the classifier, we select the positive features from the pixels surrounding the gaze locations in training sequences corresponding to a. We select the negative features from pixels far from the gaze point in training sequences corresponding to a and all the pixels in training sequences of other actions. A few representative results are shown in Fig 6. We learn the posterior for $p(a, g_t|x_t)$ by fitting a sigmoid function to the output of the SVM classifier learned for action a [20], similar to Lester et al. [16]. We can estimate the $p(x_t|a, g_t) \propto \frac{p(a,g_t|x_t)}{p(a,g_t)}$ from the output of SVM classifiers by assuming a uniform probability for $p(a, g_t)$.

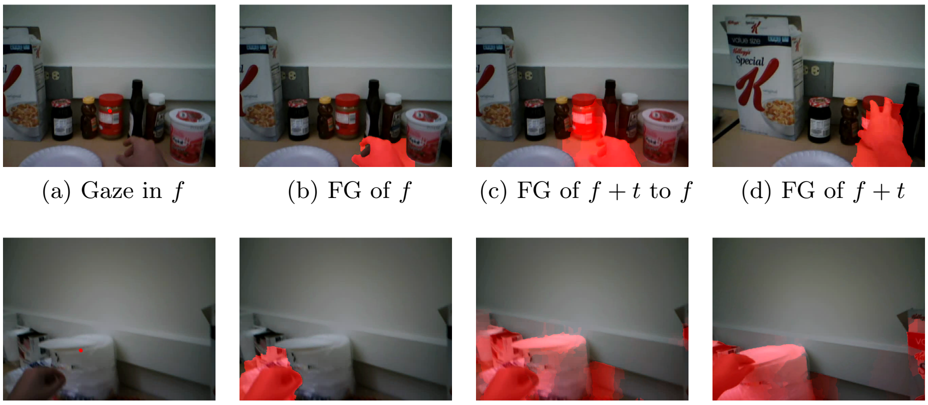
Our goal is to infer the action as well as the most likely sequence of gaze points in a test image sequence. The posterior probability of action a given the sequence of image features $X = \{x_1, ..., x_N\}$ is

$$p(a|X) \propto p(a, X) = \sum_G p(a, G, X) \approx p(a, G_a^*, X) \qquad (1)$$

Since integration over all values of G is not practical, in Eq 1 we approximate $\sum_G p(a, G, X)$ with $p(a, G_a^*, X)$, where G_a^* is the most likely sequence of gaze locations given action a. If the action a is given, the graph in Fig 2(a) becomes an HMM in which the most likely sequence of gaze locations G_a^* can be computed using the max-product (Viterbi) algorithm. Given the computed most likely sequence of gaze locations for action a, $G_a^* = \{g_1^a, ..., g_N^a\}$, we have

$$p(a|X) \propto p(a) \prod_{t=1}^{N} p(x_t|a, g_t^a) \qquad (2)$$

where we assume $p(a)$ to be a uniform distribution and $p(x_t|a, g_t^a)$ are estimated from the output of SVM classifier at location g_t^a as described above. Note



(a) Gaze in f (b) FG of f (c) FG of $f+t$ to f (d) FG of $f+t$

(e) Gaze in f (f) FG of f (g) FG of $f+t$ to f (h) FG of $f+t$

Fig. 3: This picture is best viewed in color. The gaze is usually a few frames ahead of hands. As a result, the foreground region a few frames later can provide a valuable cue for determining the gaze location in the current frame. We show two examples from initial frames of "take peanut-butter" and "take plate". The gaze falls on the object, while the hands have not reached to the object yet. In (a,e) the ground-truth gaze location in frame f of the action is shown. The computed foreground region in the frame f only contains the hand (b,f). However, when the foreground region from t frames later is transferred to this frame, it contains the gazed object (peanut-butter jar or plate) as well (c,g). The foreground region of frame $f+t$ is shown in (d,h).

that if the gaze locations were observed during the test, we could replace g_t^a in Eq 2 with observed gaze locations to compute $p(a|X)$.

4 Dataset

In this section we present two new datasets which we believe are the first of their kind. These datasets contain gaze location information associated with egocentric videos of daily activities. Our datasets are recorded from the first-person point of view and contain the subjects' gaze location in each frame of the video and are publicly available[3].

GTEA Gaze: We use the Tobii eye-tracking glasses to record this dataset. The Tobii system has an outward-facing camera that records at 30 fps rate and 480×640 pixel resolution. The glasses use an infrared inward-facing gaze sensing camera to output the 2D location of the eye gaze in each frame of the video. We setup a kitchen table with more than 30 different kinds of food and objects on it. Once each subject wore the eye-tracking glasses and the system was calibrated, we took the subject to the table, and asked them to make a meal for themselves that they can take and have if they like. We didn't put any constraints on their options. Based on the time of the day at which the subject was performing

[3] http://cpl.cc.gatech.edu/projects/GTEA_Gaze/

the meal preparation task and their personal preferences, they made different kinds of meal. The two most common meals made by the subjects were turkey sandwich and peanut-butter and jelly sandwich.

We collected 17 sequences of meal preparation activities performed by 14 different subjects. Each sequence took about 4 minutes on average. In our experiments, we use 13 sequences for training and 4 sequences for testing. We make sure that none of the sequences in the test are performed by a subject from training sequences. We annotated all the actions existing in each sequence. Each sequence contains about 30 actions on average. Each action contains a verb (for example “pour”), a set of nouns (like “milk, cup”) and a starting and an ending frame number. There exists 94 unique actions (unique combination of verbs and nouns) in our dataset. However, many of these actions only take place one or two times through out all sequences. In our experiments we prune the rare actions and only focus on the 25 remaining ones that at least take place two times in training sequences and once in testing sequences. Our set of actions contain the following verbs: take, open, close, pour, sandwich, scoop, spread.

GTEA Gaze+: We collected this dataset based on our experience in collecting the first one, in order to overcome some of its short comings. The video quality in this dataset is HD (1280×960), tasks are more organized, activities are performed in a natural setting, and the number of tasks and the number of objects used in each task are significantly bigger. The dataset is collected in Georgia Tech’s AwareHome, which is an instrumented house with a kitchen that contains all of the standard appliances and furnishings. We used SMI eye-tracking glasses to record this dataset.

We have collected data from 10 subjects, each performing a set of 7 meal preparation activities. Activities are performed based on the following food recipes: American Breakfast, Turkey Sandwich, Cheese Burger, Greek Salad, Pizza, Pasta Salad, and Afternoon Snack. Each activity (sequence) takes around 10-15 minutes on average, resulting in more than one hour of data per subject. Gaze location at each frame is recorded. We have annotated the beginning and end of different actions in each activity. Each sequence contains around 100 different actions. Actions in this dataset are associated with the following verbs: taking, putting, pouring, cutting, opening, closing, mixing, transfering, turning on/off, washing, drying, flipping, dividing, spreading, compressing, cracking, peeling, squeezing, filling, reading, moving around, distributing, draining and reading.

5 Results

In this section we present experimental results on our first dataset (GTEA Gaze). Results on the second dataset (GTEA Gaze+) can be found in the following url: `http://cpl.cc.gatech.edu/projects/GTEA_Gaze/`.

Here we first describe the details of our object detector and then we demonstrate results on our dataset that show the effectiveness of our method in gaze prediction and action recognition during daily actions.

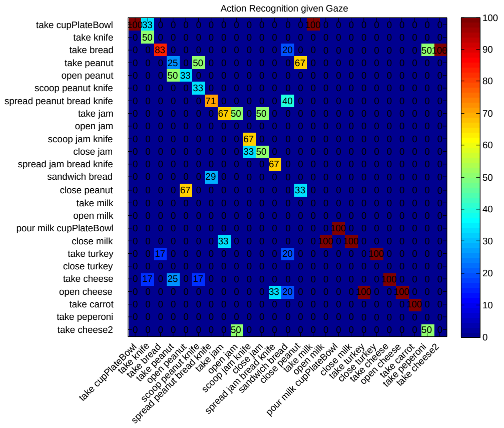


Fig. 4: This figure is best viewed in color. Confusion matrix for recognizing actions given the gaze locations in each frame. Gaze information significantly improves action recognition. The average accuracy is 47% which is significantly higher than 27% accuracy achieved by Fathi et al. [5] method. Random classification chance is 4%.

5.1 Object Detection and Segmentation

Here we describe the details of the method we use for object detection and segmentation. Our framework is not dependent on the choice of object detector and can be applied to any possible object detection and segmentation method. However, to be clear about the details of our implementation here we describe the method used in this work.

We first use [1] to extract contours and use multiple thresholds to segment each frame into layers of regions. The lowest layer contains small super-pixels. Each super-pixel is included in bigger regions in the upper levels. In order to detect and segment the objects in each image, we learn a super-pixel classifier using SVM for each object type. For each super-pixel we concatenate the color and texture histogram of its containing regions, and the color and texture histogram of multiple circles with various radiuses around its center. We compute texture descriptors using the method of [25] and quantize them to 256 kmeans centers. We further extract color descriptors for each pixel and quantize them to 128 kmeans centers. We use a few manually segmented images from training set to learn a SVM super-pixel classifier for each object type. We learn 33 object classifiers in total, including a classifier for detecting the hands. As described in Sec 3.1, we use the learned object classifiers to build the object-based feature vector that captures the object context around a potential gaze point g_t. For each pixel in image, we concatenate the maximum scores of different object classifiers in its local neighborhood to build a feature vector.

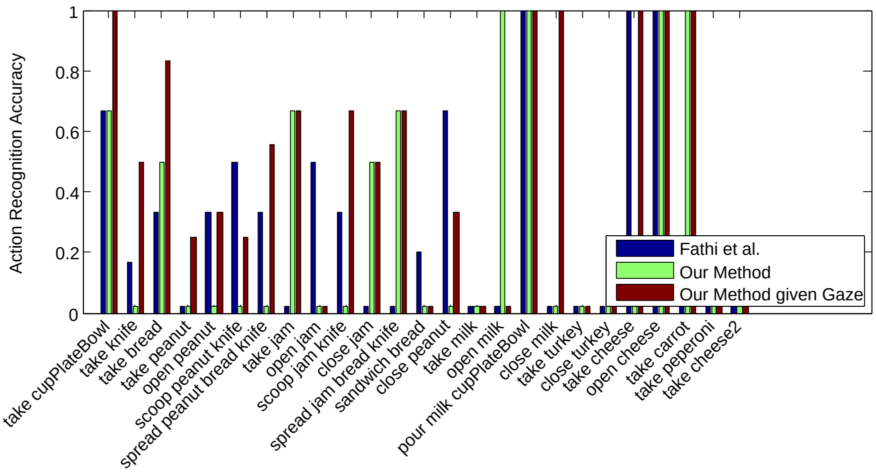


Fig. 5: The figure is best viewed in color. We compare our action recognition results with and without gaze observed during the test with results of Fathi et al. [5]. Our method with observed gaze achieves 47% average accuracy. Our method that simultaneously infers gaze and action reaches 29% accuracy. The method of Fathi et al. [5] gets 27% accuracy. The classification accuracy by chance is 4% for 25 classes.

5.2 Action given Gaze

Recognition of daily actions has its own challenges that are different than those in traditional action recognition settings. Daily actions consist of a verb and one or more object names. As a result, object context plays an important role in discriminating different actions. This makes the recognition task easier since the action verb and objects can provide context for each other [5], but at the same time the task becomes harder since miss detection of an object can result in a wrong action label. Furthermore, detection of objects in the background as part of the foreground can lead to wrong action labels. Another challenge in recognizing daily actions is that a simple action like "open peanut-butter jar" can be performed by completely different motion patterns. One might hold the jar by left hand and open it with right hand, one might leave the jar on the table and use one hand to open it, etc. Given all these variations in ways of performing an action, still the appearance of the area around the gaze point is usually consistent between different subjects performing the same action. Focusing at the neighborhood of the gaze location lets us get rid of those variations and leads to significant improvement in action recognition accuracy.

As described in Sec 3.2, for the case of observed gaze, we compute the probability of $p(a|X)$ using Eq 2 by replacing g_t^a with given gaze locations in frame t. Our method achieves 47% accuracy on action recognition compared to 27% accuracy of Fathi et al. [5]. Random classification chance for 25 classes is 4%. We show the confusion matrix for recognition of different actions in Fig 4. We compare our results to [5] in Fig 5. Fathi et al. [5] first segment the foreground

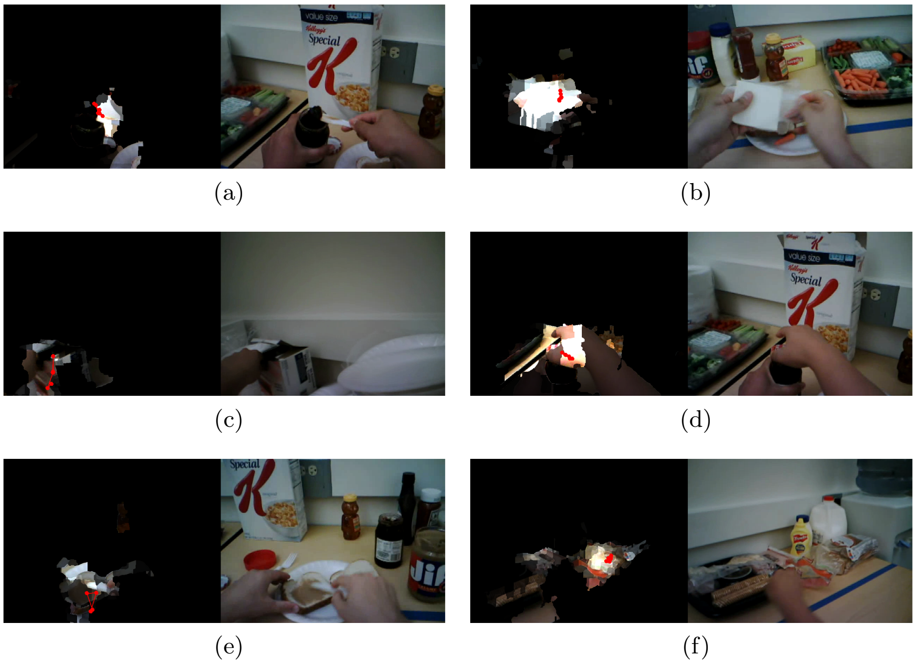

(a) (b) (c) (d) (e) (f)

Fig. 6: Our method predicts fixation locations in images for each particular action. The right hand side pictures show the frame and the left hand side images show our prediction results. The brighter the pixels are it means the higher the score returned by our algorithm is. The red dots show the ground-truth gaze locations from few adjacent frames. The actions are (a) scoop jam using knife, (b) open cheese, (c) take knife and (d) open jam, (e) spread peanut on bread using knife and (f) take bread.

from background, then use a semi-supervised learning method to detect objects, and then extract features from hands and objects to perform action recognition. To make the comparison fair, since we use pre-learned object classifiers, we provide their method with our object classifiers as well. In Sec 5.4 we show that our method of simultaneous gaze prediction and action recognition also achieves better results than [5].

5.3 Gaze given Action

The task provides a rich context for prediction of gaze location in images and video. Different subjects have a very consistent gaze pattern while performing the same action. We build a classifier that predicts human attention during performance of a particular action. We compute the likelihood of every pixel in image corresponding to gaze location by applying the classifier to feature vector extracted for that pixel location. In Fig 6 we show example outputs of our classifier. The pixels belonging to the action are scored higher than background pixels. In Fig 7, we show that our method significantly achieves better results

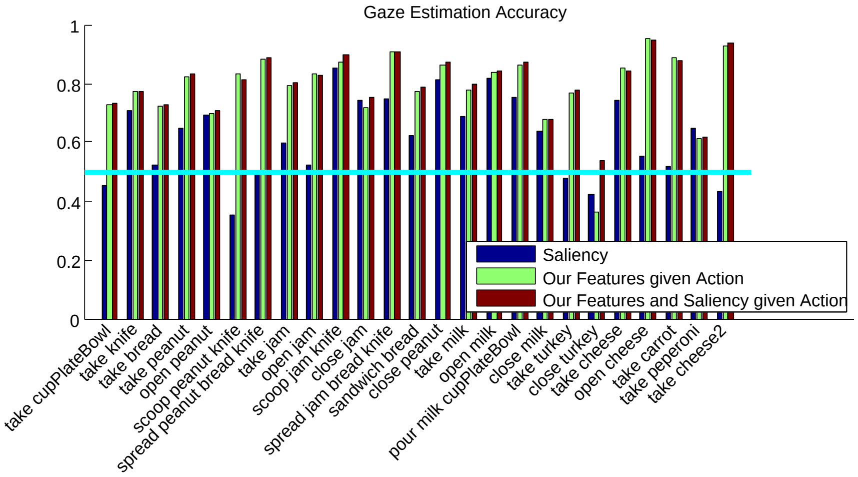


Fig. 7: This figure is best viewed in color. The task plays an important role in predicting the gaze behavior. Saliency based methods which only use low-level features are not able to capture the task related attention. Knowing the action significantly improves the results of gaze prediction. The saliency [11] at gaze location is on average higher than the saliency at 60% of the other points in image. Our classifier's score at gaze location is on average better than 81.3% of the classification scores at other image locations. Combination of low-level features used by [11] with our features only slightly improves results to 81.9%, which means the higher level action knowledge plays a more important role for predicting where humans attend. Random chance is 50% shown by the cyan line.

in comparison to general saliency methods [11] that only use low-level image features. Note that we understand that this might not be a fair comparison since our results are generated by knowing the action label for the image. The main point of our results is that (1) if the action is known, the gaze can be predicted with a good precision and (2) we show an evidence that gaze and action are closely tied together, and use this finding to justify our framework.

Each gaze prediction method in Fig 7 outputs a saliency map, in which each pixel location in the image is assigned a score. We measure the accuracy of a method by computing the percentage of the pixel scores that are lower than the pixel score of ground-truth gaze location. For example, assume the ground-truth gaze location falls at a pixel with score 0.9. If 75% of the pixels in the image are assigned scores less than 0.9, then the accuracy of the gaze prediction method for that frame is 75%. We average the accuracy over all frames belonging to the action and report them in Fig 7.

5.4 Simultaneous Inference

There are multiple reasons that motivate us to develop a method that works without having gaze data as well: (1) eye-tracking glasses are very expensive, need calibration , and still are not user friendly enough to be put on for more than a few minutes. We can learn parameters of our model from the data captured by eye-tracking glasses and then apply it to the data captured by cheap wearable cameras as well, (2) comparison of computed gaze locations with actual human data might lead to diagnosis of attention problems, measure the level of expertise and be used for human computer interaction and (3) simultaneous prediction of gaze and action demonstrates the close relationship between the two.

We use the inference method described in Sec 3.2 to recognize actions and estimate the gaze location in each image sequence. We show our results in Fig 5. Our method achieves 29% accuracy compared to the method of [5] that achieves 27%. The accuracy of random classification by chance is 4%.

6 Conclusion

We have described a novel approach to exploiting gaze measurements for action recognition in egocentric videos. Our research is motivated by the recent availability of wearable gaze tracking glasses, which make it possible to obtain continuous gaze measurements from subjects performing activities of daily living under real-world conditions. Our goal is to explore the utility of these continuous gaze measurements in solving classical vision tasks such as action recognition. We focus on classes of actions requiring hand-eye coordination which arise frequently in daily activities, such as cooking a meal, putting toothpaste on a toothbrush, etc. For such actions, we have demonstrated that the sequence of gaze fixation points within egocentric video effectively indexes the key visual properties of the image frames. We have shown that the sequence of indexed visual features is consistent across multiple users performing the same action, and is discriminative across different actions. We have introduced a generative probabilistic model for gaze behavior which combines fixation, visual features, and action labels in a simple but effective manner. We have demonstrated that our model produces more accurate predictions of gaze location and action labels than several state-of-the-art methods.

7 Acknowledgement

Portions of this research were supported in part by NSF Awards 0916687 and 1029679, ARO MURI award 58144-NS-MUR, and by the Intel Science and Technology Center in Pervasive Computing.

References

1. P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: an empirical evaluation. In *CVPR*, 2009.
2. A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *CVPR*, 2012.
3. M. Devyver, A. Tsukada, and T. Kanade. A wearable device for first person vision. In *3rd International Symposium on Quality of Life Technology*, 2011.

4. W. Einhauser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. In *Journal of Vision*, 2008.
5. A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011.
6. A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.
7. A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.
8. J.M. Findlay and I.D. Gilchrist. *Active Vision: The Psychology of Looking and Seeing*. Oxford Psychology Series. Oxford University Press, 2003.
9. A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: using spatial and functional compatibility for recognition. In *PAMI*, 2009.
10. M. Hayhoe and D. Ballard. Eye movements in natural behavior. In *TRENDS in Congnitive Sciences*, 2005.
11. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. In *PAMI*, 1998.
12. T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.
13. K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.
14. M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41:3559–3565, 2001.
15. Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.
16. J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *IJCAI*, 2005.
17. R. Mann, A. Jepson, and J. M. Siskind. Computational perception of scene dynamics. In *ECCV*, 1996.
18. J. B. Pelz and R. Consa. Oculomotor behavior and perceptual strategies in complex tasks. In *Vision Research*, 2001.
19. H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.
20. J. Platt. Probabilities for sv machines. In *Advanced in Large Margin Classifiers, MIT Press*, 1999.
21. X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010.
22. B. Schiele, N. Oliver, T. Jebara, and A. Pentland. An interactive computer vision system - dypers: dynamic personal enhanced reality system. In *ICVS*, 1999.
23. E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Egovision Workshop*, 2009.
24. A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features on object search. In *Psychological Review*, 2006.
25. M. Verma and A. Zisserman. A statistical approach to texture classification from single images. In *IJCV*, 2005.
26. J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *CVPR*, 2007.
27. A. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.
28. W. Yi and D. Ballard. Recognizing behavior in hand-eye coordination patterns. In *International Journal of Humanoid Robots*, 2009.