Learning to Predict Gaze in Egocentric Video

Yin Li, Alireza Fathi, James M. Rehg
Center for Behavior Imaging, School of Interactive Computing
Georgia Institute of Technology

Abstract

We present a model for gaze prediction in egocentric video by leveraging the implicit cues that exist in camera wearer’s behaviors. Specifically, we compute the camera wearer’s head motion and hand location from the video and combine them to estimate where the eyes look. We further model the dynamic behavior of the gaze, in particular fixations, as latent variables to improve the gaze prediction. Our gaze prediction results outperform the state-of-the-art algorithms by a large margin on publicly available egocentric vision datasets. In addition, we demonstrate that we get a significant performance boost in recognizing daily actions and segmenting foreground objects by plugging in our gaze predictions into state-of-the-art methods.

1. Introduction

With the advent of wearable cameras, such as GoPro and Google Glass, there has been an increasing interest in egocentric vision— the automatic analysis of video captured from a first-person perspective. A key component in egocentric vision is the egocentric gaze [13]. Because a person senses the visual world through a series of fixations, egocentric gaze measurements contain important cues regarding the most salient objects in the scene, and the intentions and goals of the camera-wearer. Previous works have demonstrated the utility of gaze measurements in object discovery [17] and action recognition [7].

This paper addresses the problem of egocentric gaze prediction, which is the task of predicting the user’s point-of-gaze given an egocentric video. Previous work on gaze prediction in computer vision has primarily focused on saliency detection [2]. Previous saliency models can be roughly categorized into either (1) bottom-up approaches [11] where the gaze is attracted by the discontinuities of low level features, such as color, contrast and edge; or (2) top-down approaches [24, 7, 3] where the gaze is directed by high level semantics, such as tasks, objects or scene. However, none of these approaches seem to be sufficient to predict egocentric gaze in the context of hand-eye coordination tasks. Saliency detection can be effective for visual search, but does not identify the key regions in a manipulation task. Task-driven methods can be effective, but require the identification of current activity, which is an open problem in itself. In this work we explore a third alternative: We address the question of whether measurements of head and hand movements can be used to predict gaze, without reference to saliency or activity models.

Egocentric gaze in a natural environment is the combination of gaze direction (the line of sight in a head-centered coordinate system), head orientation, and body pose. Especially during object manipulation tasks, eye, head and hand are in continual motion, and the coordination of their movements is requisite [19]. For example, large head movement is almost always accompanied by a large gaze shift [14]. Also, the gaze point tends to fall on the object that is currently being manipulated by the first person [14]. These evidences suggest that we can model the gaze of the first person by exploring the coordination of eye, hands and head, using egocentric cues alone.

The first part of our paper focuses on gaze prediction. Our major contribution is leveraging the implicit cues that are provided by the first person, such as hand location/pose, and head/hand motion, for predicting gaze. We begin with an analysis of gaze tracking data from a wearable eye tracker and demonstrate that: (1) egocentric gaze is statistically different from on-screen eye-tracking; (2) there exists a strong coordination of eye, head and hand movements in the object manipulation tasks; (3) this coordination can be used for predicting gaze in the egocentric setting. We then build a graphical model for gaze prediction that accounts for eye-hand and eye-head coordinations, and combines the temporal dynamics of gaze. The model requires no information of task or action, predicts gaze position at each frame and identifies moments of fixation. Our gaze prediction results outperform all state-of-the-art bottom-up and top-down saliency detection algorithms by a large margin on two publicly available datasets.

The second part of our paper explores applications of
gaze prediction in egocentric vision. We provide extensive experimental results on two important applications in egocentric vision: (1) foreground object segmentation and (2) egocentric action recognition. Simply by plugging in our gaze prediction, we observe a significant performance boost in comparison to several state-of-the-art methods. In object segmentation, the performance of our model is even comparable to alternative approaches that use ground-truth human gazes. We conclude that our gaze prediction model promises a great prospect for egocentric vision.

2. Related Work

2.1. Saliency Detection

Most of the bottom-up saliency models are based on the Feature Integration Theory [11]. Low level features such as color, intensity and orientation across scales are aggregated to measure the distinctiveness of a local region. Several methods also model saliency in a probabilistic framework, where the rarity of the feature defines its saliency [2]. Other approaches include spectral method [9], efficient coding [10], graphical model [8] and learning based model [12]. A recent review of saliency detection can be found in [2].

Very few computational methods have addressed the top-down saliency model. Torralba et al. [24] considered the contextual guidance of eye movement by combining low level features with high level scene semantics. Our method is related to Borji et al. [3]. They considered a driving simulation scenario where motor action is available as the top-down feature, and learned a direct mapping from both bottom-up and top-down features to fixations. Instead, we address real object manipulation tasks using egocentric vision, assume no additional information other than the video and utilize only egocentric cues for gaze prediction.

2.2. Eye, Hand and Head Coordination

Eye-hand and eye-head coordination have been studied for decades in the psychology literature. These works are mostly descriptive. Land and Hayhoe [15] studied gaze behavior in natural tasks such as tea making. They found eye fixation usually precedes hand movement by around 600ms. Pelz et al. [19] explored the temporal coordination of eye, head, and hand movements while subjects performed a simple block-copying task, where they discovered regular, rhythmic patterns of eye, head, and hand movements. These studies suggest a strong coordination of eye, hand and head movements in natural tasks. However, the coordination is variable under different situations [14].

In the computer vision community, Ba and Odobez [1] presented a model for the recognition of people’s visual focus of attention in meetings by approximating gaze direction with head orientation. Yu and Ballard [26] proposed a HMM model for action recognition based on eye-head-hand coordination in an egocentric setting by tracking gaze, head and hands using additional sensors.

2.3. Egocentric Vision

Egocentric vision is an emerging field in computer vision. The first person perspective provides a consistent viewpoint, a high quality image measurement and minimum amount of occlusion for objects. Spriggs et al. [23] addressed the segmentation and classification of activities using the first-person sensing. Fathi et al. [6] proposed to model object, action and activity jointly in egocentric vision. Other egocentric applications include object and activity detection [20] and video summarization [16]. Yamada et al. [25] combined bottom-up visual saliency with ego-motion information for egocentric gaze prediction in a walking or sitting setting. The most relevant work is Fathi et al. [7]. They presented a joint method for egocentric gaze prediction and action recognition. However, their
3. Egocentric Cues for Gaze Prediction

We focus on object manipulation tasks in a meal preparation setting, and explore the possibility of gaze prediction using egocentric cues, including hand/head movement and hand location/pose. The coordination of eye, head and hand, as we show in this section, bridges the gap between these egocentric cues and gaze prediction.

Throughout the paper, we use publicly available dataset GTEA Gaze and a subset of GTEA Gaze+ (first 15 videos of 5 subjects and 3 recipes) from [7]. Both datasets contain egocentric videos of meal preparation with gaze tracking results and action annotations. We also consider MIT eye tracking dataset [12] for comparing gaze statistics. The MIT dataset includes gaze points from 15 subjects watching 1003 images on a screen.

3.1. Eye-Head Coordination

Several psychophysical experiments have indicated that eye gaze and head pose are coupled in various tasks [19, 14, 15]. For example, large head movement is almost always accompanied by a large gaze shift. We explore the eye-head coordination in the object manipulation task using a data driven approach. The gaze statistics suggest a sharp center bias and a strong correlation between head motion and gaze shifts. These findings thus provide powerful cues for gaze prediction.

Egocentric Head Cues: In the egocentric setting, the camera is mounted on the first-person’s head, continuously capturing the scene in front of the first-person. Instead of the absolute head orientation, we model the relative orientation of the eyes with respect to the first-person’s head. This can be described by a 2D distribution of gaze over the image plane in an egocentric video. We also estimate the head movement by a global motion vector. Due to the substantial motion in egocentric videos, we apply Large Displacement Optical Flow (LDOF) [4] between two consecutive frames to get the motion field. Flows on each non-hand pixel are averaged into a 2D global motion vector. Head movements along horizontal and vertical directions are then approximated by the inverse tangent of the global motion vector divided by camera focal length. The approximation, albeit simple, provides a reasonable estimate.

Center Bias: Our first observation is a sharp center bias in egocentric gaze points. We fit a 2D Gaussian as the center prior to all gaze points in GTEA Gaze and GTEA Gaze+ dataset, respectively, as shown in Fig 2. In comparison, we also visualize the center prior as a 2D Gaussian over the MIT eye tracking dataset [12]. Egocentric gaze has a much smaller spatial variance. This is due to the fact that egocentric vision captures a first-person’s perspective in 3D world, where the gaze often aligns with the head orientation. In this case, the needs of large gaze shifts are usually compensated by head movements plus small gaze shifts. Thus, head orientation is a good approximation of gaze. Note that the preference of gaze towards the bottom part of the image is influenced by table-top object manipulation tasks.

Correlation between Gaze Shifts and Head Motion: We also observe a tight correlation between head motion and gaze shift in the horizontal direction. A scatter plot of gaze shifts (from the center) against head motion for GTEA Gaze+ dataset is shown in Fig 2b. The plot suggests a linear correlation in the horizontal direction, especially for large gaze shifts. Intuitively, when looking to the right, one tends to turn his head to the right side. This is again in consistent with the empirical finding of head eye coordination [19]. The correlation, therefore, allows us to predict gaze location from head motion.

3.2. Eye-Hand Coordination

Eye-hand coordination is the key to good performance in object manipulation tasks. Eye gaze generally guides the movement of the hands to a target [15]. Moreover, it has also been shown [21] that the proprioception of limbs may influence gaze shift, where the hands are used to guide eye movements. We introduce the concept of manipulation point. We align gaze points with respect to the first person’s hands. And we discover clusters in the aligned gaze density map, suggesting a strong eye-hand coordination. The result suggests that we can predict egocentric gaze by leveraging the first-person’s hand information.
Figure 3. Top row: Hand segmentation and manipulation points (red dots). We present four different hand configurations and the corresponding manipulation points. The hands are colored by their configurations. Bottom row: Aligned gaze density map. We align the gaze points into the hand’s coordinates by selecting the manipulation points as the origin, and projecting the gaze point into the new coordinate system every frame. We then plot the density map by averaging the aligned gaze points across all frames within the dataset. High density clusters can be found around the manipulation points, indicating spatial structures for eye-hand coordination.

Egocentric Hand Cues: Information about the hands, including their locations, poses and movements are important egocentric cues for understanding object manipulation. However, accurate tracking of hands in egocentric video is a nontrivial task. We seek to segment the hands from the video and discriminate between left/right/intersecting hands, which provides a rough hand pose estimation. We apply textonBoost [22] with CRF to segment hands in each frame. For each region, we extract spatial and shape features (centroid, orientation of major axis, eccentricity and the area of the hand masks) and train SVM to assign it to one of the three choices mentioned above. In addition, we assume there are at most two hands from the first person in a single frame. We greedily select at most two confident hand regions (with areas larger than a threshold). We also force mutual exclusiveness between region labels. For example, we can not assign a same label (single left/right hand/intersection hands) to more than one of the hand regions. And intersecting hands and single left/right hand can not show up simultaneously. Hand detection provides hand masks and configuration for each frame. We also extract hand motion by averaging optical flow vectors within the hand mask.

Manipulation Point: A major challenge for modeling eye-hand coordination is how to represent hands. Instead of tracking the hand pose, we introduce the concept of manipulation point by analyzing hand shapes at each frame. A manipulation point is defined as a control point where the first person is mostly likely to manipulate an object using his hands. For example, for a single left hand, manipulation usually happens on right tip of the hand. For two intersecting hands, the manipulation point is centered around the intersecting part. To find the manipulation point, we match the hand’s boundary to configuration-dependent templates. Examples can be found in Fig 3. A manipulation point provides an anchor with respect to the current hand pose, and allows us to align gaze points with the hand’s coordinates.

Gaze around Hands: We align the gaze points to the first-person’s hands by setting the manipulation point as the origin (Fig 3). The density maps of the aligned gaze points for four different hand configurations are plotted in Fig 3. For both GTEA Gaze and GTEA Gaze+ datasets, we observe high density around the manipulation point. The data suggest an interesting spatial relationship between manipulation points and gaze points. For single left/right hand, the gaze tends to fall on top right/top left region, where taking/putting actions might happen. For two separate hands, subjects are more likely to look in the middle, where the object is often located. For two intersecting hands, the gaze point shifts towards the bottom, partly due to opening/closing actions. These spatial distributions are consistent with the observation that people tend to look at the object they are manipulating. Thus, the combination of hand configuration and manipulation point offers a simple cue for gaze prediction.

4. Gaze Prediction in Egocentric Video

We have observed strong cues for gaze by the coordination of eye, hand and head movement. However, the flexility of the coordination makes it hard to design a hand-crafted model. Therefore, we present a learning based framework to incorporate all these egocentric cues for gaze prediction. The core of our method lies in a graphical model that combines egocentric cues at a single frame with a temporal
model of gaze shifts.

Our gaze prediction consist of two parts: predicting the gaze position at each frame and identifying the fixations among all gazes. Fixation is defined as the dwelling of the gaze within a spatially limited region (0.5-1 degree) for a minimum period of time (80-120ms) \cite{18}. The modeling of fixations captures the temporal dynamics of gaze. In this section, we discuss the egocentric features, design our model and provide our inference algorithm.

4.1. Features

We extract egocentric features regarding the first person’s hand and head cues. The feature vector \( z_t \) for frame \( t \) contains the manipulation point (2D), the global motion vector (2D), the hand motion vector (2D), the hand configuration (1D categorical). Therefore, for every frame, we get a 7 dimensional feature if hands are detected or a 2 dimensional feature if no hands are presented.

4.2. The Model

Denote the gaze point at frame \( t \) as \( g_t = [g_t^x , g_t^y]^T \in \mathbb{R}^2 \) and its binary label as \( m_t = \{0, 1\} \), where \( m_t = 1 \) denotes \( g_t \) is a fixation. Given egocentric cues \( \{z_t\} \) for all frames \( t = 1...K \), our goal is to infer the gaze points \( \{g_t\} \) and its label \( \{m_t\} = \{0, 1\} \}. We model the conditional probability \( P(\{g_t, m_t\}_{t=1}^{K}|\{z_t\}_{t=1}^{K}) \) as

\[
P(\{g_t, m_t\}_{t=1}^{K}|\{z_t\}_{t=1}^{K}) = \prod_{t=1}^{K} P(g_t|z_t) \prod_{t=1}^{K} P(m_t|g_N(t)), \tag{1}
\]

where \( g_N(t) \) are the temporal neighbors of \( g_t \). In our model, we set neighborhood to be two consecutive frames (133ms for GTEA Gaze and 80ms for GTEA Gaze+). This choice corresponds to the minimum duration of an eye fixation \cite{15, 18}. The model consists of 1) \( P(g_t|z_t) \) a single frame gaze prediction model given \( z_t \); 2) \( P(m_t|g_N(t)) \) a temporal model that couples fixation \( m_t \) and gaze prediction \( g_N(t) \). The graphical model is shown in Fig 4.

**Single Frame Gaze Prediction:** We use random regression forest for gaze prediction in a single frame. A random regression forest is an ensemble of decision trees. For each branch node, a feature is selected from a random subset of all features and a decision boundary is set by minimizing the Minimum Square Error. The leaf nodes keep the mean value of all training samples that end up in the node. And the final result is the weighted average of all leaf nodes that a testing sample reaches. We choose random forest since our feature vector \( z_t \) contains categorical data, which is easy to handle in a decision tree. We train two separate models for gaze prediction, one with both hand and head cues and one with only head cues. Our model will fall back to head motion cues if no hands are detected.

For simplicity, we train two separate regression forests for horizontal and vertical directions. The regression builds a map \( f \) between feature vector \( z_t \) and the 2D image coordinates \( \tilde{g}_t = f(z_t) \), i.e. the prediction of the gaze point at frame \( t \). The probability \( P(g_t|z_t) \) is then modeled as a Gaussian centered at \( \tilde{g}_t \) with covariance \( \Sigma_{g} \in \mathbb{R}^{2\times 2} \)

\[
P(g_t|z_t) \propto \exp(-\|g_t - \tilde{g}_t\|_2^2), \tag{2}
\]

where \( \|g_t - \tilde{g}_t\|_2^2 = (g_t - \tilde{g}_t)^T \Sigma_{g}^{-1} (g_t - \tilde{g}_t) \) is the Mahalanobis distance.

**Fixations and Gazes:** Gaze prediction and fixation detection are tightly coupled. On one hand, fixation \( m_t \) can be detected given all gaze points. On the other hand, there is a strong constraint over gaze locations if we know current gaze point is a fixation. For example, \( g_t \) should be close to \( g_{t-1} \) if \( m_t = 1 \). Therefore, we model the conditional probability \( P(m_t|g_N(t)) \) as

\[
P(m_t|g_N(t)) \propto \exp \left(-m_t \sum_{i \in N(t)} \|g_i - g_t\|_2^2\right), \tag{3}
\]

where \( m_t \) can be obtained by a fixation detection algorithm given gaze points \( g_N(t) \). Here we use a velocity-threshold based fixation detection \cite{18}: a fixation is detected if the velocities of the gaze points are below a threshold \( c \) over a minimum amount of time (two frames in our case).

\[
m_t = \prod_{i \in N(t)} \frac{-\text{sign}([g_i - g_t]_2^2 - c) + 1}{2}, \tag{4}
\]

where \( \text{sign}(x) = -1 \) if \( x < 0 \) and \( \text{sign}(x) = 1 \) if \( x \geq 0 \).

4.3. Inference and Learning

**Inference:** To infer the gaze points \( \{g_t\}_{t=1}^{K} \) and fixations \( \{m_t\}_{t=1}^{K} \), we apply Maximum Likelihood (ML) estimation to Eq (1). The minimization of the negative log likelihood function is given by

\[
\min_{\{g_t, m_t\}_{t=1}^{K}} - \log(P(\{g_t\}_{t=1}^{K}, \{m_t\}_{t=1}^{K}|\{z_t\}_{t=1}^{K}))
\]

\[
= - \log \left( \prod_{t=1}^{K} P(g_t|z_t) \prod_{t=1}^{K} P(m_t|g_N(t)) \right)
\]

\[
= \sum_{t=1}^{K} \|g_t - \tilde{g}_t\|_2^2 + \lambda \sum_{t=1}^{K-1} m_t \|g_{t+1} - g_t\|_2^2
\]

\[
s.t. \quad m_t = -\text{sign}(\|g_{t+1} - g_t\|_2^2 - c) + 1 \quad \forall t
\]
Projected gradient descent is used to obtain a local minimum of Eq (5). We first perform gradient descent over the object function assuming $m_t$ is known and ignore the constraints. $m_t$ is then updated to make all of the constraints feasible. These two steps are iterated until convergence. Intuitively, the optimization follows an EM-like updating by (1) identifying fixations $m_t$ by velocity-thresholds given all gaze predictions $g_t$ and (2) smoothing the gaze points $g_t$ given fixation labels $m_t$.

Updating $m_t$ given $g_t$ is straightforward, we estimate $m_t$ using Eq (4). Updating $g_t$ given $m_t$ is more challenging, since $g_t$ and $g_{t+1}$ are coupled together with $m_t$. Given $m_t$, we can rewrite Eq (5) using its matrix form. Let $G = [g_1 ... g_K]^T$, $\tilde{G} = [\tilde{g}_1 ... \tilde{g}_K]^T$ and $m = [m_1 ... m_K]^T$. Denote matrix $A$ as the Toeplitz matrix correspondent to the convolution kernels $[-1 \ 1]^T$. The updating of $G$ is equal to

$$\min_{G} \|G - \tilde{G}\|^2_2 + \lambda \|m^T AG\|^2_2 \quad (6)$$

The solution of Eq (6) is given by setting the first order derivative to zero

$$G^* = \left(\Sigma_s + \lambda A^T m m^T A\right)^{-1} \Sigma_s \tilde{G}. \quad (7)$$

**Learning:** We train the single frame random regression forest, using 40 trees. The parameters that need to be determined are the velocity threshold $c$, the covariance matrix $\Sigma_s$ and the constant $\lambda$. We select $c$ to be the distance of 1 degree of angular error (50/80 pixels for GTEA Gaze and GTEA Gaze+ respectively). $\Sigma_s$ defines the Mahalanobis distance between gaze points, and is learned by re-classifying training samples using random forest and re-estimating the error covariance. We empirically set $\lambda = 0.4$.

### 4.4. Gaze Prediction

We use two standard, complementary measures to assess the performance of our gaze prediction method: Area Under (ROC) Curve (AUC) and Average Angular Error (AAE). AUC measures the consistency between a predicted saliency map and the ground truth gaze points in an image, and is widely used in the saliency detection literature. AAE measures the angular distance between the predicted gaze point (e.g. the most salient point) and the groundtruth gaze, and is widely used in the gaze tracking literature. Since our method outputs a single predicted gaze point, we generate a saliency map that can be used for AUC by convolving an isotropic Gaussian over the predicted gaze.

#### 4.4.1 Results

Both the GTEA Gaze and the GTEA Gaze+ dataset contain gaze data from eye tracking glasses, which are used as ground truth for gaze prediction. We compare our method with five competing methods: a baseline center prior prediction using 2D Gaussian, three bottom-up saliency detection algorithms (Itti and Koch [11], GBVS [8], Hou et al. [10]) and one top-down saliency algorithm [7]. For all of the previous methods, we use the authors’ own implementations for benchmarking purposes. The motion cues in [11, 8] are enabled for fair comparison. One issue is that our previous method [7] requires action labels for gaze prediction. We supply it with ground truth action labels in all of our experiments. We want to emphasis that our method does not use either bottom-up image features or top-down action labels.

For GTEA Gaze dataset, we use the same training (13 videos) and testing (4 videos) split as [7] for fair comparison. For GTEA Gaze+ dataset, we perform a five-fold cross validation by leave-one-subject-out. For all our results, we average over 10 runs of random forest. We cannot compare to the results of [7] on the GTEA Gaze+ dataset, since their method requires object annotations for training, and there are no annotations currently available.

Fig. 5 shows the quantitative comparison of AUC, AAE and ROC curve in GTEA Gaze. Our method with head cues achieved AUC score of 82.3% and AAE of 10.68 degree. Adding hand cues significantly improved the AUC score (86.7%) and reduced the AAE (8.85 degree). Our temporal model added another 1% of AUC and 0.5 degree of AAE. Overall, our combined model achieved AUC score of 87.8%, where the state-of-the-art [7] gives 83.6% by using the ground truth action labels in testing. Our method also ranks highest for AAE with 8.35 degree, where the second best is 2D Gaussian (10.16 degree).

Our method works surprisingly well and outperforms a
Figure 6. Left: AUC scores and AAE for 7 different methods in GTEA Gaze+ dataset. Again, our combined model outperforms all other methods in both AUC score and AAE. Our method has less AAE than second best (2D Gaussian) for 69% of all frames. Right: ROC curves for different methods. It is interesting to find that the 2D Gaussian consistently outperforms bottom-up methods.

more complex top-down method [7] by 4.2%. Our method benefits from using the strong egocentric cues (head, hand and eye coordination) for gaze prediction and bypasses the challenging object segmentation step required by [7]. Another interesting finding is that the center prior gives better accuracy than all of the bottom-up results in AUC with a reasonable AAE. These results suggest that egocentric cues can provide a reliable gaze estimate without low-level image features or high-level task constraints.

We also tested our method in GTEA Gaze+ dataset. The results, including AUC, AAE and ROC curve are shown in Fig 6. Our final method has the best AUC of 86.7% and the best AAE of 7.93 degree, outperforming the second best (2D Gaussian) by 4.8% and 0.7 degree respectively. Using head motion already outperforms the center prior, and adding hand cues further improved the results. Again, the center prior performs better than bottom-up methods. One possible explanation for these findings is that bottom-up saliency may be an effective predictor for visual search tasks, where image features may naturally draw the viewer’s attention during a scan. However, for hand-eye coordination tasks the gaze is naturally-coordinated with the head, making the head orientation a far more useful cue. In addition, GTEA Gaze+ dataset provides ground truth labels for fixations from the eye tracking glasses. Our temporal model for fixation detection achieved 84.7% accuracy.

4.5. Object Segmentation

We further demonstrate that gaze prediction can be used to segment task-relevant foreground objects. Each video (of the 6 testing videos) is split into non-overlapping 1.5-second clips. We selected the video clips that have an action label which involves object manipulation. Two annotators were asked to select a frame within the clip and manually segment the foreground object that is involved in the action. We obtained 234 object masks from 300 video clips selected from 6 of the videos in GTEA Gaze+. More details can be found in our project webpage.

4.5.1 Results

We used both our gaze prediction results and the ground truth gaze point to seed two different methods for extracting foreground object regions: ActSeg [17] and CPMC [5]. Given the ground truth segmentation, we score the effectiveness of object segmentation under both predicted and measured gaze, thereby obtaining an alternate characterization of the effectiveness of our gaze prediction method. ActSeg [17] takes gaze points as the input and outputs one object segment per gaze point. It assumes that the gaze point always lies within the object boundary and segments the object by finding the most salient boundary. CPMC [5] uniformly samples seed points across the entire image, then generates object hypothesis and ranks them. We modified the implementation of CPMC to put the same number of dense seeds only in the vicinity of the gaze point.

We studied the relationship between foreground objects and fixation points in our dataset. We found that 82.9% (194/224) of our object annotations contain a fixation in the same frame. And 75.2% of the fixations lies within the foreground object boundary. Moreover, 94.3% of the fixations lie within the 80 pixels (1 degree) margin around the nearest foreground object boundary. These statistics suggest that human gaze tends to focus on task-relevant objects [15]. However, it is not always true that the fixation always lies within the object contour [17]. This can be explained by micro saccade [15] or gaze tracking error.

We score the segmentation results by the mean best overlapping scores, defined as the average of best overlap

(interaction over union) between a segment and the ground truth. We measure the performance of CPMC by selecting the top \( K \) candidates and varying the number \( K \). The results are reported in Fig 7(a). For ActSeg, human gaze gives 31.5\% and our gaze prediction gives 21.7\% in mean best overlapping score. For CPMC, we get equivalent performance to ActSeg from the first 6 segments, and then improve the results by 2.6\% by using predicted gaze with the first 100 segments. The performance using our gaze prediction method is comparable to that using ground truth gaze.

4.6. Action Recognition

Egocentric gaze is not only useful for foreground object segmentation, but also helps to recognize first-person’s action. We report action recognition results on GTEA Gaze dataset by substituting our predicted gaze into our previous method [7], giving the results in Fig 7(b). This method only extracts motion and appearance features from a small region around a gaze point for action recognition. In contrast to [7], using our gaze prediction improves the action recognition result to 32.8\% over the previous best result at 29\%. The upper bound of the method is given by human gaze, yielding an accuracy of 47\%. For 7 out of 25 classes, we perform better than [7]. Again, we can not report results on GTEA Gaze+ due to the lack of object annotations. We notice the large gap between our gaze prediction and real human gaze. This is left as our future work.

5. Conclusion

We described a novel approach to gaze prediction in egocentric video. Our method is motivated by the fact that in an egocentric setting, the behaviors of the first-person provide strong cues for predicting the gaze direction. We presented a model that both describes the dynamic behavior of the gaze and also reliably predicts the locations of the gazed points in video. Our gaze prediction results outperform the state-of-the-art algorithms by a large margin on GTEA Gaze and GTEA Gaze+ datasets. Finally, we demonstrate that we get a significant performance boost in recognizing daily actions and segmenting foreground objects by plugging in our gaze predictions into state-of-the-art methods.

6. 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