

Digging Deeper into Egocentric Gaze Prediction

Hamed R. Tavakoli¹, Esa Rahtu², Juho Kannala¹, and Ali Borji

¹Department of Computer Science, Aalto University

²Department of Signal Processing, Tampere University of Technology

hamed.r-tavakoli@aalto.fi, esa.rahtu@tut.fi, juho.kannala@aalto.fi, aliborji@gmail.com

Abstract

This paper digs deeper into factors that influence egocentric gaze. Instead of training deep models for this purpose in a blind manner, we propose to inspect factors that contribute to gaze guidance during daily tasks. Bottom-up saliency and optical flow are assessed versus strong spatial prior baselines. Task-specific cues such as vanishing point, manipulation point, and hand regions are analyzed as representatives of top-down information. We also look into the contribution of these factors by investigating a simple recurrent neural model for ego-centric gaze prediction. First, deep features are extracted for all input video frames. Then, a gated recurrent unit is employed to integrate information over time and to predict the next fixation. We also propose an integrated model that combines the recurrent model with several top-down and bottom-up cues. Extensive experiments over multiple datasets reveal that (1) spatial biases are strong in egocentric videos, (2) bottom-up saliency models perform poorly in predicting gaze and underperform spatial biases, (3) deep features perform better compared to traditional features, (4) as opposed to hand regions, the manipulation point is a strong influential cue for gaze prediction, (5) combining the proposed recurrent model with bottom-up cues, vanishing points and, in particular, manipulation point results in the best gaze prediction accuracy over egocentric videos, (6) the knowledge transfer works best for cases where the tasks or sequences are similar, and (7) task and activity recognition can benefit from gaze prediction. Our findings suggest that (1) there should be more emphasis on hand-object interaction and (2) the egocentric vision community should consider larger datasets including diverse stimuli and more subjects.

1. Introduction

Gaze prediction in egocentric (first person) vision, contrary to traditional gaze prediction in free-viewing setups, is an unconstrained challenging problem in which many factors including bottom-up saliency information, task spe-

cific dependencies, individual subject variables (e.g., fatigue, stress, interest) contribute. This paper evaluates various components of visual attention, including top-down and bottom-up elements, that may contribute to the prediction of gaze location in egocentric videos.

Egocentric vision considers the analysis of the visual content of daily activities from mass-marketed miniaturized wearable cameras such as cell phones, GoPro cameras, and Google glass. Analysis of images captured by egocentric cameras can reveal a lot about the person recording such images, including, intentions, personality, interests, etc. First-person gaze prediction is useful in a wide range of applications in health care, education and entertainment, for tasks such as action and event recognition [35], recognition of handled objects [37], discovering important people [16], video re-editing [26], video summarization [45], engagement detection [42], and assistive vision systems [18].

To date, in computer vision community, the study of gaze behavior has been mainly focused on understanding free-viewing gaze guidance. Consequently, while it is possible to accurately measure the gap between human inter-observer model and computational models in this task [8], our understanding of top-down and task driven gaze, despite its prevalence in daily vision [28], is relatively limited.

This paper explores bottom-up and top-down attentional cues involved in guiding first person gaze guidance. The role of spatial biases are studied using a central Gaussian map, the average fixation map, and a fixation oracle model. Further, the contribution of bottom-up saliency, vanishing point, and optical flow are investigated. as representatives of bottom-up cues. The studied task specific cues include manipulation point and hand regions. A deep model of gaze prediction is also developed as a proxy to deep models such as [49] integrating multiple cues implicitly. A set of extensive experiments is conducted to determine the contribution of each factor as well as their combination.

2. Related Works

With the increasing popularity of egocentric vision in recent years, numerous research work is being focused to solve the

computer vision problems from the first person perspective. First person vision problems are unlike classic computer vision problems since the person whose actions are being recorded is not captured. Egocentric vision poses unique challenges like non-static cameras, unusual viewpoints, motion blur, variations in illumination with the varying positions of camera wearer, real time video analysis requirements, etcetera. Tan *et al.* [43] demonstrate that challenges posed by egocentric vision can be handled in a more efficient manner if analyzed differently than exocentric vision. Substantial research has tried to address various computer vision problems such as object understanding, object detection, tracking, and activity recognition, from the egocentric perspective. We refer the readers to [6] for a recent review on the applications of first person vision.

2.1. Attention in Egocentric Vision

Yamada *et al.* [46] found that conventional saliency maps can predict egocentric fixations better than chance and that the accuracy decreases significantly with an increase in ego-motion. Matsuo *et al.* [30] proposed to combine motion and visual saliency to predict egocentric gaze. Park *et al.* [33] introduced a model to compute social saliency from head-mounted cameras to recognize gaze concurrences. Li *et al.* [29] proposed to predict gaze in egocentric activities involving meal preparation by combining implicit cues from visual features such as hand location and pose as well as head and hand motion (see also [31]). Camera motion has been shown to represent a strong cue for gaze prediction in [30]. Polatsek *et al.* [36] present a model based on spatiotemporal visual information captured from the wearer’s camera, specifically extended using a subjective function of surprise by means of motion memory. Su and Grauman [42] proposed a learning-based approach that uses long-term egomotion cues to detect user engagement during an activity (e.g., shopping, touring). Yonetani *et al.* [47] proposed a method to discover visual motifs, images of visual experiences that are significant and shared across many people, from a collection of first-person videos. Bertasius *et al.* [5] proposed the *EgoNet* network and the idea of *action-objects* to approximate momentary visual attention and motor action with objects, without gaze tracking or tactile sensors. Zhang *et al.* [49] proposed training a model for predicting the gaze on future frames. They initially generate several future frames given a latent representation of the current frame using an adversarial architecture as in video generation techniques [11]. Then, a 3D convolutional neural network is employed on the generated frames to estimate the gaze for the 50-th frame from the current frame. Recently, Huang *et al.* [23] proposed a hybrid model based on deep neural networks to integrate task-dependent attention transition with bottom-up saliency. Notice that here we do not intend to benchmark these models, rather our main goal is

to understand factors that influence gaze in daily tasks.

2.2. Top-down Visual Attention and Video Saliency

Navalpakkam and Itti [32] proposed a cognitive framework for task-driven attention using four components: 1) determining task-relevance of an entity, 2) biasing attention towards target features, 3) recognizing targets using the same low-level features, and 4) incrementally building a task-relevance map. Some models have incorporated Bayesian and reinforcement learning techniques including (e.g., [41]). Peters and Itti [34] and Borji *et al.* [9] used classification techniques such as Regression, SVM, and kNN to map a scene gist, extracted from the intermediate channels of the Itti saliency model [25], to fixations.

Some studies have investigated eye-hand coordination during grasp and object manipulation. For example, [4] studied the bidirectional sensori-motor coupling of eye-hand coordination. In a task where subjects were asked to either pretend to drink out of the presented object or to hand it over to the experimenter, they found that fixations show a clear anticipatory preference for the region where the index finger is going to be placed. Task-driven attention has also been studied in the context of joint attention during childhood as an important cue for learning (e.g., [48]). In addition, several works have studied gaze guidance during natural behavior in tasks such as sandwich making, walking, playing cricket, playing billiard, and drawing.

A tremendous amount of research has been conducted on predicting fixations over still images and videos (See [8] for a review). Traditionally, spatial saliency models have been extended to the video domain by adding a motion channel. Some models have computed video saliency in the frequency domain (e.g., [17]). Seo and Milanfar [39] utilized self similarities of spatio-temporal volumes to predict saliency. Itti and Baldi defined video saliency as Bayesian Surprise [24]. Rudoy *et al.* [38] proposed a learning-based framework for saliency prediction. A number of recent models have utilized deep learning for this purpose (e.g., [3]). For instance, Cagdas *et al.* [2], inspired by [40], proposed a two stream CNN for video saliency, one stream built on appearance and another on motion.

3. Methods

Our approach to understand the egocentric gaze guidance consists of two steps, (1) model-free evaluation to assess contribution of each cue separately and in conjunction with other cues, and 2) a model-based analysis by building computational models. For each cue, a specific computational model is developed. The computational methods are based on (1) regression from feature domain to saliency domain, (2) traditional bottom-up saliency prediction models, and (3) deep learning models. We will discuss the details of

regression and deep models of gaze prediction next. Details of feature cues will be discussed in section 5.

3.1. Regression

Here, the ground-truth fixation map is initially smoothed in order to reduce (a) the randomness of landing eye movements by viewers, and (b) eye tracking error. Then, a regressor from the feature space to fixations is learned. Assume each frame is encoded by a feature vector of size $1 \times m$. Vectors of all n frames are vertically stacked leading to the $n \times m$ matrix M . Each ground-truth fixation map has one at the location of the gaze and zeros, elsewhere (over a $k \times k$ grid map, here $k=20$). This map is first convolved with a small isotropic Gaussian (width 5, sigma 1) and is then linearized. By vertically stacking these vectors over all n frames (as above) we will have the matrix X of size $n \times k^2$. Our goal is to find vector W (of size $m \times k^2$) to minimize $\|MW - X\|_2^2$. This is a least square problem and can be solved through SVD decomposition as,

$$M \times W = X, W = M^+ \times X \quad (1)$$

where M^+ is the pseudo-inverse of matrix M (i.e., $(M^T M)^{-1} M^T$). For a test frame, we first extract feature map F and then generate the prediction map as $P = F \times W$ which is then reshaped to a $k \times k$ gaze probability map.

3.2. Deep Models

Deep Regression To investigate the power of features obtained from CNNs, we learn a regressor from frames encoded using 3 architectures, namely: Inception, ResNet and VGG16. It is worth noting that such features also encode the global context of a frame.

Gated Recurrent Units (GRUs) A deficiency of deep regression model is overlooking temporal information. Such information can be utilized using recurrent models. The input egocentric video frames are fed to a pretrained CNN (over ImageNet [14]) and then extracted features from different layers are used to train a GRU [13] to predict fixations. The task of gaze prediction at time T is to estimate the following probability,

$$p(g_1, \dots, g_T, x_1, \dots, x_T) = \prod_{t=1}^T p(g_t | g_{t < T}, x_{\leq T}) \quad (2)$$

where g_i and x_i are the 2D gaze location and feature representation of the i -th frame in the video, respectively. Given previous fixations and frames, the goal is to predict the gaze location over the current time T . Here, we assume that previous fixation data is not available so the goal is to predict the fixation location given only video frames up to the current time (i.e., offline case similar to Li *et al.* [29]). In

this case, the above joint probability distribution reduces to $\prod_{t=1}^T p(G_t | x_{\leq T})$. We utilize a GRU architecture with the following formulation:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \quad (3)$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \quad (4)$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \quad (5)$$

where x_t is the input, h_t is the output, z_t and r are the update and reset gates, respectively. W , U and b are the weights and bias to be learned. σ_g and σ_h are sigmoid and hyperbolic tangent functions, respectively.

Each video frame (RGB images with resolution 640×480) is first fed to a pretrained CNN and a feature vector x is extracted from either the fully connected layers or the final class label layer. The corresponding 2D gaze vector for each video frame is extracted and converted to a 20×20 sparse binary map with a 1 at the location of gaze and 0, elsewhere. This map is then linearized to a 400D vector and is used as the output vector in training GRU.

Network Training The proposed architecture consists of three stacked GRU units. Each GRU has 20 hidden states, and has a step size of 6. We implemented the architecture with tensorflow [1]. We train the model using cross entropy loss with softmax activation functions to discriminate between fixated locations and non-fixated ones. That is,

$$\mathcal{L}(y, x) = - \sum_i y^{(i)} \log(p(\hat{y}|x)), \quad (7)$$

where x is the input feature, y is the ground-truth indicating if there exists a fixation or not, and \hat{y} is associated with the predicted fixation. We train the model to predict which of the 400 possible locations is fixated. In other words, during the training, the ground truth data is treated as a one-hot vector defining which location is fixated at each frame. We then employed Adam optimizer [27] with learning rate of 0.0001 for 25 epochs. Although we follow a classification scheme to form a saliency map, we adopt a regression interpretation of the output of the model.

4. Data and Evaluation Criteria

4.1. Datasets

We utilize 3 datasets. The sequences consist of video game playing [9], cooking and meal preparation [15] tasks. Table 1 shows summary statistics of these datasets.

USC Video Games Data, including frames, fixations, and motor actions were collected by Borji *et al.* [9] using an infrared eye tracker while subjects played video games. We

Table 1. Summary statistics of the utilized datasets. USC videos are cartoonic outdoor while GTEA videos are natural and indoor. Hands are often visible in GTEA but not over USC videos.

Dataset	Game/ Task	Frames	Avg. Video Duration (min)	Size MB	Hands Visible?
USC	3DDS	90000	11.59 ± 0	433	N
	HDB	45000	5.59 ± 0	216	"
GTEA	Pizza	75161	10.18 ± 2.22	3630	Y
	Snack	69775	7.68 ± 0.76	2741	"
Gaze+	American	96297	13.06 ± 1.01	4651	"
GTEA Gaze	Sandwich making	35730	3.21 ± 1.28	391	Y

use data of two games. The first one called *3D Driving School (3DDS)* is a driving emulator with simulated traffic conditions. An instructor will tell the players the direction in a semi-transparent text box above the screen and/or a small arrow on the top-left corner. The second game called *Hot Dog Bush (HDB)* is a 2D time management game. Players are supposed to serve customers hot-dogs by assembling ingredients placed at different stations. Later in the game, customers can also order drinks. Players should trash burned sausages and collect the payments.

GTEA Gaze/Gaze+ Datasets We also utilize two datasets collected by Fathi *et al.* [15]. The first one is *GTEA Gaze+* dataset. We chose a subset of this dataset (as in [29]; first 15 videos of 5 subjects and 3 recipes including Pizza, American breakfast, and Afternoon snack). We report accuracies over each recipe. The second dataset, known as *GTEA Gaze*, includes 17 sequences performed by 14 different subjects. Both datasets contain videos of meal preparation (the first one includes sandwich making from ingredients on a table and the second one is cooking in a kitchen) with head-mounted gaze tracking and action annotations. All videos involve sequential object interaction.

4.2. Evaluation Criteria

We utilize two scores, Area Under the Curve (AUC) and Normalized Scanpath Saliency (NSS), to measure the consistency between observers’ fixations and models’ predictions. Please refer to [10] for detailed definitions.

5. Analysis

5.1. Spatial Biases

The spatial bias is a strong baseline in predicting the fixation locations [44]. To investigate its role in egocentric vision, we employed three baselines (1) **Central Gaussian Map (Gauss)**, (2) **Average Fixation Map (AFM)**, and (3) **Fixation Oracle Model (FOM)**. The central Gaussian model is motivated by the human tendency to look at the

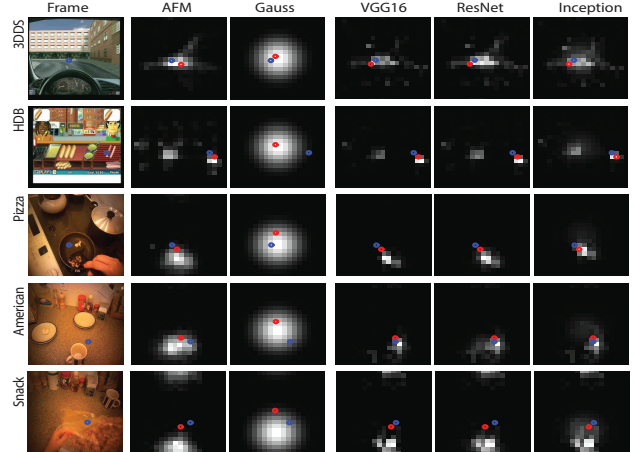


Figure 1. Sample frames of video games and egocentric videos along with predictions of recurrent deep model using various CNNs. Blue and red circles denote the ground truth fixation location and maximum of each map, respectively. Notice a sharp bias towards the bottom part of the scene over the GTEA dataset induced by table-top objects and hands.

center of images in free-viewing. The AFM is the average of all training fixations, which forms a spatial prior. The FOM is the upper-bound, obtained by convolving the ground-truth fixation maps with a Gaussian kernel. Fig. 1 visualizes sample video frames, the spatial biases and example predictions with the actual gaze point overlaid.

Table 2 summarizes the performance of spatial bias models. The AFM model is the best model and even exceeds the performance of [29]. This indicates a strong central bias in egocentric gaze estimation and is alerting for the current computational models of gaze estimation as they can be replaced by a simple Gaussian model, learned from the same sequences of data used for training such models. The FOM model scores AUC of 0.97 and NSS of 19.95 over two datasets (same on each video). This suggests that there is still a large gap between existing models and human performance in predicting egocentric gaze.

5.2. Bottom-up Saliency and Optical Flow

We computed the maps from 3 classic saliency models including *Itti* [25], *GBVS* [19], and *SR* [21]. We, then, used the saliency maps for predicting the gaze. The saliency maps were further complemented with motion features as an important source of information that influences the attention in videos. For motion features, we computed the optical flow (OF) magnitude using the Horn-Schunck algorithm [20] as a cue that captures both ego-motion and global motion. The optical flow magnitude map was then employed for predicting egocentric fixations.

We also combined the saliency maps from saliency models and optical flow to predict gaze. To this end, we trained

Table 2. The performance of spatial biases in comparison to bottom-up models and Li *et al.* [29]. The 1st row is NSS and the 2nd row is AUC. The AFM outperforms All-BU, a mixture of bottom-up models and Li *et al.* [29].

Video	AFM	Gauss	All-BU	Li <i>et al.</i> [29]
3DDS	1.588 0.796	1.705 0.814	1.112 0.768	-
HDB	2.052 0.740	0.731 0.700	0.902 0.675	
Pizza	1.682 0.885	1.467 0.829	1.064 0.767	AUC = 0.867
Snack	2.040 0.893	1.076 0.784	1.176 0.775	
American	1.986 0.888	1.164 0.803	1.101 0.774	

Table 3. Gaze prediction accuracy of BU models and their combination (1st is NSS). OF stands for optical flow magnitude. Top winner is shown in bold. GBVS does better than other models (except combination of all) due to its smoother maps.

Video	OF	Itti	GBVS	SR	All-BU
3DDS	0.229	0.834	1.067	0.160	1.112
	0.623	0.723	0.760	0.572	0.768
HDB	0.981	0.235	0.722	0.334	0.902
	0.618	0.581	0.667	0.541	0.675
Pizza	0.479	0.806	1.064	0.774	1.064
	0.675	0.719	0.764	0.740	0.767
Snack	0.486	0.872	1.157	0.816	1.176
	0.689	0.734	0.772	0.722	0.775
American	0.454	0.875	1.086	0.607	1.101
	0.702	0.737	0.774	0.711	0.774

a regressor that combines the feature maps from the three saliency models and the motion features. We will refer to this model as “all bottom-up” (All-BU) model in the rest of this paper as the representative of the bottom-up features.

Table 3 shows the results of saliency models, optical flow magnitude, and the combination of all of them. There is no single model that outperforms the combination of all models. The optical flow is not also a strong predictor alone except for sequences with highly moving objects like HDB (in HDB the scene is static, but it includes several moving objects). Nevertheless, the combination of optical flow and all other saliency models improves the results. This indicates that BU saliency models fail to capture egocentric gaze.

Table 2 shows that BU models underperform spatial biases. Even a state of the art saliency model known as SAL-ICON [22], did not perform well on this data. It achieves NSS of 0.98 and 0.81 over 3DDS and HDB games (almost as good as GBVS). These results indicate that low level saliency only weakly contributes to egocentric gaze.

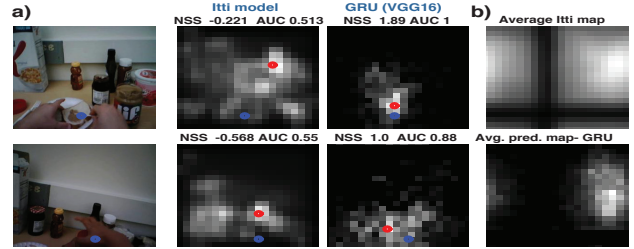


Figure 2. a) Two sample frames from the GTEA Gaze dataset along with maps from Itti and the recurrent deep model (indicated GRU), b) Average bottom-up saliency and average prediction map of our model using VGG16 features. Our model generates more focused maps. Blue and red circles denote gaze location and map maximum, respectively.

5.3. Deep Models

The performance of deep models is summarized in Table 4. The GRU model with inception features outperforms all other models over all videos in terms of AUC, except for Snack video where ResNet features perform better.

Compared to the spatial biases, in particular the AFM, deep features perform better for both architectures. The deep regression model scores well in terms of NSS. In terms of AUC score, however, the recurrent model performs the best. NSS score is a crude measure. To dig deeper into results, we calculated the percentage of video frames for which our model produces positive NSS scores. We found that on average about 75% of the frames have NSS values above zero, and approximately 69% of the frames have NSS values above one. There are only 25% of frames with negative NSS. Considering both scores, this analysis indicates that the recurrent model approximates the fixated regions better than baselines.

Table 4. Performance of deep models. NSS (1st row) and AUC (2nd row) scores of regression and recurrent model. For the ResNet and Inception CNNs, we use the class layer probabilities (a 1000D vector) to represent the video frames while for VGG16, we use the output of the last fully connected layer. The best accuracy in each row is shown in **bold**.

Video	Regression			Recurrent		
	VGG16 2048D	ResNet 1000D	Inception 1000D	VGG16 2048D	ResNet 1000D	Inception 1000D
3DDS	1.498	1.588	1.588	1.317	1.548	1.530
	0.805	0.797	0.796	0.752	0.810	0.815
HDB	2.665	2.129	2.111	1.692	1.748	1.798
	0.790	0.746	0.752	0.800	0.807	0.822
Pizza	1.387	1.720	1.640	1.650	1.748	1.696
	0.833	0.875	0.869	0.842	0.857	0.877
Snack	1.759	2.011	1.992	1.604	1.827	1.709
	0.879	0.882	0.882	0.833	0.865	0.864
American	1.412	2.045	1.884	1.702	1.717	1.984
	0.868	0.881	0.868	0.837	0.868	0.890

The deep models not only have a better performance in comparison to [29], but also are much simpler as they do not need hand detection or head motion estimation. The model by Li *et al.* [29] scores an average AUC of 0.867 over the GTEA gaze+ dataset which is below 0.877, the best average AUC of deep models. This somewhat indicates that deep models may implicitly capture some top-down factors. Fig. 2.a shows sample frames from the GTEA gaze dataset along with their corresponding maps from Itti and deep models. Itti model scores AUC of 0.749 and NSS of 1.064 on this dataset. Contrary to bottom-up saliency models, our deep models successfully highlight task-relevant regions. For example, as depicted in Fig. 2.b, the deep recurrent model predicts the gazed regions more effectively.

5.4. Task-specific cues

Here, we look into the utility of task-specific factors, including (a) *vanishing point*, (b) *manipulation point*, and (c) *hand regions*. Certain cognitive factors are believed to guide attention in specific tasks [28]. For example, drivers pay close attention to the road tangent or vanishing point while driving [28]. During making coffee attention is sequentially allocated to task-relevant entities such as coffee mug, coffee machine, and object manipulation points because hands are tightly related to objects in manipulation, reaching and grasping [4].

Vanishing points (VP) To assess whether and how much this cue can improve accuracy of gaze prediction, we ran the vanishing point detection algorithm of [7] on 3DDS frames. We chose 3DDS as it is a driving game task with vanishing points in the sequence. This algorithm outputs a 20×20 binary map with 1 at the VP location and zeros, elsewhere. We then convolved this map with a small Gaussian kernel to obtain a vanishing point map. This map scores AUC of 0.763 and NSS of 1.443 which are much higher than chance but still below spatial biases.

Hand regions To investigate whether hands predict gaze, we manually annotated hand regions over first 4 videos of the GTEA gaze dataset as depicted in Fig. 3, 6-th row. Then, we employed the binary hand mask to predict fixations in the frames with hands. NSS scores over 4 videos in order are: $-0.37, -0.28, -0.31$, and -0.22 . The negative NSS values indicate that the hand masks predict regions with no fixation because $NSS(1-S) = -NSS(S)$ where $1-S$ is the complement of map S . Thus, fixations often fall outside hand regions which means that their complement map is predictive of fixations. This indicates that hands by themselves are not informative of gaze, rather their presence in conjunction with the manipulated object is useful.

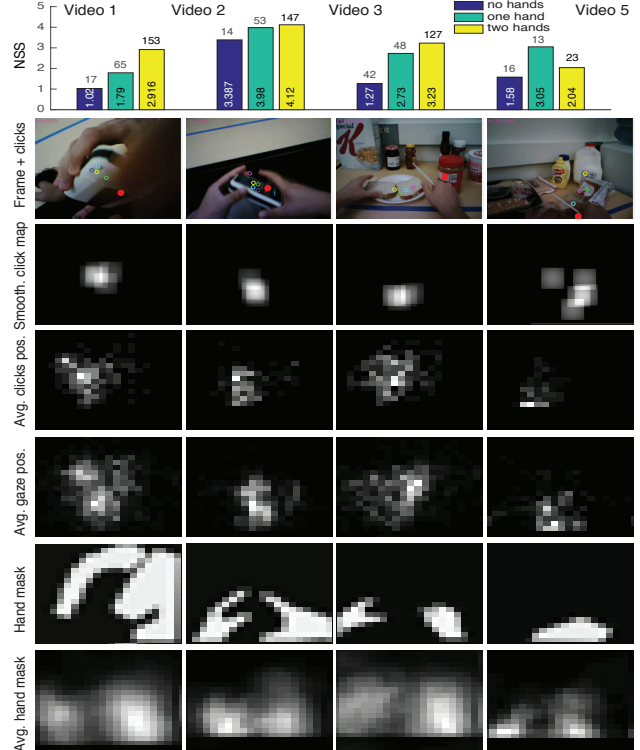


Figure 3. The role of manipulation point and hand regions on fixation prediction. Rows from top to bottom: **First:** Results of the behavioral experiment showing NSS scores for frames with no-hand, one hand and two-hands for first four videos of GTEA Gaze dataset. **Second:** Sample frames along with clicked locations by five subjects. **Third:** Smoothed click maps. **Fourth:** Average click locations over each video. **Fifth:** Ground-truth fixation locations over all frames of each video. **Sixth:** Annotated hand regions. **Seventh:** Average hand masks for each video.

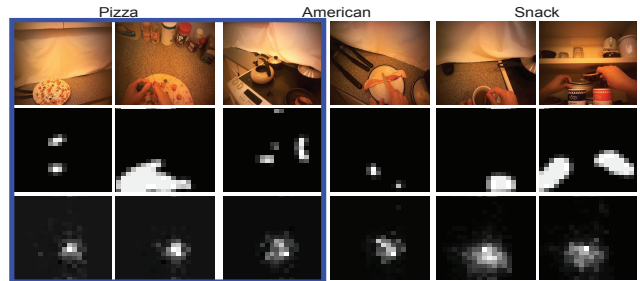


Figure 4. Sample frames from GTEA gaze+ dataset and hand segmentations using DeepLab [12] and our learned manipulation maps. Three wrong predictions are shown with the blue box.

Manipulation point (MP) To answer whether manipulation cues (during reaching, moving, or grasping objects) can improve gaze prediction, we conducted a behavioral experiment. That is, over the first 4 videos of the GTEA gaze dataset (i.e., vid1, vid2, vid3, and vid5; sampled every 10 frames), 5 subjects were asked to click where they thought the egocentric viewer has been looking while manipulating

Table 5. Accuracy of deep prediction maps augmented with manipulation point (MP) maps over the GTEA Gaze dataset. Augmenting Recurrent model (VGG16 features) with MP (4-th row) has AUCs higher than 0.878 reported for [29].

Metric	Augment	GRU (VGG16)	AFM	Gauss	MP
NSS	w/o MP	1.542	1.470	0.950	2.207
	w MP	2.293	2.310	2.171	-
AUC	w/o MP	0.856	0.836	0.748	0.802
	w MP	0.887	0.871	0.858	-

objects. The 2nd row of Fig. 3 is showing the clicked points by subjects. The average pairwise correlation among subjects, in terms of their clicked locations, over four videos in order are: 0.626, 0.678, 0.613, and 0.775. The high correlation values indicate strong agreement among subjects in predicting gaze.

We checked the agreement of manipulation map locations with fixation locations. To this end, we computed a smoothed clicked map by convolving the binary click map on each frame, made of 5 clicks, with a small Gaussian and generated a heatmap. We, then, computed the NSS over 4 videos. The results in order are: 1.90, 3.82, 2.4, and 2.22 (mean=2.6). It means that subjects are much better than chance and any of the other models in guessing fixation locations conditioned on where the hands touch the object.

We further looked into the role of hands in conjunction with manipulation point. For this purpose, we classified frames into 3 categories: *no-hands*, *one-hand*, and *two-hands* and measured NSS values over each category. The results for each video and frame category is summarized in the 1st row of Fig. 3. Mean NSS scores over 4 videos for 3 cases in order are: 1.82, 2.89, and 3.08. Results demonstrate that subjects did the best when both hands were visible. The high correlation among clicks and fixations when hands are visible indicates that hands can be strong cues for predicting where one may look.

To see if manipulation points can further contribute to models, we asked a new subject to guess gaze locations over all test frames of the GTEA dataset. We then built a manipulation point map and an MP-augmented map by adding the MP map to the prediction maps of the recurrent model. Results are shown in Table 5 using the recurrent model with VGG16 features. We find that (1) the manipulation point map does better than any of the spatial biases and the recurrent model in terms of NSS, and (2) the recurrent model augmented with manipulation map performs better than the original recurrent model. This further corroborates the fact that manipulation points are strong features for predicting where a person may look during daily tasks.

Table 6. Accuracy of the final combined model. On USC videos (3DDS and HDB), the model includes, vanishing points (only 3DDS), All-BU and recurrent deep model. On GTEA gaze+, the mean performance (Pizza, American, and Snack sequences) and the results of Li *et al.* [29] are provided for better comparison. The human upper-bounds are AUC=0.97 and NSS=19.95, showing a significant gap between human and machine.

Model	Score	3DDS	HDB	Pizza	Amer.	Snack	GTEA Gaze+
Integrated model	NSS	1.797	2.09	2.258	2.296	2.271	2.275
	AUC	0.82	0.79	0.91	0.90	0.90	0.90
Deep recurrent	NSS	1.530	1.798	1.696	1.709	1.984	1.796
	AUC	0.82	0.82	0.87	0.86	0.89	0.87
Deep regression	NSS	1.588	2.11	1.640	2.045	2.011	1.899
	AUC	0.80	0.75	0.88	0.88	0.88	0.88
All bottom-up	NSS	1.112	0.902	1.064	1.101	1.176	1.114
	AUC	0.77	0.77	0.78	0.77	0.78	0.78
Average Fixation map	NSS	1.588	2.052	1.682	1.986	2.040	1.902
	AUC	0.80	0.74	0.89	0.89	0.88	0.89
Li <i>et al.</i> [29]	AUC	—	—	—	—	—	0.87

5.5. Cue Combination

Finally, all attention maps including Recurrent model (VGG16), All-BU, vanishing points, and learned manipulation maps (MP) are combined via regression. Results are shown in Table 6. This final model does better than other models on both databases, i.e., USC database (3DDS and HDB sequences) and GTEA Gaze+ (Pizza, American, and Snack sequences). Obviously, the combination of All-Bottom-up model (combination of several bottom-up models) does not show a significantly better result as the nature of the databases are task-specific and such models lack information regarding top-down attentional cues. In comparison to Li *et al.* [29], the deep recurrent model and deep regression models achieve a similar performance, indicating that the deep features are powerful enough to achieve a performance as good as a relatively complicated probabilistic model in combination of several different feature cues.

The proposed integrated model outperforms all the models and improves over Li *et al.* [29] with accuracy of 0.90 versus 0.87. This is consistent with the fact that manipulation point is a strong predictor of gaze location in egocentric vision and potentially several top-down and task specific cues are playing a major role in egocentric gaze prediction.

5.6. Number of Subjects and Frames

To understand the effect of the number of subjects on the recurrent model, we trained the recurrent deep model from data of m subjects and tested it over the remaining subjects (using all combinations i.e., $\binom{m}{i}$, $i = 1..4$) for each video sequence. Fig. 5.a shows average scores over 5 videos of

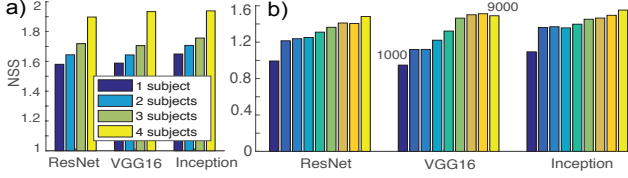


Figure 5. The effect of the number of subjects and frames

USC and GTEA Gaze+. As it is depicted, increasing the number of subjects improves the performance.

We further looked into the effect of the number of frames. To this end, we increased the number of training frames in steps of 1000 (selected from each train video; several runs) and trained a model over each video. The learned model was then applied to the whole test video. The results are summarized in Fig. 5.b. As depicted, higher number of training frames results in better accuracy.

5.7. Knowledge Transfer

To assess the generalization power of the proposed recurrent deep model, we trained the model using all the data of one video and applied it to all the data of another video. We repeated the task across databases. In other words, we trained on a sequence from USC database (3DDS and HDB sequences) and employed the model to a sequence from GTEA Gaze+ (Pizza, American, and Snack) and vice versa.

The results are summarized in Fig. 6 in terms of a confusion matrix of NSS and AUC scores. As depicted, all the results are above chance using both NSS and AUC scores. The confusion matrices show a cluster around Pizza, Snack and American video sequences (GTEA Gaze+) indicating higher similarity among them. This is not surprising as they have been following a similar task (cooking) in a similar environment (kitchen) where just a different meal is prepared.

To the contrary, the models trained on the HDB sequence generalize least to other sequences. A possible reason could be that the HDB task is significantly different than the tasks of other sequences. An example of each task in the corresponding sequences provided in Fig. 1. Further, it has fixed background, small center-bias, and no self-motion,

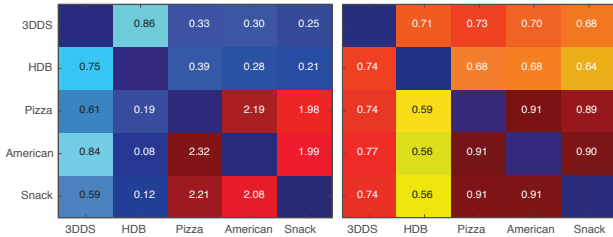


Figure 6. Confusion matrices of applying a model trained over one task to another using VGG16 features (left NSS, right AUC)

This experiment shows that it is possible to learn a model and successfully apply it to tasks that have a generic simi-

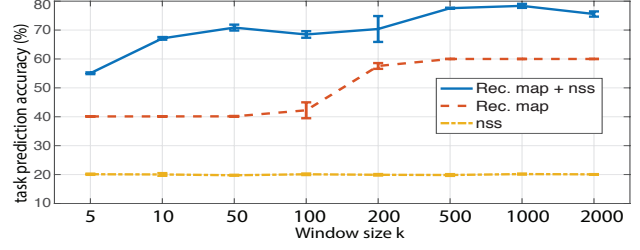


Figure 7. Task prediction over USC and GTEA Gaze+ videos

larity, e.g., cooking (American, Snack, and Pizza sequences are all involved with cooking different recipes).

5.8. Activity and Task Prediction

Facilitating task and activity recognition is one of the main motivations behind gaze prediction in egocentric vision. We, thus, investigated the use of the recurrent model for activity and task prediction. For this purpose, we chose 2000 windows of frames, each of size k , from each train video, randomly. Three types of features were computed, including, (a) *average maps of the recurrent model with VGG16 over the window*, (b) *concatenation of NSS values over frames* (a kD vector), and (c) *augmentation of a & b*. The latter is motivated by the fact that stimulus plus behavior improves the result since behavior (here gaze) offers additional information regarding the task. We trained a linear SVM to map features to video labels (1 out of 5) and plotted the performance as a function of window size. Fig. 7 shows the accuracies over 2000 test windows from each test video. The results show that decoding activity using only NSS vector is about chance. The recurrent model, however, does well specially for larger windows where the average map approaches the AFM of each video.

Augmenting the predictions with NSS values (case *c*) corresponds to the best results for activity prediction. Notice that subjects have different gaze behaviors while executing tasks. Thus, the average is obtained with respect to the prediction and each subject's NSS value. This analysis shows that gaze can be used to further improve activity and tasks recognition and has applications in egocentric vision.

6. Discussion and Conclusion

We learn the following lessons from our investigation.

a) Gaze control is based on prediction and hence inherently tied to task- and goal-relevant regions and the relative knowledge of the observer about the scene.

This is even more the case in egocentric vision, where the subject needs visual information to coordinate her manual actions in dexterous and task-effective ways.

b) The central spatial bias is due to the tendency we have to align head and eyes with our hands when manipulating something. Due to our biomechanics we also tend to use

our visual bottom hemifield more than the top one. Spatial biases are indeed completely uninformed by content, but perform well in predicting egocentric gaze.

c) Bottom up saliency is of limited relevance in strongly task-driven eye movements. Moreover considering the optical flow feature, external motion is certainly a salient feature, but in egocentric vision indeed motion is mostly self-determined and likely ignored by the subject.

d) Hands are hardly a predictor of gaze, if not for their vicinity to the objects of interest, indeed hands are barely looked at during object manipulation and are under tightly linked to fixations which in turn are under top-down control. Similarly, manipulation points are targeted when a grasp is planned and started but once the object is in the hand the gaze moves on to where the action is going to take place.

e) The deep learning model implicitly learns relevant features, which are not just the hands or the fingers but also object affordances or task-specific characteristics.

Based on these findings, we foresee several future steps, including, (1) further investigation of top-down factors, in specific manipulation point, (2) building egocentric vision databases with a diverse set of stimuli and larger number of subjects, and (3) studying the robustness of algorithms.

× Authors thank NVIDIA for the GPUs used in this work. ×

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] Ç. Bak, A. Erdem, and E. Erdem. Two-stream convolutional networks for dynamic saliency prediction. *arXiv*, 2016.
- [3] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent mixture density network for spatiotemporal visual attention. *arXiv*, 2016.
- [4] A. Belardinelli, M. Y. Stepper, and M. V. Butz. It’s in the eyes: Planning precise manual actions before execution. *Journal of vision*, 16(1):18–18, 2016.
- [5] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi. First person action-object detection with egonet. *arXiv*, 2016.
- [6] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rautenberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.
- [7] A. Borji. Vanishing point detection with convolutional neural networks. *arXiv preprint arXiv:1609.00967*, 2016.
- [8] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2013.
- [9] A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *CVPR*, pages 470–477. IEEE, 2012.
- [10] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.
- [11] V. Carl, P. Hamed, and T. Antonio. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621. NIPS, 2016.
- [12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.
- [13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.
- [15] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, pages 314–327. Springer, 2012.
- [16] J. Ghosh, Y. J. Lee, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 1346–1353. IEEE, 2012.
- [17] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, pages 1–8. IEEE, 2008.
- [18] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006.
- [19] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.
- [20] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.
- [21] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8. IEEE, 2007.
- [22] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, 2015.
- [23] Y. Huang, M. Cai, Z. Li, and Y. Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. *arXiv preprint arXiv:1803.09125*, 2018.
- [24] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *NIPS*, pages 547–554, 2005.
- [25] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 1998.
- [26] E. Jain, Y. Sheikh, A. Shamir, and J. Hodgins. Gaze-driven video re-editing. *ACM Transactions on Graphics (TOG)*, 34(2):21, 2015.
- [27] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [28] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25), 2001.
- [29] Y. Li, A. Fathi, and J. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, pages 3216–3223, 2013.

- [30] K. Matsuo, K. Yamada, S. Ueno, and S. Naito. An attention-based activity recognition for egocentric video. In *CVPR Workshops*, pages 551–556, 2014.
- [31] R. Nakashima, Y. Fang, Y. Hatori, A. Hiratani, K. Matsumiya, I. Kuriki, and S. Shioiri. Saliency-based gaze prediction based on head direction. *Vision research*, 117:59–66, 2015.
- [32] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *CVPR*, volume 2, pages 2049–2056. IEEE, 2006.
- [33] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *NIPS*, pages 431–439, 2012.
- [34] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*, pages 1–8. IEEE, 2007.
- [35] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR 2012*, pages 2847–2854. IEEE, 2012.
- [36] P. Polatsek, W. Benesova, L. Paletta, and R. Perko. Novelty-based spatiotemporal saliency detection for prediction of gaze in egocentric video.
- [37] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPR Workshops*, pages 1–8. IEEE, 2009.
- [38] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelink-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, pages 1147–1154, 2013.
- [39] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.
- [40] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.
- [41] N. Sprague and D. Ballard. Eye movements for reward maximization. In *NIPS*, page None, 2003.
- [42] Y.-C. Su and K. Grauman. Detecting engagement in egocentric video. *arXiv*, 2016.
- [43] C. Tan, H. Goh, V. Chandrasekhar, L. Li, and J. Lim. Understanding the nature of first-person videos: Characterization and classification using low-level features. 2014.
- [44] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4–4, 2007.
- [45] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*, June 2015.
- [46] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Can saliency map models predict human egocentric visual attention? In *ACCV Workshops*, pages 420–429. Springer, 2010.
- [47] R. Yonetani, K. M. Kitani, and Y. Sato. Visual motif discovery via first-person vision. In *ECCV*, pages 187–203. Springer, 2016.
- [48] C. Yu and L. B. Smith. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, 8(11):e79659, 2013.
- [49] M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *CVPR*, 2017.