In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video

Yin Li[1]*, Miao Liu[2], and James M. Rehg[2]

[1] Carnegie Mellon University
yinl2@andrew.cmu.edu
[2] College of Computing and Center for Behavioral Imaging
Georgia Institute of Technology
{mliu328, rehg}@gatech.edu

Abstract. We address the task of jointly determining what a person is doing and where they are looking based on the analysis of video captured by a headworn camera. We propose a novel deep model for joint gaze estimation and action recognition in First Person Vision. Our method describes the participant's gaze as a probabilistic variable and models its distribution using stochastic units in a deep network. We sample from these stochastic units to generate an attention map. This attention map guides the aggregation of visual features in action recognition, thereby providing coupling between gaze and action. We evaluate our method on the standard EGTEA dataset and demonstrate performance that exceeds the state-of-the-art by a significant margin of 3.5%.

1 Introduction

Therefore, "where we look" reveals important information about "what we do." Consider the examples in Fig 1, where only small regions around the first person's point of gaze are shown. What is this person doing? We can easily identify the actions as "squeeze liquid soap into hand" and "cut tomato," in spite of the fact that more than 80% of the pixels are missing. This is possible because egocentric gaze serves as an index into the critical regions of the video that define the action. Focusing on these regions eliminates the potential distraction of irrelevant background pixels, and allows us to focus on the key elements of the action.

There have been several recent works that use human gaze for action recognition [24, 31, 7]. Only our previous effort [7] attempted to model attention and action simultaneously. This paper is focused on the joint modeling of gaze estimation and action recognition in First Person Vision (FPV), where gaze, action and video are aligned in the same egocentric coordinate system. In this case, attention is naturally embodied in the camera wearer's actions. Thus, FPV provides the ideal vehicle for studying the joint modeling of attention and action.

A major challenge for the joint modeling task is the uncertainty in gaze measurements. Around 25% [11] of our gaze within daily actions are saccades—rapid

* This work was done when Y. Li was at Georgia Tech.

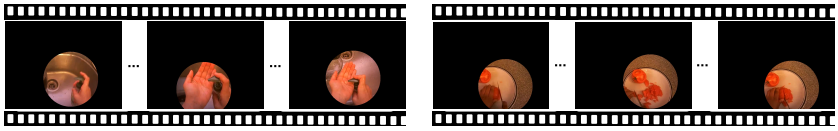

Fig. 1. *Can you tell what the person is doing?* With only 20% of the pixels visible, centered around the point of gaze, we can easily recognize the camera wearer's actions. The gaze indexes key regions containing interactions with objects. We leverage this intuition and develop a model to jointly infer gaze and actions in First Person Vision.

gaze jumps during which our vision system receives no inputs [3]. Within the gaze events that remain, it is not clear what portion of the fixations correspond to overt attention and are therefore meaningfully-connected to actions [14]. In addition, there are small but non-negligible measurement errors in the eye-tracker itself [10]. It follows that a joint model of attention and actions must account for the uncertainty of gaze. What model should we use to represent this uncertainty?

Our inspiration comes from the observation that gaze can be characterized by a *latent* distribution of attention in the context of an action, represented as an attention map in egocentric coordinates. This map identifies image regions that are salient to the current action, such as hands, objects, and surfaces. We model gaze measurements as samples from the attention map distribution. Given gaze measurements obtained during the production of actions, we can directly learn a model for the attention map, which can in turn guide action recognition. Our action recognition model can then focus on action-relevant regions to determine what the person is doing. The attention model is tightly coupled with the recognition of actions. Building on this intuition, we develop a deep network with a latent variable attention model and an attention mechanism for recognition.

To this end, we propose a novel deep model for joint gaze estimation and action recognition in FPV. Specifically, we model the latent distribution of gaze as stochastic units in a deep network. This representation allows us to sample attention maps. These maps are further used to selectively aggregate visual features in space and time for action recognition. Our model thus both encodes the uncertainty in gaze measurement, and models visual attention in the context of actions. We train the model in an end-to-end fashion using action labels and noisy gaze measurements as supervision. At testing time, our model receives only an input video and is able to infer both gaze and action.

We test our model on the EGTEA dataset—the largest public benchmark for FPV gaze and actions [19]. As a consequence of jointly modeling gaze and actions, we obtain results for action recognition that outperform state-of-the-art deep models by a significant margin (3.5%). Our gaze estimation accuracy is also comparable with strong baseline methods. To the best of our knowledge, this is the first work to model *uncertainty* in gaze measurements for action recognition, and the first deep model for *joint* gaze estimation and action recognition in FPV.

2 Related Works

First Person Vision. The advent of wearable cameras has led to growing interest in First Person Vision (FPV)—the automatic analysis of first person videos (see a recent survey in [1]). Here we focus on gaze and actions in FPV.

• **FPV Gaze**. Gaze estimation is well studied in computer vision [2]. Recent works have addressed egocentric gaze estimation. Our previous work [18] estimated egocentric gaze using hand and head cues. Zhang et al. [47] predicted future gaze by estimating gaze from predicted future frames. Park et al. [25] considered 3D social gaze from multiple camera wearers. However, these works did not model egocentric gaze in the context of actions.

• **FPV Actions**. FPV action has been the subject of many recent efforts. Spriggs et al. [37] proposed to segment and recognize daily activities using a combination of video and wearable sensor data. Kitani et al. [17] used a global motion descriptor to discover egocentric actions. Fathi et al. [6] presented a joint model of objects, actions and activities. Pirsiavash and Ramanan [27] further advocated for an object-centric representation of FPV activities. Other efforts included the modeling of conversations [5] and reactions [46] in social interactions. Several recent works have developed deep models for FPV action recognition. Ryoo et al. [30] developed a novel pooling method for deep models. Poleg et al. [28] used temporal convolutions on motion fields for long-term activity recognition. In contrast to our approach, these prior works did not consider the exploitation of egocentric gaze for action recognition.

• **FPV Gaze and Actions**. There have been a few works that incorporated egocentric gaze for FPV action recognition. For example, our previous work [19] showed the benefits of gaze-indexed visual features in a comprehensive benchmark. Both Singh et al. [35] and Ma et al. [22] explored the use of multi-stream networks to capture egocentric attention. These works have clearly demonstrated the advantage of using egocentric gaze for FPV actions. However, they all model FPV gaze and actions *separately* rather than jointly, and they do not address the uncertainty in gaze. Moreover, these methods require *side information* in addition to the input image at testing time, e.g., hand masks [35, 19] or object information [22]. In contrast, our method jointly models gaze and action, captures the uncertainty of gaze, and requires only video inputs during testing.

Our previous work [7] presented a joint model for egocentric gaze and actions. This work extends [7] in multiple aspects: (1) we propose an end-to-end deep model rather than using hand crafted features; (2) we explicitly model "noise" in gaze measurements while [7] did not; (3) we infer gaze and action jointly through a single pass during testing while [7] used iterative inference. In a nutshell, we model gaze as a stochastic variable via a novel deep architecture for joint gaze estimation and action recognition. Our model thus combines the benefits of latent variable modeling with the expressive power of a learned feature representation. Consequently, we show that our method can outperform state-of-the-art deep models [4] for FPV action recognition.

Action Recognition. There is a large body of literature on action recognition (see [41] for a survey). We discuss relevant work that targets the development of deep models and the use of attentional cues for recognizing actions.

• **Deep Models for Actions**. Deep models have demonstrated recent success for action recognition. Simonyan and Zisserman [34] proposed two-stream networks that learn to recognize an action from both optical flow and RGB frames. Wang et al. [44] extended two-stream networks to model multiple temporal segments within the video. Du et al. [40] replaced 2D convolution with spatiotemporal convolution and trained a 3D convolutional network for action recognition. Carreira and Zisserman further proposed two-stream 3D networks for action recognition [4]. A similar idea is also explored in [42]. Our model builds on the latest development of two-stream 3D convolutional networks [4] to recognize actions in FPV. Our technical novelty is to incorporate stochastic units to model egocentric gaze.

• **Attention for Actions**. Human gaze provides useful signals for the location of actions, and this intuition has been explored for action recognition in domains outside of FPV. Mathe and Sminchesescu [24] proposed to recognize actions by sampling local descriptors from a predicted saliency map. Shapovalova et al. [31] presented a method that uses human gaze for learning to localize actions. However, these methods did not use deep models. Recently, Shikhar et al. [32] incorporated soft attention into a deep recurrent network for recognizing actions. However, their notion of attention is defined by discriminative image regions that are not derived from gaze measurements, and therefore they can't support the joint inference of egocentric gaze and actions.

Our method shares a key intuition with [24, 31]: the use of predicted gaze to select visual features. However, our attention model is built within a deep network and trained from end-to-end. Our model is similar to [32] in that we also design a attention mechanism that facilitates end-to-end training. However, attention is modeled as stochastic units in our network and receives supervision from noisy human gaze measurements.

3 Method

We denote an input first person video as $x = (x^1, ..., x^t)$ with its frames x^t indexed by time t. Our goal is to predict the action category y for x. We assume egocentric gaze measurements $g = (g^1, ..., g^t)$ are available during training yet need to be inferred during testing. g^t are measured as a single 2D gaze point at time t defined on the image plane of x^t. For our model, it is helpful to reparameterize g^t as a 2D saliency map $g^t(m, n)$, where the value of the gaze position are set to one and all others are zero. And thus $\Sigma_{m,n} g^t(m, n) = 1$. In this case, $g^t(m, n)$ defines a proper probabilistic distribution of 2D gaze.

Fig 2 presents an overview of our model. We'd like to draw an analogy between our model and the well-known R-CNN framework for object detection [9, 29]. Our model takes a video x as input and outputs the distribution of gaze q as an intermediate result. We then sample the gaze map g from this pre-

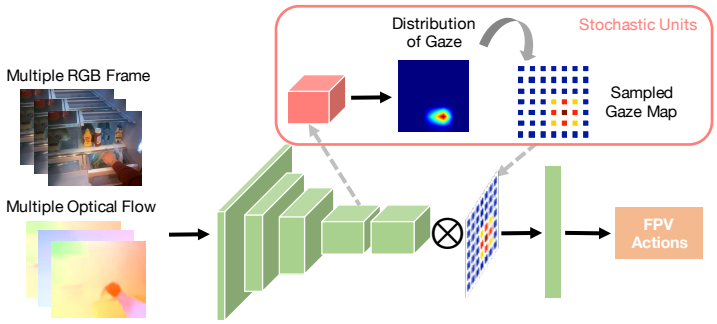

Fig. 2. Overview of our model. Our network takes multiple RGB and flow frames as inputs, and outputs a set of parameters defining a distribution of gaze in the middle layers. We then sample a gaze map from this distribution. This map is used to selectively pool visual features at higher layers of the network for action recognition. During training, our model receives action labels and noisy gaze measurement. Once trained, the model is able to infer gaze and recognize actions in FPV. We show that this network builds a probabilistic model that naturally accounts for the uncertainty of gaze and captures the relationship between gaze and actions in FPV.

dicted distribution. g encodes location information for actions and thus can be viewed as a source of action proposals—similar to the object proposals generated in R-CNN. Finally, we use the attention map to select features from the network hierarchy for recognition. This can be viewed as Region of Interest (ROI) pooling in R-CNN, where visual features in relevant regions are selected for recognition.

3.1 Modeling Gaze with Stochastic Units

Our main idea is to model $g(m, n)$ as a probabilistic variable to account for its uncertainty. More precisely, we model the conditional probability of $p(y|x)$ by

$$p(y|x) = \int_g p(y|g, x)p(g|x)dg. \tag{1}$$

Intuitively, $p(g|x)$ estimates gaze g given the input video x. $p(y|g, x)$ further uses the predicted gaze g to select visual features from input video x to predict the action y. Moreover, we want to use high capacity models, such as deep networks, for both $p(g|x)$ and $p(y|g, x)$. While this model is appealing, the learning and inference tasks are intractable for high dimensional video inputs x.

Our solution, inspired by [16, 36], is to approximate the intractable posterior $p(g|x)$ with a carefully designed $q_\pi(g|x)$. Specifically, we define $q(m, n)$ on a 2D image plane of the same size $M \times N$ as x. q is parameterized by $\pi_{m,n}$, where

$$q(m, n) = q(g_{m,n} = 1|x) = \frac{\pi_{m,n}}{\sum_{m,n} \pi_{m,n}}. \tag{2}$$

$\pi = q_\psi(X)$ is the output from a deep neural network q_ψ. $q(g|x)$ thus models the probabilistic distribution of egocentric gaze. Thus, our deep network creates a

2D map of $\pi_{m,n}$. π defines an approximation q_π to the distribution of the latent attention map. Specifically, $q(m,n)$ can be viewed as the expectation of the gaze g at position (m,n). We can then sample the gaze map $\tilde{g}$ from q_π for recognition.

Given a sampled gaze map $\tilde{g}$, our attention mechanism will selectively aggregate visual features $\phi(x)$ defined by network ϕ. In our model, this is simply a weighted average pooling, where the weights are defined by the gaze map $\tilde{g}$. We then send pooled features to the recognition network f. We further constrain f to have the form of a linear classifier, followed by a softmax function. This design is important for approximate inference. Now we have

$$p(y|g,x) = f(\Sigma_{m,n}\tilde{g}_{m,n}\phi(x)_{m,n}) = softmax\left(W_f^T(\Sigma_{m,n}\tilde{g}_{m,n}\phi(x)_{m,n})\right). \quad (3)$$

The sum operation is equivalent to spatially re-weighting individual feature channels. By doing so, we expect that the network will learn to attend to discriminative regions for action recognition. Note that this is a soft attention mechanism that allows back-propagation. Thus, top-down modulation of gaze can be achieved through gradients from action labels.

Our model thus includes three sub-networks: $q_\psi(x)$ that outputs parameters for the attention map, $\phi(x)$ that extracts visual representations for x, and $f(g,x)$ that pools features and recognizes actions. All three sub-networks share the same backbone network with their separate heads, and thus our model is realized as a single feed forward deep network. Due to the sampling process introduced in modeling, learning the parameters of the network is challenging. We overcome this challenge by using variational learning and optimizing a lower bound. We now present our training objective and inference method.

3.2 Variational Learning

During training, we make use of the input video x, its action label y and human gaze measurements g sampled from a distribution $p(g|x)$. Intuitively, our learning process has two major goals. First, our predicted gaze distribution parameterized by $q_\psi(x)$ should match the noisy observations of gaze. Second, the final recognition error should be minimized. We achieve these goals by maximizing the lower bound of $\log p(y|x)$, given by

$$\log p(y|x) \geq -\mathcal{L} = E_{g\sim q(g|x)}[\log p(y|g,x)] - KL[q(g|x)||p(g|x)], \quad (4)$$

where $KL(p||q)$ is the Kullback-Leibler (KL) divergence between distribution p and q, and E denotes the expectation.

Noise Pattern of Egocentric Gaze. Computing $KL(p||q)$ requires the prior knowledge of $p(g|x)$. In our case, given x, we observe gaze g drawn from $p(g|x)$. Thus, $p(g|x)$ is the noise pattern of the gaze measurement g. We adapt a simple noise model of gaze. For all tracked fixation points, we assume a 2D isotropic Gaussian noise, where the standard deviation of the Gaussian is selected based on the average tracking error of modern eye trackers. When the gaze point is a saccade (or is missing), we set $p(g|x)$ to the 2D uniform distribution, allowing attention to be allocated to any location on the image plane.

Loss Function. Given our noise model of gaze $p(g|x)$, we now minimize our loss function as the negative of the empirical lower bound, given by

$$-\sum_g \log p(y|g,x) + KL[q(g|x)||p(g|x)]. \tag{5}$$

During training, we sample the gaze map $\tilde{g}$ from the predicted distribution $q(g|x)$, apply the map for recognition ($p(y|\tilde{g},x) = f(\tilde{g},x)$) and compute its negative log likelihood—the same as the cross entropy loss for a categorical variable y. Our objective function thus has two terms: (a) the negative log likelihood term as the cross entropy loss between the predicated and the ground-truth action labels using the sampled gaze maps; and (b) the KL divergence between the predicted distribution $q(g|x)$ and the gaze distribution $p(g|x)$.

Reparameterization. Our model is fully differentiable except for the sampling of $\tilde{g}$. To allow end-to-end back propagation, we re-parameterize the discrete distribution $q(m,n)$ using the Gumbel-Softmax approach as in [15, 23]. Specifically, instead of sampling from $q(m,n)$ directly, we sample the gaze map $\tilde{g}$ via

$$\tilde{g}_{m,n} \sim \frac{\exp((\log \pi_{m,n} + G_{m,n})/\tau)}{\sum_{m,n} \exp((\log \pi_{m,n} + G_{m,n})/\tau)}, \tag{6}$$

where τ is the temperature that controls the "sharpness" of the distribution. We set $\tau = 2$ for all of our experiments. The softmax normalization ensures that $\sum_{m,n} \tilde{g}(m,n) = 1$, such that it is a proper gaze map. G follows the Gumbel distribution $G = -\log(-\log U)$, where U is the uniform distribution on $[0,1)$. This Concrete distribution separates out the sampling into a random variable from a uniform distribution and a set of parameters π, and thus allows the direct back-propagation of gradients to π.

3.3 Approximate Inference

During testing, we feed an input video x forward through the network to estimate the gaze distribution $q(g|x)$. Ideally, we should sample multiple gaze maps $\tilde{g}$ from q, pass them into our recognition network $f(g,x)$, and average all predictions. This is, however, prohibitively expensive. Since $f(g,x)$ is nonlinear and g has hundreds of dimensions, we will need many samples $\tilde{g}$ to approximate the expectation $E_g[f(g,x)]$, where each sample requires us to recompute $f(\tilde{g},x)$. We take a shortcut by feeding q_π into f to avoid the sampling. We note that q_π is the expectation of $\tilde{g}$, and thus our approximation is $E_g[f(g,x)] \approx f(E[g],x)$.

This shortcut does provide a good approximation. Recall that our recognition network f is a softmax linear classifier. Thus, f is convex (even with the weight decay on W_f). By Jensen's Inequality, we have $E_g[f(g,x)] \geq f(E[g],x)$. Thus, our approximation $f(E[g],x)$ is indeed a lower bound for the sample averaged estimate of $E_g[f(g,x)]$. Using this deterministic approximation during testing also eliminates the randomness in the results due to sampling. We have empirically verified the effectiveness of our approximation.

3.4 Discussions

For further insights, We connect our model to the technique of Dropout and the model of Conditional Variational AutoEncoder (CVAE).

Connection to Dropout. Our sampling procedure during learning can viewed as an alternative to Dropout [38], and thus helps to regularize the learning. In particular, we sample the gaze map $\tilde{g}$ to re-weight features. This map will have a single peak and many close-to-zero values due to the softmax function. If a position (m, n) has a very small weight, the features at that position are "dropped out". The key difference is that our sampling is guided by the predicted gaze distribution of q_ψ instead of random masking used by Dropout.

Connection to Conditional Variational Autoencoder. Our model is also connected to CVAE [36]. Both models use stochastic variables for discriminative tasks. Yet these two models are different: (1) our stochastic unit—the 2D gaze distribution, is discrete. In contrast, CVAE employs a continuous Gaussian variable, leading to a different reparameterization technique. (2) our stochastic unit—the gaze map is physically meaningful and *receives supervision* during training, while CVAE's is latent. (3) our model approximates the posterior with $q_\psi(x)$ and uses one forward pass for approximated inference, while CVAE models the posterior as a function of both x and y and thus requires recurrent updates.

3.5 Network Architecture

Our model builds on two-stream I3D networks [4]. Similar to its base Inception network [39], I3D has 5 convolutional blocks and the network uses 3D convolutions to capture the temporal dynamics of videos. Specifically, our model takes both RGB frames and optical flow as inputs, and feeds them into an RGB or a flow stream, respectively. We fuse the two streams at the end of the 4th convolutional block for gaze estimation, and at the end of the 5th convolutional block for action recognition. The fusion is done using element-wise summation as suggested by [8]. We used 3D max pooling to match the predicted gaze map to the size of the feature map at the 5th convolutional block for weighted pooling.

Our model takes the inputs of 24 frames, outputs action scores and a gaze map at a temporal stride of 8. Our output gaze map will have the spatial resolution of 7×7 (downsampled by 32x). During testing, we average the clip-level actions scores to recognizing actions in a video.

4 Experiments

We now present our experiments and results. We first introduce the dataset, the evaluation criteria and implementation details for FPV gaze estimation and action recognition. We then present our experiments on gaze and actions. Our main results are divided into three parts. First, we present an ablation study of our model. Second, we demonstrate our main results on FPV action recognition and compare our results to several state-of-the-art methods. Finally, we show results on gaze estimation and compare to a set of strong baselines. Our model achieves strong results for both action recognition and gaze estimation.

4.1 Dataset and Benchmark

Dataset. We use the Extended GTEA Gaze+ dataset.[3] This dataset contains 29 hours of first person videos from 86 unique sessions. These sessions come from 32 subjects performing 7 different meal preparation tasks in a naturalistic kitchen environment. The videos have a resolution of 1280×960 at 24Hz with gaze tracking at every frame. The dataset also comes with action annotations of 10321 instances from 106 classes with an average duration of 3.2 seconds.

EGTEA poses a challenge of fine grained action recognition in FPV. Example action categories include "Move Around pot", "Spread condiment (on) bread (using) eating utensil". Moreover, these action instances follow a long-tailed distribution. The frequent classes, such as "open fridge" have a few hundred samples and the classes on the tail, such as "crack egg" have only around 30 samples. We use the first split (8299 for training, 2022 for testing) of the dataset and evaluate the performance of gaze estimation and action recognition.

Evaluation Metric. We use standard metrics for both gaze and actions.

• **Gaze**: We consider gaze estimation as binary classification. We evaluate all fixation points and ignore untracked gaze or saccade in action clips. We report the Precision and Recall values and their corresponding F1 score.

• **Action**: We treat action recognition as multi-class classification. We report mean class accuracy at the clip level (24 frames) and at the video level.

Note that our gaze output is down-sampled both spatially (x32) and temporally (x8). When evaluating gaze, we aggregate fixation points within 8 frames and project them into a downsampled 2D map. This time interval (300ms) is equal to the duration of a fixation (around 250ms) and thus this temporal aggregation should preserve the location of gaze.

Implementation Details. We downsample all video frames to 320×256 and compute optical flow using FlowNet V2 [12]. We empirically verify that FlowNet V2 gives satisfactory motion estimation in egocentric videos. The flow map is truncated in the range of $[-20, 20]$ and rescaled to $[0, 255]$ as $[44, 34]$. During training, we randomly crop 224×224 regions from 24 frames. We then feed the RGB frames and flow maps into our networks. We also perform random horizontal flip and color jittering for data augmentation. For testing, we send the frames with a resolution of 320×256 and their flipped version. For action recognition, we average pool scores of all clips within a video. For gaze estimation, we flip back the gaze map and take the average.

Training Details. All our models are trained using SGD with momentum of 0.9 and weight decay of 0.00004. The initial weights for 3D convolutional networks are restored from Kinectcs pre-trained models [4]. For training two stream-networks, we use a batch size of 40, paralleled over 4 GPUs. We use a initial learning rate of 0.032, which matches the same learning rate from [4]. We decay the learning rate by a factor of 10 at 40th epoch and end the training at 60 epochs. We enable batch normalization [13] during training and set the decay rate for its parameters to 0.9, allowing faster aggregation of dataset statistics.

[3] Available at http://cbi.gatech.edu/fpv.

Table 1. Ablation study on backbone networks and probabilistic modeling. We show F1 scores for gaze estimation and mean class accuracy for action recognition.

Networks	Action Acc (Clip)	Action Acc (Video)
I3D RGB	43.69	47.26
I3D Flow	32.08	38.31
I3D Fusion	N/A	48.84
I3D Joint	46.42	49.79

(a) **Backbone Network**: We compare RGB, Flow, late fusion and joint training of I3D for action recognition. Joint training works the best.

Methods	Gaze F1	Action Acc
I3D Joint	N/A	49.79
Gaze MLE	24.68	51.12
Soft-Atten	10.27	50.30
Ours (Prob.)	32.97	53.30
Ours w. Dropout	32.66	52.12

(b) **Probabilistic Modeling**: We compare our model to its deterministic version (Gaze MLE). We also study the effect of Dropout.

By default, dropout with rate of 0.7 is attached for fully connected layer during training, as suggested in [44]. We disable dropout for our proposed model.

4.2 Ablation Study

We start with a comprehensive study of our model on the EGTEA Gaze dataset. Our model consists of (1) the backbone network for feature presentation; (2) the probabilistic modeling; and (3) the attention guided action recognition. We separate out these components and test them independently.

Backbone Network: RGB vs. Flow. We evaluate different network architectures on EGTEA dataset for FPV action recognition. Our goal is to understand which network performs the best in the egocentric setting. Concretely, we tested RGB and flow streams of I3D [4], the late fusion of two streams, and the joint training of two streams [8]. The results are summarized in Table 2a. Overall, EGTEA dataset is very challenging, even the strongest model has an accuracy below 50%. To help calibrate the performance, we note that the same I3D model achieved 36% on Charades [33, 45], 74% on Kinetics and 99% on UCF [4].

Unlike Kinetics or UCF, where flow stream performs comparably to RGB stream, the performance of I3D flow stream on EGTEA is significantly lower than its RGB counterpart. This is probably because of the frequent motion of the camera in FPV. It is thus more difficult to capture motion cues. Finally, the joint training of RGB and flow streams performs the best in the experiment. Thus, we choose this network as our backbone for the rest of our experiments.

Modeling: Probabilistic vs. Deterministic. We then test the probabilistic modeling part of our method. We focus on the key question: "What is the benefit of probabilistic modeling of gaze?" To this end, we present a deterministic version of our model that uses maximum likelihood estimation for gaze. We denote this model as *Gaze MLE*. Instead of sampling, this model learns to directly output a gaze map, and apply the map for recognition. During training, the gaze map is supervised by human gaze using a pixel-wise sigmoid cross entropy loss. We

keep the model architecture and the training procedure the same as our model. And we disable the loss for gaze when fixation is not available.

We compare our model with Gaze MLE for gaze and actions, and present the results in Table 2b. Our probabilistic model outperforms its deterministic version by 2.2% for action recognition and 8.3% for gaze estimation. We attribute this significant gain to the modeling. If the supervisory signal is highly noisy, allowing the network to adapt the stochasticity will facilitate the learning.

Regularization: Sampling vs. Dropout. To further test our probabilistic component, we compare our sampling of gaze map to the dropout of features. As we discussed in Sec 3.4, the sampling procedure in our model can be viewed as a way of "throwing away" features. Thus, we experiment with enabling Dropout directly after the attention pooled feature map in the model. Specifically, we compare two models with the same architecture, yet one trained with Dropout and one without. The results are in Table 2b. When Dropout is disabled, the network performs slightly better for action recognition (+1.2%) and gaze estimation (+0.3%). In contrast, removing Dropout from the backbone I3D will slightly decrease the accuracy [44]. We postulated that with regularization from our sampling, further dropping out the features will hurt the performance.

Attention for Action Recognition. Finally, we compare our method to a soft attention model (*Soft-Atten* in Table 2b) using the same backbone networks. Similar to our model, this method fuses the two streams at the end of the 4th and 5th conv blocks. Soft attention map is produced by 1x1 convolution with Sigmoid activations from the fused features at the 4th conv block. This map is further used to pool the fused features (by weighted averaging) at the 5th conv block for recognition. Thus, this soft attention map receives no supervision of gaze. A similar soft attention mechanism was used in a concurrent work [20].

For action recognition, *Soft-Atten* is worse than gaze supervised models by 0.8-3%, yet outperforms the base I3D model by 0.5%. These results suggest that (1) soft attention helps to improve action recognition even without explicit supervision of gaze; and (2) adding human gaze as supervision provides a significant performance gain. For gaze estimation, *Soft-Atten* is worse (-14%) than any gaze supervised models, as it does not receive supervision of gaze.

4.3 FPV Action Recognition

We now describe our experiments on FPV action recognition. We introduce our baselines and compare their results to our full model. These results discussed.

Baselines. We consider a set of strong baselines for FPV action recognition.

• **EgoIDT+Gaze [19]** combines egocentric features with dense trajectory descriptors [43]. These features are further selected by gaze points, and encoded using Fisher vectors [26] for action recognition.

• **I3D+Gaze** is inspired by [19, 7], where the ground truth human gaze is used to pool features from the last convolutional outputs of the network. For this method, we use the same I3D joint backbone and the same attention mechanism as our model, yet use human gaze for pooling features. When human gaze is not available, we fall back to average pooling.

Table 2. Action Recognition and Gaze Estimation. For action recognition, we report mean class accuracy at both clip and video level. For gaze estimation, we show F1 scores and their corresponding precision and recall scores.

Methods	Clip Acc	Video Acc
EgoIDT+Gaze† [19]	N/A	46.50
I3D+Gaze†	46.77	51.21
EgoConv+I3D [35]	N/A	48.93
Gaze MLE	47.41	51.12
Our Model	**47.71**	**53.30**

(a) **Action Recognition Results**: Our method outperforms previous methods by at least 3.5%, and even beats the deep model that uses human gaze at test time (I3D+Gaze) by 2.1%. †: methods use human gaze during testing.

Methods	F1	Prec	Recall
EgoGaze [18]	16.63	16.63	16.63*
Simple Gaze	30.10	25.14	37.48
Deep Gaze [47]	**33.51**	**28.04**	41.62
Gaze MLE†	24.68	18.55	36.86
Our Model†	32.97	27.01	**42.31**

(b) **Gaze Estimation Results**: Our model is comparable to the state-of-the-art methods. †: methods jointly model gaze and actions. This joint modeling does not benefit gaze estimation. *: see Sec 4.4 for discussions.

• **EgoConv+I3D [35]** adds a stream of egocentric cues for FPV action recognition. This egocentric stream encodes head motion and hand masks, and its outputs are fused with RGB and flow streams. We use Fully Convolutional Network (FCN) [21] for hand segmentation, and late fuse the score of egocentric stream with I3D for a fair comparison. This model is trained from scratch.
• **Gaze MLE** is the same model in our ablation study, where the gaze is estimated using maximum likelihood. It provides a simple baseline of multi-task learning of gaze and actions within a single deep network.

Unfortunately, we are unable to compare against relevant methods in [22, 7]. These methods require additional object annotations for training, which is not presented in EGTEA dataset. And we do want to emphasis that our method does not need object or hand information for training or testing.

Results. Our results for action recognition are shown in Table 3a. Not surprisingly, all deep models outperform EgoIDT+Gaze [19], which uses hand crafted features. Moreover, EgoConv+I3D only slightly improves the I3D late fusion results (+0.1%). This is because [35] was designed to capture actions defined by "gross body motion", such as "take" vs. "put". And our setting requires fine grained recognition of actions, e.g., "take cup" vs. "take plate".

Surprisingly, even using human gaze at test time, I3D+Gaze is only slightly better than Gaze MLE (+0.1%). And finally, our full model outperforms all baseline methods by a significant margin, including those use human gaze during testing. Our model reaches the accuracy of **53.30%**. We argue that these results provide a strong evidence to our modeling of uncertainty in gaze measurements. A model must learn to account for this uncertainty to avoid misleading gaze points, which will distract the model to action irrelevant regions.

Analysis and Discussion. Going beyond the scores, we provide more results to help understand our model. Specifically, we compare the confusion matrix of our model with backbone I3D Joint network and second best I3D+Gaze in Fig 4. Our model achieves the highest accuracy among the three for 31 out of 106

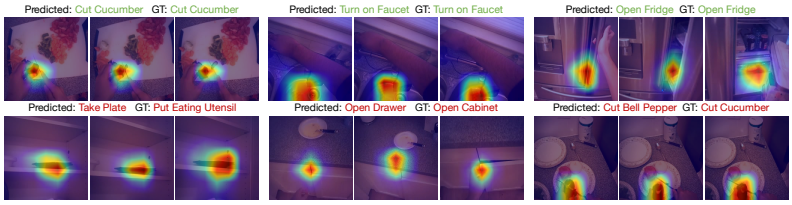

Fig. 3. Visualization of our gaze estimation and action recognition. For each 24-frame snippet, we plot the output gaze heat map at a temporal stride of 8 frames. We also print the predicted action labels and ground-truth labels above the images. Both successful (first row) and failure cases (second row) are presented.

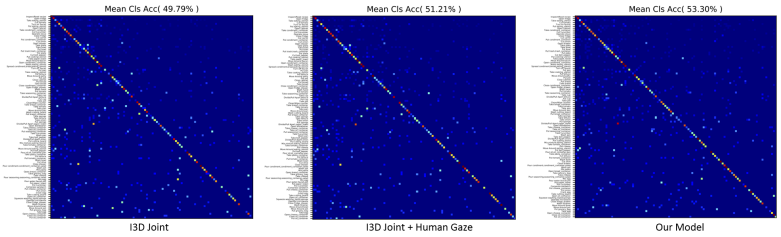

Fig. 4. Confusion matrix for action recognition. We compare our model (right) with I3D Joint (left) and I3D+Gaze (middle). Among the three methods, our method achieves the best accuracy of 31 out of 106 classes, including the challenging cases that require distinguishing actions like "turn on faucet" vs. "turn off faucet".

classes. These classes include actions where the temporal sequencing is critical, such as "take / turn on object" vs "put / turn off object". These actions have been previously considered challenging in FPV setting [19].

Finally, we visualize the outputs of gaze estimation and action labels from our model in Fig 3. Our gaze outputs often attend to foreground objects that the person is interacting with. We believe this is why the model is able to better recognize egocentric actions. Moreover, we find these visualizations helpful for diagnosing the error of action recognition. A good example is the first failure case in the second row of Fig 3, where our model outputs the action of "take plate" when the actually action has not happened yet. Another example is the last failure case in Fig 3, where the recognition model is confused due to the appearance similarity between cucumbers and bell peppers.

4.4 FPV Gaze Estimation

We now present our baselines and results for FPV gaze estimation.

Baselines. We compare our model to the following baseline methods.

• **EgoGaze [18]** makes use of hand crafted egocentric features, such as head motion and hand position, to regress gaze points. For a fair comparison, we use the FlowNet V2 for motion estimation and hand masks from FCN for hand

positions (same as our method). EgoGaze outputs a single gaze point per frame. With a single ground-truth gaze, EgoGaze will have equal numbers of false positives and false negative. Thus, its precision, recall and F1 scores are the same.
• **Simple Gaze** is a deep model inspired by our previous work [7]. Specifically, we directly estimate the gaze map using maximum likelihood (sigmoid cross entropy loss). We use the same backbone network (I3D Joint) as our model and keep the output resolution the same.
• **Deep Gaze [47]** is the FPV gaze prediction module from [47], where a 3D convolutional network is combined with a KL loss. Again, we use I3D Joint as the backbone network and keep the output resolution. Note that this model can be considered as a special case of our model by removing the sampling, the attention mechanism and the recognition network.
• **Gaze MLE** is the deterministic version of our joint model.
Results. Our gaze estimation results are shown in Table 3b. We report F1 scores and their corresponding precision and recall values. Again, deep models outperform hand crafted features by a large margin. We also observe that models with KL loss are consistently better than those use cross entropy loss. This impact is more significant for joint modeling, mostly likely due to the difficulty in balancing between the losses for two tasks. Moreover, the joint models slightly decrease the gaze estimation performance when compared to gaze-only models.
Discussion. Our results suggest that the top-down, task-relevant attention is not captured in these models, even though the top-down modulation can be achieved via back-propagation. This is thus an interesting future direction for the community to explore. Finally, we note that the benchmark of gaze estimation uses noisy human gaze as ground truth. We argue that even though these gaze measurements are noisy, they largely correlate with the underlie signal of attention. And thus the results of the benchmark are still meaningful.

5 Conclusion

We presented a novel deep model for jointly estimating gaze and recognizing actions in FPV. Our core innovation is to model the noise in human gaze measurement using stochastic units embedded in a deep neural network. Our model predicts a probabilistic representation of gaze, and uses it to select features for recognition. The method thus learns to account for the uncertainty within the supervisory signal of human gaze. We provide extensive experiments that demonstrate the effectiveness of our method. Our results surpass state-of-the-art methods for FPV action recognition and remain on par with strong baselines for gaze estimation. Going beyond FPV, our method provides a novel means of encoding the uncertainty present in training signals. We believe this capability is an important step in developing more expressive probabilistic deep models.

Acknowledgments. This research was supported in part by the Intel Science and Technology Center on Pervasive Computing (ISTC-PC).

References

1. Betancourt, A., Morerio, P., Regazzoni, C.S., Rauterberg, M.: The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology **25**(5), 744–760 (2015)
2. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 185–207 (2013)
3. Bridgeman, B., Hendry, D., Stark, L.: Failure to detect displacement of the visual world during saccadic eye movements. Vision research **15**(6), 719–722 (1975)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
5. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR (2012)
6. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV (2011)
7. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV. pp. 314–327. Springer Berlin Heidelberg (2012)
8. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
10. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(3), 478–500 (2010)
11. Henderson, J.M.: Human gaze control during real-world scene perception. Trends in cognitive sciences **7**(11), 498–504 (2003)
12. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: ICCV (2017)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
14. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience **2**(3), 194 (2001)
15. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
17. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR (2011)
18. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV (2013)
19. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: CVPR (2015)
20. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. arXiv preprint arXiv:1803.10704 (2018)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
22. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR (2016)
23. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: ICLR (2017)

24. Mathe, S., Sminchisescu, C.: Dynamic eye movement datasets and learnt saliency models for visual action recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV. pp. 842–856. Springer Berlin Heidelberg (2012)
25. Park, H.S., Jain, E., Sheikh, Y.: 3D social saliency from head-mounted cameras. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS. pp. 422–430. Curran Associates, Inc. (2012)
26. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV. pp. 143–156. Springer Berlin Heidelberg (2010)
27. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012)
28. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact CNN for indexing egocentric videos. In: WACV (2016)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) NIPS, pp. 91–99. Curran Associates, Inc. (2015)
30. Ryoo, M.S., Rothrock, B., Matthies, L.: Pooled motion features for first-person videos. In: CVPR (2015)
31. Shapovalova, N., Raptis, M., Sigal, L., Mori, G.: Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS. pp. 2409–2417. Curran Associates, Inc. (2013)
32. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. In: ICLR Workshop (2016)
33. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV. pp. 842–856. Springer International Publishing (2016)
34. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) NIPS. pp. 568–576. Curran Associates, Inc. (2014)
35. Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: CVPR (2016)
36. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) NIPS. pp. 3483–3491. Curran Associates, Inc. (2015)
37. Spriggs, E.H., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: CVPR Workshops (2009)
38. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014)
39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
40. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
41. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology **18**(11), 1473–1488 (2008)

42. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1510–1517 (2017)
43. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR (2011)
44. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV. pp. 20–36. Springer International Publishing (2016)
45. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
46. Yonetani, R., Kitani, K.M., Sato, Y.: Recognizing micro-actions and reactions from paired egocentric videos. In: CVPR (2016)
47. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: CVPR (2017)