Supplementary Materials for: "Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video"

Miao Liu[1], Siyu Tang[3], Yin Li[2], and James M. Rehg[1]

[1] Georgia Institute of Technology, Atlanta, United States
[2] University of Wisconsin-Madison, Madison, United States
[3] ETH Zürich, Switzerland

This is the supplementary material for our submission to ECCV 2020, titled "Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video". The contents are organized as follows.

A Network Architecture

Network Architecture. We present our network architecture using 3D Res50 backbone in Table 3. A similar architecture is also used for CSN-152 backbone. We followed [3] to use the features from bottom layers of the network for motor attention prediction and interaction hotspots estimation, and the features from the top layer for action anticipation. Our model jointly predicts motor attention, interaction hotspots and future actions, and thus is conceptually similar to multi-task learning e.g., [4]. The key difference is that outputs of our model depends on each other. For example, motor attention is used for interaction hotspots estimation and both motor attention and interaction hotspots are used for action anticipation.

B Mathematical Derivation for Equation 7

We present the derivation of our variational learning as discussed in Sec 3.5. Specifically, we inject posterior $p(\mathcal{A}, \mathcal{M}|x)$ into $p(y|x)$ and optimize the resulting latent variable model by maximizing the Evidence Lower Bound (ELBO). However, the prior distribution of $Q(\mathcal{A}, \mathcal{M}|x)$ is not available for training. Hence, we further approximate $p(\mathcal{A}, \mathcal{M}|x)$ by factorizing it into $p(\mathcal{A}|x)$ and $p(\mathcal{M}|x)$.

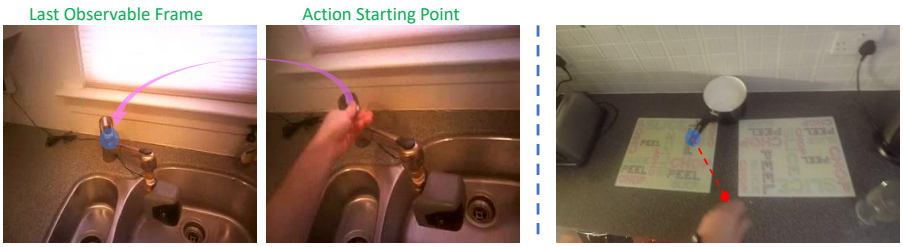

Fig. 1. (a) illustrates the interaction hotspots annotation process. (b) illustrates the approximation of the future hand trajectory on the Epic-Kitchens dataset.

Namely, we assume that $\mathcal{A}$ and $\mathcal{M}$ is conditionally independent given the input x. Thus, we have

$$\begin{aligned} &KL[p(\mathcal{A},\mathcal{M}|x)||Q(\mathcal{A},\mathcal{M}|x)] \\ =&KL[p(\mathcal{A}|\mathcal{M},x)||Q(\mathcal{A}|\mathcal{M},x)] + KL[p(\mathcal{M}|x)||Q(\mathcal{M}|x)]. \end{aligned}$$

The ELBO of our proposed joint model can be derived as

$$\begin{aligned} \log p(y|x) &\geq E_{p(\mathcal{A},\mathcal{M}|x)}[\log p(y|\mathcal{A},\mathcal{M},x)] - log(p(\mathcal{A},\mathcal{M}|x))] \\ &= \sum_{\mathcal{A},\mathcal{M}} \log p(y|\mathcal{A},\mathcal{M},x) - KL[p(\mathcal{A},\mathcal{M}|x)||Q(\mathcal{A},\mathcal{M}|x)] \\ &= \sum_{\mathcal{A},\mathcal{M}} \log p(y|\mathcal{A},\mathcal{M},x) - KL[p(\mathcal{A}|x)||Q(\mathcal{A}|x)] - KL[p(\mathcal{M}|x)||Q(\mathcal{M}|x)]. \end{aligned}$$

C Details on Data Annotation

We provide additional details on data annotation. In Sec. 4.1, we introduced how we obtain the prior distribution of motor attention and interaction htospots. Here we provide a visual illustration of our efforts on the data annotation. As shown in Fig. 1 (a), we compare the last observable frame with the first frame of action segment. If the active object presents in the last observable frame, we annotate the corresponding contact point and enforce a 2D Gaussian distribution to imitate the stochastic patterns of human-object interaction. Since the hand mask is absent from EPIC-Kitchens dataset, we adopt a 2D interpolation between the the finger tip annotation and interaction hotspots annotation to generate the pseudo ground truth of future hand trajectory (Take Fig. 1 (b) for an instance). Note that we use a smaller anticipation time (0.5s) on the EGTEA dataset. This is because the EGTEA dataset has a smaller angle of view in comparison with the EPIC-Kitchens dataset. A large anticipation time will reduce the number of samples that have next-active objects on the last observable frame. To summarize, there are 14951 annotated sample on the EPIC-Kitchens Dataset, and 7381 annotated samples on the EGTEA dataset. We believe those additional annotations can facilitate future research of human-object interaction in FPV.

Table 1. Comparison between our methods and previous state-of-the-art results RULSTM. See Sec.4.2 of our submission for discussion of Ous+Obj.

Method	Tasks	Training Supervision	Testing Inputs	End-to-End
RULSTM [15]	Action Anticipation	Action Labels Object Cls & Boxes	RGB + Object Feat. + Flow	No
Ours	Action Anticipation Visual Affordance Motor Attention Pred	Action Labels Hand & Hotspots	RGB	Yes
Ours+Obj	Action Anticipation Visual Affordance Motor Attention Pred	Action Labels Object Cls & Boxes Hand & Hotspots	RGB + Object Feat.	No

D Comparison of Experiment Setup to RULSTM

We highlight our comparison to RULSTM. In Sec.4.2 of our submission, we contrast our method with previous state-of-the-art result RULSTM [1]. Here, we draw a more clear boundary between our method and RULSTM. In Table 1, we present the experiment setup of our method and RULSTM. Both RULTM and our model (Ours) use various supervisory signals for training, yet our model only needs RGB frames for inference and is end-to-end trainable. Ours+Obj model does require more training signals in comparison to RULSTM, yet it does not need optical flow for two-stream architecture. We have to point out that, from practical prospective, we care more about the data modality during testing time. Therefore, using more supervisory signals for training does not compromise the contribution of our method. Moreover, our method also address the challenging problem of motor attention prediction and interaction hotspots estimation.

E Epic-Kitchens Challenge Leaderboard

Fig 3 presents a screenshot of the leaderboard from the EPIC-Kitchens Action Anticipation Challenge (`https://epic-kitchens.github.io/`).[4] The screenshot was acquired on the last day of supplementary material deadline. To date, our proposed method outperforms all published results by a large margin. Several very recent unpublished work (user id: "action_banks", "reza_zlf", "hepic", "prefact", "root" in Fig. 3) also attempted at the EPIC-Kitchens Action Anticipation Challenge. The only work that outperforms our method is "action_banks". Although "action_banks" slightly outperforms our method for action prediction, their results are worse than our method in terms of the verb and noun prediction.

F Experiment on Gaze Fixation Model

In this section, we present additional results on using gaze as attention distribution for visual anticipation. We follow [31] to replace motor attention and hotspots modules with a gaze module. We denote the reuslting model as Gaze

[4] Retrieved at March 13th, 2020.

Table 2. Additional results on fixation based model. We contrast Gaze Only Model with baseline I3D model and our full model.

Methods	Verb	Noun	Action
I3D-Res50	48.01/31.25	42.11/30.01	34.82/23.20
Gaze Only†	47.88/31.79	43.83/33.42	35.31/24.51
Ours†	48.96/32.48	45.50/32.73	36.60/25.30

Only model. The experiments are conducted on EGTEA dataset, as gaze is not available on EPIC-Kitchens dataset. Gaze Only model improves the I3D-Res50 baseline by a notable margin. However, it lags behind our full model. This is because our model explicitly reasons about the future representation by making motor attention a first class player.

G Additional Qualitative Results

Finally, we provide additional qualitative results. We included a video demo of our results as part of our supplementary materials. In this document, we also present more samples of predicted motor attention, interaction hotspots, and action labels in Fig 2. The figure follows the same format as Fig. 3 in the submission. These results further show that our proposed motor attention module has the remarkable ability of "imagining" possible hand movements even without the presence of hands in the observed video segments. Another interesting observation is that the predicted distribution of interaction hostpots can be sparse in certain circumstances (e.g., "Open Fridge" or "Take Condiment"). This is because of the stochastic patterns of human-object interaction: There might be multiple valid contact regions for interaction, especially when the future active object has a relatively large scale. This again shows the necessity of the stochastic units in our proposed method.

As discussed in our submission, the occlusion and absence of active objects might make the anticipation problem extremely challenging even for humans. The failure cases in Fig. 2 also suggest that the anticipation model can be biased by on-going action. This is because current FPV datasets (especially EPIC-Kitchens) segment a continuous action into several same atomic actions to ensure all action segments have similar temporal dimension. For instance, A video clip of "cutting onions" for 20 seconds is segmented into 7 or 8 shorter clips all having the same "cutting onions" label. This increases the transition probability of staying in current action state, and thereby biases the model. Therefore, the ability of predicting when exactly the action will end is important for more accurate action prediction model. This task is also related to the action localization problem in the literature [2].

References

1. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: ICCV (2019) 3

2. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018) 4
3. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: ECCV (2018) 1
4. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: CVPR (2016) 1

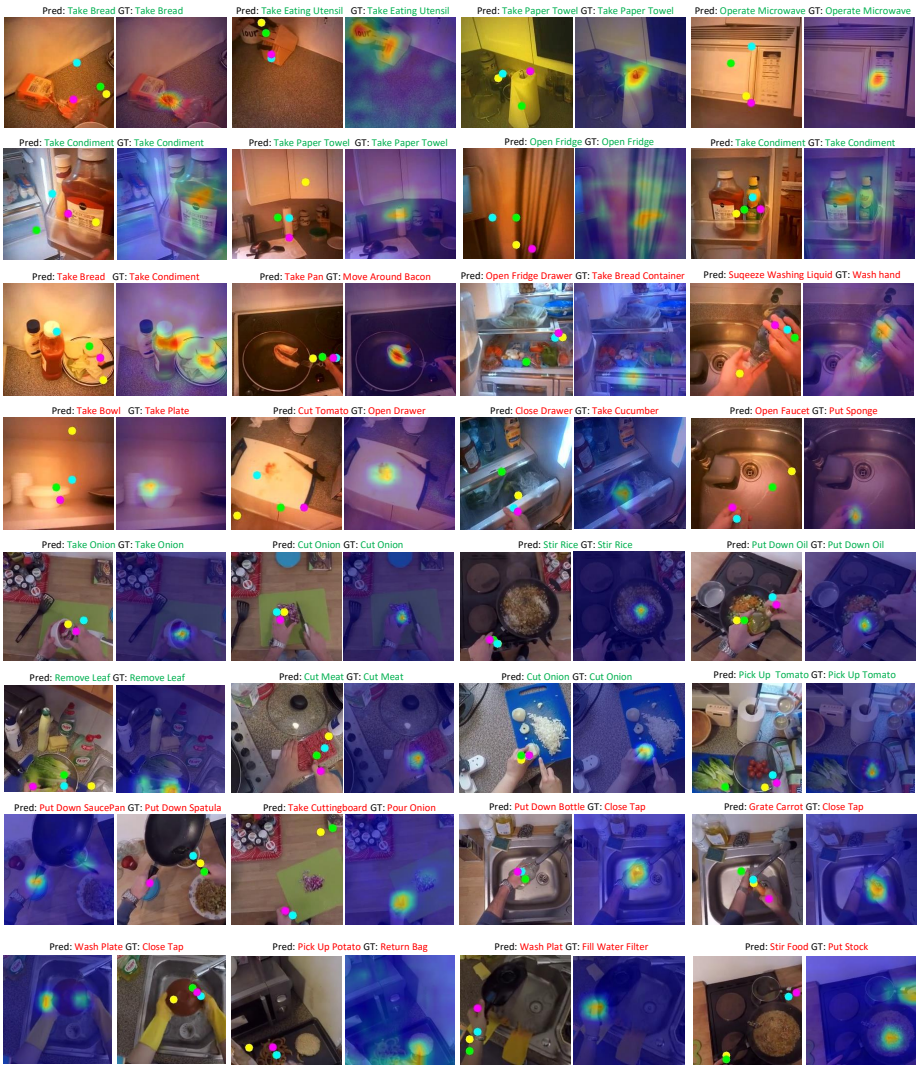

Fig. 2. Additional visualization of predicted motor attention (left), interaction hotspots (right), and action labels (top) from the EGTEA (row 1-4) and EPIC-Kitchens (row 5-8). Both successful cases (green label) and failure cases (red label) are presented. Future hands position are downsampled by a temporal factor of 8, and forecasted to the last observable frame in the order of yellow , green, cyan, and magenta.

Seen Kitchens (S1)

#	User	Entries	Date of Last Entry	Top-1 Accuracy (%)			Top-5 Accuracy (%)		
				Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲
1	action_banks	15	03/02/20	31.35 (5)	22.64 (6)	16.54 (1)	75.15 (10)	47.16 (6)	36.36 (1)
2	aptx4869lm	20	11/12/19	36.25 (1)	23.83 (3)	15.42 (2)	79.15 (3)	51.98 (3)	34.29 (3)
3	antoninofurnari	1	07/19/19	31.13 (6)	22.93 (4)	15.25 (3)	78.03 (6)	51.05 (4)	35.13 (2)
4	root	4	01/09/20	33.52 (3)	22.88 (5)	14.56 (4)	79.17 (2)	50.83 (5)	33.83 (4)
5	hepic	19	11/28/19	30.76 (7)	17.40 (7)	8.98 (5)	77.25 (7)	42.70 (7)	24.15 (6)
6	reza_zlf	3	10/28/19	29.45 (10)	24.78 (2)	8.30 (6)	79.50 (1)	53.16 (2)	25.55 (5)
7	prefact	13	11/12/19	26.32 (11)	25.60 (1)	7.44 (7)	79.10 (4)	53.18 (1)	23.10 (7)
8	EPIC_TSN_RGB	1	09/05/18	31.81 (4)	16.22 (8)	6.00 (8)	76.56 (8)	42.15 (8)	18.21 (8)
9	yassersouri	1	08/19/19	34.94 (2)	13.06 (10)	5.23 (9)	79.07 (5)	37.00 (10)	16.27 (9)
10	EPIC_TSN_Fusion	1	09/05/18	30.66 (8)	14.86 (9)	4.62 (10)	75.32 (9)	40.11 (9)	16.01 (10)
11	EPIC_TSN_Flow	1	09/05/18	29.64 (9)	10.30 (11)	2.93 (11)	73.70 (11)	30.09 (11)	10.92 (11)
12	jianjiangkcl	13	08/21/19	21.55 (13)	5.52 (13)	2.04 (12)	68.81 (13)	19.17 (13)	5.85 (12)
13	masterchef	2	09/06/18	25.30 (12)	7.78 (12)	1.58 (13)	69.98 (12)	22.07 (12)	5.85 (12)

Unseen Kitchens (S2)

#	User	Entries	Date of Last Entry	Top-1 Accuracy (%)			Top-5 Accuracy (%)		
				Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲
1	action_banks	15	03/02/20	27.48 (3)	16.63 (2)	10.00 (1)	66.81 (11)	32.84 (6)	23.39 (2)
2	aptx4869lm	20	11/12/19	29.87 (2)	16.80 (1)	9.94 (2)	71.77 (2)	38.96 (2)	23.69 (1)
3	antoninofurnari	1	07/19/19	26.63 (6)	15.47 (4)	9.12 (3)	68.11 (9)	35.27 (4)	21.88 (3)
4	root	4	01/09/20	27.42 (4)	15.30 (5)	8.02 (4)	69.14 (6)	35.06 (5)	20.52 (4)
5	prefact	13	11/12/19	23.22 (10)	16.18 (3)	4.81 (5)	71.01 (3)	39.23 (1)	14.27 (5)
6	hepic	19	11/28/19	27.01 (5)	9.76 (8)	4.64 (6)	69.48 (4)	27.55 (9)	14.27 (5)
7	yassersouri	1	08/19/19	32.37 (1)	9.66 (9)	3.52 (7)	73.51 (1)	30.83 (7)	12.67 (7)
8	reza_zlf	3	10/28/19	22.29 (12)	14.10 (6)	3.18 (8)	69.38 (5)	35.47 (3)	12.91 (6)
9	EPIC_TSN_RGB	1	09/05/18	25.30 (9)	10.41 (7)	2.39 (9)	68.32 (7)	29.50 (8)	9.63 (8)
10	EPIC_TSN_Flow	1	09/05/18	25.61 (7)	8.40 (10)	1.78 (10)	67.57 (10)	24.62 (11)	8.19 (10)
11	EPIC_TSN_Fusion	1	09/05/18	25.37 (8)	9.76 (8)	1.74 (11)	68.25 (8)	27.24 (10)	9.05 (9)
12	masterchef	2	09/06/18	22.36 (11)	6.59 (11)	0.89 (12)	63.43 (12)	19.73 (12)	3.72 (11)
13	jianjiangkcl	13	08/21/19	16.08 (13)	3.28 (12)	0.68 (13)	59.88 (13)	13.28 (13)	2.77 (12)

Fig. 3. Screenshot from Epic-Kitchens Anticipation Challenge. The user name of our proposed method is "aptx4869lm". The current rank1 team "action_banks" is unpublished work, and lags behind of our method for both verb and noun prediction on both sets. Note that user "antonionfurnari" refers to RULSTM in our main submission. They further improved the results reported in their paper.

ID	Branch	Type	Kernel Size THW,(C)	Stride THW	Output Size THWC	Comments (Loss)
1	Backbone (shared)	Conv3D	5x7x7,64	2x2x2	16x112x112x64	
2		MaxPool1	2x3x3	2x2x2	8x56x56x64	
3		Layer1 Bottleneck 0-2	3x1x1,64 1x3x3,64 (3 times) 1x1x1,256	1x1x1 1x1x1 (3 times) 1x1x1	8x56x56x256	
4		MaxPool2	2x1x1	2x1x1	4x56x56x256	Addition Pooling Reduce Memory Usage
5		Layer2 Bottleneck 0	3x1x1,128 1x3x3,128 1x1x1,512	1x1x1 1x2x2 1x1x1		
6		Layer2 Bottleneck 1-3	3x1x1,128 1x3x3,128 (3 times) 1x1x1,512	1x1x1 1x2x2 (3 times) 1x1x1	4x28x28x512	
7		Layer3 Bottleneck 0	3x1x1,256 1x3x3,256 1x1x1,1024	1x1x1 1x2x2 1x1x1		
8		Layer3 Bottleneck 1-5	3x1x1,256 1x3x3,256 (5 times) 1x1x1,1024	1x1x1 1x1x1 (5 times) 1x1x1	4x14x14x1024	
9		Layer4 Bottleneck 0	3x1x1,128 1x3x3,128 1x1x1,512	1x1x1 1x2x2 1x1x1		
10		Layer4 Bottleneck 1-2	3x1x1,128 1x3x3,128 (2 times) 1x1x1,512	1x1x1 1x2x2 (2 times) 1x1x1	4x7x7x2048	
11	Motor Attention Module	Conv3d 1 (on Layer 2 feature)	1x3x3,128	1x1x1	4x28x28x128	
12		Conv3d 2	1x3x3,1	1x1x1	4x28x28x1	KLD Loss
13		Maxpool 1	1x2x2	1x2x2	4x14x14x1	Guiding Interaction Hotspots
14		Gumbel Softmax 1 (Sampling)			4x14x14x1	Sampling Motor Attention
15		Maxpool 2	1x4x4	1x4x4	4x7x7x1	Guiding Action Anticipation
16		Gumbel Softmax 2 (Sampling)			4x7x7x1	Sampling Motor Attention
17	Interaction Hotspots Module	Weighted Pooling			4x14x14x256	With Sampled Motor Attention
18		Conv3d 1 (on Layer 3 Feature)	1x3x3,256	1x1x1	4x14x14x256	
19		Conv3d 2	1x3x3,1	1x1x1	4x14x14x1	KLD Loss
20		Maxpool 1	1x2x2	1x2x2	4x7x7x1	Guiding Action Anticipation
21		Gumbel Softmax (Sampling)			4x7x7x1	Sampling Interaction Hotspots
22	Action Anticipation Module	Weighted Avg Pool (on Final Feature)	4x7x7	4x7x7	1x1x1x1024	With Sampled Motor Attention and Interaction Hotspots
23		Fully Connected			1x1x1xN	
24		Softmax			1x1x1xN	Cross Entropy Loss (Action Anticipation)

Table 3. Network architecture of our proposed model. We omit the residual connection in backbone ResNet-50 for simplification.