Interaction-aware Joint Attention Estimation Using People Attributes (Supplementary Material)

Chihiro Nakatani[1] Hiroaki Kawashima[2] Norimichi Ukita[1]
[1] Toyota Technological Institute, Japan [2] University of Hyogo, Japan

1. Implementation Details

This section shows the implementation details that are not mentioned in the main paper.

1.1. Architecture of Fusion Modules

We employ a Convolutional Neural Network (CNN) as a fusion module for comparison, as mentioned in the main paper. The architecture of the CNN module is shown in Fig. 10. In the CNN module, each of H_{JA} and H_{AT} is fed into the CNN consisting of three convolutional layers separately. The channel size of the output feature map is 8. The two feature maps extracted from H_{JA} and H_{AT} are concatenated into a feature map. The concatenated feature map is fed into the CNN consisting of three convolutional layers followed by the Sigmoid activation at the final layer. Finally, the concatenated feature map is transformed into a one-channel heatmap representation.

1.2. Training Conditions

The parameters of our network are optimized by Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Weight decay is set to 0 for both datasets. We trained all of our models on the Nvidia A100 GPUs with batch size 4. σ^2 is empirically determined to be 10. All input images are resized to 320x640 and 320x480 for Volleyball [18] and VideoCoAtt [9] datasets, respectively. Note that evaluation metrics about distance (i.e., Dist and Thr) are calculated on the original image size.

For the experiments on the Volleyball and VideoCoAtt datasets, the gaze estimation network is trained on Volleyball and GazeFollow [43] datasets, respectively. Annotated gaze direction of each person is provided only in the GazeFollow dataset. For the Volleyball dataset, we automatically annotate the gaze directions as the directions of the person's head to the ball position.

1.3. Implementation of Previous Methods

In our experiments, ISA [9], DAVT [6], and HGTD [50] are used for comparison. Codes of these methods are prepared as follows:

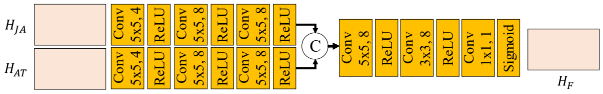

Figure 10. Fusion module. Each conv layer is shown with its kernel size and output channel size. In all the conv layers, strite = 1 is used for maintaining the xy size of the feature map.

- ISA ··· While the source code is available at https://github.com/LifengFan/Shared-Attention/tree/master/src, training and evaluation processes are not mentioned clearly. We therefore implement the method based on the paper and the source codes. Finally, the models are trained on the VideoCoAtt dataset as with the original paper in our experiments.
- DAVT ··· The Code and trained models are available at https://github.com/ejcgt/attention-target-detection. We use the trained model for our experiments.
- HGTD ··· The Code is not available. Therefore, we implement the method based on the paper. We use the model trained on GazeFollow and VideoAttentionTarget dataset as with the original paper.

Note that all methods, including our method, are trained and evaluated on the same settings (i.e., Ex.1 and Ex.2) for a fair comparison, as mentioned in the main paper.

2. Experimental Details

2.1. Volleyball Dataset

For our proposed network, people attributes (i.e., $\boldsymbol{l}$, $\boldsymbol{g}$, and $\boldsymbol{a}$) are prepared for the following two settings: (Ex.1) full-body bounding boxes given by YOLOv5 [21] and actions given by ARG are used in the training and test phases; and (Ex.2) ground-truth full-body bounding boxes and actions are used in the training and testing phases. In both experimental settings, head bounding boxes detected by

Table 9. Ablation studies in Ex.2 on the Volleyball dataset. Ablated components about the people attributes and the network architectures are separated by double lines. Each metric is evaluated with two results, namely H_{JA} in branch (α) and H_F in fusion module (γ). While smaller values are better in Dist, larger values are better in the other metrics.

Method	Dist (α)	Dist (γ)	Thr=30 (α)	Thr=60 (α)	Thr=90 (α)	Thr=30 (γ)	Thr=60 (γ)	Thr=90 (γ)
Ours w/o $\boldsymbol{l}$	83.4	60.3	16.6	44.9	70.0	63.6	74.2	81.5
Ours w/o $\boldsymbol{g}$	58.9	41.6	49.2	78.2	90.5	71.9	85.3	91.7
Ours w/o $\boldsymbol{a}$	32.5	39.2	74.0	92.2	97.0	75.5	86.2	91.3
Ours w/o (α)	-	77.4	-	-	-	59.7	69.7	76.6
Ours w/o (β)	14.3	14.3	95.1	98.6	99.2	95.1	98.6	99.2
Ours	11.6	11.4	96.3	98.8	99.6	96.3	98.9	99.6

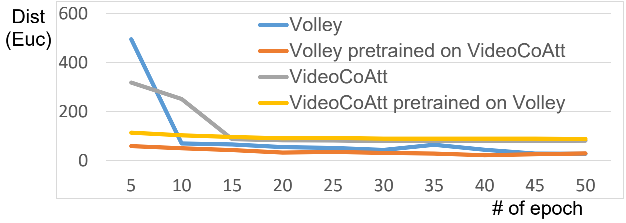

Figure 11. Learning curve of our models with pretrained models.

YOLOv5 are used because they are not annotated. In Ex.1, head bounding boxes in a whole image are used, which may cause miss detection of non-players. In Ex.2, head bounding boxes inside the ground-truth full-body bounding boxes are used.

2.2. VideoCoAtt Dataset

While ISA and DAVT are used for both dataset evaluations, HGTD is only used for the VideoCoAtt dataset because ground-truth head bounding boxes are required to train the network. These ground-truth head bounding boxes are not annotated in the Volleyball dataset. Different from the Volleyball dataset, the detected and ground-truth bounding boxes are used in Ex.1 and Ex.2, respectively.

3. Additional Experiments

Additional experimental results not included in the main paper for the page limitation are presented in this section.

3.1. Initialization by Pretrained Model

The model trained on one group activity dataset (e.g., Volleyball dataset) can be generalized to other group activities by finetuning. Figure 11 shows that the results obtained by the finetuning of a pretrained model are better than randomly initialized models in the early epochs.

Table 10. Comparison of different heatmap generators in the branch (α) in Ex.2 on the Volleyball dataset. Ours uses "(iii) J_{JA} only" for pixelwise estimation. The results obtained by H_{JA} are evaluated.

Method	Dist (α)	Thr=30	Thr=60
(i) F_{JA} only	15.6	95.0	98.2
(ii) F_{JA} and J_{JA}	17.2	93.9	97.9
J_{JA} only for imagewise	116.5	8.7	28.7
(iii) J_{JA} only (Ours)	14.3	95.1	98.6

Table 11. Comparisons of different fusion modules in Ex.2 on Volleyball dataset. The results obtained by H_F are evaluated.

Fusion	Dist	Thr=30	Thr=60
CNN	11.6	96.2	98.9
Average	12.7	95.4	98.9
Weighted (Ours)	11.4	96.3	98.9

3.2. Volleyball Dataset

3.2.1 Ablation Studies

Ablation studies in Ex.2 on the Volleyball dataset are shown in Table 9. In contrast to the results in Ex.1 (shown in the main paper), the results are improved in all metrics. It is natural because there is no error in ground-truth people attributes used in Ex.2.

3.2.2 Detailed Analysis

PJAT architecture comparison. The comparison of PJAT architecture is shown in Table 10. As with the main paper, the results obtained by "J_{JA} only (Ours)" is slightly better than the others.

Pixelwise vs. imagewise. Pixelwise estimation with PJAT is compared with general imagewise heatmapping in Ex.2. The results in Ex.1 are shown in the Table 7 of the main paper. Imagewise heatmapping means that a high-dimensional heatmap is directly estimated from a low-dimensional feature. As shown in Table 10, "J_{JA} only (Ours)" outperforms "J_{JA} only for imagewise" because our pixelwise estimation

Table 12. Ablation studies in Ex.1 on the VideoCoAtt dataset. Ablated components about the people attributes and the network architectures are separated by double lines. Each metric is evaluated with two results, namely H_{JA} in branch (α) and H_F in fusion module (γ).

Method	Dist (α)	Dist (γ)	Thr=40 (α)	Thr=40 (γ)	Acc. (α)	Acc. (γ)	F-score (α)	F-score (γ)
Ours w/o $\boldsymbol{l}$	95.6	68.5	14.7	54.6	0.50	0.54	0.35	0.36
Ours w/o $\boldsymbol{g}$	122.1	74.3	13.3	54.6	0.50	0.52	0.35	0.36
Ours w/o (α)	103.0	68.1	25.6	58.6	0.50	0.52	0.35	0.32
Ours w/o (β)	97.9	97.9	36.8	36.8	0.52	0.52	0.34	0.34
Ours	96.6	66.5	15.4	59.1	0.50	0.52	0.35	0.36

Table 13. Ablation studies in Ex.2 on the VideoCoAtt dataset. Ablated components about the people attributes and the network architectures are separated by double lines. Each metric is evaluated with two results, namely H_{JA} in branch (α) and H_F in fusion module (γ).

Method	Dist (α)	Dist (γ)	Thr=40 (α)	Thr=40 (γ)	Acc. (α)	Acc. (γ)	F-score (α)	F-score (γ)
Ours w/o $\boldsymbol{l}$	80.4	46.4	33.0	72.0	0.60	0.72	0.29	0.31
Ours w/o $\boldsymbol{g}$	89.3	45.8	38.8	72.9	0.61	0.66	0.29	0.30
Ours w/o (α)	89.3	46.6	31.0	72.9	0.61	0.61	0.29	0.30
Ours w/o (β)	80.1	80.1	40.2	40.2	0.69	0.69	0.30	0.30
Ours	81.3	45.0	39.3	74.3	0.60	0.57	0.29	0.37

Table 14. Comparison of different heatmap generators in the branch (α) in Ex.1 on the VideoCoAtt dataset. Ours uses "(iii) J_{JA} only" for pixelwise estimation. The results obtained by H_{JA} are evaluated.

Method	Dist (α)	Thr=40	Thr=80
(i) F_{JA} only	98.7	33.6	49.2
(ii) F_{JA} and J_{JA}	93.5	37.2	53.6
J_{JA} only for imagewise	131.3	13.2	31.1
(iii) J_{JA} only (Ours)	97.9	36.8	50.3

Table 15. Comparison of different heatmap generators in the branch (α) in Ex.2 on the VideoCoAtt dataset. Ours uses "(iii) J_{JA} only" for pixelwise estimation. The results obtained by H_{JA} are evaluated.

Method	Dist (α)	Thr=40	Thr=80
(i) F_{JA} only	80.6	39.8	66.4
(ii) F_{JA} and J_{JA}	80.4	40.7	67.3
J_{JA} only for imagewise	159.6	5.6	19.2
(iii) J_{JA} only (Ours)	80.1	40.2	67.8

with positional information avoids an ill-posed problem in imagewise estimation.

Fusion module comparison. As with the fusion module comparison in Ex.1 shown in Table 8 of the main paper, the fusion module comparison in Ex.2 is also shown in Table 11. While the results obtained by "CNN" is worse than the others in Ex.1, "CNN" archives the same performance as the others in Ex.2. The differences come from the reliability of branches (α) and (β). While the performance gap between branches (α) and (β) was not large in Ex.2, the performance of branch (α) is better than that of branch (β) in Ex.1, as shown in Table 9. Therefore, "CNN" in Ex.2 can learn the network easily compared with Ex.1. For example, only using information from H_{JA} can lead the high performance in Ex.2.

3.3. VideoCoAtt Dataset

3.3.1 Ablation Studies

The effect of each important component in our method is verified for VideoCoAtt dataset with the ablation studies in Tables 12 and 13. We ablate either of $\boldsymbol{l}$ and $\boldsymbol{g}$ (i.e., people attributes) by filling zero into ablated nodes in the first layer of the feature extractor network. We also ablate either of network branches (α) and (β). For the experiments, without branch (α) or (β), the output of each branch is regarded as the final joint attention estimation.

In Ex.1, "Ours" achieved the best performance in the results obtained by module (γ) except for Accuracy. It is not surprising because the Accuracy is optimized for F-score by validation data. Regarding the results obtained by branch (α), "Ours" achieved competitive performances. It shows that the ablated components are not usually effective for the performance of branch (α), but they can be important for the final estimation results obtained by branch (γ). As with the results in Ex.1, we can see the same trend in Ex.2.

3.3.2 Detailed Analysis

PJAT architecture comparison. The comparison of PJAT architecture is shown in Tables 14 and. 15. Different from the results on the Volleyball dataset, the results show that "F_{JA} and J_{JA}" achieve better performance than "J_{JA} only". The results imply the combination of F_{JA} and J_{JA} can be used for more robust estimation in a scene that only

Table 16. Comparisons of different fusion modules in Ex.1 on VideoCoAtt dataset. The results obtained by H_F are evaluated.

Fusion	Dist	Thr=40	Thr=80
CNN	75.5	47.6	66.0
Average	75.6	34.1	49.4
Weighted (Ours)	66.5	59.1	68.7

Table 17. Comparisons of different fusion modules in Ex.2 on VideoCoAtt dataset. The results obtained by H_F are evaluated.

Fusion	Dist	Thr=40	Thr=80
CNN	53.0	64.8	79.6
Average	44.2	74.2	83.0
Weighted (Ours)	45.0	74.3	82.3

contains a small number of people such as the VideoCoAtt dataset.

Pixelwise vs. imagewise. As with the Volleyball dataset, pixelwise estimation with PJAT is compared with general imagewise heatmapping in Tables 14 and 15. In both experimental settings, the results validate that "J_{JA} only (Ours)" outperforms "J_{JA} only for imagewise" to a certain degree, as validated on the Volleball dataset.

Fusion module comparison. The fusion module is also compared on the VideoCoAtt dataset, as shown in Tables 16 and 17. While "Weighted (Ours)" outperforms the others in Ex.1, "Average" and "Weighted (Ours)" achieve the comparative performance in Ex.2. It is not surprising because weights for branches (α) and (β) are trained as 0.6 and 0.4 in "Weighted (Ours)". These trained weights are similar to the weights of "Average" in which 0.5 is used as the static weight for both branches (α) and (β).

Threshold optimization comparison. The threshold that leads to the maximum F-score in validation data is used for our experiments. The threshold optimization should be changed according to which metrics are focused on in the task. For more detailed analysis, we also use the threshold which leads to the maximum Accuracy.

The comparison of threshold optimization is shown in Table 18. Each of "Accuracy-based threshold" and "F-score based threshold" denotes that the threshold which leads to the maximum Accuracy and F-score is used, respectively. While Accuracy obtained by "Accuracy-based threshold" archived the high performance, F-score is worse than the results obtained by "F-score based threshold". The result comes from the class imbalance of no joint AP, as mentioned in the main paper. In the class imbalance situation, threshold optimization by metrics taking into account the class imbalance (e.g., F-score) should be used, as done in the main paper.

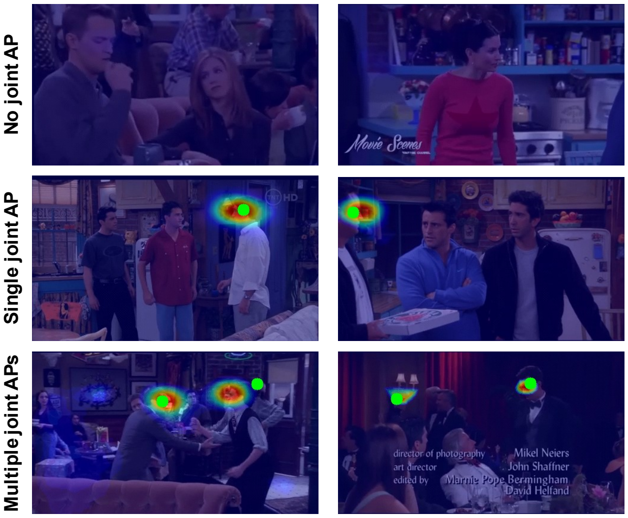

Figure 12. Visualization of joint attention estimation in no joint AP and multiple joint APs scene on the VideoCoAtt dataset.

No joint AP and multiple joint APs As mentioned in the main paper, no joint AP is observed in many frames in the VideoCoAtt dataset. The visualization of the no joint AP cases is shown in the upper examples of Fig. 12. The results show that our method successfully estimates a joint attention heatmap with low confidence values in no joint AP cases, compared with the high confidence values in single joint AP cases. The difference in confidence values is also validated by various metrics (i.e., Accuracy, F-score, and AUC) for detection accuracy in the main paper.

Furthermore, this dataset also contains several frames where multiple joint APs are observed, as shown in the bottom examples of Fig. 12. In these examples, multiple peaks are activated in the estimated joint attention heatmap. The results show that our method can estimate a joint attention heatmap even in multiple joint APs cases.

4. Future Work

Future work includes the improvement of person attributes. First, errors in predicted attributes degrade the performance, as shown in Sec. 4.5 of the main paper. For more improvement, error rectification of people attributes in our network is required. Second, actions are not used as person attributes on the VideoCoAtt since no action annotation is given to this dataset. Annotating reasonable actions for joint attention estimation is required to achieve performance, as with the Volleyball dataset.

Table 18. Comparision of the different threshold optimization on VideoCoAtt. The results obtained by Ex.1 and Ex.2 are separated by double lines.

Method	Accuracy (α)	Accuracy (γ)	F-score (α)	F-score (γ)
Ours (Ex.1, Accuracy-based threshold)	0.84	0.84	0.00	0.03
Ours (Ex.1, F-score based threshold)	0.50	0.52	0.35	0.36
Ours (Ex.2, Accuracy-based threshold)	0.84	0.84	0.05	0.16
Ours (Ex.2, F-score based threshold)	0.60	0.57	0.29	0.37