Interaction-aware Joint Attention Estimation Using People Attributes

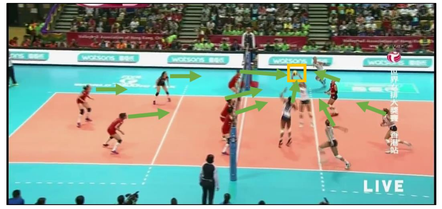
Chihiro Nakatani[1], Hiroaki Kawashima[2], Norimichi Ukita[1]

[1] Toyota Technological Institute, Japan [2] University of Hyogo, Japan

Introduction

Task

Estimating **joint attention** shared by **multiple** people



Problems

P-1. No contribution weights of people

People are equally weighted for joint attention estimation

P-2. No explicit interaction among people attributes

Interactions among people attributes are neglected

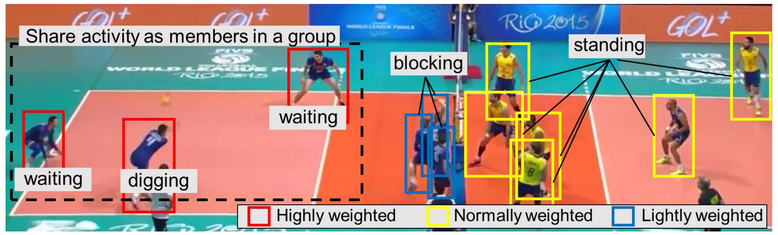
Contributions

C-1. Activity awareness

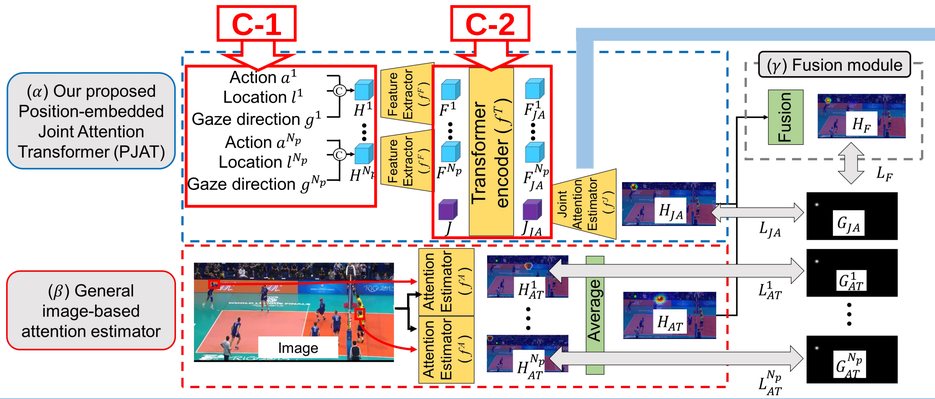
Location and action of each person are used for weighting

C-2. Interaction awareness

Interactions among people attributes are modeled by self-attention

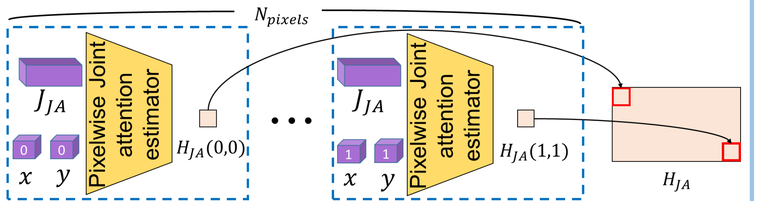


Proposed Method (Position-embedded Joint Attention Transformer: PJAT)



Pixelwise joint attention heatmapping

Estimate the joint attention confidence pixelwise with image-coordinate to avoid ill-pose problem



Experiments

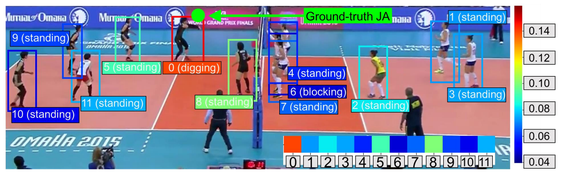
Experimental setting

		Att	Body	Head	Action
Vol	Ex.1	l, g, a	Pr	Pr in image	Pr
	Ex.2	l, g, a	GT	Pr in GT body	GT
Vid	Ex.1	l, g	–	Pr in image	–
	Ex.2	l, g	–	GT	–

Ablation study

Method	Dist ↓	Thr=30 ↑	Thr=60 ↑	Thr=90 ↑
Ours w/o l	60.3	63.6	74.2	81.5
Ours w/o g	41.6	71.9	85.3	91.7
Ours w/o a	39.2	75.5	86.2	91.3
Ours w/o (α)	77.4	59.7	69.7	76.6
Ours w/o (β)	14.3	95.1	98.6	99.2
Ours	11.4	96.3	98.9	99.6

Attention values in self-attention



Qualitative results

ISA [2] DAVT [1] Ours

Ball detection vs. Ours

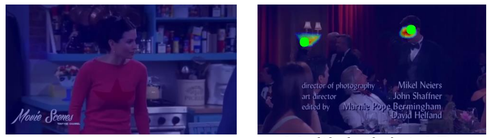
Ball detection Ours

Quantitative results

Method	Dist ↓	Thr=30 ↑	Thr=60 ↑	Thr=90 ↑
ISA [2] (Ex.1)	70.1	60.7	69.7	75.9
DAVT [1] (Ex.1)	72.0	62.0	72.8	78.6
Ours (Ex.1)	56.0	64.5	76.8	83.0
ISA [2] (Ex.2)	48.7	46.0	79.1	92.8
DAVT [1] (Ex.2)	77.4	59.7	69.7	76.6
Ours (Ex.2)	11.4	96.3	98.9	99.6

Method	Dist ↓	Thr=40 ↑	Thr=80 ↑	Accuracy ↑	AUC ↑
ISA [2] (Ex.1)	152.7	8.5	24.9	0.41	0.41
DAVT [1] (Ex.1)	68.2	58.6	68.5	0.52	0.58
Ours (Ex.1)	66.5	59.1	68.7	0.52	0.64
ISA [2] (Ex.2)	107.1	5.6	36.7	0.62	0.64
DAVT [1] (Ex.2)	46.6	72.9	80.7	0.61	0.57
Ours (Ex.2)	45.0	74.3	82.5	0.57	0.65

Flexible estimation results



No joint AP Multiple joint APs

References

1. E. Chong et al. Detecting attended visual targets in video. CVPR2020.
2. L. Fan et al. Inferring shared attention in social scene videos. CVPR2018.
3. M. Ibrahim et al. A hierarchical deep temporal model for group activity recognition. CVPR2016.

Acknowledgement (International exchange grant)

This work was supported by Tateishi Science and Technology Foundation