A*: Atrous Spatial Temporal Action Recognition for Real Time Applications

Myeongjun Kim   Federica Spinola   Philipp Benz   Tae-hoon Kim
Deeping Source Inc., Seoul, Republic of Korea
{myeongjun.kim, federica.spinola, philipp.benz, pete.kim}@deepingsource.io

1. Speed Analysis

We report the speed in Frames per Second (FPS) for different backbones and number of input frames. We compare our method to the other real-time method YOWO [2]. In fact, to the best of our knowledge, YOWO [4] is the only other method that reports their fast inference speed and only use past frames for prediction.

Table 1. Comparison of our Model's FPS Against YOWO on the JHMDB Dataset. Values are calculated using a RTX 3090 GPU.

Model	Input	Backbone	FPS
YOWO [2]	16f	RNext101	58
YOWO [2]	32f	RNext101	55
YOWO [2]	16f	SFR50	59
YOWO [2]	32f	SFR50	57
A* (Ours RTS)	16f	RNext101	32
A* (Ours RTS)	32f	RNext101	24
A* (Ours RTS)	16f	SFR50	43
A* (Ours RTS)	32f	SFR50	33

2. Failure case Analysis

2.1. Mis-labeled ground-truth problem

In the ground-truth (GT) labels of our benchmark datasets, we notice that some target actions can be labeled even though they do not occur in the video clip. Therefore, these cases can cause additional confusion in the model predictions.

2.2. Lack of focus on the target person

The proposed network architecture utilizes the entire video feature, making it easier to consider the global context. However, our method seems to lack some focus on the target person compared to using RoIAlign features. This results in our bounding-box prediction being less accurate than other methods.

3. Qualitative Results

We show multiple qualitative results of our method and compare them to previous methods. We perform a small analysis explaining A*'s strengths and failure cases.

Figures 1, 2 and 3 illustrate model predictions on the JHMDB-21 [3] dataset. Specifically, in Figure 1 we compare our A* with offline setting (A* OS), our A* with real-time setting (A* RTS), and the baseline YOWO [2], which also operates in the real-time setting. These qualitative results show that our method predicts more plausible action labels than YOWO. In Figure 2 we show our model's predictions over a sequence of frames. A* is capable of recognizing actions from their onset to their end by learning long-term temporal information and global spatial context queues. However, there are still some failure cases as illustrated in Figure 3 B. The failures can occur when the context information is occluded or poorly visible. For example, the chair in row 2 of Figure 3 B cannot easily be distinguished. Additionally, some inaccurate GT labels can confuse the network. For example, row 1 in Figure 3 B is labeled as *standing*, but the person is clearly seated at the start of the video. Therefore, the proposed A* is making appropriate predictions.

Figure 4 illustrates qualitative comparisons between our best A* OS (sub-figures A) and STMixer [5] (sub-figures B) on the AVA v2.2 [1] dataset. As can be seen in the Figures, our A* predicts comparable action labels to STMixer. However, in most cases, our bounding-boxes predictions are not as tightly-fitted to the person as STMixer's bounding-boxes. This is partly due to some ambiguities in the GT bounding boxes For example, the GT bounding box in row 3 of Figure 4 does not contain the person's hands. Another reason is explained in Subsection 2.2 above. Therefore, A*'s small deviations from the GT in bounding-box prediction contribute to lowering A*'s frame-mAP score on AVA. Nevertheless, the large proportion of correctly predicted action labels on AVA demonstrates A*'s meaningful performance in real-world action recognition applications.

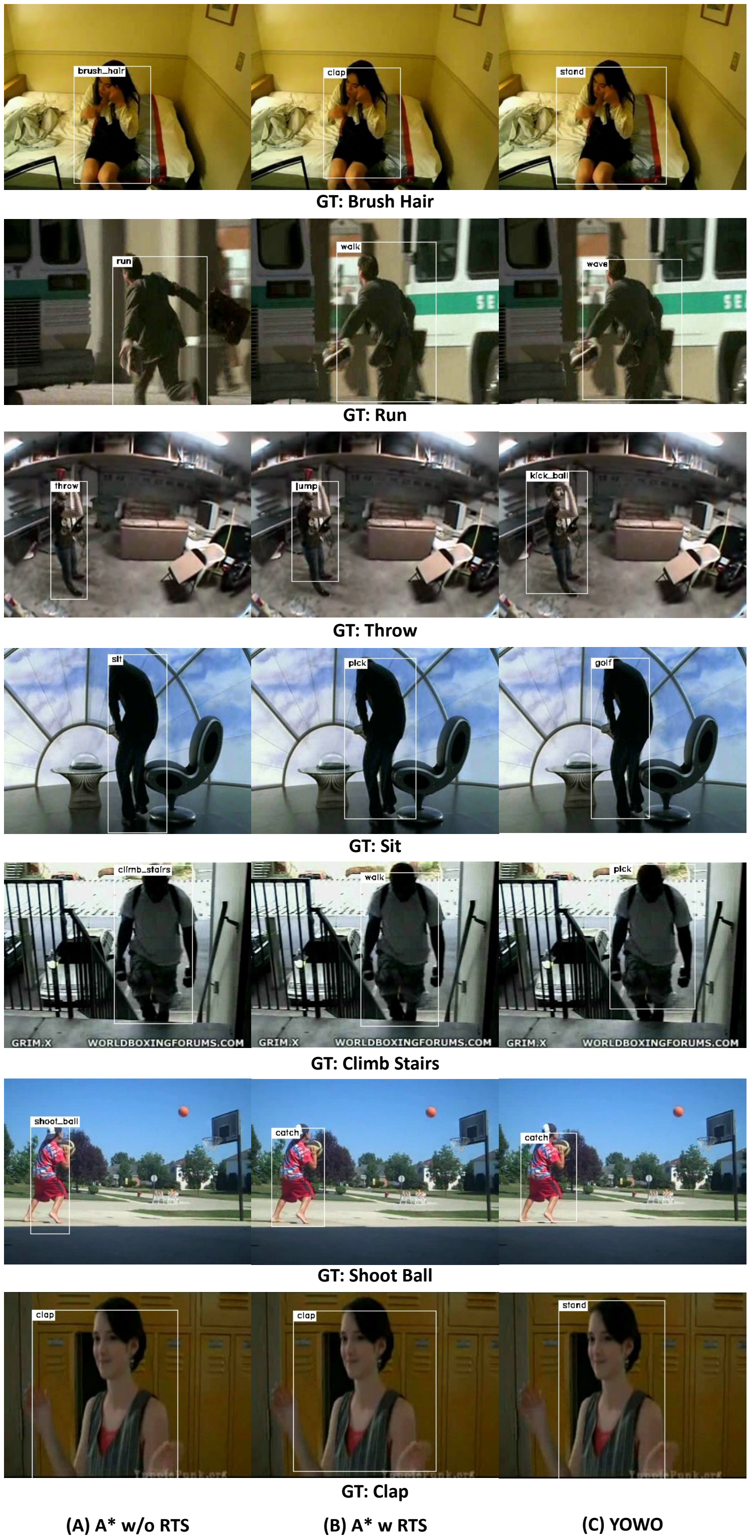


Figure 1. Qualitative results on the JHMDB-21 dataset comparing A* OS in column A, A* RTS in column B, YOWO (baseline) in column C. The network predictions are shown next to the bounding-boxes. The GT is shown below each row of images. RTS: Real-time setting. Best viewed in color.

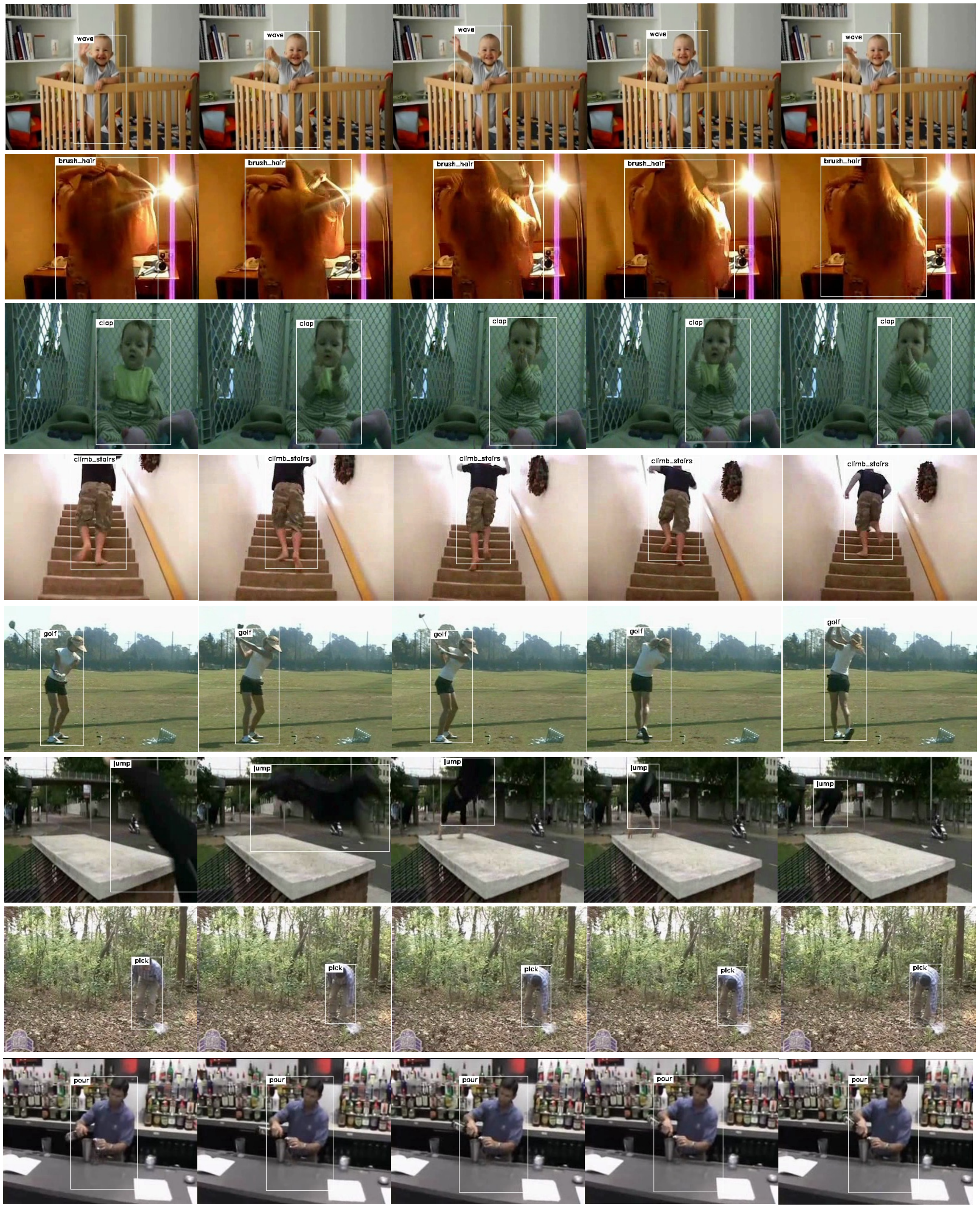
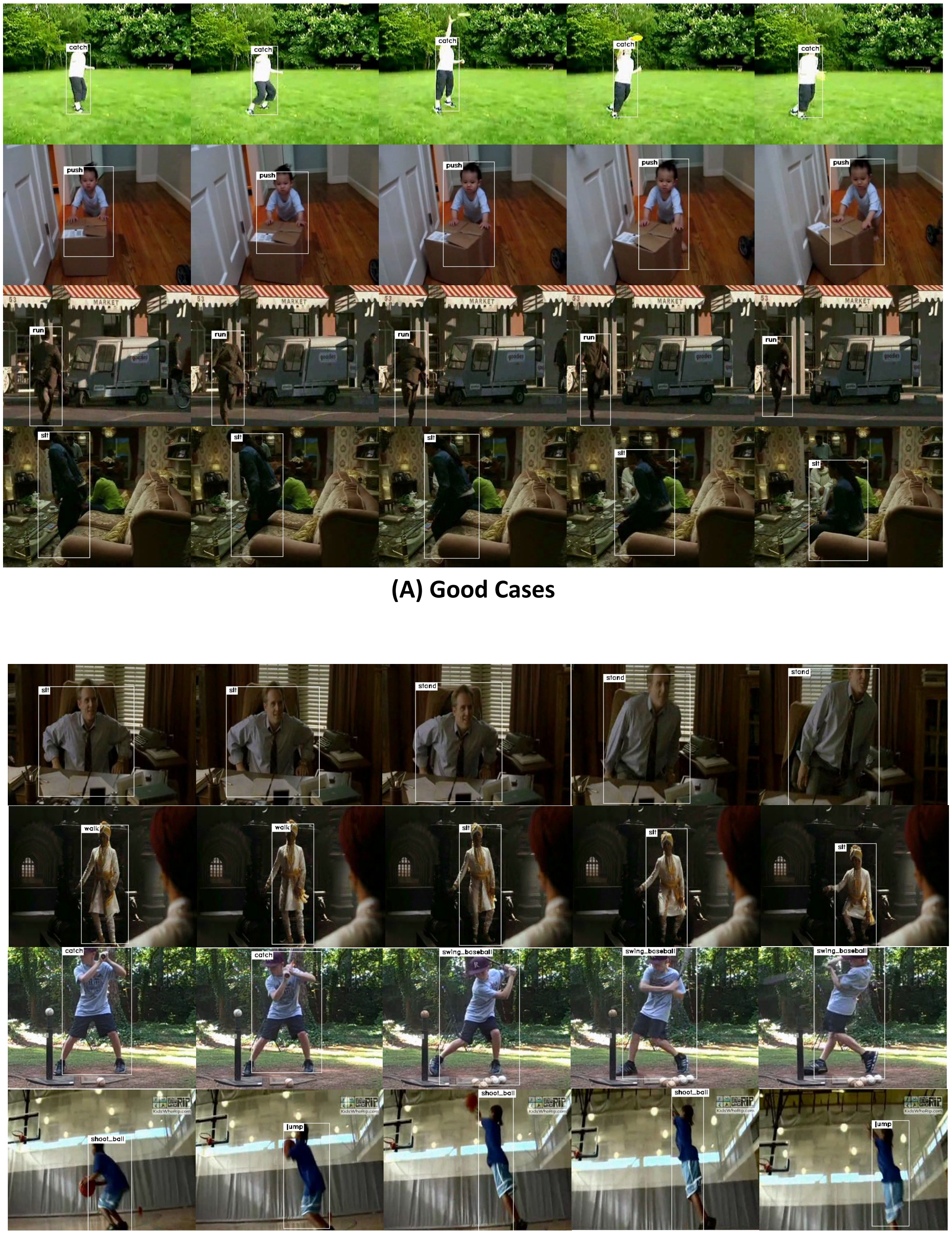


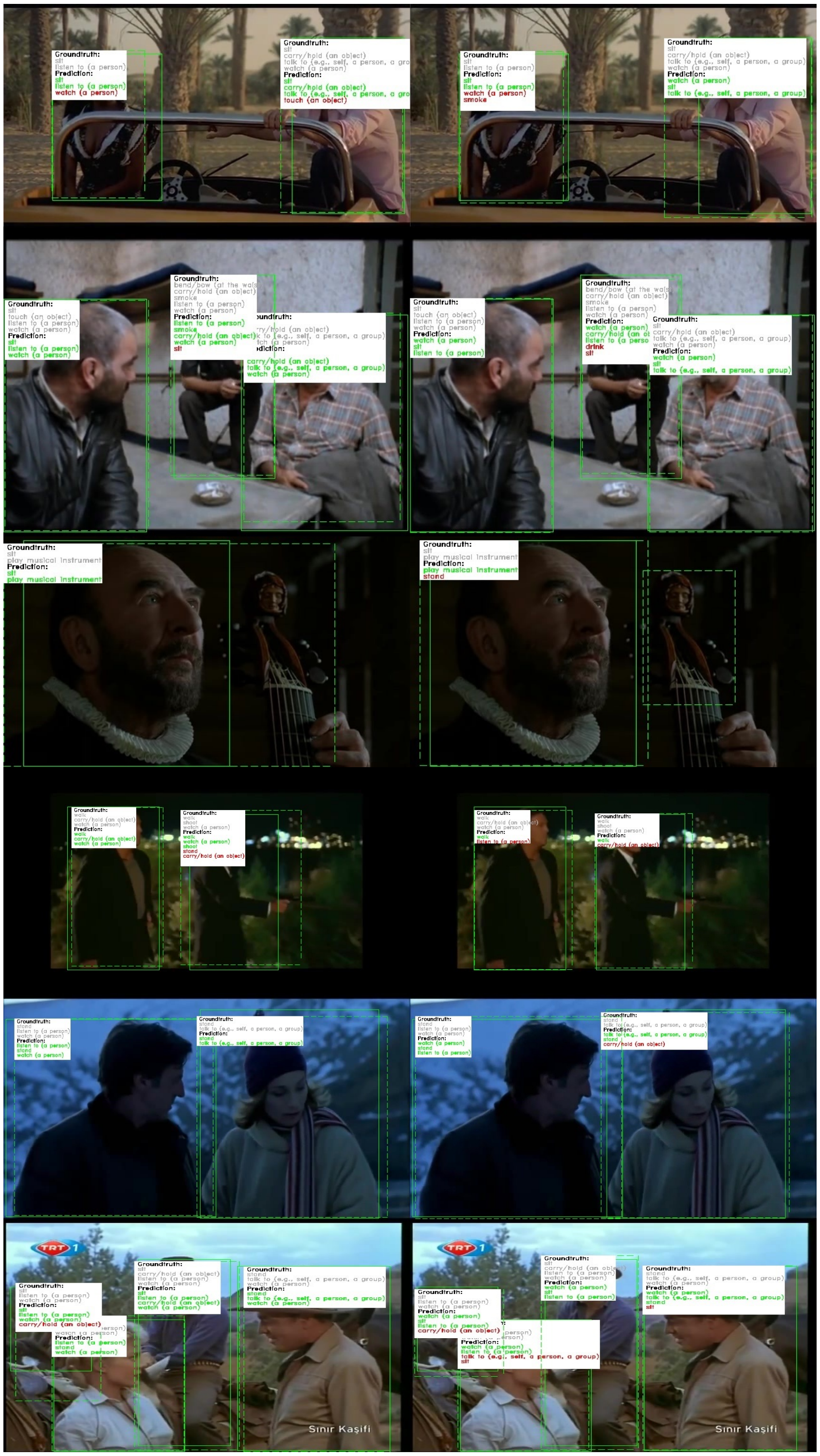
Figure 2. Qualitative results of our A* OS on JHMDB-21. These results show correct predictions over a sequence of images sampled at regular intervals from the output video. Best viewed in color.



(A) Good Cases

(B) Failure Cases

Figure 3. More Qualitative results of our A* OS on JHMDB-21. These results show additional correct predictions and failure cases over a sequence of images sampled at regular intervals from the output video. Sub-figure A shows correct predictions. Sub-figure B shows failure cases. Best viewed in color.



(A) A* **(B) STMixer**

Figure 4. Qualitative results on the AVA v2.2 dataset comparing our A* OS in column A, and STMixer in column B. The GT and network predictions are shown next to the bounding-boxes. Green action predictions are correct predictions and red action predictions are wrong predictions. Dashed-lines represent predicted bounding boxes and full-lines are GT bounding boxes. Best viewed in color.

References

[1] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[2] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.

[3] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[4] Yuanzhong Liu, Zhigang Tu, Liyu Lin, Xing Xie, and Qianqing Qin. Real-time spatio-temporal action localization via learning motion representation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[5] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. *arXiv preprint arXiv:2303.15879*, 2023.