Learning Asynchronous and Sparse Human-Object Interaction in Videos - Supplementary Material

Romero Morais*, Vuong Le, Svetha Venkatesh, Truyen Tran
Applied Artificial Intelligence Institute, Deakin University, Australia
{ralmeidabaratad,vuong.le,svetha.venkatesh,truyen.tran}@deakin.edu.au

1. Introduction

In this supplementary material, we provide:

- Qualification details of methods for the *joint segmentation and labeling* task.
- ASSIGN's training loss implementation details.
- Details about the BiRNN baselines employed in the experiments.
- Frame-level micro and macro F_1 scores of ASSIGN and related methods on the CAD-120 and Bimanual Actions datasets.
- Additional qualitative segmentation and labeling comparisons between ASSIGN and related methods.
- Source code of ASSIGN.

2. Qualification for *joint segmentation and label recognition* task

As mentioned in the main manuscript, [5, 6] implicitly or explicitly used information relevant to the ground-truth segmentation during the training or testing of their models.

For the Stochastic Grammar method [5], the model requires the computation of data statistics such as length of sub-activities and object affordances, and these were computed from the whole dataset. The experimental protocol is a leave-one-subject out cross-validation, which requires such data-dependent statistics to be computed from the training folds during each round of cross-validation. The code relies on these pre-computed statistics to execute, and in the files provided by the authors in a Github issue[1], we can verify that the lengths of sub-activities and object affordances provided are in relation to the full dataset.

For the Generalized Earley Parser method [6], the authors implicitly give segmentation information about the data to their model by repeating the segment-level features provided by Koppula *et al*. [2] as frame-level features. We can confirm that by analyzing the function collate_fn_cad[2], where the length information about the segments is used in lines 40–44 to assemble the frame-level features. This means that the segmentation is implicitly input to their model, which makes their method not suitable for this task.

Table 1. $\mathrm{F}_1@k$ results for ASSIGN with and without anticipation loss on the CAD-120 dataset.

Model	Sub-activity		Object Affordance	
	F_1@0.10	F_1@0.50	F_1@0.10	F_1@0.50
ASSIGN w/o anticipation loss	86.2	71.4	91.3	80.4
ASSIGN	**88.0**	**73.8**	**92.0**	**82.4**

3. Training loss implementation details

In addition to the segmentation and recognition losses, ASSIGN has an anticipation loss. This anticipation loss is identical to the recognition loss, but predicts the label of the next segment. All ASSIGN variations were trained with this anticipation loss. We show in Table 1 the $\mathrm{F}_1@k$ scores on the CAD-120 dataset for ASSIGN with and without the anticipation loss. Similarly to previous works [4, 8], doing anticipation helps with the recognition results.

4. BiRNN baselines

We designed two baseline models for our experiments: the Independent BiRNN and the Relational BiRNN. These baselines can be seen as single-layer dense versions of ASSIGN with restricted interaction between the entities.

The Independent BiRNN is simply a BiRNN per entity (with shared parameters for entities of the same class) followed by an MLP to recognize the sub-activity (or affordance) of the entity. We call it independent because there is no message passing between the entities. More specifically, for the e-th entity at the t-th frame we compute its BiRNN

[1] https://github.com/SiyuanQi/grammar-activity-prediction/issues/2

[2] https://github.com/SiyuanQi/generalized-earley-parser/blob/master/src/python/datasets/utils.py

Table 2. *Joint segmentation and label recognition* task with no pre-segmentation. Micro and macro F_1 performance on the CAD-120 dataset. An "*" mark methods with a different experimental protocol (see Section 2).

Model	Sub-activity F_1 (%)		Object Affordance F_1 (%)	
	Micro	Macro	Micro	Macro
Independent BiRNN	58.0	54.2	83.3	73.3
rCRF [7]	68.1	61.3	81.5	77.8
KGS [2]	68.2	66.4	83.9	69.6
Relational BiRNN	70.3	67.7	81.6	66.4
ATCRF [3]	70.3	70.2	85.4	71.9
Stochastic Grammar* [5]	76.5	76.1	82.4	69.3
Gen. Earley Parser* [6]	79.4	79.7	-	-
ASSIGN	**74.8**	**73.3**	**86.9**	**79.6**

state as

$$h_t^e = \mathrm{BiRNN}\left(x_t^e, \overrightarrow{h}_{t-1}^e, \overleftarrow{h}_{t+1}^e\right), \tag{1}$$

where x_t^e is the frame-level feature for the e-th entity at the t-th frame, and $h_t^e = \left[\overrightarrow{h}_t^e, \overleftarrow{h}_t^e\right]$ is the concatenation of the forward and backward hidden states produced by the BiRNN. We recognize the label associated with the t-th frame as

$$\hat{y}_t^e = \mathrm{Softmax}\left(\alpha\left(h_t^e\right)\right), \tag{2}$$

where α is an MLP and label recognition is done frame-wise.

The Relational BiRNN is similar to the Independent BiRNN, but it includes *inter-class* messages between the entities. The messages exchanged in the Relational BiRNN are a mean-pooling of the hidden states of the entities of the sender class. We compute the model BiRNN state as

$$h_t^e = \mathrm{BiRNN}\left(x_t^e, \overrightarrow{h}_{t-1}^e, \overleftarrow{h}_{t+1}^e\right). \tag{3}$$

The message to entity e is an *inter-class* message only, and we compute it as the average of the hidden states of the entities of class c^k

$$m_t^{inter \to e} = \frac{1}{K} \sum_{c^k \neq c^e} h_t^k, \tag{4}$$

where K is the number of entities for which $c^k \neq c^e$. Finally, we recognize the label associated with the t-th frame as

$$\hat{y}_t^e = \mathrm{Softmax}\left(\alpha\left(\left[h_t^e, m_t^{inter \to e}\right]\right)\right), \tag{5}$$

where α is an MLP. Since the Bimanual Actions dataset [1] is not annotated with object affordances, we do not include the human → object message in its Relational BiRNN.

5. Frame-level micro and macro F_1 results

We report in the main manuscript the $F_1@k$ metric for the *joint segmentation and label recognition* task. To provide a complete analysis of ASSIGN and related methods, we include here the micro and macro F_1 results on both CAD-120 (Table 2) and Bimanual Actions (Table 3) datasets.

Table 3. *Joint segmentation and label recognition* task with no pre-segmentation. Micro and macro F_1 performance on the Bimanual Actions dataset.

Model	Sub-activity F_1 (%)	
	Micro	Macro
Dreher *et al.* [1]	64.0	63.0
Independent BiRNN	76.7	74.8
Relational BiRNN	80.3	77.5
ASSIGN	**82.3**	**79.5**

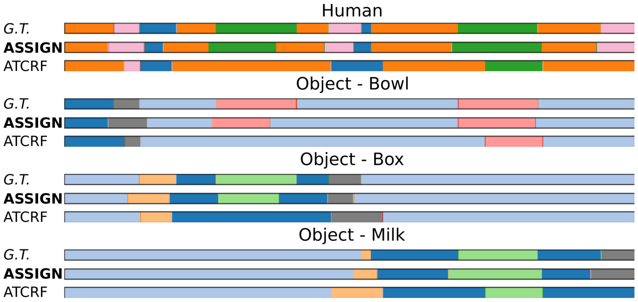

Figure 1. Segmentation and labeling results of ASSIGN and ATCRF on the CAD-120 dataset for a *making cereal* activity. In this example, ATCRF under-segments the human and the box by skipping a few segments and merging adjacent labels. Sub-activities: ■ *moving*, ■ *placing*, ■ *reaching*, and ■ *pouring*. Affordances: ■ *movable*, ■ *placeable*, ■ *stationary*, ■ *pour-to*, ■ *reachable*, and ■ *pourable*.

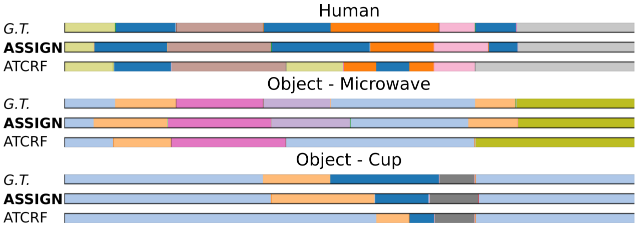

Figure 2. Segmentation and labeling results of ASSIGN and ATCRF on the CAD-120 dataset for a *taking food* activity. In this example, ATCRF over-segments the *moving* (■) sub-activity and ignores the last *reaching* (■). ASSIGN correctly segments and label the entities. Sub-activities: ■ *null*, ■ *reaching*, ■ *opening*, ■ *moving*, ■ *placing*, and ■ *closing*. Affordances: ■ *stationary*, ■ *reachable*, ■ *openable*, ■ *containable*, ■ *closeable*, ■ *movable*, and ■ *placeable*.

For both datasets, ASSIGN attains superior performance when compared to related methods and baselines. It is interesting to note that the Relational BiRNN has lower macro F_1 scores than ATCRF, even though it has higher $F_1@k$ scores. This relates to the discussion in the main manuscript that frame-level micro/macro scores are not the most appropriate metric when dealing with *joint segmentation and recognition* problems. For example, in a situation where a method over-segments a long segment, this might not reflect

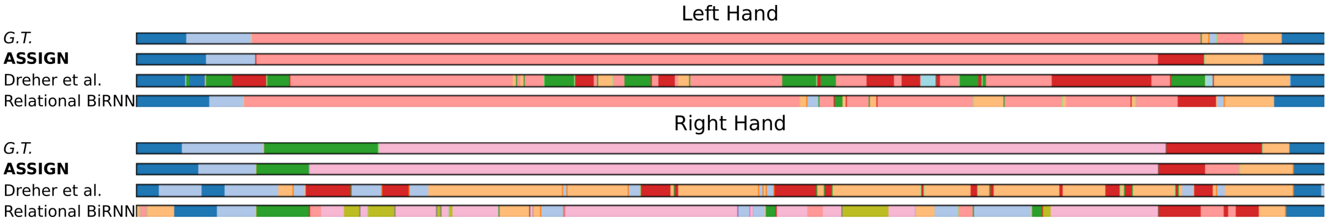

Figure 3. Segmentation and labeling results on the Bimanual Actions dataset for a *sawing* task. The main difficulty related methods have is to handle long actions, such as the left hand *hold* (■). Legend: ■ *idle*, ■ *approach*, ■ *hold*, ■ *retreat*, ■ *place*, ■ *lift*, and ■ *saw*.

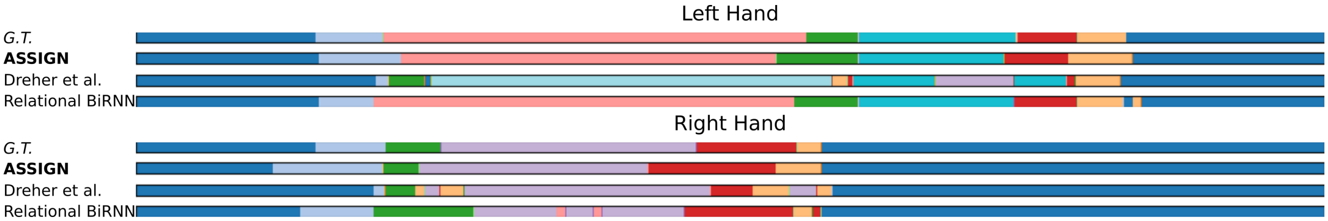

Figure 4. Segmentation and labeling results on the Bimanual Actions dataset for a *pouring* task. In this example, the Relational BiRNN has results comparable to ASSIGN but still makes several over-segmentation mistakes in both hands, such as the right hand *pour* (■). Legend:■ *idle*, ■ *approach*, ■ *hold*, ■ *lift*, ■ *drinking*, ■ *place*, ■ *retreat*, and ■ *pour*.

badly on the frame-level metrics but it will reflect badly on the $\mathrm{F}_1@k$ metric since the model effectively splits the long segment into many short segments.

As discussed in Section 2, the Stochastic Grammar [5] and Generalized Earley Parser [6] methods do not fully qualify for the *joint segmentation and label recognition* task, but for completeness we include their results in Table 2 and discuss their merits. As we can see, both methods attain high results due to their ability to regulate and generalize with the use of explicit grammar rules imposed on their generative models. ASSIGN, on the other hand, uses the activity structure as inductive biases to guide a discriminative data-driven learning. Although less regulated than the grammar-based methods, ASSIGN's approach is less sensitive to noise and more scalable with the size of the problem. In terms of segmentation functionality, ASSIGN does segmentation as a probabilistic prediction based on the observe data and in tandem with the labeling task. Such a statistic-based prediction is harder to control than a rule-guided generator and may make mistakes on patterns less observed in the training data. ASSIGN and grammar-based methods are complimentary and can also be combined for further expressiveness.

6. Segmentation and labeling extra qualitative comparisons

We further illustrate the segmentation and labeling results of ASSIGN by showing some more qualitative comparisons between ASSIGN and related methods.

For the CAD-120 dataset, we show a *making cereal* and a *taking food* activities in Figures 1 and 2, respectively. For *making cereal*, we observe that ATCRF under-segments the entities, which can happen to their model whenever their ensembling strategy agrees on a wrong label. For the *taking food* activity we observe a mixed behavior: ATCRF over-segments halfway through the video, the *moving (*■*)* sub-activity, and under-segments later the *closing (*■*)* sub-activity. For both scenarios and entities in them, ASSIGN correctly segments and label the segments.

For the Bimanual Actions dataset, we show a *sawing* and a *pouring* activities in Figures 3 and 4, respectively. For both activities, we observe that the biggest hurdle for Dreher *et al*. [1] and the Relational BiRNN is sustaining the prediction for long actions, which leads them to over-segmentation issues. For example, the long *hold (*■*)* and the long *saw (*■*)*, in Figure 3, are heavily over-segmented by them. ASSIGN, on the other hand, has no issues with that.

7. Source code

Source code, data, and pre-trained models are available[3].

References

[1] Christian R G Dreher, Mirko Wächter, and Tamim Asfour. Learning Object-Action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters*, 5(1):187–194, January 2020. 4, 3, 6

[3] https://github.com/RomeroBarata/human_object_interaction

[2] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from RGB-D videos. *The International journal of robotics research*, 32(8):951–970, July 2013. 2, 2

[3] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, January 2016. 2

[4] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11988–11996. IEEE, June 2019. 3

[5] Siyuan Qi, Siyuan Huang, Ping Wei, and Song-Chun Zhu. Predicting human activities using stochastic grammar. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1173–1181, October 2017. 2, 2, 5

[6] Siyuan Qi, Baoxiong Jia, and Song-Chun Zhu. Generalized earley parser: Bridging symbolic grammars and sequence data for future prediction. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4171–4179, Stockholmsmässan, Stockholm Sweden, 2018. PMLR. 2, 2, 5

[7] Ozan Sener and Ashutosh Saxena. rCRF: Recursive belief estimation over CRFs in RGB-D activity videos. In *Robotics: Science and Systems XI*. Robotics: Science and Systems Foundation, July 2015. 2

[8] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 843–852. PMLR, July 2015. 3