Supplementary Materials: Empowering Relational Network by Self-Attention Augmented Conditional Random Fields for Group Activity Recognition

Rizard Renanda Adhi Pramono[0000−0002−6668−5167], Yie Tarng Chen[0000−0002−7221−1603], and Wen Hsien Fang[0000−0001−6402−2688]

National Taiwan University of Science and Technology
{d10702801,ytchen,whf}@mail.ntust.edu.tw

Some supplementary materials are provided to further validate the proposed approach. The supplementary materials include:

1. the experimental details of our approach, which have not been included in our manuscript, in Sec. 1;
2. more experimental results on the New Collective Activity dataset and the comparison with the state-of-the-art works in Sec. 2.
3. the confusion matrices for group activity recognition on Volleyball, Collective Activity, and New Collective Activity are provided in Sec. 3;
4. some qualitative results of the group activity recognition by our approach in Sec. 4.

1 Experimental Settings

For the Volleyball dataset, we use the bounding box annotation provided in the middle frame in each video clip. Same as [1], the positions of all players in the other frames are obtained by an appearance-based tracker [2]. Following [1], the group activity and individual action classification results for every video are based on the ten middle frames. For the Collective Activity dataset, we use the bounding box annotation and the tracklets provided in [3]. For the intermittently observable actors, we use the tracker information provided by the datasets to compute temporal self-attention and use the duplicates of those from the previous frames in case of missed detection. If there are fewer actors than N throughout the temporal window, we use zero vectors to fill the rest of actors' features $\mathbf{B}$. In fact, instead of using a fixed number of actors, N, and a fixed number of temporal window for all actors, M, our network can use different number of actors in each frame and different temporal window for each actor by calculating the temporal and spatial self-attention node-by-node. However, it does not leverage parallel computation for faster inference and training convergence.

2 A New Dataset

Here, we also assess our approach on the less well-known New Collective Activity dataset [3], which is comprised of 32 videos. There are 6 group activity categories: gathering, talking, dismissal, walking together, chasing and queuing, and 3 individual action labels: walking, standing still, and running. This dataset provides bounding box annotation and trajectory data. The simulations mainly follow the protocols and evaluation metrics provided by the New Collective Activity dataset [3, 4].

We compare the proposed methods with the state-of-the-art works which reported their performance on this dataset, including Recurrent Modelling [5], HiRF [6], and StagNet [4], as shown in Table 1, from which we can see that Recurrent Modelling [5] is the worst as it does not consider the spatial relational structure of actors. HiRF [6] achieves better performance by constructing a hierarchical modelling of group activities based on several sub-activities. StagNet [4] improves the performance by using a structural recurrent neural network and an attention mechanism to model the spatial relationships of actors for more accurate group activity recognition. Our approach outperforms all of the aforementioned methods by leveraging the spatial relational contexts and temporal evolution of actors at various distances using self-attention augmented CRF. Moreover, it exploits the temporal dependency of the relational context and scene information using the bidirectional UTE.

3 Confusion Matrices

In this part, we provide the confusion matrices for group activity recognition with our approach on three datasets. We first consider Volleyball, as shown in Fig. 1, from which we can observe that the accuracy of our approach can achieve at least 89% for every class. Our approach can distinguish well between the group activities by the right team such as 'Right setting' and the left team such as 'Left setting'. This is because our self-attention augmented CRF considers the locality of the actors when modelling their relationships. Most of the errors come from misclassification of 'right setting' as 'right passing' or 'left setting' as 'left passing', which share similar actors' positions and appearances. However, our approach can still achieve relatively high accuracy in these hard cases, as our self-attention augmented CRF also models the temporal evolution of actors and employs the bidirectional UTE to leverage the temporal dependency of the scene and relational context information, which are crucial in classifying activities with similar spatial positions and appearances.

Next, the confusion matrix for group activity recognition on Collective Activity is shown in Fig. 2, from which we can see that our method can achieve more than 93% of accuracy, except for 'Waiting' and 'Crossing'. Since the pairwise energies of our self-attention augmented CRF models the relationships of actors based on multiple cliques with different locality scales, our approach can still differentiate 'Crossing' from 'Walking' well, which are relatively difficult

Table 1: Comparison with the state-of-the-art methods on New Collective Activity. The best results are bold-faced.

Method	Backbone	Accuracy
		Group Activity
Recurrent Modelling (RGB + Flow) [5]	AlexNet + GoogleNet	85.2
HiRF [6]	Deformable Part based Model	87.3
StagNet w/ Attention (RGB) [4]	VGG16	90.2
Ours (RGB + Flow)	**I3D + FPN**	**93.4**

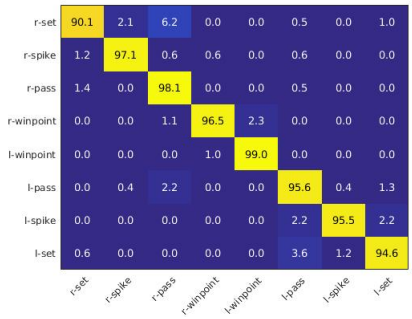


Fig. 1: Confusion matrix on the Volleyball dataset.

to distinguish as both classes have similar actors' movements and appearances. This is because, aside from modelling the spatial relationships of actors, we also model the temporal evolution of actors and simultaneously aggregate the relational contexts of actors and the scene information using the bidirectional UTE. However, 'Waiting' can be easily confused with 'Crossing' and 'Walking' as they have similar appearances and sometimes it is concurrent with those classes.

Lastly, we consider the confusion matrix on New Collective Activity, as shown in Fig. 3, from which we can see that our method can achieve more than 95% of accuracy, except for 'Gathering' and 'Dismissal'. This is because 'Gathering' can be easily confused with 'Talking' as these group activity categories have similar appearances in crowded scenes. Also, 'Dismissal' is usually concurrent with 'Talking' or 'Walking'. On the other hand, our approach can distinguish well between 'Walking' and 'Chasing' as we model the temporal evolution of the actors and their spatial relational contexts along with the scene information using the self-attention augmented CRF and bidirectional UTE.

4 Qualitative Analysis

We show some qualitative results of our group activity recognition approach on the Volleyball and Collective Activity datasets. First, some qualitative results on Volleyball are shown in Fig. 4, from which we can see that our framework can distinguish between the same group activities performed by two different teams such as 'Left setting' and 'Right setting', as shown in Fig. 4 (a) and (b).

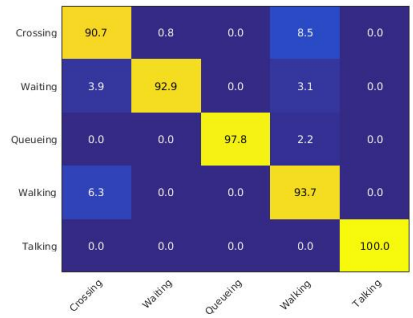


Fig. 2: Confusion matrix on the Collective Activity dataset.

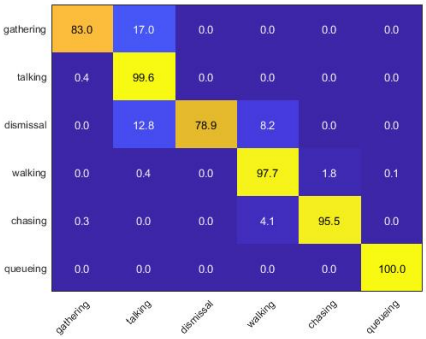


Fig. 3: Confusion matrix on the New Collective Activity dataset.

This is because our self-attention augmented CRF considers the relationships of actors with different scales of locality. Our approach can also differentiate group activities with similar actors' appearances and cues such as 'Left setting' and 'Left passing' because we also model the temporal evolution of actors and provide their spatial relational contexts to obtain a more accurate classification.

Next, some qualitative results on Collective Activity are also furnished in Fig. 5, from which we can see that our approach can differentiate 'Crossing' from 'Waiting' or 'Walking', all of which share similar appearances and movements. This is because our approach incorporates self-attention augmented CRF to model the relational contexts of actors while modelling their temporal dependency.

Frame 17 Frame 21 Frame 25

(a)

Frame 17 Frame 21 Frame 25

(b)

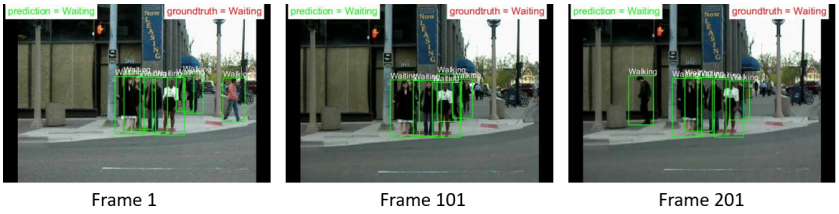
Frame 17 Frame 21 Frame 25

(c)

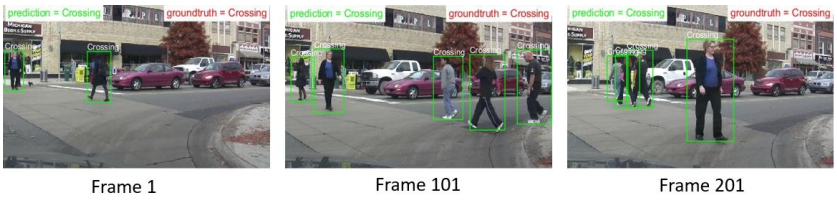
Frame 17 Frame 21 Frame 25

(d)

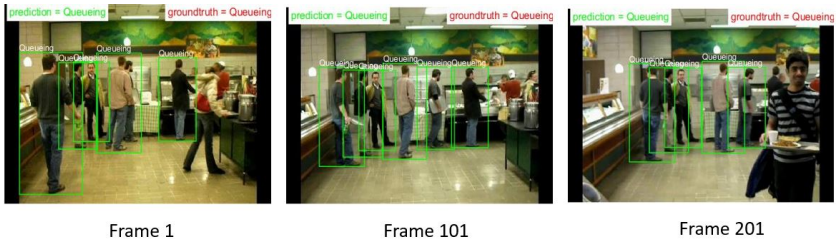
Fig. 4: Some successful group activity recognition results by our approach on Volleyball; the group activity prediction and the ground truth are in the top left and right of every image, respectively; the individual action classification is in white while the incorrect classification is in red boxes.



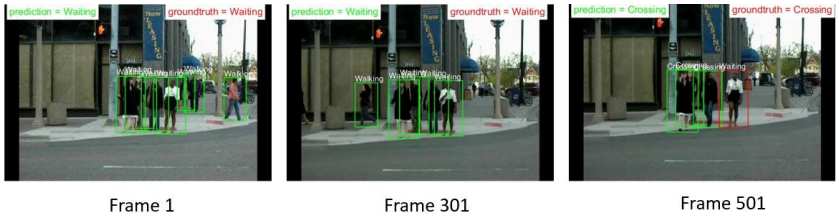
(a)



(b)



(c)



(d)

Fig. 5: Some successful group activity recognition results by our approach on Collective Activity; the group activity prediction and the ground truth are in the top left and right of every image, respectively; the individual action classification is in white while the incorrect classification is in red boxes.

References

1. Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016.
2. Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.
3. Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the European Conference on Computer Vision*, pages 215–230, 2012.
4. Mengshi Qi, Yunhong Wang, Jie Qin, Annan Li, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity and individual action recognition. *accepted by IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
5. Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3048–3056, 2017.
6. Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proceedings of the European Conference on Computer Vision*, pages 572–585, 2014.