

CTRN: Class Temporal Relational Network for Action Detection

Rui Dai¹

rui.dai@inria.fr

Srijan Das²

srijan.das@stonybrook.edu

François Brémond¹

francois.bremond@inria.fr

¹ Inria, Université Côte d'Azur,
France

² Stony Brook University,
USA

Overview

In this supplementary material, we provide more details regarding the model structure in Sec. 1, ablation studies in Sec. 2 and dataset description in Sec. 3 to complement the descriptions from the main paper.

1 Model Structure

To better illustrate our model, we further introduce the Representation Transform Module (RTM) and Temporal Convolutional Network (TCN) in this section.

1.1 Representation Transform Module (RTM)

RTM is one of the components in CTRN, which is used to filter the action-class specific feature from the I3D feature vector. This component is essential for CTM and G-Classifier to perform their functionalities. As a complementary to the main paper, here we provide the computation flow of RTM in Fig. 1. The input I3D feature is in size of $T \times D_1$. The MLP [1] is used to do the linear transformation. In practice, we set MLP as a single linear layer. Input I3D input feature is inflated to a new dimension C , for each I3D feature pertaining to class, we utilize a MLP+ReLU to filter the class-specific feature. A dropout layer is further added to prevent the over-fitting issue. The output feature of RTM is in size of $T \times C \times D_2$.

1.2 Temporal Convolutional Network (TCN)

TCN [2] layer is the 1-dimensional convolutional layer across the temporal dimension, which is popularly used in temporal modelling tasks. Similar to the 2D ConvNet, the temporal receptive field increases with the number of stacked TCN layers. As shown in Fig. 2, we present the receptive field of TCN. For simplicity, the kernel size of TCN is set to 3 and the number of layers is 3 in this figure. We find that the receptive field of 1 layer TCN is 3 time steps (in red). When stacking 3 TCN layers, the receptive field increases to 7 time steps (in yellow). In this work, multiple C-GCNs + TCNs are stacked in CTM, in which TCNs are

used to adjust the temporal scale. This structure allows the C-GCN to model short and long action relations by focusing on video features at the level of low and high temporal receptive fields.

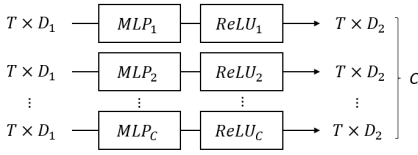


Figure 1: RTM structure.

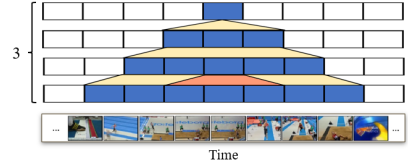


Figure 2: TCN Receptive field.

2 More Ablations

In this section, we provide more ablation results of our model. This includes scrutinizing the optimal channel size, number of blocks, and design choice of adjacency matrix of the proposed model.

2.1 Channel Size

We first analyse the impact of the channel size D_2 on the CTRN. In the representation transform module, the MLP is used to filter the class-specific feature by projecting the feature channels from D_1 to D_2 . The projection directly affects the quality of the class-specific feature. Here we compare four different dimensions for D_2 : 32, 64, 128 and 256 in Table 1. We find that increasing the channel size improves the performance. However, the improvement is marginal after 64 while adding a large number of parameters. To compromise both performance and model efficiency, we choose 64 as the class-specific feature channel size (D_2) in this work.

Table 1: Study on Channel Size D_2 . We evaluate on Charades dataset for action detection using only RGB.

Channel Size D_2	32	64	128	256
Performance (%)	23.0	25.3	25.4	25.6
Parameters (M)	5.6	10.9	22.1	46.1

2.2 Number of Blocks

We then explore the impact of the number of blocks (L) of CTM in CTRN. As mentioned in the proposed method, TCN is used to aggregate the temporal information. Thus, with more blocks, CTRN can model high level temporal information while expanding the scale across time for very long videos. Table 2 shows the results on Charades with different blocks, we find that CTRN achieves similar performance for 5 and 6 blocks. Thus, 5-block is sufficient for encoding the temporal information in complex untrimmed videos.

Table 2: Study on number of blocks L of CTM in CTRN. We evaluate on Charades dataset for action detection using only RGB.

#Blocks	4	5	6
Performance (%)	24.2	25.3	25.3

2.3 Adjacency Matrix A'_C

As mentioned in Sec.3.3.2 in the main paper, C-GCN’s graph is composed of a learnable adjacency matrix A_C and an attention mask which is superimposed on the former. Here we further analyze that both the components are complementary. In Table 3, we find that the performance declines in the absence of either A_C or the attention mask in C-GCN, reflecting both components are crucial for learning the the graph structure.

Table 3: Study on adjacency matrix in C-GCN. We evaluate on Charades dataset for action detection using only RGB. * indicates the results of CTM w/o C-GCN but only a TCN.

Adjacency Matrix A_C	Attention Mask	mAP (%)
×	×	21.4*
✓	×	24.3
×	✓	24.5
✓	✓	25.3

2.4 Modalities

As mentioned in Sec.3.2 in the main paper, our model can be used with both RGB and Optical Flow (OF). Here, we provide the results with RGB and OF. For a fair comparison, similar to the previous works [13, 14], we fuse the two modalities through a late-fusion of the logits. From Table 4, we find that: (1) For sport actions in MultiTHUMOS, Flow stream yields better performance than RGB stream (+4.3%). (2) For object-based actions with low motion in Charades, RGB stream achieves better performance (+3.8% w.r.t. Flow stream), which indicates that RGB can better model the object appearance information, especially for low motion frames.

Table 4: Study on RGB and optical flow. RGB+OF indicates the late fusion.

Modalities	RGB	OF	RGB+OF
Charades	25.3	20.3	27.8
MultiTHUMOS	44.0	47.5	51.2

3 Dataset Description

In this section, we describe the three densely annotated datasets used to evaluate our method. **Charades** [15] was recorded by hundreds of people in their private homes. This dataset consists of 9848 videos across 157 actions. The actions are mainly object-based daily living actions performed at home. Each video is about 30 seconds containing complex co-occurring actions. In our experiments, we follow the original Charades settings for action detection [15] (i.e. Charades v1 localize evaluation). The performances are measured in terms of mAP by evaluating per-frame prediction as the original paper [15] and other state-of-the-art methods [13, 14, 16].

Toyota Smarthome Untrimmed (TSU) [8] is a real-world action detection dataset, which is the untrimmed version of [8]. This dataset consists of 536 long videos (about 20 mins/video) recorded by 7 cameras with 51 densely annotated action classes. Besides long video duration, this dataset contains actions with high intra-class temporal variance. As a result, handling temporal information is critical to achieve good detection performance on this dataset. We evaluate this dataset with frame-based mAP.

MultiTHUMOS [19]: We conducted our experiments on MultiTHUMOS [19] datasets. MultiTHUMOS is an enhanced version of the THUMOS14 [10] dataset, where videos are densely annotated. The dataset consists of 65 action classes, compared to 20 in THUMOS14, and contains on average 10.5 action classes per video and 1.5 labels per frame and up to 25 different action labels in each video. THUMOS14 and MultiTHUMOS consists of YouTube videos of various sport actions like baseball games, cliff diving. We evaluate this dataset similar to the original setting, i.e. frame-based mAP.

In the main paper, we have provided a comparison of the proposed approach to the state-of-the-art methods. Here in Tables 5, 6 and 7, we provide an extension with more comparisons to the state-of-the-art methods. These tables show that our approach outperforms all the state-of-the-art methods using RGB and also the ones using RGB + OF.

Table 5: Frame-based mAP on Charades, evaluated with the localization setting. OF+RGB indicates late fusion.

	Test modality	mAP
R-C3D [13]	RGB	12.7
Asynchronous Temporal Fields [13]	RGB + OF	12.8
I3D [13]	RGB	15.6
I3D + 3 temporal conv.layers [13]	RGB + OF	17.5
TAN [8]	RGB + OF	17.6
I3D + WSGN (supervised) [8]	RGB	18.7
I3D + Stacked-STGCN [8]	RGB	19.1
I3D + Super event [13]	RGB + OF	19.4
I3D + 3 TGMs [13]	RGB	18.9
I3D + 3 TGMs + Super event [13]	RGB + OF	22.3
I3D + MLAD [13]	RGB	18.4
I3D + MLAD [13]	RGB+OF	22.9
I3D +CTRN (Ours)	RGB	25.3
I3D +CTRN (Ours)	RGB+OF	27.8

Table 6: Frame-based mAP on TSU dataset for CS protocol. Only RGB modality is used at inference time.

	Frame-mAP (CS)
Bottleneck [10]	15.7
NL block [11]	16.8
Super event [12]	17.2
LSTM [13]	22.6
Bi-LSTM [9]	24.5
Dilated-TCN [14]	25.1
MS-TCN [9]	25.9
TGM [15]	26.7
CTRN (Ours)	33.5

Table 7: Performance of the state-of-the-art methods and our approach on MultiTHUMOS. I3D model is two-stream, using both RGB and optical flow input. Note: cited papers may not be the original paper but the one providing this mAP results. For a fair comparison, we utilize the late fusion as OF+RGB.

	mAP
Two-stream [16]	27.6
Two-stream+LSTM [17]	28.1
Multi-LSTM [18]	29.6
SSN [19]	30.3
I3D [20]	29.7
I3D + LSTM [21]	29.9
I3D + temporal pyramid [22]	31.2
TAN [9]	33.3
I3D + Dilated-TCN* [23]	43.2
I3D + 3 TGMs [24]	44.3
I3D + MS-TCN* [9]	45.3
I3D + 3 TGMs + Super event [25]	46.4
I3D + MLAD [26]	49.6
I3D + CTRN	51.2

References

- [1] Luis B Almeida. C1. 2 multilayer perceptrons. *Handbook of Neural Computation C*, 1, 1997.
- [2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.
- [3] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *arXiv preprint arXiv:2010.14982*, 2020.
- [4] Xiyang Dai, Bharat Singh, Joe Yue-Hei Ng, and Larry Davis. Tan: Temporal aggregation network for dense multi-label action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 151–160. IEEE, 2019.

- [5] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [6] Yazan Abu Farha and Jurgen Gall. Ms-ten: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.
- [7] Basura Fernando, Cheston Tan, and Hakan Bilen. Weakly supervised gaussian networks for action detection. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [8] Pallabi Ghosh, Yi Yao, Larry S Davis, and Ajay Divakaran. Stacked spatio-temporal graph convolutional networks for action segmentation. *arXiv preprint arXiv:1811.10575*, 2018.
- [9] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.
- [10] Yu-Gang. Jiang, Jingen Liu, Amir Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS challenge: Action recognition with a large number of classes. <http://crcv.ucf.edu/THUMOS14/>, 2014.
- [11] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.
- [12] Behrooz Mahasseni and Sinisa Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3054–3062, 2016.
- [13] AJ Piergiovanni and Michael S Ryoo. Learning latent super-events to detect multiple activities in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [14] AJ Piergiovanni and Michael S Ryoo. Temporal gaussian mixture layer for videos. *International Conference on Machine Learning (ICML)*, 2019.
- [15] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017.
- [16] Praveen Tirupattur, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Modeling multi-label action dependencies for temporal action localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [17] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

- [18] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.
- [19] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018.
- [20] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.