Exploring Temporal Context and Human Movement Dynamics for Online Action Detection in Videos

Vasiliki I. Vasileiou, Nikolaos Kardaris and Petros Maragos

National Technical University of Athens, School of ECE
Computer Vision, Speech Communication and Signal Processing Group
silavassiliou2@gmail.com, nkardaris@mail.ntua.gr, maragos@cs.ntua.gr

29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

Table of Contents

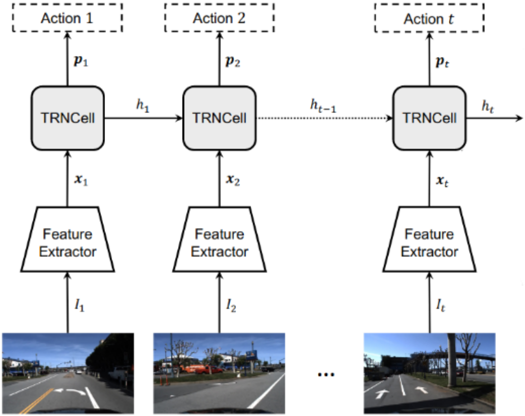
Introduction & Background

Human Action Recognition: It involves predicting the movement of a person based on sensor data and traditionally involves deep domain expertise and methods from signal processing to correctly engineer features from the raw data in order to fit a machine learning model.

- **Offline Action Recognition:** Attempt to identify the actions occurring in a short video clip given a-priori the information of future frames.

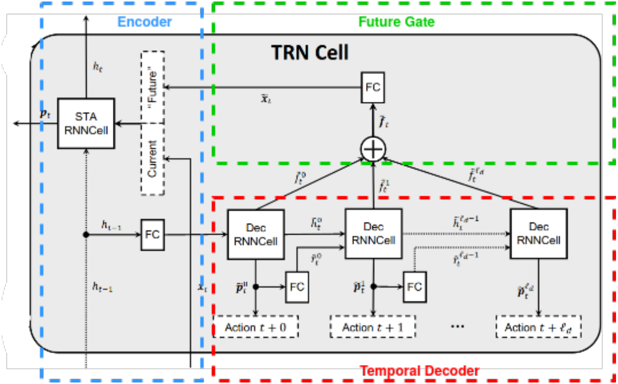
- **Online Action Recognition:** Attempt to identify the actions, performed in each frame, as soon as it arrives, without taking into account the future context.

Temporal Recurrent Networks - TRNs I



The TRN cell functions in a manner similar to any RNN cell with the only difference being the use of both current and future information generated by anticipation.

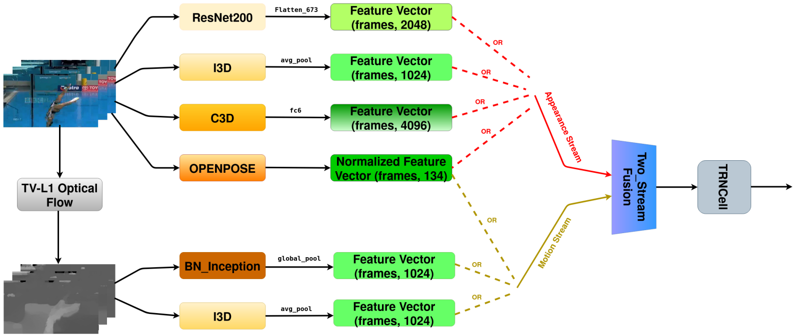
Temporal Recurrent Networks - TRNs II



- **Temporal Decoder:** Learns a feature representation and predicts actions for the future sequence.
- **Future Gate:** Embeds a hidden state vector as future context.
- **Encoder:** Estimates the action occurring in the current frame.

Architectures

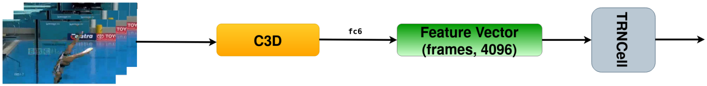
Two-Stream TRN model



- We performed in-house testing for the baseline TRN [1].
- Inspired by the two-stream baseline model baseline with the former stream consisting of the appearance features and the latter of the motion features.
- We experimented by extracting I3D features, which are low-level spatial features.

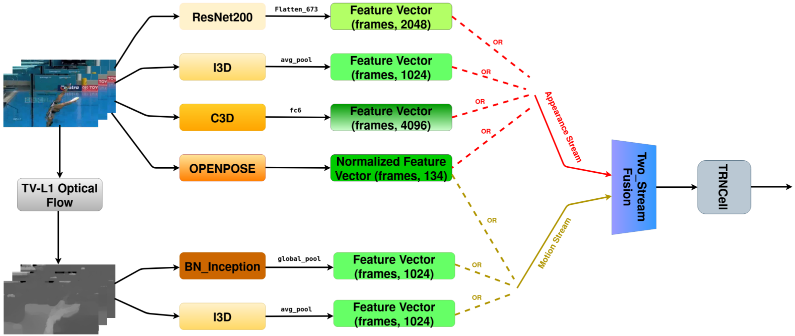
[1] M.Xu et al, in Proc. ICCV 2019

One-Stream TRN model



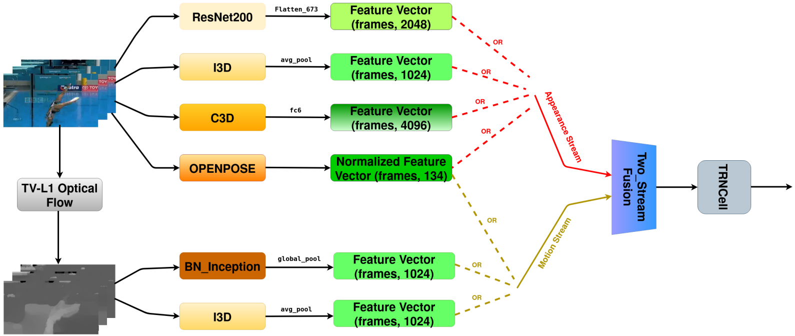
- We experimented with C3D features, being a very generic video feature representation.
- We turned the two-stream model into a one-stream model as the C3D modules can extract both spatial and temporal components.

Two-Stream TRN model



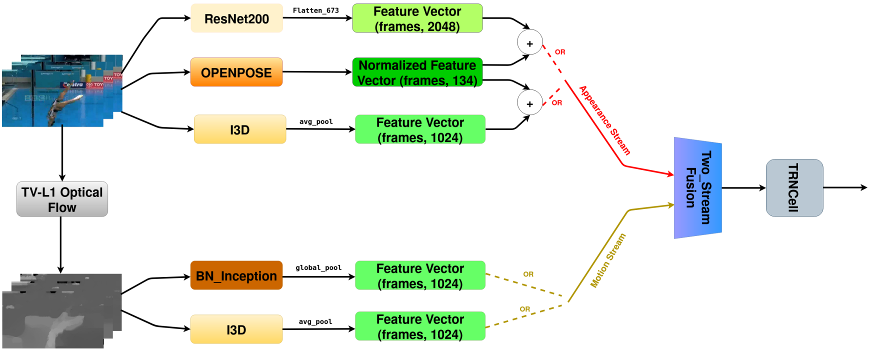
- Skeleton joint coordinates are of high precision and can accurately represent the temporal dynamics of actions.
- We experimented with 2D skeletons extracted from OpenPose, over the baseline RBG and Optical Flow features.

Two-Stream TRN model



- Skeleton features are primarily motion features. So we arranged the C3D features in the appearance stream and the pose features in the motion stream.
- We arranged the I3D RGB data in the appearance stream and the OpenPose data in the motion stream.

Fused Two-Stream TRN Model

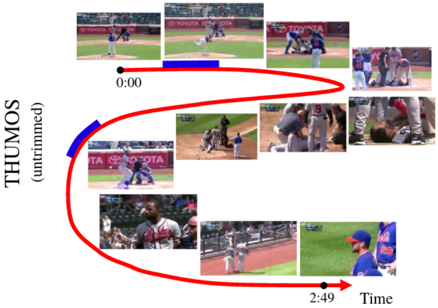


- Skeleton features sufficiently represent the temporal dynamics but the appearance or motion information is still missing.
- We attempted to combine each of our two-stream models - baseline and I3D - with the information from the skeleton.

Experimental Setup

THUMOS'14 dataset [3]: Long and untrimmed videos from various sports events. Annotated with 20 actions increased by an ambiguous class and a background class.

- **Training Set**: 200 untrimmed videos of sports events
- **Testing Set**: 213 untrimmed videos of sports events



Dataset & Tools II

Openpose [4]: 2D models are used, each keypoint consists of two spatial variables, its coordinates and a confidence parameter.

- **Human Pose**: 25 keypoints for pose/foot estimation and 2×21 keypoints for hand estimation.
- **Normalization**: We define the middle of the pelvis as the center of our coordinates and normalize with respect to the distance between the pelvis and the shoulders (average height) [5].

TV-L1 [6]: The optical flow algorithm was used to extract the optical flow frames through the Dense-Flow tool.

[3] Y.-G. Jiang et al, in Proc. ICCV 2013 [4] Z. Cao et al, in Proc. TPAMI 2019

[5]A. Shahroudy et al, in Proc CVPR 2016 [6]J. Sanchez et al, in Proc IPOL 2013

Experimental Setup & Evaluation Protocol

- **Hardware**: Nvidia GeForce RTX2080 Ti GPUs.
- **Optimizer**: Adam optimizer with learning rate and weight decay parameters set to 5×10^{-4}.
- **Loss Function**: Cross Entropy Loss.
- **Batch Size**: 2
- **Input sequence length**: 64
- **Decoder Steps**: 8
- **Frequency Rate**: We extracted video frames at 30 fps.
- **Chunk Size**: 6 & 16 frames in line with the examined set of experiment.
- **Evaluation Protocol**: We used the per-frame mean average precision (mAP) metric.

Results & Discussion

Baseline & OpenPose TRN

Method	Features Chunk size = 6 frames	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	0.50s	0.75s	1.00s	1.25s	1.50s	1.75s	2.00s	Avg
Baseline [1]	RGB – Flow	25.93	**26.15**	25.89	25.79	25.73	25.66	**25.68**	**25.66**	**25.57**	**25.77**
Ours	{RGB + OpenPose} – Flow	24.25	23.11	25.63	**26.72**	26.18	25.57	24.94	24.40	23.94	25.06
Ours	RGB – OpenPose	**37.57**	25.54	**25.93**	26.44	**26.60**	**26.28**	25.57	24.75	24.00	25.64
Ours	OpenPose – Flow	36.30	21.77	22.59	23.57	23.19	22.28	21.30	20.49	19.83	21.88

- Chunk size has been set to 6.
- Baseline exhibit the highest accuracy of 25.77% for the precision task and one of the lowest, approximately 25.93% for the classification task.
- The replacement of flow information with OpenPose features gives an increase of 11 points approximately reaching the 37.57%.
- Replacing or enhancing the RGB information with pose features does not provide any further improvement.

[1] It was re-implemented with batch_size 2 so we have a fair comparison, which dropped the accuracy to 25.93%. It was the state_of_the_ art with an accuracy of 47.2%.

C3D & OpenPose TRN

Method	Features Chunk size = 16 frames	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	0.50s	0.75s	1.00s	1.25s	1.50s	1.75s	2.00s	Avg
Ours	C3D (One-Stream)	35.43	**34.34**	**31.05**	28.22	26.46	25.37	**24.75**	**24.39**	**24.22**	**27.35**
Ours	{C3D (RBG)} – OpenPose	**36.44**	32.98	30.56	**28.37**	**26.61**	**25.38**	24.54	23.78	23.22	26.93

- Chunk size has been set to 16.
- Adding a second stream of human pose features the detection accuracy increased to 36.44% while the anticipation accuracy decreased to 26.93%.

By comparing this table to the previous one:

- Although in C3D models we observe larger anticipation accuracy, the action detection accuracy does not exceed that of the models of the previous table.

I3D & OpenPose TRN

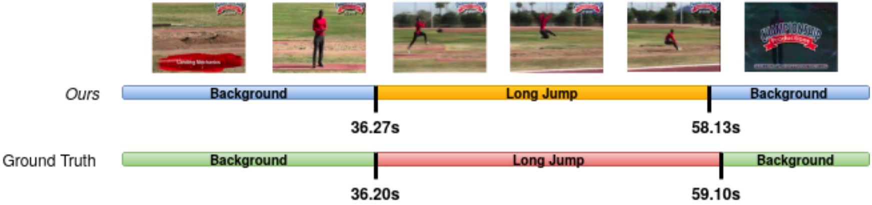
Method	Features Chunk size = 16 frames	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	0.50s	0.75s	1.00s	1.25s	1.50s	1.75s	2.00s	Avg
Ours	**I3D**	**55.25**	**52.57**	**46.69**	**41.94**	**38.39**	**35.90**	**34.22**	**33.00**	**32.08**	**39.35**
Ours	{I3D (RGB) + OpenPose} – {I3D (Flow)}	49.21	46.65	40.78	36.42	33.19	30.90	29.42	28.43	27.71	34.19
Ours	{I3D (RGB)} – OpenPose	47.43	44.59	40.08	36.77	34.24	32.37	31.29	30.56	30.06	35.00
Ours	{I3D (RGB)} – {I3D (Flow) + OpenPose}	44.47	29.55	31.92	29.62	27.21	25.63	24.78	24.20	23.68	27.07

- Chunk size has been set to 16
- Both the simple I3D model and its modifications show much better performance with the greatest reaching reaching 39.35% in the anticipation phase and 55.25% in the detection phase.
- In contrast with the previous cases, here the pose features limited its effectiveness.

Results Comparisons

Method	Features	Encoder	Decoder - Time predicted into the future (seconds)								
			0.25s	0.50s	0.75s	1.00s	1.25s	1.50s	1.75s	2.00s	Avg
Baseline	RGB – Flow	25.93	**26.15**	25.89	25.79	25.73	25.66	**25.68**	**25.66**	**25.57**	**25.77**
Ours	{RGB + OpenPose} – Flow	24.25	23.11	25.63	**26.72**	26.18	25.57	24.94	24.40	23.94	25.06
Ours	RGB – OpenPose	**37.57**	25.54	**25.93**	26.44	**26.60**	**26.28**	25.57	24.75	24.00	25.64
Ours	OpenPose – Flow	36.30	21.77	22.59	23.57	23.19	22.28	21.30	20.49	19.83	21.88
Ours	C3D (One-Stream)	35.43	**34.34**	**31.05**	28.22	26.46	25.37	**24.75**	**24.39**	**24.22**	**27.35**
Ours	{C3D (RBG)} – OpenPose	**36.44**	32.98	30.56	**28.37**	**26.61**	**25.38**	24.54	23.78	23.22	26.93
Ours	**I3D**	**55.25**	**52.57**	**46.69**	**41.94**	**38.39**	**35.90**	**34.22**	**33.00**	**32.08**	**39.35**
Ours	{I3D (RGB) + OpenPose} – {I3D (Flow)}	49.21	46.65	40.78	36.42	33.19	30.90	29.42	28.43	27.71	34.19
Ours	{I3D (RGB)} – OpenPose	47.43	44.59	40.08	36.77	34.24	32.37	31.29	30.56	30.06	35.00
Ours	{I3D (RGB)} – {I3D (Flow) + OpenPose}	44.47	29.55	31.92	29.62	27.21	25.63	24.78	24.20	23.68	27.07

Results Visualization



Contributions

Contributions & Future Work

- Explored several ways to improve online action detection, building upon Temporal Recurrent Networks.
- Highlighted the value of temporal context and human pose as useful cues for localizing action in time.
- Most of our models outperform the original TRN method.
- **Future Work:** We believe that the use of different models for anticipation and recognition could benefit the task of online action detection.

Thank You