Deep Learning + Reinforcement Learning
SUMMER SCHOOL

Recognition & Detection for Real-World Actions of Daily Living

Rui Dai, STARS Team, INRIA

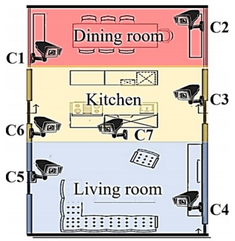
Joint work with: Srijan Das, Francois Bremond, Luca Minciullo, Lorenzo Garattoni and Gianpiero Francesca

Abstract

- We first introduce a large real-world video dataset for activities of daily living: **Toyota Smathome**. The **Toyota Smarthome** dataset poses several challenges: *high intra-class variation, high class imbalance, simple and composite activities,* and *activities with both coarse and fine-grained labels*.
- As recent action recognition/detection approaches fail to address the challenges posed by **Toyota Smarthome**, we present a novel activity detection method **PDAN** with the attention mechanism.

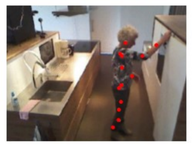
Toyota Smarthome Dataset [1,2]

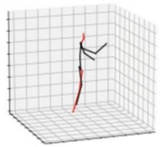
Project: https://project.inria.fr/toyotasmarthome/



Properties:
- 536 long untrimmed videos
- Unscripted behavior
- 18 subjects
- 51 action classes
- 21 mins/video
- Two-level annotation (Composite & Elementary)
- 7 camera views

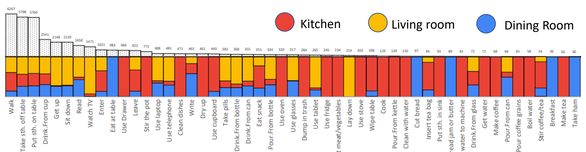
Modalities


(1) RGB

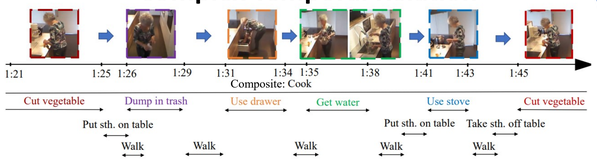

(2) 3D Skeleton


(3) Depth

Action Distributions



Complex Temporal Relation



Composite action: There are two levels of annotations: Composite (e.g. *cooking*) and elementary actions (e.g. *cutting vegetables*). We Provide temporal boundaries for both of them.

Co-occurring action: Multiple actions could happen at the same time (e.g. *taking notes whiling making a phone call*).

High Temporal Variation: Long video clips and unscripted recording leads to high temporal variation and long temporal dependencies in the dataset.

Unscripted video: actions performed in this dataset are unscripted resulting in more natural and complex relationship between the actions.

Other Challenges: Camera framing, Cross-view evaluation, Object-based actions...

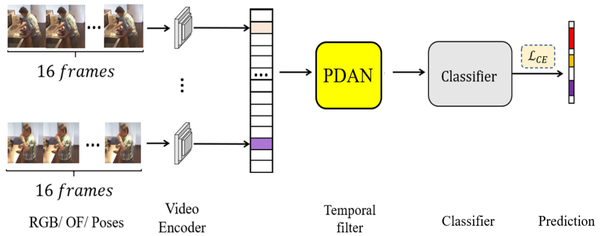
Two Version of Datasets

Dataset Version	Smarthome Trimmed	Smarthome Untrimmed
Task	Recognition	Localization
#Classes	31	**51**
#Instances	16 K	**41 K**
#Frames	3.9 M	**13.8 M**

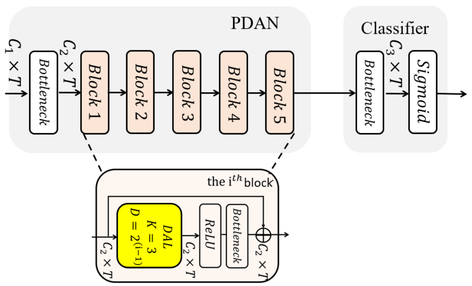
Pyramid Dilated Attention Network [3]

- We address detection of complex actions for untrimmed videos based on **Dilated Attention Layer (DAL)** and a **Pyramid structure**.
- **DAL** allocates attentional weights to each feature in the kernel, which enables DAL to learn better local representation across time.
- **Pyramid Dilated Attention Network** is built upon DAL. With the help of DAL combining with dilation and residual links, PDAN can model short-term and long-term temporal relations simultaneously by focusing on local segments at the level of low and high temporal receptive fields.
- We evaluate PDAN on three complex action datasets and achieve the SOTA performance.

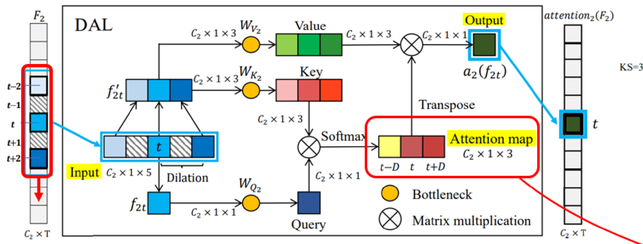
Overall Framework



PDAN Structure



Dilated Attention Layer (DAL)



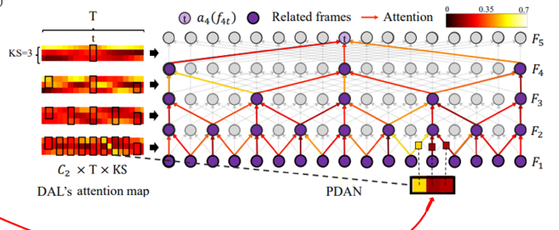


Table 1. Charades-Localization

	Modality	mAP
Two-stream [28]	RGB + Flow	8.9
Two-stream+LSTM [28]	RGB + Flow	9.6
R-C3D [36]	RGB	12.7
Asynchronous Temporal Fields [28]	RGB + Flow	12.8
I3D [24]	RGB	15.6
I3D [24]	RGB + Flow	17.2
I3D + 3 temporal conv.layers [25]	RGB + Flow	17.5
TAN [7]	RGB + Flow	17.6
I3D + WSGN (supervised) [11]	RGB	18.7
I3D + Stacked-STGCN [12]	RGB	19.1
I3D + Super event [24]	RGB + Flow	19.4
I3D + 3 TGMs [25]	RGB + Flow	21.5
I3D + 3 TGMs + Super event [25]	RGB + Flow	22.3
I3D + Dilated-TCN* [19]	RGB + Flow	23.5
I3D + MS-TCN* [9]	RGB + Flow	24.2
I3D + **PDAN**	RGB	**23.7**
I3D + **PDAN**	RGB + Flow	**26.5**

Table 2. Toyota Smarthome

	CS
R-I3D [61]	8.7
I3D(Trimmed)+Bottleneck [33]	7.4
I3D+Bottleneck [33]	15.7
I3D+Non-local block [39]	16.8
I3D+Super event [42]	17.2
I3D+LSTM [63]	22.6
I3D+Bidirectional-LSTM [59]	24.5
I3D+Dilated-TCN [41]	25.1
I3D+MS-TCN [43]	25.9
I3D+TGM [47]	26.7
I3D+PDAN	32.7

Ablation

	Dilation	Residual link	DAL in block 1	2	3	4	5	Charades	TSU
Simple(STCL)	×	×	×	×	×	×	×	17.8	15.0
Simple(DAL)	×	×	✓	✓	✓	✓	✓	18.9	16.1
Dilation (STCL)	✓	×	×	×	×	×	×	21.8	24.0
Dilation (DAL)	✓	×	✓	✓	✓	✓	✓	23.2	26.1
Residua (STCL)	×	✓	×	×	×	×	×	21.8	24.3
Residua (DAL)	×	✓	✓	✓	✓	✓	✓	23.5	26.5
PDAN (STCL)	✓	✓	×	×	×	×	×	24.1	29.0
PDAN(Low)	✓	✓	✓	✓	×	×	×	25.3	30.1
PDAN(High)	✓	✓	×	×	×	✓	✓	25.4	30.1
PDAN (DAL)	✓	✓	✓	✓	✓	✓	✓	**26.5**	**32.7**

Reference

[1] Toyota Smarthome Untrimmed: Real-World Untrimmed Videos for Activity Detection. Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, Gianpiero Francesca.. *Arxiv Pre-print*, October 2020

[2] Toyota Smarthome : Real World Activities of Daily Living. Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond and Gianpiero Francesca. *In Proceedings of the 17th International Conference on Computer Vision (ICCV)*, 2019

[3] PDAN: Pyramid Dilated Attention Network for Action Detection. Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca and Francois Bremond. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021