Institute for Infocomm Research
I²R

JOINT LEARNING ON THE HIERARCHY REPRESENTATION FOR FINE-GRAINED HUMAN ACTION RECOGNITION

Presenter: Mei Chee Leong

Authors: Mei Chee Leong[1], Hui Li Tan[1], Haosong Zhang[1;2], Liyuan Li[1], Feng Lin[2], Joo Hwee Lim[1,2]

Institute for Infocomm Research, A*STAR[1]

School of Computer Science and Engineering, Nanyang Technological University, Singapore[2]

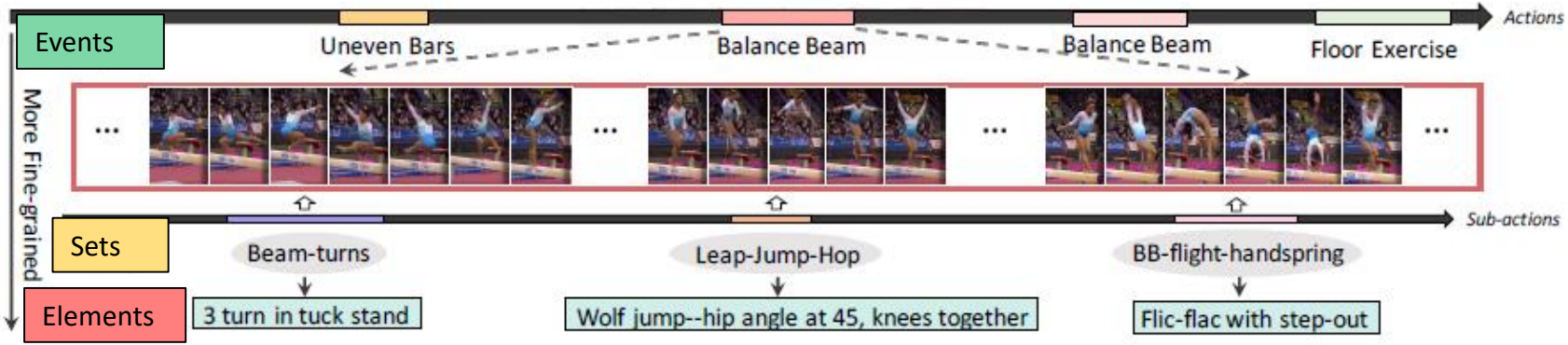
ICIP 2021

Outline

1) Fine-grained action recognition
2) Related works
3) Proposed multi-task network for joint learning
4) Experiments
5) Conclusion & future work

Fine-Grained Action Recognition

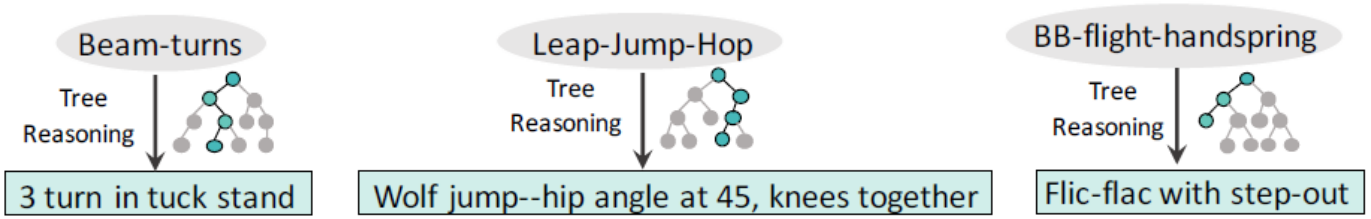
- Fine-grained action recognition – similar appearances of action and background, subtle differences in temporal dynamics and spatial semantics.
- FineGym - gymnastic video dataset with structured coarse-to-fine hierarchy annotation.
- Hierarchy representation provides discriminating and complementary information at different granularity levels.



FineGym dataset [Shao, et al. 2020]

FineGym Dataset

- A large set of fine-grained actions represented as a multi-level semantic hierarchy – provides domain constraint of actions at coarse-to-fine levels.
- Hierarchy structure derived from decision trees – encodes distinctive features of semantic meaning, visual appearance and motion information.
- Exploiting multi-level action representations – helps in better aggregation of spatial semantics and temporal dynamics, and learning of the domain class information implicitly.



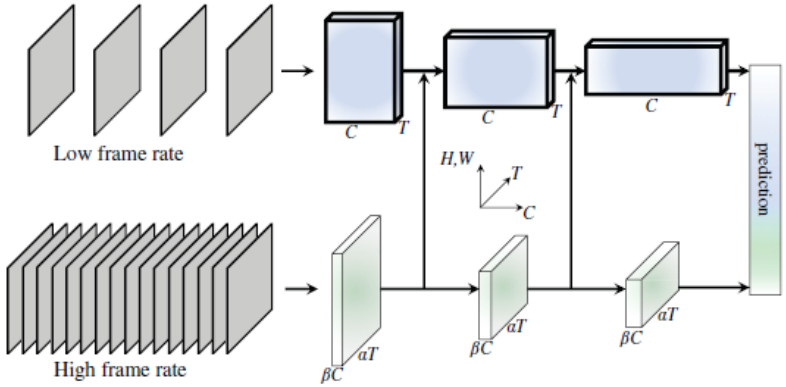
Fine-grained actions derived from decision trees. [Shao, et al. 2020]

Related Works

- **2D+1D CNN framework:**
  - a 2D CNN for extracting spatial features, followed by a 1D module for temporal aggregation.
  - E.g., TSN, TRN, TSM, ActionVLAD
- **3D CNN framework:**
  - 3D CNN to directly capture spatio-temporal features.
  - E.g., C3D, I3D, SlowOnly, SlowFast, X3D
- **Transformer-based framework:**
  - Combined CNN and Transformer, or Transformer only network
  - E.g., ViVit, TimeSformer, TQN, Swin

- Existing works mainly focus on recognizing each level of the action hierarchy separately.
- Our work investigates joint representation and prediction for hierarchical action recognition.

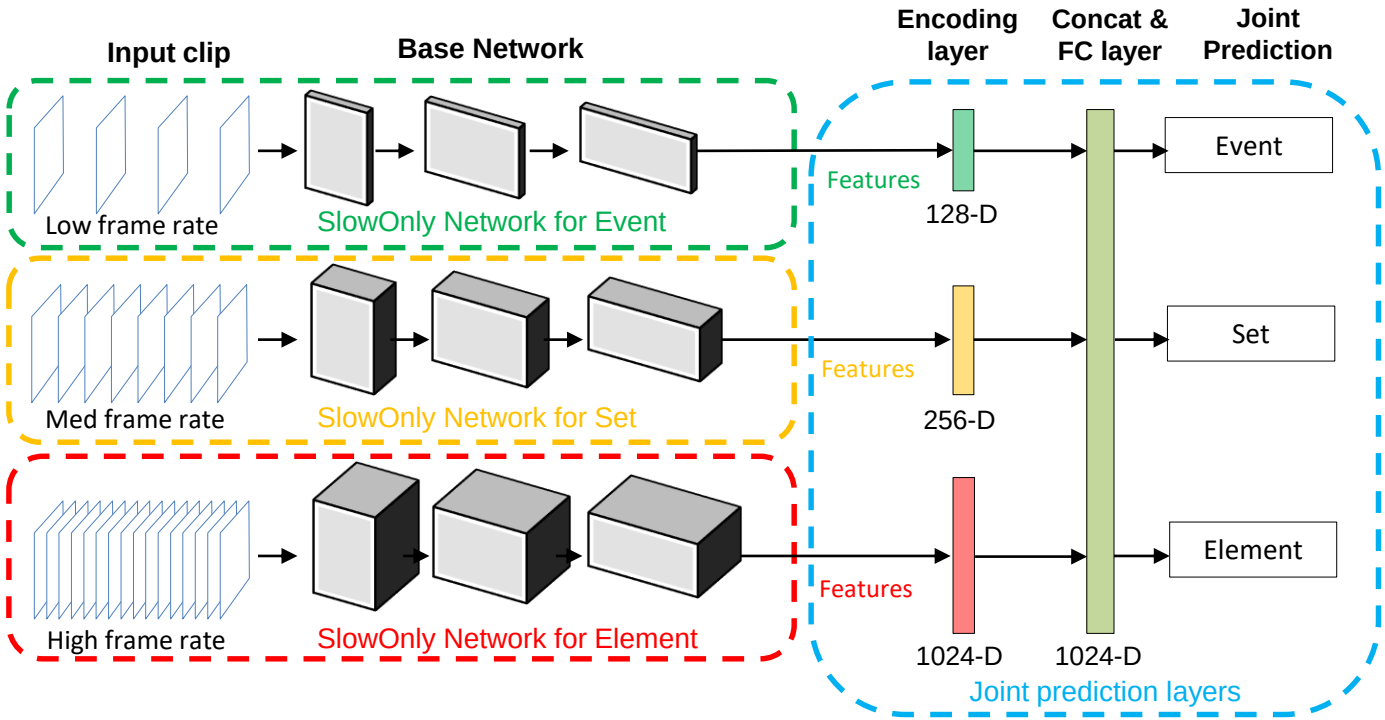
SlowFast Action Recognition Model



SlowFast network with two pathways of different temporal rates [Feichtenhofer, et al. 2019]

- Slow pathway captures spatial semantic at low frame rate.
- Fast pathway encodes motion information at high frame rate, with fewer number of channels.
- The network for Slow and Fast pathways are modified from 3D ResNet.
- Lateral connections between two pathways for feature fusion.
- SlowOnly is a variant of SlowFast without fast pathway.

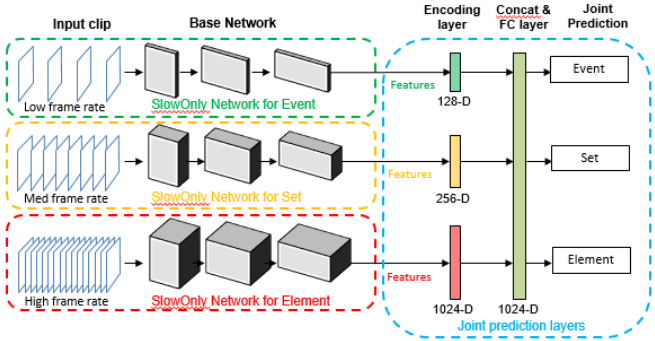
Proposed Multi-Task Network for Joint Learning



Our proposed framework for joint learning on hierarchical representation for multi-level action recognition.

Proposed Multi-Task Network

- 2-stage training:
  1) Train individual base model (SlowOnly / SlowFast) for Event, Set and Element using Element clip input.
  2) Train joint prediction layers. Freeze pre-trained base models to form network pathways.



- Encoded features from each pathway are concatenated and passed to a linear layer for joint representation learning.
- The fused feature is connected to three classifiers for joint prediction of Event, Set and Element.
- Adopt weighted cross-entropy loss function:

$$L^{total} = \lambda_1 * L^{Event} + \lambda_2 * L^{Set} + \lambda_3 * L^{Element}$$

where $\lambda_1, \lambda_2, \lambda_3 = 1, 2, 4$ respectively

Experiments

- Implemented using MMAction2.
- ResNet-50 adopted as backbone for SlowOnly and SlowFast base model.
- Experiment on Gym99 dataset, contains 4 Events, 15 Sets, and 99 Elements annotation.

Table 1. Individual model results for Event, Set and Element.

Class	Network	T	base C	Event			Set			Element		
				Top-1	Top-5	Mean	Top-1	Top-5	Mean	Top-1	Top-5	Mean
Event	SlowOnly	4	64	99.28	100	99.22	-	-	-	-	-	-
Set	SlowOnly	8	64	-	-	-	95.49	99.95	95.43	-	-	-
Set	SlowOnly	16	64	-	-	-	97.70	99.83	97.68	-	-	-
Element	SlowOnly[1]	4	64	-	-	-	-	-	-	79.30	-	70.2
Element	SlowOnly	32	64	-	-	-	-	-	-	91.05	97.82	88.40
Element	SlowFast	4 (slow) 32 (fast)	64 (slow) 8 (fast)	-	-	-	-	-	-	82.32	98.15	78.93

[1]SlowOnly results on Gym99 in MMAction2: https://github.com/open-mmlab/mmaction2/tree/master/configs/recognition/slowonly

- Utilizing high number of frame inputs generally leads to better results in Top-1 and mean accuracies.
- SlowOnly outperforms SlowFast in Element prediction due to:
  1) reduced channel dimension in the fast pathway which could not capture sufficient temporal context.
  2) temporal convolution at the later layers of SlowOnly captures more useful spatial semantics.

Experiments

Table 1. Individual model results for Event, Set and Element.

Class	Network	T	base C	Event			Set			Element		
				Top-1	Top-5	Mean	Top-1	Top-5	Mean	Top-1	Top-5	Mean
Event	SlowOnly	4	64	99.28	100	99.22	-	-	-	-	-	-
Set	SlowOnly	8	64	-	-	-	95.49	99.95	95.43	-	-	-
Set	SlowOnly	16	64	-	-	-	97.70	99.83	97.68	-	-	-
Element	SlowOnly[1]	4	64	-	-	-	-	-	-	79.30	-	70.2
Element	SlowOnly	32	64	-	-	-	-	-	-	91.05	97.82	88.40
Element	SlowFast	4 (slow) 32 (fast)	64 (slow) 8 (fast)	-	-	-	-	-	-	82.32	98.15	78.93

Table 2. Multi-task network results for joint recognition.

Combined pathways			Event			Set			Element		
Event model	Set model	Element model	Top-1	Top-5	Mean	Top-1	Top-5	Mean	Top-1	Top-5	Mean
SlowOnly	SlowOnly, $T = 8$	SlowOnly	99.50	100	99.40	98.42	100	98.16	91.58	99.64	87.50
SlowOnly	SlowOnly, $T = 16$	SlowOnly	99.54	100	99.37	98.94	100	98.87	91.80	99.69	88.46
SlowOnly	SlowOnly, $T = 8$	SlowFast	99.81	100	99.63	98.18	100	97.75	82.68	98.51	77.62
SlowOnly	SlowOnly, $T = 16$	SlowFast	99.81	100	99.49	98.78	99.98	98.47	82.95	98.73	77.76

Multi-task networks with fused features outperform their corresponding base models trained separately at each level - due to encoding and learning complementary features from multi-level action class.

Comparison with SOTA

Model	Modality	Event		Set		Element	
		Top-1	Mean	Top-1	Mean	Top-1	Mean
TSN	2 stream	99.86	98.47	97.69	91.97	86.0	76.4
TRNms	2 stream	-	-	-	-	87.8	80.2
TSM	2 stream	-	-	-	-	88.4	81.2
I3D	RGB	-	-	-	-	75.6	64.4
NL I3D	RGB	-	-	-	-	75.3	64.3
SlowFast (multi-task)	RGB	99.81	99.63	97.70	97.40	79.16	74.93
3-pathways multi-task (Ours)	RGB	99.54	99.37	98.94	98.87	**91.80**	**88.46**

- SlowFast (multi-task) – implemented as baseline by modifying its last fully-connected layer to three classifiers and adopting the same weighted loss function for training.
- Proposed 3-pathways multi-task network outperforms other action models in Element prediction, with **improvement of 3.40% and 7.26% in top-1 and mean accuracy**.
- Compared to the baseline SlowFast network, proposed network learns better joint representation that leads to improved performance in Element and Set predictions. For Event prediction, our performance is comparable to TSN and SlowFast results.

Conclusion & Future Work

- We presented a multi-task network for effective representation and learning of fine-grained actions utilizing the structured semantic hierarchy in FineGym.
- Experimental results show the effectiveness of exploiting and learning joint representation of complementary spatio-temporal features at different granularities, outperforming networks with single task prediction.
- For future work, we look into training end-to-end networks to jointly learn and refine the multi-level action representation for multi-task prediction.

CREATING GROWTH, ENHANCING LIVES

Institute for Infocomm Research
I²R

THANK YOU

www.a-star.edu.sg