‘Quo Vadis, Action Recognition?’ Towards Robust Human Action Modelling

For Interview of Research Fellow Position

Lei Wang[1,2]

[1]Australian National University
[2]Data61/CSIRO

January 6, 2023

Action Recognition, Challenges, Solutions & A Review

Action Recognition: recognize/identify actions in video

Motivations:

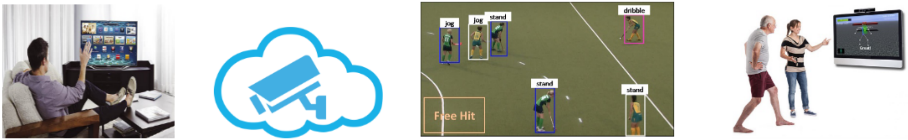

Figure 1: Many useful applications.

Challenges:

Figure 2: Many challenging issues.

Action Recognition, Challenges, Solutions & A Review

Table 1: Some benchmarks for 3D action recognition.

Datasets	Year	Classes	Subjects	#Views	#videos	Sensor	Modalities	#joints
UWA3D Activity	2014	30	10	1	701	Kinect v1	RGB + Depth + 3DJoints	15
UWA3D Multiview Activity II	2015	30	9	4	1,070	Kinect v1	RGB + Depth + 3DJoints	15
Charades	2016	157	-	-	66,500	-	RGB	-
NTU RGB+D	2016	60	40	80	56,880	Kinect v2	RGB + Depth + IR + 3DJoints	25
NTU RGB+D 120	2019	120	106	155	114,480	Kinect v2	RGB + Depth + IR + 3DJoints	25
Kinetics-skeleton	2019	400	-	-	260,232	-	2DJoints	18
Kinetics-700	2020	700	-	-	647,907	-	RGB	-

Techniques:

- Conventional RGB videos
  - handcrafted: Dense Trajectories (DT), Improved Dense Trajectories (IDT), *etc.*
  - deep-learning: two-stream networks, C3D, TSN, Inflated 3D (I3D), *etc.*
- Depth videos (*e.g.*, HON4D, HOPC, *etc.*)
- Skeleton sequences (*e.g.*, ST-GCN, *etc.*)

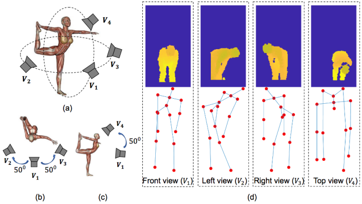

Figure 3: Camera setup, depth video frame & skeletons.

A **comparative review**[1] of recent action recognition algorithms

[1] Wang, L., Huynh, D. Q., & Koniusz, P. (2020). **A comparative review of recent kinect-based action recognition algorithms**. *IEEE TIP*, 29, 15-28.

Feature Hallucination

Motivation

- Transition: handcrafted feature → CNN models
- Handcrafted features
  - capture **domain specific information**
  - **fused with CNNs** for better performance, but **costly**

Our models

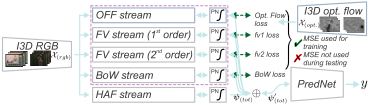

Figure 4: DEEP-HAL

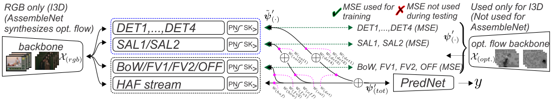

Figure 5: DEEP-HAL+SDF/ODF

- DEEP-HAL[2]
  - learn to 'translate' the CNN output to IDT
  - even 'translate' the CNN output to I3D optical flow features
- DEEP-HAL+SDF/ODF[3]
  - use detectors & saliency
  - form higher-order statistical moments (subspaces)

[2] Wang, L., Koniusz, P., & Huynh, D. Q. (2019). **Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns**. In *ICCV* (pp. 8698-8708).

[3] Wang, L., & Koniusz, P. (2021). **Self-supervising action recognition by statistical moment and subspace descriptors**. In *ACMMM* (pp. 4324-4333).

Tensor Representations & Feature Fusion

Motivation

- interactions of groups of skeletal joints
- **physical connectivity**, **limited** receptive fields
- ignore the dependency between body joints **non-connected** by body parts

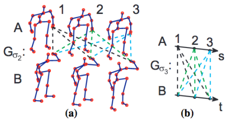

Figure 6: SCK

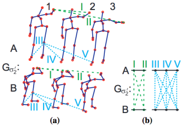

Figure 7: DCK

Our models[4]

SCK+DCK & SCK⊕+DCK⊕

- **capture complex interplay**
- incorporate **multi-modal inputs**

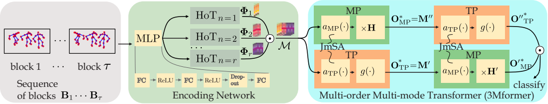

Figure 8: 3Mformer.

Multi-order Multi-mode Transformer (3Mformer)

- use skeletal hypergraph
- encode first-, second- & higher-order hyper-edges features

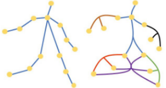

Figure 9: Skeletal graph & hypergraph.

[4] Koniusz, P., Wang, L., & Cherian, A. (2021). **Tensor representations for action recognition**. *IEEE TPAMI*, 44(2), 648-665.

Few-shot Alignment-based

Motivation

- FSL can quickly adapt to novel classes if annotations are limited
- FSL on skeletons for action recognition is underexplored

Our models

uncertainty-DTW (uDTW)[5]:

- We train the Encoding Network.
- The comparator learns the notion of similarity between query-support pairs.
- At the test time, given a set of support sequences with labels, we can decide which one matches the query.

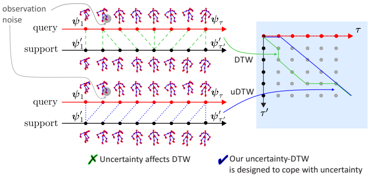

Figure 10: Soft-DTW. (top) *vs.* uncertainty-DTW (bottom).

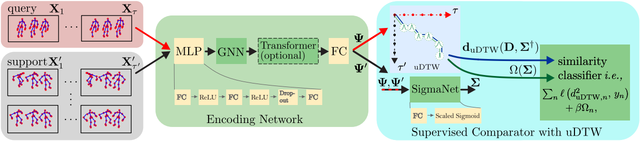

Figure 11: Supervised few-shot action recognition.

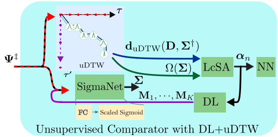

Figure 12: Unsupervised.

[5]Wang, L., & Koniusz, P. (2022). **Uncertainty-DTW for Time Series and Sequences**. In *ECCV*, oral.

Few-shot Alignment-based

Joint tEmporal and cAmera viewpoiNt alIgnmEnt (JEANIE)[6]:

- Match query-support features under varying viewpoints of 3D poses
- find a smooth joint viewpoint-temporal alignment
- minimize/maximize d_{JEANIE} for same/different support-query labels
- **JEANIE** has the transportation plan ↓, ↘, → for temporal axes & take additional steps on the viewpoint axis, *e.g.*, **step inward**, **inward-down**, *etc.*

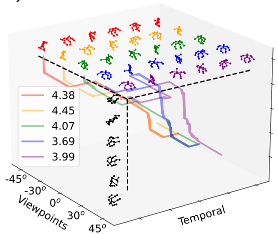

Figure 13: JEANIE.

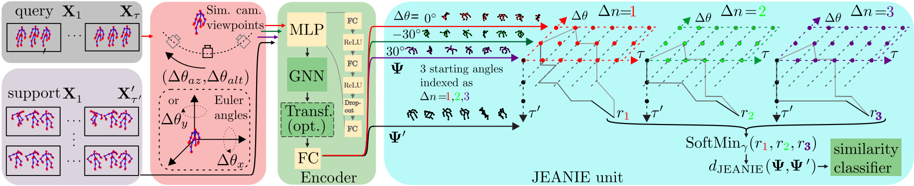

Figure 14: Our 3D skeleton-based FSAR with JEANIE.

[6] Wang, L., & Koniusz, P. (2022). **Temporal-Viewpoint Transportation Plan for Skeletal Few-shot Action Recognition**. In *ACCV*, oral, **Best Student Paper Award**.

'Quo Vadis, Action Recognition?'

Conclusion:

- Video-based:
  - self-supervision/MTL/co-regularize a CNN resembles domain adaptation
  - + easier to obtain video frames/rich visual information/robust backbones
  - – require **large-scale dataset**/computational cost/deal with redundant pixels
- Skeleton-based:
  - tensor representations & multi-order multi-mode feature fusion
  - + openpose & Kinect toolkit+OpenNI/lightweight/faster to process
  - – require **large-scale dataset**/reliability/ lack visual information
- Few-shot:
  - alignment-based/match query-support pair
  - + faster adaptation to novel classes/limited data is fine
  - – **robust data** is required to learn a good model

Remark:

- existing works all report promising results
- new and more robust algorithms are still required
- a pressing demand in real and new environments

Thank you!