Action Recognition: Past, Present and Future

Lei Wang[1,2]

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

August 12, 2023

Table of Contents

1. Action Recognition, Challenges & Benchmarks

2. Action Recognition on Videos

3. Action Recognition on Skeletons

4. Multi-modal & Multi-view Action Recognition

5. Conclusion

Action Recognition, Challenges & Benchmarks

Action Recognition, Challenges & Benchmarks

Action Recognition: recognize/identify actions in video

Motivations:

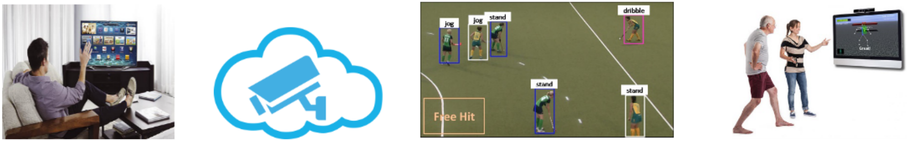


Figure 1: Many useful applications.

Challenges:

Figure 2: Many challenging issues.

Action Recognition, Challenges & Benchmarks (cont.)

Table 1: Some benchmarks for action recognition.

Datasets	Year	Classes	Subjects	#views	#video clips	Sensor	Modalities
MSRAction3D	2010	20	10	1	567	Kinect v1	Depth+3DJoints
3D Action Pairs	2013	12	10	1	360	Kinect v1	RGB+Depth+3DJoints
UWA3D Activity	2014	30	10	1	701	Kinect v1	RGB+Depth+3DJoints
UWA3D Multiview Activity II	2015	30	9	4	1,070	Kinect v1	RGB+Depth+3DJoints
MPII Cooking Activities	2012	64	12	1	3,748	-	RGB
HMDB-51	2011	51	-	-	6,766	-	RGB
EPIC-Kitchens	2018	149	32	-	39,594	-	RGB+Flow
NTU RGB+D	2016	60	40	80	56,880	Kinect v2	RGB+Depth+IR+3DJoints
Charades	2016	157	-	-	66,500	-	RGB+Flow
NTU RGB+D 120	2019	120	106	155	114,480	Kinect v2	RGB+Depth+IR+3DJoints
Kinetics-skeleton	2017	400	-	-	260,232	-	2DJoints
Kinetics	2018	400	-	-	$\sim$ 300,000	-	RGB

Figure 3: Video frame images

[A] A comparative review of recent kinect-based action recognition algorithms. TIP'20.

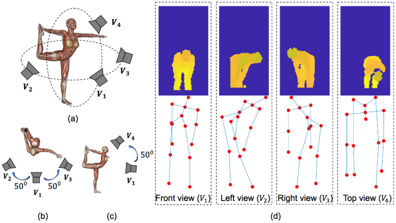
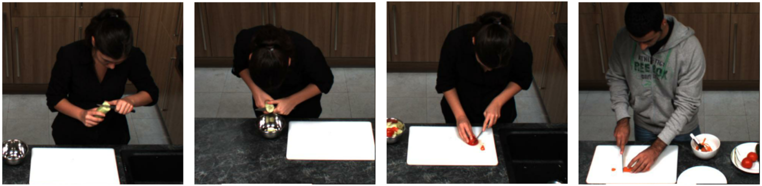


Figure 4: Setup, depth frames & skeletons[A].

Action Recognition, Challenges & Benchmarks (cont.)



(a) cut apart **(b)** cut dice **(c)** cut slices **(d)** cut slices

Figure 5: Finegrained action recognition (MPII Cooking Activities)

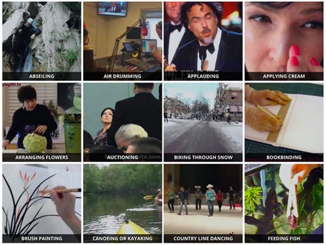


Figure 6: Video frames from Kinetics700

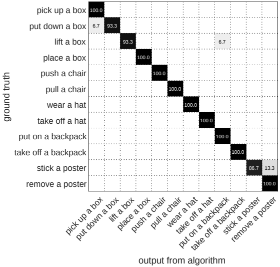


Figure 7: Confusion matrix

Action Recognition on Videos

Action Recognition on Videos

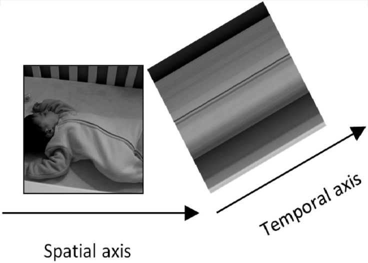


Figure 8: Spatio-temporal info.

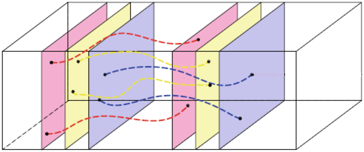


Figure 9: DT & IDT

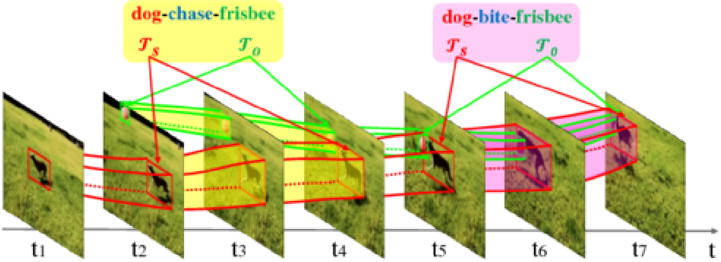


Figure 10: Video visual relation instances (ImageNet-VidVRD)

Action Recognition on Videos (cont.)

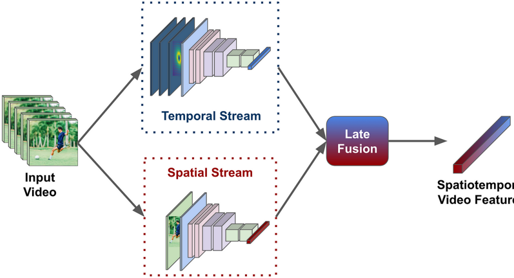


Figure 11: Two-stream networks (2D CNN)

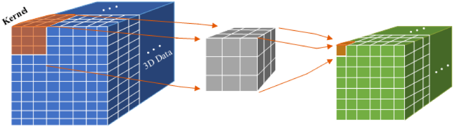


Figure 12: 3D CNN

Recent advanced AR models: AssembleNet, Video masked autoencoder, video vision transformer, video swin transformer & vision-language model e.g., CLIP, etc.

Action Recognition on Skeletons

Action Recognition on Skeletons

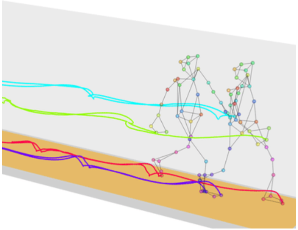


Figure 13: Skeleton sequences

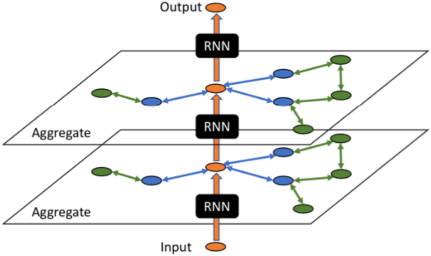


Figure 14: GNN+RNN

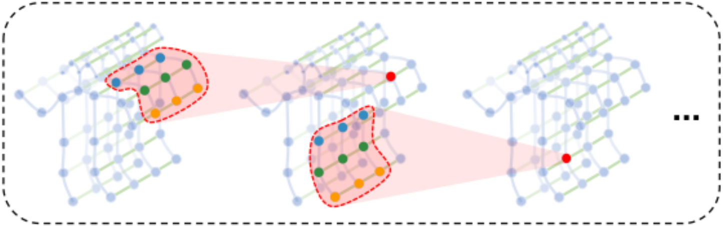


Figure 15: ST-GCN

Shift-GCN, Channel-wise Topology Refinement GCN, Efficient GCN, etc.

Action Recognition on Skeletons (cont.)

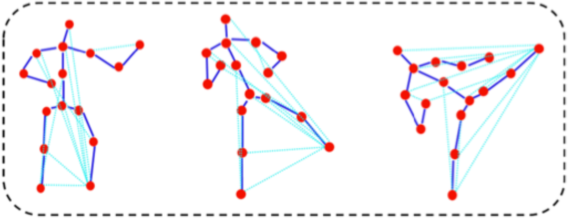


Figure 16: *Kicking* action in NTU-120 dataset

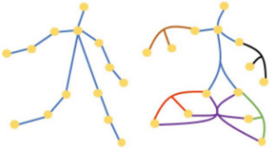


Figure 17: Skeletal graph & hypergraph.

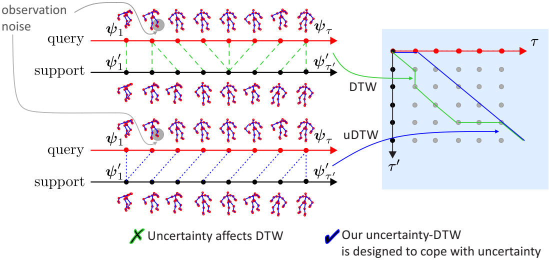


Figure 18: Temporal alignment

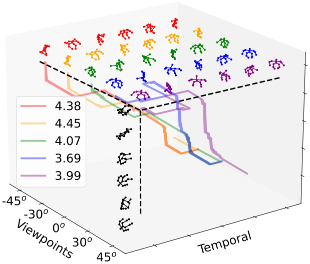


Figure 19: Joint temporal & viewpoint alignment

Multi-modal & Multi-view Action Recognition

Multi-modal & Multi-view Action Recognition

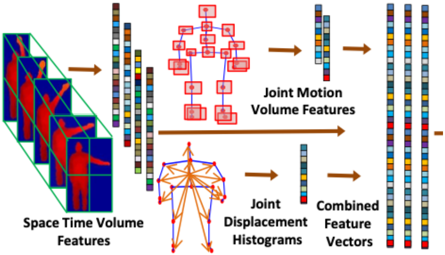

Space Time Volume Features
Joint Motion Volume Features
Joint Displacement Histograms
Combined Feature Vectors

Figure 20: Depth videos + Skeletons

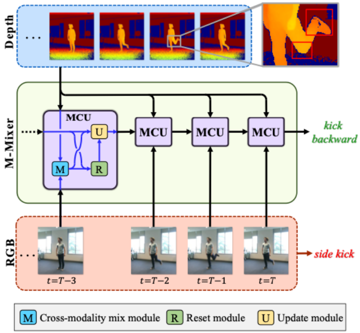


Figure 21: Modality Mixer

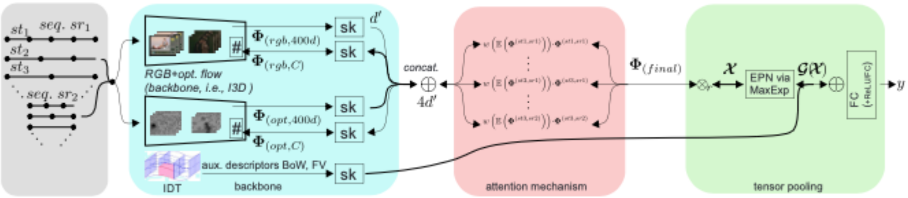
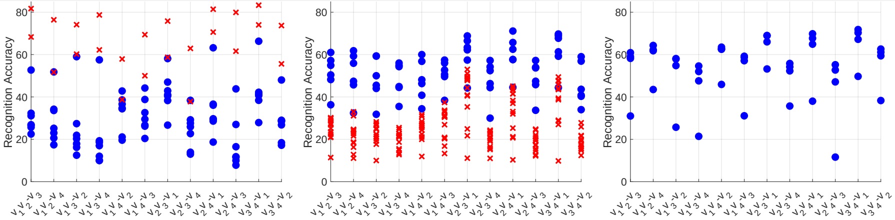


Figure 22: RGB video frames, optical flow & IDT

Multi-modal & Multi-view Action Recognition (cont.)



(a) Depth-based (b) Skeleton-based (c) Depth+Skeleton-based

Figure 23: Scatter plots of cross-view action recognition performance.

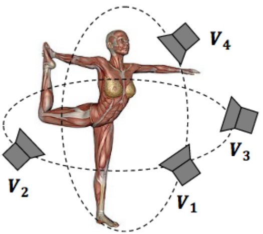


Figure 24: Camera setup.

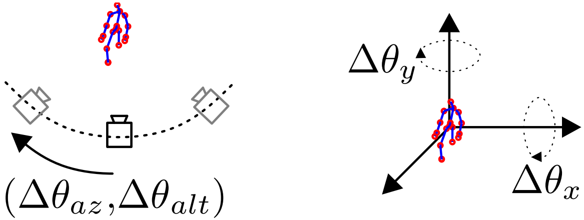


Figure 25: Viewpoint augmentation.

Multi-modal & Multi-view Action Recognition (cont.)

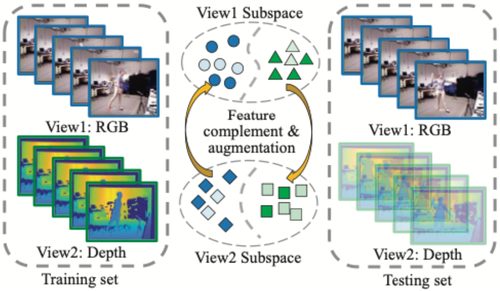


Figure 26: Generative multi-view AR

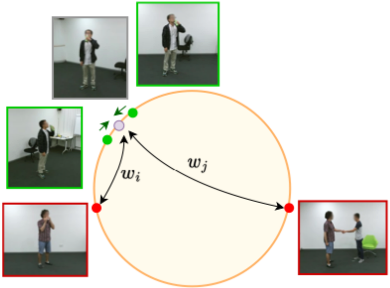


Figure 27: Cross-view Contra. Learning

View-adaptation model (learn viewpoints)

Conclusion

Conclusion

This space
intentionally
left blank.

Acknowledgment: The figures presented in the slides are obtained from either online resources or AR publications.

Thank you!