

MARMARA ÜNİVERSİTESİ

Comparison of Deep Learning Architectures for Human Action Recognition from Videos Emre Barkın BOZDAĞ Sevda DURDU emrebarkinn@gmail.com sevdaadurdu2@gmail.com

Advisor: Prof. Dr. Çiğdem EROĞLU ERDEM



Introduction

- Given a video of human actions, automatically recognize the action.
 - Video frames
 - Human poses
 - Depth maps
- Comparison of 3 different approaches.
- 2 different data sets: **NTU-RGB+D**[1] and **JHMDB**[2].

Data sets

Methodologies



Human Pose Pose Estimation Tracking





Action Recognition

Experimental Results

For **medical actions**:

- Z choices:
 - 2D (z=0)
 - 2D (z=confidence)
- Used joint numbers: **2**5

18

• First half of video

• Skipped Frame Number:

For **daily actions**:

Ο

0 2

• Last half of video

Medical Actions Results on NTU-RGB

Confusion Matrix For V/A_CNN

NTU-RGB+D data set consists of **120** classes: 82 Daily Actions, 12 Medical Actions, 26 Mutual Actions This data set contains **114,480** videos, **106** subjects,**155** different views and **3** camera angles.

Cross-subject and **Cross-view** evaluations

Chest Pain Action







Subject 7 with camera angle 2

Subject 3 with camera angle 1

Contains 3D joints and 2D Joints with on image coordinates.

• 2D coordinates generated by kinect camera in data set using rgb frames and depth maps.

Subject 7 with camera

angle

• For experiments, we generate 2D joints with OpenPose [7] using only RGB frames.





Pose Tracking [6]

• Deep Sort

Algorithm

3 different methodologies were used in this research.

1-) View Adaptive Neural Networks [3]

• Uses 3D

• VA-RNN

• VA-CNN

skeleton

2-) R(2+1)D Model [4]

weakly-supervised

• Focusing exclusively

on training data

• Uses RGB video

pre-training

• Large-scale



tation Subnetwork h ^r FC R _t Layer S _{rota} Main LSTM Network T	17 Joints 2
h ^d Layer Layer Layer Output h ^d Layer Strans iew Adaptation Subnetwork Main ConvNet I ConvNet Output	17 Joints Openpose 2
	Medical Actions
	Cross View
fc space-time pool	Cross Subje
(2+1)D conv (2+1)D conv (2+1)D conv (2+1)D conv (2+1)D conv (2+1)D conv (2+1)D conv	Cross Su 1.0 0.8 0.6 0.6
Rotation & Shifting 4 = (Actions)	0.4 → 79.8% VA → 67.9% VA → 69.5% DD → 82.0% R(2 → 81.1% VA 0.0 0
Cartesian coordinate feature: $\neq \tilde{x}$ ioints	JHMDB
JCD feature:	Accuracy

Medical				on Cross View														
Actions	VA-C	INN	CV	CS	eeze/cou	ugh 82	2.0%	0.0%	0.0%	3.5%	0.0%	0.6%	2.5%	4.4%	0.6%	0.0%	0.0%	6.3%
					stagger	ring 0	0.0%	99.7%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Baseline	31	D	90.54	82.44	falling do	wn 0).0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
					headac	che 2	2.5%	0.0%	0.0%	85.8%	0.0%	0.3%	5.7%	0.0%	1.3%	1.9%	0.0%	2.5%
25 Joints	20 (-	z=0)	83 54	79 71	chest p	ain ³	8.8%	0.0%	0.0%	0.0%	88.9%	2.8%	0.0%	3.2%	0.6%	0.0%	0.0%	0.6%
		_=0)		7 5.2 1	back p	oain 0	0.3%	0.0%	0.0%	0.0%	0.9%	93.0%	2.2%	0.0%	1.6%	0.3%	0.3%	1.3%
		\sim	07.0	04.64	neck p	ain 1	.6%	0.0%	0.0%	6.0%	0.6%	0.6%	88.0%	0.0%	0.6%	1.6%	0.0%	0.9%
25 Joints	2D (2	z=0)	87.3	81.61	fan s	self 0	2%	0.0%	0.0%	2.8%	0.9%	0.9%	0.6%	0.0%	92.7%	0.9%	0.0%	0.3%
Openpose	21	D			ya	awn 0	0.3%	0.0%	0.0%	0.6%	0.0%	0.0%	0.9%	0.0%	0.3%	87.5%	0.6%	9.7%
	(z=c	onf)	84.62	80.13	stre	tch 0	0.0%	0.0%	0.0%	0.3%	0.0%	0.6%	0.0%	0.0%	0.0%	0.9%	98.1%	0.0%
	.				blow no	ose <mark>3</mark>	8.8%	0.0%	0.0%	0.9%	0.6%	0.6%	1.9%	0.6%	0.3%	16.6%	0.0%	74.7%
17 Joints	_						ugho	ering	uwop	ache	t pain	, pain	pain	niting	n self	yawn	retch	nose
	2D (z	z=0)	87.25	80.47			eeze/c	stagg	falling	head	chest	back	neck	NON	fai		st	blow
						•						4						-
17 Joints Openpose	2D (z	z=0)	88.93	83.15												ł		
						Со	ug	h			С	hes	t Pa	in		Ne	ck l	Pain
Medical Actions		VA-	CNN	VA-R	NN	VA	\-F	505	510	N	D	N−C	JET		R(2	2+´	1)D	
Cross Vie	2W	90	0.5%	5% 86.8%		91.6%			88.1%			84.2%)				
Cross Su	ss Subject 85.2% 80.8% 86.5%				79.1% 88.1%				,									
(Cumulative Match Characteristic Curves on NTU-RGB																	
Cros	ss Subiect	Evaluati	on on 77 Ca	ategorv	D	dla	1 2)ei	Cro	ss Vie	w Ev	aluat	ion oi	n 77 (Cated	orv		

• 13

25 body joints from NTU-RGB +D data set

25 body joints from OpenPose

JHMDB data set

- A fully annotated data set
- 2D skeletons are interpreted from RGB video
- 928 clips from HMDB51 [8] comprising 21 categories.





Used Technologies



3-) DD-NET (Double-feature **Double-motion** Network) [5]

- Uses 2D skeleton
- Location-viewpoint variation
- Motion scale variation
- Related/unrelated to global trajectories
- Uncorrelated joint indices

(a) Location-viewpoint variation

Joints Collection Distances	Slow Cart motion Coord	esian linates Fast motion
	Temporal difference (stride=1)	Temporal difference (stride=2)
CNN(1,2*filters)	CNN(1,2*filters)	CNN(1,2*filters)
CNN(3,filters)	CNN(3,filters)	CNN(3,filters)
CNN(1,filters), /2	CNN(1,filters), /2	CNN(1,filters)
·,	Concatenate	
	▼ 2×CNN(3,2*filters), /2	
	↓ 2×CNN(3,4*filters), /2	
	2×CNN(3,8*filters)	
	♦ GAP	
	♦ FC(128)	
	★ FC(num_classes)	



Results on JHMDB

HMDB	VA-CNN	VA-RNN	DD-NET	R(2+1)D
Accuracy	72.76%	52.31%	76.00%	87.68%

Model Architecture Comparison

For experimental setup, Nvidia Tesla T4 is used in Google Cloud

Model Architecture Comparison	Total Number of Parameter	Run Time per Video
VA-CNN	24,157,988	4.6ms
DD-NET	1,809,246	0.212ms
R(2+1)D	63,540,197	153ms

77 Action Results on NTU RGB Data Set

				1 frame	2 frame	
77 Category	Accuracy	First Half	Last Half	skip	skip	Evaluation
	80.42%	24.47%	42.90%	72.06%	64.86%	CS
VA-CNN	86.42%	26.82%	45.12%	78.01%	70.59%	CV
	67.85%	21.16%	39.85%	66.92%	64.97%	CS
VA-LSTM	80.86%	25.94%	44.57%	77.79%	73.74%	CV



	76.70%	20.8%	31.34%	64.52%	55.67%	CS
DD-NET	83.63%	20.19%	33.71%	66.88%	57.17%	CV
R(2+1)d	81.96%	72.06%	64.86%	70.70%	68.57%	CS
	85.28%	71.39%	59.85%	81.30%	72.60%	CV

References

- [1] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding
- [2] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in International Conf. on Computer Vision (ICCV), Dec. 2013, pp. 3192–3199.
- [3] Zhang, Pengfei and Lan, Cuiling and Xing, Junliang and Zeng, Wenjun and Xue, Jianru and Zheng, Nanning View adaptive neural networks for high performance skeleton-based human action recognition 2017 IEEE Transactions on Pattern Analysis and Machine Intell [4] Fan Yang, Sakriani Sakti, Yang Wu, Satoshi Nakamura, Nara Institute of Science and Technology, Japan RIKEN, AIP, Japan Kyoto University, Japan Make Skeleton-based Action Recognition Model Smaller, Faster and Better
- [5] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, Dhruv Mahajan Facebook AI Large-scale weakly-supervised pre-training for video action recognition
- [6] Nicolai Wojke, Alex Bewley, Dietrich Paulus, "Simple Online And Realtime Tracking With A Deep Association Metric" 2017 IEEE International Conference on Image Processing (ICIP)2017
- [7] Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017b. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR. [8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. ICCV, pp. 2556 – 2563, 2011.

Conclusions

- NTU-RGB +D data set
- Skeleton based models (VA-CNN) gives better results for Cross View tests
- Video based model R(2+1)D gives better result for Cross Subject tests
- JHMDB data set
 - Video based model R(2+1)D gives better results.
- For real-time application, since every person's pose can be detected separately, with using pose based methods multiple person's actions can be recognized. For single person cases, using R(2+1)d model is better because it doesn't need pose estimation so it operates fasters as shown in our test results.