Comparison of Deep Learning Architectures for Human Action Recognition from Videos

Emre Barkın BOZDAĞ
emrebarkinn@gmail.com

Sevda DURDU
sevdaadurdu2@gmail.com

Advisor: Prof. Dr. Çiğdem EROĞLU ERDEM

MARMARA ÜNİVERSİTESİ

MARMARA ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ

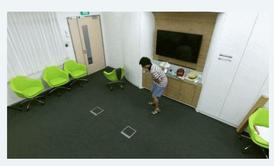
Introduction

- 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].**

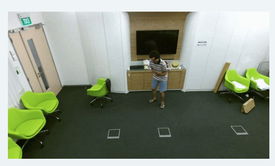
Data sets

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

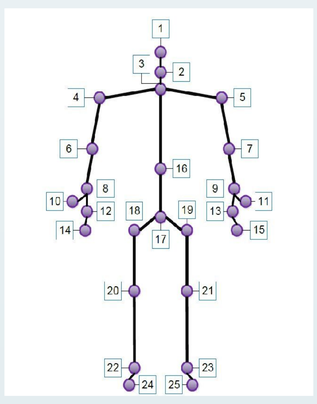


Subject 7 with camera angle 1

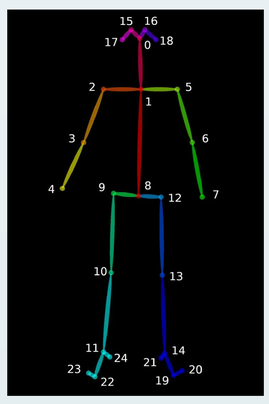
Subject 3 with camera angle 1

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



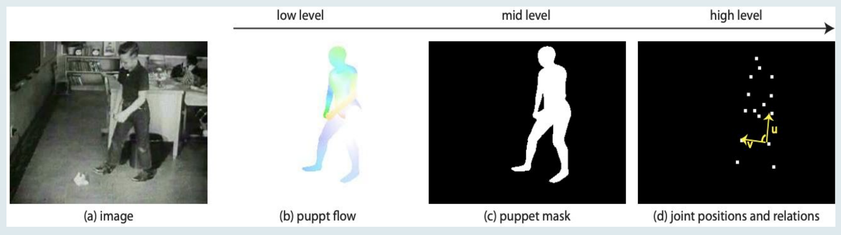
25 body joints from NTU-RGB +D data set



25 body joints from OpenPose

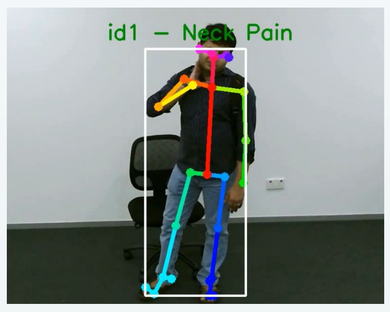
JHMDB data set

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



Brush Hair Catch Clap Climb Stairs Golf

Methodologies



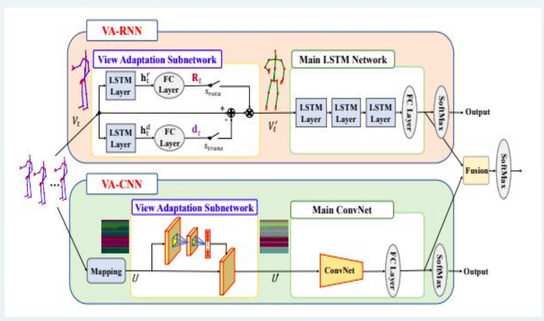
Pose Tracking [6]

- Deep Sort Algorithm

3 different methodologies were used in this research.

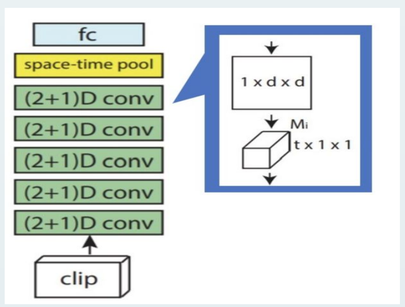
1-) View Adaptive Neural Networks [3]

- Uses 3D skeleton
- VA-RNN
- VA-CNN



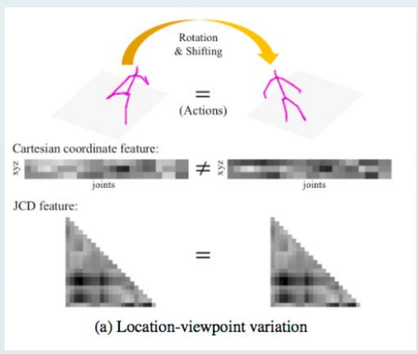
2-) R(2+1)D Model [4]

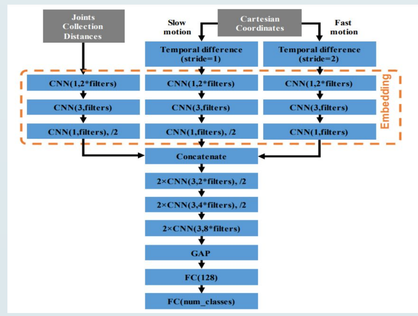
- Uses RGB video
- Large-scale weakly-supervised pre-training
- Focusing exclusively on training data



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





Experimental Results

For **medical actions**:

- Z choices:
  - 2D (z=0)
  - 2D (z=confidence)
- Used joint numbers:
  - 25
  - 18

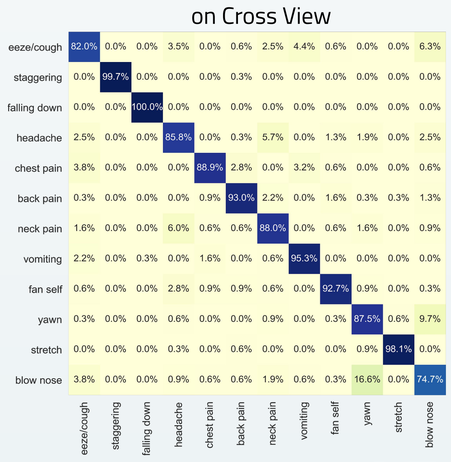
For **daily actions**:

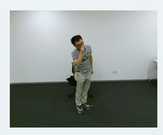
- Skipped Frame Number:
  - 1
  - 2
- First half of video
- Last half of video

Medical Actions Results on NTU-RGB

Medical Actions	VA-CNN	CV	CS
Baseline	3D	90.54	82.44
25 Joints	2D (z=0)	83.54	79.21
25 Joints Openpose	2D (z=0)	87.3	81.61
	2D (z=conf)	84.62	80.13
17 Joints	2D (z=0)	87.25	80.47
17 Joints Openpose	2D (z=0)	88.93	83.15

Confusion Matrix For VA-CNN on Cross View





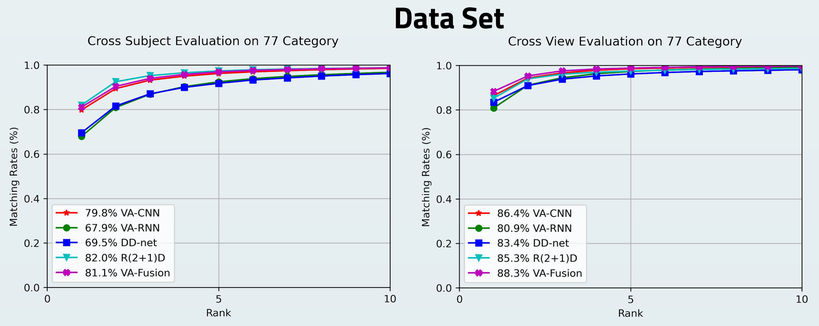
Cough

Chest Pain

Neck Pain

Medical Actions	VA-CNN	VA-RNN	VA-FUSION	DD-NET	R(2+1)D
Cross View	90.5%	86.8%	91.6%	88.1%	84.2%
Cross Subject	85.2%	80.8%	86.5%	79.1%	88.1%

Cumulative Match Characteristic Curves on NTU-RGB Data Set



Results on JHMDB

JHMDB	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

77 Category	Accuracy	First Half	Last Half	1 frame skip	2 frame skip	Evaluation
VA-CNN	80.42%	24.47%	42.90%	72.06%	64.86%	CS
	86.42%	26.82%	45.12%	78.01%	70.59%	CV
VA-LSTM	67.85%	21.16%	39.85%	66.92%	64.97%	CS
	80.86%	25.94%	44.57%	77.79%	73.74%	CV
DD-NET	76.70%	20.8%	31.34%	64.52%	55.67%	CS
	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

Used Technologies

Google Cloud

scikit learn

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.