AUTOMATIC 3D HUMAN ACTION RECOGNITION

Ajmal Mian
Associate Professor
Computer Science & Software Engineering
www.csse.uwa.edu.au/~ajmal/

Overview

- Aim of automatic human action recognition
- Applications
- Challenges
- Deep Learning
- Data generation
- Training the deep neural network models
- Results

Aim of human action recognition

- Given a video, automatically classify what action is performed in the video
- Who is performing the action is not relevant, the action is e.g.
  - Walking or running
  - Falling down
  - Struggling in a pool
  - Answering a phone
  - Holding head / chest / stomach
  - Interactions with others such as handing over a bag

Applications

- Elderly care / smart houses
- Child minding
- Swimming pool monitoring
- Surveillance & security
- Action based video search

- At a more finer level
  - Rehabilitation (e.g. quantifying progress in walking)
  - Sports action analysis (injury prevention)

Challenges

- The same action appears very different (to computers) when viewed from different angles
- Changes in illumination can completely change the appearance of videos
- The same action may be performed in different ways
- Noise due to variations in clothing, background and occlusions

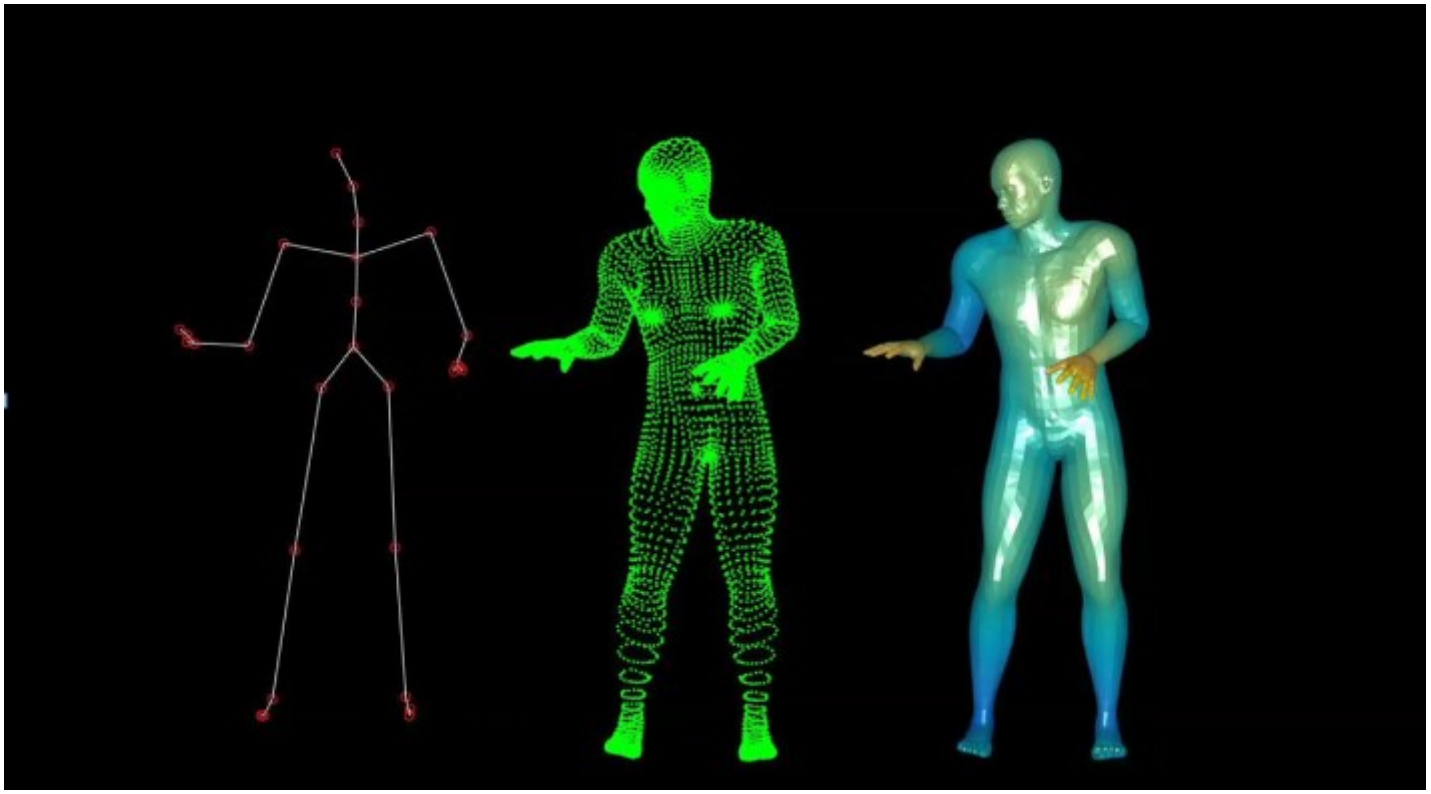
Deep Neural Networks (DNN)

- DNNs have the potential to model all these complex variations
- DNN learns a non-linear mapping between the input and output data in the form of millions of parameters
- However, to learn these parameters, DNN requires a large corpus of labelled training data
- Human action videos that are (1) multiview, (2) labelled and (3) large in quantity to train DNNs are not available and expensive to generate

Our solution to paucity of data

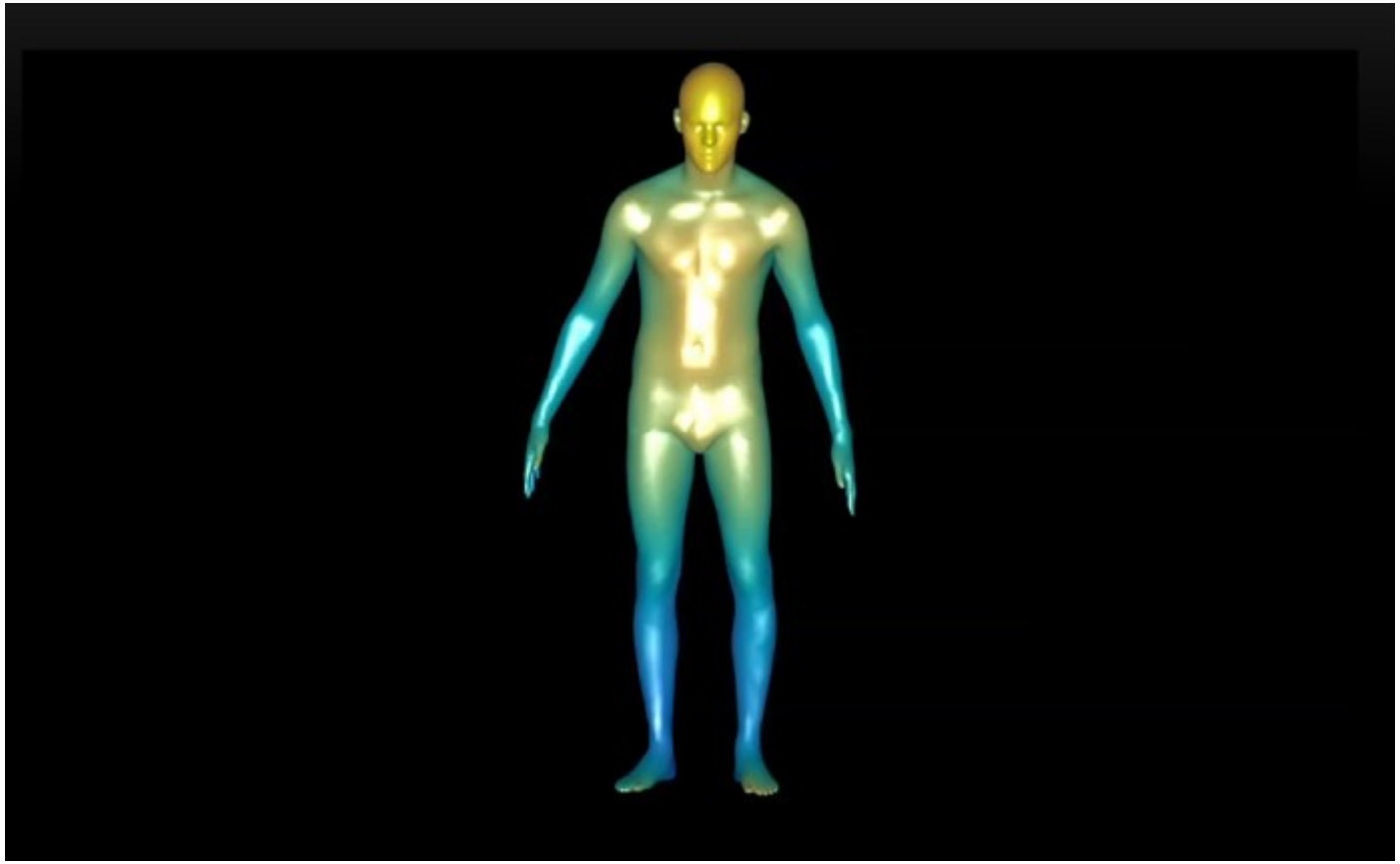
- Generate the data synthetically
  - Covers the multiview & quantity parts
- What about labels?
  - Use dummy labels.

Generating training data – Synthetically



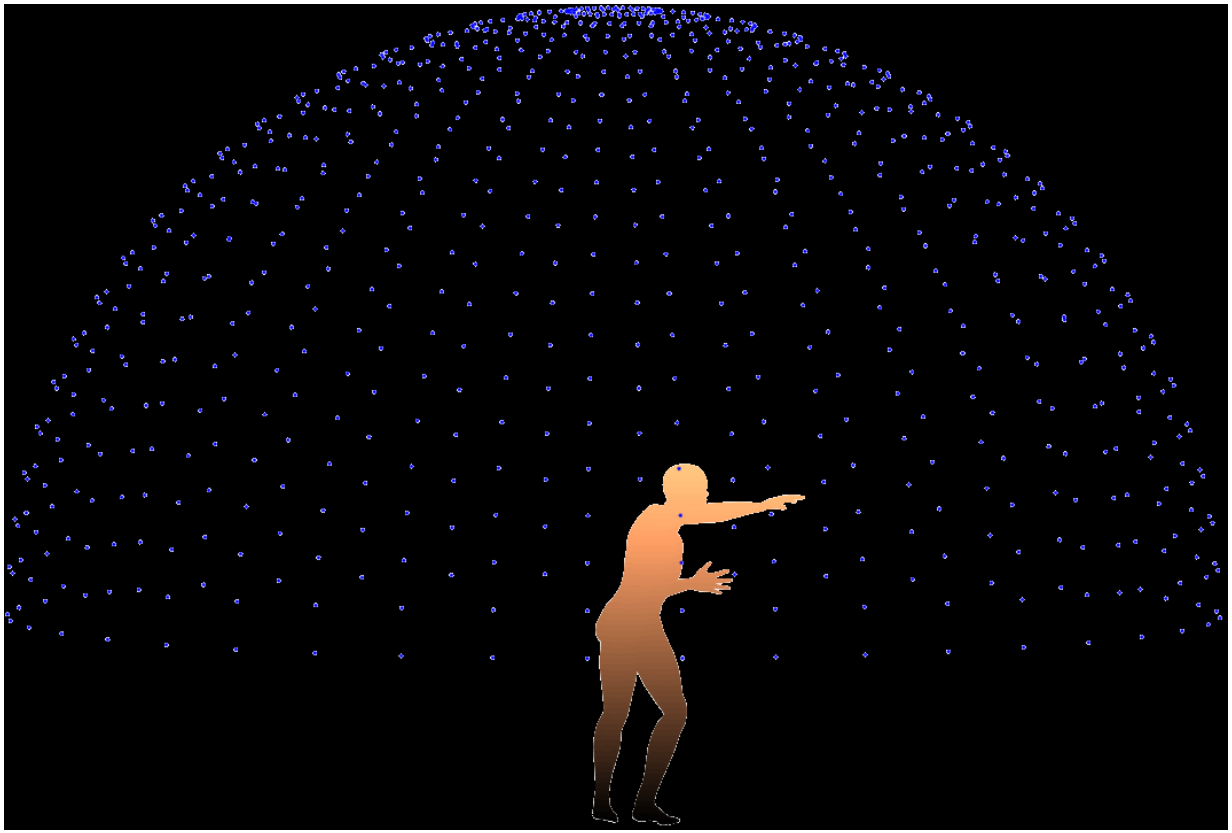
- We fit 3D human models to Motion Capture data to synthesize 3D videos

Simulating different body shapes



- We use different 3D human body shapes. Real or synthetic work equally well.

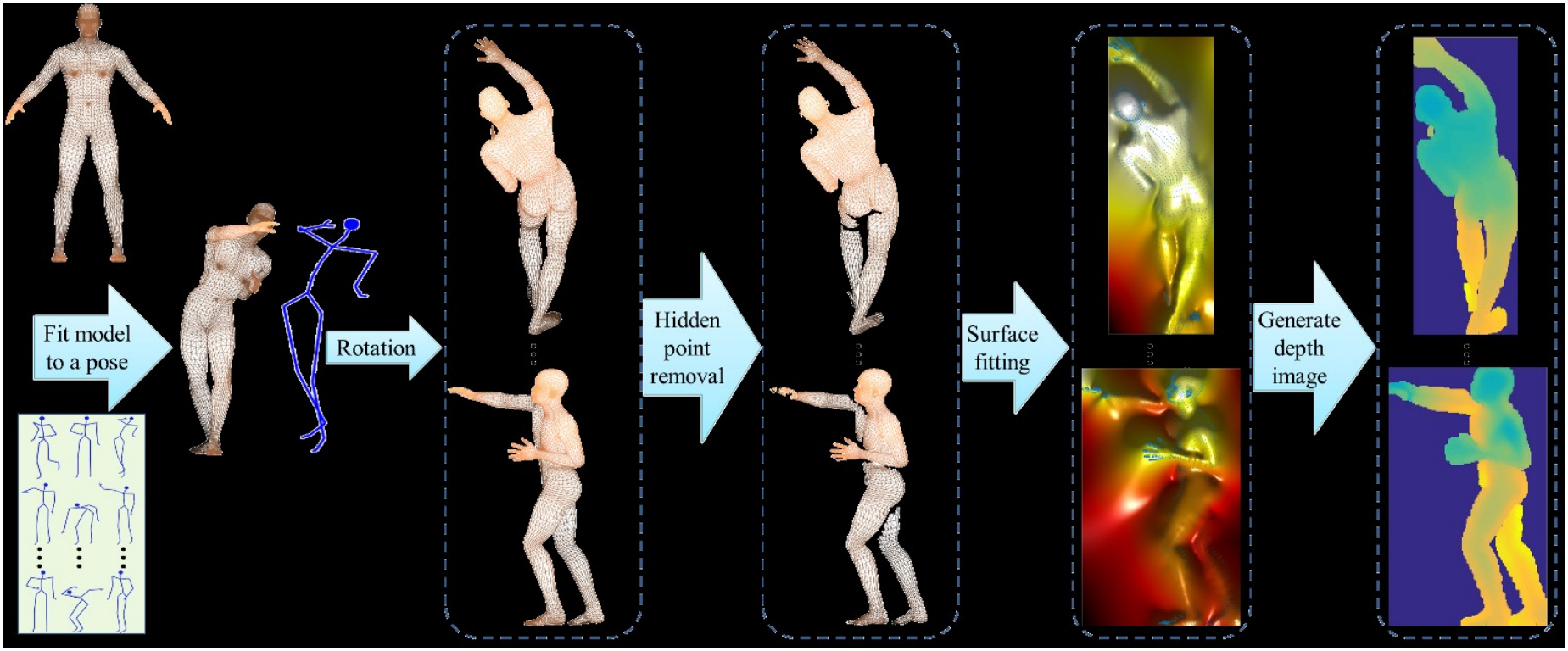
Simulating different camera veiwpoints



- We place many cameras on a hemisphere to capture different viewpoint images or videos

Feature extraction from 3d videos

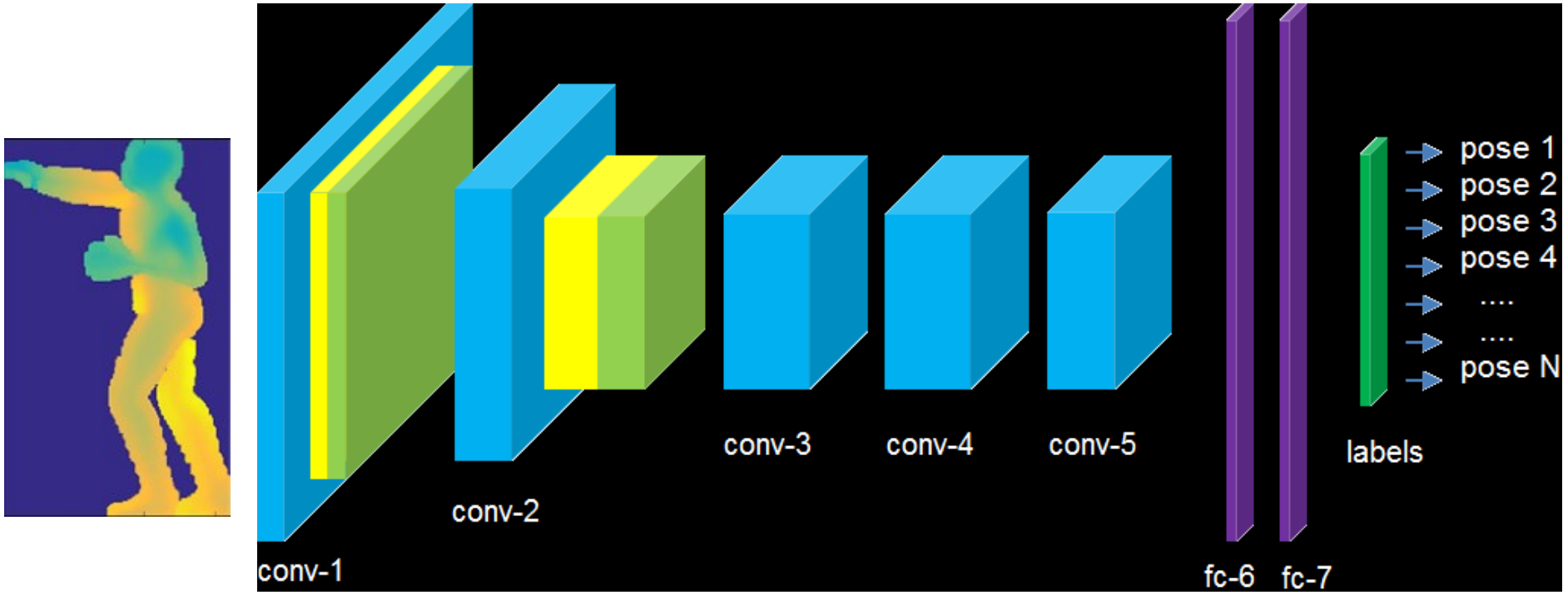
Generating Synthetic depth (3D) images



We choose typical human poses (~350) and generate depth images for each pose from a large number of camera viewpoints. These depth images resemble real data from 3D sensors.

3d Human Pose Model

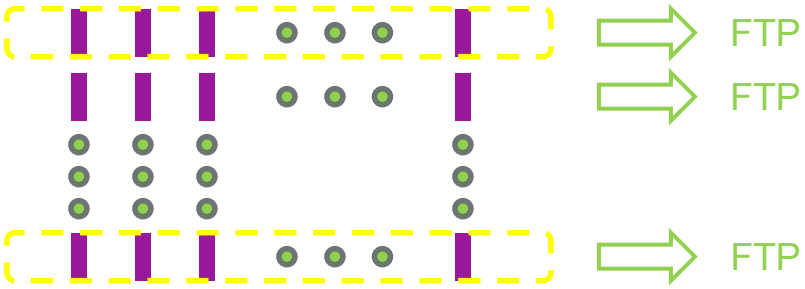
Convolutional Neural Network (CNN) is trained to map an input depth (3D) image, irrespective of the camera viewpoint, to one of N poses.



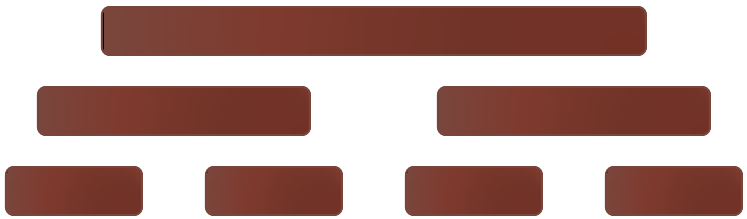
HPM: Human Pose Model

Modelling CNN features with Fourier Temporal Pyramid (FTP)

- CNN outputs a viewpoint invariant representation of the human pose
- A sequence of human poses defines action
- We perform Fourier analysis on each individual CNN fc-7 features



- FTP splits the sequence at each level, performs Fourier analysis on each individual split to find the low frequency components

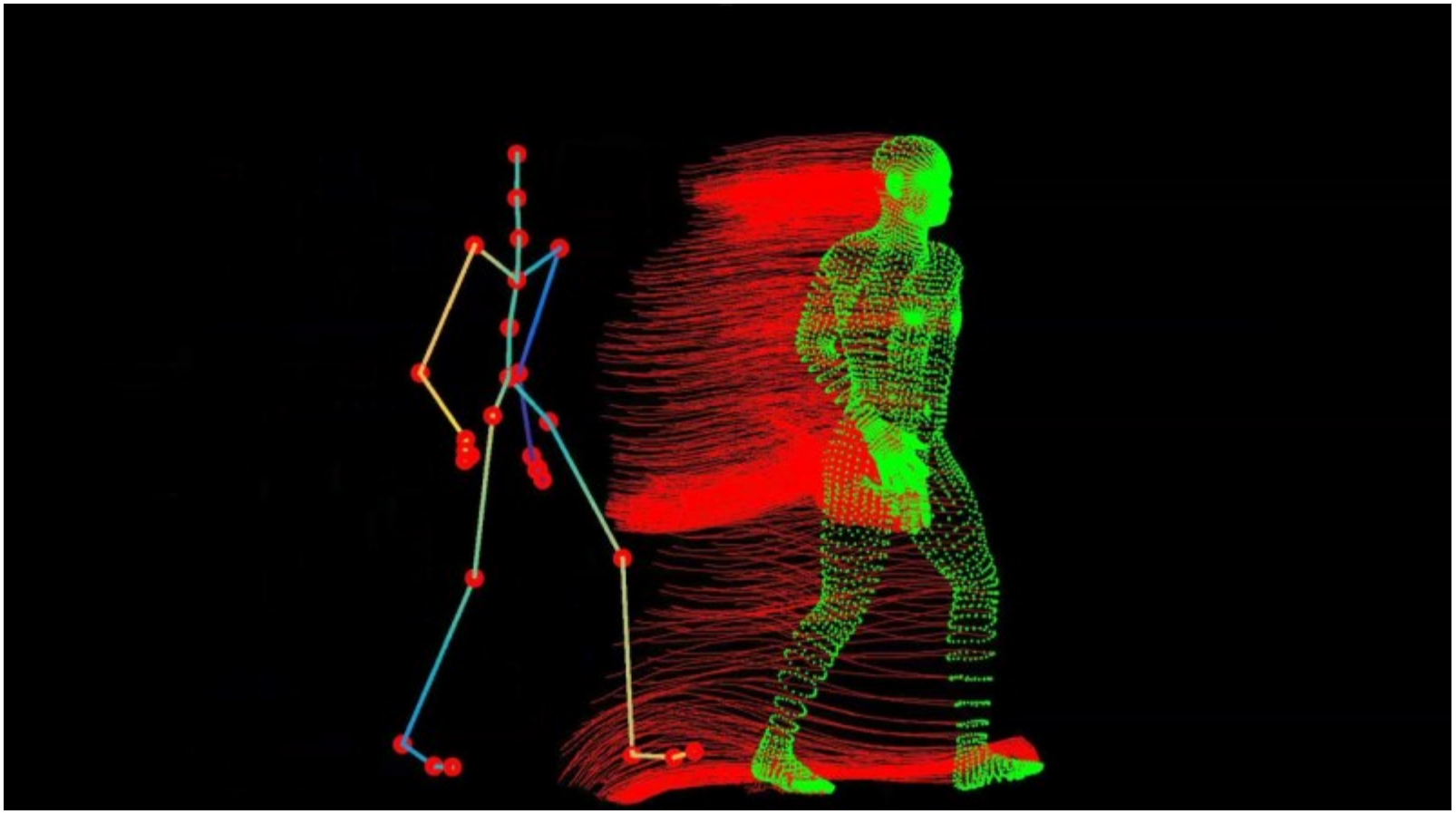


Feature extraction from 2d videos

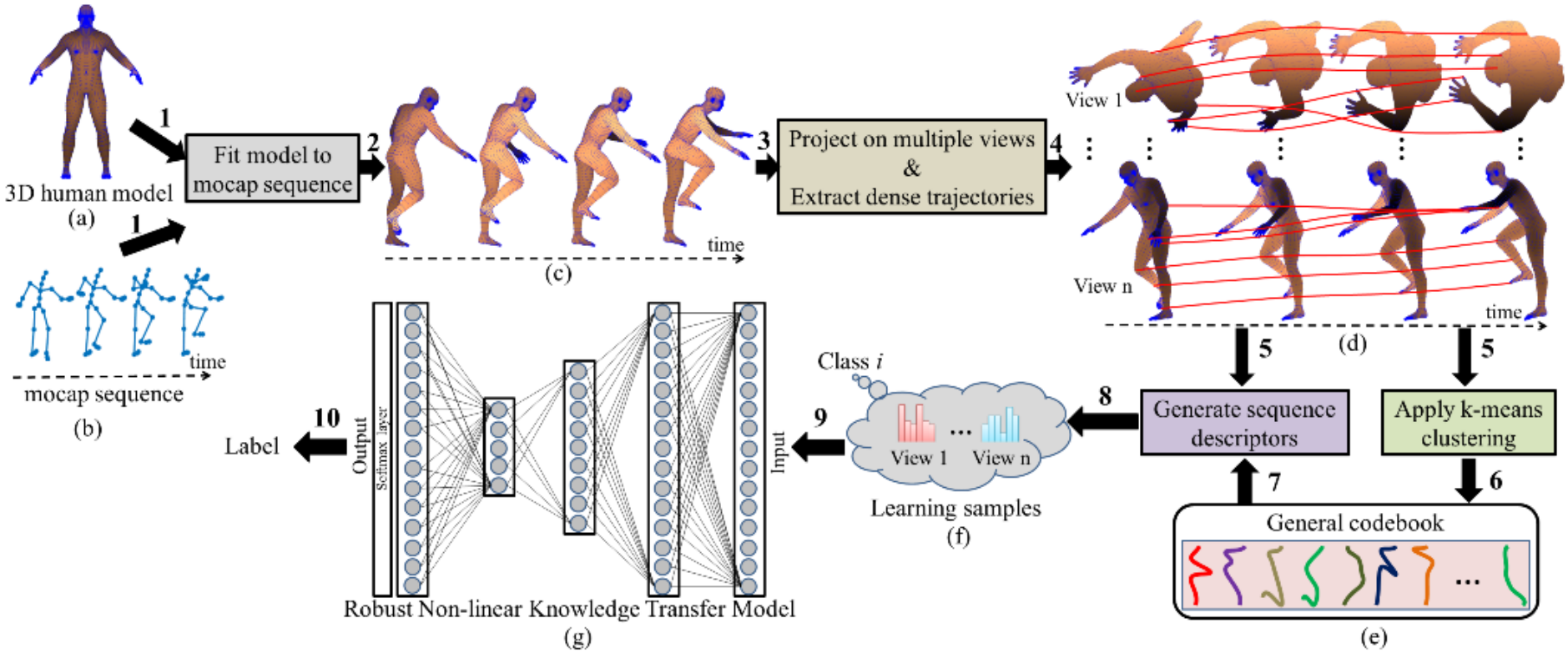
Human action recognition in conventional videos

- Can we synthesize data to train deep neural networks for action recognition in 2D videos?
- Motion trajectories in 3D videos projected to image planes resemble motion trajectories in 2D videos
- We fit 3D human models to Motion Capture (MoCap) data and find motion trajectories from 108 viewpoints
- MoCap sequences do not correspond to our actions of interest so we give them each a dummy label

From MoCap to 3D Trajectories

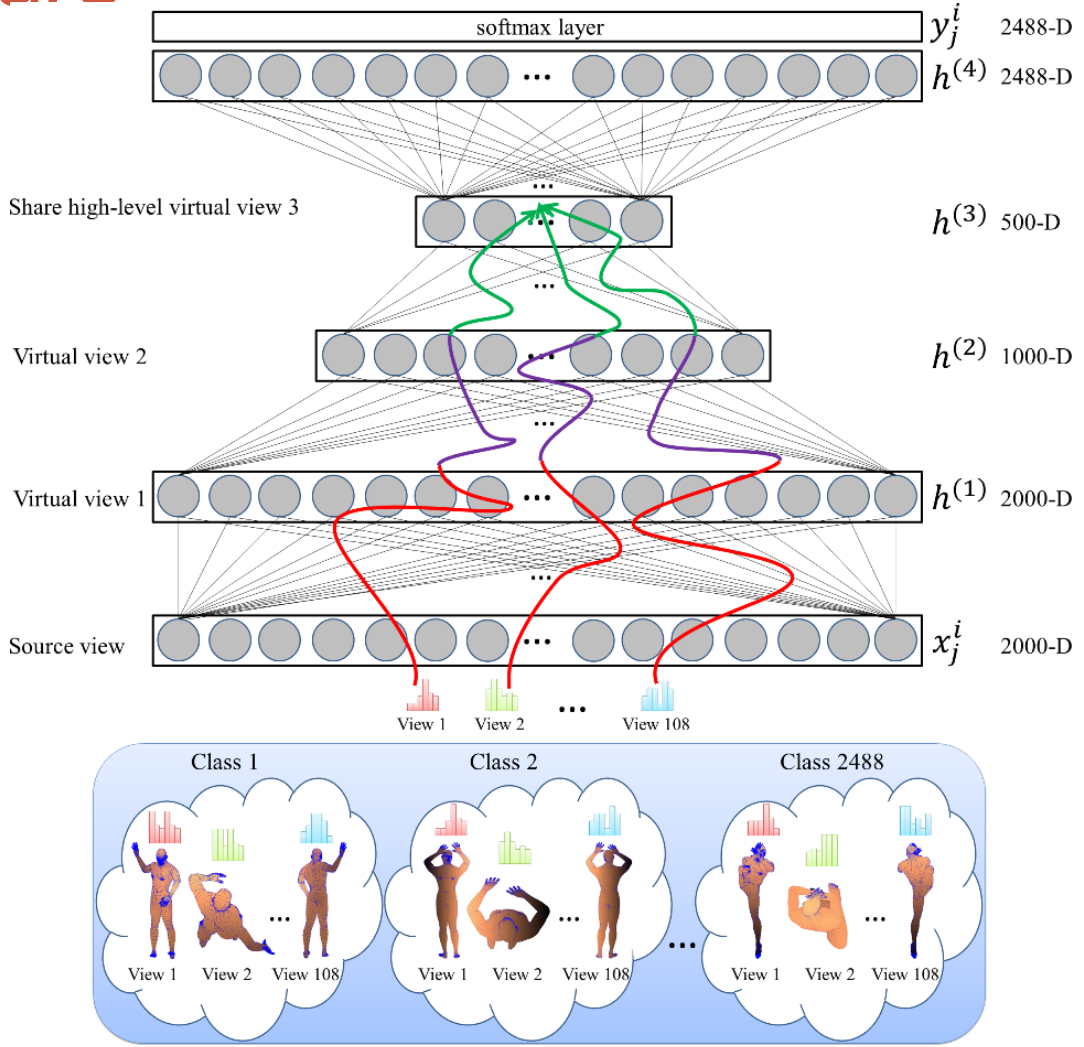


Complete pipeline for NN-Training from Synthetic dense trajectories



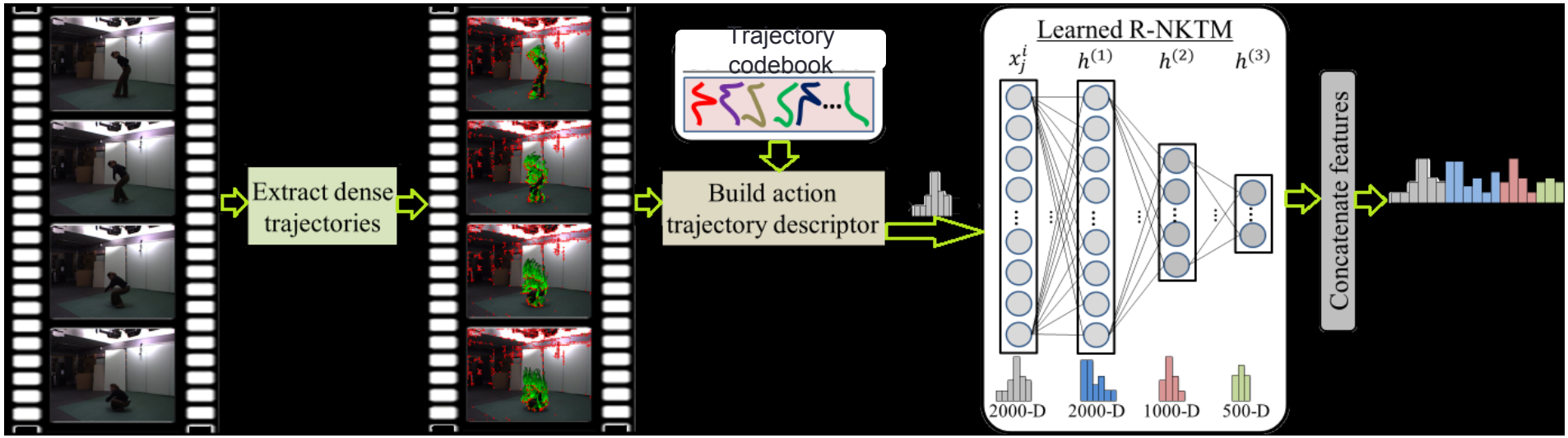
Network Architecture

- The network is trained using
  - Synthetic data
  - Dummy labels
- It learns to project each action, observed from any viewpoint, to a shared high level space
- We get the same high level features irrespective of viewpoint



R-NKTM: Robust Non-linear Knowledge Transfer Model

Feature Extraction from Real 2D Videos



Classification

- Our deep neural network models are trained on synthetic data with
  - Pose labels (image based) rather than action labels
  - Dummy labels (video based) that correspond to random actions

- Therefore, we use the deep models as feature extractors and then train another classifier on these features using real video labels
  - Support Vector Machines
  - L_1L_2 Classifier

The N-UCLA Benchmark dataset

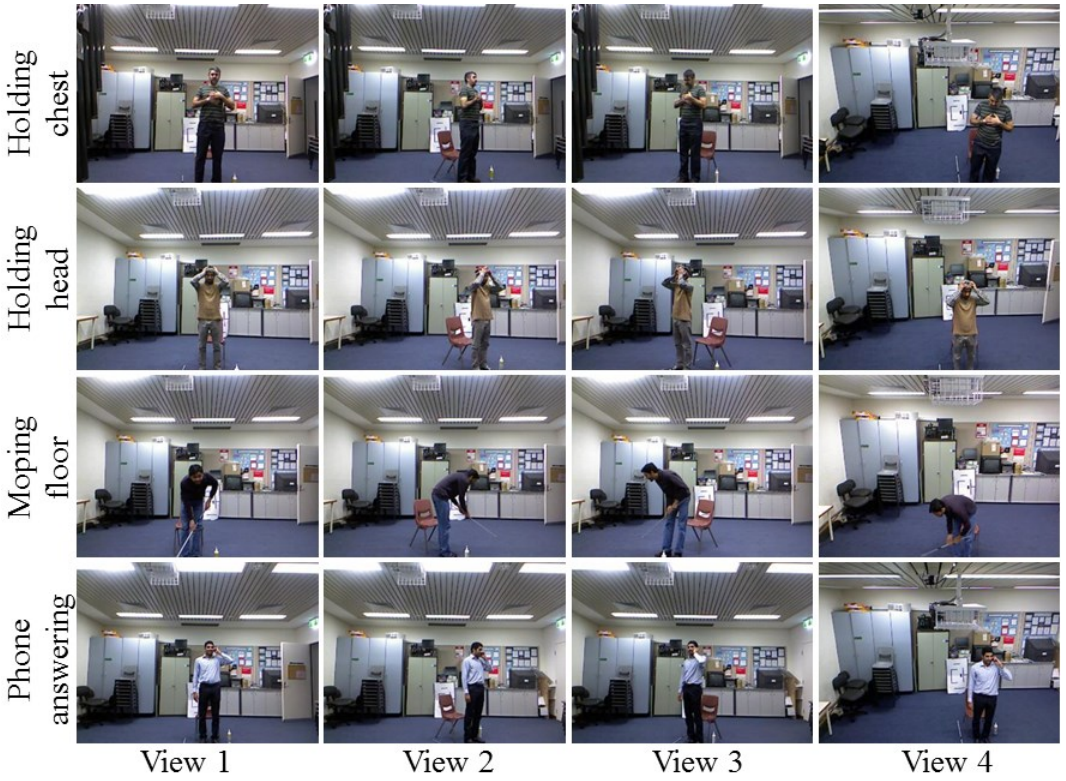
- North-Western University of California Los Angeles

- 10 subjects
- 20 actions
- 3 camera viewpoints
- RGB-D videos (Kinect-1)
- 640×480 RGB resolution
- 320×240 Depth resolution

The UWA3D-II Benchmark Dataset

- University of Western Australia

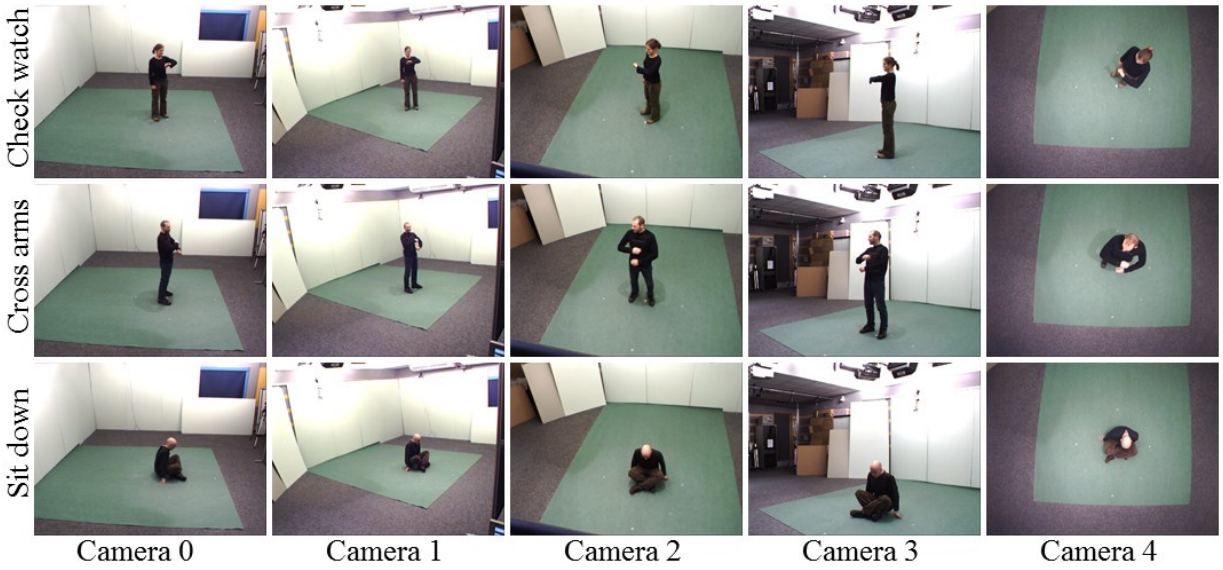
- 20 subjects
- 30 actions
- 4 camera viewpoints
- RGB-D videos (Kinect-1)
- 640×480 RGB resolution
- 320×240 Depth resolution



The ixmas Benchmark Dataset

- INRIA Xmas Motion Acquisition Sequence

- 11 subjects
- 11 actions
- 5 camera viewpoints
- RGB videos
- 390×291 resolution



Comparative Results (2D Only)

Mean accuracy (%) on the three datasets when test camera view is not included in the training data.

Method	**Published year**	**IXMAS** (all cams)	**UWA** (all cams)	**N-UCLA** (all cams)
Dense Trajectories	2011	61.7	60.4	72.7
Hankelets	2012	56.4	51.8	45.2
Non-linear Circulant Temporal Encoding	2014	67.4	61.2	68.6
3D Convolutional Neural Networks	2013	62.8	61.0	
Two-stream Convolutional Networks	2014	42.5	57.6	
Our method (RNKTM)	2016	74.1 %	67.4 %	78.1 %

Comparative results (3d only)

Mean accuracy (%) on the two datasets when test camera view is not included in the training data.

Method	Published year	Data Type	UWA (all cams)	N-UCLA (cam-3 only)
ActionLet	2013	Joints	39.8	76.0
3D Skeleton Points in Lie Group	2014	Joints	43.4	74.2
Histogram of Oriented 4D Normals	2013	Depth	28.9	39.9
Continuous Virtual Path	2013	Depth	25.6	53.5
Histogram of Oriented PCs	2016	Depth	52.2	80.0
Our method (HPM+TM)	2016	Depth	76.9	92.0

Comparative results (2D + 3d)

Mean accuracy (%) on the two datasets when test camera view is not included in the training data.

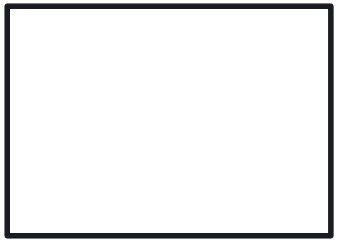
2D and 3D features were combined to train an $L_1 L_2$ classifier

Type of Data	Neural Network Model	UWA (all cams)	N-UCLA (all cams)
2D videos	NKTM (fully connected)	67.5	78.1
3D videos	Human Pose Model (CNN)	76.9	79.7
Combined	NKTM+HPM+TM	84.1	83.3

Conclusion

- Existing MoCap data can be exploited to synthesize videos for training deep neural networks
- Dummy labels can be used to train deep models that are good feature extractors
- 3D data improves human action recognition accuracy
- Future research may find new uses for legacy MoCap data that has been routinely collected for decades at universities and sports research institutes
- Our future research aims at combining MoCap data from multiple institutions to perform fine grained action analysis for athlete performance enhancement

Acknowledgements



- This research is supported by the Australian Research Council Discovery Project (DP160101458)
- MoCap data was obtained from the Carnegie Mellon University, USA

- Deep learning was performed in MATLAB using the Matconvnet library from the University of Oxford
- Related publications:

[1] Hossein Rahmani, Ajmal Mian and Mubarak Shah, "*Learning a deep model for human action recognition from novel viewpoints",* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2017.

[2] Hossein Rahmani and Ajmal Mian, "*3D Action recognition from novel viewpoints",* IEEE Conference on Computer Vision and Pattern Recognition (CVPR-Oral), 2016.