Research @ VIGIL

C. Krishna Mohan

Visual Learning and Intelligence Group (VIGIL)

Dept. of Computer Science & Engineering

Indian Institute of Technology Hyderabad

भारतीय प्रौद्योगिकी संस्थान हैदराबाद

Indian Institute of Technology Hyderabad

Human Action Recognition in videos

Research problem

Action recognition aims to identify the goals of one or more agents, from a sequence of observations of the agent(s) movements in an environment.

Challenges:

- **sequence of observations ⟶ motion information used for action recognition ⟹ *issues related to motion***
- **agent(s) movements in an environment ⟹ *issues affecting the appearance of subject(s) in the environment***
- **observations of the agent(s) ⟹ *issues related to capturing devise***

Significance of Human Action Recognition

Applications

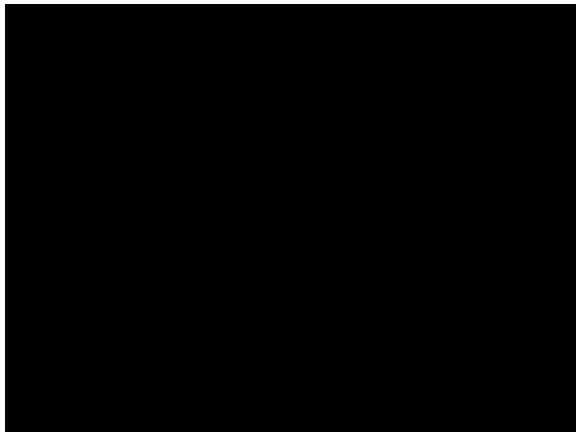
Fall detection

http://makeagif.com/e_Zwx4?

Robotic Natural User Interface

Source: https://www.youtube.com/watch?v=WPZaGLFS9fA

Human computer Interaction



Source: https://www.youtube.com/watch?v=r7WI2KhWZCk

For efficient analysis and categorization of videos

Rapid growth in video content: ~500 Hours of Video Uploaded/Minute

Efficient Analysis of videos

Source: https://people.eecs.berkeley.edu/~gkioxari/ChainModels/index.html

Video Categorization by recognizing actions like **violent scenes** and **human interactions**

Source: http://www.robots.ox.ac.uk/~alonso/tv_human_interactions.html

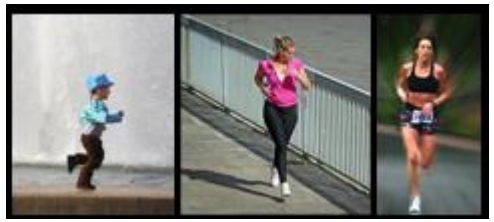
Challenges in Human Action Recognition

Motion related issues

inability to **repeat identical movement**

alternative movements to perform the same action

gait characteristics of subjects



Appearance related issues

complexity of the **background**

non-uniform **illumination** of the subject

variation in **appearance** across subjects

Sources: http://www.di.ens.fr/willow/research/stillactions/, http://www.enjoy-swimming.com/swimming-strokes.html http://vision.stanford.edu/Datasets/40actions.html

Issues due to capturing conditions

distance from the camera change in **angle-of-view**

occlusion by obstacles.

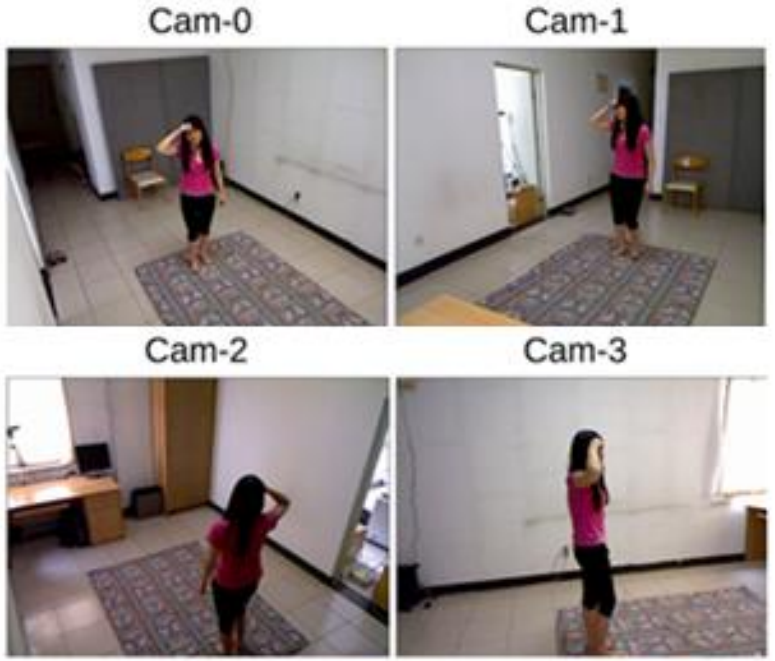
Major factors affecting human action recognition

modality in which the observations are captured & its noise.

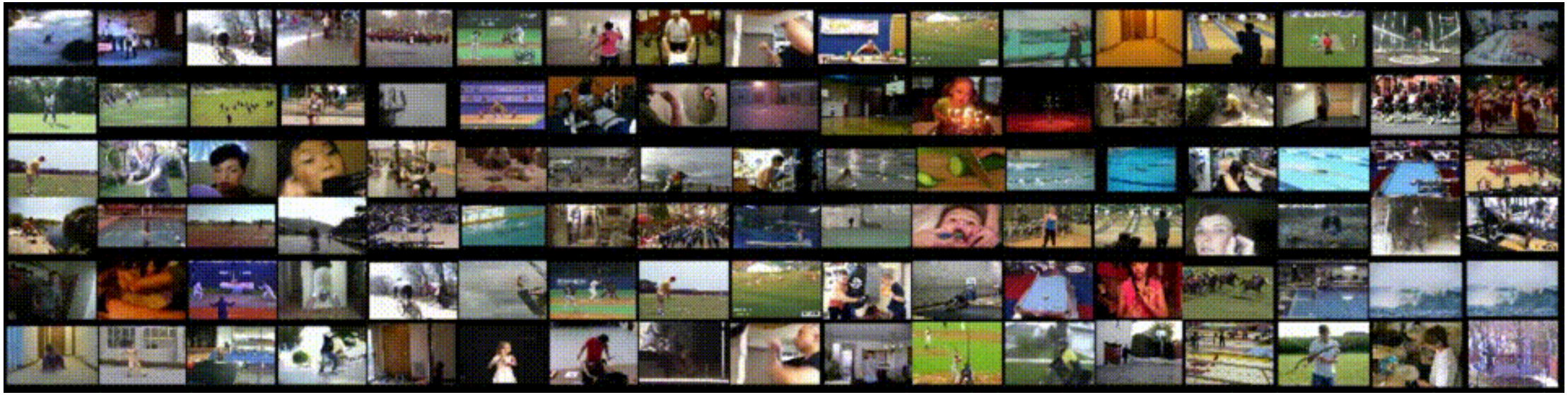
Eg: RGB, Depth, Skeletal

No. of subjects in the action & angle of video

No. of cameras used to capture observations



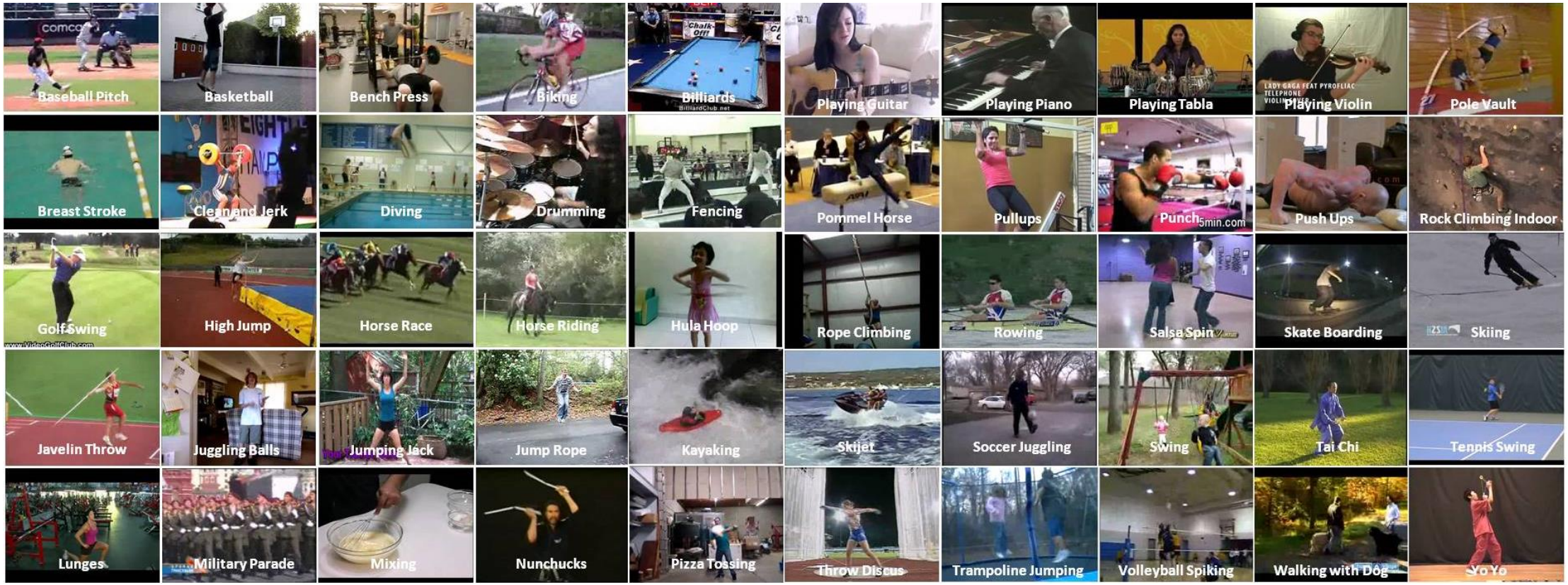
Issues of human action recognition in videos



- Variation in category of action
  - Human-Object Interaction
  - Body-Motion only
  - Human-Human Interaction
  - Playing Musical Instruments
  - Sports
- Camera motion
- Angle of view
- Region of interest
- Background
- Occlusions
- Illumination conditions
- Variation in clothing
- Low resolution of video
- Speed of action execution
- Alternative movements for action

Action Recognition Dataset: **UCF101**

- University of Central Florida
- 101 realistic human actions
- 13000+ videos are collected from YouTube
- large no. of actions



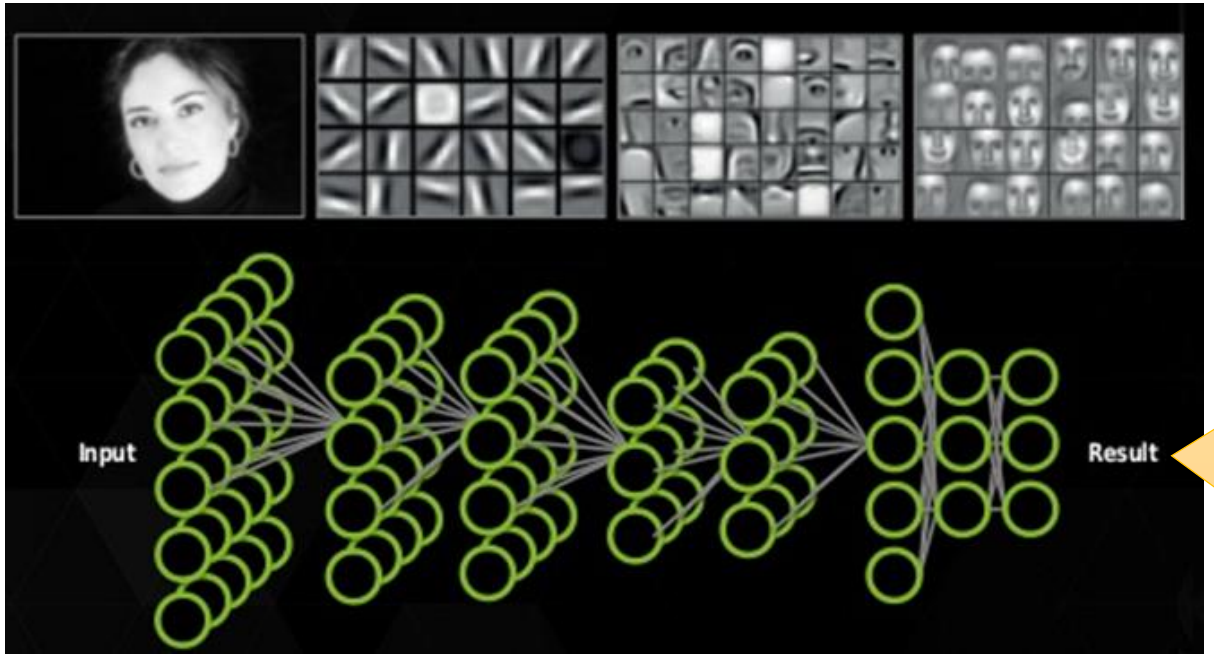
Action Recognition Dataset: **HMDB51**

- Human Motion Data Base (MHDB51)
- 51 realistic human actions
- 7000+ clips

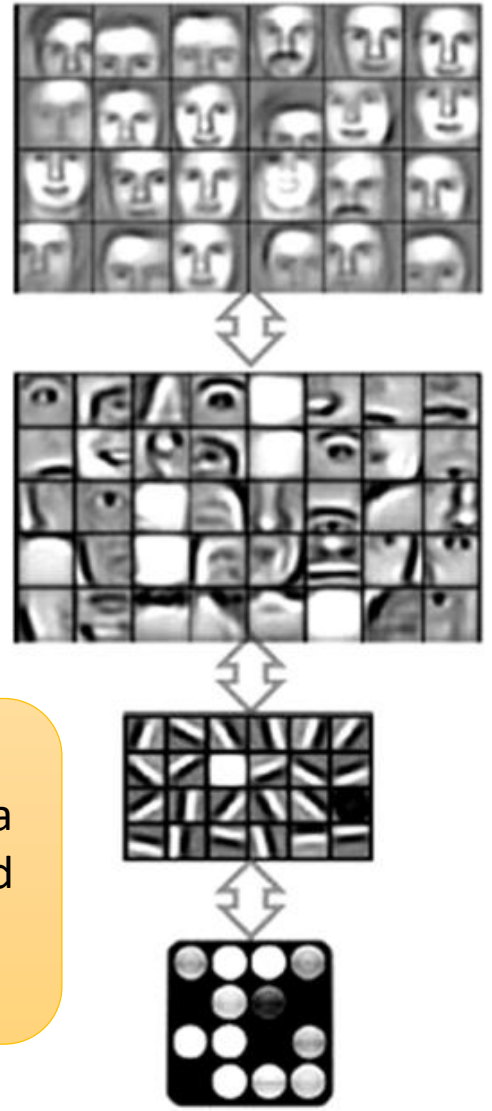
Deep learning for hierarchical features learning

Convolutional neural network (CNN)

- a variation of multi-layer neural network
- used to learn a hierarchy of features from input data
- trained using a variant of back-propagation algorithm
- convolution masks are learnt through supervised training

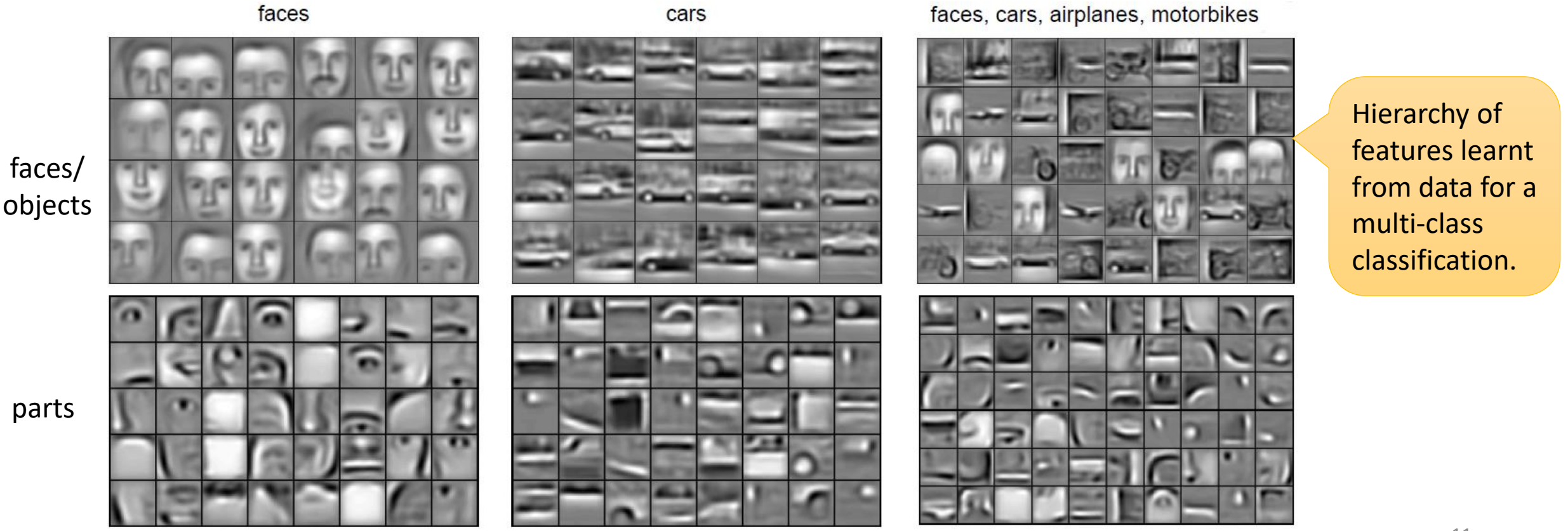


For parallel computation Tesla K40 GPU, received under NVIDIA Hardware Grant



Features for multi-class recognition

The deep learning model automatically learns the discriminative features necessary for classification of input images.



Convolution Neural Networks (CNNs)

- Focus on the structural layout of the multilayer perceptron
- These networks are inspired by the idea of existence of complex arrangement of cells within the visual cortex.
  - These cells are sensitive to small sub-regions of the input space, called a **receptive field**, and are tiled in such a way as to cover the entire visual field.
  - These filters are local in input space and are thus better suited to exploit the strong spatially local correlation present in natural images.
- Convolution networks are designed specifically to recognize 2-dimensional visual patterns with a high degree of invariance to translation, scaling, skewing and other forms of distortion.
- Constraints are imposed on the structure of the network and all the weights are learned through training (supervised learning)

Working principle of CNNs

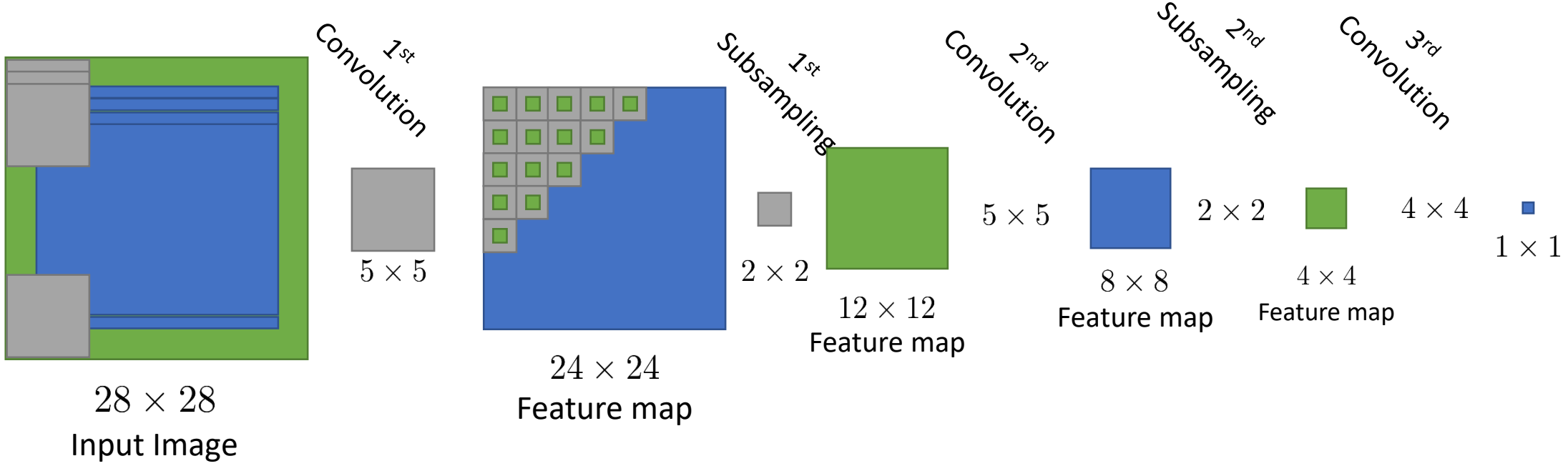
C1: Convolution with receptive field size of 5x5

S1: Subsampling and local-averaging with receptive field size of 2x2

C2: Convolution with receptive field size of 5x5

S2: Subsampling and local-averaging with receptive field size of 2x2

C3: Convolution with receptive field size of 4x4



Robustness of ConvNet features

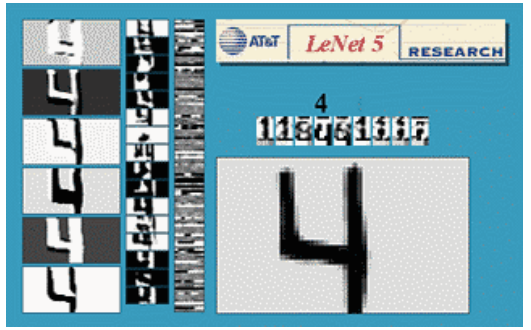
Addresses following issues in input representation for MNIST digit recognition

■ Shift ■ Rotation ■ Scale ■ Width ■ Stretch/Squeeze ■ Random noise ■ Noisy patterns

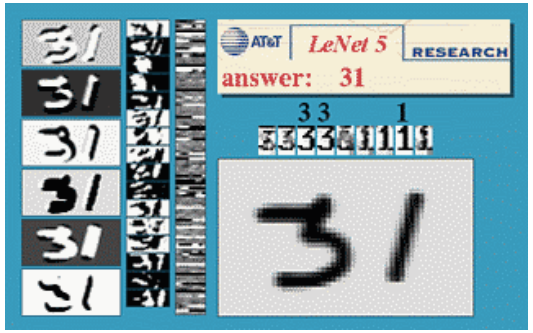
Shift invariance

Rotational invariance

Scale invariance

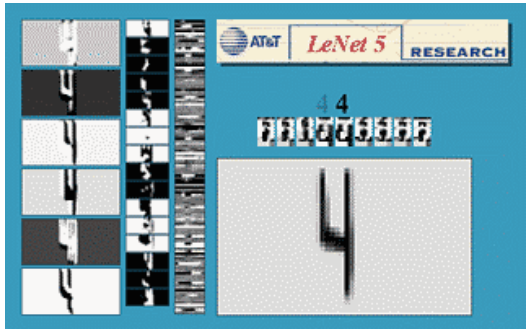


Partial patterns

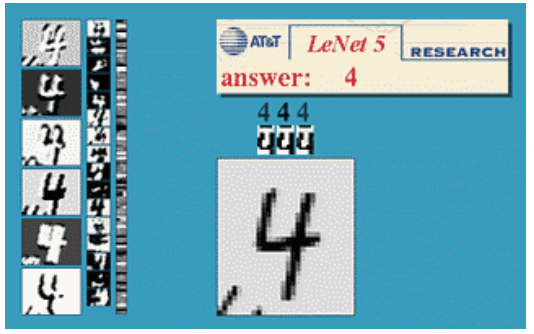


Width of pattern

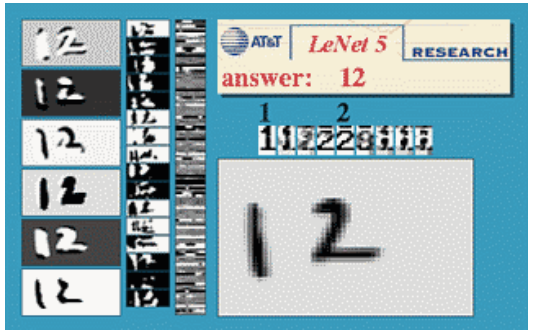
Stretch/Squeeze



Random noise



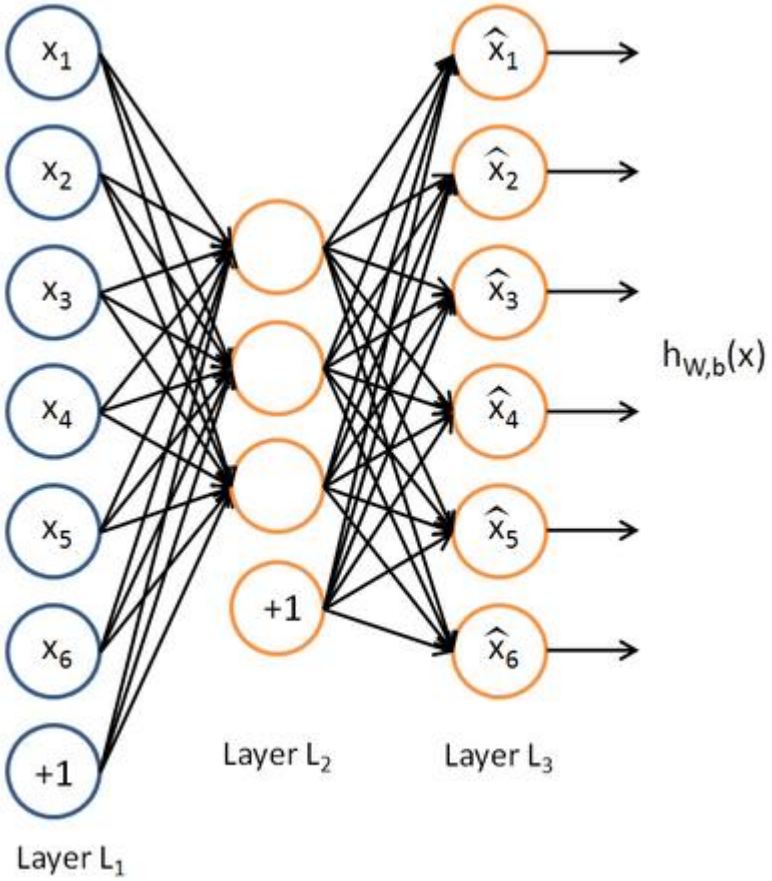
Overlapping patterns



Source: http://yann.lecun.com/exdb/lenet/

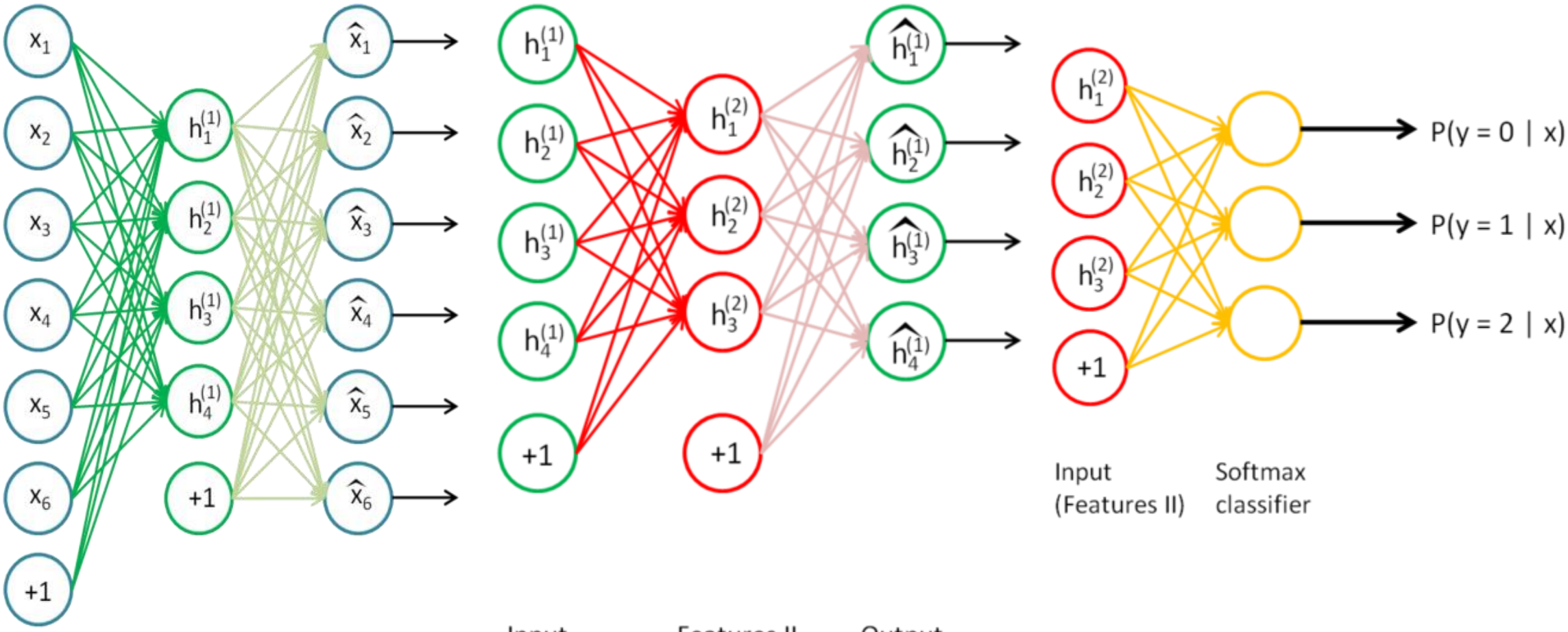
Auto-encoders

An auto-encoder neural network is an supervised learning algorithm that applies backpropagation, setting the target values to be equal to the inputs.



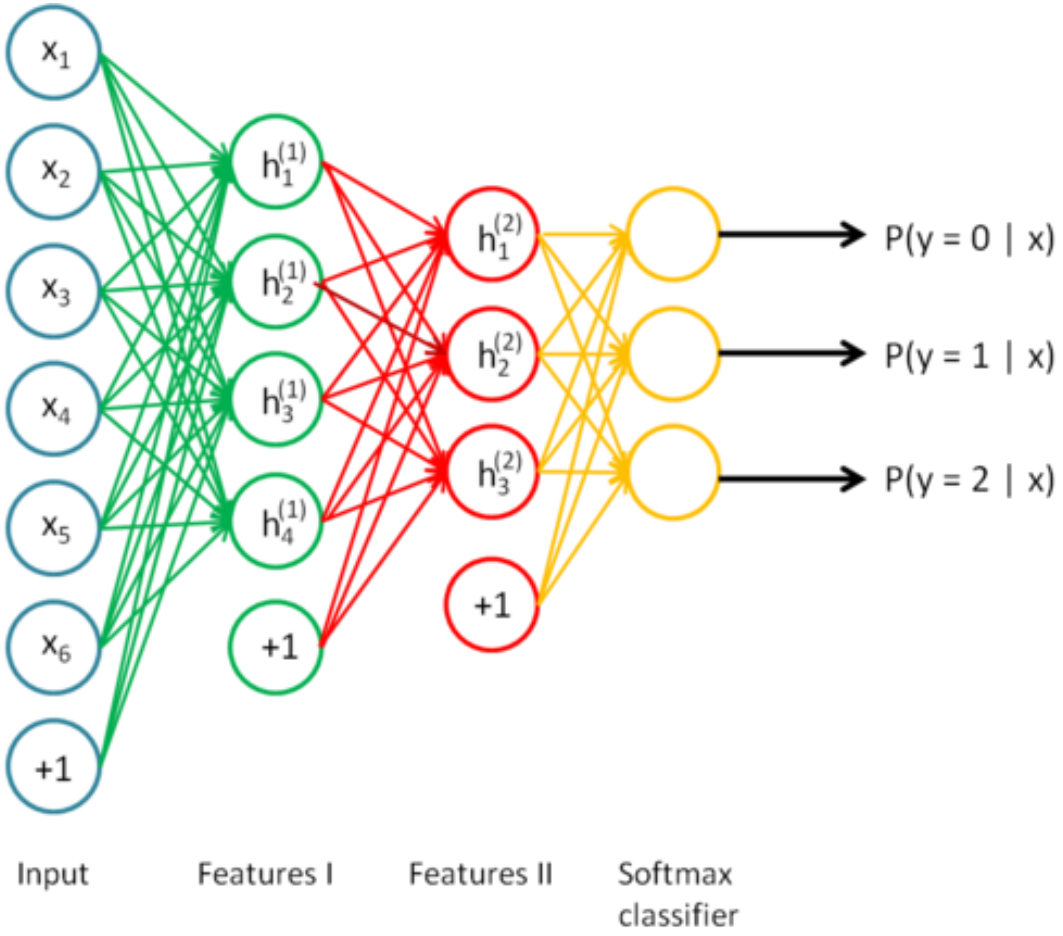
http://deeplearning.stanford.edu/wiki/index.php/Autoencoders_and_Sparsity

Greedy layer-wise training of autoencoder



http://ufldl.stanford.edu/wiki/index.php/Stacked_Autoencoders

Stacked Auto-encoder



- It captures a useful "hierarchical grouping" or "part-whole decomposition" of the input.
- The first layer of a stacked autoencoder tends to learn first-order features in the raw input (such as edges in an image).
- The second layer of a stacked autoencoder tends to learn second-order features corresponding to patterns in the appearance of first-order features (e.g., in terms of what edges tend to occur together--for example, to form contour or corner detectors).
- Higher layers of the stacked autoencoder tend to learn even higher-order features.

http://ufldl.stanford.edu/wiki/index.php/Stacked_Autoencoders

Classification of human actions using pose-based features and stacked auto encoder

Associated publications

(IDRBT Doctoral Colloquium 2016)

(PRL Nov 2016, 10.1016/j.patrec.2016.03.021)

(ICAPR 2015, 10.1109/ICAPR.2015.7050706)

(IMCLA 2014, 10.1109/ICMLA.2014.30)

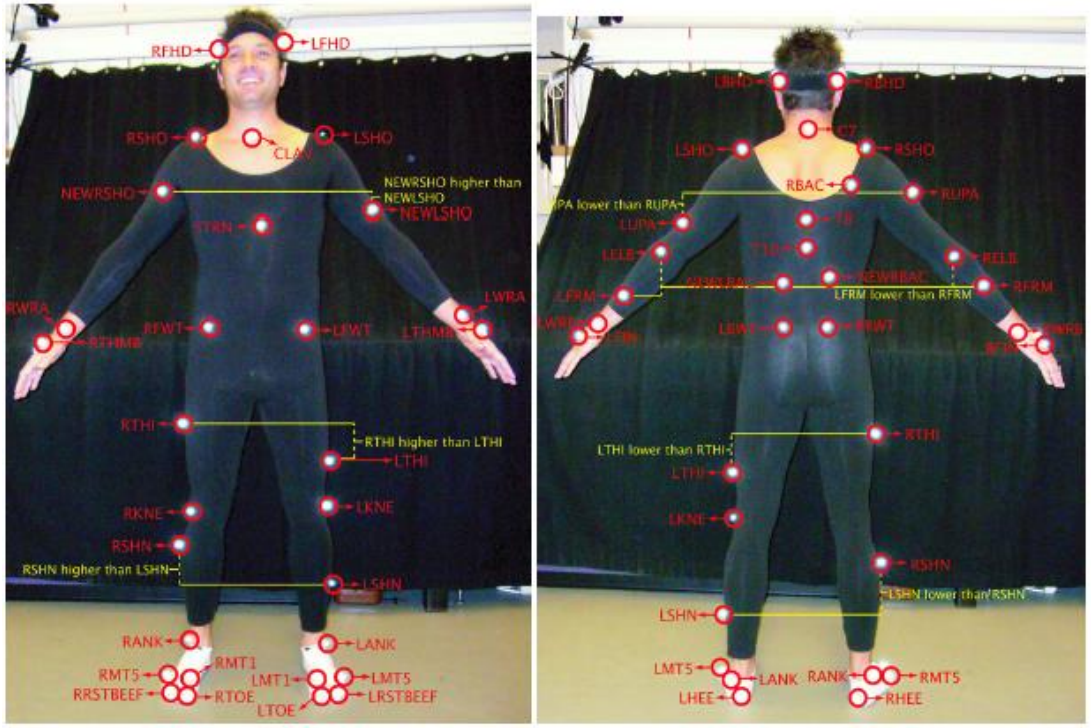
(IMCLA 2014, 10.1109/ICMLA.2014.69)

ISSUES ADDRESSED

- ✓ Motion related
- ✓ Appearance related
- ✓ Capturing related

MOCAP information

Marker based approaches:



Due to accurate tracking, marker based MOCAP information is considered for action recognition.

Marker less approaches:

DeepPose

Source: http://people.eecs.berkeley.edu/~gkioxari/ChainModels/out596.gif

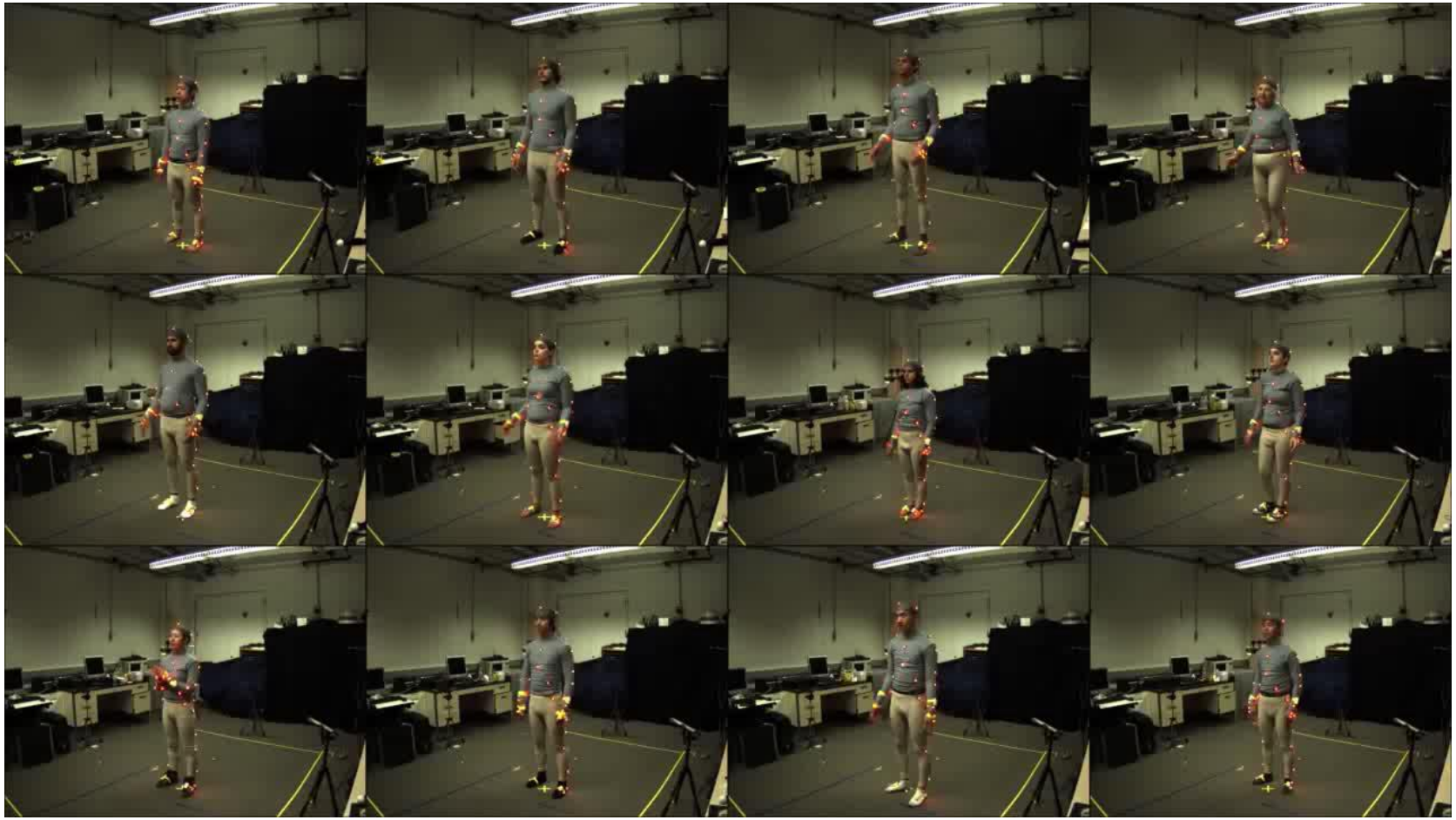
Kinect skeletal tracking

Source: https://praveenitech.wordpress.com/tag/kinect-sdk/

Source: http://thomas.lewiner.org/imgs/gesture_learning_sibgrapi.gif

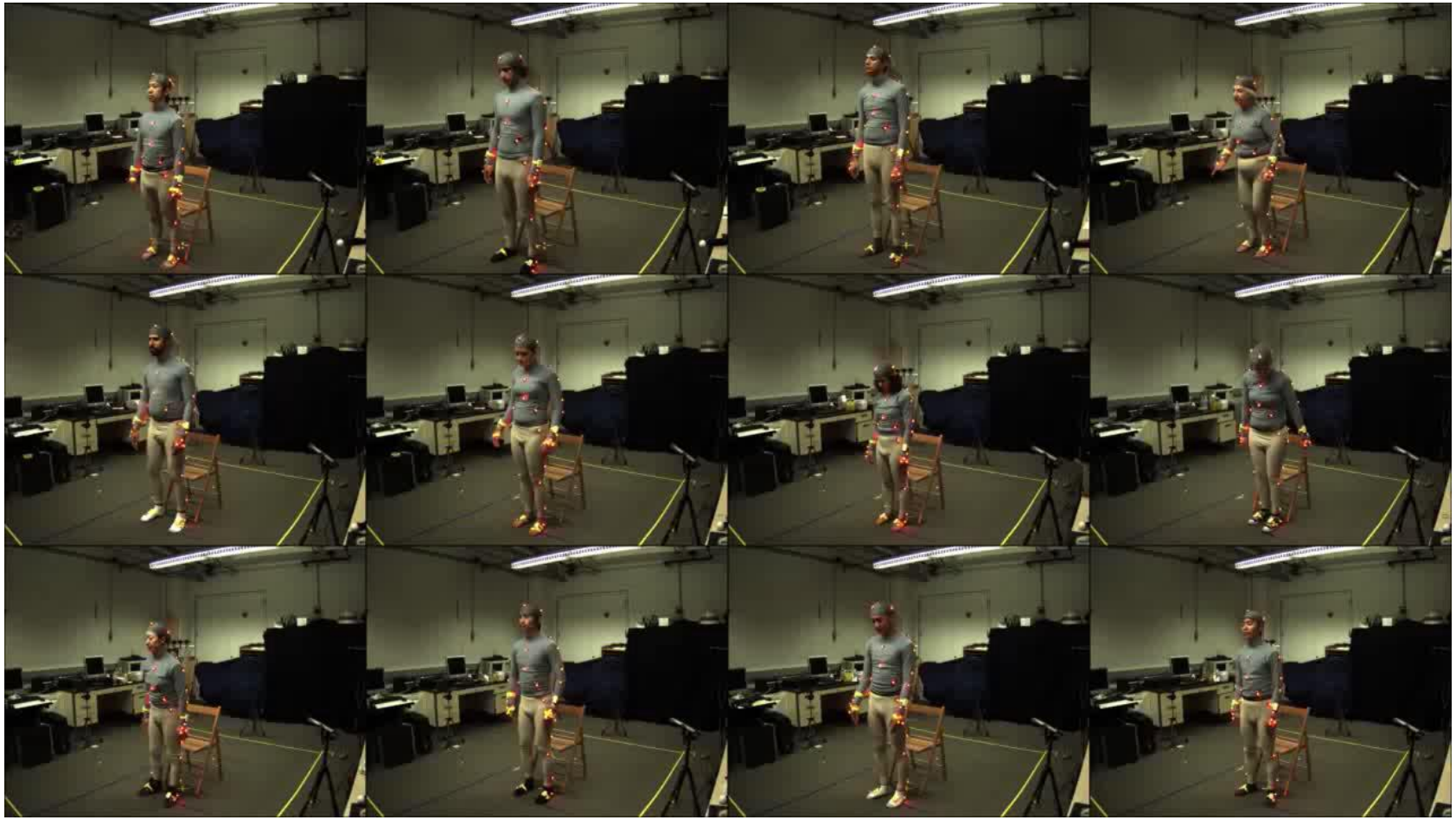
Challenges

- Height variation across subjects
- Speed of execution of action



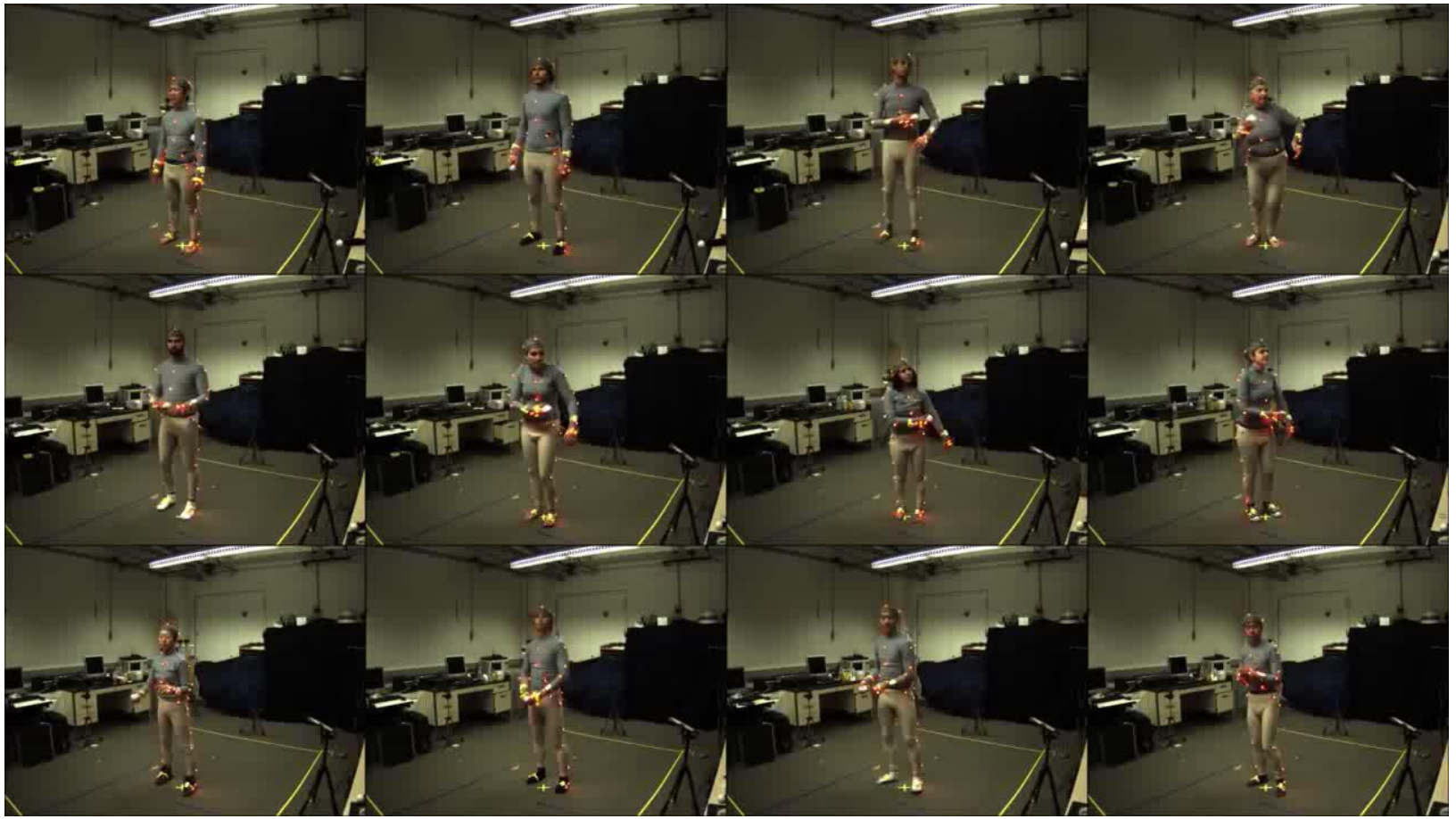
Challenges

- Variation in limb movements

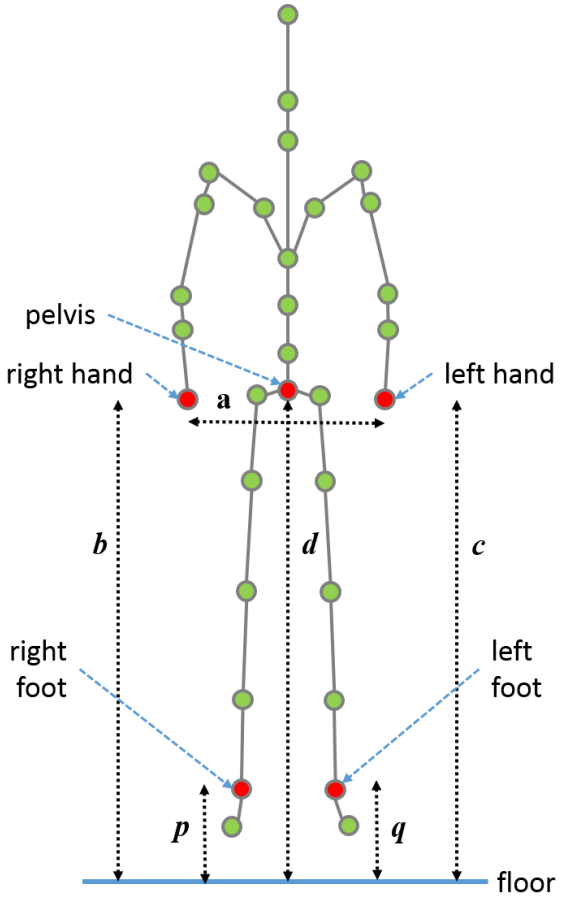


Challenges

- Beginning of action
- Duration of the action



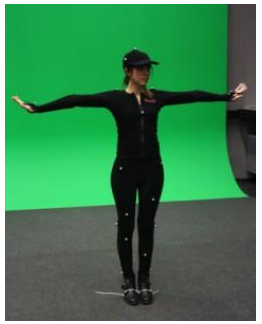
Pose-based distances and their normalization



NORMALIZATION OF MEASURES a, b, c AND d

Measure	Normalization
a	divided by the distance between the hips
b	divided by the height of left shoulder
c	divided by the height of right shoulder
d	subtract and divide by the value of d in T-pose

To address variation in subject GAIT characteristics

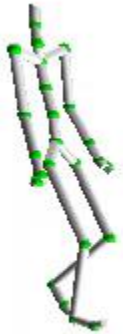


Source: http://sgruenvo.web.th-koeln.de/motion-capturing/

Recognizing locomotion actions

Running

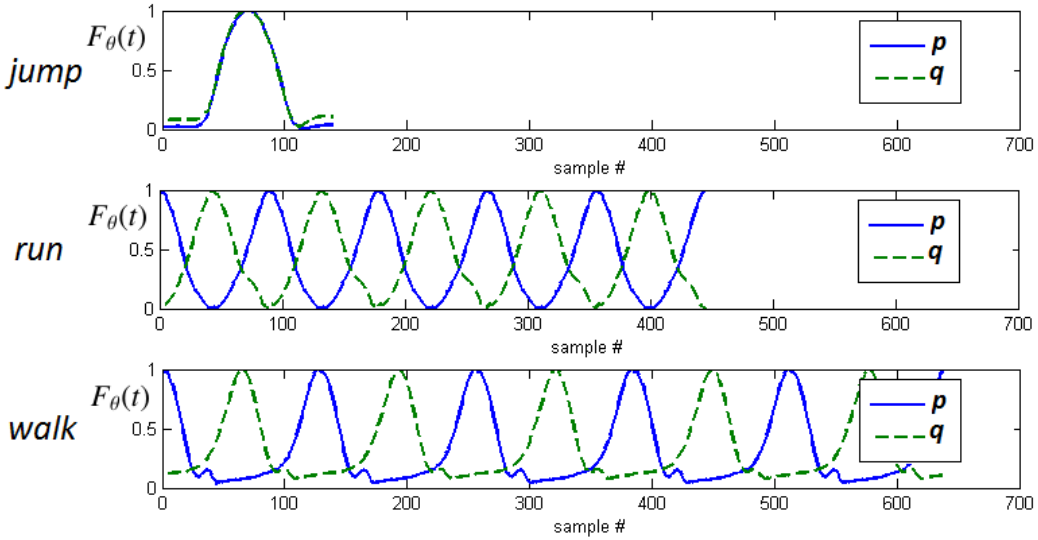
Walking



Jumping

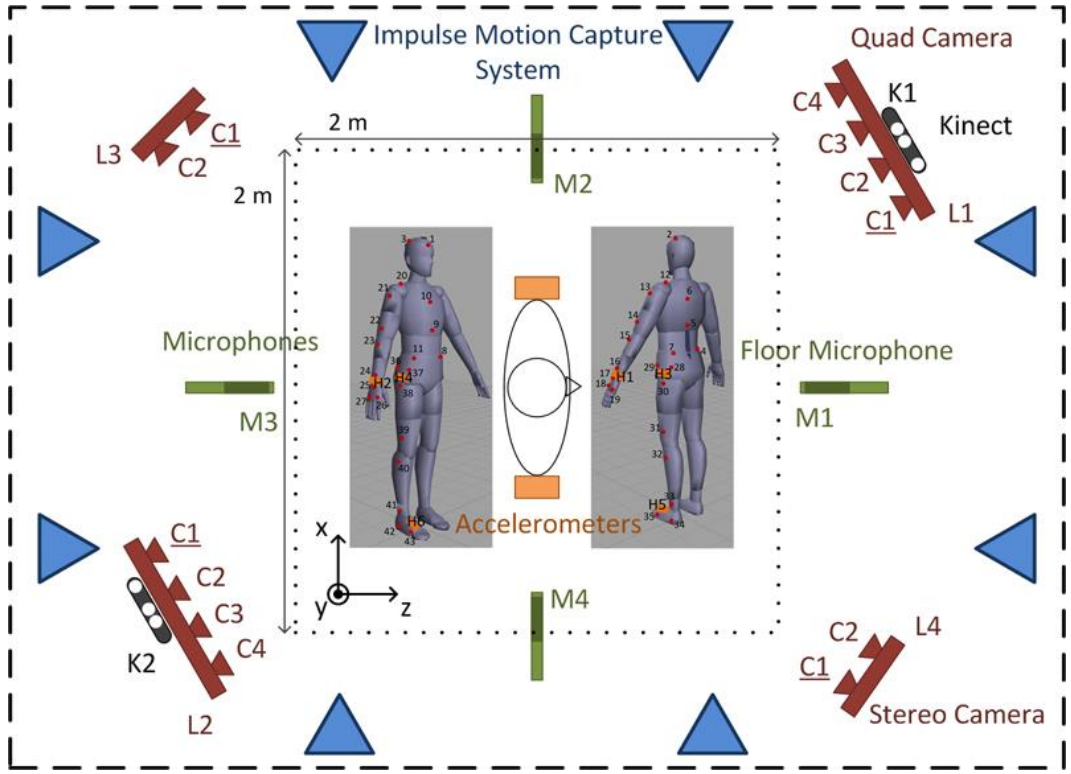
To address variations across observations

$$F_\theta(t) = \begin{cases} 0 & if,\ \theta(t) \le 0 \\ \frac{\theta(t)}{\theta_{max}} & if,\ 0 < \theta(t) < \theta_{max} \\ 1 & if,\ \theta(t) \ge \theta_{max} \end{cases}$$



Berkeley Multimodal Human Action Database (MHAD)

- 11 actions
- Performed by 12 subjects
- Modalities: RGB, Depth, MOCAP, IMU, Audio



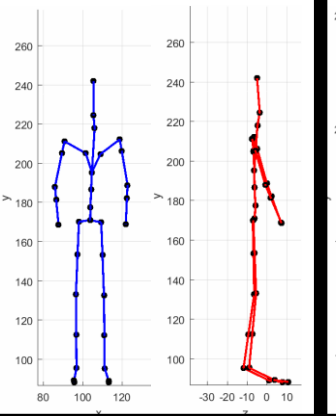
Source: http://tele-immersion.citris-uc.org/berkeley_mhad

82 minutes of 660 action sequences

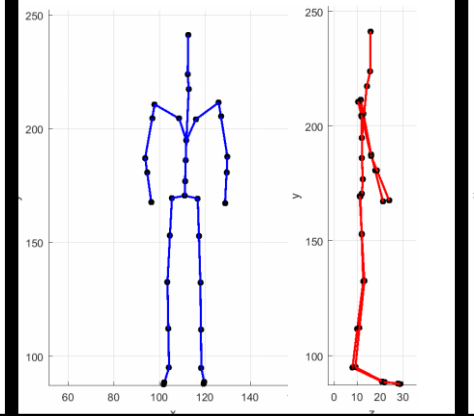
Action	# of repetitions/recording	# recordings	≈ length
Jump	5	5	5 sec
Jumping jack	5	5	7 sec
Bending	5	5	12 sec
Punching	5	5	10 sec
Wave 2 hands	5	5	7 sec
Wave 1 hand	5	5	7 sec
Clapping	5	5	5 sec
Throwing	1	5	3 sec
Sit then stand	5	5	15 sec
Sit-down	1	5	2 sec
Stand-up	1	5	2 sec

Actions in MHAD dataset

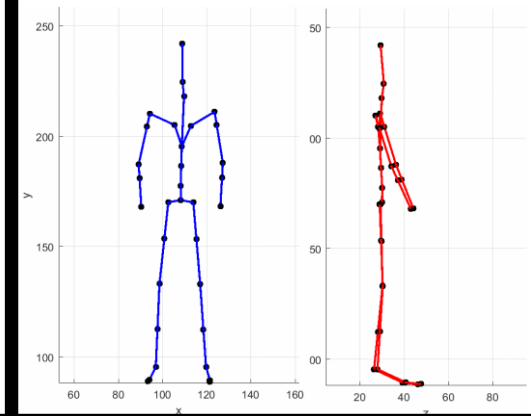
Jump



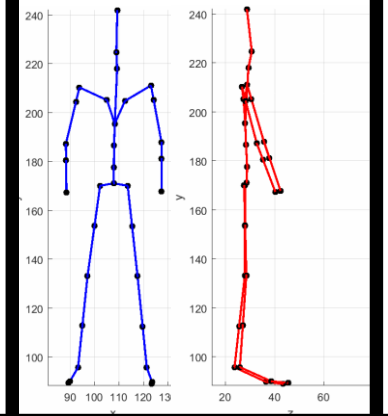
Jumping Jack



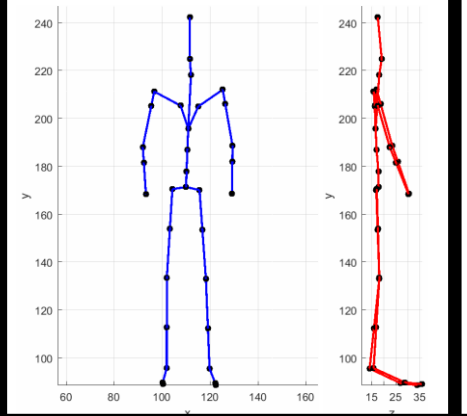
Bending



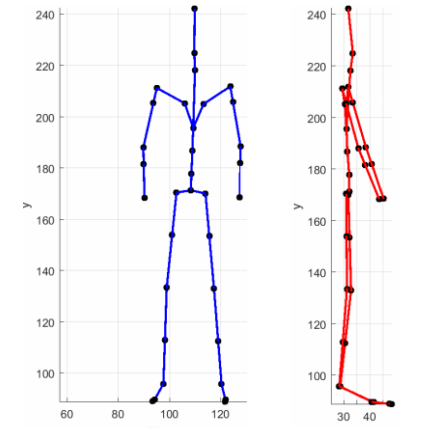
Punching



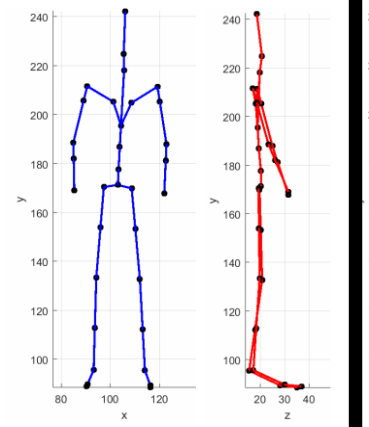
Wave 2 hands



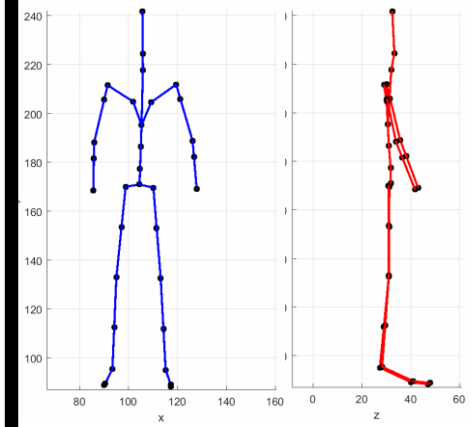
Wave 1 hand



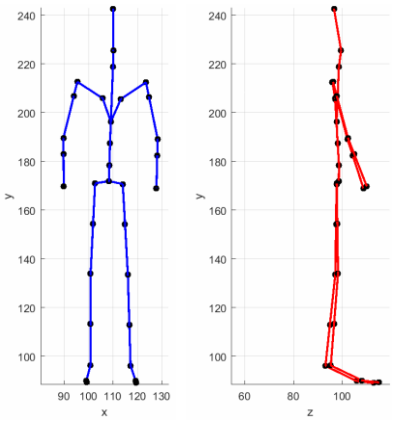
Clapping



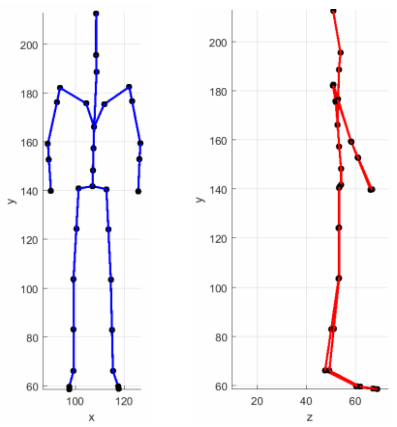
Throwing



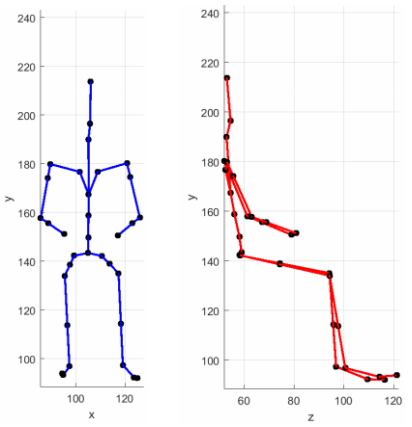
Sit & Stand



Sit-down



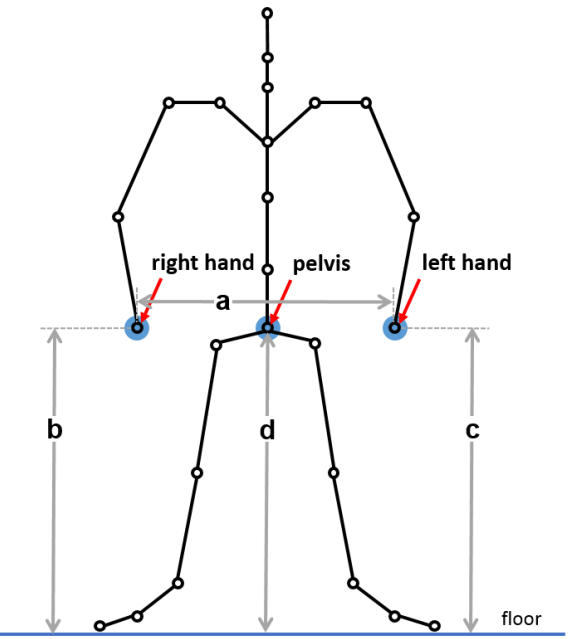
Stand-up

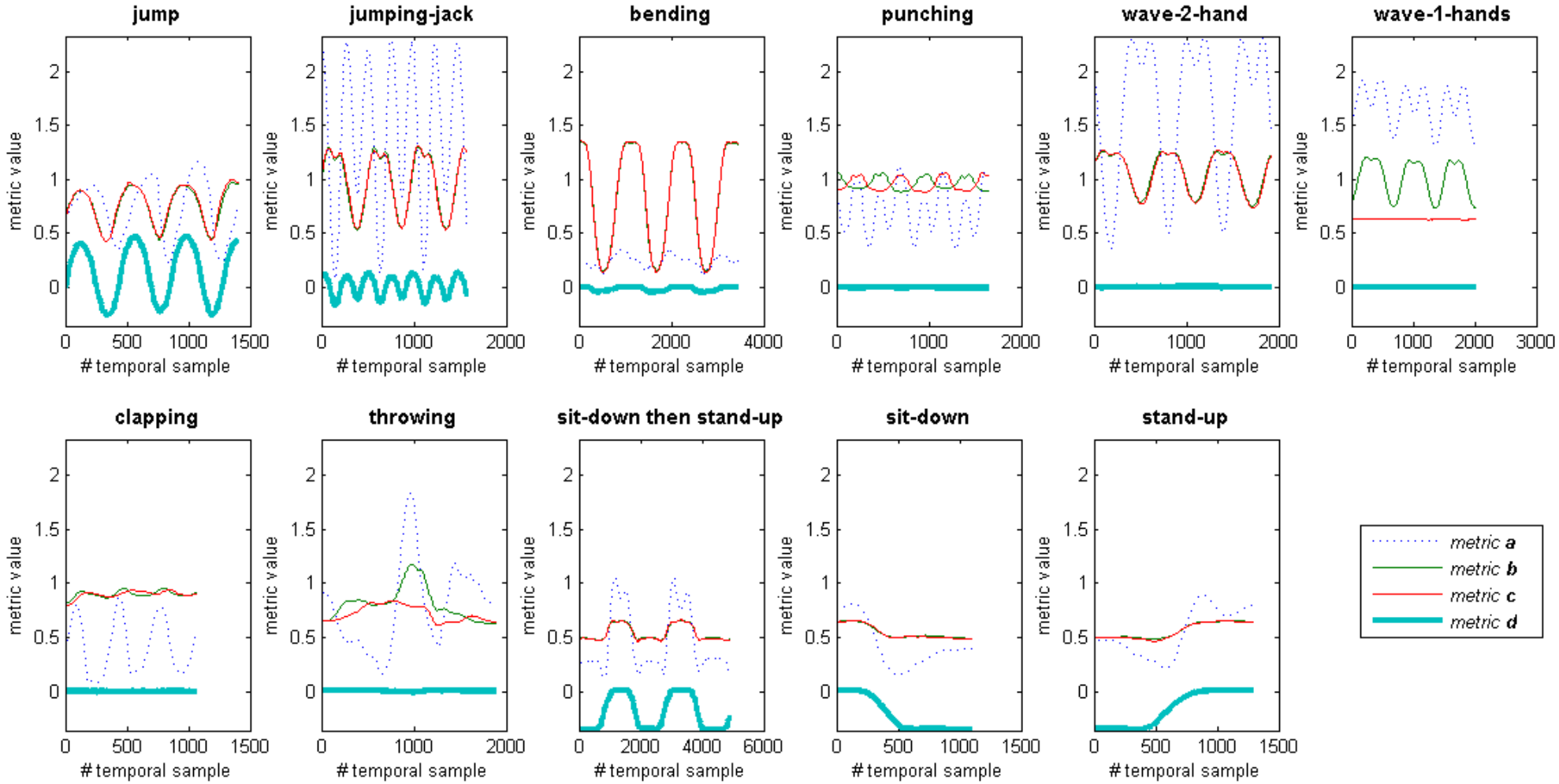


Front view

Side view

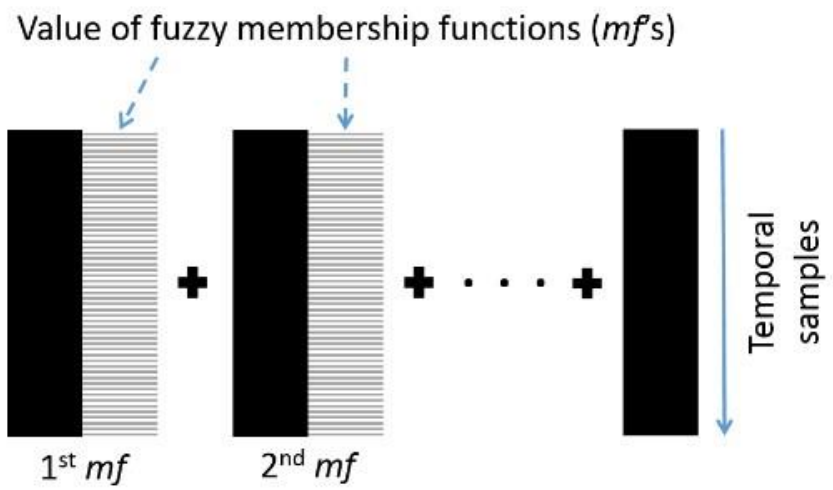
Variation of normalized distances



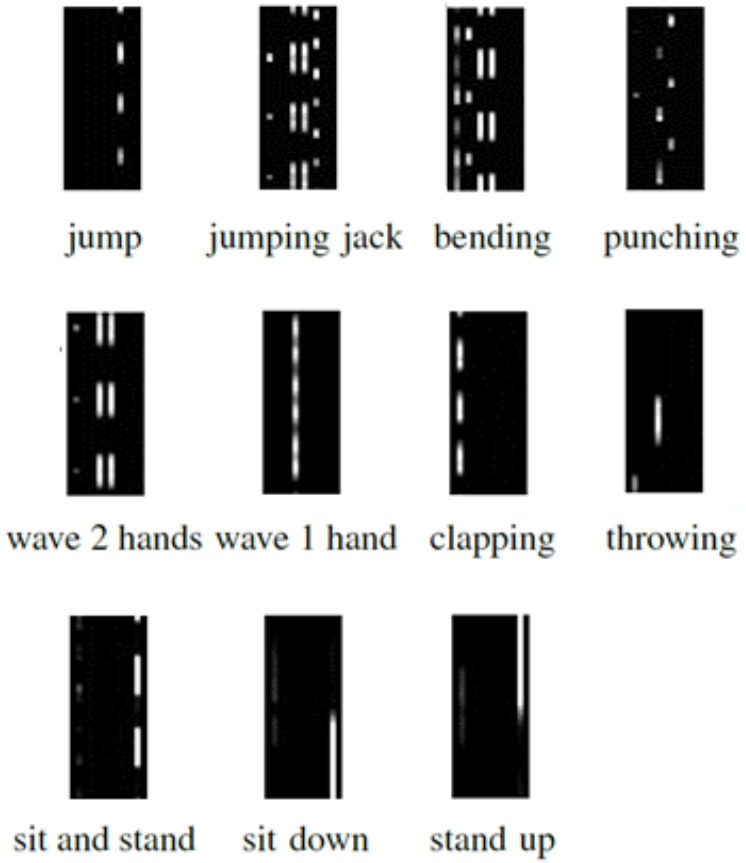


Observation: The nature and range of variation of these distances is unique for each action.

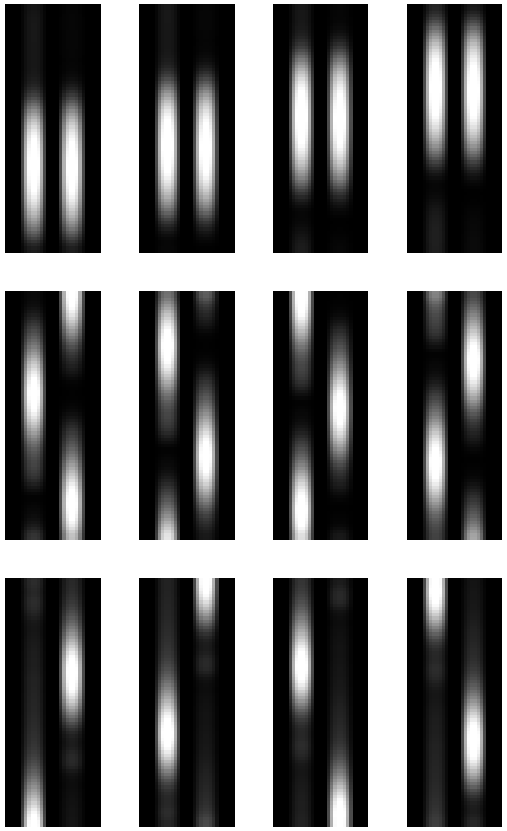
Representation of MOCAP actions



MHAD actions



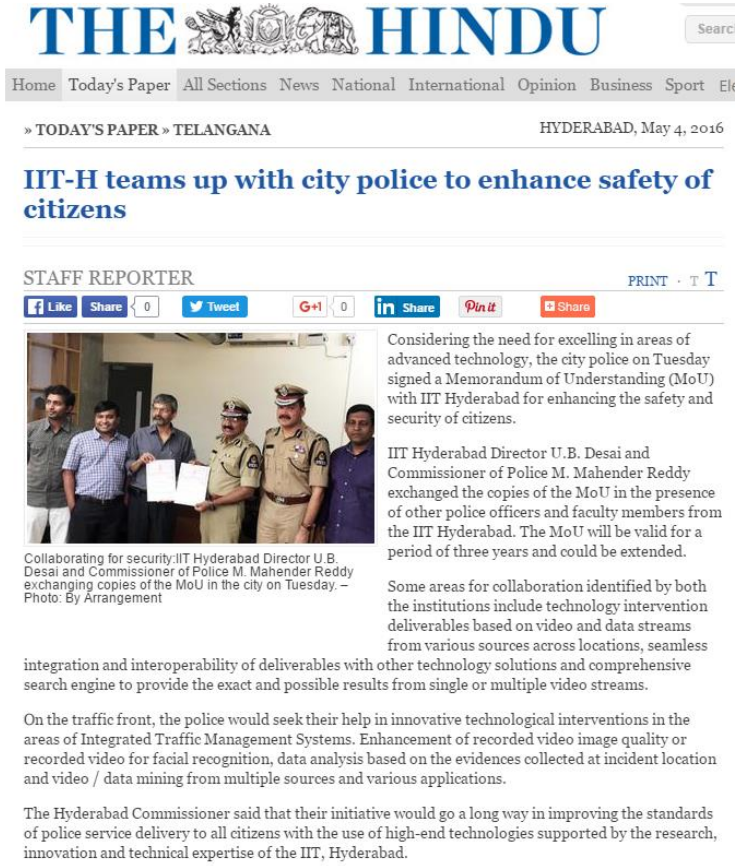
CMU actions



Observation: Distinct input representation for each action within the dataset.

Challenges in Traffic Surveillance

- Identifying common people across different religious processions
- Backtracking the path of offending vehicles to identify origin of vehicle (Use cases : Drunk driving, chain snatching)
- Monitoring sanitation state of areas for effective dispersal of municipal workers
- Monitoring for accidents for quick dispersal of emergency services
- We have signed an MOU with the Telangana State Police Department to obtain surveillance videos from traffic cameras.



THE HINDU

Home | Today's Paper | All Sections | News | National | International | Opinion | Business | Sport

» TODAY'S PAPER » TELANGANA

HYDERABAD, May 4, 2016

IIT-H teams up with city police to enhance safety of citizens

STAFF REPORTER

PRINT · T T

Like Share 0 Tweet G+1 0 Share Pin it Share

Collaborating for security:IIT Hyderabad Director U.B. Desai and Commissioner of Police M. Mahender Reddy exchanging copies of the MoU in the city on Tuesday. – Photo: By Arrangement

Considering the need for excelling in areas of advanced technology, the city police on Tuesday signed a Memorandum of Understanding (MoU) with IIT Hyderabad for enhancing the safety and security of citizens.

IIT Hyderabad Director U.B. Desai and Commissioner of Police M. Mahender Reddy exchanged the copies of the MoU in the presence of other police officers and faculty members from the IIT Hyderabad. The MoU will be valid for a period of three years and could be extended.

Some areas for collaboration identified by both the institutions include technology intervention deliverables based on video and data streams from various sources across locations, seamless integration and interoperability of deliverables with other technology solutions and comprehensive search engine to provide the exact and possible results from single or multiple video streams.

On the traffic front, the police would seek their help in innovative technological interventions in the areas of Integrated Traffic Management Systems. Enhancement of recorded video image quality or recorded video for facial recognition, data analysis based on the evidences collected at incident location and video / data mining from multiple sources and various applications.

The Hyderabad Commissioner said that their initiative would go a long way in improving the standards of police service delivery to all citizens with the use of high-end technologies supported by the research, innovation and technical expertise of the IIT, Hyderabad.

Biker/Non-biker Detection

Kunal Dahiya, Dinesh Singh, C. Krishna Mohan, "**Automatic Detection of Bike-riders without Helmet using Surveillance Videos in Real-time**", International Joint Conference on Neural Networks (IJCNN 2016), **Vancouver, Canada**

Helmetless Bikers Detection in Sparsely Crowded Environment

Kunal Dahiya, Dinesh Singh, C. Krishna Mohan, "**Automatic Detection of Bike-riders without Helmet using Surveillance Videos in Real-time**", International Joint Conference on Neural Networks (IJCNN 2016), **Vancouver, Canada**

Accident Detection in Surveillance Videos

Challenges

- No exact definition of accident
- No clear distinction between occlusion and collision
- Not many examples of accidents available for modeling
- Occlusion, lighting conditions, camera angles etc.

Accident Detection in Surveillance Videos

Snatch Theft Detection in Surveillance Videos

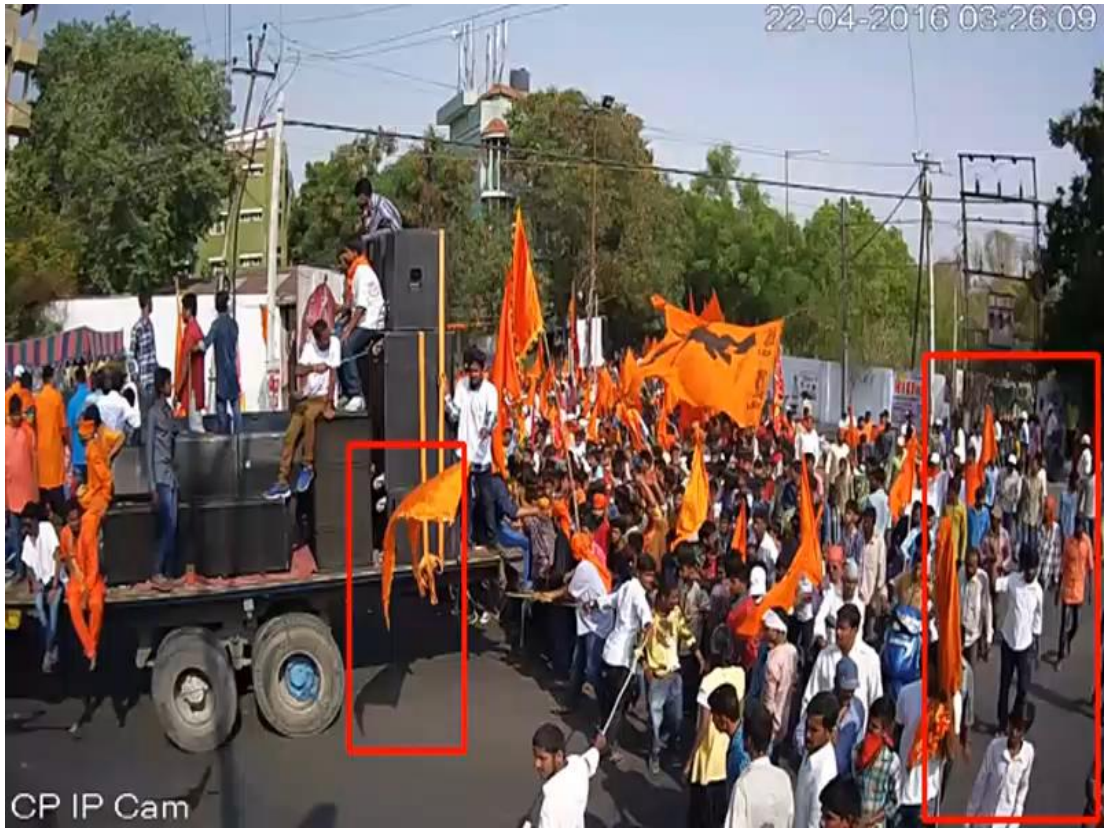
Challenges

- No exact definition of snatch theft
- Thefts may seem deceptively similar to regular interactions
- Not many examples of thefts available for modeling
- Human tracking issues: occlusion, lighting conditions, camera angles etc.
- Sequential modeling of thefts is difficult due to heterogeneity (all snatch thefts do not look same)

Snatch Theft Detection

Person re-identification in processions

- Identification of persons (probable miscreants) across different CCTV cameras deployed by the authorities
- Matching with police database for known criminals
- Tracking the activity of said persons throughout the city
- ***Issues:*** Occlusion, Camera motion, Illumination, Camera angle, Limited field of view, face may not be completely visible or may have very low resolution



Spontaneous Facial Expression Recognition In Unconstrained Videos

Spontaneous Facial Expression Recognition

To efficiently recognize spontaneous (not posed) facial expressions in wild.

Issues

Illumination

Occlusion

Issues (cont)

Pose

Group emotion recognition

Issues (cont)

- Less availability of databases, specially in spontaneous and wild environment.

Acted facial expression in wild

MIT facial expression database (AM-FED)

Action Unit 1-
Inner Brow Raiser

Action Unit 2-
Outer Brow Raiser

Action Unit 4-
Brow Lowerer

Action Unit 5-
Upper Lid Raiser

Action Unit 6-
Cheek Raiser

Action Unit 7-
Lid Tightener

Figure: Facial action unit and coding system approach for spontaneous expression recognition in wild.