HUMAN POSE RECOVERY AND ACTION RECOGNITION

Khandesh Bhange (11196) & Piyush Kumar(11496)
Advisor – Prof. Amitabha Mukerjee
Dept .of Computer Science and Engineering, IIT Kanpur

PROBLEM STATEMENT

We have implemented a way through which, given a sequence of frames of RGB image we should be able to display its skeleton for every frame and his/her action performed during this video.

Fig1. Walk

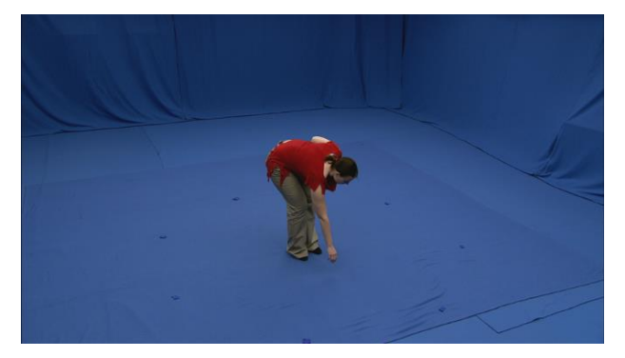


Fig2. Bend

Fig3. Wave

Fig4: Jump

Ref : Above images taken from i3D dataset.

So what's the difficulty in doing this ?

Large degree of freedom of human body, foreshortening of limbs due to change in body shapes, varying number of peoples, unavailability of depth data, low quality images, background cluttering and many more.

So How to Predict which action he is doing ???

Related Work Done

Articulated pose Estimation With Flexible Mixtures of parts (By - Yi Yang, Deva Ramanan) :

This paper describes a method for pose estimation in stationary images based on part models. In this method they have used a spring model as a human model and calculated a contextual correlation between the model parts.

One way to visualize the model is a configuration of body parts interconnected by springs. The spring like connections allow for the variations in relative positions of parts with respect to each other. The amount of deformation in the springs acts as penalty (Cost of deformation).

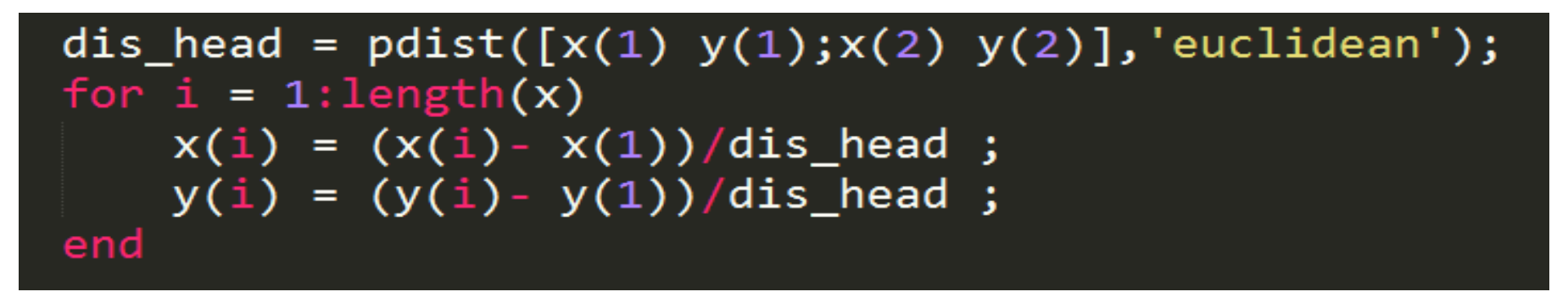
An Approach to Pose based Action Recognition (Chunyu Wang, Yizhou Wang and Alan L. Yuille):

- For representing human actions, it first group the estimated joints into five body parts namely Head, L/R Arm, L/R Leg.
- A dictionary of possible pose templates for each body parts is formed by clustering the poses of training data.
- For every Action class we distinguish some part sets (Temporal and Spatial) for representing the given action and then find the maximum intersection out of it.

OUR APPROACH

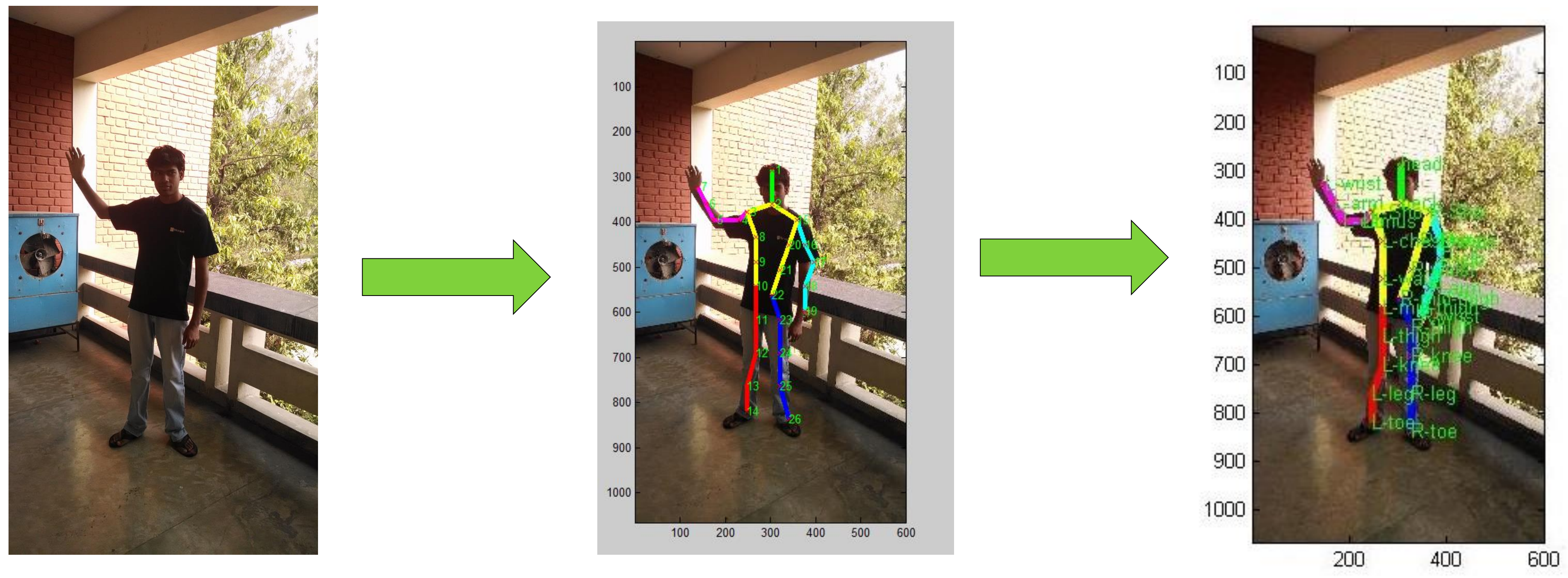
Part 1: Limb Labelling

We have clustered the human skeleton into 11 parts namely : Left Hand/Arm/Torso/Thigh/Leg, Right Hand/Arm/Torso/Thigh/Leg and Head. For this we have modified the Deva Raman's code. For calculating these clusters we have normalised the skeleton w.r.t Head-Neck length using following equation :



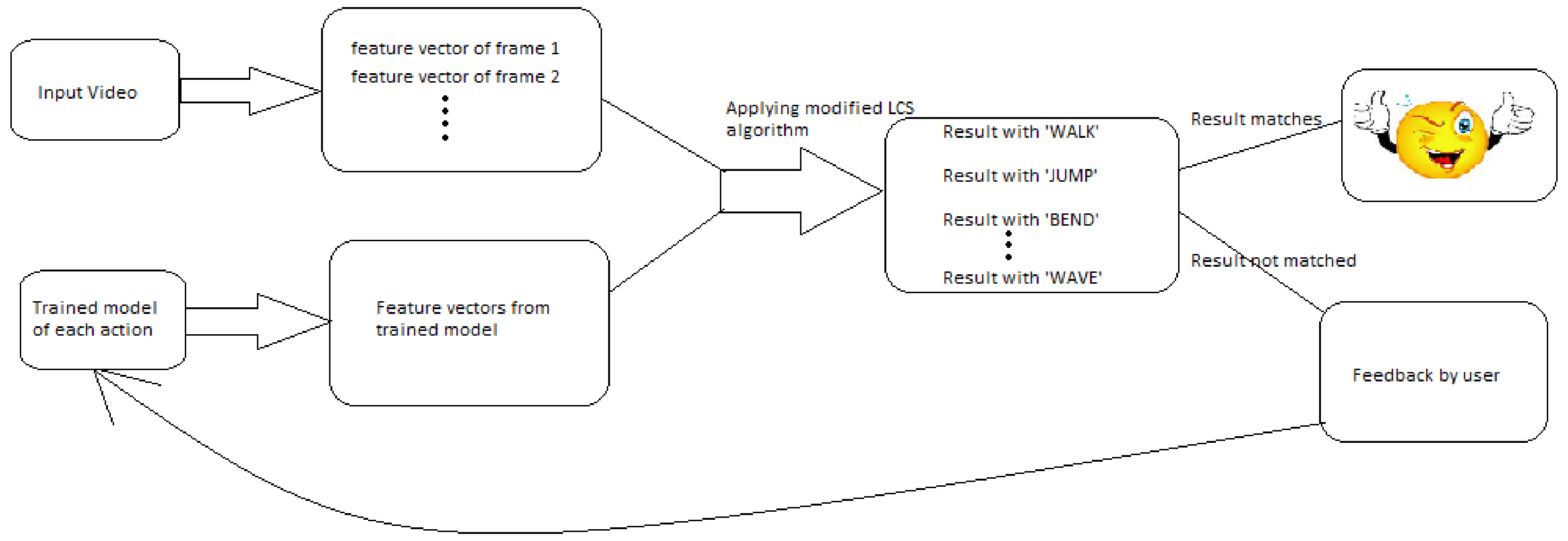
```
dis_head = pdist([x(1) y(1);x(2) y(2)],'euclidean');
for i = 1:length(x)
    x(i) = (x(i)- x(1))/dis_head ;
    y(i) = (y(i)- y(1))/dis_head ;
end
```

Now using Linear Regression we have estimated all the 11 body parts for a single frame of video.



Part 2 : Action Recognition using labelled frames from above method

- **Extracting Feature Vectors from Frames** : We calculate 8 angles defined as follows from above figure – Angles between : Left Hand and Arm, Left Arm and Torso, Left Torso and Left Thigh, Left Thigh and Left leg and similarly for Right body parts. Each angle can vary from 0 to 360 degrees.
- **Training the model :** For training the model we have used the dataset '**i3DPost Multi-view Human Action Datasets**' and used supervised learning technique for learning**.** We used idea similar to codebook generation to train our model. For every frame of each video we calculated its feature vector and stored it accordingly for each action.
- **Testing for a New Frame :**


Input Video
feature vector of frame 1
feature vector of frame 2
Trained model of each action
Feature vectors from trained model
Applying modified LCS algorithm
Result with 'WALK'
Result with 'JUMP'
Result with 'BEND'
Result with 'WAVE'
Result matches
Result not matched
Feedback by user

Modified Longest Common Subsequence (LCS) Algorithm : For sequence of frames in test data we calculate the count of maximum subsequence which matches with training model for each action. While comparing two frames, count of LCS is only increased when the '**Star distance**' is within a particular range and penalty calculated is less than a threshold. We calculated this range and threshold experimentally. We have implemented it using Dynamic-Programming method.

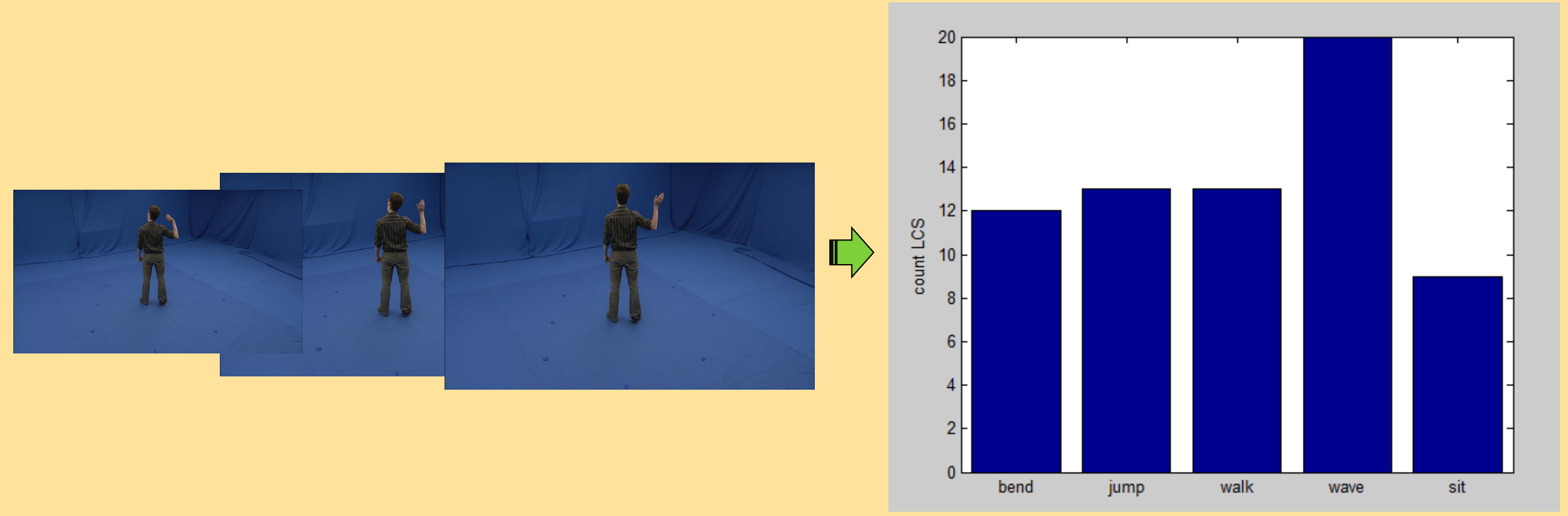
DATASET INFO

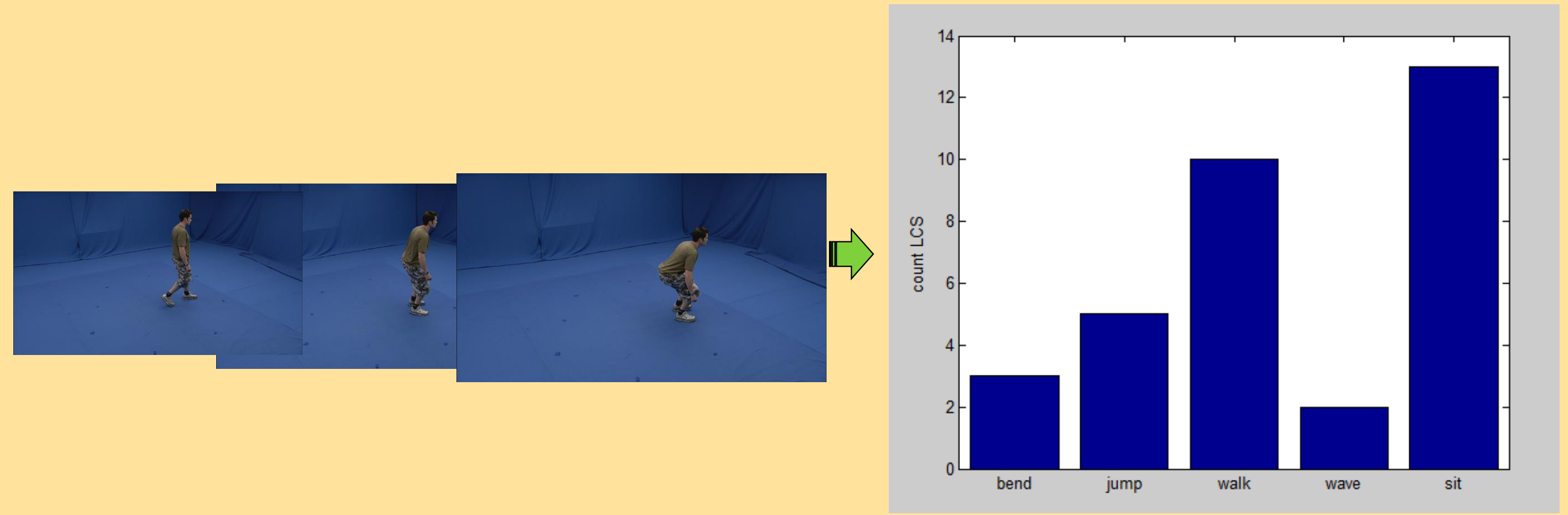
We have used '**i3DPost Multi-view Human Action Datasets**' for training and testing purposes. This provides a dataset of HD image sequences of 8 person (different camera view) doing 13 actions namely Walk, Run, Jump, Bend, Hand-wave, Jump-in-place, Sit-Stand Up, Run-fall, Walk-sit, Run-jump-walk, Handshake, Pull, Facial-expressions.

In Human pose recovery, they have used Image Buffy dataset and Parse dataset. This Parse dataset contains 305 pose-annotated images full body images of human poses.

RESULTS

We have tested our algorithm for 5 actions namely – Jump, Walk, Walk-Sit, Wave, Bend and following are the snapshots of some of the results:





CONCLUSIONS

Future work:

1. Can be trained on larger data set for each action to get better results
2. The Deva Ramanan's parse model can be trained on '**i3DPost**' dataset to improve accuracy.
3. Can give better result if RGB-D dataset is used.
4. To incorporate multiple camera views for action detection.

REFERENCES

1. "Articulated pose estimation with flexible mixtures-of-parts"
Y Yang, D Ramanan - Computer Vision and Pattern Recognition (CVPR), 2011
2. "An approach to pose –based action recognition"
Chunyu Wang, Yizhou Wang, and Alan L. Yuille (CVPR),2013
3. i3DPost Multi-view Human Action Datasets :
http://kahlan.eps.surrey.ac.uk/i3dpost_action/
4. Code provided by D. Ramanan :
http://www.ics.uci.edu/~dramanan/software/pose/