Human Pose Recovery And Gesture Recognition

CS365 : Artificial Intelligence

Khandesh Bhange(11196) And Piyush Kumar(11496)

Problem Statement

- There are two parts of this project.

  1st – Human Pose Recovery – We have to label 14 limbs (if visible) in a human body.

  2nd –Gesture Classification - The focus of this part is on gesture recognition on RGB data set, corresponding to 20 gesture categories like Wave, Point, Clap and some traffic gestures such as Turn Left, Turn Right.

Limbs Detection and Labelling

Fig1 : Chalearn Challenge
(http://gesture.chalearn.org/_/rsrc/1390380203402/mmdata/Track1.png)

Related Works

- Articulated pose estimation with flexible mixtures-of-parts

  By - Yi Yang, Deva Ramanan.

This paper describes a model for capturing contextual co-occurrence relations between parts, augmenting standard spring models that encode spatial relation. (Text taken directly from Ref -1)

Model Used

- The model represents the body as a deformable configuration of individual parts which are in turn are modelled separately in a recursive manner.
- 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(deformation cost).

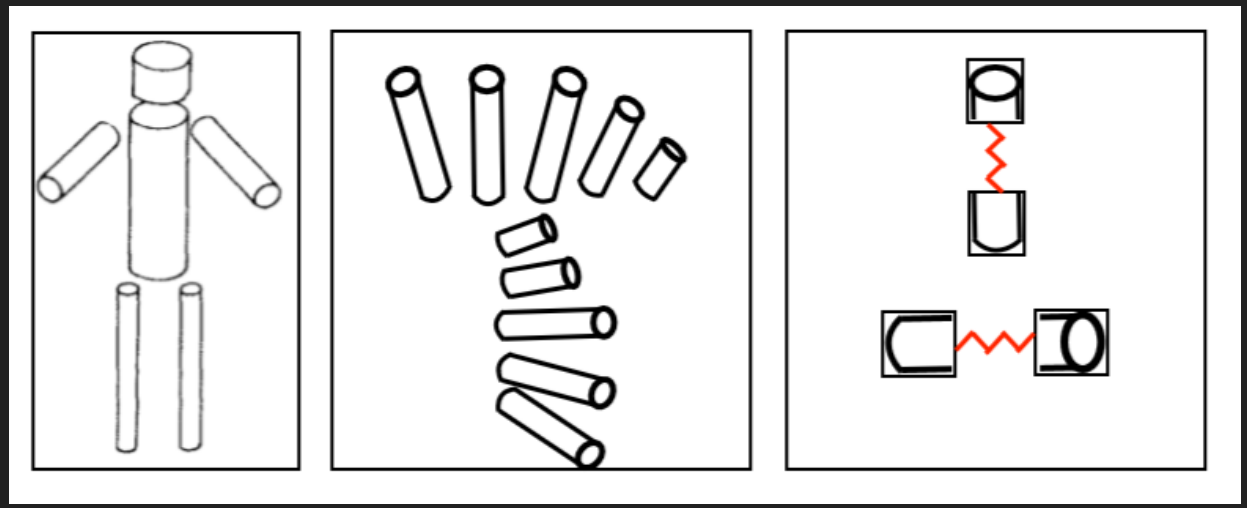

Fig2 : Spring model (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5995741

Approach Used

- For this part we are primarily going to follow the work done by Dev-Ramanan.
- Feature set – For every pixel in every image it describes a feature vector (HOG descriptor in this case) which was implemented in the paper "Histograms of Oriented Gradients for Human Detection" by Navneet Dalal and Bill Triggs.
- It defines a Co-Occurrence Model : To find score of a configuration of parts, it first defines a compatibility function for part types that factors into a sum of local and pairwise scores.
- It selects K such configurations in decreasing order of their scores.

Approach Used (contd..)

$$\text{score}_i(t_i, p_i) = b_i^{t_i} + w_{t_i}^i \cdot \phi(I, p_i) + \sum_{k \in \text{kids}(i)} m_k(t_i, p_i) \quad (6)$$

$$m_i(t_j, p_j) = \max_{t_i} b_{ij}^{t_i, t_j} + \max_{p_i} \text{score}(t_i, p_i) + w_{ij}^{t_i, t_j} \cdot \psi(p_i - p_j) \quad (7)$$

Fig 3: Mathematical Explanation of the approach (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5995741)

Result of Part 1

- It will give us K number of poses per image.
- Each pose will be modelled as a Graph with 14 nodes and each node will represent a Human body Limb(part).
- All parts will be represented as 5 × 5 HOG cells in size.

Part 2 : Gesture Classification

- Given a video, we first obtain best-K pose estimations for each frame using the above method for pose recovery.

Approach

- Determining the best pose for each frame of the video using energy function maximization.
- To represent human actions, we 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.
- Now for every test video, we count the presence of part-sets in the selected poses and form a histogram and action class with maximum intersection is selected.

Spatial And Temporal Part-sets

- Spatial-part-sets These are *part-sets* which occur frequently in one action class but rarely in other classes. This helps in discrimination as well as representation.
- Temporal-part-sets It is basically a pose sequence which occur frequently for an action class.

Step by Step Methodology

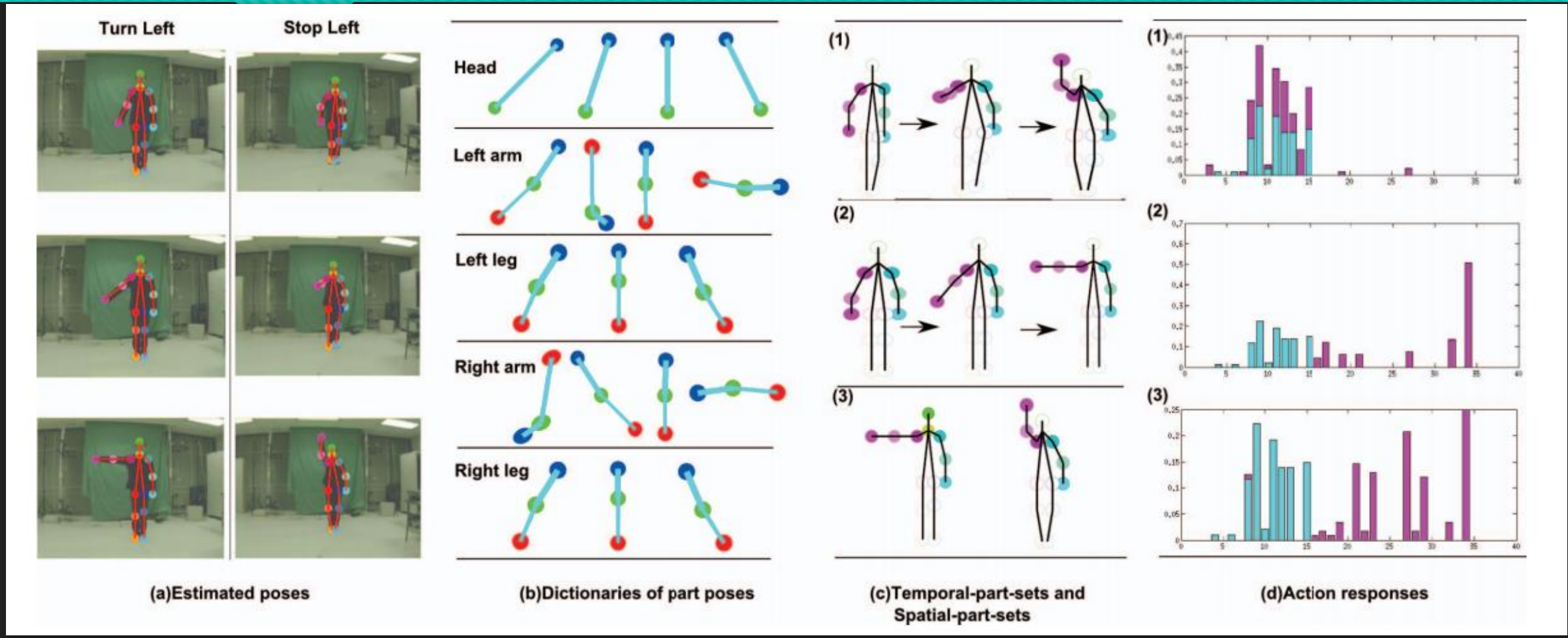

Fig 3 : http://www.cv-foundation.org/openaccess/content_cvpr_2013/papers/Wang_An_Approach_to_2013_CVPR_paper.pdf)

Estimating Best pose for a frame (optional)

- Model select best pose (P) for each frame by maximizing the Energy function

$$E_P = \sum_{i=1}^{L} \phi(P_{j_i}^i, I^i) + \sum_{i=1}^{L-1} \psi(P_{j_i}^i, P_{j_{i+1}}^{i+1}, I^i, I^{i+1})$$

Fig4 : http://www.cv-foundation.org/openaccess/content_cvpr_2013/papers/Wang_An_Approach_to_2013_CVPR_paper.pdf

- Here the first term measures the likelihood of the pose and second is a pairwise term that measures the appearance, and location consistency of the joints in consecutive frames.

Available Dataset

- Dataset for the first part : A collection of RGB image and 14 binary masks(per image) corresponding to the 14 labelled limbs (if visible): Head, Torso, R-L Upper-arm, R-L Lower-arm, R-L Hand, R-L Upper-leg, R-L Lower-leg, and R-L Foot is provided.
- Dataset for the 2nd Part : A collection of videos containing 50-70 frames per action is available (Keck Gesture dataset).
- In the above dataset number of frames per gesture/action will be given in a CSV file per video frame.

References

- Main References :

  Y. Yang, D. Ramanan. **"Articulated Pose Estimation using Flexible Mixtures of Parts"** *Computer Vision and Pattern Recognition* (CVPR) Colorado Springs, Colorado,June 2011.

  An approach to pose-based action recognition – "Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on " By – C. Wang, Y. Wang, Alan L.Yuille .

- Other Reference :

  Challenge Url : http://gesture.chalearn.org/mmdata

  Code provided by D. Ramanan : http://www.ics.uci.edu/~dramanan/software/pose/

  Histograms of Oriented Gradients for Human Detection "Published in International Conference on Computer Vision & Pattern Recognition " on Dec 2010: By – Dalal and Triggs