Human Activity Recognition

B.Tech Project Report

Satwik Kottur*
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India - 400076
satwik.kottur@iitb.ac.in

with
Neha Bhargava †
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India - 400076
neha@ee.iitb.ac.in

Guide
Prof. Subhasis Chadhuri ‡
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India - 400076
sc@ee.iitb.ac.in

Abstract

Human activity recognition is gaining importance, not only in the view of security and surveillance but also due to psychological interests in understanding the behavioral patterns of humans. This report is a study on various existing techniques that have been brought together to form a working pipeline to study human activity in social gatherings. Humans are first detected with Deformable part models and tracked as a feature point in 2.5D co-ordinate system using Lucas-Kanade algorithm. Linear cyclic pursuit model is then employed to predict short-term trajectory and understand behavior.

1 Introduction

With the increasing rise in the need for security and surveillance, particularly in crowded areas like airports, shopping malls and social gatherings, the problem of human detection and activity recognition has attained importance in the vision community. This is both

*http://www.ee.iitb.ac.in/student/~satwik
†http://www.ee.iitb.ac.in/student/~neha
‡http://www.ee.iitb.ac.in/~sc

an interesting and challenging problem due to two reasons. Firstly, the behavior can be governed by set of underlying rules which when exploited give insights into the crowd motion. For instance, motion analysis of players in a game of basketball strictly follows the rules of the game. Secondly, different levels of analysis give different inferences. Motion of people in a busy street, as seen as individual agents, might look completely uncorrelated and random. However, a crowd level analysis might reveal interesting patterns like crossing the road in opposite directions, halting at the traffic signs, etc.

This problem of human activity has been studied primarily using two approaches - holistic and reductionist. In the former approach, crowd is modeled as a single entity and behavior is analyzed. For example, [3] uses optical flow on the crowd while [2] uses the concepts of fluid flow and fluid dynamics. The latter approach considers sparse individual motions and tracks humans applying suitable filters to evaluate parameters like position and velocity ([1],[4],[5]). Srikrishnan et al. [14] use the linear cyclic pursuit (LCP)([13]) framework for modeling crowd motion. Short term trajectories of individuals agents are predicted and used for further inference. A comprehensive survey of various crowd analysis methods is present in [17] by Zhan et al.

This report is organized as follows. Section 2 explains the pipeline of the analysis framework. Section 3 discusses, in detail, various existing works that form the above pipeline. Finally, section 4 shows the results obtained and concluding remarks are listed in section 5.

2 Approach Pipeline

A single, calibrated camera is used to obtain the video from a social gathering, which is the input to the system. Starting from this video, we detect humans, track them using feature point and predict the short term trajectory. The following outlines the approach studied in this report (figure 1):

1. *Detection:*
   Human detection forms the first stage of this approach. The locations found in the image plane are further used for tracking. Since detection involves extensive searching over the image, it is computationally the most expensive process. Section 3.1 discusses the algorithm and other relevant issues.

2. *Tracking:*
   To satisfy the real-time execution of the system, detection cannot be performed on all the incoming frames due to its high computation time. Hence, detection is performed on intermittent frames while tracking is employed to evaluate the

location of humans in the other frames. The Lucas-Kanade tracking algorithm based on optical flow, is used for this purpose. Further details are contained in [11].

3. *2.5D Co-ordinates:*
   The above detection and tracking methods give the location in the image plane. To make meaningful inferences, it is necessary to convert the 2D point to the corresponding 3D point in the world co-ordinate system. Section 3.2 elaborates this idea.

4. *Prediction:*
   Once the 3D locations of all the humans (or agents) are found, Linear cycle pursuit model is used to predict the trajectory and make short-term inferences as explained in section 3.3.

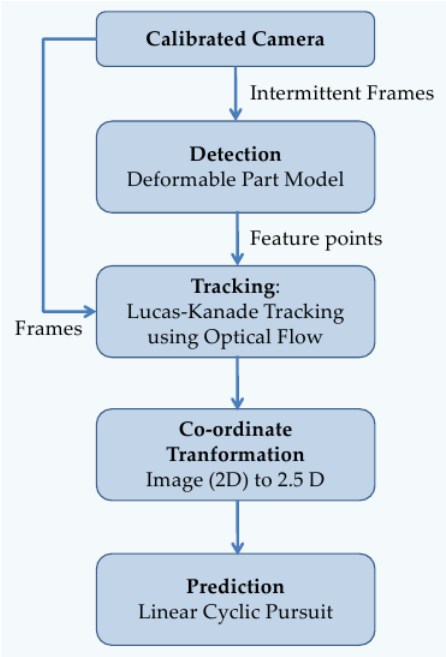

Figure 1: Pipeline of the algorithm

3 Related Work

Following the order mentioned in section 2, various works are discussed in detail that form the skeleton of the analysis.

3.1 Detection

Human detection is an interesting and well sought-after problem in itself. The difficulty is elevated due to the variability in human poses, sizes and self-occlusions. Felzenszwalb et. al proposed a deformable part model in [8] to detect objects that stands as the current

state-of-the-art. Though variations of such modeling have been proposed to slightly improve the accuracy, this report studies the original one in [8]. The training data ([10]) allows learning models for twenty different object classes out of which *person* class is used for the problem at hand.

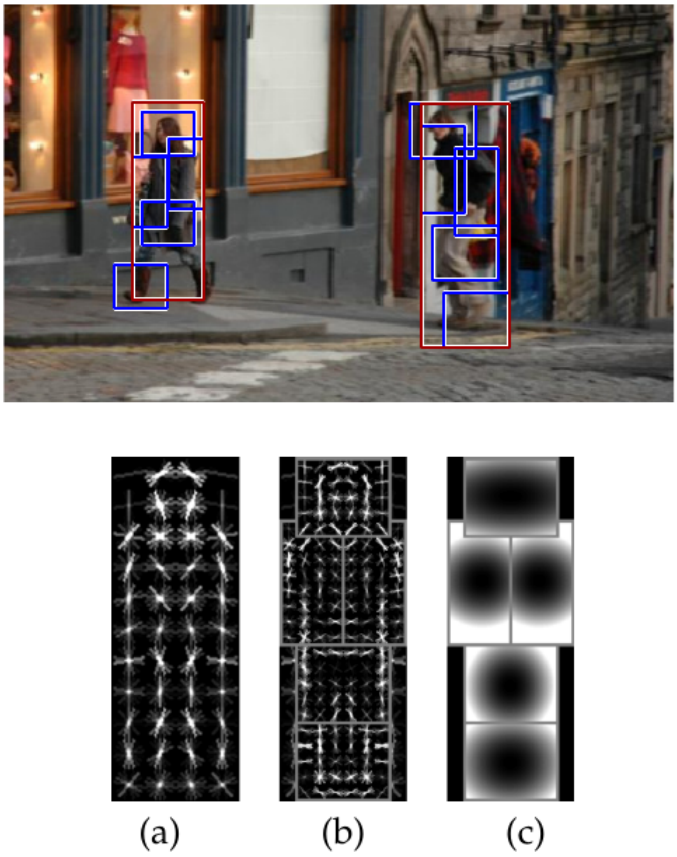

Figure 2: Detection using Deformable part models. (a) Root Filter (b) Part Filters for various parts of the human body (c) Gaussian models for part locations with respect to the root center location *Image courtesy:[8]*

The fundamental idea of part modeling is the consideration of articulated nature of human body. This condition imposes an additional constraint on the location of the detected parts, or constituent elements, of a given object. For a given image, Histogram of Gradients (HOG) features are extracted as first proposed in [6]. For each part, Support Vector Machine (SVM) classifiers are trained using training data in a supervised . A SVM classifier for the overall structure or root is also trained. Additionally, standard *anchor* positions of parts with respect to the root position are also learnt as a 2D Gaussian distribution. Figure 2 shows the above mentioned models. Multiple models, trained at various scales (image sizes) account for the pose (standing, sitting, sidewards, etc. as in case of person) and size variations.

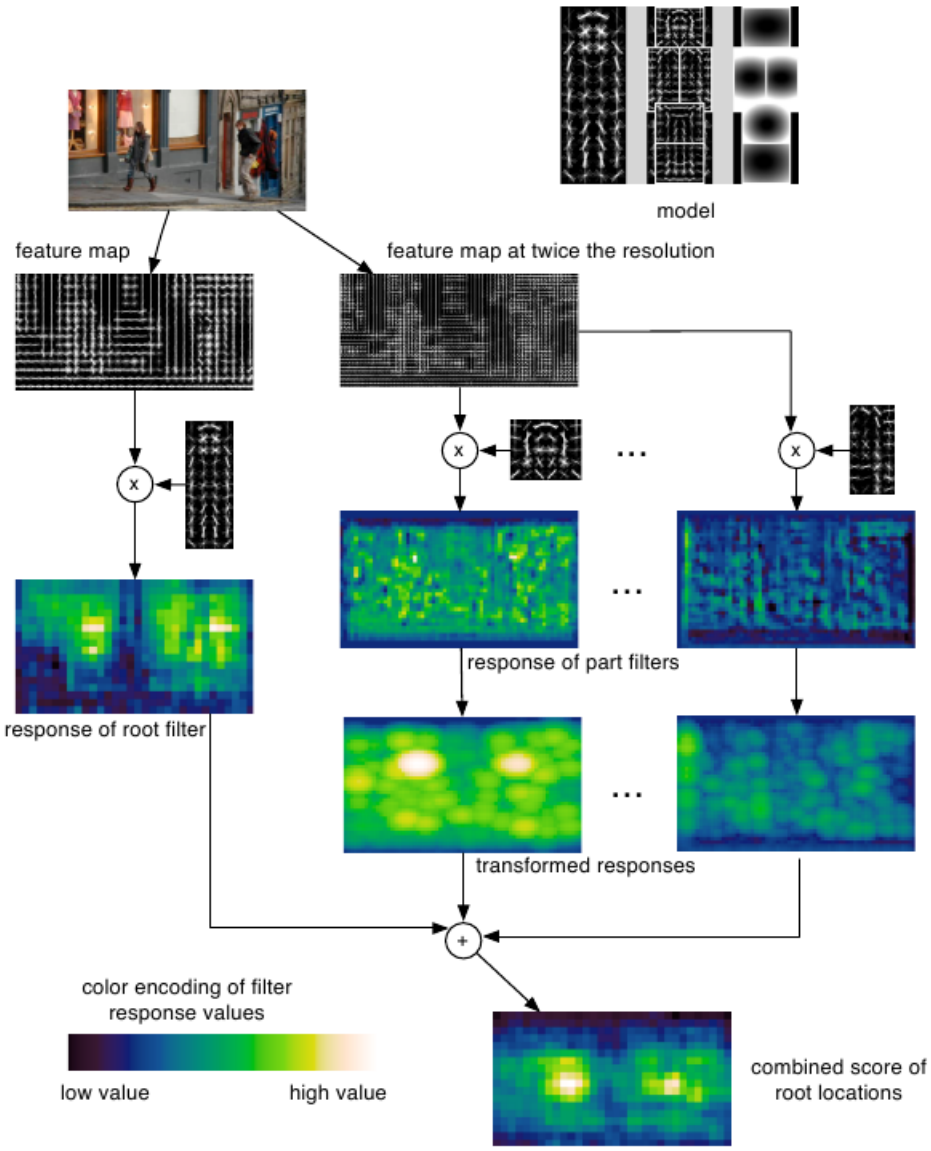

Figure 3: Overall view of detection procedure. (a) HOG features for given image is computed at various resolutions. (b) Part Models and Root Models are used to score the image (c) Displacement of parts from expected location is scored and added to overall score *Image courtesy:[8]*

The score from SVM classifiers for parts and the root is used as part detection scores and root detection score, respectively. Finally, a spring model, as shown in figure 2, is employed to score the displacement of a detected part from its anchor position and an overall detection score is computed based on displacement scores, root detection score and part detection scores. Detection is then performed using a threshold followed by Non-Maximal Suppression to eliminate false positives for optimal performance. The locations of root and parts are used as feature points for tracking. Finer mathematical details for the above modeling can be found in [8].

3.2 2.5D Co-ordinates

After solving the problem of finding the location of the humans (agents) in the image plane, it is necessary to estimate their positions in the 3D world co-ordinate system. Neha et al. have incorporated 2.5D co-ordinate system in [12] using an average height plane assumption which is detailed below.

Since the camera used is calibrated, the transformation matrix P mapping world co-ordinates to the image co-ordinates is known. The general problem of back-projecting the image co-ordinates to the world co-ordinates is ill-posed. However, [12] assumes the existence of a plane parallel to the ground plane at a height H. This plane denotes an average height for all the human agents. A nominal value of $160cm$ was used in [12]. Figure 4 shows the position of the calibrated camera along with the average height plane.

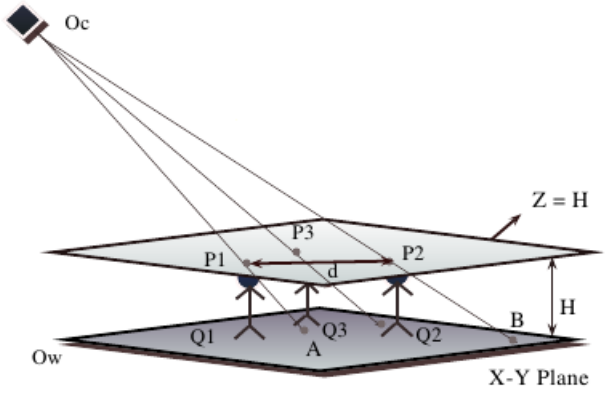

Figure 4: Camera setup for 2.5D co-ordinates using average height plane for all agents *Image courtesy:[12]*

Let O_w be the origin of the world co-ordinates and O_c be the position of calibrated camera. Consider an image point $\mathbf{x}$ corresponding to human (agent) $P1$. We can obtain the co-ordinates of A using P^+ as height of A is known (i.e. ground plane). Since $A(x_a, y_a, z_a)$, $O_c(x_c, y_c, z_c)$ and $P1(x_{p1}, y_{p1}, z_{p1})$ are collinear, the following is obtained:

$$\frac{x_{p1} - x_c}{x_a - x_c} = \frac{y_{p1} - y_c}{y_a - y_c} = \frac{z_{p1} - z_c}{z_a - z_c} \tag{1}$$

leading to the evaluation of 2.5D co-ordinates of $P1$ as:

$$x_{p1} = \frac{H - z_c}{z_a - z_c}(x_a - x_c) + x_c \tag{2}$$

$$y_{p1} = \frac{H - z_c}{z_a - z_c}(y_a - y_c) + y_c \tag{3}$$

$$z_{p1} = H \tag{4}$$

Notice that equation (4) is the direct manifestation of average height plane assumption.

3.3 Prediction

To predict the trajectory, a linear cyclic pursuit model([13]) is applied to the agents or more specifically, to a feature point corresponding to the location of a human. A brief insight given in [12] is replicated here.

For a group of N agents and their motion in $\mathbf{R}^2$, linear cyclic pursuit model works on the following principles:

1. Motion in both dimensions are independent of each other i.e. X direction motion does not have a bearing on Y direction motion. Hence, analysis can be carried out in both directions irrespective of each other.
2. Direction of motion for one agent is towards the weighted mean of positions of other agents.

For i^{th} agent, at time t_j:

$$\dot{x}_i(t_j) = k_i[\sum_{k=1}^{N-1} \eta_k x_{((i+k))_N}(t_j) - x_i(t_j)] \tag{5}$$

where k_i is the gain for the i^{th} agent, η_k is the weight assigned to k^{th} agent and $((a))_N$ denotes a mod N. From the second assumption, $\eta_k > 0$ and $\sum_{k=1}^{N} \eta_k = 1$. The parameters η_k and k_i determine the system's motion collectively and are hence called motion parameters. Equation (5) can be written as:

$$\begin{pmatrix} \dot{x_i}(t_1) \\ \dot{x_i}(t_2) \\ . \\ . \\ . \\ \dot{x_i}(t_M) \end{pmatrix} = \begin{pmatrix} x_1(t_1) & x_2(t_1) & . & . & . & x_N(t_1) \\ x_1(t_2) & x_2(t_2) & . & . & . & x_N(t_2) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_1(t_M) & x_2(t_M) & . & . & . & x_N(t_M) \end{pmatrix} \begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ . \\ . \\ . \\ \alpha_{Mi} \end{pmatrix} \tag{6}$$

Using M incoming frames ($M > N$) and knowing the initial positions (from detection and tracking), the first order differential equations in X and Y are solved to yield a set of N linear equations for each $\dot{x}$ (equation 6). It is important to realize that these parameters are valid only for short-time due as humans are intelligent agents. Using the incoming frames, motion parameters are estimated and used to predict trajectories. A thorough study of linear cyclic pursuit model with other pursuit models can be found in [13] along with discussions on stability and convergence.

4 Results

Video sequences from a calibrated, surveillance camera at VMCC, IIT Bombay are used to test the studied framework. For discriminative part models, a MATLAB implementation ([10]) is available. But the run times by using above source do not allow scaling down to a real-time application. Hence, an accelerated implementation proposed in [7] is used. This uses FFTW library ([9]) for faster convolution thereby reducing algorithm run-time. The former results in an execution time of around $6s$ while latter around $0.6s - 1s$ on a quad-core processor.

The results of this study and [12] for various stages of the pipeline, are shown in figures 5, 6 and 7.

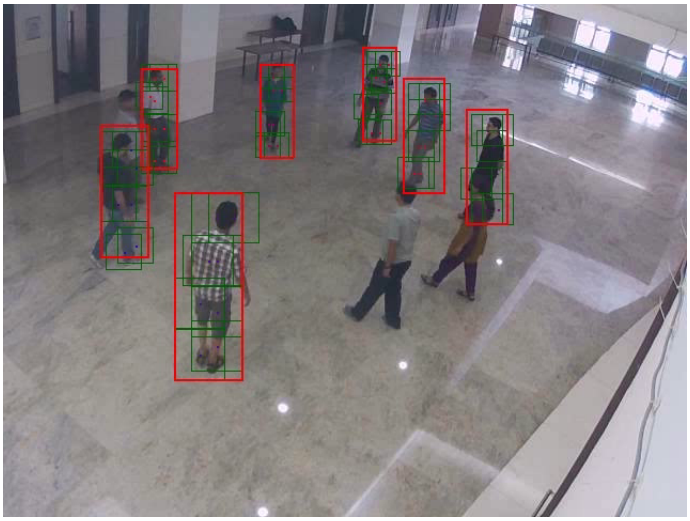

Figure 5: Human Detection using Deformable part models (Root shown in *red* box and parts in *green* boxes)

5 Conclusion

This report focuses on studying various existing techniques by getting them into a closely-knit pipeline, similar to [12]. However, the manual labeling required in [12] has been done way with in this study, using Deformable part model based detection. The problem of human activity recognition and analysis is no where close to completion. A lot of scope exists to model and understand the behavioral pattern of human agents along with predictions for security and surveillance purposes.

Further study is aimed towards recognizing activity like talking, eating, drinking, etc. by understanding the head and body pose of humans in the given video. Though a hard problem, this will be open up many prospects towards understanding and modeling human actions for various real-life applications.

Figure 6: Tracking of part locations using Lucas Kanade algorithm (Colors corresponds to an agent. Part locations are tracked are multiple feature points)

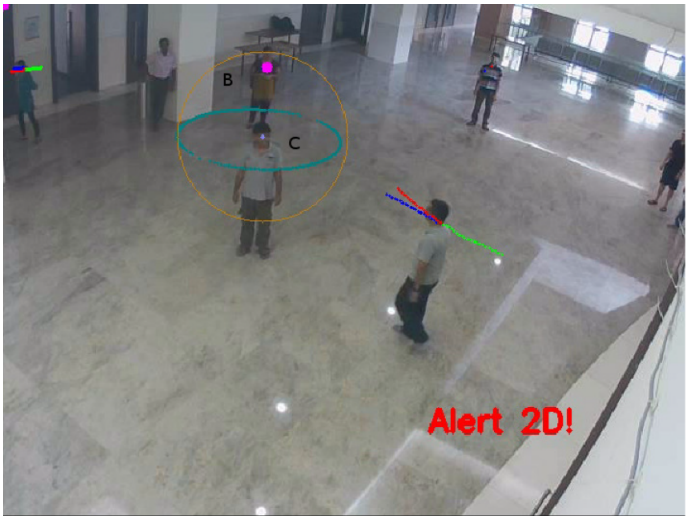

Figure 7: Trajectory Prediction using Linear Cyclic Pursuit model, necessary security action can be taken accordingly *Image courtesy:[12]*

Acknowledgments

I thank Prof. Subhasis Chaudhuri for guiding me towards the completion of my B.Tech Project Stage−1. I would also like to thank Neha Bhargava for her valuable inputs and suggestions.

References

[1] Ahmed Fouad Mohamed Soliman Ali and Kenji Terada. A general framework for multi-human tracking using kalman filter and fast mean shift algorithms. *Journal of Universal Computers Science*, 16(6):921–937, 2010.

[2] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[3] Ernesto L Andrade, Scott Blunsden, and Robert B Fisher. Modelling crowd scenes for event detection. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 175–178. IEEE, 2006.

[4] Robert Bodor, Bennett Jackson, and Nikolaos Papanikolopoulos. Vision-based human tracking and activity recognition. In *Proc. of the 11th Mediterranean Conf. on Control and Automation*, volume 1. Citeseer, 2003.

[5] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289. IEEE, 2009.

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[7] Charles Dubout and Franois Fleuret. Exact acceleration of linear object detectors. In *In ECCV, 2012. 7*.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[9] Matteo Frigo, Steven, and G. Johnson. The design and implementation of fftw3. In *Proceedings of the IEEE*, pages 216–231, 2005.

[10] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[11] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.

[12] Subhasis Chaudhuri Neha Bhargava and Guna Seetharaman. Linear cyclic pursuit based prediction of personal space violation in surveillance video. Technical report, Applied Imagery Pattern Recognition Workshop, 2013.

[13] Arpita Sinha. Multi-agent consensus using generalized cyclic pursuit strategies. 2009.

[14] Viswanthan Srikrishnan and Subhasis Chaudhuri. Crowd motion analysis using linear cyclic pursuit. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3340–3343. IEEE, 2010.

[15] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, International Journal of Computer Vision, 1991.

[16] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[17] Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, 2008.