Asian Journal of Information Technology 15 (2): 287-297, 2016
ISSN: 1682-3915

# Human Activity Detection from UTI Dataset

[1]C.G. Ravichandran and [2]P. Sivaprakash
[1]SCAD Institute of Technology, 641664 Palladam, Tirupur, Tamil Nadu, India
[2]Department of Electronics and Communication Engineering,
RVS College of Engineering and Technology, 624005 Dindigul, Tamil Nadu, India

**Abstract:** In today's digital age, law enforcement officials and even employers may find it easier than ever to take advantage of camera and wiretap surveillance. Surveillance cameras now line many public streets and workplace locations in an attempt to monitor activity and law enforcement agencies continue to use wiretapping to aid in investigations. Even with the advancement of technology, we have always resorted to manned surveillance techniques which will require impeccable human attention to the video feed received from the surveillance cameras. The orthodoxical way of surveillance system can be automated by spatially identifying the human body and the vital body parts namely head, torso, etc. from live video frames such as CCTV camera footage. With this spatial information from the video frames, we estimate the poses held with the temporal association between the successive and the previous video frames. The system works well with unconstrained backgrounds and without any premeditation of the clothing, brightness of the video frame. We also do not impose any constraints on the position of a person in the video frame. The only constraint that is imposed by our system is that people should be in a head-over-torso position with either near-frontal or near-rear viewpoint for greater accuracy of the estimation However, the system responded with a considerable accuracy for side poses as well, during testing. The poses gleaned are used for detecting punching activity.

**Key words:** Human pose estimation, UTI dataset, activity detection, upper body detection and tracking, India

## INTRODUCTION

A automating a surveillance mechanism often has a range of challenges posed by several factors. The video feed from the surveillance camera often lacks clarity. We can eliminate this issue by assuming that the feed is at least legible with definite edges around objects that is non-pixel related. The other problems include natural cluttered background, varying lighting conditions throughout the video, position of the people in the frame if the person is near or far/spatial positioning of the human being in a single frame. Video frames often include motion blur. Keeping all this under our consideration, in this paper we primarily achieve to detect and estimate the pose of human, provided a video feed. Our approach we adopted is, initially a generic upper body detector will detect the coordinate position of the human body (upper body) in a video frame. This module provides us a bounding box coordinates of the detected upper body. These localized rectangular coordinates help us eliminate majority of the background clutter and focus only on the human being under question. Taking bounding box coordinates as input, the next module approximates a 'stickman' model of the human body. This module is based on Image Parsing technique proposed by Ramanan *et al*. (2005) and Ferrari's articulated human pose estimator (Eichner *et al*., 2012) which will be elaborately discussed in the following sections. The 'stickmen' coordinates consists of 10 fundamental body parts which are vital in defining a specific body pose. These include-head (1), torso (1), upper-arm (2), lower-arm (2), upper-leg (2), lower-leg (2). Figure 1 shows working of human pose estimator module (clockwise from top).

These stickmen coordinates are supplied as input to the last module which takes advantage of the temporal association in a video multiple action that occur more or less at the same time which may or may not be related at all and estimates the stances held by the human body in a video. The final extrapolation is based on how each stickmen coordinates move relatively across the video frame. Precisely, based on how quick (speed) and how wide (angle) these 10 fundamental sticks move across frames the 2D pose can be estimated. The 2D poses are fundamental in surveillance because, a definitive pose sequence across the video length ultimately decides the person's attitude or action. Also, the 2D poses cover a wide spectrum of applications ranging from comprehending a video to automating manned

**Corresponding Author:** C.G. Ravichandran,SCAD Institute of Technology, 641664 Palladam, Tirupur, Tamil Nadu, India

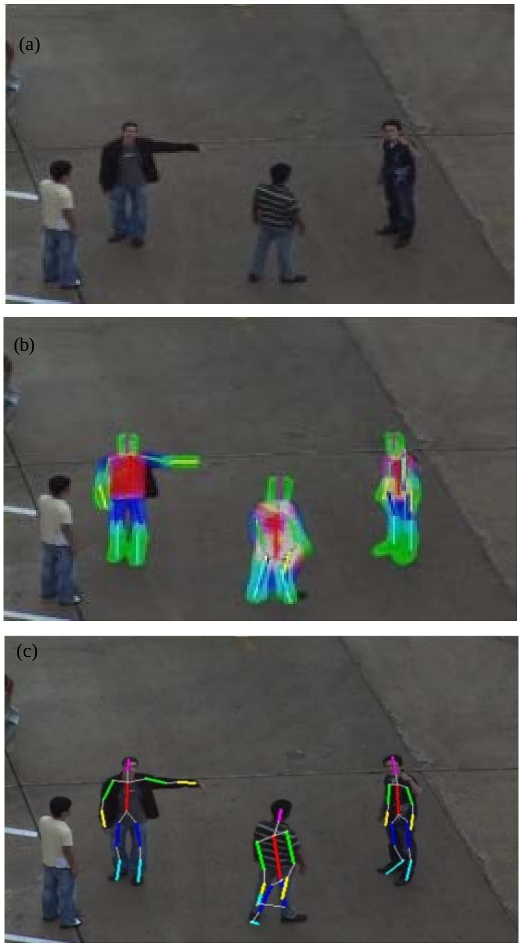

Fig. 1: Working of human pose estimator module (clockwise from top): a) Input image a frame from the UT Interaction Dataset; b) Pixels softened corresponding to different body parts; Red = Torso; Purple = Head; Green = Upper-arm; Yellow = Lower-arm; Dark Blue = Upper-leg (thigh); Light-blue = lower-leg. The more brighter the pixel the more probable it is to belong to a body part. (c) Stickmen deliverable with 10 vital parts, characterizing the pose

surveillance. Further, 2D poses forms the building blocks for determining 3D pose individual frames. The initial upper body detector is used because, human head, torso, arms majorly classify poses and provide sufficient information to detect the actions impending. The assumptions and the requirements imposed on the nature of the video feed are very trivial, in the sense that we require human beings to appear in a head-over-torso position. We do not impose any constraints on the clothing, sartorial choices or the background they appear in.

**Literture review:** The studies on human pose estimation in still images and videos are prodigious. Wide range of literature dating back as far as (Ioffe and Forsyth, 1999; Forsyth and Fleck, 1997) emphasizes the elusive nature of this topic. Major approaches to this can be broadly classified as bottom-up approaches (Hua *et al*., 2005; Lee and Cohen, 2004; Mori *et al*., 2004) and top-down approaches (Felzenszwalb and Huttenlocher, 2005; Mikolajczyk *et al*., 2004). In this study, we mainly concentrate on the past works which more or less had the same essence as ours pictorial structure. Ferrari *et al*. (2001) are some of the notable authors of literature on this topic based on which we build directly. Technologies as early as (Ioffe and Forsyth, 1999) achieved pose estimation and pictorial structure applications on naked human beings with uncluttered background. This success, though remarkable was not something we could benefit from. Since, the background had to be meticulously taken care of to be uncluttered, natural background has always been a challenge to this field as were people with unknown clothing. These two challenges took in a wide range of works trying to overcome them (Felzenszwalb and Huttenlocher, 2005; Buehler *et al*., 2008; Ramanan *et al*., 2005). Now, plainly thinking as to how to overcome these problems, in a brute-force manner-the least automatic yet highly credible-is to deduce the appearance models from segmented parts in a video (Buehler *et al*., 2008) where the segmentation was done manually. Alternative to this least automatic approach would be to carry out background subtraction and use the foreground pixels as a unary potential (Felzenszwalb and Huttenlocher, 2005; Lan and Huttenlocher, 2004, 2005). The famous (Ramanan *et al*., 2005) researches the frames throughout a video trying to match a predefined characteristic pose also known as strike-a-pose approach. The approaches discussed above cannot be applied to a single image as they require video. By far the only renowned approach to research with a single image with an unknown (as in natural) background is that of Ramanan's image parsing technique. It iteratively matches the appearance models starting out with just generic features such as edges and goes on incrementally improvising the appearance model with the estimation from previous step as input in every step. This was a big leap towards estimating poses of people with unrestricted and common sartorial as well as a flexible background from a single image rather than a video. As to the very recent literatures in human pose estimation include using:

- Adaptive pose priors
- Gradient based sophisticated features for detecting body parts (Dalal and Triggs, 2005; Tran and Forsyth, 2010)
- Colour segmentation

Our approach is a combination of Ramanan's research as well as Ferrari's work in unconstrained still images employing pictorial structures as in (Felzenszwalb and Huttenlocher, 2005).

## MATERIALS AND METHODS

### Implementation concepts

**Frame grabber:** Our system works primarily with images as the basic unit of processing. Therefore, we grab the frames from the video under consideration. MATLAB provides an off the shelf frame grabber algorithm. This method uses a VideoReader object to read individual frames. The algorithm can be tweaked a bit to modify the number of frames collected per second of the video.

**Key frame extractor:** The video frames grabbed passes through a series of computation intensive steps to detect punching activity. With 25 frames per second this computation puts a huge overhead on the system. Intuitively, the poses held by a human in a video changes very negligibly across a second and it is redundant to calculate for all 25 frames per second. Therefore, we extract the key frames of a video and only use those frames for the subsequent steps. We can exploit the temporal and spatial redundancy of a video for this step. We compare the histogram of a frame with the last key frame considered and if the change in histogram is past a threshold value we consider that a key frame. The threshold value is computed by considering all the frames, thus providing a reliable key frame extractor. The key frames extracted doesn't resemble each other thus eliminating redundant calculations while, at the same time, maintains the temporal order of the frames.

**Upper body detector:** The foremost step of this work is achieving a reliable upper body bounding box from a generic upper body detection algorithm. This bounding box greatly helps in restricting the search space for the possibilities of body parts in the immediate and subsequent steps of this research. We exploit the fact that in surveillance videos such as CCTV footages-people appear upright, as in the head over torso position, thereby avoiding having to account for every possibility just on pre-processing stages of our researches. These stages include Upper-body detection, which restricts the search space by approximating the position of human beings in the image and Foreground highlighting which nullifies the background clutter from affecting our processing. Detailed explanations on foreground highlighting are to follow this study. Thus, the ultimate benefit of a generic and a credible upper body detector is to restrict the position and appearance of body parts of a person in a particular image and thereby allows us to apply Ramanan's image parsing techniques.

**Create a detector object and set properties**

```
bodyDetector = Vision Cascade object detector('UpperBody');
bodyDetector. Minsize = [60 60];
bodyDetector, MergeThreshold = 10;
```

**Read input image and detect upper body**

```
12 = imread('visionteam.jpg')
bbox body = step(bodyDetector, 12);
```

**Annotate detected upper bodies**

```
IBody = insert objectAnnotation(12, 'rectanle', bbox
body, 'upper Body');
figure, Inshow (IBody), title ('Detected Upper bodies');
```

Fig. 2: Vision's Cascade object detector for detecting human upper body based on Viola-Jones algorithm (Viola and Jones, 2004)

Surveillance video feeds are typical, in the sense that human beings are mostly upright with upper body conspicuous. Thus, implementing a felicitous upper body detector will provide us with an edge in the subsequent stages of the estimator. Here, we critically weighed a variety of upper-body detectors based on factors vital to surveillance-accuracy and detection time. The upper body detectors which we considered used (Dalal and Triggs, 2005), sub-partitioning a frame into tiles based on Histogram of Oriented Gradients (HOG) using a linear SVM classifier. Another upper body detector which implemented (Felzenszwalb *et al*., 2008) Part Based Model (PBM) approach to detect upper bodies. We observed these detectors are reliable to a degree but they were considerably slow. Open CV provides an upper body detector complemented with a face detection provided remarkable reliability but lagged more than the previous detectors, owing to additional computations for face detection. The speed factor is almost indispensible in live applications such as surveillance. A cascaded Object detector system provided by the Vision package in MATLAB$^{TM}$ implements the Viola-Jones' algorithm (Viola and Jones, 2004) to detect upper bodies. This Vision's Upper Body detector, not as accurate as the HOG and PBM based detectors but notably faster within few hundred milliseconds-was also capable of obtaining multiple upper bodies from a single frame instantaneously. Therefore, an apt upper body detector to deliver the task is achieved by a tradeoff between the detection time and accuracy. Figure 2 shows the Vision's Cascade Object Detector for detecting human upper body based on Viola-Jones algorithm.

**Algorithm A: Pseudo code for the activity detector:**

```
    HumanPoseEstimator (VideoFile) do:
    Frames [] = Splitter (videoFile) //Split into distinct frames
    For each frame in Frames [] do:
boundingBox = UpperBodyDetector (frame)
    //find the location of upperbody
    fghigh = ForegroundHighlighter (boundingBox)
    // highlight foreground and eliminate background clutter
    poses = EstimPoses (fghigh)
    //soft pixelate the highlighted foreground
    segments [] = segmentPoses (poses)
    // get distinct line coordinates
    for each line(x, y) in segments [] do:
adjustedLine(x, y) = FilterAlgo (line(x, y))
    done
    keyFrames [] = ExtractKeyFrame (Frames [])
    for each kf in keyFrames [] do:
for each line(x, y) in segments [] do:
for corresponding cor_line(x, y) in other keyFrames [] do:
estimated_pose = compare (line(x, y), cor_line(x, y));

done
done
    done
```

**Median filter:** We deploy a running median filter with a window value of 5 frames. The pose estimator struggles to map the stickmen skeleton when there's motion blur in the video frames due to swift action. This might cause a sudden deflection in the stickmen coordinates.

This noise might give rise to false positive results. To avoid this we stabilize the noise in the coordinate values (x, y) by averaging across a window of 5 frames. It stabilizes the noisy coordinates of the stickmen's sticks using windowed mean procedure. The typical complementary filter algorithm was adopted for stabilization but we couldn't achieve stabilization of the highest degree until we applied windowed mean. Windowed mean takes up a constantly moving window for finding mean instead of the whole set of values. It works out really well for cases when we're unaware of the fluctuations in the future values. This module takes up the floating point coordinates of the stickmen. Each skeleton coordinates get stabilized on its own inside the module. The output of this module is the stabilized and noise free coordinates for the stickmen which can be used for consequent stages.

**Temporal association:** After applying the upper body detection to the frames in the video, we perform a temporal association of the resulting bounding boxes. That is, we associate the resulting bounding box coordinates of a particular frame across nearby time frame-both preceding and succeeding-to achieve continuity maximization. This temporal association is effectively viewed as a grouping problem where the entities to be grouped are the bounding-box coordinates.

The grouping has to be achieved across the video frames throughout the length. This is based on the trivial fact that the human upper body's bounding-box once detected, doesn't move abruptly across frames, rather smooth transition takes place between frames. We solve the grouping problem by using Clique Partitioning Algorithm of (Ferrari *et al*., 2001). We group the bounding boxes across nearby time frames maximizing the Intersection over union effectively. The algorithm is very flexible and fast. The main goal of temporally associating is to increase the intersection over union between frames. Often the upper-body detector produces false positives.

Since, the upper body detector processes one single frame at a time, it might produce False Positives. These are detections which turned out positive in a frame but do not occur for more than half a second in the overall video. These false positives tend to falter subsequent processes by leading on in an erroneous path. Temporal association helps in filtering these false positives. Thus, this method proves to be more substantial than the regular tracking which is also in accordance with (Ozuysal *et al*., 2006; Sivic *et al*., 2005). Figure 3 shows the Block diagram for the entire activity detection system.

**Foreground highlighting:** The bounding-box coordinates localizes the spatial possibility of the human body. With the upper body detections we can estimate the scale of human body in the video frames. The 2D poses with arms stretched out are detected as wider rectangular boxes. The wide bounding-box often has an ambiguous background giving rise to false positive detections from the upper body detector. We can overcome this issue, taking advantage of the knowledge, we have about the pictorial overlay of each area. For example we can localize the head somewhere along the middle of the bounding-box's upper-half and torso right below it but the arms cannot be precisely localized. With these regional localization denoting the probability of the person's presence in the bounding-box and using the initial foreground/background color models we start GrabCut (Rother *et al*., 2004). The GrabCut delivers a bounding box with a green overlay thereby nullifying the background clutter thus substantially assuaging the load on the subsequent steps. Figure 4 represents Foreground highlighting results on UT Interaction dataset.

**Computation time:** We present here a breakdown of the runtime of our HPE pipeline. The results are averaged over 10 runs on an Intel®Core™ i5-2450M CPU@ 2.50 GHz. The implementation is a mix of C++ and MATLAB code. Our proposed method attains the human detection at 2.3 sec, respectively. All further processing stages are repeated independently for the each detection:

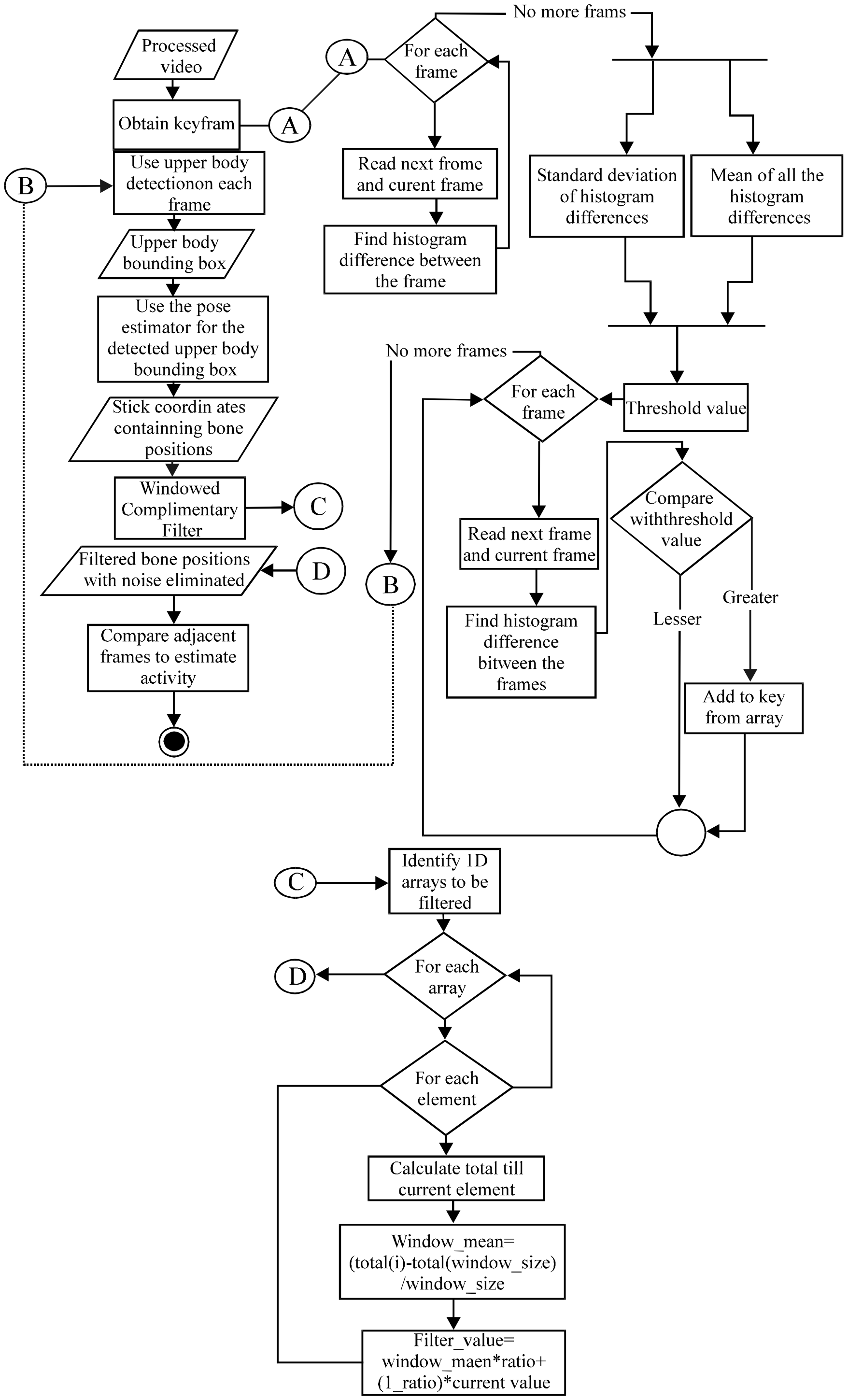
Processed video
Obtain keyfram
A
A
For each frame
No more frams
Read next frome and curent frame
Find histogram difference between the frame
Standard deviation of histogram differences
Mean of all the histogram differences
B
Use upper body detectionon each frame
Upper body bounding box
Use the pose estimator for the detected upper body bounding box
Stick coordin ates containning bone positions
Windowed Complimentary Filter
C
Filtered bone positions with noise eliminated
D
Compare adjacent frames to estimate activity
No more frames
For each frame
Threshold value
Read next frame and current frame
Find histogram difference bitween the frames
Compare withthreshold value
Lesser
Greater
Add to key from array
B
C
Identify 1D arrays to be filtered
D
For each array
For each element
Calculate total till current element
Window_mean= (total(i)-total(window_size) /window_size
Filter_value= window_maen*ratio+ (1_ratio)*current value

Fig. 3: Block diagram for the entire activity detection system

- Foreground highlighting 2.3 sec
- Estimating appearance models: 0.6 sec
- Parsing: Computing unary terms: 1.5 sec Inference: 0.8 sec
- Overhead of loading models, image resizing, etc.: 1.4 sec. After human detection, the total time for HPE on a person is 6.6 sec. The total time for an image is 3.3+6.6P sec with P the number of detections

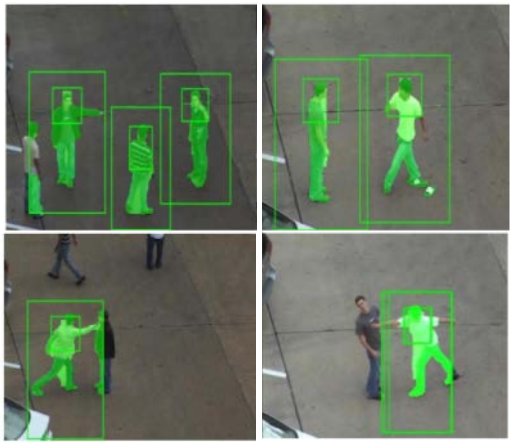

Fig. 4: Foreground highlighting results on UT Interaction dataset: The enlarged detection window area is shown by the green bounding box. Green patches depict the foreground segments picked by the algorithm. These foreground segments effectively remove majority of the background clutter from the box

This speed can be boosted by implementing our project on a high end sophisticated processor especially, with a higher frequency of execution and capable of running more threads.

## RESULTS AND DISCUSSION

**Evaluation:** Here, we present a comprehensive evaluation of our Human Pose Estimation algorithm (HPE). We start by describing the datasets used for training and testing. Then, we critically evaluate the different mechanisms adopted by the upper body detectors and try to establish a generic comparison between them.

**Datasets used:** The upper-body detectors are trained on a single set of images from BUFFY and ETHZ PASCAL dataset, manually annotated with bounding-boxes enclosing upper-bodies. The detectors' evaluation is carried out on a test set of video frames from UT Interaction Dataset (especially for the Punching and Kicking). This dataset contains 193 frames of which 85 are negative images (i.e., no frontal upper-bodies are visible) and the remaining ones contain 108 instances of frontal upper-bodies. The upper body detector and the pose estimator were trained on all the episodes of BUFFY dataset. Final evaluation is supposed to be carried out on the UT interaction dataset for punching. The Buffy and ETHZ PASCAL datasets were released by Martin Eichner Publications on behalf of Calvin and it contains a set of 96 images meticulously selected, keeping in mind not to compromise on the diversity and poses, also this training dataset contains near frontal and near rear views of human beings in a variety of sartorial choices, thereby challenging our estimator in a constructive manner. The UT interaction dataset contains a set of 7 classes of different actions and poses both near frontal and near rear-view. Additionally our software was also vigorously tested on the Perona November 2009 Challenge which is a set of images captured by Pietro Perona and his coworkers in order to challenge pose estimator after critically examining the factors influencing human pose estimation.

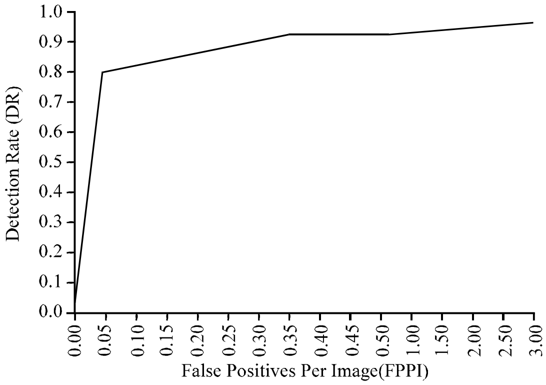

Fig. 5: Detection rate VS false positives per image trace on part based upper body detector (Felzenszwalb *et al*., 2008)

**Evaluating upper-body detectors:** Upper body detector was evaluated with 187 images from the UT Interaction Dataset for punching and kicking. The set was diverse in its own way containing 102 positive images and 85 negative images where the upper body was either hindered or the human is posed with his side facing the camera. Figure 5 and 6 shows the comparison (detection rate (DR) versus False Positives Per Image (FPPI)) between two frontal view upper-body detectors: HOG-based upper body detector (Ferrari *et al*., 2008) Part-Based (PBM) (Felzenszwalb *et al*., 2008) upper body detector Practically both detectors work well for all viewpoints in a 30 degree pan on both sides of the straight-on frontal and back views. We observe that the detection rate for the HOG based upper body detector is almost 90% if we accept 1 false positive every 10 images, i.e., 0.1 FPPI which is evidently more than that of Part-Based Model Upper Body detector at the same specifications. Calvin Upper Body detector, an HOG based detector was used for evaluation. Faster detections can be possible with the MATLAB's vision packages inbuilt detector but the detection rate is substantially low for the same specifications. Figure 7 show the running median filter applied for the lower limb coordinates and Fig. 8 show the running median filter with a window of 5 video frames applied for the lower limb coordinates. We count the detection as legitimate if it crosses the PASCAL VOC criterion (beyond 0.5 detection with ground-truth bounding box).

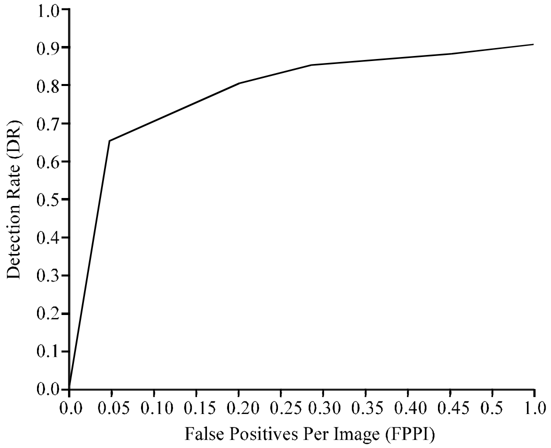

Fig. 6: Detection rate vs false positives per image trace on Histogram of gradients based upper body detector (Dalal and Triggs, 2005)

**Evaluating activity detector for punching activity:** The videos used for evaluating this module consisted of a mixture of activities including punching, pointing, kicking, pushing and handshaking. The activity detector was evaluated for punching activity detection. With minor changes in the threshold values, the pose estimator can be used to estimate any activity. Table 1 comprehends the test runs of the activity detection system.

- Video file specifies the source of the frames
- Range specifies the frame range from start to end
- Bounding box specifies the coordinates of the rectangular bounding box which encloses the human body of our interest
- Left specifies the percentage confidence for punching activity in left limb
- Right specifies the percentage confidence for punching activity in right limb
- Actual activity performed specifies the original activity in the frame range

**Results of HPE on UTI datasets:** Figure 9-12 shows the pointing sequence, kicking sequence, punching sequence, poses obtained from UTI dataset.

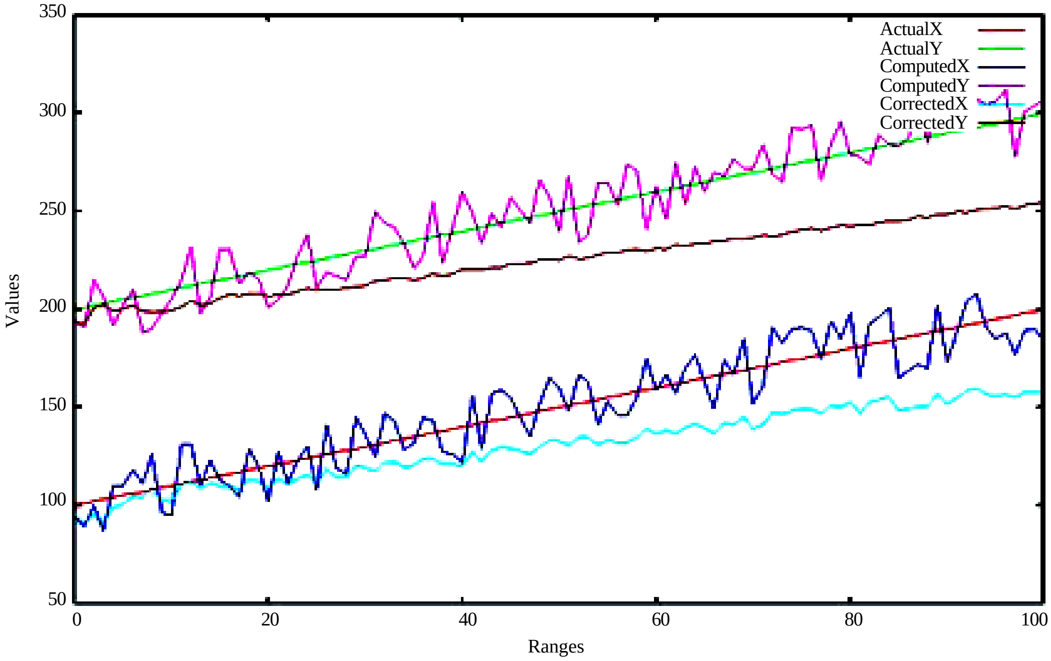

Fig. 7: Running median filter applied for the lower limb coordinates

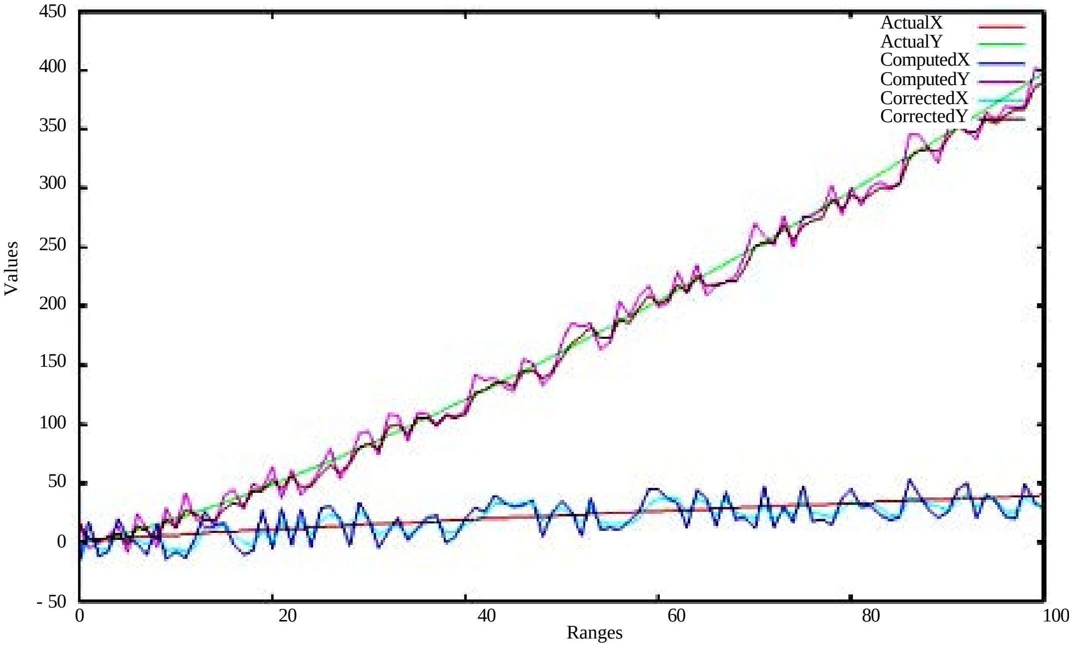

Fig. 8: Running median filter with a window of 5 video frames applied for the lower limb coordinates

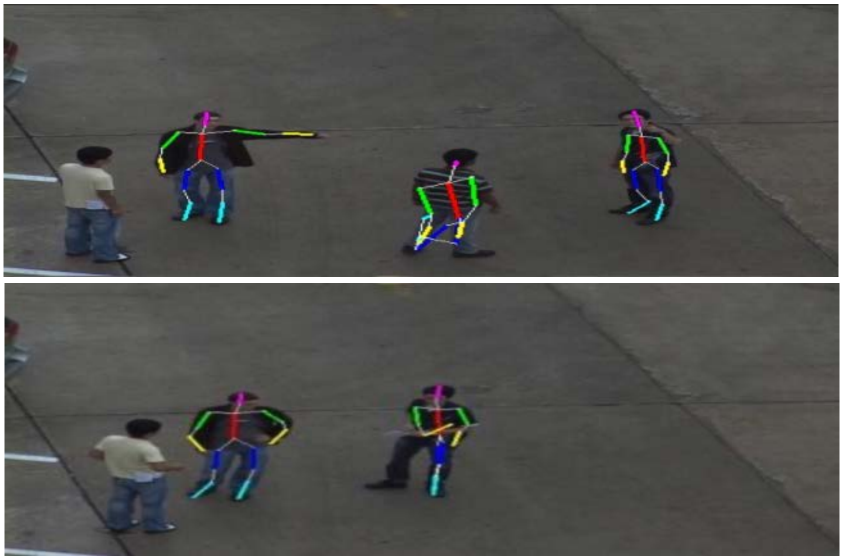

Fig. 9: Pointing sequence from UTI dataset

Table 1: Comprehends the test runs of the activity detection system.

| Video file frame range | Bounding box | Left | Right | Actual activity performed |
|---|---|---|---|---|
| 733-738 (seq5.avi) | [298 154 113 106] | None | 95.82 | Punching |
| 588-600 (seq5.avi) | [232 154 120 96] | none | none | Handshake |
| 876-900 (seq5.avi) | [257 161 162 105] | 217.36, 50.69% | 181.15, 15.87% | Kicking |
| 460-480 (seq2.avi) | [364 120 142 90] | none | 93.23% | Punching |
| 664-684 (seq2.avi) | [377 140 101 82] | none | none | Pushing |
| 771-799 (seq2.avi) | [464 121 128 90] | none | none | Handshaking |
| 830-840 (seq16.avi) | [310 166 108 70] | none | 93.56% | Punching |
| 350-365 (seq16.avi) | [201 164 95 75] | none | none | Pointing |
| 2675-2705 (seq9.avi) | [172 188 88 60] | none | 86.16% | Punching |
| 2785-2800 (seq9.avi) | [535 140 95 70] | none | none | Pointing |
| 2920-2940 (seq9.avi) | [185 191 97 70] | none | 22.43% wo filter 16.6% w filter | Kicking |

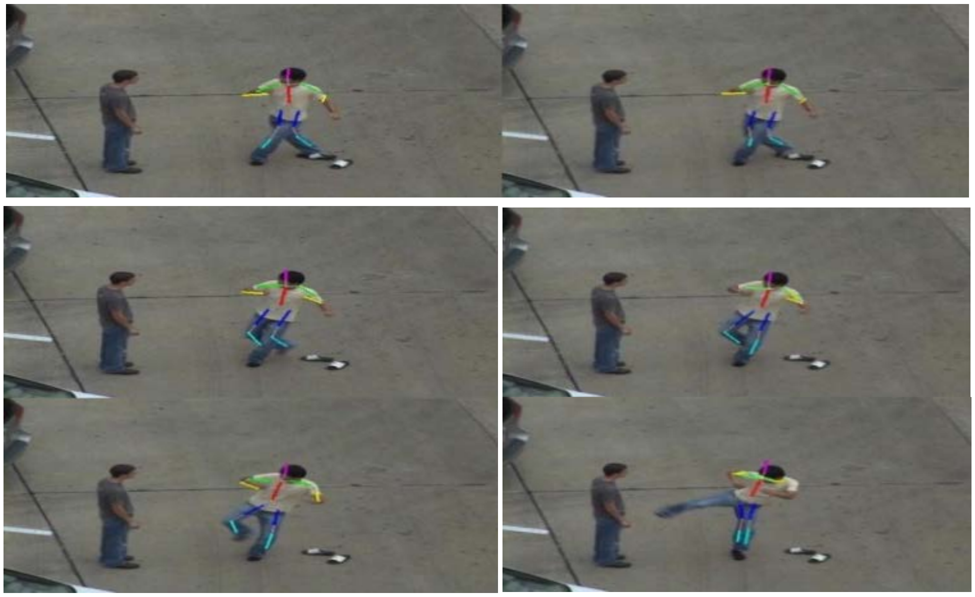

Fig. 10: Kicking sequence from UTI dataset

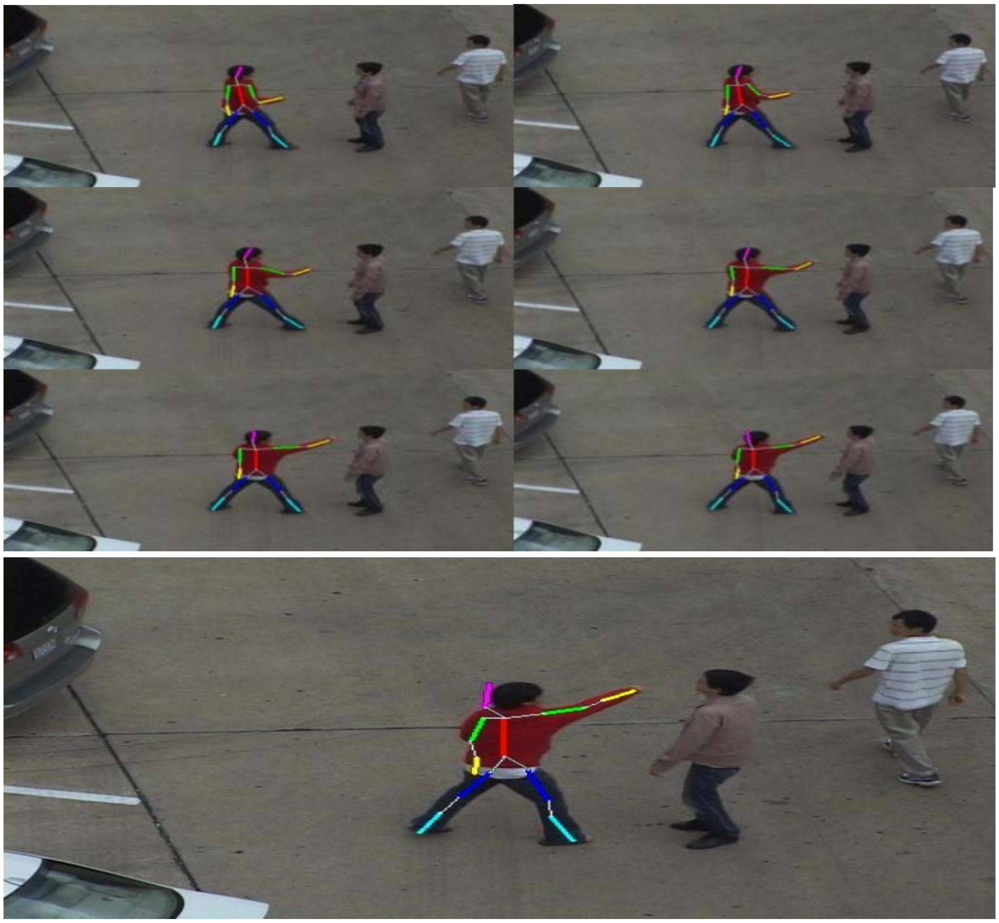

Fig. 11: Punching sequence from UTI dataset

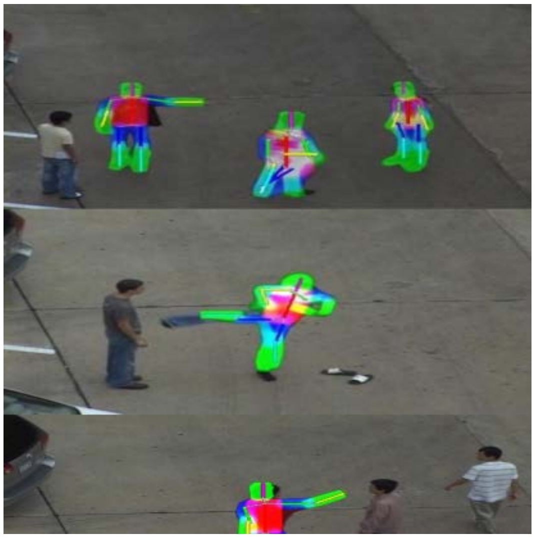

Fig. 12: Poses obtained from UTI dataset

## CONCLUSION

We have presented a Human pose estimating system which can work efficiently on even cluttered background without any prior knowledge or assumption as to human's sartorial choices. It works through a video, one frame at a time, temporally associating and exploiting the fact that people are positioned upright to estimate the 2D pose. The system is stable and reliable only for near-frontal and near-back viewpoints. The lighting and the quality of the video feed are tolerable as the system is fairly flexible for those factors. Basically, the system contains distinct modules which are linked serially and they are mutually exclusive. The first module is a generic upper body detector which gives a bounding-box roughly estimating the position of the human body. Temporally, associating this bounding-box through a Clique Partitioning and casting it as a grouping problem we can filter off the possible false positives. Foreground highlighting on these bounding boxes and soft pixelating the possible positions of body parts forms the final modules. This modularity of the system makes it pliable for the future extensions and scalability.

## RECOMMENDATIONS

The system as of now cannot reliably take over manned surveillance system owing to the fact that it fails to work efficaciously with the presence of occlusions or from side view points. These are potential future extensions for our system. Further, the computation time for the system as a whole is slow than expected which makes it impractical for real-time applications. The system can be more rapid by enhancing it module wise. For example, the upper body detector can be optimized in a way which provides a faster as well as a credible output for the input video frames. Also, as mentioned previously these 2D poses form a basic building block for estimating 3D poses from individual frames. This system can be extended to be used for automated surveillance which sets off an alarm if a specific prohibited activity is detected.

## REFERENCES

Buehler, P., M. Everingham, D.P. Huttenlocher and A. Zisserman, 2008. Long term arm and hand tracking for continuous sign language TV broadcasts. Proceedings of the 19th Conference on British Machine Vision, September 1-4, 2008, BMVA Press, London, UK., ISBN 978-1-901725-36-0, pp: 1105-1114.

Dalal, N. and B. Triggs, 2005. Histograms of oriented gradients for human detection. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition, 1: 886-893.

Eichner, M., M.M. Jimenez, A. Zisserman and V. Ferrari, 2012. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. Int. J. Comput. Vision, 99: 190-214.

Felzenszwalb, P., D. McAllester and D. Ramanan, 2008. A discriminatively trained, multiscale, deformable part model. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, Anchorage, AK., 1-8.

Felzenszwalb, P.F. and D.P. Huttenlocher, 2005. Pictorial structures for object recognition. Int. J. Comput. Vision, 61: 55-79.

Ferrari, V., M.M. Jimenez and A. Zisserman, 2008. Progressive search space reduction for human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, June 23-28, 2008, IEEE, Anchorage, AK., pp: 1-8.

Ferrari, V., T. Tuytelaars and L.V. Gool, 2001. Real-time affine region tracking and coplanar grouping. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001, December, 8-14, 2001, IEEE, London, USA., pp: 226-233.

Forsyth, D.A. and M.M. Fleck, 1997. Body plans. Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1997, June 17-19,1997, IEEE, San Juan, Puerto Rico, pp: 678-683.

Hua, G., M.H. Yang and Y. Wu, 2005. Learning to estimate human pose with data driven belief propagation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, June 20-25, 2005, IEEE, USA., pp: 747-754.

Ioffe, S. and D. Forsyth, 1999. Finding people by sampling. Proceedings of the Seventh IEEE International Conference on Computer Vision 1999, September 20-27, 1999, IEEE, Kerkyra, Greece, pp: 1092-1097.

Lan, X. and D.P. Huttenlocher, 2004. A unified spatio-temporal articulated model for tracking. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004, 27 June-2 July 2004, IEEE, London, USA., ISBN: 0-7695-2158-4, pp: 722-729.

Lan, X. and D.P. Huttenlocher, 2005. Beyond trees: Common-factor models for 2d human pose recovery. Proceedings of the Tenth IEEE International Conference on Computer Vision ICCV 2005, October 17-21, 2005, IEEE, London, USA., ISBN: 0-7695-2334-X, pp: 470-477.

Lee, M.W. and I. Cohen, 2004. Proposal maps driven mcmc for estimating human body pose in static images. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004, June 27-2 July, 2004, IEEE, London, USA., pp: 334-341.

Mikolajczyk, K., C. Schmid and A. Zisserman, 2004. Human detection based on a probabilistic assembly of robust part detectors. In: Computer Vision-ECCV 2004. Tomas, P. and M. Jiri (Eds.). Springer Berlin Heidelberg, Berlin, Germany, pp: 69-82.

Mori, G., X. Ren, A.A. Efros and J. Malik, 2004. Recovering human body configurations: Combining segmentation and recognition. IEEE Comput. Vision Pattern Recogn., 2: 326-333.

Ozuysal, M., V. Lepetit, F. Fleuret and P. Fua, 2006. Feature Harvesting for Tracking-by-Detection. Computer Vision-ECCV 2006. Ales, L., B. Horst and P. Axel (Eds.). Springer Berlin Heidelberg, Berlin, Germany, pp: 592-605.

Ramanan, D., D.A. Forsyth and A. Zisserman, 2005. Strike a pose: Tracking people by finding stylized poses. Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, June 20-25, 2005, IEEE, London, USA., ISBN: 0-7695-2372-2, pp: 271-278.

Rother, C., V. Kolmogorov and A. Blake, 2004. GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Trans. Graphics, 23: 309-314.

Sivic, J., M. Everingham and A. Zisserman, 2005. Person Spotting: Video Shot Retrieval for Face Sets. In: Image and Video Retrieval. Leow, W.K., M.S. Lew, T.S. Chua, W.Y. Ma and L. Chaisorn *et al.* (Eds.). Springer Berlin Heidelberg, Berlin, Germany, pp: 226-236.

Tran, D. and D. Forsyth, 2010. Improved Human Parsing with a Full Relational Model. In: Computer Vision-ECCV 2010. Kostas, D., M. Petros and P. Nikos (Eds.). Springer Berlin Heidelberg, Berlin, Germany, pp: 227-240.

Viola, P. and M.J. Jones, 2004. Robust real-time face detection. Int. J. Comput. Vision, 57: 137-154.