I J C T A, 9(2) 2016, pp. 455-466
© International Science Press

Human Pose Estimation from UT Interaction Dataset

Ravichandran C. G.[1] and Sivaprakash P.*[2]

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 orthodox call 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 try to roughly estimate the poses held with the temporal association between the successive and the previous video frames. One of the several challenges we face is the cluttered and natural background of the CCTV footages. Manned surveillance should be replaced only by a credible and reliable system. The system should be able to work 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 viewpoints.

Keywords: Human pose estimation, UTI dataset, Activity detection, Upper body detection and tracking

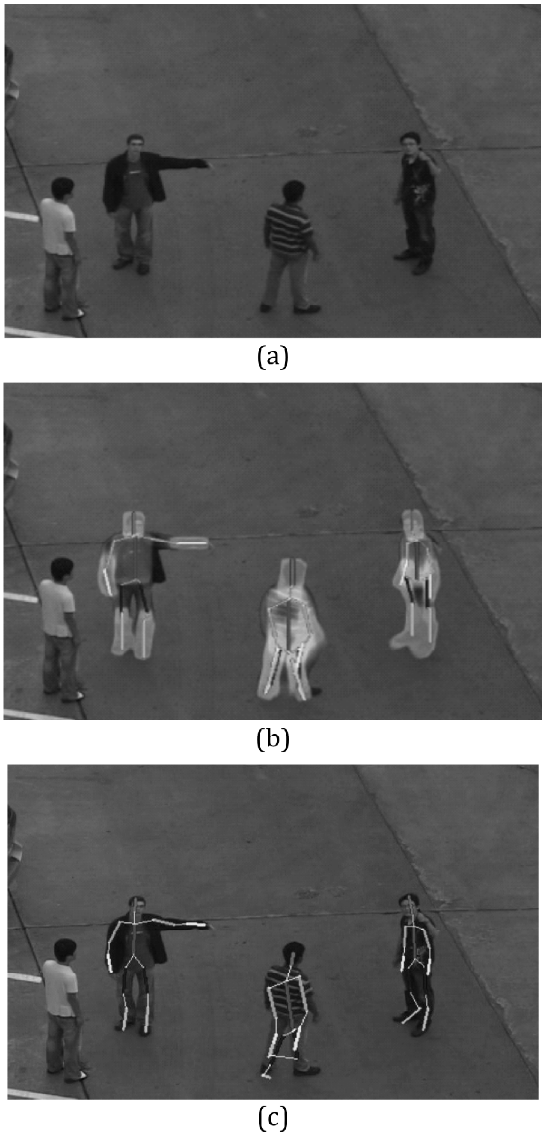
INTRODUCTION

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-pixelated. 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 (2006) and Ferrari's articulated human pose estimator Ren (2005) 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). 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

[1] Principal, SCAD Institute of Technology, Palladam - 641 664, Tirupur, India Tamilnadu, India

[2] Dept of Electronics and Communication Engineering, RVS College of Engineering and Technology, Dindigul 624005,Tamilnadu, India, *E-mail: siva465prakash@gmail.com*

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. 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 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.


(a)

(b)

(c)

Figure 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.

BACKGROUND AND RELATED WORK

The studies on Human pose estimation in still images and videos are prodigious. Wide range of literature dating back as far as [Felzenszwalb, 2005., 10] emphasizes the elusive nature of this topic. Major approaches to this can be broadly classified as bottom-up approaches [Mikolajczyk, 2004.,Sapp, 2010., andFelzenszwalb, 2008] and top-down approaches [Ioeffe, 1999., Dalal, 2005]. In this section we mainly concentrate on the past works which more or less had the same essence as ours - pictorial structure. Ramanan(2006) and Ferrari (2008) are some of the notable authors of literature on this topic based on which we build directly. Technologies as early as Eichner (2012) 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 [Ferrari 2008.,Ioeffe, 1999., Kumar, 2009., and Tran, 2009]. 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 (Ferrari, 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 [Ioeffe, 1999.,Lan, 2005., and Lan, 2004]. The famous Tran(2010) searches 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 work with a single image with an unknown (as in natural) background is that of Ramanan's(2006) 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:

(1) Adaptive pose priors (Ferrari, 2001)

(2) Gradient based - sophisticated features for detecting body parts [Forsyth, 1997.,Buehler, 2008., Ramanan, 2005., Özuysal, 2006]

(3) Colour segmentation (Ferrari, 2001).

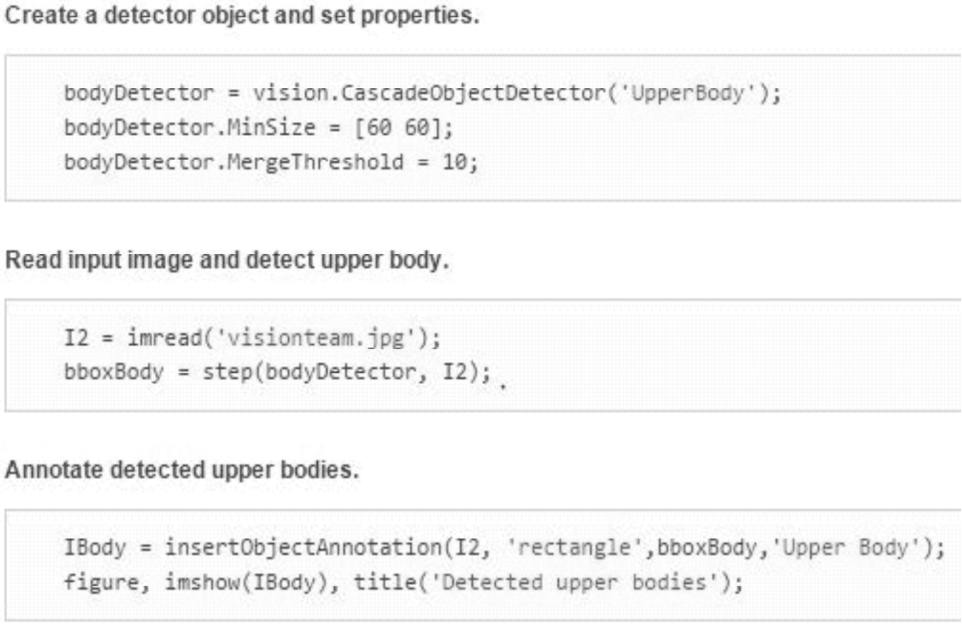
Our approach is a combination of Ramanan's work (2006) as well as Ferrari's work in unconstrained still images employing Pictorial Structures (Ioeffe, 1999).

IMPLEMENTATION CONCEPTS

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 work. 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 works. 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 section. 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(2006) Image parsing techniques.

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 (Forsyth, 1997), sub-partitioning a frame into tiles based on Histogram of Oriented Gradients (HOG) using a linear SVM classifier. Another upper body detector which implemented (Hua, 2005) Part Based Model (PBM) approach to detect upper bodies. We observed these detectors are reliable to a degree, but they were considerably slow. OpenCV provides a 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 (2004) to detect upper bodies.



Create a detector object and set properties.

```
bodyDetector = vision.CascadeObjectDetector('UpperBody');
bodyDetector.MinSize = [60 60];
bodyDetector.MergeThreshold = 10;
```

Read input image and detect upper body.

```
I2 = imread('visionteam.jpg');
bboxBody = step(bodyDetector, I2);
```

Annotate detected upper bodies.

```
IBody = insertObjectAnnotation(I2, 'rectangle',bboxBody,'Upper Body');
figure, imshow(IBody), title('Detected upper bodies');
```

Figure 2: Vision´s Cascade Object Detector for deteting Human Upper Body based on Viola-Jones Algorithm (2004)

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.

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 Lee, 2004. 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 a 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 [Johnson, 2010., Sivic, 2005].

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. 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

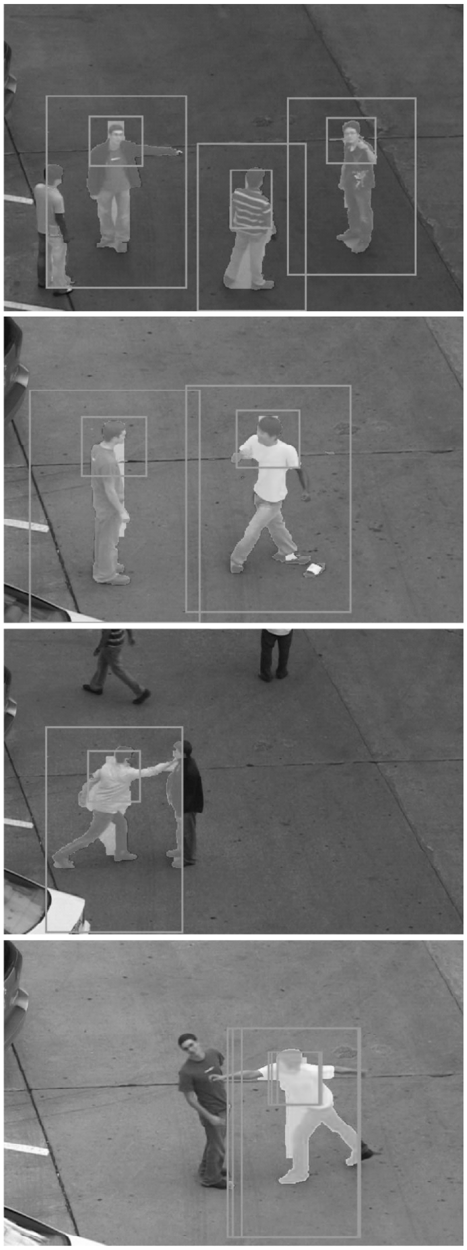


Figure 3: Foreground Highlighting results on UT Interaction dataset: The enlarged detection window area is shown by the green bounding box. Green patches depict the foreground segments picked by the algorithm. These foreground segments effectively remove majority of the background clutter from the box.

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, 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.

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.

Human detection takes 2.3 sec.All further processing stages are repeated independently for each detection:

(1) Foreground highlighting 2.3 sec.

(2) Estimating appearance models: 0.6 sec.

(3) Parsing:

Computing unary terms: 1.5 sec.

Inference: 0.8 sec.

(4) 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.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.

CONCLUSION

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. 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 PietroPerona and his coworkers in order to challenge pose estimator after critically examining the factors influencing Human Pose Estimation.

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 4 and 5 shows the comparison (detection rate (DR) versus false positives per image (FPPI)) between two frontal view upper-body detectors:

(i) HOG-based upper body detector Ren (2005)

(ii) Part-based (Hua, 2005) (PBM) upper body detector

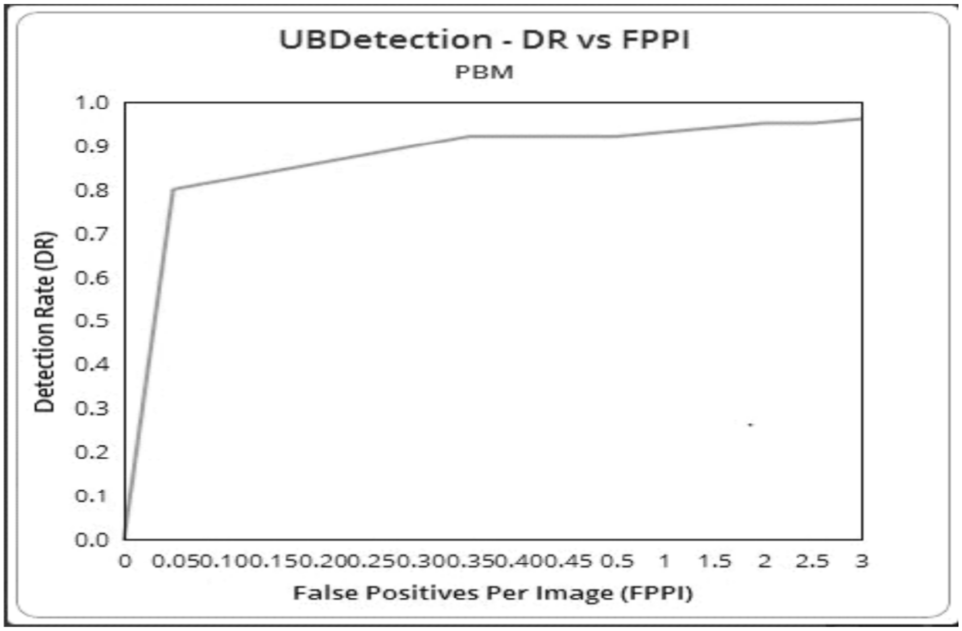


Figure 4: Detection Rate VS False Positives per Image trace on Part based upper body detector(Hua, 2005)

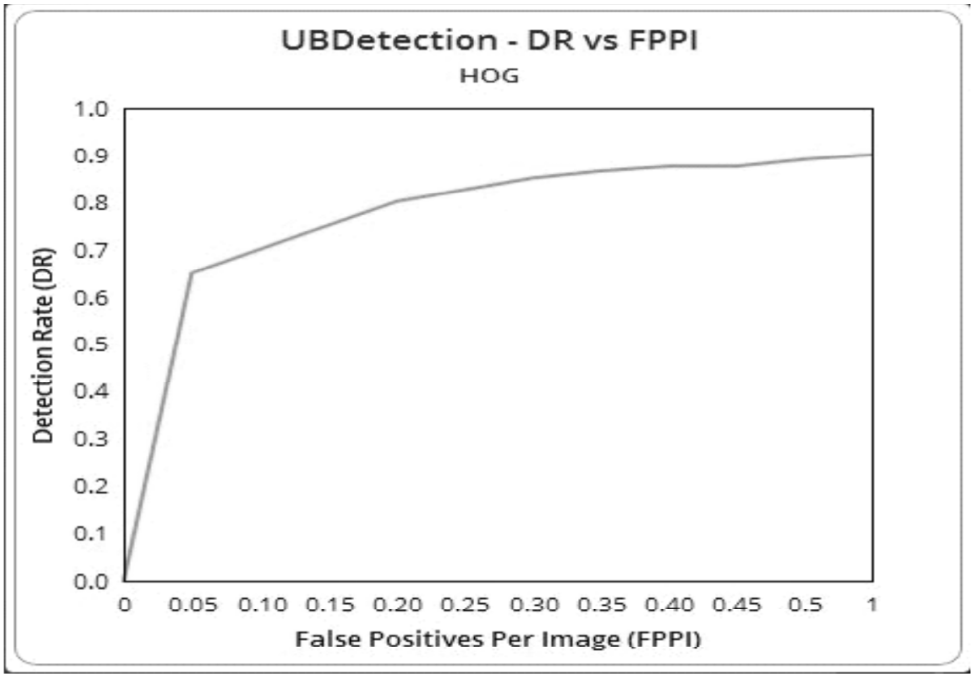


Figure 5: Detection Rate VS False Positives Per Image trace on Histogram of Gradients based Upper Body detector (Forsyth, 1997)

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. We count the detection as legitimate if it crosses the PASCAL VOC criterion (beyond 0.5 detection with ground-truth bounding box). Figures 6, 7 and 8 give the HPE results for pointing, kicking and punching sequences respectively. Figure 9 shows the poses obtained for each sequence.

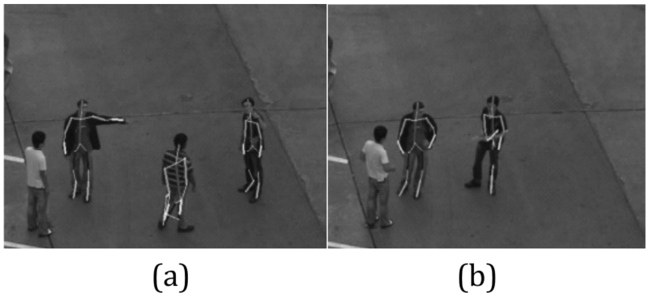
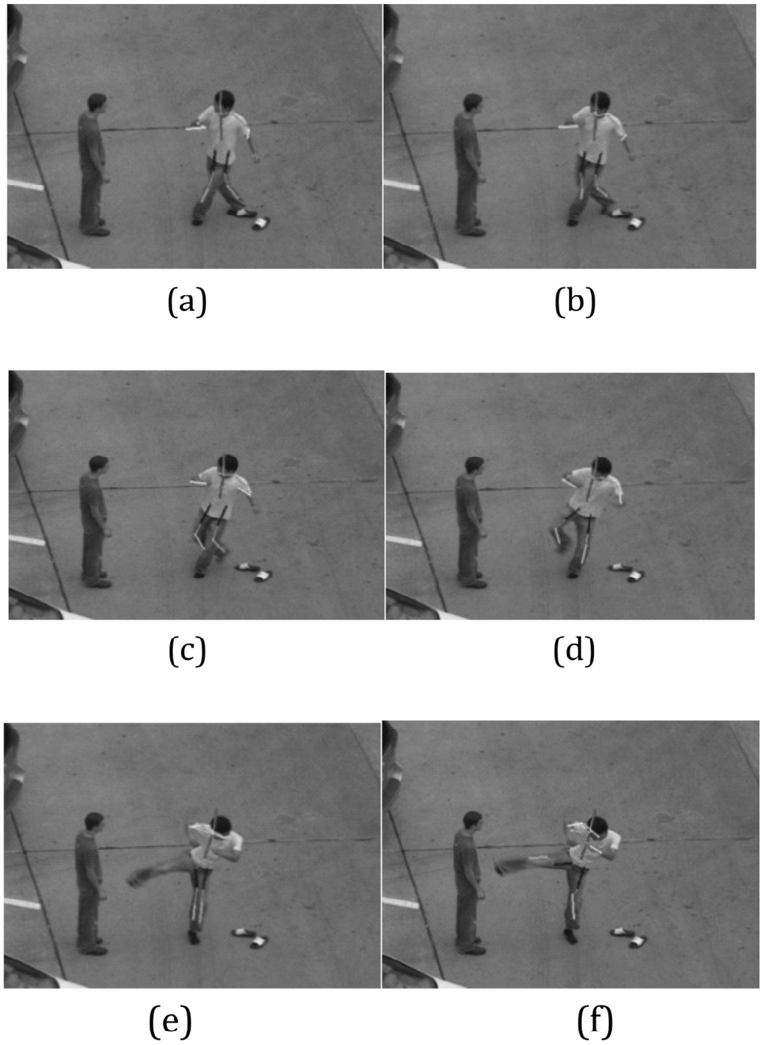

(a) (b)

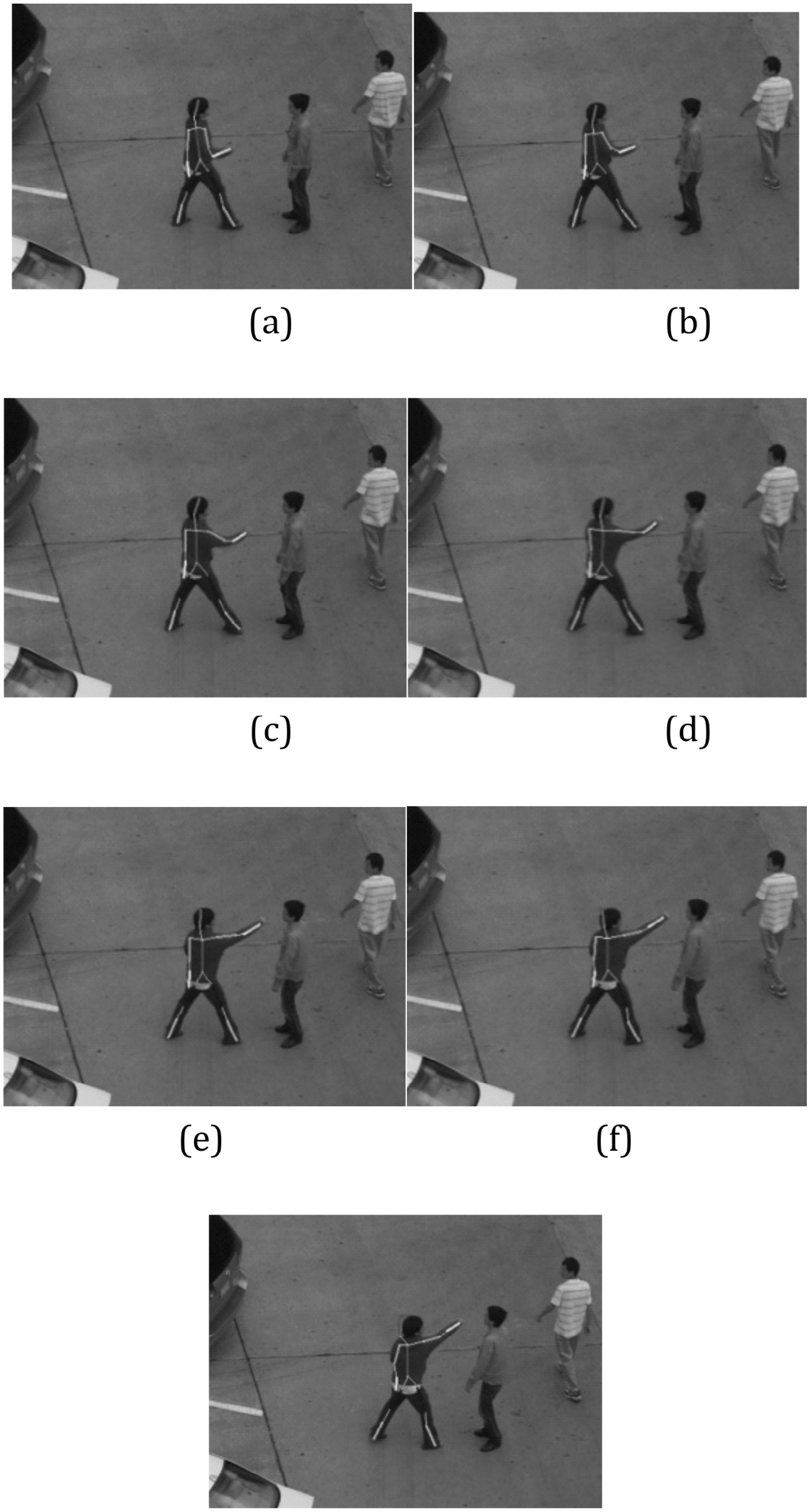
Figure 6: Pointing Sequence from UTI Dataset


(a) (b)

(c) (d)

(e) (f)

Figure 7: Kicking Sequence from UTI Dataset



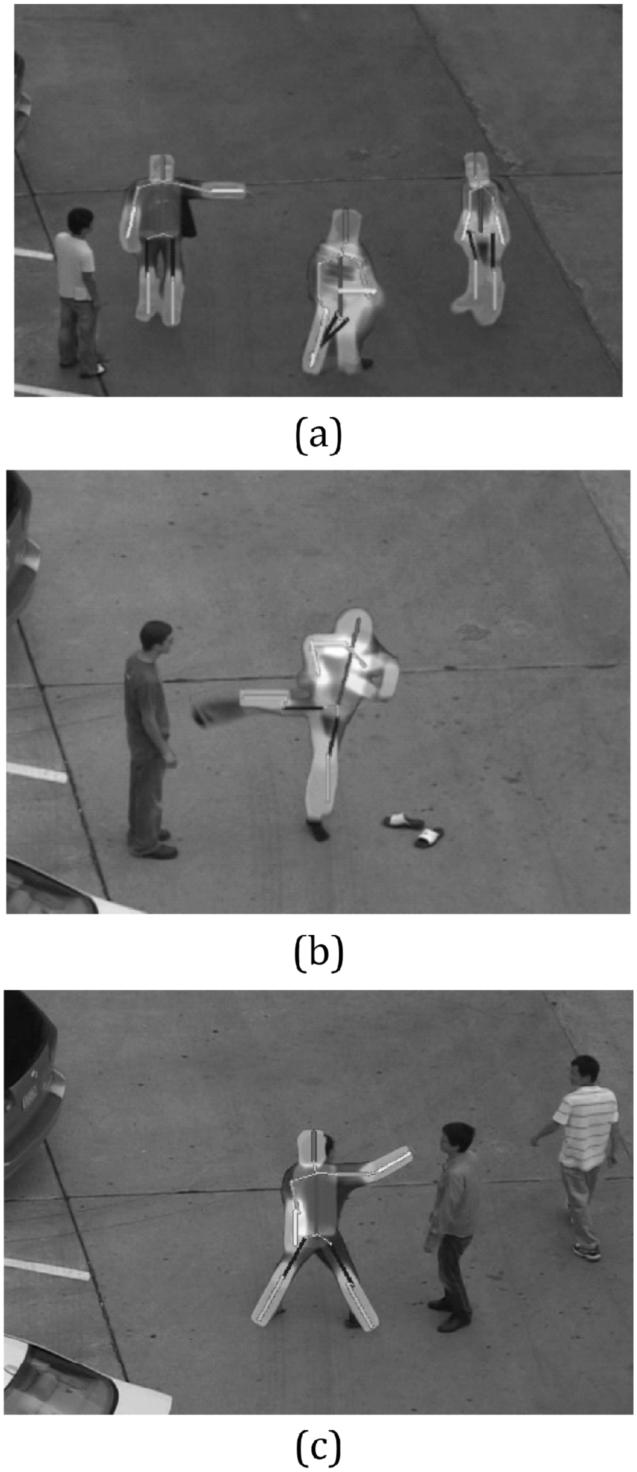
(a) (b)

(c) (d)

(e) (f)

(g)

Figure 8: Punching sequence from UTI Dataset


(a)

(b)

(c)

Figure 9: Poses Obtained from UTI Dataset for (a) Pointing (b) Kicking (c) Punching

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.

FUTURE WORK

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 viewpoints. 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.

REFRENCES

[1] Buehler, P., Everingham, M., Huttenlocher, D.P. and Zisserman, A., 2008.Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the 19th British Machine Vision Conference*: 1105-1114.

[2] Dalal, N. and Triggs, B., 2005, June.Histograms of oriented gradients for human detection.In *Computer Vision and Pattern Recognition, 2005.CVPR 2005. IEEE Computer Society Conference,*1: 886-893.

[3] Eichner, M., Marin-Jimenez, M., Zisserman, A. and Ferrari, V., 2012.2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, *99*(2):190-214.

[4] Felzenszwalb, P., McAllester, D. and Ramanan, D., 2008, June.A discriminatively trained, multiscale, deformable part model.In *Computer Vision and Pattern Recognition, 2008.CVPR 2008. IEEE Conference :* 1-8.

[5] Felzenszwalb, P.F. and Huttenlocher, D.P., 2005. Pictorial structures for object recognition. *International Journal of Computer Vision*, *61*(1): 55-79.

[6] Ferrari, V., Marin-Jimenez, M. and Zisserman, A., 2008, June.Progressive search space reduction for human pose estimation.In *Computer Vision and Pattern Recognition, 2008.CVPR 2008. IEEE Conference :*1-8.

[7] Ferrari, V., Tuytelaars, T. and Van Gool, L., 2001.Real-time affine region tracking and coplanar grouping.In *Computer Vision and Pattern Recognition, 2001.CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference*, 2: II-226.

[8] Forsyth, D.A. and Fleck, M.M., 1997, June. Body plans. In *Computer Vision and Pattern Recognition, 1997.Proceedings., 1997 IEEE Computer Society Conference :* 678-683.

[9] Hua, G., Yang, M.H. and Wu, Y., 2005, June.Learning to estimate human pose with data driven belief propagation.In *Computer Vision and Pattern Recognition, 2005.CVPR 2005. IEEE Computer Society Conference,* 2: 747-754.

[10] Ioffe, S. and Forsyth, D., 1999.Finding people by sampling.In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference*, 2 :1092-1097.

[11] Johnson, S. and Everingham, M., 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC* 2(4): 5.

[12] Kumar, M.P., Zisserman, A. and Torr, P.H., 2009, September.Efficient discriminative learning of parts-based models. In *Computer Vision, 2009 IEEE 12th International Conference:* 552-559.

[13] Lan, X. and Huttenlocher, D.P., 2004, July. A unified spatio-temporal articulated model for tracking.In *Computer Vision and Pattern Recognition, 2004.CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference* , 1: I-722.

[14] Lan, X. and Huttenlocher, D.P., 2005, October. Beyond trees: Common-factor models for 2d human pose recovery. In *Computer Vision, 2005.ICCV 2005. Tenth IEEE International Conference*, 1: 470-477.

[15] Lee, M.W. and Cohen, I., 2004, July. Proposal maps driven mcmc for estimating human body pose in static images. In *Computer Vision and Pattern Recognition, 2004.CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference*, 2: II-334.

[16] Mikolajczyk, K., Schmid, C. and Zisserman, A., 2004. Human detection based on a probabilistic assembly of robust part detectors. In *Computer Vision-ECCV 2004* : 69-82.

[17] Özuysal, M., Lepetit, V., Fleuret, F. and Fua, P., 2006. Feature harvesting for tracking-by-detection. In *Computer Vision–ECCV 2006*: 592-605.

[18] Ramanan, D., 2006. Learning to parse images of articulated bodies. In*Advances in neural information processing systems* : 1129-1136.

[19] Ramanan, D., Forsyth, D.A. and Zisserman, A., 2005, June. Strike a pose: Tracking people by finding stylized poses. In *Computer Vision and Pattern Recognition, 2005.CVPR 2005. IEEE Computer Society Conference,* 1: 271-278.

[20] Ren, X., Berg, A.C. and Malik, J., 2005, October.Recovering human body configurations using pairwise constraints between parts.In *Computer Vision, 2005.ICCV 2005. Tenth IEEE International Conference,* 1: 824-831.

[21] Rother, C., Kolmogorov, V. and Blake, A., 2004.Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG), 23*(3): 309-314.

[22] Sapp, B., Jordan, C. and Taskar, B., 2010, June. Adaptive pose priors for pictorial structures. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference:* 422-429.

[23] Sivic, J., Everingham, M. and Zisserman, A., 2005. Person spotting: video shot retrieval for face sets. In *Image and Video Retrieval* : 226-236.

[24] Tran, D. and Forsyth, D., 2010. Improved human parsing with a full relational model. In *Computer Vision–ECCV 2010*: 227-240.

[25] Viola, P. and Jones, M.J., 2004. Robust real-time face detection.*International journal of computer vision*, *57*(2):137-154.