Body Pose based Pedestrian Tracking in a Particle Filtering Framework

Md. Junaedur Rahman[1], Jesus Martinez Del Rincon[2], Jean-Christophe Nebel[1] and Dimitrios Makris[1]
1.Digital Imaging Research Centre, Kingston University
Kingston-upon-Thames KT1 2EE, UK
2. The Institute of Electronics, Communications and Information Technology (ECIT)
Queen's University Belfast, NI Science Park, Queen's Road, Queen's Island, Belfast,BT3 9DT
[j.rahman@kingston.ac.uk, j.martinez-del-rincon@qub.ac.uk, {j.nebel, d.makris}@kingston.ac.uk]

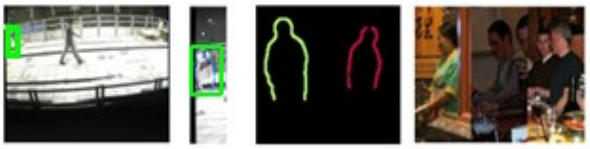
Abstract

A novel body pose based human tracking model is proposed for pedestrian tracking. This work investigates the challenges of reliable pedestrian tracking and proposes an improved model under challenging environments. Specifically, it claims that it is useful to exploit the curvature information of different body poses in tracking framework to overcome general tracking problems. In this paper different body pose detectors are combined as a useful feature for tracking. Performance has been evaluated in a rich evaluation framework. Result shows that poselet based features are more suitable for tracking than just detecting the person over the frames.

1. Introduction

It has become more and more demanding nowadays to process the surveillance videos and obtain useful information. An efficient human tracking model is therefore essential to overcome common tracking challenges and assist in higher level surveillance tasks such as activity recognition. In video sequences, pedestrian detection in a single frame may assist the localization of the pedestrian in subsequent frames exploiting the temporal coherence. In recent studies, the poselet feature detector has been proved successful even in most of the challenging situations. However, false detections may arise when near metameric shapes are present or detections may be missed due to low contrast. A tracking framework may significantly improve the pedestrian localization task. The pedestrian tracking task in this paper is focused on videos captured with a single surveillance camera.

Object detectors are based on discriminative features for reliable performance. In challenging environments the Histogram of Oriented Gradients feature has been proved successful and vastly used in recent days because of its rotational invariance, scale invariance and easy computation [1, 4, 5, 6, 7]. Grabner et al. has proposed a pedestrian scene specific whole body detector [7]. It takes the whole body patches for training, so when in real life one dynamic appearance of pedestrian is found which does not match the pattern, it fails. Due to the number of missed detections and false positives in HOG detection the system was considered unreliable for tracking solutions. The main challenge of using HOG detector for tracking is to correlate the detection responses which has included missing and inconsistent detections over the consecutive frames. Until recently not many people have tried to use the part based body pose information to detect the object except Lubomir et al. 2009 [2]. The Poselet feature uses collections of body poses instead of considering only the shape of the whole body.



A B C D

Figure 1: False positive on "Near-Metameric" shape. Here, (A) false positive, (B) detected poselet bounds, (C) activated poselet and (D) corresponding training patches.

Mostly, the problem is occurred when "Near-Metamerics" objects in the background is found in a scene [3]. The word metamerics comes from biology and it means body segments of an object which are fundamentally similar to the basic structure. Here, in tracking it is common to have human body or part like shapes in the background. When the detector is applied it creates false positives on those objects. In tracking it is severe because it affects the performance of the detector considerably. Another problem is the temporal information of the subsequent frames is not often used for detection. So the detector search space is the whole image, not temporally constrained by previous frames, causing a high number of false detections. Generally, part-based detectors have satisfactory performance in moderately high resolutions. However, they may fail when the camera is placed too high or too far or at a position from where the parts are not clearly identified in one frame.

There are some problems addressed generally in the tracking solutions. Approaches which are not using human detectors but used to track human in videos are analyzed. Sometimes, to aid the detection process, a combination of features is used which provides more reliability and robustness than a single cue [9, 10, 18]. But if the features or one of the features of the combination is not reliable then the overall performance of tracking goes down. This happens mainly if the feature does not suit the target object. Buehler et al. [11] tried to track the upper body part of human for an hour in a signing video. Issues that are mostly addressed for the combinatorial model is self occlusion of hands and the problem of tracking only hands or head or combination of them which become

unreliable for a long sequence of time mainly because it loses track when the model gets self occluded or position of the limb or angle is changed significantly. Kaaniche et al. [17] chose corner point detector. They have applied HOG for all the corner points which are concatenated to comprise the feature vector and used on them for tracking using Kalman filter. However, a suitable object tracker is needed which will consider the dynamics of the whole object and use the observation from the previous frame.

Several problems are identified in the tracking pipelines which use human detectors. A good number of methods are now using reliable object detector or a group of detectors to have satisfactory detection of pedestrian at the very beginning. Yang et al. [12] have used colour and elliptical head model feature and combine them for detection for providing supervision to the adaptation of the multi cue tracking. Breitenstein et al. [13] have used whole body HOG detector for multi person tracking with particle filter. However, the object's track is lost very often due to the inconsistency of the detector and it has to be restored based on the observation of the previous frame. In summary, tracking by detection demands an efficient detector with two characteristics. Firstly, the feature should be distinguishable from the background, fast and easy to compute. Secondly, the appearance model of the body or the parts should be efficiently clustered to boost the efficient detection. In our approach in accordance with the gradient information of different body poses it considers the combination of different key points of the human body joints.

This paper shows a new way of body pose detection based on pedestrian tracking. As the poselet feature is a very strong distinguishable feature for detecting human and its poses, the idea has been developed to use it for tracking the person over a period of time. The information about the different body parts are often a good key when the tracking faces challenges in difficult environment. The tracking performance is well evaluated by using different metrics to demonstrate how the temporal information is well utilized for pedestrian detection and tracking in every frame of the sequence.

2. Methodology

The proposed tracking process seen in figure 2 consists of four major phases. First, the input video frames are initialized based on the pedestrian location. Particle filter populates the particle set based on this initialization. Secondly, a poselet body part based model which is created from the H3D [2] dataset is been used to locate the probable appearance of pedestrian. Then, for each particle a set of detected poselet bounds, which is a square covering the detected part, are selected. Thirdly, the similarity between each particle and the model is measured by calculating the SVM score [2]. The final observation is a voting probability score which comes from the sum of the multiplication of activated poselets detected within the particle. Particles' weights are calculated from the Max Margin Hough Transform used by Maji et al. [20]. The score is computed based on the Hough votes casted by each cluster for a given activated poselet location. Finally, we combine all the scores to rate the hypothesis and use it as a weight to select particles for the next frame.

Bayesian sequential estimation i.e. particle filter has been a popular method to estimate time evolving posterior distribution of the target object. It offers a framework for representing the tracking uncertainty in a Markovian manner by only considering information from past frames. Therefore, such an approach is more suitable for time-critical, online applications. A particle filter creates the probability density cloud which avoids expensive iterative state space estimation and makes the algorithm considerably fast.

We initialize the particle filter from the known information of the first frame. A Gaussian distribution is used for the particle distribution in the problem space. The problem of determining a posteriori density is in general referred to as the Bayesian approach. The state vector X_t

$$X_t = \{X_{cen}\ Y_{cen}\ w\ h\ s\} \quad (1)$$

is formulated where X and Y represent the object centre position w is width of the bounding box and h is the height where s is the scale of the image. In the proposed method seen in Figure 2, the video captured by a surveillance camera is considered as input of the system.

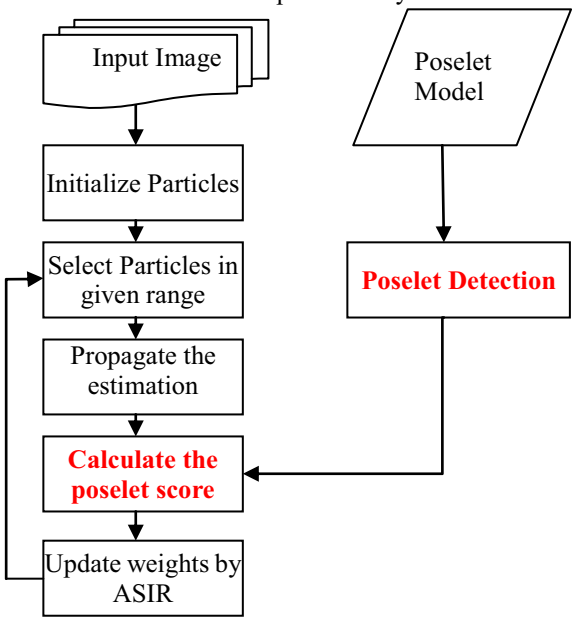


Figure 2: Block diagram of the proposed method

Not using the measurement at the time of resampling makes the tracker inefficient and unreliable. Another variant of the particle filter is Auxiliary Sampling Importance Resampling (ASIR) which was implemented with the same appearance modeling [23]. It is more accurate than SIR when the previous observation is taken into account [22].

In the observation phase, we take the detections of body parts inside a particle. HOG is calculated for each particle and a scanning window technique over the particle bounding box is employed. H3D trained SVM classifier of different body parts and calculated windows are compared for that particle. The products between the blocks of the model and the window are summed up to calculate SVM score. If the score passes a threshold then the corresponding poselet is activated. The outcome of the poselet detector for each particle is a group of bounding boxes selected for different body parts.

For each particle an overall confidence score is calculated by taking the sum between all the activated poselet

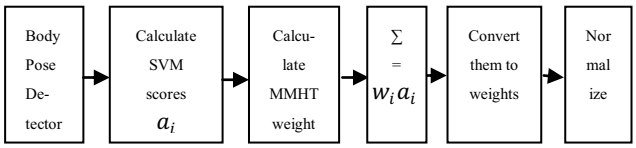


Figure 3: Processing of each particle in PF

score multiplied with the MaxMarginHoughTransform weight shown in figure 3. This activation score a_i in eq. 2 comes from the product of every block between the H3D trained SVM classifier weight and the observed HOG for each poselet. The MMHT weight u_i of the poselet is calculated based on the idea that poselet may have unequal significance in pedestrian detection. After the detection of poselets, the hits are clustered by mean-shift to cast a vote for a specific object location x. Maji et al. 2009 [20] had employed Max-Margin Hough transform to optimize the responses of different activations at different locations by learning the weights minimizing the effects of negative examples and maximizing the positives. The idea is to find a peak in the hough space which shows the probable object location. In figure 4, an example shows how the poselets are activated for a particle.

A B

Figure 4: Poselet convergence example for a particle. In (A) selected particle, in (B) activated poselets

Similarity between each poselet and the pre-trained poselet model is measured as follows.

$$w_k \propto \sum_i u_i a_i(x) \quad (2)$$

$$p(Z_t|X_t) \approx w_k \quad (3)$$

Considering the activation of i^{th} poselet score a_i at location x in the particle with MMHT weight u_i, the probability of the weight for that particle w_k can be found in eq. 2. The weight is normalized and used as measurement likelihood in eq. 3.

In the proposed method using particle filter predicts the location of the pedestrian in the next frame by using the information of the previous location of the same pedestrian. In the scanning window approach the whole image is searched and no temporal information is used. Therefore, in table 1 we often see it fails to detect the pedestrian when any change in appearance occurs. On the other hand, the PF tracker always maintain a good track of a particular target object all over the scene not only by using the poselet feature but also by exploiting the probability of having the detection in the most probable space.

3. Experimental Results and Evaluation

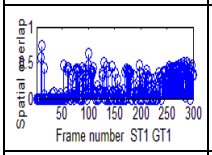
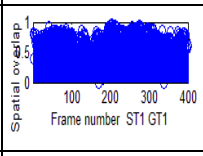
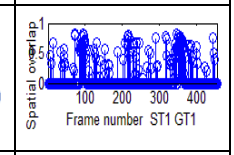
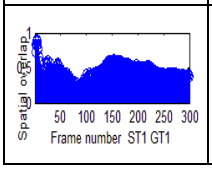
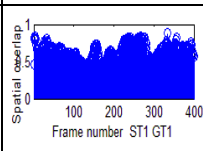
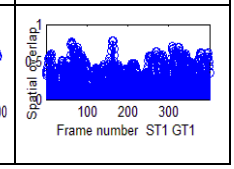
In this section, we present some real-scene object tracking results using the proposed algorithm. The algorithm was implemented in Matlab with C++ and run on a 2.66 GHZ Pentium Core 2 Duo PC with 3 GB memory. The tracking algorithm was initialized with a manually selected region in the first frame. The number of particle was selected as 250 and the standard deviation of Gaussian function in prediction model equal to 45 pixel area.

Frame 10 Angular view | Frame 109 Reflection | Frame 291 Non linear motion | Frame 140 Multimodality

Figure 5: Various tracking problems in different frames in PETS Dataset.

The datasets which are chosen for this project contain a good set of pedestrian images in different poses, background, color and different illumination conditions. For testing the performance of the tracking algorithm three datasets are chosen namely HumanEva, PETS 2007 and Muhavi, for their rich and challenging human motion content and cluttered environment. Every training image of the H3D dataset (person category) is in upright and the camera is placed in more or less chest height position.

Table 1: Spatial overlap and distance error comparison.

PETS	HumanEva	Muhavi
		
		

In the table above, first row is the datasets tested, second row shows the original poselet detection rate per frame and third row is the detection rate using poselet based particle filter.

In the first PETS scene, the camera is mounted on a pole at the top left position of the tracked object. The challenges here in this sequence are: a. pedestrian view is

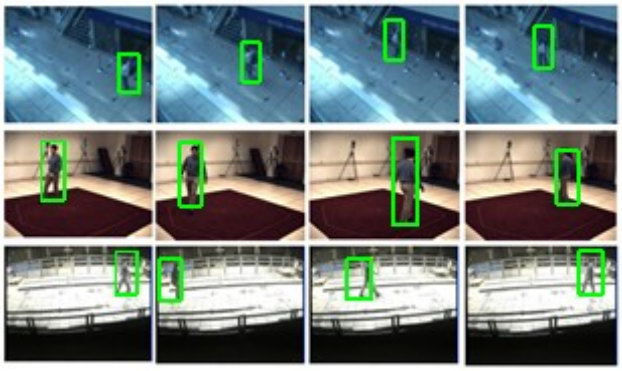


Figure 6: Tracking result in every 50 frames. First row from PETS, second from HumanEva and third row is from Muhavi dataset.

not upright all the time b. The motion is non linear i.e. his speed is quite variable and he stopped couple of times c. Reflection of the pedestrian in mirror d. Interaction with another pedestrian.

It is very important to find an appropriate measurement function for the likelihood between the model and the local feature response, the optimized Max-Margin Hough Transformed activation weight score in this case. In the table below the distribution of particles in different measurement metrics are shown. The color of the columns denotes different datasets.

Table 2: Sample tracking results in different datasets

Dataset / Metric	PETS		HumanEva		Muhavi	
	Tracking by Poselet	PF Tracking by Poselet	Tracking by Poselet	PF Tracking by Poselet	Tracking by Poselet	PF Tracking by Poselet
Average track closeness (%)	0.1792	0.4644	0.7520	0.6382	0.1368	0.4409
Deviation of closeness (%)	0.1795	0.1073	0.1187	0.1054	0.2493	0.1164
Average distance error	153.674	52.59	16.471	25.851	253.119	32.133
Deviation of distance error	131.737	14.51	17.180	11.676	189.377	14.733

To assess the tracking performance the evaluation framework of Yin et al. [24] is used. Specifically, the spatial overlap and distance error metric have been calculated as seen in table 2. Other important metrics which give us an implicit idea about the general performance of the tracker are also obtained using the same framework. Average track closeness tells us how close the tracking is to the actual ground truth. The table above shows that the combination particle filter with poselet performs better than detecting the feature in every frame. This approach is applied on the standard datasets.

4. Conclusion

A poselet based pedestrian tracker has been developed to solve non linear pedestrian tracking problem with different measurement systems. Unlike other methods, here no background model is learned. The challenges of pedestrian detection and tracking were discussed and analyzed. The local feature selection process was investigated and problems and advantages were discussed from practical observation. The necessity of having an offline training of pedestrian pose based body part images is investigated and an algorithm is proposed which improves reliability. The novelty lies in the combination of a discriminative feature for detecting human which is beneficiary to the existing tracking trend and adds significant value. Results have shown that the combined framework of the model and filter is able to track objects undergoing complex deformation of shape with small changes in inter frame object position. However, the method may still be sensitive to background clutter to some extent.

References

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In CVPR, volume 2, June 2005.

[2] Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. ICCV (2009)

[3] Bourdev, Lubomir: Poselets and Their Applications in High-Level Computer Vision. PhD thesis, University of California, 2011.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan: Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep. 2010.

[5] Wang, X., Han, T., Yan, S: An HOG-LBP human detector with partial occlusion handling. Computer Vision, 2009 IEEE 12th International Conference (32 - 39), 2009.

[6] R. Poppe. Evaluating example-based pose estimation: Experiments on the humaneva sets. CVPR 2nd Workshop on EHuM2, 2007.

[7] H. Grabner, P. M. Roth, and H. Bischof: Is pedestrian detection really a hard task? In Proc. IEEE Intern. Workshop on Performance Evaluation of Tracking and Surveillance, pages 1–8, 2007.

[8] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool: Robust tracking-by-detection using a detector confidence particle filter. In ICCV, 2009.

[9] J. Zuo, C. Zhao, Y. Cheng, H. Zhang, Particle filter based visual tracking using new observation model, in: Proceedings of the IEEE International Conference on Automation and Logistics, 2007, pp. 436–440.

[10] B.Han, C.J.Yang, R.Duraiswami and L.Davis: Bayesian filtering and integral image for visual tracking. The International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, Switzerland, 2005.

[11] Buehler, P., Everingham, M., Huttenlocher, D.P. and Zisserman, A.: Upper Body Detection and Tracking in Extended Signing Sequences. In: International Journal of Computer Vision (IJCV), pp. 1–18, Springer (2011).

[12] M. Yang, F. Lv, W. Xu, and Y. Gong:Detection driven adaptive multi-cue integration for multiple human tracking. In Proceedings of the 12th International Conference on Computer Vision, 2009, pp. 1554–1561.

[13] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, ¥Online multi-person tracking-by-detection from a single, uncalibrated camera," IEEE Trans. on Pattern Analysis and Machine Intell. (PAMI), vol. 33, no. 9, 2011.

[14] Wei-Lwun Lu. Tracking and Recognizing Actions of Multiple Hockey Players using the Boosted Particle Filters. Masters thesis., The University of British Columbia, 2007.

[15] W.-Lwun Lu, K. Okuma, and J. J. Little: Tracking and recognizing actions of multiple hockey players using the boosted particle filter. Image and Vision Computing, vol. 27, no. 1-2, pp. 189–205, January 2009.

[16] Tang, F., Brennan, S., Zhao, Q., Tao, H.: Co-tracking using semi-supervised support vector machines. In: Proc. ICCV. (2007) 1-8.

[17] M. B. Kaaniche and F. Bremond: Tracking hog descriptors for gesture recognition. Advanced Video and Signal Based Surveillance, IEEE Conference on, 0:140–145, 2009.

[18] A.P.Li, Z.L.Jing and S.Q.Hu, ”Particle filter based vision tracking with multi-cue adaptive fusion,” Chiness Optics Letters, vol. 3, no. 6, pp. 326- 329, June, 2005.

[19] N.T. Siebel, S.J. Maybank. “Real-Time Tracking of Pedestrians and Vehicles”, Proc PETS 2001.

[20] Subhransu Maji and Jitendra Malik. Object detection using a max-margin hough tranform. In CVPR, 2009.

[21] B. Han and L. Davis. On-line density-based appearance modeling for object tracking. In Proc. ICCV, volume 2, pages 1492–1499, 2005.

[22] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. 50(2):174–188, February 2002.

[23] M. Pitt and N. Shephard, “Filtering via simulation: Auxiliary particle filters,” J. Amer. Statist. Assoc., vol. 94, no. 446, pp. 590–599, 1999.

[24] Fei Yin, D. Makris, Sergio A Velastin, James Orwell “Quantitative evaluation of different aspects of motion trackers under various challenges” Annals of the BMVA, 2010(5), pp. 1-11.