JACET Journal of Advances in Computer Engineering and Technology, 1(2) 2015

Online multiple people tracking-by-detection in crowded scenes

Sahar Rahmatian[1], Reza Safabakhsh[2]

Received (2015-01-23)
Accepted (2015-03-19)

Abstract **– Multiple people detection and tracking is a challenging task in real-world crowded scenes. In this paper, we have presented an online multiple people tracking-by-detection approach with a single camera. We have detected objects with deformable part models and a visual background extractor. In the tracking phase we have used a combination of support vector machine (SVM) person-specific classifiers, similarity scores, the Hungarian algorithm and inter-object occlusion handling. Detections have been used for training person-specific classifiers and to help guide the trackers by computing a similarity score based on them and spatial information and assigning them to the trackers with the Hungarian algorithm. To handle inter-object occlusion we have used explicit occlusion reasoning. The proposed method does not require prior training and does not impose any constraints on environmental conditions. Our evaluation showed that the proposed method outperformed the state of the art approaches by 10% and 15% or achieved comparable performance**

Keywords- detection; tracking; crowded-scenes; online tracking

Index Terms **— detection; tracking; crowded-scenes; online tracking**

I. INTRODUCTION

People detection and tracking is a crucial pre-processing step in a wide range of video analysis applications such as video surveillance, human-computer interaction, robotics, entertainment, intelligent control and so on. Although extensive research has been done in this field, due to several challenges, it is still an active and challenging field of research. The main difficulties and challenges in this field are changes in illumination, lack of variability in different objects clothing, inter-object occlusion and object occlusion caused by other scene objects. It should also be noted that most of the above applications require online detection and tracking which use information from the past and present. Our goal in this paper is to detect and track multiple objects online and improve the existing results.

To overcome the mentioned difficulties we have proposed an online tracking-by-detection approach which uses information from the detection bounding box and trains a person specific classifier which yields useful discriminative information for each target and predicts the targets states using this information, spatial information, and an inter-object occlusion reasoning. We have not imposed any constraints on the environmental conditions such as the targets appearance, background and foreground illumination, the objects pose, etc. Our approach has the following strengths:

- Handling inter-object occlusion
- Handling lack of variation between different targets clothing and appearance
- Identifying targets with no movement
- Identifying targets with different poses

1- Department of Computer Engineering, Amirkabir University of Technology (srahmatian@aut.aci.ir)
2- Department of Computer Engineering, Amirkabir University of Technology (safa@aut.ac.ir)

- Handling short-term occlusion between the moving objects and other scene objects

It also shows the following drawbacks:

- Inability to handle long-term occlusions between the object and other scene objects
- The target has to be present in a certain number of frames to be tracked
- High resolution imagery is required for the detection phase
- If there is no detection associated to the tracker, it can only detect the target for a limited number of frames

This paper is organized as follows: Section II gives a brief overview of the related and state of the art work in the area of pedestrian detection and tracking. Sections III and IV describe the proposed detection and tracking methods respectively. Section IV gives a summary of the detection and tracking algorithm, and describes all the components used in our tracking algorithm. Section V introduces the employed datasets, evaluation methodology and metrics, and the results. Finally, conclusions and future work are given in Section VI.

II. RELATED WORK

Human detection and tracking has been an active research area for decades. A complete review of this field is beyond the scope of this paper. From one aspect we can divide human tracking into two groups: single people tracking and multiple people tracking. Single people tracking can be performed by adaptive visual tracking based on structured output predictions [1]. Another approach for single tracking is by estimating the target location and motion in every frame and building the trajectory through interpolation and based on P-N learning [2]. Tracking multiple people is much more complex and complicated due to the data association problem, interactions between different targets, and lack of knowledge about the number of targets a-priori and its changes over time.

Tracking multiple people can be performed online or offline. Online tracking considers past states and present observations while offline tracking also considers future information. It is obvious that since offline tracking uses more information, it can yield better and more improved results. Offline tracking can be performed by minimizing a continuous energy function [3], or using a discrete-continuous Conditional Random Field (CRF) for multi target tracking that handles inter-object occlusion [4]. In [5], Zamir et al. use a generalized minimum clique problem to solve the formulated optimization problem. The approach in [6] is a type of tracking-by-detection which uses [3] as the baseline tracker and [7] as the baseline detector, and proposes a joint detector by combining the baseline detector with a detector for pairs of people. Also it investigates tracking failure cases and trains a detector to overcome these cases and at last trains the detector with the people tracker in the loop. Li et al [8] learn affinity models by a Hybrid Boost algorithm for tracking multiple targets.

Although offline tracking can yield more robust results, time critical and security applications require online tracking. In [9], Benfold et al. use Histogram of Oriented Gradients (HOG) detections with Kanade-Lucas-Tomasi (KLT) tracking and Markov-Chain Monte-Carlo Data Association (MCDMA) for real time tracking. In [10], online multi person tracking is performed in a particle-based framework in which a combination of final detections, continuous detector confidence and classifier output are used to guide particles. In [11], Kuo et al. use detection responses and learn online discriminative appearance models for tracking multiple targets. In [12], Shu et al. propose an online multi person tracking-by-detection method which is based on the Deformable Part Model (DPM) detector [7] and a tracking method based on dynamic occlusion handling and Support Vector Machine (SVM) detectors for each tracker. Although all the above methods achieve relatively good results, they all have weaknesses and can be improved. For example, [11] suffers when faced with appearance changes and occlusion and [9] only uses head detections which is not always appropriate and sufficient.

III. DETECTION

The detection we have used in this paper is a combination of the DPM detector and a background subtraction algorithm called VIsual Background Extractor (ViBe) [13]. First we briefly explain the above algorithms and then we introduce the proposed detection approach.

A. DPM

The DPM detector is based on a set of star-

structured part-based models. Each model is a combination of a "root" filter, a set of part filters, and deformation models associated to them. The part filters can adjust their positions with respect to the root filter in order to capture these possible deformation models. The detection score of the model is the sum of the root filter response, part filter responses, and a deformation cost that measures the difference of the part filters from their ideal positions relative to the root. To be more precise, the detection score for a detection hypothesis h=(p_0,…p_n), where p_ i=(x_i,y_i,l_i) is the ith part which is specified by its position and level, is computed as:

$$\text{score}(h) = \sum_{i=0}^{n} F_i' . \emptyset(H, p_i) - \sum_{i=1}^{n} d_i . \emptyset_d(dx_i, dy_i) + b \quad (1)$$

where F`*i* is the score of the *i*th part filter, *H* is the HOG feature pyramid, $\emptyset(H,Pi)$ is the vector obtained by concatenating the feature vectors from *H* at the subwindow of part p_i , d_i is a four dimensional vector which specifies the coefficients of the deformation features, (dx_i , dy_i) is the displacement of part *i* relative to its anchor position and $\emptyset_d$ (dx_i , dy_i) are deformation features [7].

B. ViBe

The ViBe background subtraction algorithm combines different methods for motion detection. This technique is seen as a classification problem, in which each new pixel value is classified with respect to its neighborhood in the specified color space. The technique stores a set of values taken in the past at the same location of the pixel or its neighborhood. In the next step, this set is compared to the current value of the pixel and it is determined if this pixel belongs to the background or not. A difference between this algorithm and other existing algorithms is that a new value when compared to background samples should be close to some of the samples and not the majority of them. This difference makes the approach more reliable since it is better to estimate the background pixels with a small number of close values rather than just a large number of values.

This method uses the first frame to initialize the background model. The next step for this algorithm is updating the model. A conservative update policy is used, meaning that a foreground pixel is included in the background model only if it is classified as a background sample. This policy gives sharp detections, but when a background sample is incorrectly classified as foreground it prevents the background pixel model to be updated and creates a deadlock. Therefore, a solution is needed to update background pixel models which are covered with foreground pixels. A new background sample of a pixel should also update the models of its neighboring pixels to solve this problem [13].

C. Combining DPM and ViBe

First we have used the pre-trained DPM detector with specific and fixed settings to give us the initial detections. The detection obtained from this detector includes false positives due to structures that resemble the human body, and false negatives due to occlusion, etc. To reduce false positives, first we remove bounding boxes which contain a specific percentage of another bounding box. In other words, we remove detections which are falsely detected as overlapping detections. For the second step, we obtain the ViBe results and compare them to the detections. If the bounding box contains more foreground pixels than a pre-defined threshold, we consider it as a valid detection; if not we consider it as a false positive and remove it. For the last step, bounding boxes with overlap are considered such that one is the base detection and all the other overlapping boxes are considered as background pixels, if the base detection has more foreground pixels than the defined threshold, we consider it as a true positive; otherwise, we remove it from the valid detections. Fig. 1 shows the results of each step on frames of the TUD-Stadmitte dataset.

IV. MULTI TARGET TRACKING

In this section we have described the different components of our tracking algorithm. Fig. 2 shows the summarized algorithm.

A. Person-specific classifiers

Similar to [12] we have trained online person classifiers for each target. For each frame, the detections are classified by the classifiers. We extract features from the bounding boxes and train with a linear SVM classifier. There are different types of features which can be used to train the classifier.

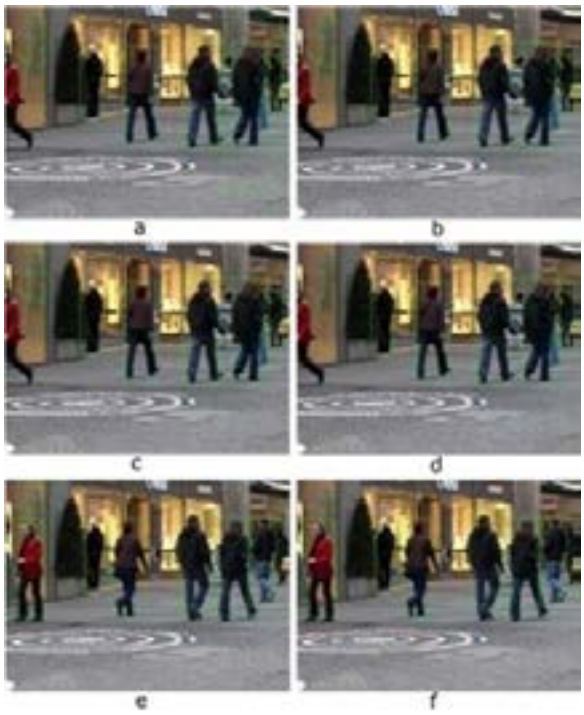

Figure 1. Results of the different stages of our detection approach. (a) The result of the DPM detector on frame 7043, (b) The result of stage 1 applied on (a), which shows that the overlapping detection has been removed. The result of stage 2 which is performed on frame (c) is shown in (d), it can be seen that the falsely detected background has been removed. Frame 7035 after stage 2 is shown in (e), and (f) is the result of stage 3 on (e) which we can see the falsely detected object has been successfully removed.

Input: New Image Frame at time t and existing trackers: $\{tr_k\}_{k=1}^{K}$
Perform Human detection: $\{d_n\}_{n=1}^{N}$
for *all trackers* **do**:
  for *all detections* **do**:
    Compute person specific classification scores between the existing tracker and detection $S(d_n, tr_k)$.
    Compute similarity score between detection and existing tracker using equation (2).
  end
end
Solve the Hungarian algorithm to determine the detection assignments to the existing trackers.
Check if there is a new target.
for *all targets* **do**:
  if *new target* **then**
    Initialize a new tracker
    Increase total number of trackers: $K = K + 1$
  else
    Predict trackers velocity: $v_{k,pred}$
    Predict the trackers state using equation (4)
    Update the trackers state using equation (3)
  end
  Perform occlusion reasoning and save trackers model.
  Update person-specific classifiers.
end
Check if any of targets need to be deleted
Output: updated trackers: $\{tr_k\}_{k=1}^{K}$

Figure 1. Algorithm summary

Based on [14] these features are: gradient histogram, gradients, grayscale, color, texture, self-similarity and motion. Among these features the HOG features have shown the best performance. The HOG feature is used in the detection step, so it is better to use a different feature to compliment this feature in tracking. We have used the HSV (Hue, Saturation, and Value) color space histogram of the image as features for the classifier.

The detections in the trajectory of the current tracker are used as the trackers positive examples and detections in the trajectory of other trackers are used as negative examples.

B. Data Association

Data association in this paper refers to matching the detection responses to existing human trackers. Since we are tracking multiple people online and we use only the information from the past and present, our data association uses Markov models to build the trajectories based on the observations. For this task, we have used the Hungarian algorithm [15] which is a well-known algorithm for solving the weighted bipartite matching. The Hungarian algorithm solves the assignment problem by assigning the present detections to the existing trajectories in a way that the sum of the similarity scores is maximized. Unlike the greedy algorithm which picks the highest scoring pair in each step, this algorithm solves the problem so that the sum of all the picked pairs is maximized.

In time t, if we suppose to have N detection responses $(d_1, d_2, \ldots, d_N)$ and K trackers $(t_1, t_2, \ldots, t_K)$, we compute an $n \times k$ similarity matrix S whose rows are the detections and its columns are the trackers.

To compute the similarity score between the detection n and the tracker k, we use the following equation based on the svm classifier score, spatial proximity, overlap and predicted state,

$$S(d_n, tr_k) = A_{svm}(d_n, tr_k) \times A_{pos}(d_n, tr_k) \times A_{area}(d_n, tr_k) \times A_{pred}(d_n, tr_k) \quad (2)$$

The term A_{svm} is the output of the person-specific classifier trained for target k. A_{pos} is the Euclidean distance between the center of the detection n and tracker k. A_{area} is the overlap area between detection n and tracker k.A_{pred} is the overlapping area between tracker k and the position of the tracker if the detection n is selected. All the terms are normalized between 0 and 1; thus the similarity score is a number between 0 and 1. Similarity scores which are below a defined threshold are assumed to be zero. Matching pairs are selected from the similarity

matrix with the Hungarian algorithm. This does not mean that there is always a pair for each detection or tracker; there might be a detection to which no tracker is assigned to and vice versa.

C. Tracker initialization and termination

If for T consecutive frames we have a detection of the same person which is not assigned to any of the existing trackers, we initialize a new tracker. To see if the detections belong to the same person, we compute a similarity score between them and if it is above a defined threshold, we assume the detections belong to the same person.

A tracker is terminated in two cases: The first case uses the same strategy used in initialization; if for T consecutive frames a tracker is lost, we terminate the tracker. In the second case, we compare the tracker with the vibe result; if the tracker contains background pixels above the defined threshold, we terminate it.

D. Tracker Updating

The tracker is updated based on two factors: the trackers predicted velocity and the detection assigned to it, as follow:

$$tr_{k,t} = \left(1 - S(d_{assign}, tr_k)\right) \times tr_{k,pred} + S(d_{assign}, tr_k) \times d_{assign} \quad (3)$$

where $S(d_{assign},tr_k)$ is the similarity score between tracker k and the detection assigned to it (if there is no assigned detection this score will be zero), d_{assign} is the assigned detections state, $tr_{k,t}$ is tracker k's updated state at time t. Now, $tr_{k,pred}$ is computed as:

$$tr_{k,pred} = tr_{k,t-1} + v_{k,pred} \quad (4)$$

The predicted velocity is based on the trackers position in the F previous frames. In other words, the trackers velocity is the average velocity computed in the F previous frames. The second factor of (3) might exist and might not. If no detection is assigned, the updated state only depends on the predicted state which is only based on the predicted velocity.

E. Inter-object occlusion handling

We have tried to detect inter-object occlusion by computing the overlap between our tracker and the other existing trackers. When a new tracker is initialized, we save its image in that frame as its model. In each frame, we update this model. If the tracker does not have overlap with the other existing trackers, the new tracker's image is replaced as the new model. If the tracker has overlap with other trackers, the occluded area is replaced with the corresponding area of the saved model. Consequently, the new tracker's image which has been reconstructed is saved as the new model, and is used in retraining the person-specific classifier.

V. EVALUATION AND EXPERIMENT RESULTS

We have evaluated our proposed method with three publicly available datasets; the Parking Lot dataset [12], Pets S2.L1 dataset [16], and the TUD-Stadmitte dataset [17]. All the datasets include semi-crowded to crowded scenes, occlusions and are all outdoor scenes captured with static cameras. We have not used information from the camera, ground plane or obstacles in the image.

For the DPM detector, we have used voc-release 4.01, and we have set the detection score threshold and the NMS overlap to -0.6 and 0.4, respectively, for all three datasets. Since this detector requires high quality imagery for this part, we have upsampled the frames for the Pets S2.L1 and TUD-Stadtmitte datasets. For the background subtraction the radius R has been set and fixed to 20, the threshold T to 2, and N has been set to 20.

In the implementation of tracking, we have used 125 bin HSV color histograms as feature vectors for the person-specific classifiers. The results have shown better performance for the histogram-based classifiers in HSV color space than in RGB color space. The training data for each person-specific classifier consists of up to 100 positive samples and 100 negative samples. When the number of collected samples exceeds this limit, we delete the oldest ones to ensure the model is up to date.

A. Datasets

The parking lot sequence is a modestly crowded scene including groups of pedestrians walking in queues. The challenges in this dataset include long-term inter-object occlusion, and similarity of appearance among the people in the scene. This sequence consists of 1,000 frames of a relatively crowded scene with up to 14

pedestrians. This dataset's frame resolution is 1920 ×1080, and the frame rate is 29 fps. The TUD-Stadmitte dataset contains 200 consecutive frames taken in a typical pedestrian area. This sequence has a low camera angle and frequent occlusions. The frame resolution is 640×480 and frame rate of 13-14 fps. The Pets S2.L1 dataset is filmed from an elevated viewpoint and is 795 frames long, showing up to 8 people. This dataset's frame resolution is 768×576 and the frame rate is 7 fps.

B. Metrics

We evaluate our tracking results using the standard CLEAR MOT (Multiple Object Tracking) metrics [16,18], MOTP (multiple object tracking precision), MOTA (multiple object tracking accuracy), MODP (multiple object detection precision), MODA (multiple object detection accuracy), precision and recall. TP measures the precision of true positive tracked object positions, while TA considers false negatives, false positives, and ID-switches. Therefore, TA is a more effective factor in people detection and tracking and is our main focus in this paper. Recall measures the number of correctly marched detections divided by the total number of detections in the ground truth. Precision measures the number of correctly marched detections divided by the number of output detections.

For all datasets, we have considered the detection a true positive if it has at least 50% overlap with the ground truth. In computing MOTA and MODP, we have assumed c_m=1, c_f=1 and $c_s=log_{10}$ *ID-SWITCH*.

C. Results

The detection results for the Parking Lot, TUD-Stadmitte and Pets datasets are shown in Table 1, Table 2, and Table 3, respectively. The results have been improved for all three datasets, which is predictable and reasonable. The percentage of improvement is different for the three datasets. This is due to the nature of the datasets and also gives us information about the videos. For the parking lot dataset, the first stage has 2.84% improvement which shows that we have a small number of false positives due to falsely detecting two objects instead of one object. Stage 2 and stage 3 have a small improvement. For the TUD-Stadmitte dataset, we have a relatively large improvement in all three stages. We can conclude that the DPM detection has a large number of false positives due to falsely detecting two objects instead of one object, and falsely detecting the background as an object. This means the background is similar to the objects. The Pets dataset percentages are similar to the Parking lot dataset, meaning stage 1 has a high improvement in comparison to the other two stages. Table 1 has a 5th column which compares our detection results with the results in [12]. We can see that our algorithm has improved the accuracy, but the precision has decreased due to the fact that [12] considers an occlusion model and their detection is based on part models which is more precise. It should be noted that [12] focuses on improving the false negative rate by detecting humans which have not been detected due to occlusion and only their head or upper body is visible, while our method improves the false positives by identifying the falsely detected humans which have similar features to the background.

The tracking results are shown in Table 4. Our method has improved the accuracy in comparison with [12] on the Parking Lot dataset. MOTP and MOTA are significantly increased in comparison to [3] and [4] on the TUD-Stadmitte dataset. This shows that our method has the state of the art performance in tracking. Our method does not have improvement on the Pets dataset in comparison to [3]; but our results are comparable and the algorithm has the capability to achieve better results.

Table 1.Detection results on the Parking Lot dataset

Parking Lot	DPM[7]	Stage1	Stage2	Stage3	Shu et al.[12]
MODP	71.56	71.58	71.58	71.6	74.4
MODA	77.35	80.19	80.36	81.74	79.8

Table 2. Detection results on the TUD-Stadmitte dataset

TUD-Stadmitte	DPM	Stage1	Stage2	Stage3
MODP	76.16	76.11	76.99	76.8
MODA	48.53	55.96	76.04	79.07

Table 3. Detection results on the Pets dataset

Pets	DPM	Stage1	Stage2	Stage3
MODP	75.2	75.2	75.2	75.25
MODA	55.97	79.26	80.09	83.95

Table 4. Comparison of tracking results

Method	MOTP	MOTA	Precision	Recall
Parking Lot	72.63	86.91	94.22	92.65
Parking Lot[12]	73.77	77.1	-	-
TUD-Stadmitte	75.65	74.65	97.37	76.82
TUD-Stadmitte[3]	65.8	60.5	-	-
TUD-Stadmitte[4]	61.6	56.2	-	-
Pets	73.6	80	93.12	87.39
Pets[3]	76.1	81.4	-	-

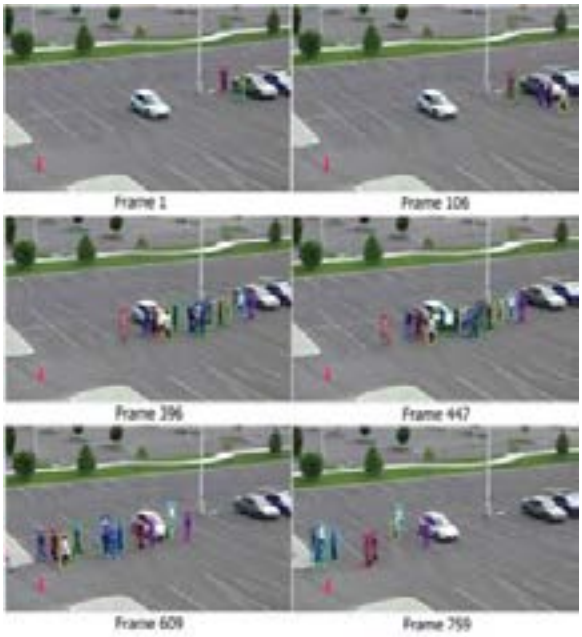

Figure 3- Selected frames of the tracking results from the Parking lot dataset

Figure 4- Selected frames of the tracking results from the TUD-Stadmitte dataset

We have used fixed parameters for all three datasets and since the frame rate is lower for the Pets dataset in comparison to the other two datasets; we should set the parameters so that they would be suitable for this dataset. Also our method shows that it can achieve better results for videos which have frequent inter-object occlusion since its main strength is explicit inter-object occlusion reasoning. Fig 3. and Fig 4.show tracking results on selective frames from the Parking Lot and TUD-Stadmitte datasets.

VI. CONCLUSIONS

In this paper, we have presented a multi target tracking algorithm which is based on detection. Our detection approach is a DPM based approach which removes false positives by using a background subtraction algorithm and a heuristic for identifying falsely overlapping humans and removing them. We have trained a person specific classifier based on HSV histogram features, and velocity prediction for determining the targets next state. Explicit occlusion reasoning has been employed for detecting inter-object occlusion and improving the results. Our results show that we have outperformed or achieved comparable performance to the state of the art approaches. For future work, we plan to use information from super-pixel segmentation of the object instead of information from the object bounding box. Also we can achieve better results of we employ a parameter learning algorithm.

References

[1] Hare, Sam, Amir Saffari, and Philip HS Torr. “Struck: Structured output tracking with kernels.” In Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 263-270. IEEE, 2011.

[2] Kalal, Zdenek, Jiri Matas, and Krystian Mikolajczyk. “Pn learning: Bootstrapping binary classifiers by structural constraints.” In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 49-56. IEEE, 2010.

[3] Andriyenko, Anton, and Konrad Schindler. “Multi-target tracking by continuous energy minimization.” In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 1265-1272. IEEE, 2011.

[4] Milan, Anton, Konrad Schindler, and Stefan Roth. “Detection-and trajectory-level exclusion in multiple object tracking.” In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 3682-3689. IEEE, 2013.

[5] Zamir, Amir Roshan, Afshin Dehghan, and Mubarak

Shah. “Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs.” InComputer Vision–ECCV 2012, pp. 343-356. Springer Berlin Heidelberg, 2012.

[6] Tang, Siyu, Mykhaylo Andriluka, Anton Milan, Konrad Schindler, Stefan Roth, and Bernt Schiele. “Learning people detectors for tracking in crowded scenes.”ICCV’13 (2013).

[7] Felzenszwalb, Pedro F., Ross B. Girshick, David McAllester, and Deva Ramanan. “Object detection with discriminatively trained part-based models.”Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, no. 9 (2010): 1627-1645.

[8] Li, Yuan, Chang Huang, and Ram Nevatia. “Learning to associate: Hybridboosted multi-target tracker for crowded scene.” In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 2953-2960. IEEE, 2009.

[9] Benfold, Ben, and Ian Reid. “Stable multi-target tracking in real-time surveillance video.” In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3457-3464. IEEE, 2011.

[10] Breitenstein, Michael D., Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. “Robust tracking-by-detection using a detector confidence particle filter.” In Computer Vision, 2009 IEEE 12th International Conference on, pp. 1515-1522. IEEE, 2009.

[11] Kuo, Cheng-Hao, Chang Huang, and Ram Nevatia. “Multi-target tracking by on-line learned discriminative appearance models.” In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 685-692. IEEE, 2010.

[12] Shu, Guang, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. “Part-based multiple-person tracking with partial occlusion handling.” InComputer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1815-1821. IEEE, 2012.

[13] Barnich, Olivier, and Marc Van Droogenbroeck. “ViBe: A universal background subtraction algorithm for video sequences.” Image Processing, IEEE Transactions on 20, no. 6 (2011): 1709-1724.

[14] Dollar, Piotr, Christian Wojek, Bernt Schiele, and Pietro Perona. “Pedestrian detection: An evaluation of the state of the art.” Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, no. 4 (2012): 743-761

[15] Kuhn, Harold W. “The Hungarian method for the assignment problem.” Naval research logistics quarterly 2, no. 1-2 (1955): 83-97.

[16] Ellis, Anna, Ali Shahrokni, and James Michael Ferryman. “Pets2009 and winter-pets 2009 results: A combined evaluation.” In Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on, pp. 1-8. IEEE, 2009.

[17] Andriluka, Mykhaylo, Stefan Roth, and Bernt Schiele. “Monocular 3d pose estimation and tracking by detection.” In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 623-630. IEEE, 2010.

[18] Kasturi, Rangachar, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. “Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol.”Pattern Analysis and Machine Intelligence, IEEE Transactions on 31, no. 2 (2009): 319-336.