MJEN

MANAS Journal of Engineering

Volume 6 (Issue 1) (2018) Pages 45-55

Gelecek elimizde...

Outdoor People Tracking

Venera Adanova*[1], Maksat Atagoziev[2]

[1]Department of Computer Engineering, Middle East Technical University, 06800 Ankara
[2]Department of Electrical and Electronics Engineering, Middle East Technical University, 06800
Ankara, venera@ceng.metu,edu.tr ; maks.atagoziyev@gmail.com

Received: 03-12-2017; Accepted: 19-03-2018

Abstract: *We address the problem of automatically detecting and tracking outdoor people. We propose to combine background subtraction based tracking with tracking-by-detection approach in a Kalman filtering framework. For the detection purpose, we use HOG detector, which detects people only in those regions suggested by background subtraction model. Since detectors may give false positives or miss targets, background subtraction is used to complement the detector outcomes.*

Keywords: *people detection, people tracking, tracking-by-detection, surveillance, Kalman filtering.*

Алгоритм отслеживания путем наружного наблюдения

Абстракт: *В данной работе была рассмотрена проблема автоматического обнаружения и отслеживания людей в открытой местности. Мы предлагаем использовать комбинацию метода отслеживания путем отсечения фоновых изображений и метода отслеживания и обнаружения в рамках фильтрации Калмана. Для обнаружения людей на видео, мы используем детектор HOG, который обнаруживает людей только в тех регионах, которые предлагаются моделью отсечения фона. Так как при работе детекторов могут происходить ложные срабатывания или ситуации, когда цель пропущена, фоновое отсечение используется для дополнения результатов детектора*

Ключевые слова: *Обнаружение людей, отслеживание людей, отслеживание по обнаружению, наблюдение, фильтрация Калмана.*

* Corresponding Author.

1. INTRODUCTION

The efficient online people detection and tracking is a challenging and important task in computer vision. It has many applications in surveillance, traffic safety and sport analysis. The task is challenging because it requires for data streams to be processed in real-time, which would reduce the manual effort that is still applied for video analysis. The success lies in accurate detection and tracking of moving objects and on the relationship of their trajectories to the background. However, this is not a trivial task because there are many sources of uncertainty for the object locations, such as measurement of noise, occlusions, and changing background. Moreover, since tracking is modeled such that object locations at time step *t* are predicted from those predicted at time step *(t-1)*, one wrong prediction propagates to subsequent frames leading to wrong results.

Many approaches rely on *background modeling*-based trackers [1, 2, 3, 4, 5], where moving object detection requires good background estimation. In those approaches, background is a focus of attention and good tracking results are obtained via sophisticated algorithms for background modeling. Recently, *tracking-by-detection* approaches [6, 7, 8] have become popular. This is because of a latest progress made in object detection research [9, 10, 11], which enables fairly good object detection even in complex outdoor scenes.

The detection stage is followed by the tracking stage. Generally tracking relies on Markovian assumption, where the current state depends on the previous step. Classic tracking approaches are Extended Kalman Filtering [2], Particle Filtering [6, 7] and Mean-Shift tracking [12]. In [13] a multi-object tracking by coupling object detection and tracking in a non-Markovian hypothesis selection framework is presented.

This work is a combination of tracking-by-detection approach and background modeling-based approach. In contrast to works that completely rely on the object detector results, our approach aims to close the gaps introduced by the detector. Background modeling is used to show the detector possible locations of foreground objects, hence, eliminate misses introduced by detector.

The paper is structured as follows. Section 2 describes our approach. Section 3 provides illustrative experimental results. Finally, Section 4 gives a brief conclusion.

2. PEOPLE TRACKING MODEL

Our people-tracking model consists of detection and data association between two consecutive frames. The detection process is specific to each frame, meaning that detection in each frame does not use temporal information. Once a person is detected in one frame, it is associated with the person detected from the previous frame via tracking process.

People Detection

The first step in tracking people is detecting people in each frame. Detection of people in video streams is known to be a significant and difficult research problem. In most known applications, segmenting means segmenting video streams into moving and background components. Detecting moving blobs provides a focus of attention for the subsequent tracking process. Two common detection methods are frame differencing and background subtraction. Frame differencing method takes a difference between two consecutive frames, and is very adaptive to dynamic environment. However, it does a poor job in extracting all relevant feature pixels, e.g., it cannot detect objects that stop moving or moving towards or away from camera. The background subtraction method estimates the background from several frames and subtracts the foreground objects from background. It provides the most complete feature data, but is extremely sensitive to dynamic scene changes due to lighting and extraneous events. Note that background subtraction method also needs to update the background estimation from time to time, since some objects that were considered stationary may start to move.

Needless to say that both of the methods are not able to classify moving objects to people and other possible objects. People that stay very close to each other or a person standing in front or attached to some object are not separated from each other.

In order to cope with problems mentioned above we decided to use a sliding-window based detector, namely Histogram of Oriented Gradients (HOG) [9]. Sliding the detection window over the entire image, however, takes a lot of time, hence, is not applicable for real-time applications. Since most part of the images generally consists of background objects, the detection can be performed only on the objects other than the background. To find possible foreground objects we decided to use background subtraction to estimate the background of an image from L first frames. Once the background is estimated we can estimate possible foreground object regions and apply HOG detection only on those regions. Keep in mind that the background estimation is done only once at the beginning of a video-stream and is never updated again. We use the background subtraction model used in [1].

A pixel-wise median filter with L frame length is employed to build the background. The background model for pixel $x_t\,(m, n)$ using a length L median filter is

$$x_t = median_L(x_1(m,n), \dots, x_L(m,n))$$

This retains the stationary pixels in the background. A median filter can estimate the background even when the objects are moving, but it usually requires a large amount of memory to save L frames at a time. That is why it is applied only once.

As the background is estimated, foreground extraction is done by background subtraction. It is applied to each incoming frame. Foreground objects are subtracted from background in each RGB color channel and only maximum absolute values are taken as shown below.

$$Diff_c = max\{|R_f - R_b|, |G_f - G_b|, |B_f - B_b|\}$$

Maximum absolute values are also calculated using edge density values, *Dens*:

$$Diff_g = max\{|Dens_{Rf} - Dens_{Rb}|, |Dens_{Gf} - Dens_{Gb}|, |Dens_{Bf} - Dens_{Bb}|\}$$

where the edge density *Dens* is

$$Dens(m,n) = \frac{1}{(2w+1)^2} \sum_{i=m-w}^{i=m+w} \sum_{j=n-w}^{j=n+w} |G_x(i,j) + G_y(i,j)|$$

Here, $G_x(i,j)$ and $G_y(i,j)$ are the horizontal and vertical edge magnitudes and $2w+1$ is the size of an average window filter. Foreground pixels are obtained as following, where Th_1 and Th_2 are determined empirically.

$$F(x,y) = 1 \; if \; Diff_c > Th_1 \; or \; Diff_g > Th_2$$

$$F(x,y) = 0 \; otherwise$$

To combine color and gradient information for background modeling has also been proposed by Shah et al. [5]. The reason is that the gradients are less sensitive to changes in illumination, providing illumination invariance to some degree.

The HOG detector scans only those regions that are considered as foreground regions. Figure 1 shows detection results of both background subtraction and HOG detector. The first column is the result of the background subtraction method, whereas, the second column is the result of the HOG detector, which is applied after the background subtraction method retrieves possible foreground regions. Each row depicts a different scene. In (a), the man standing in front of a car is not detected by the background subtraction method, but going further one more level and applying HOG detector over already detected foreground regions separates a man from the car. In (b), the two men due to perspective transformation, appearing as if they are on top of each other are considered to be one blob by the background subtraction method, while HOG detector is able to separate them.

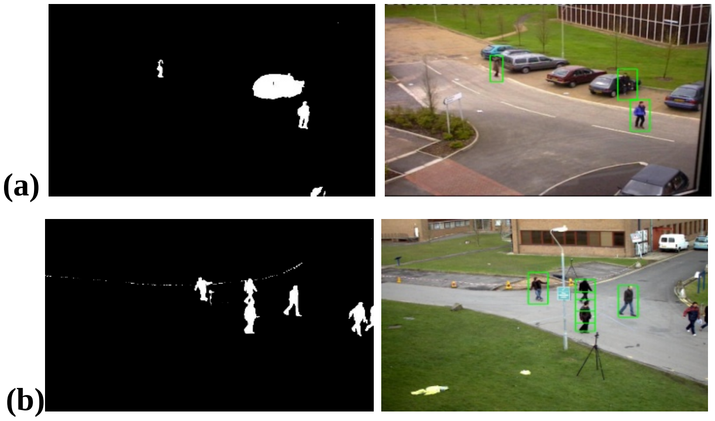


Figure 1. Results of background subtraction (first column) and HOG detector applied on moving regions (second column).

2. OBJECT TRACKING

HOG detector gives incomplete detection list introducing missing detections. A person detected in one frame is not necessarily detected in the next frame. The tracking step must not only match the detections from two consecutive frames but it must also try to close the gaps introduced by HOG detector. Objects detected by HOG detector are tracked over time by matching them between frames of the video sequence. We use the extended notion of Kalman filter used in [2] which contains a list of multiple hypotheses to handle cases where there is a matching ambiguity between multiple objects. We use both HOG detector results and foreground objects obtained by background subtraction. For the sake of clarity, we will call them detections and blobs respectively.

Each object in each frame is represented by several arguments: 1) p = position in image coordinates; 2) δp = position uncertainty; 3) $\vec{v}$ = image velocity; 4) $\delta\vec{v}$ = uncertainty in velocity; 5) Object bounding box in image coordinates; 6) image intensity template.

For each new frame, the possible future position of each object that was obtained in previous frame is predicted as following.

$$p_{n+1} = p_n + \hat{v}_n \Delta t$$

$$\delta p_{n+1} = \delta p_n + \delta \hat{v}_n \Delta t$$

Predicted position and position uncertainty help to decrease the number of candidate blobs that are considered in the matching stage. This is done by moving the bounding box of the object to predicted position and enlarging it by the predicted position uncertainty. Blobs whose centers lie within this bounding box are candidate blobs. The object scans trough candidate blobs and finds the best match via correlation. The correlation function is given below.

$$C(d) = \sum_{x \in R} \frac{W(i,j)|I_n(x) - I_{n+1}(x+d)|}{\|W\|}$$

where I is pixel intensity at the current position and d is a potential object displacement. The weighting function W is calculated as following, where $r(x)$ is the radial distance from x to the center of the object region and r_{max} is the largest radial distance.

$$W(x) = \frac{1}{2} + \frac{1}{2}(1 - \frac{r(x)}{r_{max}})$$

$\|W\|$ is a normalization and given by

$$\|W\| = \sum_{x \in R} W(x)$$

The displacement of an object from previous frame in current frame is the argmin or the correlation function.

$$\hat{d} = min_d C(d)$$

Once we know the best displacement a new position and velocity estimate of an object can be obtained by

$$p_{n+1} = p_n + d$$

$$\hat{v}_{n+1} = \frac{d}{\Delta t}$$

The new velocity and velocity uncertainty estimates are filtered through IIR filter to provide the final velocity estimate of an object for the following frame.

$$\vec{v}_{n+1} = \alpha \hat{v}_{n+1} + (1-\alpha)\vec{v}_n$$

$$\delta \vec{v}_{n+1} = \alpha |\vec{v}_{n+1} - \hat{v}_{n+1}| + (1-\alpha)\vec{v}_n$$

There are several cases that may arise during tracking:

- An object does not match any blob: Either the object has left the scene or has been occluded. The solution would be to keep the track of an object in background for *T* frames. If no match is found after *T* frames, delete the object.
- An object matches exactly one blob: Best case. All parameters of an object are updated to those of a blob.

- Multiple objects match a single blob: This might occur if two objects occlude each other or merge. In this case the intensity of an object is not updated.
- Once all the objects from previous frame find their matches in the current one, the new detections obtained by HOG detector are considered. A new object is created only if it is detected by HOG detector and does not coincide with any existing objects.

3. EXPERIMENTAL RESULTS

Dataset

There is no publicly accepted benchmark available for multi-person tracking. Nevertheless, there are several commonly used sequences, such as TUD Campus, PETS'09 S2.L2-S2.L3, and PETS2000. We will use such sequences in our experiments.

Detector

For all experiments, HOG detector that scans only those regions that are considered to be moving according to background subtraction is used. It is trained for INRIA Person Dataset by Matlab linear SVM classifier. According to the frame size and camera position the detection window size changes but all of them are resized to 64 × 128 before calculating the HOG descriptor. Fig.2 shows estimated backgrounds of three sequences. For S2.L2 and PETS2000 we used L = 40 frames to build the background. Since TUD Campus sequence is crowded the number of frames is L = 60. We can see from the third column in Figure 2 that even with an increased number of frames, some parts of foreground objects still remain on the background.

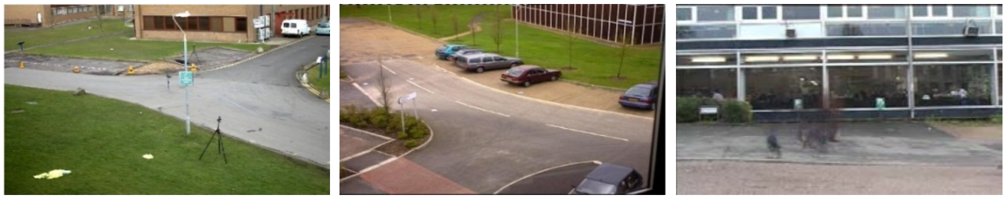
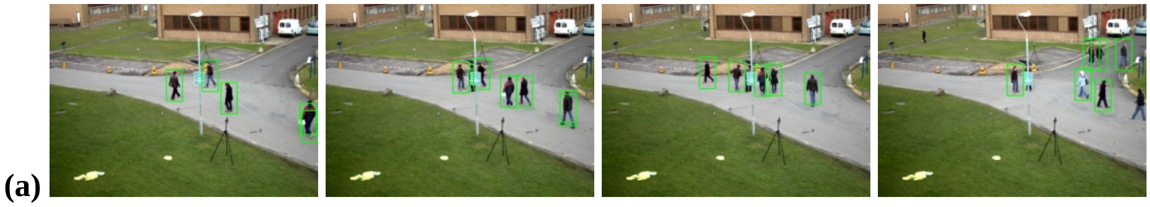


Figure 2. Estimated backgrounds of S2.L2, PETS2000 and TUD Campus.

Figure 3 demonstrates some HOG detector results. Since in S2.L2 no objects other than people exist, the detection gives good results as shown in Figure 3 (a). In PETS2000, however, there are moving cars, introducing several false positives during detection. We believe that by estimating possible positions of foreground objects we are decreasing the possibility of HOG detector to give false positives. This is like directing the detector where to look.



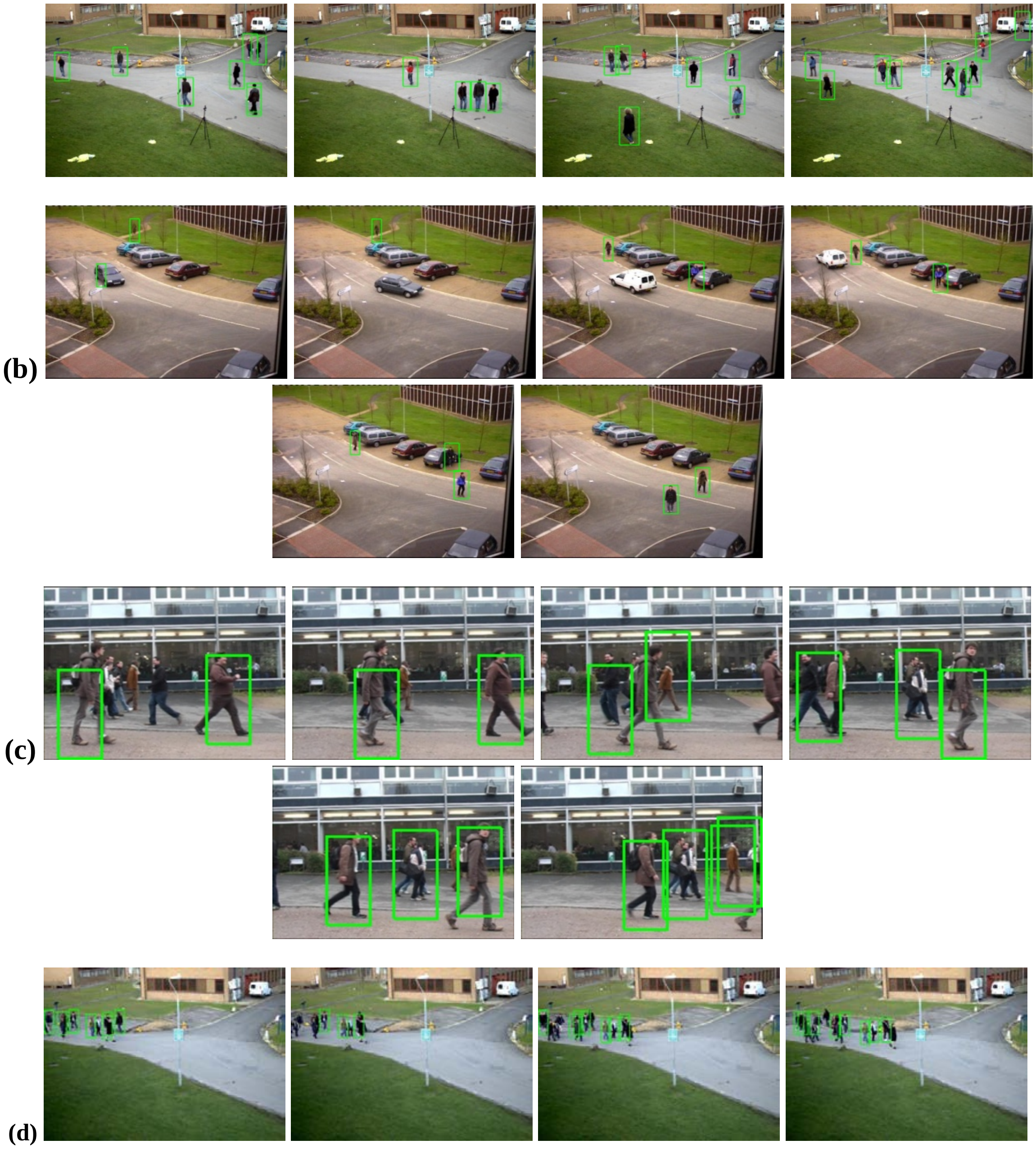
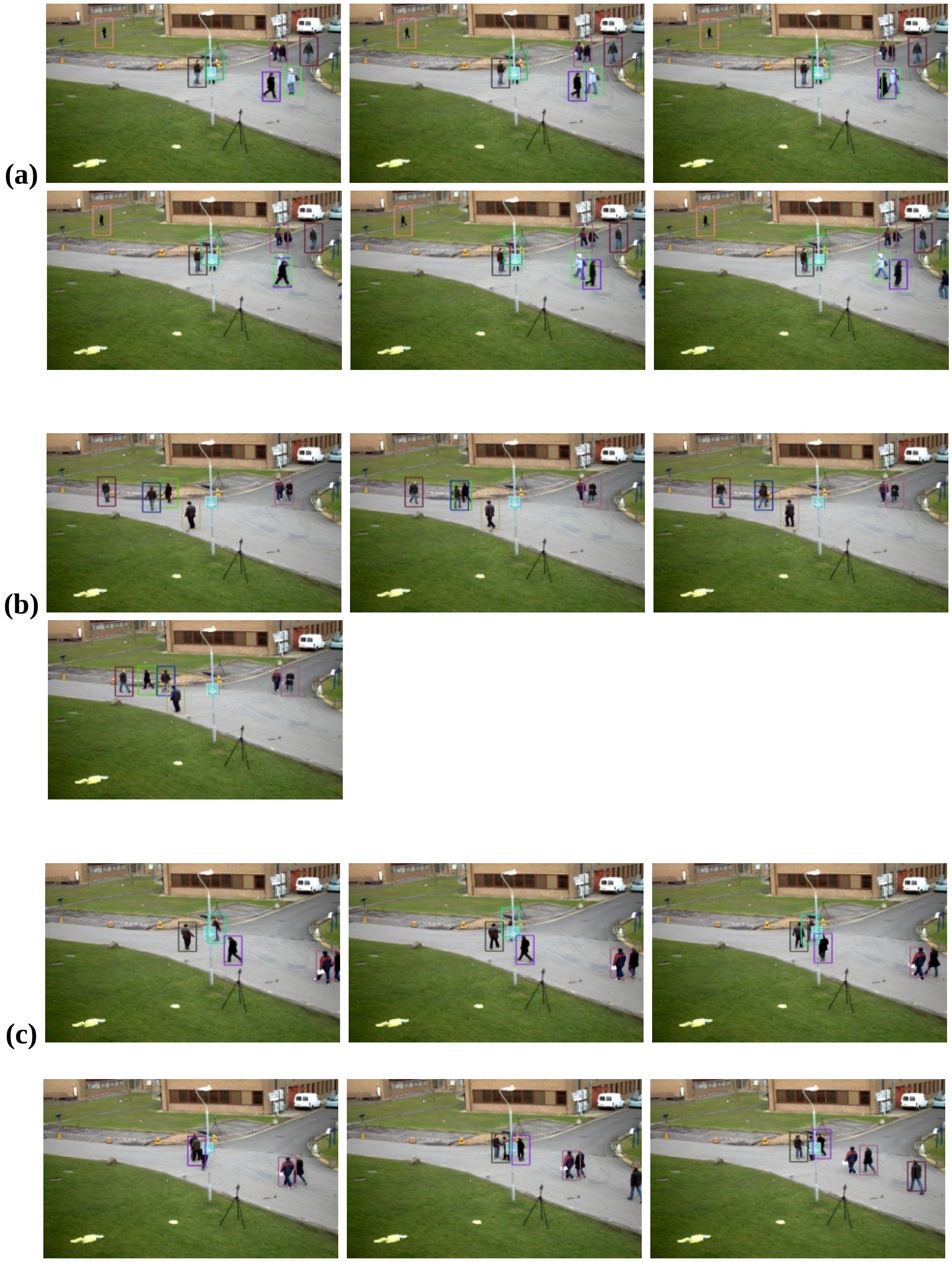


Figure 3. Some HOG detector combined with Background subtraction results a) S2.L2, b) PETS2000, c) TUD Campus, d) S2.L3.

Tracking

The main problem of tracking is occlusions; they must be handled such that after the occlusion a person's identity is still known. Figure 4 shows some occlusions and how the algorithm handles them. In illustrations of tracking, each person being tracked is depicted using its own individually colored bounding box. The longest duration of occlusion is shown in Figure 4(c) where three men come together behind a street lamp and then one of them passes to the back of the other two and leaves them with the same color of its bounding box.


(a)
(b)
(c)

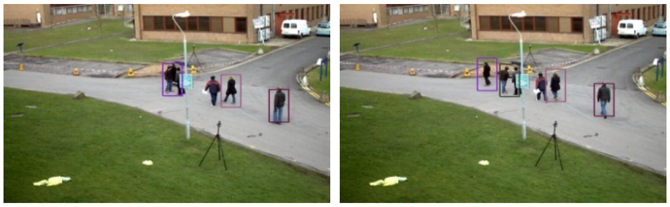


Figure 4. Occlusion handling cases in S2.L2

Tracking of TUD campus is shown in Figure 5. There are no identity changes in tracking of TUD campus sequence but the number of missing detections is large. This is because people close to the camera completely overlap people farther from camera preventing them from being detected.

The Table 1 shows number of false positives, misses and ID switches. Although there are no false positives and ID switches for TUD Campus sequence, the number of undetected people in terms of misses is very large. This is due to compactness of people standing very close to each other overlapping one another and moving approximately with the same speed. S2.L2 introduces 1 miss per frame with average of 6 people in each frame. The number of false positives is seen once per frame for about half of the sequence. For S2.L3 sequence (Figure 6) the performance evaluation is conducted only on 100 frames (frames from 50 to 150).

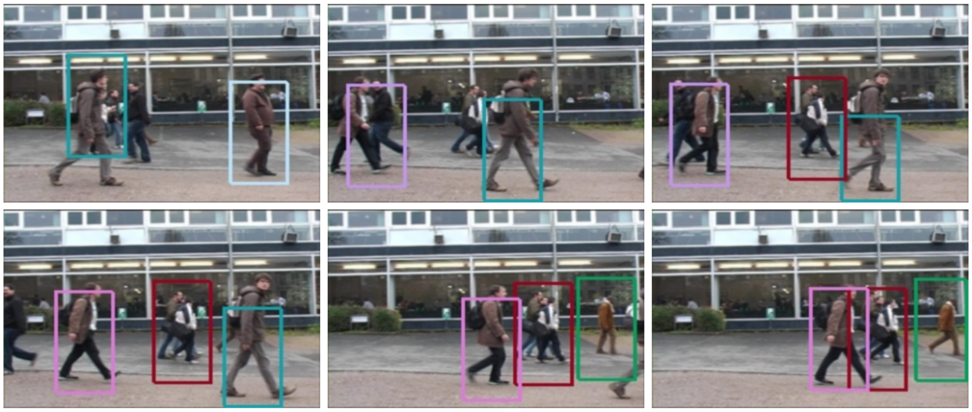


Figure 5 TUD Campus tracking results

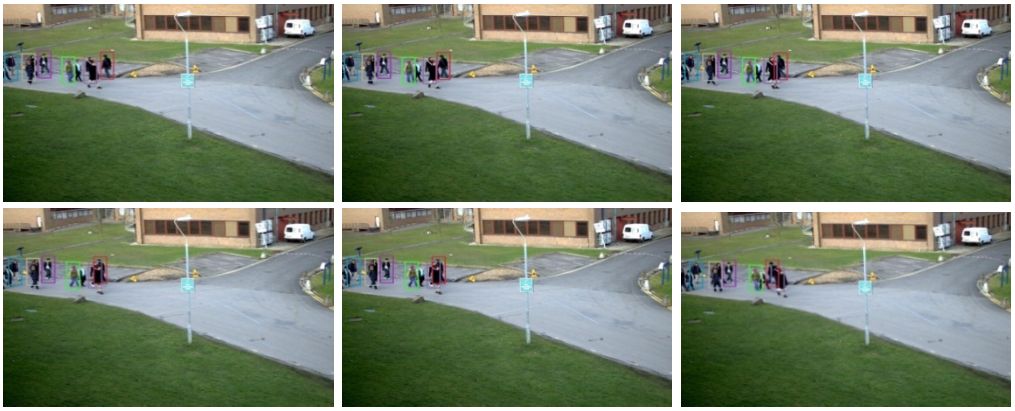


Figure 6. PETS'09 S2.L3 tracking snapshot of an occlusion handling.

Table 1. Performance Evaluation.

	False Positives	Misses	ID Switches
TUD Campus	0	65%	0
PETS'09 S2.L2	9%	17%	20
PETS'09 S2.L3	6%	30%	30

4. CONCLUSION

We proposed to combine background subtraction based tracking with tracking-by-detection approach in a Kalman filtering framework. The HOG detector is used to detect people in an outdoor scene. However the HOG detector can introduce some faulty results or detection misses. For this reason it is combined with background subtraction method. Without background subtraction method being used there would be more false positives because HOG detector can erroneously confuse trees or building parts as people. By directing the detector where to look false positives are reduced.

The algorithm works fairly well if we can get more or less accurate background estimation. In very crowded regions people will come together and form large blobs thus introducing difficulties both in detection and subsequent matching. The algorithm is using one view of a scene obtained by a single camera making the detection of people in a crowded region incomplete. As it is seen from results people walking behind other people are often not detected. For better results sequences from different view points for the same sequence can be considered.

REFERENCES

[1] Zhou Q., Aggarwal J.K., "Tracking and classifying moving objects from video", in Proc. of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, (2001), 52-59.

[2] Collins R.T., Lipton A.J., Kanade T., Fujiyoshi H., Duggins D., Tsin Y., Tolliver D., Enomoto N., Hasegawa O., Burt P., Wixons L., "A system for video surveillance and monitoring", Technical Report, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, (2000).

[3] Chen B., "Indoor and outdoor people detection and shadow suppression by exploiting HSV color information", Frontiers of Electrical and Electronic Engineering in China, Vol. 3, (2008), 406-410.

[4] Gucchiara R., Grana C., Piccardi M., Prati A., "Detecting objects, shadows and ghosts in video streams by exploiting color and motion information", in Proc. of 11th International Conference on Image Analysis and Processing, (2001), 360-365.

[5] Shah M., Javed O., Shafique K., "Automated visual surveillance in realistic scenarios", IEEE Multimedia, Vol. 14, (2007), 30-39.

[6] Breitenstein M.D., Reichlin F., Leibe B., Koller-Meier E., Gool L.V., "Robust tracking-by-detection using a detector confidence particle filter", IEEE 12th International Conference on Computer Vision, (2009), 1515-1522.

[7] Breitenstein M.D., Reichlin F., Leibe B., Koller-Meier E., Gool L.V., "Online multi-person tracking-by-detection from a single, uncalibrated camera", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, (2011), 1820-1833.

[8] Kuo C., Neviata R., "How does person identity recognition help multi-person tracking", IEEE Conference on Computer Vision and Pattern Recognition, (2011), 1217-1224.

[9] Dalal N., Triggs B., "Histogram of oriented gradients for human detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, (2005), 886-893.

[10] Leibe B., Seemann E., Schiele B., "Pedestrian detection in crowded scenes", IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, (2005), 878-885.

[11] Viola P., Jones M., "Robust real-time face detection", International Journal of Computer Vision , Vol. 57, (2004), 137–154.

[12] Comaniciu D., Ramesh V., Meer P., "Kernel-based object tracking", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, (2003), 564–575.

[13] Leibe B., Schindler K., Gool L.V., " Coupled detection and trajectory estimation for multi-object tracking", in Proc. of 11th IEEE International Conference on Computer Vision, (2007), 1-8.