International Journal of Computational Intelligence and Informatics, Vol. 3: No. 4, January - March 2014

Adaptive Multiple Kernels with SIR-Particle Filter Based Multi Human Tracking for Occluded Environment

T Karpagavalli
Department of Electronics and Communication
KLN College of Information Technology
Sivagangai, Tamilnadu, India
tkarpagam08@gmail.com

S Appavu alias Balamurugan
Department of Information Technology
KLN College of Information Technology
Sivagangai, Tamilnadu, India
app_s@gmail.com

***Abstract*- This paper proposes a new technique to build a fully automatic tracking system which handles occlusion problem in a complex environment. In multiple human tracking, handling of occlusion is the challenging issue. When occlusion occurs, kernel based tracking was proven to be the promising approach. Hence, to overcome the occlusion problem the human body was considered to have multiple kernels. In this paper, SIR-Particle filter tracking was embedded with multiple kernels that build a fully automatic tracking system. The accuracy of the tracking system was evaluated by using Multiple Object Tracking Accuracy (MOTA) metric. Our tracking system was experimented using PETS benchmark dataset and found that the accuracy was computed as 97%.**

Keywords-Human tracking, MOTA, Multiple kernels, Occlusion, SIR-Particle filter

I. INTRODUCTION

Multi human tracking in complex environments has become more and more important in video surveillance. In modern society, the automatic surveillance system becomes more popular in a large variety of applications. For example, CCTV surveillance is used to record and counteract criminal acts in town centre's, public buildings and transport termini, for road planning, and to observe shopping patterns in a supermarket.

The process of locating the moving object in sequence of frames is known as tracking. This tracking can be performed by using the feature extraction of objects and detecting the objects in sequence of frames. Several approaches have been proposed to deal with the tracking problems. The three major category of tracking systems are as follows [1].

a) Point tracking: The target in the frame is expressed as a point, and the previous target state is utilized to make the association between targets and points. Particle filter and kalman filter are widely used in this category.

b) Kernel tracking: Targets are tracked by computing the motion of the kernels which represent the appearance of the targets. Mean shift tracker is a kind of kernel tracking.

c) Silhouette tracking: If the contour or the shape of the target is required, then this method will be adopted. Given the target model, the target is tracked by matching the contour region in each frame.

Among the above tracking methods, kernel based tracking was known to be the popular method for better and more robust tracking. The basic idea of the kernel-based tracking is to minimize the difference between the target and the candidate appearance models, which are constructed by spatially masking the object with a kernel [2]. When occlusion occurs the tracker couldn't locate target in consecutive frames and loses the target. This paper focuses on dealing with the occlusion problem, which is the major challenging issue in video tracking. Single kernel does not locate target in each frame, since the visual information is not sufficient for kernel usage. So multiple kernels are used which locates target's position in each frame even after occlusion. Thus, in order to track multiple humans simultaneously and automatically, this paper proposes a new technique in which SIR-Particle filter tracking system was embedded with multiple kernels. The obtained tracking system was fully automatic without any manual intervention.

II. REVIEW OF THE LITERATURE

In video object tracking, there are many techniques existed for handling occlusion. Here we are going to see the review on kernel based tracking system. Mean shift method is one of the kernel based tracking; it finds the most similar location around the local neighbourhood area [2], [3]. In [4], difference of Gaussian is used which tracks the object in scale space. In [5], sample based similarity measure is combined with fast Gaussian

transform to fulfil the mean shift tracking. All the above work uses single kernel which doesn't locate the targets position in each consecutive frames. This was not suitable when occlusion occurs.

They all lose its target during occlusion. So in order to provide better tracking system, multiple kernels are used. In [6], different similarity measures are used for multiple kernel tracking. A two-step approach was used to determine the multiple kernels in [7]. With the help of multiple kernels, the performance was enhanced during fast motion [8]. Fragments-based multiple kernel tracking, considers the target as several fragments and was tracked by a kernel [9]. In [10], it uses feature representation which considers the multiple kernel tracking.

In general, there are two necessary activities, detection and tracking. For this, determining the state vectors is essential. There are several techniques which could be suitable for real time implementation for automatic tracking of multiple humans. State estimators can be classified into three categories; maximum likelihood (ML), maximum a posterior (MAP), Bayesian estimation. The ML method estimates the state from the current observations without prior knowledge e.g., [11]. The state vector is estimated by finding the optimal state which maximises a likelihood function which includes the appearance descriptor. The MaP method, e.g., [12], [13], is similar to ML but the estimated state is computed by optimization, and prior knowledge of the state is used to calculate the posterior state variables. Bayesian estimation preserves both the prior and posterior probability distributions, and is better able to predict future probability distributions and sampling strategies. This has been used extensively in visual tracking, e.g., [14-16], notably through particle filtering.

For multiple subject tracking, detection in space and time can be linked to previously known subjects using trajectories by the shortest path, e.g., [17], or by appearance, e.g., [18], [19]. There are many forms of appearance descriptor, based for example on silhouettes, edges, and boundaries, e.g., [20], [21] pattern descriptors, e.g., [22], [23] or color features and histograms, e.g., [11]. 3-D descriptors allow the system to better combine likelihoods from multiple cameras, e.g., [24].

Based on the review, the multiple kernel tracking was considered to be the better approach for multi-human tracking under occlusion and the bayesian approach, SIR-Particle Filter was used to obtain the fully automatic tracking system. Then a new form of appearance model is used, based on a colored, textured 3-D ellipsoid that is progressively learned as the subject moves.

III. OUR PROPOSED SYSTEM

The tracking system in our approach was to track humans continuously even after occlusion. This was achieved by using multiple kernel adaptive filter. This would locate the target's position in each consecutive frame after occlusion. Then the SIR-Particle filter tracking was embedded with this to obtain the fully automatic tracking system. The figure.1 shows the entire tracking system of our approach. Each block in the system was explained in each section.

In our approach the humans were projected using ellipsoidal model. Since ellipses can provide more precise shape information than bounding boxes, simple 3D parametric ellipse was used in order to decrease computational costs, and at the same time, it is sufficient to represent the tracking results.

There are several benchmark datasets available in video surveillance. For our experiment, the benchmark dataset PETS was used. To track humans from video, first the video has to be converted to frames. The pre-processing on those frames should be done to filter the noise, here Gaussian filter was used. Then the frames can be used for further processing which tracks human accurately.

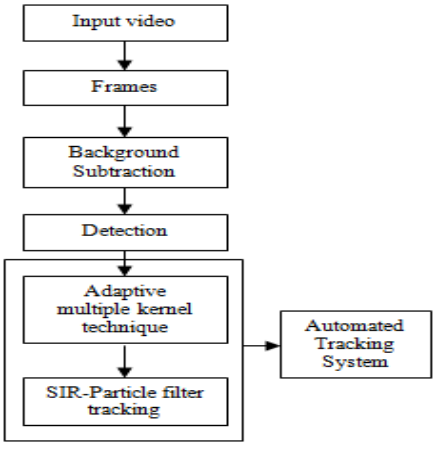

Figure 1. Our Proposed system

A. Background Subtraction

At each frame background subtraction using mixture of Gaussian (MoG) model [25] extracts active pixels that differ from the background. Thus, the foreground pixels obtained are used for detection. The Gaussian distribution is given by,

$$g(x) = \frac{1}{\sqrt{2\pi}.\sigma}.e^{\frac{-x^2}{2\sigma^2}} \quad (1)$$

where σ – standard deviation, x- variable.

To detect the human correctly from the foreground image, the morphological operations open and close has been done. So that the blob can be identified correctly from the foreground and can track the humans correctly.

(a)

(b)

Figure 2. Input frame and Background subtracted frame

The Fig.2 shows the input frame and the foreground frame in which morphological operation has been done.

B. Detection of a subject

It is necessary to detect a new presence, caused for example by entry into the field of view or re-emergence from occlusion or blind areas. Detection must search the entire observable space, and can have considerable complexity. It is the initiation of a subject track performed independently in each frame. The first four blocks in the figure.1 represents the detection process. In each frame the position of the human might change. So it is necessary to detect the presence of human in each consecutive frame to track them continuously.

The detected object was projected by using ellipse.

C. Multiple Kernel Tracking System

a) Single Kernel Tracking

In conventional kernel based tracking [2], a model is represented as the probability density function in the feature space, i.e., the histogram. During the histogram extraction, the amount of the contribution of a pixel is controlled by the value of a kernel function, and the value is determined based on the distance between the pixel and the kernel centre. In [2], the tracking procedure for maximizing the similarity is formulated as finding x that maximizes the density estimator $f(x)$,

$$f(x) = \frac{\sum_{i=0}^{N_k} w_i k\left(\left\|\frac{x - z_i}{h}\right\|^2\right)}{\sum_{i=0}^{N_h} k\left(\left\|\frac{x - z_i}{h}\right\|^2\right)}, \quad (2)$$

Where x is the center of the kernel, which is normally the centroid of the object; z_i is the pixel location to be considered and N_h is the number of pixels; w_i is the weight for each pixel; $k(.)$ is the kernel function.

If we use a single kernel to track the object, the mean-shift tracking can be adopted. However, when the target is occluded, the error may occur. This can be avoided by applying multiple kernels as shown in figure.3, where a kernel is expressed as an ellipse. If occlusion happens, the tracking performance can be severely affected since the kernel 1 incorporates a large portion of irrelevant information (obstacle) to the target in figure.1 (a). However, once an additional kernel is added in figure.1 (b), although kernel 2 is nearly non-observable, the well-observable kernel 1 can still be used to compensate the adverse effect resulted from the occlusion after some constraints which link the two kernels are introduced.

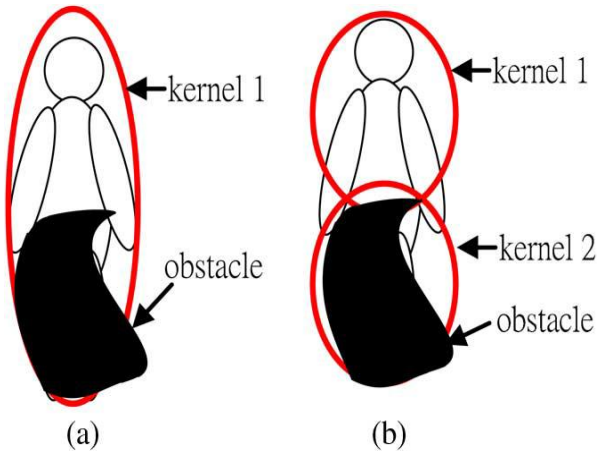

Figure 3. Red ellipses represent kernels. (a) Single kernel with occlusion. (b) Two kernels with occlusion.

b) Multiple Kernel Tracking

In multiple-kernel tracking module, each object is represented by multiple inter-related kernels, and the overall state vector x is composed of the centroids of all the kernels $[x_1 y_1 .... x_N y_N]^T$, where N is the number of kernels, and (x_i, y_i) is the centroid of the $i-th$ kernel. Theoretically, more kernels will give better results since it can handle occlusion from various directions, especially in a crowded scene environment. However, due to the different sizes of the targets, the number of kernels should be limited. For example, if the size of the object is small, each kernel may cover small portion of the body. The extracted histograms may not be informative enough due to insufficient pixels. In this paper, based on the symmetry and characteristics of human body, we use two kernels or four kernels to represent an object, and the layouts are shown in figure.4. These kernels are generated systematically and automatically once the multiple-kernel tracking is activated. The layout and constraints can be designed to reflect the target's physical characteristics. For instance, in the two kernels setting, no matter the viewing angle is from front or side, the system automatically generates the first kernel to represent the upper 55% while the other one covers the lower 55% of the object, so there is some overlapping between them. Since the upper and lower body parts of human normally have rigid geometrical relationship, it is reasonable to formulate the constraints as,

$$c_1(x_i, x_j) = \frac{((x_i - x_j)^2 - L^2_{x,ij})}{(L^2_{x,ij} + 1)} \tag{3}$$

$$c_2(y_i, y_j) = \frac{((y_i - y_j)^2 - L^2_{y,ij})}{(L^2_{y,ij} + 1)} \tag{4}$$

where $L_{x,ij}$ and $L_{y,ij}$ are the distances which remain constant after the initialization, unless the object size changes. At the first frame, each kernel initializes its own target model based on its covering area, and the $L_{x,ij}$ and $L_{y,ij}$ values are automatically determined by setting (3) and (4) to zero; that is, the constraint functions are satisfied in the beginning: $c_1(x_i^{initial}, x_j^{initial}) = 0$ and $c_2(y_i^{initial}, y_j^{initial}) = 0$. The $L_{x,ij}$ and $L_{y,ij}$ values are adjusted in proportional to the scale change only when the size of the object changes. As shown in figure.4, the extremes of the kernels $x_{\min}, x_{\max}, y_{\min}, y_{\max}$, form the bounding box.

The centroid of the tracking result is obtained based on (5),

$$x_c = \left(\frac{x_{\min} + x_{\max}}{2} \right),\ y_c = \left(\frac{y_{\min} + y_{\max}}{2} \right) \tag{5}$$

The target's model will only be updated if the average of all kernels similarity values is above 3.0. We use K-L distance [26] for all the similarity-related computations through the implementation. To construct the histogram of the object, the HSV color space and roof kernel are employed.

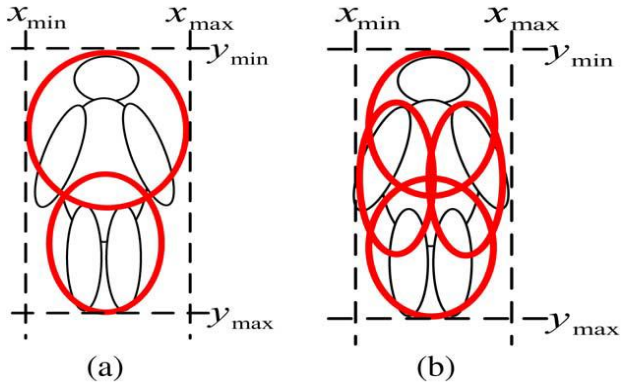

Figure 4. Layout for the multiple kernels (a) 2 kernels. (b) 4 kernels. Kernels are represented as red ellipses.

c) SIR- Particle Filter tracking

Sequential Importance Resampling (SIR) particle filter is in a family of Monte Carlo estimation. An objective is to approximate a posterior density function when a prior and a likelihood function are available. To achieve an effective sampling method, the particle exploits prediction or a prior density. The prior density is expressed by a set of points in state space; a point is called a particle. Then a weight of each particular particle is computed from the likelihood function (or similarity between observation and a synthetic image generated from the particle). The resulting weight and the particle distribution represent a mass distribution of the particle cloud and the mass distribution represents posterior density.

A particle in a high density region of the posterior represents high confidence in the state. Once the posterior is computed, the system can determine the output state from the expectation value or the maximum posterior method. In order to prepare the prediction for the next generation, the set of particles representing the posterior density is re-sampled and then transformed by a transition function.

An objective of re-sampling is to generate samples proportional to the density because a particle in the dense area will contribute more accuracy. The transition function can be modeled by the physical and statistical dynamics of the subject. Finally, all weights of particles are equalized so the particle cloud represents a prior density of next generation and the process repeats again.

Likelihood Function: The similarity value between the synthesis image I_s and the observation image I_o and save the similarity to w_n. A prior density function $P(S_t)$ and the observation image I_t are fed to the likelihood function to posterior density $P(S_t/I_t)$.

Resampling Function: The posterior density is resampled by inversion sampling.

$$u = csw(n)\ ,$$

$$n_i^{'} = csw^{-1}(u_i) \tag{6}$$

Where, csw - cumulative sum of weights, u - dummy variables.

Transition Function: Transformation is done by the transition function to generate the prior density of the next generation.

Apply deterministic transformation, $S_n \rightarrow AS_{n;t}^{'}$

Add noise to state vector, $S_n \rightarrow S_n + \sigma w$

where, S_n - Set of particles, σ - standard deviation.

When we execute the algorithm, once the human is detected, the particles are generated based on the motion of the person. Then with the help of particles, the human was tracked continuously.

IV. EXPERIMENTAL RESULTS

The experiment was analyzed using PETS dataset. In the dataset, a small clip is taken for experiment which is occluded. For tracking the human continuously, first the human was detected in the observable space then the human was tracked. If the human was occluded or if disappears for short time, the human has to be detected first then only it can be tracked continuously. So detection module is essential for tracking multiple humans.

The Fig.5 shows the detection of multiple humans who are occluded in the video sequence. The ellipse projected on the human shows the presence of the subject on that frame. The Fig.6 shows the generation of particles based on the motion of humans. To track multiple humans continuously, the motion of the human is important.

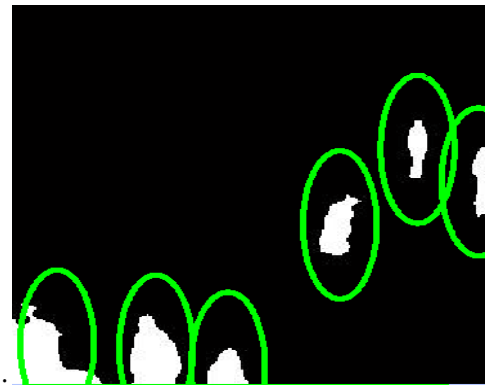

Figure 5. Detection of multiple humans

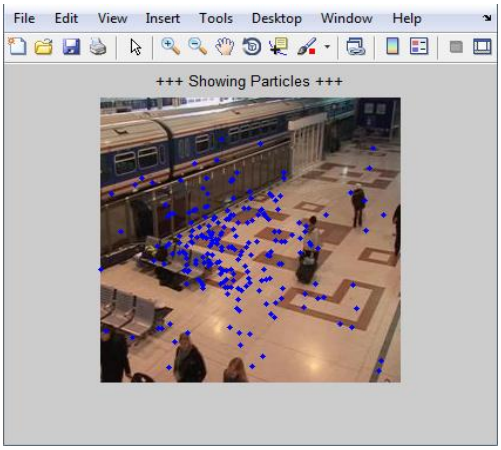

Figure 6. Generation of particles

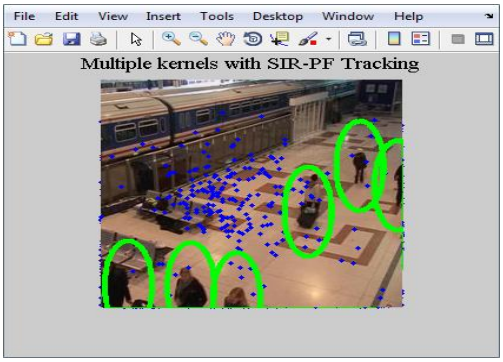

Figure 7. Tracking of multiple humans

Thus multiple humans are tracked continuously who are occluded in the video. The projected ellipse shows the continuous tracking of humans.

A. Evaluation of the System

The precision and computation time of the system was evaluated to find the accuracy of the technique used. The detection and tracking frameworks which are described can be implemented as sequential or parallel versions. A correct implementation must have identical results in both versions. The Multiple Object Tracking Accuracy (MOTA) was applied for evaluating our technique [27].

$$MOTA = \left(1 - \frac{\sum_t (f_t + m_t + s_t)}{\sum_t g_t}\right) \times 100\% \quad (7)$$

TABLE I. MOTA EVALUATED

DATASET	MOTA (%)
PETS	97.25
EPFL	97.15
PASCAL	97.09
SVS	96.90
APIDIS	97.05
AVERAGE	**97.08**

The MOTA is computed from the summation for all frames t of f_t false alarm, m_t miss detection, s_t switch events and divided by g_t total number of subjects in the reference view. When a tracker is away from its subject by more than 1m the miss detection and false alarm are counted. The table 1 shows the accuracy evaluated when tested with different benchmark datasets and found that the accuracy was improved up to 97%.

V. CONCLUSION

Thus, the proposed automated tracking system which handles the occlusion problem was obtained from the use of adaptive multiple kernels with SIR-Particle filter approach. From the experimental results, it was clearly found that our tracking system performs better for occluded environment. The accuracy of the tracking system was evaluated by MOTA and found that it was increased up to 97%.

In future, the algorithm can be modified to handle long time occlusion in crowded environment.

ACKNOWLEDGMENT

I sincerely thank all my professors and the institution for helping me to complete the project.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey”, ACM Comput Surveys, vol. 38, no. 4, 2006.

[2] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 5, pp. 564–577, May 2003.

[3] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 5, pp. 603–619, May 2002.

[4] R. T. Collins, “Mean-shift blob tracking through scale space”, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003, vol. 2, pp. 234–240.

[5] V. Yang, R. Duraiswami, and L. Davis, “Efficient mean shift tracking via a new similarity measure”, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005, vol. 1, pp. 176–183.

[6] G. D. Hager, M. Dewan, and C. V. Stewart, “Multiple kernel tracking with SSD”, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004, vol. 1, pp. 790–797.

[7] B. Martinez, L. Ferraz, X. Binefa, and J. Diaz-Caro, “Multiple kernel two-step tracking”, in Proc. IEEE Int. Conf. Image Processing, 2006, pp. 2785–2788.

[8] F. Porikli and O. Tuzel, “Multi-kernel object tracking”, in Proc. IEEE Int. Conf. Multimedia and Expo., 2005, pp. 1234 1237.

[9] J. Fang, J. Yang, and H. Liu, “Efficient and robust fragments-based multiple kernels tracking”, Int. J. Electron. Commun., vol. 65, pp. 915–923, 2011.

[10] V. Parameswaran, V. Ramesh, and I. Zoghlami, “Tunable kernels for tracking”, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006, vol. 2, pp. 2179–2186.

[11] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 5, pp. 603–619, May 2002.

[12] A. Jepson, D. Fleet, and T. El-Maraghi, “Robust online appearance models for visual tracking”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 10, pp. 1296–1311, Oct. 2003.

[13] L. Ma, J. Liu, J. Wang, J. Cheng, and H. Lu, “A improved silhouette tracking approach integrating particle filter with graph cuts”, in IEEE Int. Conf. Acoust. Speech Signal Process., Mar. 2010, pp. 1142–1145.

[14] G. Li, W. Qu, and Q. Huang, “A multiple targets appearance tracker based on object interaction models”, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 3, pp. 450–464, Mar. 2012.

[15] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map", IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 2, pp. 267–282, Feb. 2008.

[16] Z. Husz, A. Wallace, and P. Green, "Tracking with a hierarchical partitioned particle filter and movement modelling", IEEE Trans. Syst., Man, Cybern., Part B: Cybern., vol. 41, no. 6, pp. 1571–1584, Dec. 2011.

[17] F. Yan, A. Kostin, W. Christmas, and J. Kittler, "A novel data association algorithm for object tracking in clutter with application to tennis video analysis", in Proc IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Washington, DC, 2006, vol. 1, pp. 634–641.

[18] Z. Kalal, K. Mikolajczyk, and J. Matas, "Face-TLD: Trackinglearning-detection applied to faces", in Proc. 17th IEEE Int. Conf. Image Process., Sep. 2010, pp. 3789–3792.

[19] G. Cielniak and T. Duckett, "People recognition by mobile robots", J.Intell. Fuzzy Syst., vol. 15, pp.21–27, 2004.

[20] H.-B. Kim and K.-B. Sim, "A particular object tracking in an environment of multiple moving objects", in Proc. Int. Conf. Control Autom. Syst., Oct. 2010, pp. 1053–1056.

[21] Y. Tang, Y. Li, T. Bai, X. Zhou, and Z. Li, "Human tracking in thermal catadioptric omnidirectional vision", in Proc. IEEE Int. Conf. Inf. Autom., Jun. 2011, pp. 97–102.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, vol. 1, pp. 886–893.

[23] H. Ma, C. Zeng, and C. X. Ling, "A reliable people counting system via multiple cameras", ACM Trans. Intell. Syst. Technol. vol. 3, no. 2, pp. 31:1–31:22, Feb. 2012.

[24] J. Yao and J. M. Odobez, Multi-Person Bayesian Tracking With Multiple Cameras. New York: Elsevier, 2009, ch. 15, p. 363.

[25] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real time tracking", in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 1999, vol. 2, p.252.

[26] C. M. Bishop, Pattern Recognition and Machine Learning, New York, NY, USA: Springer, 2006.

[27] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear MOT metrics", J. Image Video Process., vol. 2008, pp. 1–10, 2008.