Journal of Real-Time Image Processing (2021) 18:1803–1814
https://doi.org/10.1007/s11554-021-01144-5

SPECIAL ISSUE PAPER

A real-time person tracking system based on SiamMask network for intelligent video surveillance

Imran Ahmed[1] · Gwanggil Jeon[2]

Received: 22 February 2021 / Accepted: 7 June 2021 / Published online: 28 July 2021

Abstract
Real-time video surveillance systems are widely deployed in various environments, including public areas, commercial buildings, and public infrastructures. Person detection is a key and crucial task in different video surveillance applications, such as person detection, segmentation, and tracking. Researchers presented different image processing and artificial intelligence-based approaches (including machine and deep learning) for person detection and tracking, but mainly comprised of frontal view camera perspective. A real-time person tracking and segmentation system is introduced in this work, using an overhead camera perspective. The system applied a deep learning-based algorithm, i.e., SiamMask, a simple, versatile, fast, and surpassing other real-time tracking algorithms. The algorithm also performs segmentation of the target person by combining a mask branch to the fully convolutional twin neural network for target or person tracking. First, the person video sequences are obtained from an overhead perspective, and then additional training is performed with the help of transfer learning. Finally, a comparison is performed with other tracking algorithms. The SiamMask algorithm delivers good results, with a tracking accuracy of 95%.

Keywords Smart video surveillance · Image processing · Deep learning · Overhead view · Person tracking · SaimMask

1 Introduction

Real-time video surveillance extends the importance of the intelligent world, allowing sensory connections at a worldwide level, playing as the joining point between the digital and real worlds, and serving as a convincing catalyst for the smart and digital transformation for various surveillance applications. These applications are widely expanded in various public environments, applied for real-time monitoring of physical assets, locations, analysis of video information obtained to identify security indicators, and security planning. The advent of machine, deep learning, and image processing techniques has commenced new research possibilities in this field. Deep learning has enabled the automated extraction and analysis of information from images and video sequences. The convergence of deep learning with image processing is valuable in a variety of security applications. Detection of high risk situations before they heighten is one of the essential motivations behind the advancement of artificial intelligence-based security and surveillance applications. With these advancements, operators can expand surveillance solutions beyond mere monitoring to leverage every video frame and piece of data available to identify threats and inform the emergency response.

In real-time video surveillance systems, detecting a person is essential for diverse applications, including person identification, person tracking, person counting, unusual event detection, and crowd monitoring [7]. Along with a wide range of applications, it is a challenging task for researchers because of the variable visual features of a person's body, including appearances (scale and size) and deformable poses. The complex and cluttered backgrounds or scenes, lighting conditions, different kinds of occlusions, abrupt variations in motion, and camera perspectives also affect the efficiency and performance of tracking algorithms. Researchers proposed different person tracking techniques,

✉ Gwanggil Jeon
gjeon@inu.ac.kr

Imran Ahmed
imran.ahmed@imsciences.edu.pk

[1] Centre for Excellence in Information Technology, IMSciences Peshawar, 1-A, Sector E-5, Phase VII, Hayatabad, Peshawar, Pakistan

[2] Department of Embedded Systems Engineering, Incheon National University, Incheon, Korea

mostly employed conventional handcrafted features and machine learning-based approaches [8, 9], which are computationally expensive and require extra background training to learn person features. In contrast, advanced deep learning-based methods e.g., [10–13] presents effective solutions for person detection and tracking which are effective in terms of efficiency and computation speed [9] and tried to overcome aforementioned challenges.

Mostly advanced machine learning and deep learning-based tracking and detection algorithms are often used frontal or asymmetric camera viewpoints [10–12], where images are obtained from frontal view as presented in Fig. 1. The person's body, visual features in the images vary in terms of body orientation, different poses, movements, and body articulations. The example images also highlight the occlusion problem that occurs when the other person and object overlap each other. Some researchers, e.g., [6, 14, 15], considered an overhead camera perspective for person tracking and detection as illustrated in Fig. 2. It is noticeable that the person's body's visual features are different from such an extreme view than the frontal view, usually depend on the local rotations, the movements of the body, and its position with respect to the position of the camera. In an overhead view, the problem of occlusion is considerably less as compared to the frontal view, where cross object occlusion problem can occur when the scene/environment turns crowded, as illustrated in Fig. 1.

With above mentioned motivations, researchers preferred to use an overhead camera in different surveillance applications including person detection [16–18], person counting [4, 5, 19, 20], person tracking [20–24], action recognition, crowd analysis [25], behaviour understanding [26], and human posture identification [27]. Moreover, managing the occlusion dilemma, it also overcomes privacy issues [28], reduce computation and installation expenses [29]. The contrast between both camera perspectives is highlighted in Figs. 1 and 2. One can easily observe that the person's body features depend upon different camera perspective, both camera perspective results unusual modification in the person's body visual characteristics (pose, size, shape, scale, and body orientation). This perspective might overcome the challenges of occlusion and allows broader coverage of the view.

A real-time intelligent surveillance system is presented for an overhead view person tracking and segmentation. For person tracking and segmentation, a deep learning-based algorithm, i.e., SaimMask [30], is explored. The algorithm can produce both video target tracking and video target segmentation in real time. The algorithm is simple, versatile, and fast and delivers good results compared too other real time tracking systems [31, 32], and pand provides the state-of-the-art in the target tracking field. Meanwhile, it obtained competitive performance and the fastest speed on the DAVIS-2016, DAVIS-2017 video segmentation data set

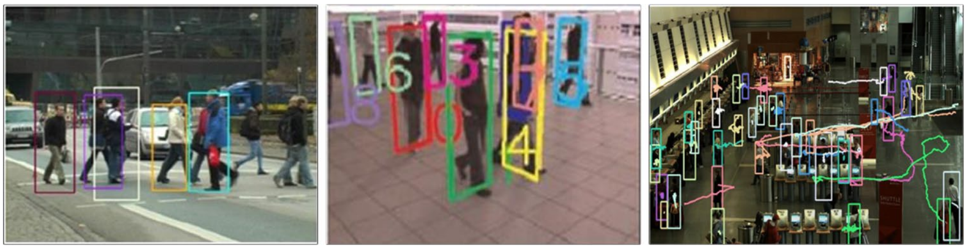


Fig. 1 Frontal or asymmetric camera view: sample images taken from [1–3]. The first two sample images are recorded from the frontal view camera, while the third image is captured from the asymmetric view. The body features of individuals vary in terms of appearance (scale, shape, size, and pose). The sample images also highlight the problem of occlusion

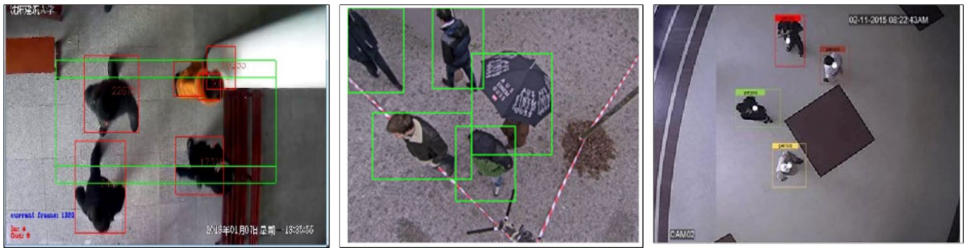


Fig. 2 Overhead camera view: example images taken from [4–6]. All sample images are captured from an overhead view with different camera heights. It can be recognized that the visual features of the person's body are different from a frontal view; the occlusion problem is reduced, and more coverage of the scene is obtained

[30]. This tracking algorithm mainly realizes the segmentation of the target by combining a mask branch to the fully convolutional neural network (twin) for person (target) tracking. The algorithm is first experimented with an overhead view data set. While considering the overhead view, there are significant variations in visual features of the person; for that reason, the network is additionally trained with the overhead view person data set, and the improved trained feature layer is added with the existing network applying transfer learning. The experimental results reveal that the accuracy results of the person tracking and segmentation algorithm are improved after training. In general, the principal objectives of the paper are given as:

- Real-time intelligent surveillance system is introduced for overhead view person tracking and segmentation.
- The system utilized a deep learning-based tracking algorithm, i.e., SiamMask, for an overhead view person target tracking and segmentation in video sequences.
- The performance of the network architecture is investigated by testing the tracking algorithm with pre-trained and trained network architecture.
- The tracking performance is also compared with different bounding box representation strategies.
- The tracking accuracy results are compared with other tracking algorithms.

The work presented in this paper is arranged as follows: a review of the related work is provided in Sect. 2. Section 3 explains a real-time smart surveillance system for overhead perspective person tracking and segmentation. The implementation, and experimental assessment of the system, are reported in Sect. 4. Finally, the conclusion and possible prospective directions of the work are presented in Sect. 5.

2 Related work

Person tracking from an overhead perspective is considered a challenging task for researchers in various surveillance applications. In this section, we presented a review of some recent overhead view-based person tracking methods.

Migniot et al. [15] presented a hybrid 3D–2D tracking approach using particle filtering for human tracking. Authors in [20] presented a graph structuring technique for overhead view person tracking. Most of the techniques developed by researchers are mainly focused on the head, head–shoulder, or sometimes, on the entire body information of the person. Few researchers like [21, 31–33], also applied particle filter for person tracking using overhead perspective. A good review of different overhead view detection and tracking techniques utilized for people is provided in [34].

Vera et al. [4] adopted a Hungarian approach for people tracking in overhead view video sequences. Gao et al. [35] practiced median filter and [36] adopted mean shift algorithm for person tracking. Bagaa et al. [37] provided an effective tracking system for 5G networks. Nakatani et al. [28] considered head region information as Region of Interest (RoI) and applied hair texture information for person tracking. Authors in [22, 36, 38–40] assumed head– shoulder information as RoI for person detection. Researchers in [32, 36, 41, 42] examined the full human body as RoI.

Some researchers, e.g., [28, 32, 35, 39], applied color-based information, while few researchers utilized edge information's, such as canny edge detector with SIFT features [32], and Sobel filters [43], for person detection and tracking. In [44], authors utilized a feature-based method, e.g., Histogram of Oriented Gradient (HoG) and presented an efficient person detection system. In [24], authors studied local ternary patterns and support vector machine classifiers for person detection and tracking. Ozturk et al. [32] analyzed the shape of individual body as an elliptical blob and introduced a tracking and detection system. Wu et al. [45] and Wetzel et al. [46] designed person tracking and detection method utilizing depth images captured from an overhead camera.

Ahmed et al. [47] introduced a rotated HoG method to recognize people in a complex industrial setting using an overhead camera. In [14], authors proposed a robust algorithm for person detection that utilized variable/different sized bounding boxes with different angles. Authors in [24] assumed a fixed sized detection bounding box and method based on feature for person detection. Authors, in [23, 48] offered another feature-based approach for person detection and tracking in indoor and industrial conditions. Ullah et al. [49] performed a comparison and investigated some conventional tracking algorithms for the person using an overhead camera. Ullah et al. [50] further implemented a blob-based strategy and offered a rotation invariant system for person tracking. A rotated feature and classification- based method is presented by [17] for person detection.

Deep learning methods [5] are also utilized for person tracking. The majority of the advanced studies' practiced the frontal perspective data set. Many scholars [15, 51–53] offered target detection and tracking utilizing aerial and satellite data set. Authors in [54, 55] studied pre-trained deep learning approaches for detecting and tracking persons using an overhead camera perspective. Authors in [18] studied Mask-RCNN and Faster-RCNN for multiple overhead view object segmentation and detection. Ahmed et al. [7] presented multiple people tracking framework based on deep learning-based tracking and detection model using 5G infrastructure.

Ahmad et al. [16] selected a deep learning model to detect and track multiple people in an overhead view outdoor and indoor scenes. In another work, authors [56] studied different deep learning-based segmentation methods for people using an overhead view. Ahmed et al. [57] implemented two separate deep learning models for multiple object detection coupled with various tracking algorithms. Authors, in [60], presented a real-time IoT-based framework for overhead view person detection by utilizing a deep learning model.

From the above review, it is concluded that significant work has been performed by different researchers for overhead view person tracking. Researchers employed color, texture, and shape-based information for person tracking. Mostly researchers used different handcrafted feature-based approaches, while few of them also practiced different deep learning-based models. In this work, we explored the deep learning-based SiamMask algorithm for overhead view person tracking and segmentation.

3 Real-time overhead view surveillance system for person tracking and segmentation

A real-time person tracking and segmentation system is presented for an intelligent surveillance application. The technical description of the introduced system is presented in Fig. 3, which includes an image processing unit, mainly comprised of a deep learning algorithm. The image processing unit is connected with the cloud server and internet connection to enhance the efficiency of the developed system by decreasing the computational expense and processing high-resolution video sequences over the cloud in real time. The recorded video sequences are collected at the cloud storage and image processing unit with the help of the internet connection. The image processing unit consists of artificial intelligence or a deep learning-based algorithm to process or analyze the high-resolution video sequences at high processing and computation speed. The network architecture is also trained for person video sequences captured using an overhead camera. For person tracking, a deep learning algorithm, SiamMask [30], is applied. As the visual features of a person's body from an overhead view are different, thus to enhance the system's accuracy for person tracking in an overhead view, additional training is performed. The improved, learned features are added with pre-trained weights using transfer learning, as depicted in Fig. 3. The results of the image processing unit are further transmitted to the monitoring and surveillance unit, where it might assist monitoring operators for different surveillance applications. The developed system details are given as follows.

The algorithm tries to shorten the space between binary object tracking and object segmentation. It is also known as a multi-task learning process that can be applied to resolve object tracking and segmentation problems. The algorithm's primary innovation is that if an object rotates, the appearance of a single box typically creates a significant loss, which is actually a defect in the representation itself. SiamMask instantly predicts the mask of an object, which allows to obtain the most accurate box. The algorithm follows offline training and online speed method and efficiently improves

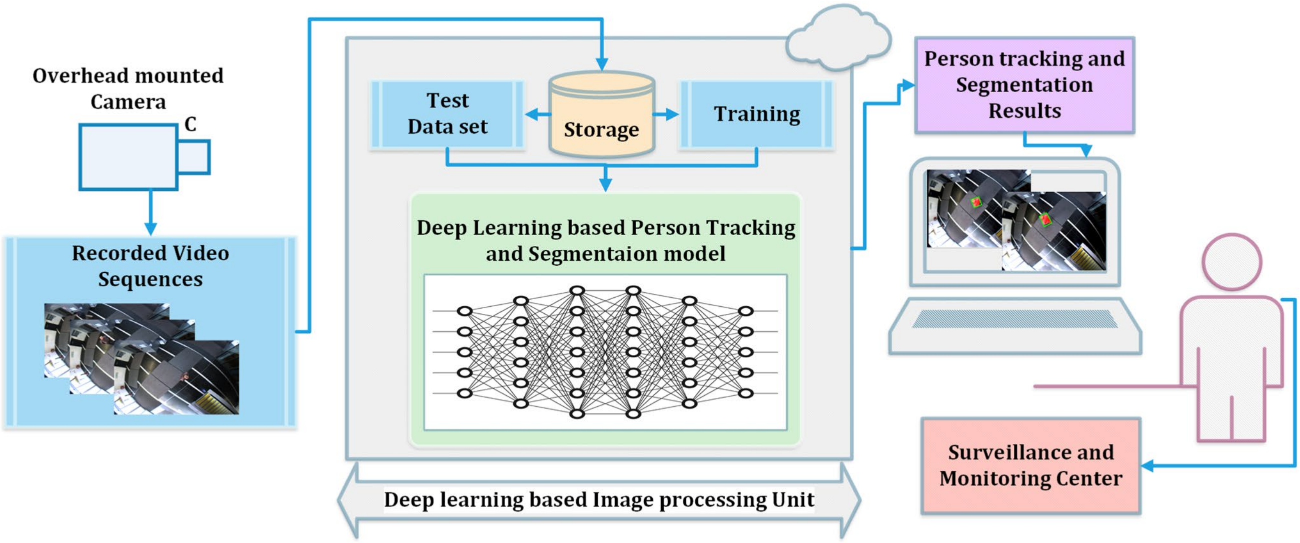

Overhead mounted Camera
C
Recorded Video Sequences
Test Data set
Storage
Training
Deep Learning based Person Tracking and Segmentaion model
Deep learning based Image processing Unit
Person tracking and Segmentation Results
Surveillance and Monitoring Center

Fig. 3 Real-time person tracking and segmentation system for overhead view surveillance. The recorded overhead view video sequences are sent to the cloud storage and image processing unit. The image processing unit utilized a deep learning-based algorithm for person tracking and segmentation. The final tracking and segmentation results are transferred to the monitoring and surveillance unit for further processing

the representation of the object (target) while confined to a simple axis-aligned boundary box.

The schematic representation of the architecture is presented in Fig. 4. It can be seen that, the input is classified into two parts, the top one is used for target image z, and the lower part is used for the search image x (a larger then Z). To have fast speed and online operability, the algorithm adopted the fully convolutional Siamese network (SiamFC). It resembles a target image z against image x to get a feature map (dense response map), as illustrated in Fig. 4. The flow chart of the SiamMask tracking algorithm/model is explained in Fig. 5.

As the size of z is smaller than x, the obtained feature map $f_\theta(z)$ is also smaller than the feature map $f_\theta(x)$. The $f_\theta(z)$ is then slid over $f_\theta(x)$, and a similarity matrix is applied to join the two matrices into a particular scoring matrix. Lastly, the large value in the scoring matrix is the point with the highest confidence, the region corresponding to image x is the predicted region of the frame image. These two input images are processed by the corresponding CNN, to obtain features of the image and producing cross-correlated feature maps [58]:

$$g_\theta(z;x) = f_\theta(z) \star f_\theta(x). \tag{1}$$

Each spatial component of the feature map $g_\theta(z;x)$ is referred as the response of a candidate window (RoW). It means that the highest value of the feature map corresponds to the target area in the image x. Alternatively, to enable every RoW to encode more valuable data regarding the target object, the cross-correlation feature map in Eq. 1 is replaced by depthwise cross-correlation [58], and a response/feature map with multiple channels are produced. The SiamFC network was trained on millions of frames using the logistic loss [59]. To enhance the efficiency of SiamFC, a regional recommendation network is applied, which enables a variable sized bounding box to determine the target position. In particular, every RoW encodes a collection of anchor boxes (k) proposals and similar background or object scores.

As viewed from the schematic diagram, the mask generated by each RoW is a vector, which means that the resulting mask image is very rough, and its size is smaller than the initial image. Hence, it is often flatten with a process of up-sampling and adjustment. The accuracy of predicting the mask is not so high; therefore, refine module u-shape structure is used, which combines the feature map of the backbone, and performs up-sampling to get more fine segmentation results. Along with bounding box coordinates and similarity scores, the RoW of a fully convolutional Siamese network is applied to further encode the information needed to generate a pixelwise binary mask. Thus, the Siamese network is continued with an additional branch and loss. The binary masks are predicted $w \times h$ (one for each RoW) utilizing a single two layers neural network $h\phi$ with ϕ learnable parameters. Throughout the training, every RoW is labeled with $y_n \varepsilon \{\pm 1\}$ a true binary label and corresponds with c_n pixelwise true mask having a size of $w \times h$. The mask prediction loss function L_{mask} is a binary logistic regression of all RoWs, which is given as [30]:

$$L_{\text{mask}}(\theta, \phi) = \sum_n \left(\frac{1 + y_n}{2wh} \sum_{ij} \log \left(1 + e^{-c_n^{ij} m_n^{ij}} \right) \right). \tag{2}$$

In Eq. 2, $c_n^{ij} \varepsilon \{\pm 1\}$ indicates the label of the corresponding pixel i, j of the object in the RoW. Therefore, the h_ϕ,

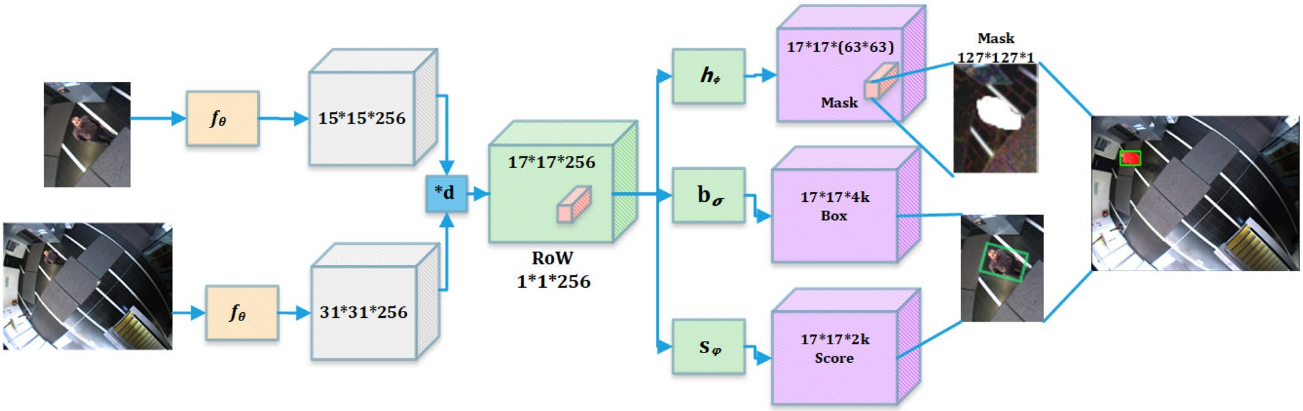
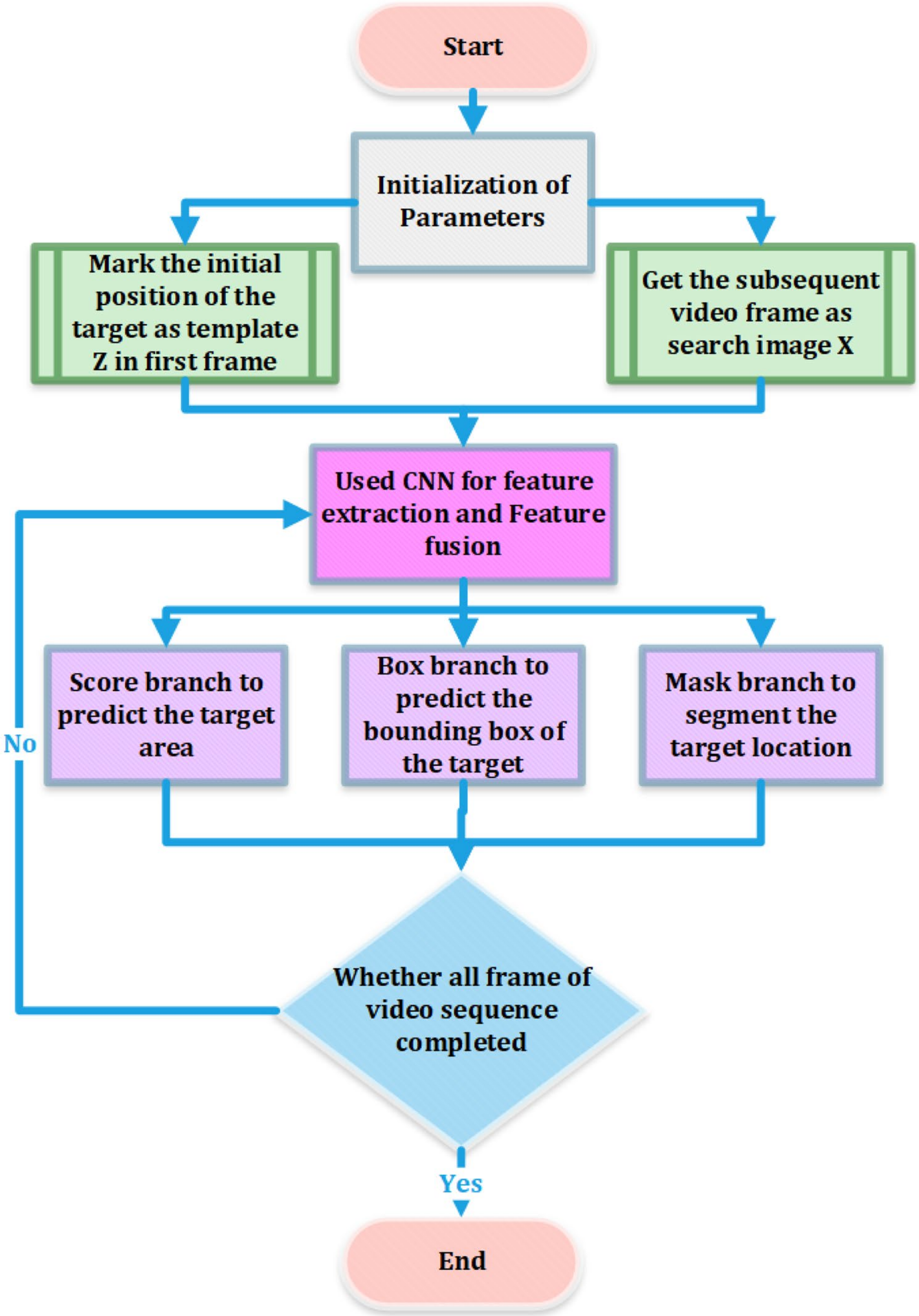


Fig. 4 Schematic illustration of SiamMask [30], the input image is processed through backbone convolutions layers "Conv" ResNet-50 for feature extraction (three-branch architecture). The architecture is based on twin networks. *d is a depthwise cross-correlation process, which indicates that the correlation computation is done channel-by-channel basis. The middle as RoW (response of candidate window) is responsible for keeping the number of channels unchanged and then divides three categories or branches based on this RoW, segmentation, regression, and classification

Fig. 5 Flow chart of SiamMask algorithm for tracking and segmentation of person in overhead videos


Start
Initialization of Parameters
Mark the initial position of the target as template Z in first frame
Get the subsequent video frame as search image X
Used CNN for feature extraction and Feature fusion
Score branch to predict the target area
Box branch to predict the bounding box of the target
Mask branch to segment the target location
No
Whether all frame of video sequence completed
Yes
End

classification layer comprised of $w \times h$ classifiers, all showing either a given pixel relates to the target in the candidate window or not. In original work, the L_{mask} is estimated only for positive RoWs. In general, the training process is end-to-end training, which means that all three branches are trained at the same time. Thus, for all training samples, the labels of the three branches are provided. To utilize the smooth L1 and the cross-entropy losses, the output branches are trained. For bounding box regression and classification score, L_{box} and L_{score} is used, respectively. The total loss function is calculated as:

$$L_{3B} = \lambda_1 \cdot L_{\text{mask}} + \lambda_2 \cdot L_{\text{score}} + \lambda_3 \cdot L_{\text{box}}. \tag{3}$$

In Eq. 3, L_{3B} describe the three-branch network and the two-branch variant (for more details of the equation we refer reader for original work [30, 59, 60]). The mask branch simply determines the loss of the positive sample. The sample is referred to as a single RoW; when there is an anchor box and ground truth in an RoW and its intersection over union IOU is greater than 0.6, it is recognized as a positive sample. For the scoring and the bounding box branch, the SiamFC [60] and SiamRPN [59] method is used, respectively. In Fig. 5, the general flow of the overall algorithm is presented. It can be seen that two images, z and x, are given to the CNN model. The CNN model performed feature extraction and output three different branch results, including classification score, bounding box, and segmentation for the target object in image x. The process is performed for all frames of the video sequences. The bounding box of all feature proposals is achieved according to the bounding box and mask branch, and the final results are obtained by applying NMS (non-maximal suppression).

4 Experimental results

Different experimentation and performance evaluation are developed in this section. A comprehensive explanation of the data set utilized for experimentation is also discussed. The experiments are implemented using the Python programming language. For testing videos, we utilized SaimMask [30] without any modification. For both variants, same as [30], a ResNet-50 architecture is utilized. During tracking, SiamMask simply assessed one frame for once, without any modification. The output mask is obtained for both variants utilizing the location and highest value in the classification branch. Furthermore, the mask branch's output is binarised, after utilizing a per-pixel sigmoid, at the threshold of 0.5. The data set utilized for person tracking, and segmentation has also been discussed. The performance assessment is made using different quantitative tests. For initial experiments, we apply mean intersection over union (mIoU) and average precision (AP) at $\{0.5;0.7\}$. The tracking accuracy of the algorithm is analyzed with different state-of-the-art tracking algorithms.

4.1 Data set

A real-world data set, recorded using an overhead camera at the Southampton University, United Kingdom [48] is utilized for obtaining person video sequences. The Point Grey Flea camera with a Fujinon wide-angle lens is utilized for recording purposes. The video sequences are converted into a video frame that has a frame resolution of 1024×768 pixels and PNG format. The video sequences are recorded utilizing a single camera installed at an altitude of about 4 m from the ground.

The positions and locations of the person are manually annotated to determine the ground truth information.

4.2 Tracking and segmentation results

To improve the results, the architecture is additionally training with an overhead data set. Despite being different in nature, the original architecture gives failure, because it unambiguously separated objects from the foreground. The results of the tracking algorithm are elaborated in Fig. 6. The tracking and segmentation results of the SiamMask for the overhead person data set have been visualized for different test video frames. It can be observed that extra training enhances the overall performance accuracy of the algorithm. Persons with different visual features, as shown in Fig. 6 are now correctly classified, segmented, and tracked in subsequent frames. The red color in the sample frames shows the segmented mask, while the green color box represents an automatically rotated bounding box that is used for tracking the target person in the video frame.

The experimental results show that the SiamMask algorithm achieves good results for moving target tracking results; from the first frame to the last frame (01–2000), in overhead view video frames, the target is always in the tracking state. In Fig. 6, we show the output results for few frames, for example, in the first-row sample frames, it can be seen that a person at different locations is kept tracked from the center of the scene to the upper left corner. The algorithm segment the target region along with adjusting the bounding box. Similarly, in subsequent frames of row two, the bounding box and segmentation mask of the algorithm is accurately obtained, although the shape of the person is significantly varying in the scene when the person moves away from the camera position.

In third-row sample frames, the person suddenly changes its direction and body angle from the overhead perspective, but still, the algorithm keeps its segmentation mask and bounding box according to its detected shape. In the fourth row, the person at the center of the images is accurately tracked without any failure. The same visual effect can be seen in the last row of the image in which the position and the location of the person are varying, but the deep learning-based algorithm accurately tracked it without any failure.

In Fig. 6, we show the results for few sample frames, mostly it can be seen that the person visual features are changing in subsequent frames in terms of size, scale, pose, and orientation. The overall tracking results are good, as compared to other tradition tracking algorithms.

4.3 Performance evaluation

The surveillance system's performance is evaluated using different performance parameters, i.e., mAP and mIOU. The tracking algorithm outputs a set of bounding boxes and segmentation masks for the person in video frames. For evaluation

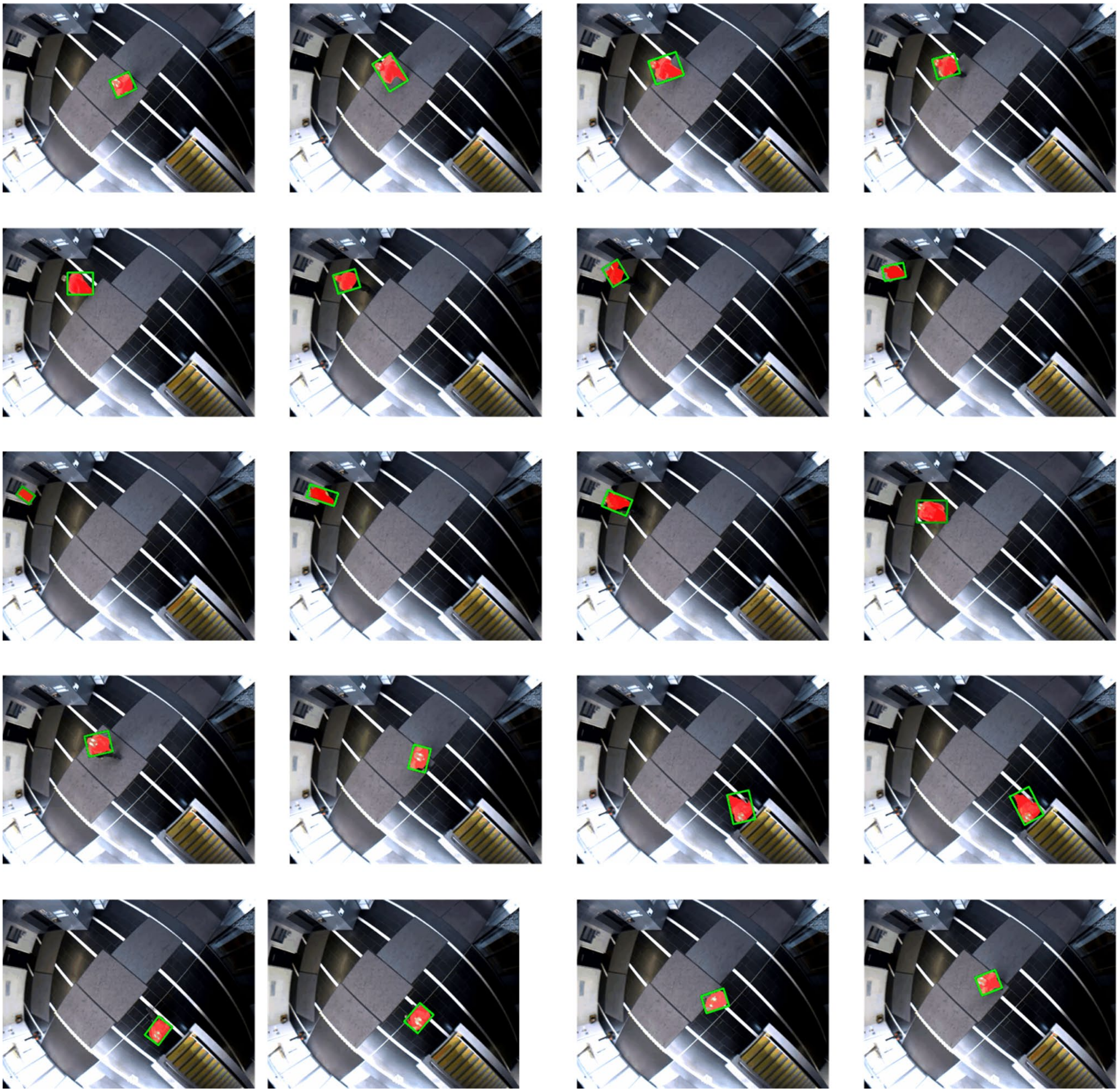


Fig. 6 Visualization results of SiamMask used for person tracking and segmentation, the results are presented for the few frames, the movement, visual characteristics, and person location are varying in the scene. In the green rectangular bounding boxes, the correctly detected target object as a person label class is shown, while the red color is used for the predicted segmented mask

of classification, the predicted bounding box is matched with the ground truth bounding box, and IoU is calculated given as:

$$\mathrm{IoU} = \frac{b_{\mathrm{pred}} \cap b_{\mathrm{groundtruth}}}{b_{\mathrm{pred}} \cup b_{\mathrm{groundtruth}}}. \quad (4)$$

In Eq. 4, the predicted bounding box is expressed with b_{pred} and ground truth bounding boxes are indicated with $b_{\mathrm{groundtruth}}$. For person classification, IoU threshold is defined as IoU ≥ 0.5 and 0.7. In addition, if multiple detections occur, then the first one is counts as positive and the rest as negatives. The precision p, recall r, and accuracy acc values are also applied for determining the mAp values and formulated as follows:

$$p = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \quad (5)$$

$$r = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \tag{6}$$

$$\mathrm{acc} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fp} + \mathrm{fn}}. \tag{7}$$

In the above equations, tp represents true positive are correctly classified bounding boxes as the person, fp shows false positives the number of bounding boxes inaccurately classified, tn true negative indicates those bounding boxes which are correctly recognized as background, and fn false-negative means those bounding boxes which are incorrectly recognized other objects or background as a person.

The mAP mean average precision for N classes is given as:

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i. \tag{8}$$

In Eq. 8, the interpolated average precision AP value is given as:

$$\mathrm{AP} = \frac{1}{11} \sum_{r \varepsilon (0, 0.1, \ldots, 1.0)} \max_{r' \geq r} p(r'). \tag{9}$$

The mAP and mIOU values of the algorithm are provided in Table 1. We present the results for generated bounding box from a binary mask, as original work [30] three different approaches, namely, Min–max (the object in an axis-aligned rectangle), MBR (the minimum bounding box rectangle), and Opt: (the rectangle obtained via the optimization) are used. It is worth to note here that tracking results are almost consistent with original work [30]. The results show that the algorithm produces the best mIOU, no matter which bounding box generation approach is utilized.

The threshold precision curve is also plotted in Fig. 7. Moreover, the results are further compared with some other tracking algorithms. We have experimented algorithms on the overhead person data set, and for a fair illustration of comparison results, the same experimental parameters are used. It can be seen that the SiamMask performs adequately and give good results with a precision rate of 0.95 as compared to other tracking algorithms.

The tracking accuracy results compared with other algorithms are shown in Fig. 8. We experimented KCF, Median Flow, TLD, and SiamMask for the overhead view person data set. It can be observed from the experimental results that the tracking accuracy rate of SiamMask is 0.95, while mostly tracking algorithms not performed well and give a tracking accuracy rate around 0.80.

Table 1 Performance of bounding box representation approaches on overhead view person data set

Method	mIoU	mAP at 0.5 IOU %	mAP at 0.7 IOU %
SiamMask–Min–max	64	82	43
SiamMask–MBR	67	85	50
SiamMask–Opt	71	91	62

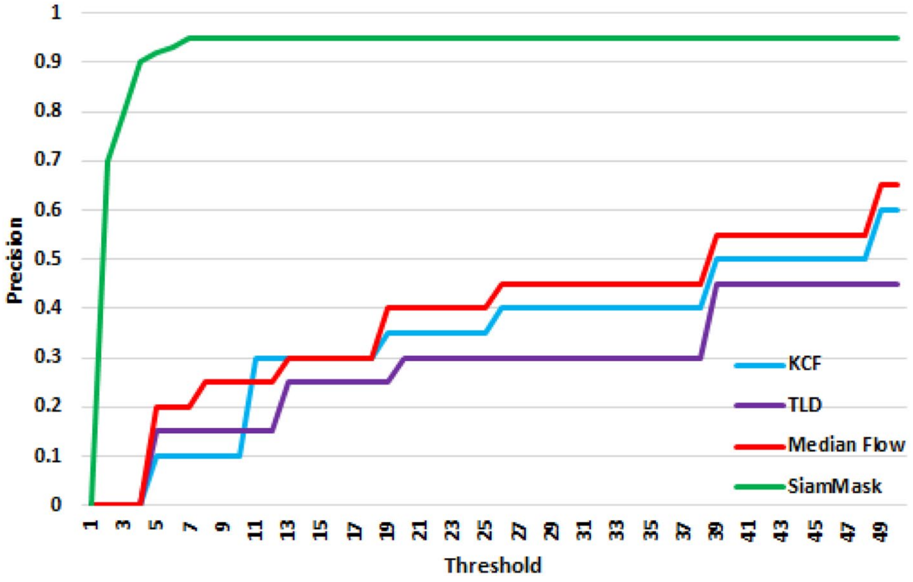


Fig. 7 Threshold–precision curve of the tracking algorithms

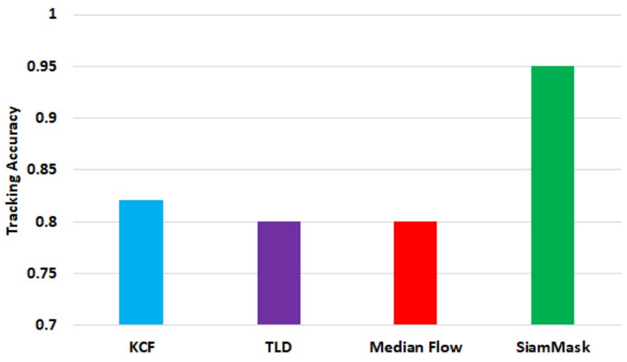


Fig. 8 Tracking accuracy of different algorithms and SaimMask

5 Conclusion and future work

In this work, a real-time person tracking and segmentation system has been introduced for an overhead view. The introduced system consists of an image processing unit that utilized a deep learning algorithm named SiamMask for real-time tracking and segmentation applications. We presented an intelligent real-time surveillance system by integrating it with the cloud and the internet server. The deep learning-based algorithm performed segmentation of the target person by combining a mask branch to the fully convolutional twin neural network. We performed additional training to enhance the performance of the system. Finally, the tracking results of the SiamMask algorithm are compared with other tracking algorithms, namely, KCF, TLD, and Median flow. Experimentation reveals that the SiamMask algorithm delivers better results than the other algorithms. The tracking accuracy rate of the SiamMask algorithm is 0.95. Moreover, the threshold–precision curve is also plotted and compared with other tracking algorithms. The algorithm's accuracy might be enhanced with new data sets recorded in various situations against multiple backgrounds and illumination situations in future work. In addition, the work might be continued for detection, tracking, and segmentation of multiple objects.

Acknowledgements This work was supported by Incheon National University Research Concentration Professors Grant in 2019.

References

1. Choi, J.W., Moon, D., Yoo, J.H.: Robust multi-person tracking for real-time intelligent video surveillance. ETRI J. **37**(3), 551 (2015)
2. Liu, P., Li, X., Liu, H., Fu, Z.: Multidisciplinary Digital Publishing Institute: online learned Siamese network with autoencoding constraints for robust multi-object tracking. Electronics **8**(6), 595 (2019)
3. Potdar, K., Pai, C.D., Akolkar, S.: A convolutional neural network based live object recognition system as blind aid (2018). arXiv preprint. arXiv:1811.10399
4. Vera, P., Monjaraz, S., Salas, J.: Counting pedestrians with a zenithal arrangement of depth cameras. Mach. Vis. Appl. **27**(2), 303 (2016)
5. Ertler, C., Possegger, H., Opitz, M., Bischof, H.: Pedestrian detection in RGB-D images from an elevated viewpoint. In: Kropatsch, W., Janusch, I., Artner, N. (eds.) Proceedings of the 22nd Computer Vision Winter Workshop. TU Wien, Pattern Recognition and Image Processing Group, Vienna (2017)
6. Ahmad, M., Ahmed, I., Ullah, K., Khan, I., Adnan, A.: Robust background subtraction based person's counting from overhead view. In: 9th IEEE Annual Ubiquitous Computing. Electronics Mobile Communication Conference (UEMCON), pp. 746–752 (2018)
7. Ahmed, I., Ahmad, M., Ahmad, A., Jeon, G.: Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: within 5G infrastructure. Int. J. Mach. Learn. Cybern. 1–15 (2020)
8. Nguyen, D.T., Li, W., Ogunbona, P.O.: Human detection from images and videos: a survey. Pattern Recognit. **51**, 148 (2016)
9. Buongiorno, A., Trotta, G.F., Bevilacqua, V.: Computer vision and deep learning techniques for pedestrian detection and tracking: a survey. Neurocomputing **300**, 17 (2018)
10. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey (2019). arXiv preprint arXiv:1905.05055
11. Yao, R., Lin, G., Xia, S., Zhao, J., Zhou, Y.: Video object segmentation and tracking: a survey (2019). arXiv preprint. arXiv:1904.09172
12. Zhou, S., Ke, M., Qiu, J., Wang, J.: A survey of multi-object video tracking algorithms. In: Abawajy, J., Choo, K.K.R., Islam, R., Xu, Z., Atiquzzaman, M. (eds.) International Conference on Applications and Techniques. Cyber Security and Intelligence ATCI 2018, pp. 351–369. Springer International Publishing, Cham (2019)
13. Li, P., Wang, D., Wang, L., Lu, H.: Deep visual tracking: review and experimental comparison. Pattern Recognit. **76**, 323 (2018)
14. Ahmed, I., Adnan, A.: A robust algorithm for detecting people in overhead views. Clust. Comput. **21**(1), 633 (2018). https://doi.org/10.1007/s10586-017-0968-3
15. Migniot, C., Ababsa, F.: Hybrid 3D–2D human tracking in a top view. J. Real Time Image Process. **11**(4), 769 (2016)
16. Ahmad, M., Ahmed, I., Khan, F.A., Qayum, F., Aljuaid, H.: Convolutional neural network-based person tracking using overhead views. Int. J. Distrib. Sens. Netw. **16**(6), 1550147720934738 (2020)
17. Ahmed, I., Ahmad, M., Nawaz, M., Haseeb, K., Khan, S., Jeon, G.: Efficient topview person detector using point based transformation and lookup table. Comput. Commun. **147**, 188 (2019)
18. Ahmed, I., Din, S., Jeon, G., Piccialli, F.: Exploring deep learning models for overhead view multiple object detection. IEEE Internet Things J. **7**(7), 5737 (2020)
19. Kristoffersen, M., Dueholm, J., Gade, R., Moeslund, T.: Pedestrian counting with occlusion handling using stereo thermal cameras. Sensors **16**(1), 62 (2016)
20. Burbano, A., Bouaziz, S., Vasiliu, M.: 3D-sensing distributed embedded system for people tracking and counting. In: 2015 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 470–475 (2015)
21. Tseng, T., Liu, A., Hsiao, P., Huang, C., Fu, L.: Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4077–4082 (2014)
22. García, J., Gardel, A., Bravo, I., Lázaro, J.L., Martínez, M., Rodríguez, D.: Directional people counter based on head tracking. IEEE Trans. Ind. Electron. **60**(9), 3991 (2013)

23. Ahmed, I., Ahmad, A., Piccialli, F., Sangaiah, A.K., Jeon, G.: A robust features-based person tracker for overhead views in industrial environment. IEEE Internet Things J. **5**(3), 1598 (2018)
24. Rauter, M.: Reliable human detection and tracking in top-view depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 529–534 (2013)
25. Ryan, D., Denman, S., Sridharan, S., Fookes, C.: An evaluation of crowd counting methods, features and regression models. Comput. Vis. Image Underst. **130**, 1 (2015)
26. Lin, Q., Zhou, C., Wang, S., Xu, X.: Human behavior understanding via top-view vision. AASRI Procedia **3**, 184 (2012)
27. Hsu, T.-W., Yang, Y.-H., Yeh, T.-H., Liu, A.-S., Fu, L.-C., Zeng, Y.-C.: Privacy free indoor action detection system using top-view depth camera based on key-poses. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 004058–004063 (2016)
28. Nakatani, R., Kouno, D., Shimada, K., Endo, T.: A person identification method using a top-view head image from an overhead camera. JACIII **16**(6), 696 (2012)
29. Ahmad, M., Ahmed, I., Ullah, K., Khan, I., Khattak, A., Adnan, A.: Energy efficient camera solution for video surveillance. Int. J. Adv. Comput. Sci. Appl. **10**(3) (2019). http://dx.doi.org/10.14569/IJACSA.2019.0100367
30. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: a unifying approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1328–1338 (2019)
31. Iguernaissi, R., Merad, D., Drap, P.: People counting based on kinect depth data. In: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods—Volume 1: ICPRAM. INSTICC (SciTePress, 2018), pp. 364–370. https://doi.org/10.5220/0006585703640370
32. Ozturk, O., Yamasaki, T., Aizawa, K.: Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1020–1027 (2009)
33. Snidaro, L., Micheloni, C., Chiavedale, C.: Video security for ambient intelligence. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **35**(1), 133 (2005)
34. Ahmad, M., Ahmed, I., Ullah, K., Khan, I., Khattak, A., Adnan A.: Int. J. Adv. Comput. Sci. Appl. **10**(4) (2019). https://doi.org/10.14569/IJACSA.2019.0100470
35. Gao, C., Liu, J., Feng, Q., Lv, J.: Person detection from overhead view: a survey. Multimedia Tools Appl. **75**(15), 9315 (2016). https://doi.org/10.1007/s11042-016-3344-z
36. Velipasalar, S., Tian, Y., Hampapur A.: Automatic counting of interacting people by using a single uncalibrated camera. In: 2006 IEEE International Conference on Multimedia and Expo, pp. 1265–1268 (2006)
37. Bagaa, M., Taleb, T., Ksentini, A.: Efficient tracking area management framework for 5G networks. IEEE Trans. Wirel. Commun. **15**(6), 4117 (2016)
38. Yu, S., Chen, X., Sun, W., Xie D.: A robust method for detecting and counting people. In: 2008 International Conference on Audio, Language and Image Processing, pp. 1545–1549 (2008)
39. Wateosot, C., Suvonvorn, N., et al.: Top-view based people counting using mixture of depth and color information. In: The Second Asian Conference on Information Systems, ACIS (Citeseer, 2013)
40. Perng, J., Wang, T., Hsu, Y., Wu B.: The design and implementation of a vision-based people counting system in buses. In: 2016 International Conference on System Science and Engineering (ICSSE), pp. 1–3 (2016)
41. Yahiaoui, T., Meurie, C., Khoudour, L.: A people counting system based on dense and close stereovision. In: Cabestaing, F., Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) Image and Signal Processing, pp. 59–66. Springer, Berlin (2008)
42. Cao, J., Sun, L., Odoom, M.G., Luan, F., Song X.: Counting people by using a single camera without calibration. In: 2016 Chinese Control and Decision Conference (CCDC), pp. 2048–2051 (2016)
43. Mukherjee, S., Saha, B., Jamal, I., Leclerc, R., Ray N.: Anovel framework for automatic passenger counting. In: 2011 18th IEEE International Conference on Image Processing, pp. 2969–2972 (2011)
44. Pang, Y., Yuan, Y., Li, X., Pan, J.: Efficient HOG human detection. Signal Process. **91**(4), 773 (2011)
45. Wu, C.J., Houben, S., Marquardt, N.: EagleSense: tracking people and devices in interactive spaces using real-time top-view depth-sensing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery, New York, NY, USA, 2017). CHI '17, pp. 3929–3942. https://doi.org/10.1145/3025453.3025562
46. Wetzel, J., Laubenheimer, A., Heizmann, M.: Joint probabilistic people detection in overlapping depth images. IEEE Access **8**, 28349 (2020)
47. Ahmed, I., Carter, J.N.: A robust person detector for overhead views. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1483–1486 (2012)
48. Ahmed, I., Ahmad, M., Adnan, A., Ahmad, A., Khan, M.: Person detector for different overhead views using machine learning. Int. J. Mach. Learn. Cybern. **10**(10), 2657 (2019). https://doi.org/10.1007/s13042-019-00950-5
49. Ullah, K., Ahmed, I., Ahmad, M., Khan, I.: Comparison of person tracking algorithms using overhead view implemented in OpenCV. In: 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON) (IEEE, 2019), pp. 284–289
50. Ullah, K., Ahmed, I., Ahmad, M., Rahman, A.U., Nawaz, M., Adnan, A.: Rotation invariant person tracker using top view. J. Ambient Intell. Humaniz. Comput. 1–17 (2019)
51. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: object detection and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
52. Zhu, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., Wu, H., Nie, Q., Cheng, H., Liu, C., et al.: VisDrone-VDT2018: the vision meets drone video detection and tracking challenge results. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
53. Qi, Y., Zhang, S., Zhang, W., Su, L., Huang, Q., Yang, M.H.: Learning attribute-specific representations for visual tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8835–8842 (2019)
54. Ahmad, M., Ahmed, I., Adnan, A.: Overhead view person detection using YOLO. In: 2019 IEEE 10th Annual Ubiquitous Computing. Electronics Mobile Communication Conference (UEMCON), pp. 0627–0633 (2019)
55. Ahmad, M., Ahmed, I., Ullah, K., Ahmad, M.: A deep neural network approach for top view people detection and counting. In: IEEE 10th Annual Ubiquitous Computing. Electronics Mobile Communication Conference (UEMCON), pp. 1082–1088 (2019)
56. Ahmed, I., Ahmad, M., Khan, F.A., Asif, M.: Comparison of deep-learning-based segmentation models: using top view person images. IEEE Access **8**, 136361 (2020)
57. Ahmed, I., Din, S., Jeon, G., Piccialli, F., Fortino, G.: Towards collaborative robotics in top view surveillance: a framework for multiple object tracking by detection using deep learning. IEEE/

CAA J. Autom. Sin. 1–18 (2020). https://doi.org/10.1109/JAS.2020.1003453

58. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P.H., Vedaldi, A.: Learning feed-forward one-shot learners (2016). arXiv preprint arXiv:1606.05233
59. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8971–8980(2018)
60. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional Siamese networks for object tracking. In: European Conference on Computer Vision, pp. 850–865. Springer, Berlin (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Imran Ahmed (Senior Member, IEEE) received a Ph.D. degree in Computer Science from the University of Southampton, Southampton, U.K. Currently, he is working as an Assistant Professor with the Institute of Management Sciences Peshawar, Hayatabad, Peshawar, Pakistan. His research interests include deep learning, machine learning, data science, computer vision, feature extraction, digital image and signal processing, medical image processing, biometrics, pattern recognition, and data mining. He has also attended several international conferences in these areas. He has published numerous articles in refereed journals and conference proceedings, including IEEE Access, IEEE Transactions on Industrial Informatics, IEEE Internet of Things Journal, International Journal of Machine Learning and Cybernetics, IEEE/CAA Journal of Automatica Sinica, Journal of Ambient Intelligence and Humanized Computing, Multimedia Tools and Applications, Cluster Computing, IEEE Annual Ubiquitous Computing, Computer Communications, Electronics & Mobile Communication Conference (UEMCON), ACM/SIGAPP Symposium On Applied Computing at Pau, France, and many others. He has also been acting as a reviewer of journals, such as IEEE Transactions on Industrial Electronics, IEEE Access, Journal of Ambient Intelligence, and Elsevier, etc. He is also Associate Editor of well reputed international journals, including IEEE Access.

Gwanggil Jeon received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2003, 2005, and 2008, respectively. From 2009.09 to 2011.08, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. From 2011.09 to 2012.02, he was with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, as an Assistant Professor. From 2014.12 to 2015.02 and 2015.06 to 2015.07, he was a Visiting Scholar at Centre de Mathématiques et Leurs Applications (CMLA), École Normale Supérieure Paris-Saclay (ENS-Cachan), France. From 2019 to 2020, he was a Prestigious Visiting Professor at Dipartimento di Informatica, Università degli Studi di Milano Statale, Italy. He is currently a Full Professor at Incheon National University, Incheon, Korea. He was a Visiting Professor at Sichuan University, China, Universitat Pompeu Fabra, Barcelona, Spain, Xinjiang University, China, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, and University of Burgundy, Dijon, France. Dr. Jeon is an Associate Editor of Sustainable Cities and Society, IEEE Access, Real-Time Image Processing, Journal of System Architecture, and MDPI Remote Sensing. Dr. Jeon was a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and Industry-Academic Merit Award by Ministry of SMEs and Startups of Korea Minister in 2020.