Integrated Metric Learning Based Multiple Object Tracking Method under Occlusion in Substations

Pengfei Zhang¹, *, Zhongyang Chen¹, *, Xinyue Zhang², b, Zhongguang Yang³, c and Wenbin Shi⁴, d

¹Shanghai Electric Power Research Institute, State Grid Shanghai Electric Power Company, Shanghai 200063, China
²School of Computer Science and Cyberspace Security, Hainan University, Haikou 570228, Hainan, China
³College of Computer and Science, Shanghai University of Electric Power, Shanghai 200082, China
⁴School of Information and Electronic Engineering, Zhejiang University of Science & Technology, Hangzhou 310023, China

*Corresponding author e-mail: yuedomore@163.com, ¹574977207@qq.com, ²648705322@qq.com, ³857322130@qq.com, ⁴yuexinuai@163.com

Abstract. Nowadays most of background subtraction based algorithms lack the robustness to handle tracking multiple objects with specific situations such as heavy occlusion in intelligent video surveillance system. This paper proposed an integrated metric learning based multiple object tracking method in smart substations. First tracks personnel obtaining bounding box, then integrates obtained mahalanobis distance and cosine distance of personnel. When occlusion occurs, proposed method compares the integrated metric vale with threshold so as to track personnel in different frames. Experiments in the smart substation confirm that our method can track multiple personnel under occlusion effectively and reliably.

1. Introduction
We have witnessed in recent years an unprecedented explosion in the availability of and access to the smart grid. With the development of the smart substation’s construction requirements, traditional monitoring methods have been unable to meet the design requirements. Therefore, many works apply computer vision technology in smart substations [1, 2]. Multiple object tracking is an essential component of computer vision. The task of the Multiple Object Tracking is that automatically identifies the target and reliably estimates its location [6]. Accurate and effective multiple object tracking will improve the surveillance performance and ensure the security of staff in the substations. Nowadays, due to the rapid development of image based detection methods, detection based tracking methods [3, 4] have gradually become an important part of multiple object tracking methods. Detection based methods first detect objects using background subtraction, then establish correspondence from frame to frame in order to find the tracks of the objects. However, the performance of the detection based tracking method highly depends on the target detector. When the detector handles the complex situations such as heavy occlusion, false and missing detection will
frequently occur. Thus, one of the important issues which come across in the video tracking systems for the surveillance applications, especially for the substation surveillance, is the occlusion of the being tracked personnel. We consider the problem of simultaneously tracking multiple objects in multiple video sequences. In particular, we focus on the cases where complex background occludes target personnel. In Figure 1, we simulate the occlusion scene in substations. The background of substations usually are full of chaotic facilities, which is difficult for surveillance. Figure 1 consists of three pictures in chronological order. Figure (a) contains one personnel, the detector provides a blue bounding box to him. In figure (b), due to occlusion, the personnel in blue bounding box disappears. In figure (c), the personnel should have been in blue bounding box now in a red bounding box, that is false detection, which degrades the tracking performance, especially the online tracking algorithm that only uses the current frame information to obtain the trajectory. Therefore, preventing occlusion correctly is important for enhancing the reliability of multiple object tracking algorithms.

![Figure 1. Occlusion scene in substation.](a) (b) (c)

Existing video tracking frameworks adopt various of models to deal with the occlusion problem, which mainly contains generative models and discriminant models. Generative models are established through online learning, then search and reconstruct the image area with the least error to complete the target location. However, generative models do not have ideal performance in speed and accuracy [7, 8]. Discriminant models regard the tracking as a binary classification problem, and among these models, tracking method based on deep learning has become the mainstream. Gao et al. [9] proposed an online specific model to deal with the obstacle occlusion. Cheong et al. [10] proposed a system, which includes foreground object segmentation, color tracking, object specification and occlusion handing to track occluded objects correctly. Tang et al. [11] proposed a multiple object tracking model to handle occlusion and nearby objects that are sharing a similar appearance. Judd et al. [12] extended the multimotion visual odometry pipeline to consistently estimate the full trajectory of every motion in a scene and recognize when temporarily occluded motions become unoccluded. With the development of multiple object tracking under occlusion, it has gained interest in the specific domain, such as sports, to capture highlights and to gather statistics. Kim [13] proposed a novel multiple object tracking method in soccer videos. However, a very few efforts have been done to alleviate occlusion issues in the substation. Therefore, we are concentrated on occlusion issues in the smart substations so as to enable optimal moving personnel detection and tracking.

What is more, in the field of deep neural network, usually extract the eigenfeatures and construct the eigenvectors to represent samples. Metric learning basically learns samples to obtain metric matrix that can reflect the similarity. Unlike most existing tracker, Hu et al. [14] introduced a deep metric learning approach for tracking under the particle filter framework, which learns nonlinear distance metric to classify targets. Zhao et al. [15] tracking objects with structured metric learning through making use of the constraints from the target and its neighbor training samples. Above tracking methods are not effectively due to the limitation of the single metric learning method.

Based on aforementioned approaches but go further, in this paper, we propose an integrated metric learning based multiple object tracking method to deal with occlusion in substations. The proposed tracking method uses multiple surveillance cameras. First, the tracker provides a bounding box for target personnel for each video frame sequence. Secondly, the method combines Mahalanobis Distance [5] predicted by Kalman Filter Algorithm and Cosine Distance obtained via our proposed deep neural network. When occlusion occurs, the method compares the integrated metric value with threshold to measure the similarity of personnel in different frames. The rest of the paper is organized
as follows. In Section II, we present the integrated metric learning based multiple object tracking method. Simulation results of the proposed method is provided in Section III. Finally, some concluding remarks are presented in Section IV.

2. Integrated metric learning based multiple object tracking method

Tracking objects in complex scenes naturally leads to the problem of occlusion. To deal with occlusion, we propose here a simple method named integrated metric learning based multiple objects tracking method. The precondition of proposed method is a frame set as threshold. If the time difference between current frame and the last successfully matching frame is bigger than the threshold, then we regard the track as stop, and we delete later trajectories of targets. If the time difference is smaller than the threshold, then we keep tracking. When tracking stops, our method starts to deal with occlusion. The overview of our architecture is shown in Figure 2.

![Figure 2. Overview of the proposed method.](image)

The method consists of three parts (shown in Figure 2): (1) object tracking, (2) predict motion and extract appearance information, (3) integrated metric learning algorithm. In part one, the tracking detector provides a bounding box for personnel for each video frame sequence, and track the trajectory of target personnel. In part two, we predict motion information as mahalanobis distance and design a deep neural network named EAF network learning the appearance features to obtain cosine distance. In part three, after occlusion, we calculate the similarity of personnel in different video sequences with integrated metric learning method.

2.1. Object tracking

In substations, due to the high height of the surveillance cameras, the tracking method needs to ensure not only the requirements of real-time but tracking small-size targets. Meanwhile, in order to compare fairly with other state-of-the-art models, we use YOLOv3 network [16] as the backbone network of the object detection network. As shown in Figure 2, the whole YOLOv3 network structure only consists of convolutional layers. The input image can complete the target classification and location by only going through the network once. And YOLOv3 has reached the most advanced level in both speed and accuracy, while adopting multi-scale prediction to improve the identification ability of small-size targets.
Since we only detect personnel in substations, we modify the classifier and change the output dimension of the last layer. Therefore, we obtain the coordinate of personnel as \((x, y, w, h)\) and confidence as \((\text{confidence}, c)\), where \((x, y, h, w)\) is the normalized offset, height and width of the center position of the person relative to the grid position. Our method uses multiple cameras (only 2 are shown in Figure 2). Each camera records video frames as \(F_i = (f_1, f_2, \ldots, f_n)\). \(i\) represents the camera’s identification. In this stage, our method provides personnel with specific identification, and obtain their trajectories. After tracking, a person was represented as \(P_i = (F_i, (x, y, h, w), \text{ID}_i)\), \(\text{ID}_i\) is the person’s identification. We use Kalman Filter Algorithm \([17, 18]\) to estimate optimal state of personnel at the next frame with current frame. The midpoint \((d_x(k), d_y(k))\) on the bottom edge of the detection box is taken as the tracking feature point, and the length \(w\) and height \(h\) of the detection box are selected as the other two feature variables to form a four-dimensional state variable. Then, the workflow of Kalman Filter Algorithm is used to predict the four-dimensional state variables of the midpoint on the bottom edge and the length \(w\) and height \(h\) of the detection box.

### 2.2. Integrated metric learning

We regard the target personnel in two continuous frames as two independent objects. If there is occlusion, we combined the motion information with appearance information, then calculate the relevance of two separate objects.

#### 2.2.1. Motion information

Considering that the detection target in current frame is composed of four-dimensional vector, we adopts mahalanobis distance to measure the similarity between current and historical trajectory of the target personnel. Mahalanobis distance represents the covariance distance of data, and it effectively calculates the similarity between two multi-dimensional personnel through taking into account the correlation between various features of the target.

In this paper, we calculate the Mahalanobis Distance \(M(i-1, i)\) of target personnel in \(\#i\) frame and in \(\#i-1\) frame, which represents as equation (1).

\[
M(i-1, i) = \sqrt{(t_i - g_{i-1})^T S_{i-1}^{-1} (t_i - g_{i-1})}
\]  

(1)

Where \(t_i\) represents the personnel state \((d_x(k), d_y(k), w, h)\) in \(\#i\) frame. \(g_{i-1}\) denotes the prediction of \(\#i\) frame in \(\#i-1\) frame. \(S_{i-1}\) is the covariance matrix of target trajectory predicted by the Kalman Filter Algorithm in \(\#i\) frame. Since the motion in the video frame is continuous, we use \(M(i-1, i)\) to filter the target, and set 3.08 as the threshold of filtering. Equation (2) shows the filtering rule, \(\text{filter}\) represents the filtering function.

\[
a(i-1, i) = \{ \text{filter}[M(i-1, i) \leq 3.08], i \geq 2, \\
0, 0 < i \leq 0\}
\]  

(2)

#### 2.2.2. Appearance information

In order to extract appearance information, we design a simple deep neural network named EAF network. To give an intuitive understanding, we visualize the EAF network architecture in Figure 3. The construction of EAF consists of ten blocks including conv1, conv2, conv3, maxpool4, residual5, residual6, residual 7, residual 8 and dense 9. The details of EAF network is shown in table 1. After ten blocks, the input of personnel appearance in bounding box finally converts into 256-dimension.
Different from other distance algorithm that directly measures between two points, the cosine distance algorithm calculates the cosine value of the angle of two vectors, which is not related with absolute value. And due to long-duration tracking, continuous tracking will change the appearance of personnel. Therefore, we use Cosine Distance Algorithm as the measurement function, in order to calculate the cosine distance between feature vector of the personnel \( b \) in \( #i \) frame and average value of \( f \) feature vectors before \( #i \) frame. Let \( r_i = \{r^{(b)}_1, r^{(b)}_2, \ldots, r^{(b)}_k\} \) denote feature vectors of personnel \( b \) before \( #i \) frame. The cosine distance calculation is in equation (3):

\[
cosdis(k, i) = 1 - r_i^T \frac{\sum_{k=1}^{f} r^{(b)}_k}{f}
\]

(3)

Here we set a threshold \( s \), when \( \cosdis(k, i) \leq s \), we regard the correlation of personnel between two frames as successful. We set a weight coefficient \( \kappa \), and obtain weighted average of mahalanobis distance and cosine distance through equation (4).

\[
u = \kappa a(i - 1, i) + (1 - \kappa) c(k, i)
\]

(4)

Finally, the Hungarian algorithm [19] was used for optimal allocation of track and detection.

**Table 1.** The details of EFA network.

| Layer     | Patch Size | Stride | Output Size      |
|-----------|------------|--------|------------------|
| Conv1     | 3×3        | 1      | 32×128×64        |
| Conv2     | 3×3        | 1      | 32×128×64        |
| Conv3     | 3×3        | 1      | 32×128×64        |
| Maxpool4  | 3×3        | 2      | 32×64×32         |
| Residual5 | 3×3        | 1      | 32×64×32         |
| Residual6 | 3×3        | 2      | 64×32×16         |
| Residual7 | 3×3        | 2      | 128×16×8         |
| Residual8 | 3×3        | 2      | 256×8×4          |
| Dense9    | -          | -      | 256              |
| Batch and \( \ell_2 \) normalization | -        | -      | 256              |

3. Experiments
Since constantly updating the training samples is time-consuming, we choose a graphics processing server with high performance. The proposed method is implemented on a PC with Intel(R) Core(TM) i7-8086k CPU @4.00ghz, x64-based processor, 16G memory, and GPU RTX2080 Ti.
3.1. Implementation details
As a concrete illustration, we consider the particular application of video tracking in smart substations. In order to detect the location of personnel in image accurately and ensure the performance of proposed method, we transform surveillance videos sequence into frames. Meanwhile, we select the same amount pictures from public data set to guarantee the generalization of proposed method. We mix collected frames and pictures from public data set as our data set. The data set contains 10073 images, then divide the data set into training set, which contains 6011 images, validation set, which contains 2021 images and test set, which contains 2041 images. We annotate the images according to the image annotation criteria of the Pascal VOC dataset, and output only the person category, while counting the person (Target ID is labeling at the top left corner of the bounding box). As the person ID for the initial frame trace, partial images of annotated dataset are shown in Figure 4. Left annotated image is the example of pascal VOC dataset, and the right one is the example of our dataset.

![Figure 4. Example of annotated images.](image)

3.2. Results and Analysis
We collect videos in substations, and the collected substation monitoring videos are divided into two groups. One group evaluates the tracking performance under the condition of less occlusion. The other group tests the robustness and accuracy in the case of severe occlusion. Figure 5 shows the results in two conditions respectively. Figure (a) is the #2 frame in a less occluded environment, and figure (b) is the #82 frame in the same environment as figure (a). Five personnel are in the image, proposed method accurately detects and tracks the 5 personnel. Figure (c) is the #352 frame in a severe occluded environment, and figure (d) is the #544 frame in the same environment as figure (c). Two personnel are in the image, and proposed method accurately detects and tracks the 2 personnel. The experimental result shows proposed method with integrated metric learning and kalman filter can effectively detect and track multiple people in the substation.

![Figure 5. Tracking result in the substation.](image)
Proposed method realizes multi-target tracking online, in order to evaluate the algorithm, we test following indicators of the proposed method: multiple target tracking algorithm evaluation indexes: multiple object tracking accuracy (MOTA) multiple object tracking precision (MOTP), the total number of identity switches (IDS) and Frames per second (FPS). The performance results of different multi-target trackers in a less occlusion environment and a heavy occlusion environment are listed in Table II and Table III. “↑” means the higher the score, the better; “↓” means the lower the score, the better.

Table 2 and Table 3 represent the extensive experiments with video sequences under occlusion with two degrees, which shows our method is effective and efficient in multiple object tracking in not only less occlusion but also heavy occlusion scenes. It is accurate yet highly computationally effective.

### Table 2. The performance result of different trackers under light occlusion.

| Tracker  | MOTA (↑) | MOTP (↑) | IDS (↓) | FPS |
|----------|----------|----------|---------|-----|
| OVBT [19]| 70.1     | 81.4     | 7       | 7   |
| POI [20] | 79.8     | 83.2     | 5       | 10  |
| EAMTT [21]| 70.5   | 81.1     | 8       | 8   |
| Our Method | 80.6  | 85.4     | 3       | 20  |

### Table 3. The performance result of different trackers under heavy occlusion.

| Tracker  | MOTA (↑) | MOTP (↑) | IDS (↓) | FPS |
|----------|----------|----------|---------|-----|
| OVBT [19]| 57.3     | 70.9     | 8       | 7   |
| POI [20] | 63.1     | 73.4     | 7       | 10  |
| EAMTT [21]| 59.6   | 70.1     | 9       | 8   |
| Our Method | 72.1  | 78.2     | 4       | 20  |

4. Conclusion
This paper proposed an integrated metric learning based multiple object tracking method dealing with occlusion in substations. Experiments show that the method can deal with not only less but also heavy occlusion in the substations. Moreover, it can track multiple personnel with long-duration and complete occlusion. While the method is highly computationally cost effective and accurate. Future work includes merging image processing, in order to study multiple personnel tracking under the poor visual conditions in the substations.

References
[1] Y. Liu, J. Liu, and Y. Ke. "A detection and recognition system of pointer meters in substations based on computer vision." Measurement, vol. 152, 2020.
[2] J. Li, H. Wang, Y. Zhao, R. Huang, and S. Yang. "Application Research of Artificial Intelligent Technology in Substation Inspection Tour." In 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). May 2019.
[3] P.M. Raju, D. Mariam, and R.K.S.S Gorthi. "Detection based long term tracking in correlation filter trackers." Pattern Recognition Letters, vol. 122, 2019.
[4] J. Nigam, K. Sharma, and R. M. Rameshan. "Detection-Based Online Multi-target Tracking via Adaptive Subspace Learning" In International Conference on Smart Multimedia (ICSM). August 2018.
[5] R. D. Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. "The mahalanobis distance." Chemometrics and intelligent laboratory systems, vol. 50, no.1, 2000.
[6] Y. Xu, X. Zhou, S. Chen, and F. Li. "Deep learning for multiple object tracking: a survey." IET Computer Vision, vol.13, no.4, 2019.
[7] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Liu. "Visual object tracking using adaptive correlation filters." In 23rd IEEE Conference on Computer Vision and Pattern
Recognition, (CVPR 2010), June 2010.

[8] M. Danelljan, G. Hager, F. Khan, and M. Felsberg. "Accurate scale estimation for robust visual tracking." In British Machine Vision Conference, September 2014.

[9] X. Gao and T, Jiang. "Osmo: Online specific models for occlusion in multiple object tracking under surveillance scene." In Proceedings of the 26th ACM international conference on Multimedia. October 2018.

[10] Y. Z. Cheong and W. J. Chew. "The Application of Image Processing to Solve Occlusion Issue in Object Tracking." In 9th Eureca 2017 International Engineering Research Conference. February 2018.

[11] Z. Tang and J. N. Hwang. "MOANA: An online learned adaptive appearance model for robust multiple object tracking in 3D." IEEE Access, vol.7, 2019.

[12] K. M. Judd, J. D. Gammell, and P. Newman. "Multimotion visual odometry (mvo): Simultaneous estimation of camera and third-party motions." In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2018.

[13] W. Kim. "Multiple object tracking in soccer videos using topographic surface analysis." Journal of Visual Communication and Image Representation, vol.65, 2019.

[14] J. Hu, J. Lu, and Y. P Tan. "Deep metric learning for visual tracking." IEEE Transactions on Circuits and Systems for Video Technology, vol.26, no.11, 2015.

[15] X. Zhao, Z. Xu, B. Zhao, X. Chen and Z. Li. "Object Tracking With Structured Metric Learning." IEEE Access, vol.7, 2019.

[16] M. Ju, H. Luo, Z. Wang, B. Hui and Z. Chang. "The application of improved YOLOv3 in multi-scale target detection." Applied Sciences, vol.9, no.18, 2019.

[17] F. Ebert and H. J. Wuensche. "Dynamic Object Tracking and 3D Surface Estimation using Gaussian Processes and Extended Kalman Filter." In 2019 IEEE Intelligent Transportation Systems Conference (ITSC). October 2019.

[18] E. Hamuda, B. McGinley, and M. Glavin. "Improved image processing-based crop detection using Kalman filtering and the Hungarian algorithm." Computers and Electronics in Agriculture, vol.148, 2018.

[19] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horard. "Tracking multiple persons based on a variational bayesian model." In 14th European Conference on Computer Vision (ECCV), October 2016.

[20] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan. "Poi: Multiple object tracking with high performance detection and appearance feature." In 14th European Conference on Computer Vision (ECCV), October 2016.

[21] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. "Online multi-target tracking with strong and weak detections." In 14th European Conference on Computer Vision (ECCV), October 2016.