Training configuration analysis of a convolutional neural network object tracker for night surveillance application

Zulaikha Kadim[1], Mohd Asyraf Zulkifley[2], Nor Azwan Mohamed Kamari[3]
[1,2,3]Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, Bangi 43650, Malaysia
[1]MIMOS Berhad, Technology Park Malaysia, 57000 Bukit Jalil, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Jan 18, 2020
Revised Apr 3, 2020
Accepted Apr 23, 2020

Keywords:

Convolutional object tracking
Night surveillance
Tracking algorithm
Training data analysis

ABSTRACT

Automated surveillance during the night is important as it is the period when crimes usually happened. By providing continuous monitoring, coupled with a real-time alert system, appropriate action can be taken immediately if a crime is detected. However, low lighting conditions during the night can degrade the quality of surveillance videos, where the captured images will have low contrast and less discriminative features. Consequently, these factors contribute to the problem of bad appearance representation of the object of interest in the tracking algorithm. Thus, a convolutional neural network-based object tracker for night surveillance is proposed by exploiting the deep feature strength in representing object features spatially and semantically. The proposed convolutional network consists of six layers that consist of three convolutional neural networks (CNN) and three fully connected (FC) layers. The network will be trained by using a binary classifier approach of objects and its background classes, which is updated on a fixed interval so that it fully encapsulates the changes in object appearance as it moves in the scene. The algorithm has been tested with different sets of training data configurations to find the best optimum ones with regards to VOT2015 evaluation protocols, tested on 14-night surveillance videos. The results show that the configuration of a total of 250 training samples with a sample ratio of 4:1 between positive and negative data delivers the best performance for the sequence length of [1,550]. It can be inferred that more information on the object is required compared to the background, where the background might be homogeneous due to low lighting conditions. In conclusion, this algorithm is suitable to be implemented for night surveillance application.

Corresponding Author:

Mohd Asyraf Zulkifley,
Department of Electrical, Electronic and Systems Engineering,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, Bangi 43650, Malaysia.
Email: asyraf.zulkifley@ukm.edu.my

1. INTRODUCTION

Visual object tracking has been an active research area in computer vision and image processing fields because of its wide application in numerous real-world problems including visual surveillance [1], robotics [2], human-computer interactions [3], traffic analysis [4], physiotherapy [5], and autonomous vehicles [6]. In this paper, we propose an object tracking algorithm based on a convolutional neural network (CNN) approach for the night surveillance application. The role of the object tracker is to estimate movement trajectories throughout image sequences by assigning a consistent label to the tracked object. The tracker task is also to provide object-centric information such as orientation, area or shape of the tracked object. This tracking information serves as low-level input for higher-level applications such as behavior analysis of the object of interest and condition monitoring of physiotherapy patients. There are numbers of tracking algorithms that have been proposed in the literature, however, most of them are developed for the application in a bright environment, with little emphasis on dark surroundings.

Videos from night security cameras pose a more difficult challenge to the standard object tracking algorithms. Due to low lighting conditions during the night time, the quality of captured images is normally not that good with characteristics of low brightness, low contrast, and almost no distinguishable color information. This condition is even worse if the target object is small in size [7].

One of the approaches in handling tracking performance for the night sequence is by enhancing the image quality first before detection and tracking are done. This enhancement can be achieved by applying preprocessing steps such as histogram equalization, histogram specification, and intensity mapping. Huang et al. [8] analyze the object's local contrast changes to improve object detection accuracy in the night video application. Local contrast is computed by finding the ratio between the local standard deviation of image intensity with local mean intensity, which is the basis for Huang's Contrast Change (CC) model. Objects are then detected by thresholding the contrast change values in the successive frames. The computation speed is relatively fast; however, it is prone to the problem of similar appearance between the object and its surrounding background. Huang et al. [9] further improve the detection accuracy by utilizing motion prediction and spatial nearest neighbor data association. Wang et al. [10] further improve Huang’s CC model by introducing a salient contrast change (SCC) model, which requires online learning and analysis of the detected object trajectories. The work in [11] introduces illumination invariant representation by multiplying Shahnon’s entropy estimation with their own contrast estimation.

In general, tracking requires the target to be represented by a model that might include information about the shape or appearance of the object. The model will be used as a reference in finding the most probable location of the object in the next frame. Object appearance can be represented using global or local features, in which some are more suitable for certain tracking challenges such as illumination variation, background clutter, and occlusion. Yang et al. [12] and Li et al. [13] provide a good overview of local and global feature representations for tracking purposes and a summary of target appearance models, respectively. A good object representation method should be able to clearly distinguish the target from other background objects.

The authors of [14] categorize tracking algorithms based on feature representation methods, in which they are divided into two groups namely handcrafted and deep features. Effective feature representations should be discriminative while maintaining the geometric, structural and spatial target information. Structural, geometrics and spatial target information encode the appearance variation, shapes, and location of different object parts, respectively. Some of the handcrafted features capture this low-level information but encode only a small fraction of semantic information. Deep features, on the other hand, are able to encode low-level spatial and high-level semantic information which are essential components in locating the objects precisely. These abilities make deep learning approach popular in many image processing tasks; including object detection [15-17], classification [18], and tracking.

In [19], Jong et al. propose a CNN based human detection, which serves as an input to an object tracker for the night time conditions. The input images are resized to 183x119 and its histogram representation is equalized first before passing the images to the CNN model to improve human detection accuracy. In [20], Ham and Han propose an online tracking framework based on multi-domain representations. Their network architecture consists of multiple shared layers known as domain-independent layers and classification layers which is known as domain-specific ones. Domain independent layers are trained offline using multiple annotated video sequences, while classification layers are trained separately based on the specific new image sequences. In [21], multiple CNNs is maintained in a tree structure to represent multi-modal target appearance. A general tracking framework for thermal infrared videos has also been proposed in [22]. Thermal images have the most similar features to the night surveillance images, spefically in terms of low contrast information and negligable textures. Multiple CNNs approach to model the target appearance in different cases is also proposed. During network updates, parent nodes will be replaced by the new node so that there is no redundancy in the pool of target object appearance models. In [23], a Siamese approach is utilized in which pair of patches are compared to find the most likely location of the target object.

In view of that, this paper proposes a convolutional neural network-based object tracker algorithm for night surveillance application by exploiting the strength of deep features in representing object appearance. Various training data configurations are also examined to find the best setting for the proposed tracker. The paper is organized as follows: Section 2 summarizes the proposed tracker algorithm for night surveillance, whose performance is evaluated and discussed in Section 3. Finally, conclusion is given in Section 4.

2. CONVOLUTIONAL NEURAL NETWORK OBJECT TRACKER

The proposed convolutional object tracker algorithm is based on the concept of tracking by detection, whereby the tracked object location is inferred based on the best matched samples in each frame. In this work, the detection is achieved by constructing an object model to classify either the input samples belong to the background or object of interest. The top five samples that are best classified to be the object of interest are used to construe the final object location. The overall flow of the tracker algorithm is illustrated in Figure 1.

During the first frame, the model will be trained using sampled candidates that are generated based on the initial user-specified object location. N number of positive and m number of negative samples are generated randomly around the initial object location, governed by a certain percentage of overlapping area between the sample and the initial object location box. These annotated samples are then used to train the first object appearance model. The network architecture consists of three convolutional neural network (CNN) and three fully connected (FC) layers. This model architecture will be further discussed in section 3.1.

For the subsequent frames, previously known object location will be used as the pivot point for generating sample candidates. This approach assumes that the object in the current frame does not move too far from the previous known location. These samples are then matched with the trained model by classifying each sample as background or object of interest. The weighted average of the boundary points of the top-5 samples, which have beed classified as the object is then used to generate the smoothed object location in the current input frame. Finally, the model will be updated periodically to capture the object's appearance changes as it moves around.

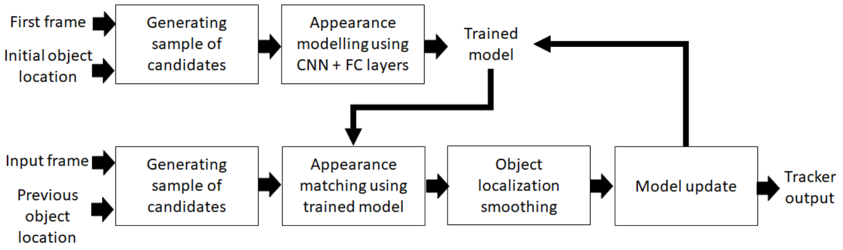
First frame
Initial object location
Generating sample of candidates
Appearance modelling using CNN + FC layers
Trained model
Input frame
Previous object location
Generating sample of candidates
Appearance matching using trained model
Object localization smoothing
Model update
Tracker output

Figure 1. Algorithm flow of the proposed convolutional object tracker

2.1. Network architecture

Figure 2 illustrates the network architecture used in the proposed tracker. The network consists of three CNN and three FC layers. Table 1 details out the number of filters, filter size, stride and padding of the receptive field in each CNN layer. The size of the input image is set to 75x75, thus the output from the conv3 layer is a flat layer of 512x1x1 feature map. The final output of the network is a binary classification probability that the input image belongs to the object-of-interest or background. The weights of the three CNN layers are ported directly from trained convolutional weights of the first 3 CNN layers of the Vgg-m model [24] that were trained on the ImageNet ILSVRC-2012 dataset [25]. Vgg-m model was first introduced by [24] as CNN-m that consists of five CNN layers and three FC layers as shown in Table 2. Thus, there is a clear difference between ours and the Vgg-m network in terms of the number of CNN layers (3 vs. 5). Apart from that, the convolutional stride and spatial padding in conv2 and conv3 layers also different compared to Vgg-m, as well as less dimensionality for fc1 to fc3 (512 vs. 4096).

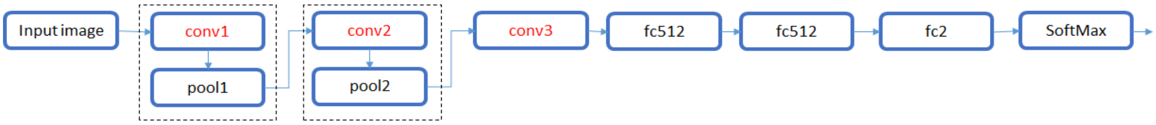

Figure 2. Network architecture of the convolutional object tracker with three CNN layers and three FC layers

Table 1. Number of filters, filter size, stride and padding of receptive fields in each layer of proposed model

Layer	Conv1	Pool1	Conv2	Pool2	Conv3
Number of filters	96	96	256	256	512
Filter size	7x7	3x3	5x5	3x3	3x3
Stride	2	2	2	2	2
Padding	0	0	0	0	0

Table 2. VGG-M architecture as proposed in [22]

Arch	Conv1	Conv2	Conv3	Conv4	Conv5	Full6	Full7	Full8
CNN-M	96x7x7 st. 2, pad 0 LRN, x2 pool	256x5x5 st. 2, pad 1 LRN, x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1, x2 pool	4096 dropout	4096 dropout	1000 Softmax

2.2. Samples generation

The proposed tracker algorithm tries to find the best candidate representations of the tracked object from a set of samples using a deep-learning model that acts as a binary classifier. Positive and negative candidates refer to the sub-image that corresponds to a tracked object or background area, respectively. These samples will then serve as the training data used to train the network in a supervised manner. For the initial training phase in the first frame, more samples are generated since the actual object location is known to be true. However, during online network updates, a lesser number of samples are created to reduce the computation burden, as well as to minimize the possibility of background inclusion in the model.

To generate the samples, object location in the previous frame is used as the pivot point to spawn positive and negative candidate locations. To ensure the variability of the candidates, the samples are spawned randomly according to Gaussian distribution, whereby the samples are concentrated closer to the pivot point. The generated positive sample should have at least 80% overlapping areas with the previously known object area, while negative samples should also have an overlapping area with at most 20%. Sample size is one of the most important factors in training a deep-learning model. A higher number of training samples is better for generalizing the model. Thus, in this work, the selection of the number of samples and the ratio between positive and negative samples are analyzed to determine the best combination for the proposed object tracker algorithm.

2.3. Network learning and update

Training of the network in the first frame will dictate how well the tracker will perform in the subsequent frames. The three CNN layer's weights are transferred from the pre-trained weights of the first 3 CNN layers in the Vgg-m network, while the weights and biases in the FC layers are initialized randomly and fixed to 0.05, respectively. Only these FC layers will go through learning stage using positive and negative samples generated from our specific testing data.

Choosing a learning rate for optimal learning is a challenging task as it may cause slow convergence or frequent fluctuation of the loss function, which will lead to the model divergent. Thus, in this work, we have selected Adam optimizer [26] as the backpropagation optimization algorithm. Adam is an adaptive moment estimation that computes different learning rates for different parameters (i.e. model weights and biases) by using the estimates of their first and second-order moments of the gradient. The final model parameter updates will be based on the newly adjusted learning rate, which equals to initial learning rate multiplies with the ratio of its first and second-order moments of the gradients. The initial learning rate is set as 0.00075 based on the finding in [27], and Adam's hyperparameters exponential decay rate for the first-order (β1), exponential decay rate for the second-order (β2) and small values constants to prevent division by zero (ε), which are set to 0.9, 0.999 and 1e-08 respectively. Adam has been gaining popularity since its introduction in 2015, where it has been widely used in others network learning [28].

It is important to update the network to ensure that the model captures the changes in object appearance as it moves around under different lighting conditions and background scenes. In this work, our model will be updated at a fixed interval of every 10 frames. Positive and negative samples that are generated in the period of 10 frames will be accumulated and used to retrain the model. However, the training is considered as weak supervision as the true object location is not known and it may lead to background inclusion in the model.

3 RESULTS AND DISCUSSION

3.1. Experimental setup

14-night surveillance videos are used to evaluate the proposed object tracker algorithm, which has been recorded from three different surveillance cameras that cover challenging night scenarios of low lighting, low contrast between object and background, small object size, object occlusion, and move-stop-move object movement. Some sample images of these scenarios are shown in Figure 3. The resolution of the video is only 352x288 with the object of interest size of approximately 30x70 pixels. Video length varies from 33 to 550 frames each. Groundtruths are generated by an expert in computer vision using an annotation tool to draw the bounding box that surrounds the object in each frame for each video. The tracker code is developed in Python with Tensorflow as the main library to perform model training and testing. Seven different combination of training samples size and ratios are configured for tracker evaluation as listed in Table 3.

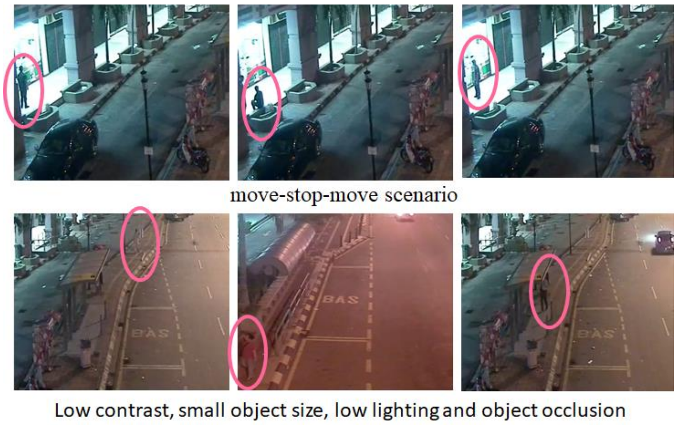
move-stop-move scenario
BAS
Low contrast, small object size, low lighting and object occlusion

Figure 3. Sample images with challenging night scenario in testing dataset

Table 3. Training sample configuration for tracker evaluation

Configuration	1:1	1:2	1:3	1:4	2:1	3:1	4:1
Positive sample size	100	50	50	50	100	150	200
Negative sample size	100	100	150	200	50	50	50
Total sample size	200	150	200	250	150	200	250

3.2. Performance evaluation measure

Three VOT2015 [29] (Visual Object Tracking) evaluation metrics are used to quantify the performance of our tracker. These metrics are accuracy (*Acc*), robustness (*Ro*) and expected area overlap (*EAO*). Accuracy and robustness require the tracker to be re-initialized once it drifts off the target. Accuracy measures how well the tracked bounding box relative to the ground truth box by computing the intersection over union (*IOU*). The higher the *IOU*, the better the tracking accuracy is. On the other hand, robustness measures the number of tracking failures in a video, which is triggered when zero *IOU* occurs. To reduce the bias in robustness measurement, the tracker is re-initialized five frames after the failure, while to reduce bias in accuracy calculation, the accuracy values from the first 10 frames after the re-initialization process are ignored from overall performance computation [30].

EAO does not require re-initialization of the tracker and it is used to rank the tracker. It averages the *IOU* over a range of frames between lower (*Nlo*) and upper limit (*Nhi*) of the testing sequence. In this work, the lower and upper limit is based on the range of [1, 550]. The *Acc*, *Ro* and *EAO* calculation are as follows:

$$\text{Accuracy, } Ac = \frac{1}{\varphi}\sum_{i=1}^{\varphi}\frac{s_{i,output}\cap s_{i,gt}}{s_{i,output}\cup s_{i,gt}} \quad (1)$$

where φ denotes the number of frames in the test video, while $s_{i,output}$ and $s_{i,gt}$ are the bounding boxes of object in frame *i* from the tracker output and ground truth, respectively.

$$\text{Robustness, } Ro = \sum_{i=1}^{\varphi} F^{i} \quad (2)$$

where $F^i = \begin{cases} 0\ if\ IOU > 0 \\ 1\ if\ IOU \leq 0 \end{cases}$

$$\text{Expected average overlap, } EAO = \frac{1}{N_{hi}-N_{lo}} \sum\nolimits_{N_s=N_{lo}:N_{hi}} IOU^i \qquad (3)$$

3.3. Results

Accuracy and robustness values for different training configurations are summarized in Table 4 and Table 5, respectively. From Table 4, there is no single configuration works the best for all test sequences. All different configurations perform almost relatively the same in terms of accuracy, however in terms of robustness, configuration 1:4 shows significantly better performance compared to the others. For no-reinitialization protocol evaluation, the *EAO* curve is shown in Figure 4. The curve shows that the configuration of 4:1 works consistently better than other configurations approximately after the first 150 frames, followed by configuration of 1:3, thus makes both configurations ranked as the top two as demonstrated in Table 6. This explains that different training sample sizes and ratios affect tracking performance. The tracker performs better when there are more positive samples used to represent the object compared to the negative samples. This indicates that although the object-background contrast is low, it also looks homogeneous due to low lighting, thus a lower number of samples are sufficient to represent them.

Table 4. Accuracy results of different training sample configuration per each test sequence

No	Method	Number of frames	1:1	1:2	1:3	1:4	2:1	3:1	4:1
1	Cam01-video01	271	0.649	0.618	0.565	0.580	0.159	0.165	**0.651**
2	Cam01-video02	454	0.562	0.452	0.529	**0.595**	0.585	0.576	0.528
3	Cam01-video03	70	0.142	0.213	0.088	**0.284**	0.143	0.159	0.065
4	Cam01-video04	130	**0.650**	0.622	0.599	0.573	0.604	0.583	0.598
5	Cam01-video05	33	0.680	0.751	0.673	0.616	0.773	0.593	**0.777**
6	Cam01-video06	224	0.560	0.605	0.619	0.538	0.620	0.614	**0.627**
7	Cam01-video07	460	0.153	0.101	0.401	0.121	0.445	0.108	**0.637**
8	Cam01-video08	124	0.548	0.314	0.517	**0.562**	0.420	0.237	0.208
9	Cam02-video01	400	0.257	**0.463**	0.259	0.332	0.343	0.326	0.350
10	Cam02-video02	459	0.535	0.466	**0.570**	0.540	0.355	0.502	0.563
11	Cam03-video01	343	0.262	0.552	0.547	**0.564**	0.437	0.498	0.182
12	Cam03-video02	150	0.410	0.169	0.217	**0.529**	0.368	0.547	0.413
13	Cam03-video03	150	0.644	0.611	0.622	**0.682**	0.585	0.587	0.595
14	Cam03-video04	550	**0.521**	0.492	0.497	0.339	0.507	0.461	0.465
	Average accuracy		0.469	0.459	0.479	**0.490**	0.453	0.425	0.476

Table 5. Raw robustness results of different training sample configuration per each test sequence

Configuration	1:1	1:2	1:3	1:4	2:1	3:1	4:1
Total samples	200	150	200	250	150	200	250
Average robustness	6.643	6.214	6.071	**3.929**	7.929	8.857	6.071

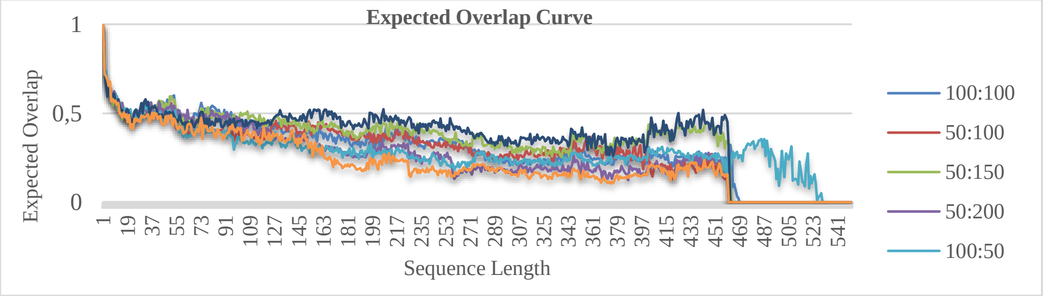

Figure 4. Expected average overlap curve for different training sample configuration tested on all test sequences

Table 6. Expected average overlap results and the tracker rank of different training sample configuration using a range of [1,550]

Training sample configuration	1:1	1:2	1:3	1:4	2:1	3:1	4:1
EAO	0. 296745	0. 288171	0. 34047	0. 253524	0. 286333	0. 220636	0. 360529
Rank	3	4	2	6	5	7	1

4 CONCLUSION

In this paper, we have proposed a convolutional object tracker for the night video surveillance. The tracker requires the model to learn the object's appearance and its background properties as the object moves around in the scene. The algorithm was tested with different training data configurations to find the most suitable ones with the highest accuracy and robustness. The results show that the ratio of 4:1 between positive and negative samples produces the best performance for the sequence range of [1,550], followed by the 1:3 configuration. Both configurations also produce the second-best performance in terms of tracking robustness. This explains that different training sample sizes and ratios affect tracking performance differently and as more positive samples are used to represent the object compared to the negative samples, the better the tracker performance is. This indicates that although the object-background contrast is low, the background is also homogeneous due to low lighting condition, thus a lower number of samples are sufficient to represent them.

ACKNOWLEDGEMENTS

The authors would like to acknowledge fundings from Ministry of Education Malaysia (FRGS/1/2019/ICT02/UKM/02/1) and Universiti Kebangsaan Malaysia (GUP-2019-008). The graphic processing unit used in this research is donated by NVIDIA Corporation (KK-2019-005).

REFERENCES

[1] A. Ali, A. Jalil, J. Niu, X. Zhao, S. Rathore, J. Ahmed, and M. A. Ikhar. “Visual object tracking: classical and contemporary approaches,” in *Frontiers of Computer Science: Selected Publications from Chinese Universities,* vol. 10, no. 1, pp. 167-188, 2016.

[2] L. Zhang, C. Lim, Y. Chen, and H. R. Karimi, "Tracking Mobile Robot in Indoor Wireless Sensor Networks," in *Mathematical Problems in Engineering*, Volume 2014 https://doi.org/10.1155/2014/837050

[3] J. Severson, “Human-digital media interaction tracking”, *US Patent* 9,713,444, 2017

[4] P. R Iyer, S. R. Iyer, R. Ramesh, Anala, K. N. Subramanya, *“Adaptive real time traffic prediction using deep neural networks,”* in *IAES International Journal of Artificial Intelligence* (IJ-AI), vol. 8, no. 2, June 2019, pp. 107-119.

[5] M. A. Zulkifley, N. A. Mohamed, and N. H. Zulkifley, “Squat Angle Assessment Through Tracking Body Movements,” in *IEEE Access,* Vol. 7, pp. 48635-48644, 2019.

[6] V. A Laurense, J. Y Goh, and J C. Gerdes, “Path-tracking for autonomous vehicles at the limit of friction,” In *American Control Conference*, pp. 5586-5591, 2017.

[7] M. A. Zulkifley, N. S. Samanu, N. Zulkepeli, Z. Kadim, and H. H. Woon, “Kalman filter-based aggressive behaviour detection for indoor environment,” in *Lecture Notes in Electrical Engineering*, Vol. 376, pp. 829-837, 2016.

[8] K. Huang, L. Wang, and T. Tan, “Detecting and tracking distant objects at night based on human visual system,” *Lect. Notes Comput. Sci.*, vol. 3852 LNCS, pp. 822-831, 2006.

[9] K. Huang, L. Wang, T. Tan, S. Maybank, “A real-time object detecting and tracking system for outdoor night surveillance,” in *Pattern Recognition*, vol. 41, np. 1, pp. 432-444, 2008.

[10] L. Wang, K. Huang, Y. Huang and T. Tan, "Object detection and tracking for night surveillance based on salient contrast analysis," *IEEE International Conference on Image Processing,* pp. 1113-1116, 2009.

[11] A. Nazib, C. Oh and C. W. Lee, "Object detection and tracking in night time video surveillance," *10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 629-632, 2013.

[12] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, “Recent advances and trends in visual tracking: A review,” in *Neurocomputing,* vol. 74, no. 18, pp. 3823-3831, 2011.

[13] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V Den Hengel, “A survey of appearance models in visual object tracking,” in *ACM Transactions on Intelligent Systems and Technology,* vol. 4, no. 4, 2013.

[14] M. Fiaz, A. Mahmood, S. Javed, and S. K. Jung, ‘‘Handcrafted and deep trackers: Recent visual object tracking approaches and trends,’’ in *ACM Computing Survey,* vol. 52, no. 2, pp. 1-43, 2019.

[15] C. Szegedy, A. Toshev, and D. Erhan, “Deep Neural Network for Object Detection,” *Neural Inf. Process. Syst.*, vol. 7, no. 5, pp. 200-205, 2013.

[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, 2017.

[17] M. A. Zulkifley, S. R. Abdani and N. H. Zulkifley, "Pterygium-Net: a deep learning approach to pterygium detection and localization," in *Multimedia Tools and Applications*, vol. 78. no. 24, pp. 34563-34584, 2019.

[18] S. B. Jadhav, V. Udupi, S. Patil, "Convolutional neural networks for leaf image-based plant disease classification," *IAES International Journal of Artificial* Intelligence (IJ-AI), vol. 8, no. 4, pp. 328-341, 2019.

[19] J. H. Kim, H. G. Hong, K. R Park, "Convolutional Neural Network-Based Human Detection in Nighttime Images Using Visible Light Camera Sensors", in *Sensors* (Switzerland), vol. 17. no. 5, pp. 1-26, 2017.

[20] H. Nam and B. Han, ''Learning multi-domain convolutional neural networks for visual tracking,'' *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 4293-4302, 2016.

[21] H. Nam, M. Baek, and B. Han. ''Modeling and propagating CNNs in a tree structure for visual tracking,'' in *ArXiv CoRR,* vol. abs/1608.07242, pp. 1-10, 2016

[22] M. A. Zulkifley and N. Trigoni, "Multiple-Model Fully Convolutional Neural Networks for Single Object Tracking on Thermal Infrared Video," in *IEEE Access*, vol. 6, pp. 42790-42799, 2018.

[23] M. A. Zulkifley, "Two Streams Multiple-Model Object Tracker for Thermal Infrared Video", in *IEEE Access*, vol. 7, pp. 32383-32392, 2019.

[24] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Networks", in *British Machine Vision Conference,* pp. 1-12, 2014.

[25] O. Russakovsky et al., ''ImageNet large scale visual recognition challenge'', in *International J. Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.

[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ArXiv CoRR*, vol. arXiv:1412.6980, pp. 1-15, 2014.

[27] Z. Kadim, M. A. Zulkifley, N. Hamzah, "Deep-learning based single object tracker for night surveillance," in *IAES International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, 2020.

[28] Y. Wu, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," in *ArXiv CoRR*, vol. abs/1609.08144, pp. 1-23, 2016

[29] M. Kristan and others. "The Visual Object Tracking VOT2015 Challenge Results," in *Visual Object Tracking Workshop ICCV 2015*, 2015.

[30] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, and T. Vojir. "The VOT2013 challenge: overview and additional results," In *Computer Vision Winter Workshop*, 2014.

BIOGRAPHIES OF AUTHORS

Zulaikha Kadim received her B.Eng degree in electronics engineering majoring in telecommunication from Multimedia University, Malaysia in 2000 and currently persuing her Ph.D degree in Universiti Kebangsaan Malaysia. She is also a researcher in national R&D institution. Her current research interests are in video analytics, behavioral analysis and visual object tracking.

Mohd Asyraf Zulkifley (M'18) received the B.Eng. degree in mechatronics from International Islamic University Malaysia in 2008 and the Ph.D. degree in electrical and electronic engineering from The University of Melbourne in 2012. He was a sponsored Researcher at the Department of Computer Science, University of Oxford from 2016 to 2018. He is currently an Associate Professor at the Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia. His current research interests are visual object tracking and medical image analysis.

Nor Azwan Mohammed Kamari received B.Eng. from Meiji University, Japan, M.Sc. degrees from Ehime University, Japan and PhD degree from Universiti Teknologi Mara, Malaysia 2000, 2004 and 2016, respectively. He has authored and co-authored more than 30 technical papers published in the international journals, national and international conferences. His research interests include power system stability and artificial intelligence.