Smart Structures and Systems, *Vol. 28, No. 6 (2021) 811-825*
https://doi.org/10.12989/sss.2021.28.6.811

Anomaly detection of isolating switch based on single shot multibox detector and improved frame differencing

Yuanfeng Duan [a], Qi Zhu, Hongmei Zhang*, Wei Wei and Chung Bang Yun

College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China

(*Received April 22, 2021, Revised August 9, 2021, Accepted August 16, 2021*)

Abstract. High-voltage isolating switches play a paramount role in ensuring the safety of power supply systems. However, their exposure to outdoor environmental conditions may cause serious physical defects, which may result in great risk to power supply systems and society. Image processing-based methods have been used for anomaly detection. However, their accuracy is affected by numerous uncertainties due to manually extracted features, which makes the anomaly detection of isolating switches still challenging. In this paper, a vision-based anomaly detection method for isolating switches, which uses the rotational angle of the switch system for more accurate and direct anomaly detection with the help of deep learning (DL) and image processing methods (Single Shot Multibox Detector (SSD), improved frame differencing method, and Hough transform), is proposed. The SSD is a deep learning method for object classification and localization. In addition, an improved frame differencing method is introduced for better feature extraction and a hough transform method is adopted for rotational angle calculation. A number of experiments are conducted for anomaly detection of single and multiple switches using video frames. The results of the experiments demonstrate that the SSD outperforms the You-Only-Look-Once network. The effectiveness and robustness of the proposed method have been proven under various conditions, such as different illumination and camera locations using 96 videos from the experiments.

Keywords: anomaly detection; high voltage isolating switch; Hough Transform; improved frame differencing; object tracking; rotating angle; single shot multibox detector; vision-based method

1. Introduction

With the rapid increase in electricity consumption, ensuring a reliable power supply becomes a paramount task in society. High-voltage isolating switches are important equipment for power supply systems. Thus, the anomaly detection of isolating switches is necessary to improve the safety and reliability of power supply systems. Recent studies have shown great potentials of computer vision and image processing methods for anomaly or movement detection of industrial equipment and civil structure (Wang *et al*. 2009, Zhu *et al*. 2010, Ye *et al*. 2016, Yu *et al*. 2017, Li *et al*. 2019). Advanced auxiliary equipment, such as unmanned aerial vehicles, have also been adopted for the same purpose (Guo *et al*. 2020, Mondal and Jahanshahi 2020). Traditional methods such as manual inspection are generally time-consuming and dangerous. Consequently, many researchers have been working on image-based techniques to detect faults in electrical facilities such as power lines and insulators (Siddiqui *et al*. 2018, Lei and Sui 2019).

The fault detection of isolating switches has focused on the static state of switches. Image processing techniques based on color and shape features of the switches have been commonly used (Shi *et al*. 2007, Chen *et al*. 2012, Zhao *et al*. 2016, Fang *et al*. 2017). However, these techniques mainly depend on manually extracted features, and only the static open and closed states of the switches can be detected. With the development of artificial intelligence techniques, machine learning (ML) and deep learning (DL) methods have been gradually adopted in this field (Nassu *et al*. 2018, Wanyan *et al*. 2019) and achieved relatively high accuracy. Nevertheless, these methods can still only detect the static state of switches. The moving object tracking method was proposed recently (Wang *et al*. 2017) by processing video frames of an isolating switch, and the distance between two knife switches was detected during the opening and closing process. Notwithstanding, this method is unsuitable for real cases due to the limitation of the camera location (the camera needs to be fixed right below the switch and located in the middle of the two knife switch parts), resulting in only a single switch can be detected.

Although many studies have been conducted on the fault detection of switches, most of them are not robust enough and can only detect the static states of the switch, which made them less effective and more expensive in field applications. The development of effective systems for isolating switch movement detection is a challenging task, considering the different locations of switches during the movement and the computational cost for training of the DL network and complex image processing algorithms. Therefore, research on movement detection is inadequate.

*Corresponding author, Associate Professor,
E-mail: 3140102972@zju.edu.cn
[a] Professor, E-mail: ceyfduan@zju.edu.cn

Copyright © 2021 Techno-Press, Ltd.

ISSN: 1738-1584 (Print), 1738-1991 (Online)

Recently, pre-trained DL networks have shown great potential in computer vision tasks such as object detection and tracking. That is the main reason for adopting a single shot multibox detector (SSD) method (Liu *et al*. 2016) and the frame differencing (FD) method (Fei *et al*. 2015, Paul *et al*. 2017) in this study.

Moving object tracking—the movement detection of objects—is an important task in the field of computer vision. At present, various applications, such as traffic surveillance (Jun *et al*. 2018, Sharma *et al*. 2017, Zhao *et al*. 2018), robot visual navigation (Geiger *et al*. 2013), and person tracking (Li *et al*. 2017, 2018, Tang *et al*. 2017, Ran *et al*. 2019), have been put into use. The main studies in moving object tracking can be categorized into four: optical flow (Choi *et al*. 2015), background subtraction (Supreeth and Patil 2018), FD (Fei *et al*. 2015, Paul *et al*. 2017), and ML (Jun *et al*. 2018, Sharma *et al*. 2017, Zhao *et al*. 2018, Geiger *et al*. 2013, Li *et al*. 2017, 2018, Tang *et al*. 2017, Ran *et al*. 2019). Considering the complex characteristics of isolating switches, the optical flow and background extraction methods are generally computationally expensive and not robust to change in illumination.

In this study, an improved frame differencing (IFD) method is presented for improved feature extraction. The conventional FD method conducted using direct frame subtraction operations on inter-frames, and the noise due to camera and background cannot be fully alleviated and may result in wrong feature extraction. In the proposed method, a movement judgment and area filter methods are used to isolate the noise effect in rotational angle detection.

The main contributions of the proposed method are as follows:

(1) DL method is adopted for multiple object detection: The SSD is adopted to effectively localize the multiple knife switch parts of isolating switches in each video frame. And another two object detection networks named YOLO and YOLOv5s are used to compare with the SSD;
(2) IFD method is introduced to eliminate the impact of the noise in the feature extraction: Improvements of the FD method including movement judgment and area filter operations are introduced to eliminate the noise and increase the accuracy in rotational angle detection;
(3) Rotational angle of the knife switch boundary using the Hough transform (HT) as an index for motion detection, so no need for the reference object: Motion detection of the structure can give out more information and achieve high success ratio for anomaly detection. By adopting the rotational angle as the index for motion, reference object is not needed. Also, Hough transform (HT) is a mature method and can tell the small difference of the rotational angles;
(4) The robustness and effectiveness of the system are explored regarding the illumination and camera locations: For anomaly detection of isolating switches, the robustness and effectiveness of the method regarding illumination and camera locations have not been examined before. Also, the background of the experiment is much complex than in other researchers' work as summarized in Table 1, which also shows that only the proposed method can solve all five problems in anomaly detection.

The rest of this paper is organized as follows. Section 2 describes the details of the proposed method combining SSD, IFD, and HT methods. Section 3 presents the experiments, and Section 4 describes the results and discussion for single and multiple isolating switches.

Table 1 Summary of related work on object tracking

Methods applied	Problems solved					Major drawback
	State**	M**	MO**	CB**	Video	
Edge detection (Shi *et al*. 2007)	√	×	×	×	×	The angle of the closed state needs to be formerly detected.
Image processing, SIFT*, HT* (Chen *et al*. 2012)	√	×	×	×	×	Template images of switch need to be prepared and cannot adapt to severe weather.
Ant colony algorithm, image segmentation (Zhao *et al*. 2016)	√	×	×	×	×	The algorithm mainly depends on the quality of the image acquired and can be affected heavily by the background.
Image processing (Fang *et al*. 2017)	√	×	×	×	×	The whole technique depends upon the color threshold.
Feature extraction, AdaBoost classifier (Wang *et al*. 2017)	√	√	×	×	√	The camera needs to be placed right under the switch, and the reference should be selected before.
CNN*, SVM*, descriptors (Nassu *et al*. 2018)	√	×	×	√	×	Cannot detect multiple switches in an image, and only two states are considered.
SSD* (Wanyan *et al*. 2019)	√	×	√	√	√	Only a small dataset used for training and two states of the switches are considered.
SSD*, IFD, HT* (Proposed method)	√	√	√	√	√	The performance of the result depends on camera pixels.

Notes: * SIFT - scale-invariant feature transform; * HT - Hough transform; * CNN - convolutional neural network; SVM* - support vector machine; SSD* - single shot multibox detector; ** State - open and closed states of the switch; **M - movement; **MO - multiple objects; CB** - complex background

Finally, Section 5 presents the conclusions and future works.

2. Proposed method

Fig. 1 shows the isolating switch system with knife switches at two kinds of states (open and closed). Damage to the isolating switches usually occurs during the opening and closing operations and is manifested in the pause of the switches. The proposed method detects such defects by tracking the motion of the knife switches. As to the index for the moving object tracking, the rotational angle of the knife switch is proposed, which can be directly obtained using HT.

The current system for anomaly detection consists of a video camera and computer. The overall operation of the proposed method is illustrated in Fig. 2. The whole procedure can be divided into two parts: SSD network training for object and anomaly detection. For the SSD network training, datasets for training are acquired by the camera and processed using Fast Forward Mepg (FFmpeg) software for segmenting and LabelImg software for labeling. After obtaining a trained SSD network, anomaly detection can be performed. At the first stage, videos of the isolating switches are processed using the trained SSD

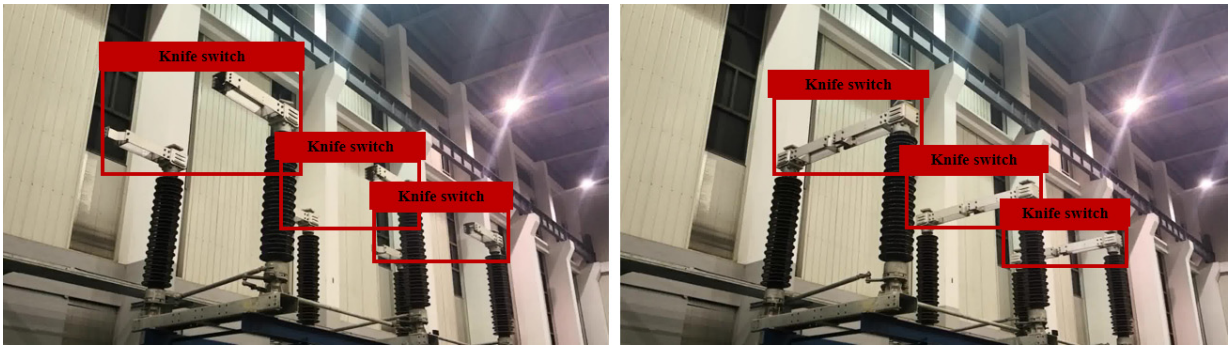


Fig. 1 Knife switch parts in isolating switches (open and closed states)

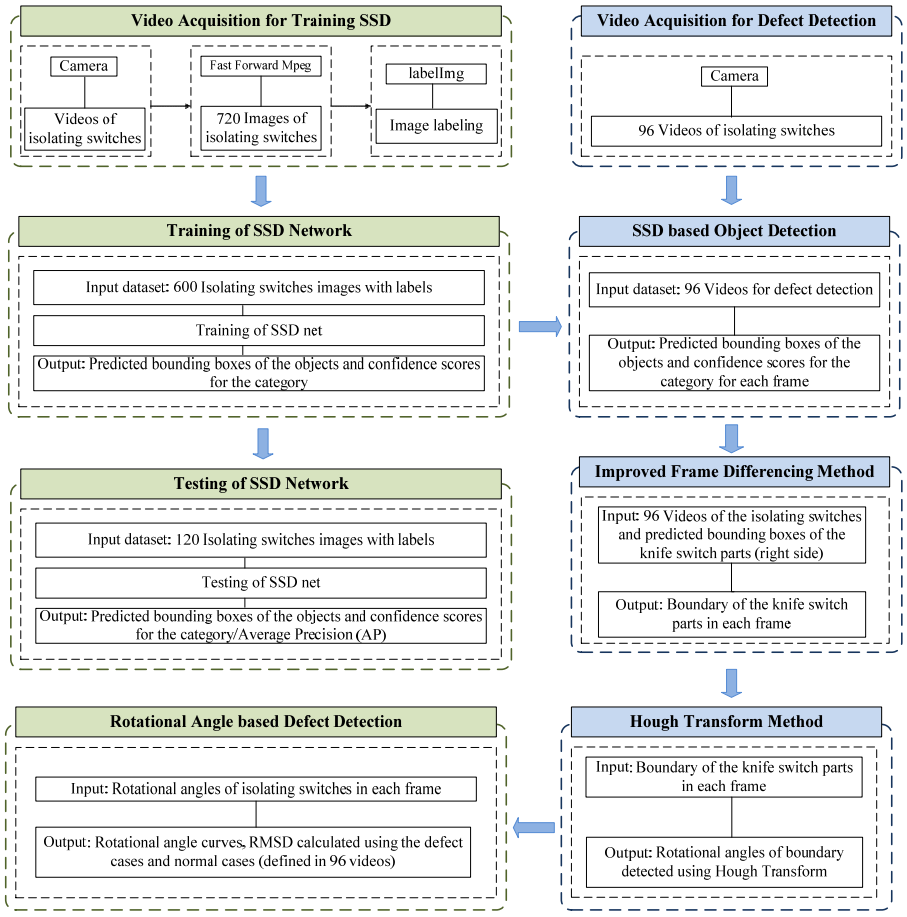


Fig. 2 Schematic of the proposed vision-based method

network. Then, the coordinates of the detected bounding boxes of the knife switch parts are transferred to the IFD module, where the boundary feature of the right-side knife switch is detected at each video frame, and the rotational angle of the knife switch is evaluated using HT. After all rotational angles of the knife switches are obtained during the motion, defects can be recognized using the rotational angle curves and the root mean square deviation (RMSD). In this section, the methods adopted will be briefly introduced.

2.1 SSD-based region of Interest detection

Objects need to be segmented from the background and distinguished from each other in each video frame for later feature extraction. In this study, SSD is adopted for object detection, which is a DL network for visual recognition developed by Liu *et al*. (2016). It was reported that SSD outperformed the fast region-based convolutional neural network (Ren *et al*. 2015) and the You-Only-Look-Once (YOLO) model (Redmon *et al*. 2016) both in accuracy and computing time for training. In this study, the performance of the SSD is compared with YOLO and YOLOv5s (YOLOv5 2020).

The overall procedure of the region of interest (ROI) detection for the knife switch is composed of SSD training and testing. Video segmentation and image labeling (the ground-truth boxes for the knife switch and category) will be done manually in the training and testing dataset preparation. The ground-truth box represents the real location of the knife switch in the image using four parameters (two coordinates of the center, height, and width), and only one category (switch) is labeled, whereas the background category is added automatically during the training process. After the network has been trained, the testing dataset is used to verify its performance. With the trained SSD network, only videos of switches are needed for the knife switch detection regardless of the camera location and illumination condition in the field applications.

The standard structure of the SSD is illustrated in Fig. 3. The input to the SSD are images of size [300, 300] with three channels (pixel values in range 0–255) and labels. The main parts of the SSD are a base network and an auxiliary structure consisting of extra convolution layers and a non-maximum suppression layer (NMS). The image labels are fed to the output of the extra convolution layers for loss calculation during the training. The final outputs of the SSD consist of the predicted bounding boxes and category

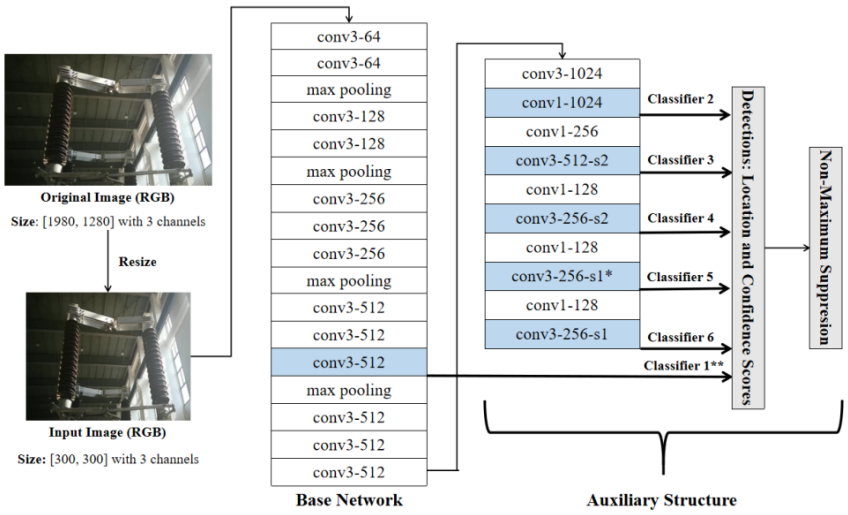


Fig. 3 Structure of the SSD

Notes: * conv3-256-s1 - Convolution operation using 256 kernels, and the kernel size of 3*3 with stride as 1;
** Classifiers 1–6 - Each classifier is composed of two convolution layers for shape offsets and confidence score prediction. Default boxes are also generated for final bounding box generation.
Detailed information for convolution operations in each classifier is illustrated in Table I-1 in Appendix I.

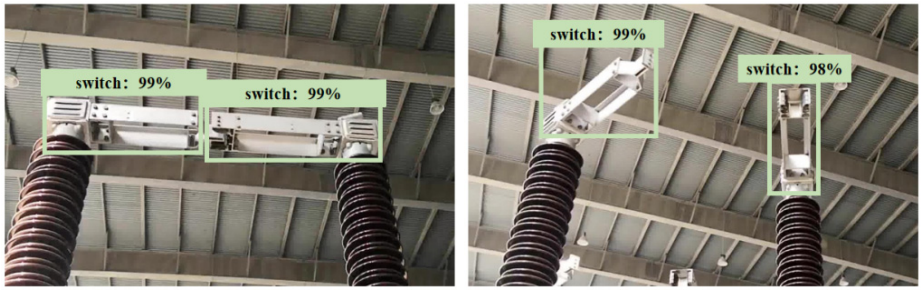


Fig. 4 Outputs of the SSD network for single isolating switch detection

confidence scores, as shown in Fig. 4. The structure of the SSD is described in Appendix I, and the training procedure of the SSD is as shown in Appendix II.

2.2 Improved FD method

2.2.1 Conventional FD method

FD is a method that employs the subtraction operation to adjacent frames in videos to detect the moving object with varying pixel values between the two frames. In this study, the boundary of the knife switch part can be detected in that way.

Before performing the subtraction operation, gray processing is adopted to change 3D RGB images into 2D gray images with pixel values in the range (0–255); the gray processing algorithm is as follows

$$I_t(x,y) = 0.299 \times R(x,y) + 0.587 \times G(x,y) + 0.114 * B(x,y) \quad (1)$$

where $I_t(x,y)$ is the gray value at point (x,y) and time step t; $R(x,y)$, $G(x,y)$, and $B(x,y)$ are the pixel values in red, green, and blue channels, respectively.

If the pixel value of the subtraction result is below the threshold, then there is almost no change at this pixel. The subtraction result at point (x,y) can be simply expressed as

$$I_{result}(x,y) = |I_{t+1}(x,y) - I_t(x,y)| \quad (2)$$

In some studies, binary operation (Otsu 1979) or morphological filter (Serra 1982, Li *et al*. 2017) was incorporated into the method to extract the object more accurately. Since foreground aperture and ghosting problems exist in the FD results, some studies also presented new FD methods such as the adaptive FD (Mittal 2013) to find the optimal number of skipped frames (two or three frames) for subtraction operation to overcome FD problems. In this study, besides basic image processing methods, such as image binarization and morphological filter, more attention is paid to noise alleviation for better extraction of the boundary feature of objects by adopting movement judgment and area filter methods.

2.2.2 IFD method

The noise from the camera and environment may cause significant errors in estimating the rotational angle at the latter part. Therefore, an IFD method is proposed to better extract the boundary of the knife switch part. The improvements consist of movement judgment and area filter operations.

Table 2 Procedure of the IFD method

Input: Input videos (N frames)
For k = 3:2: N do

Step 1. Using the trained SSD network to obtain the ROI of the object at time k

Step 2. Conduct subtraction operation to the right part of knife switch for boundary feature extraction as

$$I = \left|I_k^{ROI}(x,y) - I_{k-2}^{ROI}(x,y)\right|$$

Step 3. Movement judgment
Set angle_flag = 1
If the pixel value $I(x,y)$ ≥ threshold: Object is moving
Else if the pixel value $I(x,y)$ < threshold: Set rotational angle as the value in the previous step
End if

Step 4. Image binarization
Image binarization is to change the gray values in a range of 0–255 to be 1 or 0 to simplify the values using the threshold determined by the Ostu method (Otsu 1979) as follow

$$I_{binary}(x,y) = \begin{cases} 1, & I_{result}(x,y) \geq threshold \\ 0, & I_{result}(x,y) < threshold \end{cases}$$

Step 5. Morphological filter operation (Serra 1982, Li et al. 2017)
Morphological filter operation is to segment the noises to the main parts in the obtained gray image, and the equation is as

$$I_{mor} = I_{binary} \Theta g_1 \oplus g_2$$
$$g_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad g_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

where g_1 and g_2 are morphological filters; Θ and $\oplus$ are corrosion and dilation operations achieved by operating the two morphological filters to the input image and reserve the intersection and union parts, respectively.

Step 6. Area filter operation (Liu *et al*. 2017)
Sort the areas of connected regions in the I_{mor} and reserve the top 10 connected regions.
The connected regions are obtained by searching the 8 pixels connected.

Output: The boundary feature of the knife switch

To speed up the FD processing, a subtraction operation is conducted at frames and time t and t+2. After obtaining the subtraction result, the movement judgment is added to reduce the error in the estimation of the rotational angle in static frames. The threshold for movement judgment is set to 20% of the maximum gray value of the processed frame. The pixel values of the subtraction result are ranked. If the pixel values after the top 20 are still larger than the threshold, then the frame contains a moving object. Otherwise, the rotational angle is set to the value in the previous step.

For the filter operation, an area filter (Liu *et al*. 2017) is added to keep the main area of the subtraction result, which can also reduce the influence of the noise caused by the environment and camera. In the area filter operation, connected regions in the subtraction results are ranked by the area sizes, and regions after the top 10 are excluded to keep the main boundary feature as well as exclude the noise in small size. The overall IFD procedure is summarized in Table 2. The right and left knife switch parts move simultaneously due to the mechanism of the device, so the right knife switch part is adopted in this paper for motion tracking.

2.3 HT Method

The HT method (Hough 1962, Duda and Peter 1972) is a linear and curve detection method. The concept of the method was first presented by Hough and adopted in image processing by Duda. The HT method is used to obtain the rotational angle of the object to move trajectory tracing based on the boundary line of the isolating switch extracted by the IFD method in this study.

The theory of the rotational angle detection using HT is illustrated in Fig. 5. Points along a straight line in the Cartesian coordinates can be represented by a point with two parameters r_0 and θ_0 in the parameter space as shown in Fig. 5(b). The equation corresponding to this transform can be written as

$$x\cos(\theta_0) + y\sin(\theta_0) = r_0 \tag{3}$$

The angle (θ_0) search can be performed by a least-squares fit of the points on the line using Eq. (3). θ_0 is restricted to [−90, 90], hence ensuring that the angle for a line is unique. To simplify the procedure in latter steps, θ_0 less than 0 in the multiple isolating switch cases are added by 180.

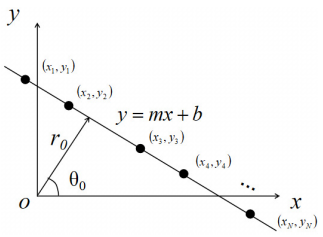
3. Experiments on isolating switches

3.1 Equipment and software

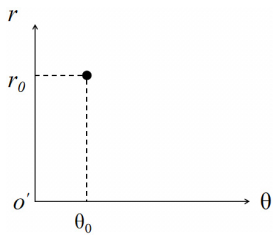
The camera equipment used to collect videos has a focal length of 12 mm and a diaphragm of 1.4. The sampling frequency is set to 30 frames per second (fps). The datasets are processed using the FFmpeg software and LabelImg software. Analysis of the image data is conducted by Python 3.7.7 (a programming language) and Pytorch 2.8.1 (a DL framework) using a high-performance server, including GeForce RTX 2070 and Cuda 10.1 with Intel (R) Core (TM) i7-9700 CPU @ 3.00GHz.

3.2 Database preparation

In this study, a large size of dataset with 720 images for training SSD was obtained from 12 videos of isolating switches taken under various illumination and camera



(a) Points along a line in Cartesian coordinate (x, y)



(b) Points for the line (a) in parameter space (r, θ)

Fig. 5 HT for rotation angle detection

Table 3 Detail information of videos for training of SSD

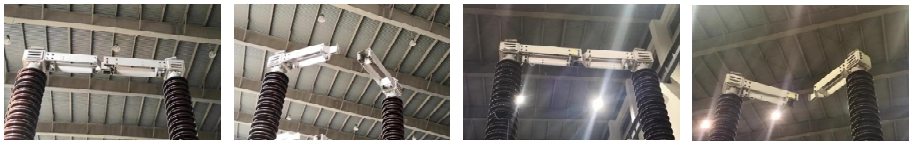
Name of dataset	Number of samples	Description of the sample			
		Camera location	Pause	Objects	Illumination
Case I	2	Same*	Without	Single	Day
Case II	2	Same	Without	Single	Day
Case III	2	Same	Without	Single	Day
Case IV	2	Same	Without	Multiple	Night (with light)
Case V	2	Same	Without	Multiple	Night (with light)
Case VI	2	Same	Without	Multiple	Night (with light)

Note: * - The camera locations are the same for two samples, while different for others in different conditions

Table 4 Detail information of videos for anomaly detection

Name of dataset*	Number of samples	Description of the sample			
		Camera location	Pause	Objects	Illumination
Cases 1-2	6	Same**	Without	Single	Day
	6		With		
Cases 3-4	6	Same	Without	Single	Day
	6		With		
Cases 5-6	6	Same	Without	Single	Night (with light)
	6		With		
Cases 7-8	6	Same	Without	Single	Night (with light)
	6		With		
Cases 9-10	6	Same	Without	Multiple	Day
	6		With		
Cases 11-12	6	Same	Without	Multiple	Day
	6		With		
Cases 13-14	6	Same	Without	Multiple	Night (with light)
	6		With		
Cases 15-16	6	Same	Without	Multiple	Night (with light)
	6		With		

Notes: * - Odd number Cases: without pause, and Even number Cases: with pause;
** - The camera locations are the same for two cases with and without pause, while different for other pairs in different conditions



(a) Single isolating switch: day and night with different camera locations

(b) Multiple isolating switch: day and night with different camera locations

Fig. 6 Example frames of photos in video samples

location conditions as shown in Table 3. For image dataset generation, 12 videos were firstly segmented into images using FFmpeg software (about 160 images for each), and 60 images were randomly selected for each video and then labeled with the LabelImg software for the SSD training (isolating switch images and labels). Among them, 600 images were used for training and the other 120 images for testing.

Transfer learning is adopted in this study by initializing from the source network of SSD, YOLO and YOLOv5s, and updating all neural parameters for the new training project (known as fine-tuning, Yin *et al*. 2018), making the procedure more efficient and thus no need for large dataset like 100,000 images. Since various conditions have been considered in dataset generation, and the detection result of the testing data*et al*so indicates the good performance of the trained SSD as discussed later, it could be concluded that 600 frames are enough for this study.

To examine the efficiency and robustness of the proposed method for anomaly detection, 96 videos (each video contains about 160 frames) of isolating switches under various conditions are collected as shown in Table 4. Six videos under the same conditions are taken for each case (three videos for the opening process and three for the closing process). Fig. 6 shows several frames in the video samples, which show switches at different states (open and closed) and under various camera locations and illumination. The proposed method mainly detects the defect that occurred during the opening and closing processes, which is manifested in the pause of the switches.

Table 5 Definitions of TP, FP, TN, and FN

Cases	Explanation
TP (True Positive)	Truly detected as positive for a positive sample.
FP (False Positive)	Falsely detected as positive for a negative sample.
TN (True Negative)	Truly detected as negative for a negative sample.
FN (False Negative)	Falsely detected as negative for a positive sample.

4. Results and discussion

4.1 Isolating switch detection by SSD

In this study, the effectiveness and accuracy of the SSD method are evaluated using the Precision (P), Recall (R) and average precision (AP) parameters (Everingham *et al.* 2015). Detection is assigned to ground-truth objects and judged to be true or false positives by measuring the Intersection over Union (IoU) defined in Fig. I-2 in Appendix I. For a true positive sample, the IoU must exceed 0.5 (Rahman and Wang 2016). Four evaluation indexes are described in Table 5.

In most ML methods, precision and recall values are two main indicators to show the detection performance of the network, which are defined as

$$Precision(P) = \frac{TP}{TP + FP}, \quad Recall(R) = \frac{TP}{TP + FN} \tag{4}$$

where precision is the probability of the truly detected positive samples among the samples positively detected, and recall is the probability of the truly detected positive samples among all positive samples. AP is to combine two indicators (P and R) above for detection performance, defined as the area under the *Recall–Precision* curve, as illustrated in Table 6.

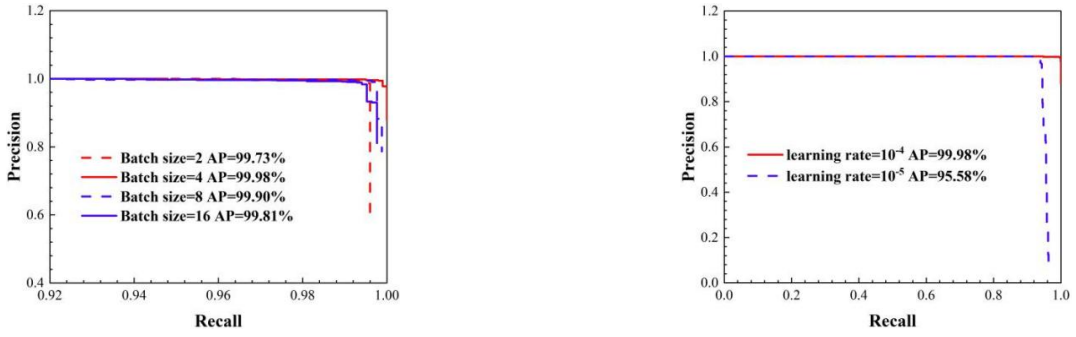
At first, the effects of the learning rate and batch size on the SSD performance are investigated, and the results are compared with those by YOLO network (Redmon *et al.* 2016) and YOLOv5s (YOLOv5 2020). YOLO network was proposed for the same purpose as the SSD, which has 24 convolution layers followed by 2 fully connected layers. The output layer of YOLO is in a shape of a matrix [7, 7, 30] to generate 98 bounding boxes as detection candidates for the non-maximum suppression (NMS) operation. YOLOv5s has a more complex structure than the SSD and YOLO, which has three components as Backbone (Focus, convolutional layer, spatial pooling layer), Neck (upsample layer, BottleneckCSP, and convolutional layer), and Head (convolutional layer).

In Fig. 7(a), AP values for the knife switch detection are shown for various batch sizes as 2, 4, 8, and 16 with a learning rate of 10^{-4}. The AP values with different batch sizes are found to be very similar, however the batch size is set to 4, which gives the largest AP. Fig. 7(b) examines the effect of different learning rates with a batch size of 4. The curve with a learning rate of 10^{-4} outperforms the others. The recall, precision, and AP values of the SSD, YOLO and YOLOv5s networks obtained using the whole testing dataset are summarized in Table 7, which shows the superiority of the SSD method. The precision and the recall of the SSD are 100% and 97.8%, whereas those of the YOLO are 99.3% and 78.6%, and those of the YOLOv5s are 97.2% and 98.8%.

Table 6 Procedure of AP calculation

Step 1. Sort the detection results of the testing dataset by the confidence scores.
Step 2. Calculate the largest precision when recall as 0, 1/*N*, 2/*N*, …, *m*/*N*.
(m is the number of the detected bounding boxes matched with the ground truth boxes in the testing dataset; and *N* is the number of the ground truth boxes)
Step 3. Plot the curve of *Recall* (x-axis)-*Precision* (y-axis) obtained in Step 2, then calculate the area between the curve and coordinate axis

$$AP = \int_0^1 P(R)dR$$



(a) Effect of batch size (b) Effect of learning rate

Fig. 7 Effects of parameters on the results of the SSD network

Table 7 Performance for object detection by SSD, YOLO and YOLOv5s

Method	Precision (%)	Recall (%)	Average precision (%)
SSD	97.8	100	99.98
YOLO	78.6	99.3	98.99
YOLOv5s	98.8	97.2	99.00

Note: - the learning rate and batch size of the YOLO and YOLOv5s are chosen under the largest AP value

4.2 Rotational angle detection by IFD and HT

The rotational angle detection procedure consists of the IFD and HT. The IFD method is adopted to extract the boundary feature of the knife switches for rotational angle detection.

The movement judgment is performed using the static frames detection as shown in Fig. 8(a), where the green dots indicate the extracted feature for movement. In the static frame detection case, no feature for movement was extracted by the IFD method, whereas many green dots (noise) were obtained by the conventional FD method. Fig. 8(b) shows the effect of the area filter for a moving object. Since area filters can keep the main feature as well as leave out the noise in a smaller size, the result obtained by the IFD method contains less green dots, apart from on the boundary of the knife switch than those by the conventional FD method. The IFD outperforms the conventional FD in minimizing the noise effect by adopting movement judgment and area filter operations (Fig. 8).

The comparison of the conventional FD and IFD methods for the rotational angle calculation for the knife switch is carried out on 48 videos with a pause during the movement described in Table 4. Various camera locations and illumination conditions are considered. The rotational angle is set to the same as the previous frame for static frames, otherwise it is calculated using HT. The average accuracy for the rotational angle detection is adopted as follows

Table 8 Average accuracies in rotational angles for cases with pauses

Name of dataset	On a single isolating switch	
	Conventional FD	IFD
Case 2	0.818	0.993
Case 4	0.762	0.995
Case 6	0.934	0.997
Case 8	0.888	0.982

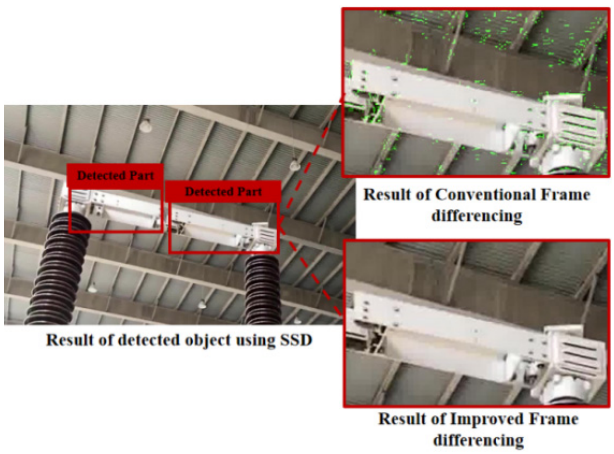
Note: Size comparison of three switches — Switch 1 > Switch 2 > Switch 3

$$Accuracy_j = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{i,j}}{A_{i,j}} \tag{5}$$

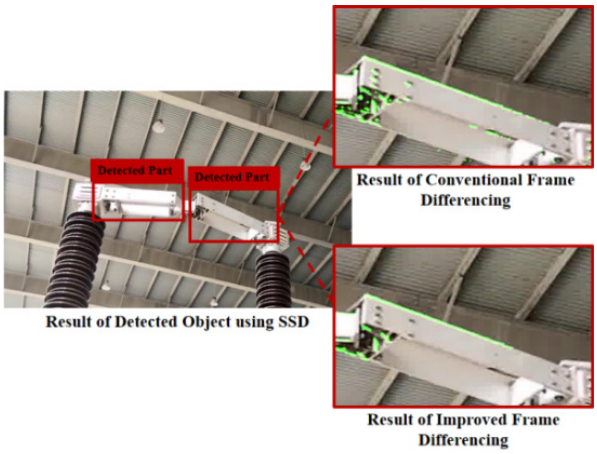
where N is the number of samples (6) for each case, $T_{i,j}$ is the number of truly detected rotational angles (in Sample i, Case j), and $A_{i,j}$ is the number of all the detected rotational angles (in Sample i, Case j). The truly detected rotational angles are defined as those angles with errors less than 3° (about 5% to the minimum rotational angle ranges). The average accuracies of two FD methods are compared for cases with pauses in Table 8. The rotational angle curves obtained by the IFD method are found to be much better in recognizing the frames with pause and the boundaries of the knife switches. The minimum average accuracies of the IFD are found to be 0.982 on a single isolating switch and 0.817 on multiple switches, while the minimum accuracies of the FD are 0.762 and 0.635.

4.3 Anomaly detection in movements

The damage of the isolating switches usually occurs during the opening and closing operations, manifested in the pause of the switches. The rotational angle is used to track the movement of the switches. Fig. 9 shows two examples of rotational angle curves in Cases 1 and 2 and Cases 13 and 14 detailed in Table 4. For each example,



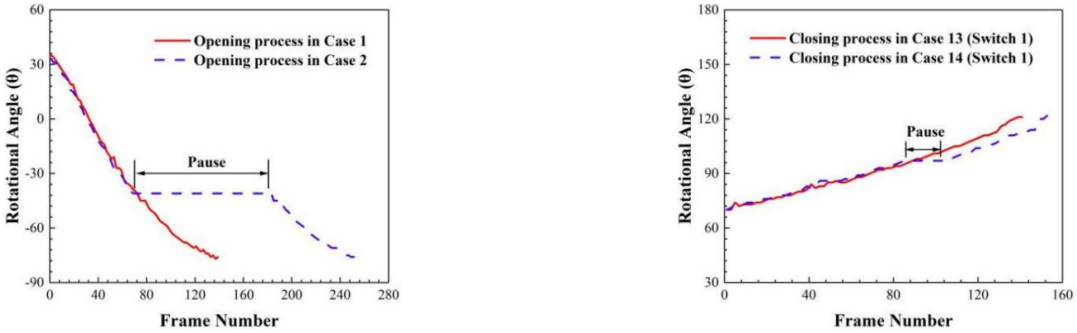
(a) Static frames: movement judgment



(b) Moving frames: the area filter effect

Fig. 8 Two IFD methods

Note: Green dots indicate movement features by the FD method



(a) Single isolating switch (b) Multiple isolating switches

Fig. 9 Rotational angle curves for normal and defect cases by HT

Note: Rotational angles smaller than 0 in multiple isolating switch cases are added by 180

videos were taken at the same camera locations with similar illumination. The only difference lies in the pause during the movement of the isolating switch in Cases 2 and 14.

The root mean square deviation (RMSD) is employed to detect defects during the movement as

$$RMSD_{k,j} = \sqrt{\sum_{i=1}^{N}\left(S_{k,j}^{i} - \frac{\bar{S}_{k-1}^{i}}{\sum_{i=1}^{N}(\bar{S}_{k-1}^{i})^{2}}\right)^{2}}, \quad k = 2, 4, \ldots, 16 \tag{6}$$

where $S_{k,j}^{i}$ is the rotational angle for the defect Sample j at frame i in Case k; and $\bar{S}_{k-1}^{i}$ is the average rotational angle for the normal cases without pause from six samples in Case (k-1); and N is the number of frames during the opening and closing movement for each case. For comparison, RMSD values for normal cases are also calculated as

$$RMSD_{k,j} = \sqrt{\sum_{i=1}^{N}\left(S_{k,j}^{i} - \frac{\bar{S}_{k}^{i}}{\sum_{i=1}^{N}(\bar{S}_{k}^{i})^{2}}\right)^{2}}, \quad k = 1, 3, \ldots, 15 \tag{7}$$

The mean RMSD results for normal cases using the rotational angles obtained from FD and IFD are summarized in Table 9 and the RMSD results for defect cases are summarized in Table 10. The RMSD values for defect cases are large enough for anomaly detection compared with normal cases. The value of the RMSD depends on the camera location, the opening and closing processes of the isolating switches, and the duration of the pause. Obviously, when the number of pause frames becomes larger, the RMSD result also becomes larger. The smallest pause frame number detected in the experiment is 11 (Samples 2 and 3 in Case 14), and the pause time for the case is about 0.37s (30 fps). In this study, the threshold for anomaly detection is taken as two times the average largest RMSD value of IFD for cases without pause which is 0.088 from Switch 3 in Case 11 at Table 9. Then, the success ratio of the anomaly detection is obtained as 97.6%, where 2 out of 84 RMSD values of the defect cases are less than the threshold as shown in Table 10. When the conventional FD is adopted, the threshold is 0.382 which is two times the largest RMSD value of FD at Table 9, and the success ratio of the anomaly detection is 79.8%.

Table 9 Average RMSD for cases without pauses

Cases	On a single isolating switch	
	Conventional FD	IFD
Case 1	0.058	0.019
Case 3	0.050	0.035
Case 5	0.041	0.014
Case 7	0.077	0.039

Cases	On multiple isolating switches					
	Conventional FD			IFD		
	Switch 1	Switch 2	Switch 3	Switch 1	Switch 2	Switch 3
Case 9	0.041	0.070	0.155	0.039	0.048	0.063
Case 11	0.061	0.134	0.196	0.042	0.076	0.088
Case 13	0.049	0.135	-	0.068	0.066	-
Case 15	0.057	0.152	-	0.069	0.074	-

4.4 Effect of illumination and camera location

The effect of the illumination and camera location is examined based on the average accuracies for defect cases and the average RMSD of normal cases in Tables 8 and 9. For single isolating switch cases, the camera is set immediately before the isolating switch in Cases 1 and 5. The illumination condition is in day time for Cases 1 and 3, whereas it is at night time with light for Cases 5 and 7. The accuracy of the rotational angle detection is not much affected by both camera location and illumination conditions, whereas the average RMSD in Cases 1 and 5 is lower than the other two cases due to the different camera locations.

For multiple isolating switch cases, the size of the knife switch has a significant influence on both accuracy of the angle detection and RMSD for anomaly detection. As for the effect of the camera location and illumination impact, the average accuracy values do not have much difference

Table 10 RMSD for anomaly detection with pauses by IFD

Cases		RMSD and Number of pause frames in parentheses					
		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Case 2		*1.152**(111)**	0.637(73)	1.085(83)	0.365(49)	1.031(59)	0.712(67)
Case 4		*1.095**(101)**	0.567(79)	1.042(99)	0.575(71)	1.074(61)	0.702(73)
Case 6		1.059(23)	0.431(21)	*1.156(31)	0.457(25)	1.038(27)	0.584**(61)**
Case 8		0.668(25)	0.381**(55)**	*0.795(43)	0.213(19)	0.689(27)	0.303(33)
Case 10	Switch1	0.664(75)	0.743(53)	0.712**(79)**	0.782(59)	0.561(65)	*0.819(63)
	Switch2	0.745(75)	0.730(53)	*0.776**(79)**	0.771(59)	0.579(65)	0.792(63)
	Switch3	0.837(75)	0.820(53)	*0.898**(79)**	0.861(59)	0.697(65)	1.103(63)
Case 12	Switch1	0.545(71)	0.732(47)	0.414(55)	0.835(57)	*0.945**(131)**	0.940(77)
	Switch2	0.649(71)	0.761(47)	0.551(55)	0.841(57)	*0.996**(131)**	0.938(77)
	Switch3	0.588(71)	0.648(47)	0.496(55)	0.740(57)	*0.931**(131)**	0.856(77)
Case 14	Switch1	0.163(13)**	0.236(11)	0.220(11)	*0.532(13)	0.268**(17)**	0.366**(17)**
	Switch2	0.199(13)	0.315(11)	0.240(11)	*0.411(13)	0.299**(17)**	0.391**(17)**
Case 16	Switch1	0.104(17)**	0.509(17)	0.228(19)	0.495(17)	0.456**(57)**	*0.618(41)
	Switch2	0.238(17)	0.513(17)	0.304(19)	0.488(17)	0.543**(57)**	*0.614(41)

Note: * - with the largest RMSD value in each case; **bold** - with the largest number of pause frames;
** - with RMSD less than the threshold of IFD

among the four cases, whereas the average RMSD values for Cases 13 and 15 are much larger than those for Cases 9 and 11. When checking the frames in the four cases, the color difference between the knife switch and background is quite obvious in Cases 9 and 11.

In summary, for accurate rotational angle detection in practical engineering conduction, it will be better to set the camera immediately before the object. Since isolating switches are normally established in outdoor environment, the color difference between switches and background is much more obvious than in the experiment, resulting in less impact on the results. For small object detection, camera equipment with higher pixels may be required.

4.5 Number of skipped frames in IFD

In the proposed method, the IFD method is performed at time t and t+2 instead of two adjacent frames (t and t+1) to speed up the processing. Samples 1 and 2 in Case 1 (each sample contains about 160 frames) are adopted to compare the performance of the rotational angle detection with different numbers of frames skipped. RMSD method is used here to compare the angles with a smoothed curve by the moving average scheme as

$$S_i^{'} = \frac{1}{5}(S_{i-2} + S_{i-1} + S_i + S_{i+1} + S_{i+2}) \tag{8}$$

where S_i is the ith rotational angle obtained and $S_i^{'}$ is the smoothed rotational angle. Then, the RMSD is calculated as

$$RMSD = \sqrt{\sum_{i=1}^{N}(S_i - S_i^{'})^2 / \sum_{i=1}^{N}(S_i^{'})^2}, \quad i = 1, 2, \ldots, N \tag{9}$$

Table 11 RMSD and time-consumed for IFD

Samples	RMSD with different number of skipped frames			
	t and t+1	t and t+2	t and t+3	t and t+4
Sample 1	0.011	0.014	0.021	0.026
Sample 2	0.014	0.016	0.018	0.028
Time-consumed (s)	25–26	14–15	10–11	8–9

where N is the number of frames. The RMSD and time consumed are listed in Table 11, which shows considerable improvements in the RMSD–based angle detection and the processing time with an increasing number of the skipped frames. The IFD with t and t+2 is chosen in this study, because the performance gets better in the angle detection and processing time.

4.6 Challenges in the proposed method

Dataset availability and preparation:

Deep learning based methods generally need a large amount of datasets for training to achieve good performance of the network. In this study, 720 samples are manually prepared for training and testing, taking about 5 hours for ground-truth box and category labeling. The samples for training need to consider the different camera locations and illumination conditions. In practical engineering, more data samples for isolating switches need to be collected outdoor under different weather conditions to improve the robustness of the method.

Training and detection time:

The SSD network adopted in this study uses the transferred network VGG-16, which takes about 2 hours to train the present isolating switch system, while networks

used in other electrical equipment detection took an average of 48 hours to train (Siddiqui *et al*. 2018). The SSD work shows a detection time of 25–27 ms per frame (about 38 frames per second), which is a little slower than the speed of 46 frames per second in a reference (Liu *et al*. 2016). However, the current speed already meets the needs of a real-time inspection system using the camera with fps of 30 (Bouvy *et al*. 1995). The speed of the IFD and HT part is at 75–80 ms per frame.

5. Conclusions

A novel system for anomaly detection for moving isolating switches has been proposed, considering different conditions such as camera location, illumination, single object, and multiple objects. The key idea is the combination of DL and image processing methods in which SSD method, FD improvement, and HT introduction are employed to compute rotational angle of knife switches in motion. The results of this study are summarized as follows:

- The SSD network is found to be excellent for detecting the isolating switch. The precision and the recall of the SSD are 100% and 97.8%, whereas those of the YOLO are 99.3% and 78.6%, and those of the YOLOv5s are 97.2% and 98.8%.
- The IFD and HT methods show good performance in the rotational angle calculation. The movement judge in the IFD improves the ability in recognizing the pause frames during the motion, while the area filter keeps the main part of the boundary feature as well as leaves out the noise in a smaller size. The minimum accuracy of the rotational angles using the IFD and HT method is 98.2% for a single isolating switch and 81.7% for multiple switches, whereas the conventional FD method are 76.2% and 63.5%, respectively.
- The proposed method using HT for the rotation angles is found to be sensitive to the defect that occurred during the movement of the isolating switch. With RMSD criteria for anomaly detection set to two times of the maximum RMSD for the case without pause, the success ratio of the anomaly detection using IFD is 97.6%, whereas the success ratio of the FD is 79.8%. The minimum duration of pause in the experiment is about 0.37 s with an RMSD of 0.22.
- The proposed method is robust under various different illumination conditions and camera positions. The average accuracy for rotational angle detection and RMSD for anomaly detection are not much affected. However, the camera location is better to be set immediately in front of isolating switches for better performance, and camera equipment with higher pixels may be required for small object detection.
- The adoptions of the transfer learning in SSD and the skipped frames in IFD result in significant improvements in the anomaly detection and the processing time in network training.

Future research and applications are suggested in the following areas: automatically labeling for the DL method, more experiments on real cases, application to other types of moving objects and civil engineering areas such as monitoring of large flexible structures and detecting/ tracking of vehicles.

Acknowledgments

The research described in this paper was financially supported by the National Key R&D Program of China (2018YFE0125400, 2019YFE0112600, 2017YFC0806100) and National Natural Science Foundation of China (U1709216).

References

Bouvy, R.J., Vincent, J., Parulski, K.A., Balch, K.S. and Erickson, G.L. (1995), "Progressive scan 30-frame-per-second megapixel camera", *Proceedings of SPIE - The International Society for Optical Engineering*, **2416**, 30-36.

Chen, A.W., Le, Q.M., Zhang, Z.Y. and Sun, Y. (2012), "Image recognition method for substation disconnecting switches state based on robots", *Automat. Electric Power Syst.*, **6**, 106-110.

Choi, J.W., Wangbo, T.K. and Kim, C.G. (2015), "A contour tracking method of large motion object using optical flow and active contour model", *Multimedia Tools Applicat.*, **74**(1), 199-210. https://doi.org/10.1007/s11042-013-1756-6

Duda, R.O. and Peter, E.H. (1972), "Use of the Hough transformation to detect lines and curves in pictures", *Commun. ACM*, **15**(1), 11-15. https://doi.org/10.1145/361237.361242

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2015), "The pascal visual object classes challenge: a retrospective", *Int. J. Comput. Vis.*, **111**(1), 98-136. https://doi.org/10.1007/s11263-014-0733-5

Fang, S., Shu, X.H. and Li, D.W. (2017), "A Method based on machine vision for opening-closing status recognition of substation disconnecting switches", *J. Hunan Univ. Technol.*, **06**, 1673-9833.

Fei, M.J., Li, J. and Liu, H.H. (2015), "Visual tracking based on improved foreground detection and perceptual hashing", *Neurocomputing*, **152**, 413-428. https://doi.org/10.1016/j.neucom.2014.09.060

Felzenszwalb, P., Girshick, R., McAllester, D. and Ramanan, D. (2010), "Object detection with discriminatively trained part based models", *IEEE Transact. Pattern Anal. Mach. Intell.*, **32**(9), 1627-1645. https://doi.org/10.1109/TPAMI.2009.167

Geiger, A., Lenz, P., Stiller, C. and Urtasun, R. (2013), "Vision meets robotics: The kitti dataset", *Int. J. Robot. Res.*, **32**(11), 1231-1237. https://doi.org/10.1177/0278364913491297

Girshick, R. (2015), *Fast R-CNN*, Computer Science.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, The MIT Press.

Guo, Y., Xu, Y. and Li, S. (2020), "Vision-based full-field panorama generation by UAV using GPS data and feature points filtering", *Smart Struct. Syst., Int. J.*, **25**(5), 631-641. https://doi.org/10.12989/sss.2020.25.5.631

Hough, V.P.C. (1962), Method and means for recognizing complex patterns.

Jun, K.S., Jae-Yeal, N. and Chul, K.B. (2018), "Online tracker optimization for multi-pedestrian tracking using a moving vehicle camera", *IEEE Access*, **6**, 48675-48687. https://doi.org/10.1109/ACCESS.2018.2867621

Lei, X.S. and Sui, Z.H. (2019), "Intelligent fault detection of high

voltage line based on the Faster R-CNN", *Measurement*, **138**, 379-385. https://doi.org/10.1016/j.measurement.2019.01.072

Li, M.H., Liu, Z.X., Xiong, Y.Y. and Li, Z. (2017), "Multi-person tracking by discriminative affinity model and hierarchical association", *Proceedings of the 3rd IEEE International Conference on Computer and Communications* (*ICCC*), Sichuan, China, December.

Li, W.B., Chang, M.C. and Lyu, S. (2018), "Who did what at where and when: Simultaneous multi-person tracking and activity recognition", University at Albany, New York, NY, USA.

Li, S., Guo, Y., Xu, Y. and Li, Z. (2019), "Real-time geometry identification of moving ships by computer vision techniques in bridge area", *Smart Struct. Syst*., *Int. J*., **23**(4), 359-371. https://doi.org/10.12989/sss.2019.23.4.359

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., and Berg, A.C. (2016), "SSD:Single shot multibox detector", *Proceedings of European Conference on Computer Vision*, Amsterdam, Netherlands, October.

Liu, Y.Q., Zhan, F.Y., Jiang, X.W. and Zhou, H.Y. (2017), MATLAB: computer vision and deep learning combat.

Mittal, S. (2013), "Object tracking using adaptive frame differencing and dynamic template matching method", Bachelor. Dissertation; National Institute of Technology Rourkela, Odisha, India.

Mondal, T.G. and Jahanshahi, M.R. (2020), "Autonomous vision-based damage chronology for spatiotemporal condition assessment of civil infrastructure using unmanned aerial vehicle", *Smart Struct. Syst*., *Int. J*., **25**(6), 733-749. https://doi.org/10.12989/sss.2020.25.6.733

Nassu, B.T., Lippmann, L., Marchesi, B., Canestraro, A. and Zarnicinski, V. (2018), "Image-based state recognition for disconnect switches in electric power distribution substations", *Proceedings of the 31st Conference on Graphics, Patterns and Images* (*SIBGRAPI*), Foz do Iguacu, Brazil, October.

Otsu, N. (1979), "A thresholding selection method from gray-level histogram", *IEEE Trans.syst.man. & Cybern*, **9**(1), 62-66.

Paul, N., Singh, A., Midya, A., Roy, P.P. and Dogra, D.P. (2017), "Moving object detection using modified temporal differencing and local fuzzy thresholding", *J. Supercomput*., **73**, 1120-1139. https://doi.org/10.1007/s11227-016-1815-7

Rahman, M.A. and Wang, Y. (2016), "Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation", *Proceedings of the 12th International Symposium on Visual Computing* (*ISVC*), Las Vegas, VA, USA, December.

Ran, N., Kong, L., Wang, Y.H. and Liu, Q.J. (2019), "A robust multi-athlete tracking algorithm by exploiting discriminant features and long-term dependencies", *Proceedings of the 25th International Conference on Multimedia Modeling* (*MMM*), Thessaloniki, Greece, January.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016), "You only look once: Unified, real-time object detection", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, June.

Ren, S., He, K., Girshick, R. and Sun, J. (2015), "Faster R-CNN:Towards real-time object detection with region proposal networks", *Proceedings of the 29th Annual Conference on Neural Information Processing Systems* (*NIPS*), Montreal, Canada, December.

Serra, J. (1982), "Image analysis and mathematical morphology", *Biometrics*, **39**(2), 536.

Shafiee, M.J., Chywl, B., Li, F. and Wong, A. (2017), "Fast YOLO: A fast you only look once system for real-time embedded object detection in video", *J. Computat. Vision Imag. Syst*., **3**(1).

Sharma, S., Ansari, J.A., Murthy, J.K. and Krishna, K.M. (2017), "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking", *Proceedings of 2018 IEEE International Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June.

Shi, Y.H., Luo, Y., Tu, G.Y. and Wu, T. (2007), "An edge detectable algorithm for high-voltage isolating switch", *Relay*, **35**(12), 23-26.

Siddiqui, Z.A., Park, U., Lee, S.-W., Jung, N.-J., Choi, M., Lim, C. and Seo, J.-H. (2018), "Robust powerline equipment inspection system based on a convolutional neural network", *Sensors*, **18**(11), 3837. https://doi.org/10.3390/s18113837

Simonyan, K. and Zisserman, A. (2014), "Very Deep Convolutional Networks for Large-Scale Image Recognition", *Proceedings of 2015 International Conference on Learning Representations* (*ICLR*), San Diego, CA, USA, May.

Supreeth, H.S.G. and Patil, C.M. (2018), "Efficient multiple moving object detection and tracking using combined background subtraction and clustering", *Signal Image and Video Processing*, **12**(6), 1097-1105. https://doi.org/10.1007/s11760-018-1259-z

Tang, S., Andriluka, M., Andres, B. and Schiele, B. (2017), "Multiple person tracking by lifted multicut and person re-identification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, USA, June.

Wang, H., Xu, F., Jin, Y.Q. and Ouchi, K. (2009), "Estimation of bridge height over water from polarimetric sar image data using mapping and projection algorithm and de-orientation theory", *IEICE Transact. Commun*., **92**(12), 3875-3882. https://doi.org/10.1587/transcom.E92.B.3875

Wang, J., Liu, Q., Zhao, K., Yi, J. and Kai, P. (2017), "Recognition of high voltage isolating switch's states based on object tracking", *Proceedings of 4th International Conference on Systems and Informatics* (*ICSAI*), Zhejiang, China, November.

Wanyan, X.X., Fu, Q.C. and Xin, J. (2019), "Research on multi - target recognition of traction substation video based on transfer learning", *Comput. Eng. Applicat*., **55**(24), 196-201.

Ye, X.W., Dong, C.Z. and Liu, T. (2016), "Image-based structural dynamic displacement measurement using different multi-object tracking algorithms", *Smart Struct. Syst*., *Int. J*., **17**(6), 935-956. https://doi.org/10.12989/sss.2016.17.6.935

Yin, X., Chen, W., Wu, X. and Yue, H. (2018), "Fine-tuning and visualization of convolutional neural networks", *Proceedings of 2017 12th IEEE Conference on Industrial Electronics and Applications* (*ICIEA*), *IEEE*.

YOLOv5 (2020), https://github.com/ultralytics/yolov5. Ultralytics LLC, CA, USA.

Yu, C., Ding, X., Zhang, G. and Chen, M. (2017), "Observing Bridge Dynamic Deformation in Vibration by Digital Photography System", *Proceedings of 2nd International Conference on Mechatronics and Information Technology*.

Zhao, R., Shang, W., Wang T. and Zou, X. (2016), "State Evaluation of Isolating Switch in Transformer Substation Based on Image Processing", *Comput. Measure. Control*, **24**, 241.

Zhao, D.W., Fu, H., Xiao, L., Wu, T. and Dai, B. (2018), "Multi-object tracking with correlation filter for autonomous vehicle", *Sensors*, **18**(7), 2004. https://doi.org/10.3390/s18072004

Zhu, D., Feng, Y., Chen, Q. and Cai, J. (2010), "Image recognition technology in rotating machinery fault diagnosis based on artificial immune", *Smart Struct. Syst*., *Int. J*., **6**(4), 389-403. https://doi.org/10.12989/sss.2010.6.4.389

Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J. and Le, Q.V. (2020), "Learning data augmentation strategies for object detection", *Proceedings of European Conference on Computer Vision*.

Appendix I: Structure of the SSD

In the SSD, the base network is a standard architecture used for high-quality image classification, namely, Visual Geometry Group Network-16 (VGG-16, Simonyan and Zisserman 2014). In the auxiliary structure, the extra convolution layers are processed with classifiers to predict bounding boxes and category confidence scores, and NMS is adopted to ensure detection for each object. The confidence scores (c^0, c^1) are the probabilities of the object (0%–100%) in the predicted bounding box belonging to the corresponding categories obtained using convolution operations and are evaluated in Appendix II: c^0 for background and c^1 for switch, while the background is defined as the region excluding the switch.

The output-bounding box represents the location parameters using two coordinates of the center, height, and width parameters. However, the outputs of the convolution operation for the location prediction in the classifiers are shape offsets (deviations in four location parameters), whereas a default box is required as a basic bounding box to obtain the location parameters for the final bounding boxes. The width and height of default boxes are obtained using scales and aspect ratios. The scales determine the relative size of the default box to the input image, whereas the aspect ratios set the ratios between width and height of the box. The detailed parameters for 6 classifiers are illustrated in Table I-1.

An example for shape offsets and confidence score prediction of a feature map cell is shown in Fig. I-1. The aspect ratios are set as $a_r \in \{1,\ 2,\ 1/2\}$, and the number of categories is two (including background). For each feature map cell, default boxes with four kinds of width and height are assigned, then four pairs of shape offsets and confidence scores from the convolution operations will match with default boxes for loss calculation detailed in Appendix II. One pair of shape offsets and confidence scores is matched with the default box shown in red.

Appendix II: Training of the SSD

During the training, default boxes are matched with the ground-truth boxes, only those default boxes with the IoU in Fig. II-1 values larger than 0.5 are chosen as the matched ones and used in the loss calculation. For each ground-truth box, only one default box with the largest IoU is assigned as positive and the rest as negative for the computation of the loss functions in Eqs. (II-2) and (II-4). Both localization and classification losses are adopted to train the SSD network. The objective loss function of SSD is a weighted sum of the two kinds of loss as follows

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \qquad \text{(II-1)}$$

where L_{loc} and L_{conf} are the localization and confidence losses, respectively; N is the number of matched default boxes;$\alpha$is a weight term, set to 1 in this study; x represents the category match of the default box to the ground-truth box; c is the confidence score; l and g are the shape offsets of the default and ground-truth boxes, respectively.

Table I-1 Detailed information for classifiers

Classifier	Input feature map size	Kernel size	Number of convolution kernels		Output feature map size	
			For shape offsets	For confidence scores	For shape offsets (with channels)	For confidence scores (with channels)
Classifier 1	[38, 38]	3 × 3	4 × 4	4 × 2	[38, 38, 16]*	[38, 38, 8]**
Classifier 2	[19, 19]	3 × 3	6 × 4	6 × 2	[19, 19, 24]	[19, 19, 12]
Classifier 3	[10, 10]	3 × 3	6 × 4	6 × 2	[10, 10, 24]	[10, 10, 12]
Classifier 4	[5, 5]	3 × 3	6 × 4	6 × 2	[5, 5, 24]	[5, 5, 12]
Classifier 5	[3, 3]	3 × 3	4 × 4	4 × 2	[3, 3, 16]	[3, 3, 8]
Classifier 6	[1, 1]	3 × 3	4 × 4	4 × 2	[1, 1, 16]	[1, 1, 8]

Notes: * [38, 38, 16] - The 3D matrix is reshaped to [38, 38, 4, 4] to match with the default box parameters for bounding box generation;
**[38, 38, 8] - The 3D matrix is reshaped to [38, 38, 4, 2] to adapt to the corresponding confidence scores

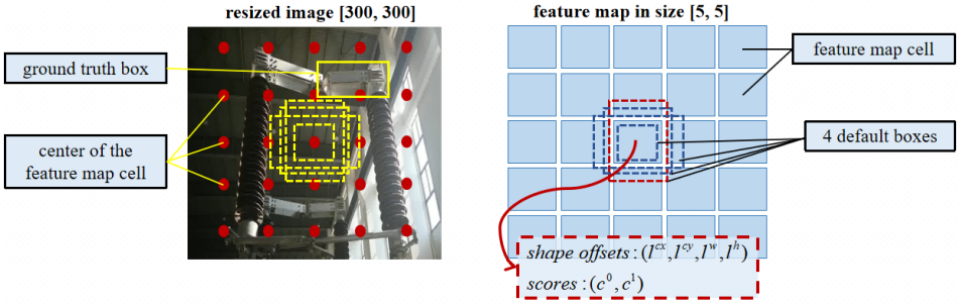
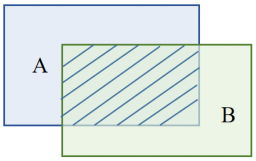


Fig. I-1 Example for shape offsets and confidence score prediction



$$IoU = \frac{area(A \cap B)}{area(A \cup B)}$$

Fig. II-2 IoU

The detailed computation of localization loss is given by (Liu *et al*. 2016)

$$L_{loc}(x,l,g) = \sum_{i \in Positive}^{N} \sum_{m \in (cx,cy,w,h)} x_{ij}\, smooth_{L1}\left(l_i^m - \hat{g}_{ij}^m\right)$$
$$\hat{g}_{ij}^{cx} = \frac{g_j^{cx} - d_i^{cx}}{d_i^w}, \qquad \hat{g}_{ij}^{cy} = \frac{g_j^{cy} - d_i^{cy}}{d_i^h}, \qquad \hat{g}_{ij}^{w} = log\left(\frac{g_i^w}{d_i^w}\right), \qquad \hat{g}_{ij}^{h} = log\left(\frac{g_i^h}{d_i^h}\right) \tag{II-2}$$

where $x_{ij} = \{1, \quad 0\}$ ($x_{ij} = 1$ if matched) is an indicator for matching the *i*th default box to the *j*th ground truth box of the switch category and $\hat{g}_{ij}$ is the shape offset between the ground-truth box *j* and default box *i*. The smooth L1 loss function is as follows (Girshick 2015)

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geq 1 \end{cases} \tag{II-3}$$

The confidence loss is the softmax loss as follows (Goodfellow *et al*. 2016)

$$L_{conf}(x,c) = -\sum_{i \in Positive}^{N} x_{ij}\, log(\hat{c}_i^1) - \sum_{i \in Negative} log(\hat{c}_i^0) \tag{II-4}$$

$$\hat{c}_i^1 = \frac{exp(c_i^1)}{exp(c_i^1) + exp(c_i^0)}, \qquad \hat{c}_i^0 = \frac{exp(c_i^0)}{exp(c_i^1) + exp(c_i^0)} \tag{II-5}$$

where $\hat{c}_i^1$ and $\hat{c}_i^0$ are the softmax calculation results for the switch and background categories, respectively, corresponding to the *i*th default box; c_i^1 and c_i^0 are the confidence scores of the switch and background categories, respectively.

To optimize the training, both hard negative mining (Felzenszwalb *et al*. 2010) and data augmentation methods (Zoph *et al*. 2020) are adopted. The hard negative mining is used to balance the negative and positive samples in a ratio of 3:1 (Liu *et al*. 2016), and data augmentation is to enlarge the dataset by cutting, rotating, and flipping operations to the images.