Object recognition for power equipment via human-level concept learning

Siheng Xiong1 | Yadong Liu1 | Yingjie Yan1 | Ling Pei1 | Peng Xu2 | Xiaofei Fu2 | Xiuchen Jiang1

1 School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
2 State Grid Shanghai Electric Power Company, Shanghai, China

Correspondence
Yadong Liu, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.
Email: lyd@sjtu.edu.cn

Funding information
National Key Research and Development Program of China; Modes and key supporting technologies for the sustainable development of smart grid based on TCP ISGAN of the International Energy Agency, Grant/Award Number: 2019YFE0102900

Abstract
Inspection robots are popularized in substations due to the lack of personnel for operation and maintenance. However, these inspection robots remain at the level of perceptual intelligence, rather than cognition intelligence. To enable a robot to automatically detect defects of power equipment, object recognition is a critical step because criteria of infrared diagnosis vary with types of equipment. Since this task is not a big-sample learning problem, prior knowledge needs to be added to improve existing methods. Here, an object recognition model based on human-level concept learning is proposed, which utilizes relationship between equipment. The proposed method is composed of three parts: Mask RCNN, Bayesian Context Network and human-level concept learning. As the backbone network, Mask RCNN, a pixel-wise segmentation network, gives preliminary recognition results. Then, based on the object relationship graph of Bayesian Context Network, human-level concept learning corrects the results in sequence by maximizing the conditional probability of an object given its neighbourhood. Experiments show that the accuracy of the proposed method increases 9.7% compared with its backbone network, making industrial application of this technology possible.

1 INTRODUCTION

The technology of replacing human with robot has become an important issue in electric power industry. Inspection robots in substation, as a typical type of this technology, can considerably reduce the workload of personnel for operation and maintenance [1–3]. However, a lot of manual labour is still required, due to the fact that these robots are not smart enough at present. For example, the work of defect detection based on RGB and infrared images of power equipment is not automatic. To identify defects which cause temperature rise, the information of type and position of equipment needs to be determined, since the temperature distribution of equipment is unobtainable without its position on the image and the normal temperature range of equipment varies with its type. As a result, automatically recognizing power equipment type and relating it to different alarm criterion of temperature become essential characteristics of the next-generation inspection robot technology.

However, power equipment recognition is not a big-sample learning problem, which makes it still a difficulty at present. To our knowledge, there is no available dataset of annotated power equipment images. Since the appearance of certain type of power apparatus is quite similar due to insulation requirements, image annotating of power equipment requires professional knowledge of electric engineering which blocks the collection work. Although many models of object recognition have been proposed recently [4–8], they generally achieved a good performance on a large dataset, e.g. the 200K-labelled COCO dataset [9] and the 11.5K-labelled Pascal VOC dataset [10]. To improve their performance in the task of power equipment recognition, more information within the images needs to be added as prior knowledge.

For a scene of power equipment, strong relationships between objects exist because of the functional connections between them. For example, bushings tend to come up in a picture with tubular bus-bars, since they are electrically connected. Similarly, tubular bus-bars and column insulators which support these bus-bars are often found together in a picture. These abundant and useful context information and object relationships are ignored in the object recognition models [4–8], which...
perform a task of classifying candidate boxes independently. As a result, it is a natural way of improving these models in the power equipment recognition task by adding the relationship between objects.

Object relationship modelling can be divided into two types: parametric and non-parametric. A parametric model tends to mapping features into higher dimension space, which are then processed by complex non-linear functions. Benefiting from a large number of parameters, this kind of model has a strong expressivity, which, in other words, performs quite well on large datasets. Besides, parametric models are still a black box at present, the mechanism of which is totally different with human perception. A non-parametric model, on the other hand, learns object relationship through probabilities, which is less dependent on data by introducing uncertainty measurements and prior knowledge. This kind of learning mechanism is better understood by human, making the combination between algorithm and human knowledge possible.

Parametric object relationship modelling has attracted the attention of researchers for its good performance on large dataset. In [11], object detection is formulated as graph structure inference, with nodes and edges denoting objects and object relationships respectively. The proposed Structure Inference Network (SIN) is based on Recurrent Neural Network [12], where object features are taken as initial state, and the message from scene features and features of other objects are taken as input. The authors of [13] proposed a relation network (RN) according to attention module [14, 15], which simulates human focusing on local information instead of whole image. In RN, object relationships consist of appearance features and geometric features which are embedded into the same high-dimension space for following operation. These methods tend to include large numbers of learnable parameters, because of which they have a strong expressivity but require a lot of data for learning.

Power equipment relationship modelling should be non-parametric for the following reasons. First, compared with parametric ones, non-parametric model has less dependency on data. Besides, non-parametric model is similar to human perception, making the introduction of human knowledge and logical rules possible. A suitable framework for non-parametric object relationship modelling is introduced. In [16], Hierarchical Bayesian Program Learning (HBPL), based on the three important principles of human learning process: compositionality, causality, and learning to learn [17], is designed for small-sample learning problems. It decomposes a complex concept into combination of simple parts which are suitable for transfer learning. Besides, causal relationships between these simple parts are helpful to the introduction of human experience.

Inspired by the framework of HBPL, a non-parametric object relationship model is proposed. Detection results of Mask R-CNN [18], which is a pixel-wise segmentation network, are used as the input of Bayesian Context Network (BCN). BCN, constructs the relationship network of detection results. Then human-level concept learning module corrects, in order of the number of reliable neighbours, these detection results by maximizing the probability of a device given its neighbourhood. Compared with those parametric models, the proposed model is more suitable for this task, making industrial application of this technology possible.

2 | NETWORK STRUCTURE

In Figure 1, whole structure of the proposed method is given, where BCN is used to model the object relationship and Human-level Concept Learning (HLCL) is utilized to correct detection results based on these relationships. Since the object relationship is modelled at the level of objects, detection results of Mask R-CNN are used as the input of BCN. Mask R-CNN, a deep convolutional neural network, is elaborated for the task of object recognition but ignores the relationship between objects [18]. BCN is actually a graph representation of the scene where nodes denote equipment and edges denote equipment relationships. HLCL is a generative method based on the framework of HBPL. Through maximizing the probability of a node given its neighbourhood, HLCL is able to infer the accurate equipment types.

Characteristics of Mask R-CNN, which is the foundation of object relationship model, is discussed below. Mask R-CNN, firstly feeds the input image into a convolutional backbone to generate region of interest (RoI) via Region Proposal Network (RPN) and RoI features via RoI Align layer. Then a R-CNN head and a Mask head are utilized respectively to classify ROIs and generate boxes and masks which represent the position of objects in the picture. The classification of ROIs is independent without consideration of their relationships. And class, box and mask compose the detection results.
Bayesian Context Network

Human-level concept learning

MODELLING

Mask R-CNN is proposed, which is named as Bayesian Context Network. A graph of the scene constructed from the detection results of Mask R-CNN is illustrated in Figure 2. Before object relationships modelling, a scene composed of objects in substations. These scenes have two obvious characteristics: (1) few labelled samples can be found for training; (2) strong relationships of object recognition in scenes like power equipment are difficult, which is done step by step in the proposed method. Figures 3 and 4 give respectively an illustration of the inference process and the correction process of the inference.

### 3 | OBJECT RELATIONSHIP MODELLING

The main contribution of the paper is to model object relationship at the level of objects, which can greatly improve the performance of object recognition in scenes like power equipment in substations. These scenes have two obvious characteristics: (1) few labelled samples can be found for training; (2) strong relationships exist between objects. The modelling process is divided into two parts: BCN and HLCL, which will be introduced in detail.

#### 3.1 | Bayesian context network

Before object relationships modelling, a scene composed of many objects and their relationships needs to be represented clearly and efficiently. As a kind of representation, graph contains rich relationship information among elements [19]. As a result, a graph of the scene constructed from the detection results of Mask R-CNN is proposed, which is named as Bayesian Context Network. Here the construction process is described in detail with the illustration shown in Figure 2.

A BCN is composed of nodes and edges which respectively denote objects and the relationship between objects. It is easy to understand that a single object is represented by one node. As for the relationship between objects, some explanations need to be given. A relationship, also an edge in the graph, is defined by the two nodes it connects. From the mathematical perspective, edges should exist between every two nodes. However, this strict rule makes the network quite complicated, and the process of learning and prediction become quite time-consuming. Besides, for a single object, not all the relationships are equally important. The closer two objects are, the stronger connection between them exists. When learning with few labelled samples, focusing on stronger and more helpful relationships rather than all of them is necessary. Consequently, a node is only connected with its neighbours in the BCN.

After the definition of BCN, attributes of the nodes are also discussed. Suppose there are N objects d(1), d(2), ..., d(N) in the detection result of a single image. For each detection object d(i), it has three attributes: class d(i)c, bbox d(i)b and score d(i)s. Bbox refers to the bounding box of the object and can be written as d(i)b = (x(i), y(i), w(i), h(i)) where (x(i), y(i)) are the central point, width and height of the box respectively. Score refers to the confidence of the detection object given by Mask R-CNN. Obviously, d(i)s ∈ (0, 1) and a detection object with higher confidence is more likely to be correct.

Based on these attributes, the definition of neighbourhood is given. A single detected object is called an instance for convenience. In an image, the neighbourhood of an instance d(i) with attribute d(i)c = (x(i), y(i), w(i), h(i)) is considered as a bounding box d(i)b = (x(i), y(i), αw(i), αh(i)) where α ≥ 1 is an adjustable parameter controlling the size of neighbourhood. Another instance d(j) is considered as a neighbour of d(i) if d(i)b and d(j)b have intersection. The set of neighbours of d(i) is denoted by ε(d(i)) = {d(1), d(2), ..., d(N)}. It should be noted that the class d(i)c of an instance d(i) can be incorrect especially when d(i)s is low. The correct class of d(i) is assumed as g(i).

#### 3.2 | Human-level concept learning

After the representation of a scene, the correction of detection results from the perspective of object relationship is done with Human-level concept learning, which is a generative model based on HBPL.

Inferring the correct class of all the instances at one time is difficult, which is done step by step in the proposed method. Suppose there is an instance d(i) = {d(i)c, d(i)b, d(i)s} with its correct class g(i) and neighbours ε(d(i)). Obviously, from the perspective of object relationship, the instance and its neighbours would be at the highest possibility given its correct class, which can be written as:

\[
\text{g}(i) = \arg \max_{g} P(d(i), \varepsilon(d(i)) | g(i))
\]

The order of correction in a single picture is discussed below. When inferring the correct class of a node, it is obvious that this inference is more likely to be accurate if more instances in its neighbourhood are reliable. That is, if ε(d(i)) = {d(1), d(2), ..., d(N)} are more reliable, g(i), inferred by maximizing P(d(i), ε(d(i)) | g(i)), is more likely to be accurate. Reliable instances come from two ways: (1) it has a high detection confidence; (2) it has been corrected by the model. Unreliable neighbouring nodes are not used in correction. Figures 3 and 4 give respectively an illustration of the inference process and the correction process of HLCL.

Suppose there are N instances d(1), d(2), ..., d(N) in a single image. For d(i)s ≥ threshold, d(i) is considered to be reliable.
For $j = 1, 2, \ldots, n$,

$$S_j = \left\{ d^{(\alpha)}, d^{(\beta)}, \ldots, d^{(\omega)}, d^{(1)}, d^{(2)}, \ldots, d^{(j-1)} \right\}$$

$$g^{(j)}_c = \arg \max P(S_j \cap E(d^{(j)}), d^{(j)} \mid g_c)$$

$$\tilde{c}^{(j)} = \arg \max_d \#(S_j \cap E(d^{(j)}))$$

$$t \in \{1, 2, \ldots, N\} - \{r_1, r_2, \ldots, r_m\} - \{c_1, c_2, \ldots, c_{j-1}\}$$

where $S_j$ is the set of reliable instances at step $j$, which is composed of the original set of reliable instances $d^{(\alpha)}, d^{(\beta)}, \ldots, d^{(\omega)}$ and corrected instances before step $j$; refers to the idea that only reliable instances in the neighborhood of $d^{(j)}$ are used for inference; in Equation (4), $d^{(j)}$ is corrected as $d$ with $d^{(j)}$ replaced by $g^{(j)}_c$.

The order of the sequence $d^{(\alpha)}, d^{(\beta)}, \ldots, d^{(\omega)}$ is determined by the number of reliable instances in the neighborhood and can be written as:

For $j = 1, 2, \ldots, n$,

$$\tilde{c}^{(j)} = \arg \max_d \#(S_j \cap E(d^{(j)}))$$

$$t \in \{1, 2, \ldots, N\} - \{r_1, r_2, \ldots, r_m\} - \{c_1, c_2, \ldots, c_{j-1}\}$$

and no correction will be made to it. For $d^{(j)} <$ threshold, $d^{(j)}$ will be corrected in sequence. Now there are a set of reliable instances $d^{(\alpha)}, d^{(\beta)}, \ldots, d^{(\omega)}$ and a sequence of unreliable instances $d^{(\alpha)}, d^{(\beta)}, \ldots, d^{(\omega)}$, which will be corrected in order. The correction will be processed as:

$$P(d_5, d_6, d_7, d_8 \mid g_c) = P(d_5, d_6, d_7, d_8)$$

Then the joint probability can be further written as:

$$P(d_5, d_6, d_7, d_8) = P(d_5 \mid d_6, d_7, g_c) P(d_6 \mid d_5, d_7, g_c) P(d_7 \mid d_5, d_6, g_c) P(d_8 \mid d_5, d_6, d_7, g_c).$$

Note that $d_6$ and $d_8$ are not connected, $P(d_5 \mid d_6, d_7, g_c)$ can be simplified as $P(d_5 \mid d_5, d_7, g_c)$. Actually, since not all the nodes are connected, all the conditional probability in correction process can be simplified in a similar way.

In the correction process, instance attributes of class $d$, and $bbox$ $d_5$ are used. The effect of location and distance changes of camera is eliminated by normalising $d_5$. Through translation.
that moves \((x^{(i)}, y^{(i)})\) to the origin of coordinates and scaling that changes \(w, h\) into the unit of \(x\)-axis and \(y\)-axis, it can be obtained that \(d_x^{(i)} = (x^{(i)} - x^{(j)}) / w^{(j)},\)
\(d_y^{(i)} = (y^{(i)} - y^{(j)}) / h^{(j)},\)
\(d_w^{(i)} = w^{(i)} / w^{(j)},\)
\(d_h^{(i)} = h^{(i)} / h^{(j)},\) for \(d^{(j)} \in \epsilon(d^{(i)}).\) To make the model agree with human cognition, bbox \(d^{(j)} = (x^{(j)}, y^{(j)}, w^{(j)}, h^{(j)})\) is divided into two parts: position \(d_p^{(j)} = (x^{(j)}, y^{(j)})\) and size \(d_w^{(j)} = (w^{(j)}, h^{(j)})\). Therefore, for an instance, three attributes of it are considered in the correction process: \(d_x, d_y, d_w, d_h.\)

Simultaneously using three attributes also causes the problem: the probability distribution is hard to assume and therefore unfriendly to small-sample learning. This problem is solved according to human cognition of relationship between power equipment. Suppose there is a picture of tubular buses and column insulators which support these buses, their attributes can be hierarchically described: (1) Class: tubular bus and column insulator tend to appear together; (2) Position: if there are two devices which are tubular bus and column insulator, it is very likely that column insulator is under tubular bus; (3) Size: if there are two devices which are tubular bus and column insulator, it is very likely that their sizes are similar. Here equation of this process is given. For an instance \(d^{(i)}\), the probability of another instance \(d^{(j)}\) appearing in its neighbourhood can be written as:

\[
P(d^{(i)}|d^{(j)}) = P(d_x^{(i)}, d_p^{(i)}, d_w^{(i)} | d^{(j)})
= P(d_x^{(i)}|d^{(j)}) P(d_p^{(i)}|d^{(j)}, d_x^{(i)}) P(d_w^{(i)}|d^{(j)}, d_x^{(i)}, d_p^{(i)})
\]

(7)

4 | EXPERIMENTS

In this section, the proposed model is utilized to recognize power equipment in substations to prove its feasibility and reliability. Besides, Mask R-CNN and relation network [13] are also evaluated on this task to compare with the proposed method.

4.1 | Data source

A dataset of pictures taken by inspection robots of a substation in Shanghai, China in August 2018, is created for the experiments. The camera of robots is a RGB-Cam with a resolution of 1920 × 1080 pixels. To guarantee the performance of image recognition, focus-failed, back-light, night-shot and severe-occlusion pictures are not selected. Eventually 751 pictures of power equipment are adopted and annotated by professionals. Typical scenes of power equipment in substations are shown in Figure 5. Through statistics, it is found that the annotation contains 1918 insulators, 330 bushings, 66 arresters, 1326 capacitors, 258 buses, 306 PT/CTs and 210 switches. Generally speaking, using higher resolution pictures as input brings more accurate recognition results at the expense of processing time. However, there will be no increase of accuracy when the resolution is greater than a certain threshold. In our experiments, these pictures are scaled to a resolution of 640 × 448 pixels, which has been tested as the resolution threshold for our dataset.

To test the model performance under the condition of small-sample learning, the size of training dataset is changed. Table 1 gives the configuration. Take Config. No.1 as an example, 600 pictures are randomly chosen from dataset for training. The rest of the pictures are used for test. Five-fold cross validation is adopted, meaning 80% samples in training set are used for model training, while 20% of them are used for validation. To eliminate the effect of sample choosing, experiments are conducted ten times for each configuration, and all the indexes used for evaluation are averaged.

4.2 | Human-level concept learning

Weights of Mask R-CNN, which is at the bottom level of the proposed method, are learned with training set and validation set. Since the dataset of power equipment images is too small to train all layers, only the RPN, RCNN and mask head of the network are trained with our data. As for other layers, trained weights based on COCO dataset are used. This is feasible because these layers are at the bottom of Mask R-CNN
An example of correction process of HLCL. In Step I, it is the original detection results of Mask RCNN. In Step II, it shows the detection confidence, where devices in green are with a high confidence, that in red is with a low confidence. From Step III to Step VIII, HLCL corrects the detection results step by step and are used to extract bottom-level features like edges, colours, curves and their combinations. The backbone network is Resnet 101 [20], which is a deep neural network proven to be effective in extracting features for object recognition. Feature pyramid networks [21] are adopted to help detection objects on different scales. Non-Maximum Suppression [22] is adopted to remove repeated detections.

Conditional probabilities in the correction process are obtained in training set. The validation set is used to obtain hyper-parameters of BCN and HLCL like (controlling the size of neighbourhood) and the threshold of confidence (determining whether an instance is reliable). The value of hyper-parameters is chosen when the model has the best performance on validation set.

An example of correction process of HLCL is given in Figure 6, which is discussed below in detail. In Step I, it is the original detection results of Mask RCNN. In Step II, it shows the detection confidence, where devices in green are with a high confidence, those in red are with a low confidence. The threshold of confidence is set as 0.9. From Step II, it can be found that three instances need correction. In order of the number of reliable neighbours, the leftmost instance is corrected first. The type of this instance given by Mask RCNN is insulator. Considering its neighbourhood, which is marked by a red box in Step III, there is an instance of bushing completely overlapping it. The statistics show that this situation seldom happen. Therefore, HLCL remove this instance as shown in Step IV. Similarly, in Step V, the wrong instance of insulator is corrected as bushing based on the instance of bushing near it; in Step VII, the wrong instance of CT is removed based on the instances of bushing above it.

Comparison of detection results between Mask R-CNN(left column) and Human-level concept learning(right column) are shown in Figure 7, where correct detections are in green, wrong detections are in red and missing objects are in white. Although almost half of the detection results of Mask R-CNN are correct, quite a few wrong detection results lower the quality of object recognition, making automatic fault detection difficult. With human-level concept learning, these wrong detections are corrected based on the relationship between power equipment.

4.3 Comparison with other methods

In this section, HLCL is compared with relation network (RN) [13] and Det59-R128 Network [23], which are parametric and non-parametric object relationship model respectively. Interaction between appearance features and geometry features of objects are considered in RN. In the experiment, two object relation modules are embedded in Mask R-CNN to improve its performance, which is shown in Figure 9. This module, which is essentially a scaled dot-product operation, includes four large weight matrices of learnable parameters, impeding its performance on small datasets. Besides, human knowledge cannot be added into the module due to these inexplicable parameters. To be specific, the structure of Mask R-CNN here is the same as that used in human-level concept learning. Two object relation modules are respectively added after the two fully connected layers in the R-CNN head. Non-Maximum Suppression [22] is adopted to remove repeated detections.

The authors of Det59-R128 Network, on the other hand, proposed a non-parametric model for object relationship. In this network, a knowledge database of probability and a region-to-region graph are built to indicate object relationship, then features of the original region of object and those of related regions are fused for recognition. Figure 10 shows the network structure used in our experiment. Even though non-parametric module reduces the dependence on data, some weaknesses still exist. First, only category of objects is considered when modelling object relationships, while other useful information such as position and size is ignored. The proposed model, however, includes all useful attributes of objects, making object relationship more accurate. Besides, the reliability of the category of detection results are not discussed. Once the category of the detection results is wrong, the object relationship based on this information becomes unreliable. This problem is solved in the proposed method by correcting object category in sequence, where unreliable results are not used in a single correction step.
Specifically speaking, the structure of Det59-R128 Network in the experiment is organized in the following way. DetNet-59 [24] Network is adopted as backbone network. Based on category-level knowledge learnt from training data, a region-to-region graph is used to find related regions from ROIs for each object. Then features of original region of the object and those of related regions are fused for further classification and localization. It should be noted that only the top 128 ROIs are included in the region-to-region graph.

Detection results of Relation Network (left column) and Det59-R128 Network (right column) are shown in Figure 8, where correct detections are in green, wrong detections are in red and missing objects are in white. It can be seen from Figures 7 and 8 that detection results of RN and Det59-R128 Network are better than those of Mask R-CNN but still worse than the proposed method. This is because the proposed method, as a non-parametric probability-learning-based model, is more suitable for small-sample learning than parametric
models like RN. Even though Det59-R128 Network is also non-parametric, it ignores useful attributes such as the position and size of objects and suffers from unreliable category information, which has been solved in the proposed method by iteration of correction.

To further compare these models, the accuracy of results under Config No.1 is given in Table 2, where F1 score is adopted as evaluation metric. For object recognition, the objective includes two aspects: class and position. While class is easy to evaluate, position is evaluated in a carefully designed way. For both ground truth and detection result, their position can be represented by the bounding box, the difference of which can be evaluated by Intersection over Union (IoU) \[ 25 \]. A detection result is judged as accurate if its class is correct and the IOU on its bounding box and that of ground truth is greater than a threshold which is commonly set as 50% \[ 18 \]. Then F1 score is defined as the harmonic mean of precision and recall \[ 26, 27 \].

Some conclusions can be made from Table 2. First, the total F1 score of HLCL is 9.7% higher than that of Mask RCNN. Compared with Mask RCNN, relationship between power equipment is considered in HLCL. Power equipment recognition is a small-sample learning problem, with the characteristics: (1) the type of power equipment is much less than that of common objects; (2) the probability of certain types of equipment appearing together is high. As a result, HLCL can be used in this task and achieves a good performance.

Second, the total F1 score of HLCL is 6.4% higher than that of RN and 5.1% higher than Det59-R128. Compared with the parametric model of RN, HLCL, based on probability learning, models equipment relationship in a human-like way, making it possible to introduce human knowledge as prior. In contrast, the large amount of inexplicable parameters of RN, which are completely learnt from data, become inaccurate with insufficient training samples. As a result, HLCL is more suitable for power equipment recognition than RN. On the other hand, HLCL is better than the non-parametric model of Det59-R128 Network, since it includes more attributes of object such as position and size and overcomes the challenge of unreliable category information by iteration of correction.

Third, F1 score of Mask RCNN varies with equipment type. In fact, there are two factors that largely affect the accuracy of recognition: (1) the number of training samples; (2) the situation of overlapping and occlusion. It can be found from Table 2 that F1 score of Arrester and Bus is comparatively low. This is because the number of training samples for Arrester is quite

---

**TABLE 2** F1 Score of Different Models under Config No.1

|       | HLCL | Mask R-CNN | RN  | Det59-R128 |
|-------|------|------------|-----|------------|
| Total | 0.7921 | 0.6951 | 0.7280 | 0.7411 |
| Insulator | 0.6673 | 0.5818 | 0.6023 | 0.6314 |
| Bushing | 0.7337 | 0.5547 | 0.5810 | 0.5823 |
| Arrester | 0.3567 | 0.3027 | 0.2841 | 0.3021 |
| Capacitor | 0.7783 | 0.4989 | 0.6147 | 0.6232 |
| Bus | 0.3863 | 0.3023 | 0.2451 | 0.3455 |
| PT/CT | 0.6526 | 0.4610 | 0.5481 | 0.5780 |
| Switch | 0.7626 | 0.5456 | 0.5974 | 0.6639 |
small; and the situation of overlapping and occlusion is quite severe for Bus. It should be noted that accuracy of all classes has been largely improved by the proposed method, meaning object relationship modelling is helpful to the recognition of all classes rather than specific ones.

Finally, the tendency of the four models’ performance changing with training-set size is shown in Figure 11. Reliability of the proposed method is demonstrated, since it has a much more stable performance than the other three models. A good model for small-sample learning problems is supposed to be less affected by training-set size within a certain range. Besides, its performance can be further improved with more training samples.

5 | CONCLUSION

HLCL is proposed to model power equipment relationships, which is necessary to improve the performance of power equipment recognition with few training samples. Inspired by human recognition mechanism, HLCL corrects detection results in sequence based on object relationship, which is implemented by maximizing the conditional probability of an object given its neighbourhood. Compared with other object relationship models, HLCL has these advantages: non-parametric, less data-dependent, easy to introduce human experience as prior as well as logical rules. Experiments on our dataset show that HLCL outperforms other three models on this task, making industrial application of this technology possible.

ACKNOWLEDGMENTS

This work is sponsored by the National Key Research and Development Program of China Modes and key supporting technologies for the sustainable development of smart grid based on TCP ISGAN of the International Energy Agency (2019YFE0102900)

REFERENCES

1. Van Nhan, N., et al.: Intelligent monitoring and inspection of power line components powered by UAVs and deep learning. IEEE Power Energy Technol. Syst. 6(1), 11–21 (2019)
2. Haipeng, W., et al.: Research on the technology of indoor and outdoor integration robot inspection in substation. In: IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference. Chengdu, China, pp. 2366–2369 (2019)
3. Rui, G., et al.: A mobile robot for inspection of substation equipments. In: 1st International Conference on Applied Robotics for the Power Industry. Montreal, QC, pp. 1–5 (2010)
4. Shaoqing, R., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2016)
5. Joseph, R., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, pp. 779–788 (2016)
6. Wei, L., et al.: SSD: Single shot multiBox detector. In: European Conference on Computer Vision. Amsterdam, pp. 21–37 (2016)
7. Jifeng, D., et al.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems. Barcelona, Spain, pp. 379–387 (2016)
8. Tsung-Yi, L., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, pp. 2980–2988 (2017)
9. Tsung-Yi, L., et al.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Zurich, pp. 740–755 (2014)
10. Mark, E., et al.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision 88(2), 303–338 (2010)
11. Yong, L., et al.: Structure inference net: object detection using scene-level context and instance-level relationships. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, pp. 6985–6994 (2018)
12. On the Properties of Neural Machine Translation: Encoder Decoder Approaches. https://arxiv.org/abs/1409.1259 (2020). Accessed 10 March 2020
13. Han, H., et al.: Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, pp. 3588–3597 (2018)
14. Massive Exploration of Neural Machine Translation Architectures. https://arxiv.org/abs/1703.03906 (2020). Accessed 9 March 2020
15. Ashish, V., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. Long Beach, CA, pp. 5998–6008 (2017)
16. Lake, B.M., et al.: Human-level concept learning through probabilistic program induction. Science 350(6266), 1332–1338 (2015)
17. Kaiming, H., et al.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, pp. 2961–2969 (2017)
18. Kaiming, H., et al.: Graph Neural Networks: A Review of Methods and Applications. https://arxiv.org/abs/1812.08434 (2020). Accessed 20 March 2020
19. Graph Neural Networks: A Review of Methods and Applications. https://arxiv.org/abs/1812.08434 (2020). Accessed 20 March 2020
20. Neubek, A., Van Gool, L.: Efficient non-maximum suppression. In: International Conference on Pattern Recognition. Hong Kong, Vol. 3, pp. 850–855 (2006)
21. Zhuangzhhuang, T., et al.: Object detection in optical remote sensing images by integrating object-to-object relationships. Remote Sens. Lett. 11(5), 416–425 (2020)
22. DetNet: A Backbone Network for Object Detection. https://arxiv.org/abs/1804.06215 (2020). Accessed 12 Oct 2020
25. Hamid, R., et al.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, pp. 658–666 (2019)

26. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. https://dspace2.flinsters.edu.au/xmlui/handle/2328/27165. Accessed 12 March 2020

27. Dell, Z., et al.: A Bayesian hierarchical model for comparing average F1 scores. In: IEEE International Conference on Data Mining. Atlantic City, NJ, pp. 589–598 (2015)