IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 1, 2017, accepted September 19, 2017, date of publication September 27, 2017, date of current version November 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2757030

Aggregating Deep Convolutional Feature Maps for Insulator Detection in Infrared Images

ZHENBING ZHAO[1], (Member, IEEE), XIAOQING FAN[1], GUOZHI XU[2], LEI ZHANG[1], YINCHENG QI[1], AND KE ZHANG[1], (Member, IEEE)

[1]School of Electrical and Electronic Engineering, North China Electric Power University, Baoding, China
[2]NetEase, Hangzhou, China

Corresponding author: Xiaoqing Fan (18730256293@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant number 61401154, in part by the Natural Science Foundation of Hebei Province under Grant number F2016502101, and in part by the Fundamental Research Funds for the Central Universities under Grant number 2015ZD20.

ABSTRACT Insulator detection using an infrared image is challenged by variance of temperature, orientations, and a cluttered background. A robust and discriminative representation of insulators in electric power systems is needed. This paper proposes a novel method for generating this type of representation in infrared images by taking advantage of high-level discriminative Convolutional Neural Networks (CNNs) to feature the extraction framework and the deformation invariant nature of the Vector of Locally Aggregated Descriptors (VLAD) aggregator. Different from existing methods, we delve deep into the convolutional feature maps. We first extract deep activation maps from convolutional layers of a pretrained deep model and replace the last three fully-connected layers with a VLAD pooling layer to generate the representation of an insulator. Then, we train a Support Vector Machine (SVM) for binary classification. To further verify the effectiveness and robustness of our proposed feature, an insulator detection pipeline based on an object proposal is introduced. The experimental results show that our proposed method can achieve an accuracy of 93%. Meanwhile, the detection results demonstrate that our insulator detection pipeline has satisfied performance goals.

INDEX TERMS Convolutional feature map, Deep Convolutional Neural Network, feature extraction, infrared image, insulators, object detection.

I. INTRODUCTION

Insulators are widely used for conductor insulation and mechanical support of electrical equipment in electric power systems. The malfunction of insulators could cause a short circuit between conductors resulting in serious damage to electrical equipment, and the subsequent failure can cause an unplanned power outage of a whole district, leading to serious collateral injuries and even death [1]–[3]. Thus, electric insulator condition evaluation is a priority in regular inspection and maintenance tasks. Fig. 1 shows the insulators in a power system.

Temperature is an important indicator of insulator conditions. Faulty machinery, improper electrical loading conditions and internal defect of insulators often present abnormal temperature patterns [4]. Fortunately, infrared thermography (IRT) which is a non-contact measuring technique, allows acquisition of the condition data remotely. IRT can instantly visualize the thermal pattern and quickly locate the hot spot, thereby enabling early detection of equipment flaws under operating condition. All these measurements can be conducted without a major shutdown. IRT helps reduce the system shutdown time and maintenance cost [5]. Thus, IRT has greatly increased efficiency in electrical equipment fault diagnosis and gained widespread applications.

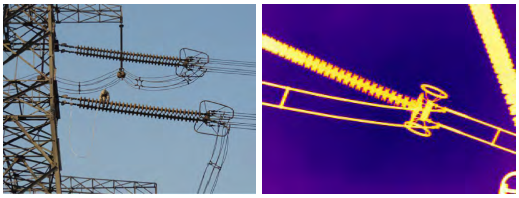


FIGURE 1. Insulator images. Left: visible light image, Right: infrared image taken from a thermal imager.

Since the conventional methods of diagnosing the condition of insulators require well-qualified and experienced personnel for the temperature data analyzing, the procedure of insulator condition evaluation is time-consuming due to the massive quantity of insulators. Recently, research effort has focused on the automated thermal condition analyzing of electrical equipment. The general procedure of electrical equipment inspection is divided into several steps. The first step is to localize the region of the equipment within the infrared image, and the second step is to extract the statistics and other information relevant to the thermal condition of the corresponding region. Finally, the extracted information will be sent to the decision procedure. This means that the proper region detection is the key to the right decision [6].

In the literature, many researchers have proposed methods for finding the regions of electrical equipment in infrared images. Jin *et al.* [7] proposed a method for the insulator segmentation based on Otsu and mathematical morphology. Almeida *et al.* [8] proposed to apply watershed transformation to detect the surge arrester. The watershed algorithm was robust to noise and non-uniform illumination. However, this algorithm needed the initial marks to be well estimated, and the electrical equipment to be detected must be located in the center of the image. Jaffery *et al.* [9] presented a color-based segmentation method to extract hot regions. All the above-mentioned segmentation-based methodologies are applied to a target region. Based on the nature of an infrared image, the visualized pattern is relying on temperature distribution. The object to be detected might be blended with surroundings. The segmentation-based methods rely on the pixel intensity values, and traditional methods like thresholding and clustering tend to be over-segmented [10]. Jadin *et al.* [6], [11] presented an approach for detecting repetitive objects based on scale invariant local features and agglomerative clustering. They applied local invariant key point detection and feature description, and combined it with clustering to detect similar objects in infrared images.

The correct detection of insulators will always contribute to accurate and precise evaluation results. The difficulties of intelligent condition evaluation when using electrical equipment to obtain infrared images is to find the right regions of interest (ROIs). Conventional methods extract these regions based directly on the raw image intensity values and ignore the higher-level object representations. A robust and discriminative representation is one possible solution to this challenging problem.

Image representation is one of the widely concerned problems in computer vision. Typically, an image is represented by a set of local descriptors [12] for various applications. The Bag-of-Features (BoF) model [13] aggregates robust local features [14], [15] like the Scale-Invariant Feature Transform (SIFT) into a compact representation after a series of coding and quantifying strategies. However, these kinds of feature pooling methods discard the spatial relationships between features, and they have limited descriptive ability in generating contour information or when segmenting an object from its surroundings. To generate a more informative image representation, Jégou *et al.* proposed the Vector of Locally Aggregated Descriptors (VLAD) [16], [17], which aggregates local features into a global descriptor based on locality criterion. Residuals are accumulated and concatenated to represent the image.

Despite the success of hand-crafted features and generic image descriptors, Convolutional Neural Networks (CNNs) have demonstrated great success in computer vision and pattern recognition [18], [19] starting from the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [18], which set the state of the art in visual recognition [20]. Many general object detection methods such as Faster R-CNN [21], SSD (Single Shot MultiBox Detector) [22], R-FCN [23] have achieved remarkable success in object detection. CNNs can learn an effective representation from raw-pixel inputs by directly optimizing the given objective. The input image is first sent through several convolutional layers, each of which consists of feature maps and then max-pools the output within local neighborhoods. After a series of sub-sampling and filtering, another series of fully connected layers process the convolutional feature maps into a fixed size feature (4096 D). The final representation with rich semantic information is highly discriminative. However, the feature still preserves spatial information that allows it to retain its sensitivity to spatial transformation. Zeiler *et al.* [24] showed how the reconstructed visualization of the fifth layer activation looks similar to the input image. This means that although max-pooling within each feature map contributes invariance to the small-scale deformations, deep features are still sensitive to affine transformation in object representation [25]. We visualized the feature maps of two insulators in Fig. 2. We can see that the corresponding activations in the intermediate layers are shown in diverse patterns, which means the deep features are sensitive to rotation changes.

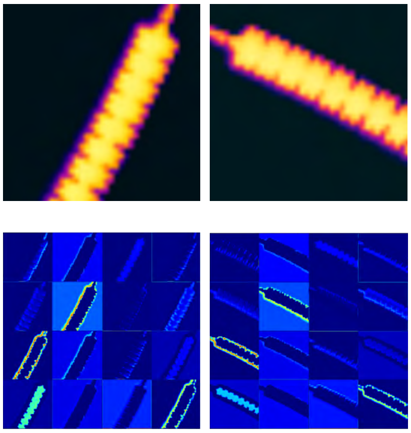


FIGURE 2. Infrared insulator images and their corresponding feature map visualizations. The deep neural activations are highly related to the deformation of the input images. Top row: Images from our infrared insulator dataset, with the same object but with a different view point. Bottom row: Visualized convolutional feature activations of the corresponding images.

When compared to general visible image classification and detection tasks, the insulators from infrared images vary extremely from aspect ratio and orientation due to their

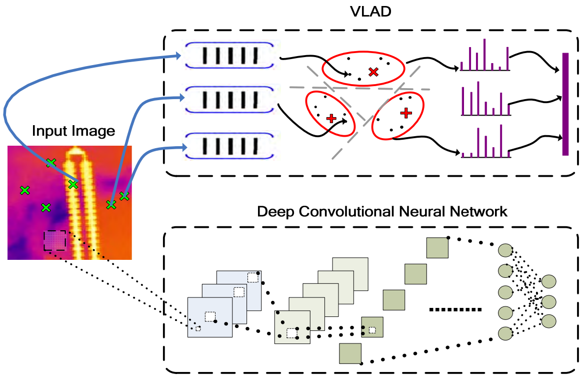


FIGURE 3. VLAD and CNN pipelines for generating global image descriptors.

special collection process, which is different from general image datasets, such as Caltech Pedestrians [26], INRIA [27], and ETH [28]. Objects of these datasets generally share a similar aspect ratio, shape and orientation. With the purpose of generating features with more deformation invariance, the VLAD aggregator [13], [16], [17], [25] comes to mind. The local features can be aggregated into a compact and discriminative global image descriptor with scale and orientation invariance by VLAD aggregation. The pipelines of CNN and VLAD are presented in Fig. 3.

By taking advantage of high-level discriminative CNN activations and the deformation invariant nature of the VLAD aggregator, we propose a method for aggregating deep convolutional feature maps for insulator detection in infrared images.

Different from the existing CNN-VLAD methods, we delved into the deep model by generating rich information from convolutional feature maps. We propose a convolutional feature map pooling method, which greatly augments the invariance of the deep features. We first extract deep activations from convolutional feature maps of infrared insulator images based on a pre-trained model. As a general classification task, a classifier like Support Vector Machine (SVM) is directly trained on these learned features from the fully-connected (*fc*) layers. Instead, we modify the deep feature extraction framework by replacing the last three *fc* layers with the VLAD pooling layer. We named this feature-generating strategy the *convolutional feature map VLAD* or ConvVLAD for short. The feature aggregated from ConvVLAD is highly discriminative and an SVM classifier is trained for infrared insulator classification and detection. The main contributions of this paper are:

- A compact and representative feature generating method is proposed based on Convolutional Neural Network activations and VLAD. In this model, the fully connected pooling is replaced by the VLAD pooling strategy.
- An object proposal-based insulator detection pipeline of the infrared image is presented, and the Edge box with a slight modification for object proposals is introduced.

The remainder of this paper is organized as follows: Section II presents the Deep Convolutional Neural Network feature extraction strategy and VLAD feature map aggregating method. In Section III, experimental results and data analysis are presented. Section IV summarizes our method.

II. PROCEDURE OF THE PROPOSED METHOD

In this section, we introduce the deep convolutional feature map aggregating method in detail. The designed feature is robust and discriminative for insulator detection. In the proposed framework, we first feed an infrared insulator image into a pre-trained CNN model for deep activation extraction. Then, the last three *fc* layers are replaced by VLAD feature pooling. The overall feature-generating framework can be divided into several steps as illustrated in Fig. 4.

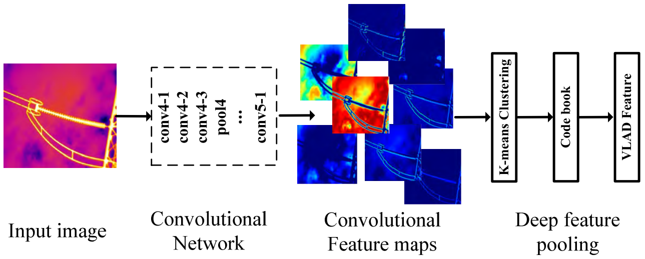


FIGURE 4. The global image descriptor generating framework. An infrared image of insulator is fed into an ImageNet pre-trained CNN model, and then deep convolutional activations are extracted. The feature maps of the convolutional layer are vectorized and pooled by VLAD coding and quantifying. Finally, the final image representation is generated.

A. Deep Convolutional Activation Extraction

Our work is inspired by the remarkable success of the deep CNN in all kinds of computer vision tasks. CNNs are biologically inspired models that simulate the human visual cortex, which can extract rich semantic information from an image or region. A CNN consists of a succession of convolutional and max-pooling layers, and each layer connects with the previous layer. All parameters are adjustable through the optimization of the minimal loss on the training set. In this paper, we employ the VGG16 architecture released by [20], which achieves the state-of-the-art accuracy on ILSVRC classification and localization tasks. The VGG16 network consists of 13 convolutional layers and three fully connected layers, and the convolutional layer is comprised of 3 × 3 small convolutional filters as well as five max-pooling layers. The detailed parameters of the network architecture are presented in Table 1. On each convolutional layer l, a convolution operation of its M^{l-1} input maps from previous layer *l-1*, with a filter of size $k_l \times k_l$, is conducted. The resulting output is the summations of the responses with a nonlinear function:

$$F_j^l = \sigma\left(\sum_{i=1}^{M^{l-1}} F_i^{l-1} * W_{ij}^l + b_j^l\right) \tag{1}$$

TABLE 1. Architecture of the deep CNN feature extractor.

N	Layer	Dimension			Kernel	Stride	Padding
		W	H	D	-	-	-
0	input	224	224	3	-	-	-
1	conv1_1	224	224	64	3	-	1
2	conv1_2	224	224	64	3	-	1
3	pool1	112	112	64	2	2	-
4	conv2_1	112	112	128	3	-	1
5	conv2_2	112	112	128	3	-	1
6	pool2	56	56	128	2	2	-
7	conv3_1	56	56	256	3	-	1
8	conv3_2	56	56	256	3	-	1
9	conv3_3	56	56	256	3	-	1
10	pool3	28	28	256	2	2	-
11	conv4_1	28	28	512	3	-	1
12	conv4_2	28	28	512	3	-	1
13	conv4_3	28	28	512	3	-	1
14	pool4	14	14	512	2	2	-
15	conv5_1	14	14	512	3	-	1
16	conv5_2	14	14	512	3	-	1
17	conv5_3	14	14	512	3	-	1
18	pool5	7	7	512	2	2	-
19	*fc*6	1	1	4096	-	-	-
20	*fc*7	1	1	4096	-	-	-
21	*fc*8	1	1	1000	-	-	-

where l indicates the layer, F_j^l and F_j^{l-1} are the feature maps from layer l and layer *l-1* respectively, and W_{ij}^l is a filter of size $k_l \times k_l$. b_j^l indicates the bias, and $\sigma(\cdot)$ is the ReLU (Rectified Linear Unit) function:

$$\sigma(x_i) = \begin{cases} x_i, & \mathit{if}\ x_i > 0 \\ 0, & \mathit{if}\ x_i \leq 0 \end{cases} \tag{2}$$

The architecture of VGG16 contains more than 138 million parameters, and it would be time-consuming to train this huge network. Instead of optimizing the whole model on the limited dataset, we adopted an ImageNet pre-trained CNN model to extract deep feature maps. We use the features from intermediate layers of CNNs rather than the fully-connected layers for representation. We mainly focus on the activations on the conv5_1 layer (the 11th convolutional layer), which is composed of $h \times w \times d$ tensor. As discussed in [29] and [30], features located in different layers have different meanings. Semantic concepts can be acquired by higher layers, while lower layers are more sensitive to the appearance of object variations. The feature map is sized $h \times w$, and d is the number of feature maps. For example, an image of size 224 × 224 is fed into a VGG16 network through forward computation, and the extracted feature maps are 14 × 14 × 512. All these feature maps can be integrated to generate a discriminative representation of the input image or region. Visualization of the feature maps from conv3_1 (256) and conv5_1 (512) of the same input image are shown in Fig. 5.

B. Deep Convolutional Feature Maps Aggregation

Let $\mathbf{F}_l = [f_1, \ldots, f_i, \ldots, f_d] \in \mathbf{R}^{h\times w}$ denote the feature activation set from the l-th convolutional layer extracted from a pre-trained model, f_i is the i-th feature map which is a 2-D matrix with the size of $h \times w$, and d indicates the total number of the feature maps. The feature map extraction is given by:

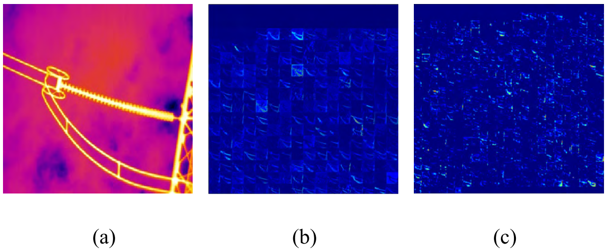


FIGURE 5. Visualization of the feature maps from (a) conv3_1 and conv5_1 of the same input original image, (b) the size of the feature map in conv3_1 is 56 × 56 with 256 dimensions, and (c) 14 × 14 with 512 dimensions from the conv5_1 layers.

$$\mathbf{F}_l = \varphi_l(I)_{l=1,\ldots,N} \tag{3}$$

where φ_l extracts l-th layer activations from image I. All these feature maps are the representations of the local regions from the original input image, as SIFT like local features [14]. To pool these feature maps into a fixed-dimension descriptor, we first vectorize $\mathbf{F}_l$ into a one-dimensional feature set $\mathbf{X}_l$, where $\mathbf{X}_l = [x_1, \ldots, x_i, \ldots, x_d]$, $x_i \in \mathbf{R}^{1\times D}$, and $\mathrm{D} = h \times w$. The transformation is represented by $\mathbf{T}$ and the process is given as follows:

$$\mathbf{F}_l \underset{\mathrm{T}}{\Rightarrow} \mathbf{X}_l, [x_1, \ldots, x_i, \ldots, x_d] \in \mathbf{R}^{1\times D} \tag{4}$$

There are several methods for integrating features into a compact high-level feature representation. For example, Li [13] proposed the Bag-of-Features method, which used an order-less collection of local features to represent an image. Benefiting from the power of local invariant features, such methods have achieved excellent performance. However, they still suffer from the drawback of the limited semantic description of local descriptors. Jégou [16] introduced the Vector of Locally Aggregated Descriptors (VLAD). This feature pooling method was designed to be low dimension yet highly distinguishing.

We first perform deep convolutional activation extraction and vectorization on all the training images. After all the deep features from the training set are extracted, the codebook is generated by a k-means clustering algorithm. When the clustering is finished, the centers are assigned as visual words. The codebook is a $k \times D$ matrix, composed of k visual words with dimension D.

Algorithm 1 explains the procedure of the VLAD feature generation. $\mathbf{F}_l$ represents the deep feature maps extract from VGG16. With a trained codebook $\mathbf{C} = \{c_1, c_2, \ldots, c_k\}$, we can aggregate deep features into a compact representation, and this is called feature pooling. Given a whole or a region of an infrared image, first, deep features are extracted. Then, each vectorized convolutional map x_n is associated to its nearest visual word $c_n = NN(x_n)$, and NN indicates the nearest neighbor search. The nearest $c(x_n)$ can be indexed

Algorithm 1 Deep Convolutional Feature Generation

```
Input:
    Deep feature maps $\mathbf{F}_l = [f_1, \ldots, f_i, \ldots f_d]$,
    Codebook $\mathbf{C} = \{c_1, c_2, \ldots, c_k\}$
Output :
    Aggregated feature vector $v = \{v_1, v_2, \ldots, v_k\}$
Procedure:
1   Vectorization $\mathbf{F}_l \to \mathbf{X}_l$
2   for $i = 1$ to $n$ do
3       $t$ = index argmin $d|c_j, x_i|, j \in \{1, 2, \ldots, k\}$
4       $v'_t = v'_t + (x_i - c_t)$
5   end for
6   $v' = \{v'_1, v'_2, \ldots, v'_k\}$
7   $v = \frac{v'}{\|v'\|^2}$
```

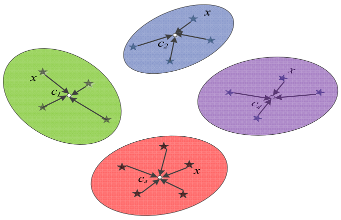


FIGURE 6. Deep feature residual computing.

by (5), where $d|\bullet|$ denotes the distance between two features.

$$c(x_n) = arg \min_{c_i} \ d\,|c_i, x_n| \tag{5}$$

Fig. 6 is a simple illustration on how the deep features are pooled. For each visual word c_i, the residuals of x_n-c_i are computed as shown in (6), where $i = 1, 2, \ldots$, and k indexes the visual words.

$$v_i = \sum_{s.t.\ NN(x_n)=c_i} x_n - c_i \tag{6}$$

The dimension of v_i is the same as deep feature x_n, and in this paper, we use the conv5_1 descriptor of 196-dimension. Then, the accumulated residuals corresponding to all the visual words are concatenated, and hence the feature vector v we get is a $196 \times k$ dimension vector, as presented in (7). Moreover, the vector v is subsequently L_2-normalized by (8).

$$v' = \left[v'_1, \ldots v'_i, \ldots v'_k\right], \quad v'_i \in R^{196} \tag{7}$$

$$v = \frac{v'}{\|v'\|^2} \tag{8}$$

The pooled vector is the representation for an input image. The VLAD vector generation is shown in Algorithm 1. The proposed deep feature aggregating architecture is illustrated in Fig. 7.

Compared with other classifiers such as neural networks [31], the classification tree model [32] and k Nearest Neighbor (kNN) [33], SVM outperformed these models on small datasets, and when using the nonlinear and high dimensional features. All these generated feature vectors were then fed into an SVM with an RBF kernel. This is an optimal problem as:

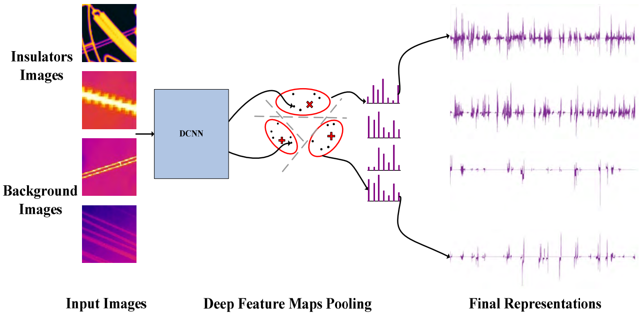


FIGURE 7. The compact representation generation of insulator images and background images are based on the proposed ConvVLAD architecture.

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{j} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K\left(x, x_j\right) - \sum_{j=1}^{l} \alpha_j$$
$$s.t. \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \ i = 1, \cdots, l \tag{9}$$

where C is a constant which controls the punishment for misclassified samples and the nonlinear decision function is shown in (10):

$$f(x) = sign\left(\sum_{i=1}^{n} a_i y_i K(x, x_i) + b\right) \tag{10}$$

where α_i is the Lagrange multiplier, y_i is the label, x_i is the support vector and x are the 4096-dimensional feature. $K(x, x_k)$ is the kernel function, and in this case, it can be formulated as:

$$K(x, x_i) = \exp\left(\frac{-\|x - x_i\|^2}{2\sigma^2}\right) \tag{11}$$

The constant σ controls the width of the RBF kernel.

III. EXPERIMENTAL RESULTS AND DATA ANALYSIS

In this section, the effectiveness of our proposed ConvVLAD strategy on the Convolutional Neural Network is demonstrated. First, the dataset of the infrared insulator is introduced. Based on this dataset, we evaluate the invariance of the proposed feature through classification accuracy; moreover, comparisons with other deep features and hand-crafted features are presented. The insulator detection results are also presented.

A. Infrared Insulator Dataset

The infrared insulator image dataset consists of 672 positive samples and 1012 negative samples. All the infrared images were acquired from the infrared thermal imager. Insulator images were manually cropped from the original images taken in the power substations.

We divided the dataset into two parts: 30% for training and the remaining 70% for testing. All the training samples were labeled with "*insulator*" and "*non-insulator*". Sample images from the dataset are shown in Fig. 8.

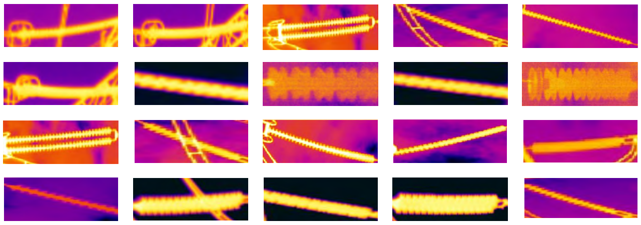


FIGURE 8. Sample images of infrared insulators.

B. Experiment Setup

All the experiments were conducted on a computer with Intel Xeon E5-2609v2 (2.6GHz), 16GB RAM, NVIDIA Tesla K20C with 5GB GDDR5. The OS is a 64-bit Ubuntu 14.04, and our method was deployed by MATLAB 2015a. We employed the famous Caffe framework for deep feature extraction, and the VGG16 deep model was pretrained on ImageNet. We removed the last three fully-connected layers and extracted the activations from the conv5_1 layer for feature aggregation.

C. Classification Results

The performance parameter of the classifier is calculated as given.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where TP is the true positive, TN the true negative, FP the false positive, FN is the false negative. And TP, TN range from 0 to 470, FP, FN range from 0 to 708 in our experiments. The proposed deep feature vector generation was conducted on each image in the training set, and an RBF SVM classifier was trained for classification. To evaluate accuracy between different classifiers, we trained the three most common classifiers: SVM, *k*NN [33], and the Classification Tree [32], and we employ 5-fold cross validation to get more accurate results. The classification accuracy on the test set is presented in Table 2.

From the table above, we can see that SVM is a suitable choice for classification of insulators, and a better classifier is of the same importance as a feature. SVM is widely used because of its ability to avoid overfitting and its excellent performance on small datasets. Different features were also compared and classification accuracy of the state-of-the-art descriptors is presented in Table 2. We evaluated the performances of the newly proposed descriptors like Speeded Up Robust Features (SURF) [15], Oriented FAST, and Rotated BRIEF (ORB) [34] and Binary Robust Invariant Scalable Keypoints (BRISK) [35] via the VLAD pooling method. Due to the deformable invariant nature, the hand-crafted features outperform the features from fully connected layers (*fc*6 and *fc*7) by a large gap. For hand-crafted features, the SURF feature has the most robust invariance, achieving 86.5874% in accuracy. ORB and BRISK achieved accuracy of 80.1358% and 80.8149%, respectively, which is inferior to SURF. We also compared this method with our recently proposed method (F-BRISK, FAST-BRISK Encoding) [36] which achieved 89.1341%. Compared with hand-crafted features, the activations from fully connected layers are less invariant to deformable transformations. We also perform the classification based on the other well-known AlexNet [18] model. Both networks cannot achieve good results, and this might reveal the limitation of the activations from the fully connected layers. The SVM classifier might be less tuned than *k*NN and Classification Tree based on fully connected features. With the VLAD pooling strategy, invariance and robustness of the deep features can be augmented, and our proposed method achieves the best accuracy of 93.4634%.

TABLE 2. Classification accuracy of different methods.

Feature representations	SVM	*k*NN	Classification Tree
SURF	86.5874%	54.4992%	76.9950%
ORB	80.1358%	82.6825%	60.9508%
BRISK	80.8149%	69.2699%	82.0882%
F-BRISK[34]	89.1341%	82.0882%	82.6825%
AlexNet(*fc*6)	54.0747%	75.8913%	73.0900%
AlexNet(*fc*7)	51.1036%	76.0611%	75.9762%
VGG16(*fc*6)	50.2546%	75.9762%	75.2122%
VGG16(*fc*7)	50.8489%	77.5891%	76.6553%
Proposed method	**93.4634%**	91.6808%	71.4770%

D. Classification Accuracy of Feature Maps from Different Layers

According to the results in Table 2, deep convolutional feature maps pooled via VLAD can remain invariant to deformable transforms. As discussed above, features located in different layers have different meanings. Semantic concepts can be acquired by higher layers, while lower layers are more sensitive to appearance variation of objects. We extract deep convolutional feature maps from different layers of the VGG16 network architecture. Activations from different depths share various semantic and deformable information [29], [30]. We evaluate the performance of four kinds of SVM classifiers: linear SVM, SVM with polynomial kernel, SVM with RBF kernel and SVM with sigmoid kernel. The SVM classifiers are implemented based on the LibSVM package.

The classification results of the four SVM classifiers proved the efficiency of the proposed convolutional feature map aggregation method. The VLAD pooling strategy outperformed the fully connected pooling by a large step, and the SVM classifier with RBF kernel achieved the best accuracy of 93.4634% on conv5_1 layer. The classification

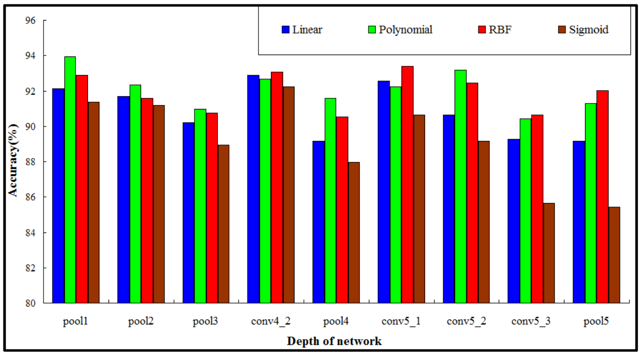


FIGURE 9. **Classification performance on test set with pooled feature maps from different layers.**

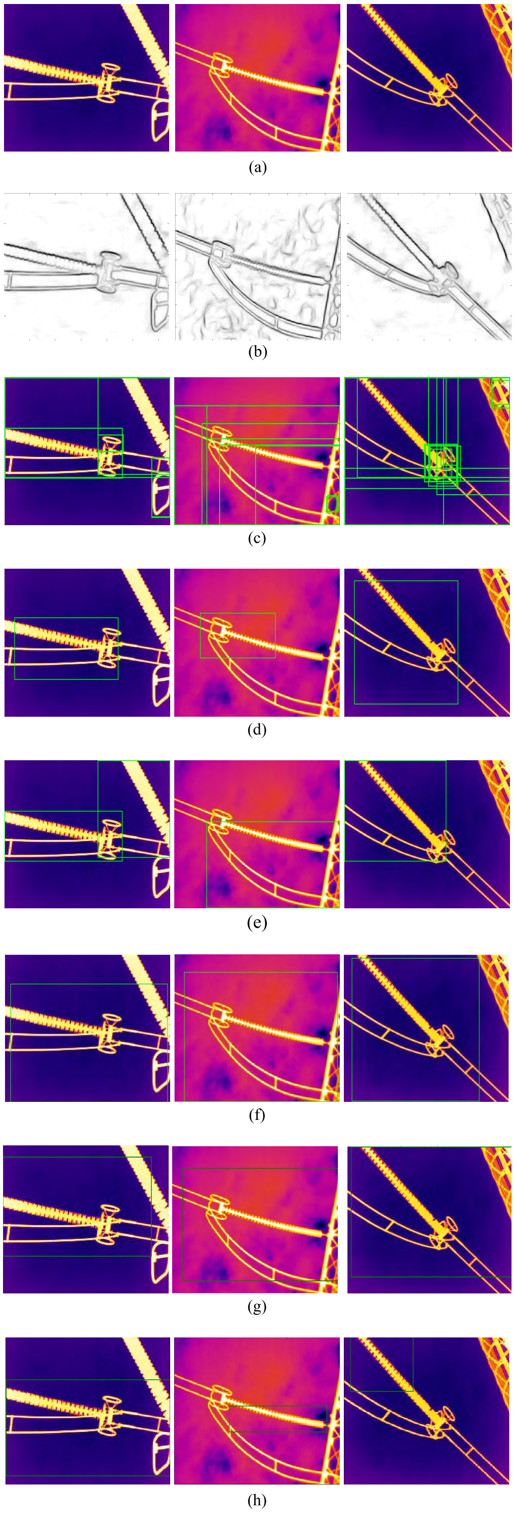
TABLE 3. **Details of the Classification Results of RBF-SVM of Different Intermediate Layers (codebook size = 100).**

Depth	Size of the feature maps	Descriptor length	Accuracy(%)
pool 1	112×112×64	1254400	92.8692
pool 2	56×56×128	313600	91.5959
pool 3	28×28×256	78400	90.7470
conv4_2	28×28×512	156800	93.0390
pool 4	14×14×512	19600	90.5772
conv5_1	14×14×512	19600	**93.4634**
conv5_2	14×14×512	19600	92.4448
conv5_3	14×14×512	19600	90.6621
pool 5	7×7×512	4900	92.0203

performance on the test set is illustrated in Fig. 9, and the detailed parameters of the descriptors are presented in Table 3. The descriptor generated based on the feature maps of the conv4_2 layer also achieved excellent accuracy at 93.0390%. However, the length of the conv4_2 descriptor is 156800, which is eight times larger than the conv5_1 descriptor. Besides, the SVM with a polynomial kernel also achieved comparable accuracy on the pool 1 layer while the length of the descriptor was 1254400. Computing and storing these large descriptors are a huge burden to a limited computation resource.

E. Insulator Detection

With the robust feature extraction strategy, we constructed a coarse insulator localization pipeline for object detection. Instead of applying a sliding-window based method for the candidate region search, we adopted the object proposal method for the generation of the regions of interest (ROIs), which is much faster than the sliding window method. For insulators, Edge boxes [37] are good for preserving the insulators' edge information which is very important for insulators detection. So in this paper, we applied the state-of-the-art Edge boxes to generate the region proposals within infrared images; then, ConvVLAD features were extracted from these regions. A trained SVM classifier was used for the coarse object localization. After all the candidate windows were classified, we used non-maxima suppression (NMS) to locate the final location of an insulator. Detection results are shown in Fig. 10.


(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

FIGURE 10. **Insulator detection pipeline: (a) original infrared images, (b) edges detected by Edge box algorithm, (c) bounding boxes generated based on edges, (d) detection results based on SURF and VLAD features, (e) detection results based on ConvVLAD features, (f) detection results based on Faster R-CNN, (g) detection results based on R-FCN, (h) detection results based on SSD.**

We constructed Edge boxes with a slight modification which are good for preserving the insulators' edge information. Since only a few number of the bounding boxes contained the desired objects, we limited the number of the generated bounding boxes by setting an empirical threshold *Th*. This simple method can effectively reduce the computation cost yet still give an excellent performance. From the result of Fig. 10, we can see that the insulators in infrared images have arbitrary scale and rotation. We compared the detection results based on Conventional SURF feature, our proposed detection pipeline based on ConvVLAD feature and effective general detection framework include Faster R-CNN, R-FCN, SSD. Our method achieves the best results. However, for other methods include SURF, Faster R-CNN, R-FCN, SSD, some of the insulators were missed or inaccurately located, and false alarm rate is higher. Except inaccuracy, the methods based on Faster R-CNN, R-FCN, SSD need to fine-tune and waste a lot of time. Our proposed method can effectively locate the insulators precisely and consume less time. The experimental result indicates the robustness and effectiveness of the proposed deep-feature aggregating strategy.

IV. CONCLUSIONS

In electrical equipment condition evaluation, a powerful and robust feature is of crucial importance, especially for the object detection tasks in infrared images. To generate a discriminative and robust feature representation, we delved deep into the convolutional activation maps for deep feature aggregation. Faced with the various deformable transformations of the insulators, our method can be summarized as follows.

1) Different from other CNN-VLAD methods, we generated rich information from mid-level convolutional feature maps. We propose a convolutional feature map pooling method, which greatly augments the invariance of the deep features. A compact and representative feature generating method is proposed based on deep convolutional neural activations.
2) In this model, we integrated the VLAD pooling method into the feature extraction architecture, and deep activation maps are vectorized and encoded with more spatial invariance.
3) To detect insulators in infrared images, we applied an Edge box with a slight modification for object proposals. The experimental results verify the effectiveness and robustness of our proposed method.

As the deep feature aggregating is time-consuming, future work can be focused on speeding up the extraction process and be extended for real-time detection applications.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NVIDIA Corporation with the kind donation of the GPU used for this research.

REFERENCES

[1] S. Han, R. Hao, and J. Lee, "Inspection of insulators on high-voltage power transmission lines," *IEEE Trans. Power Del.*, vol. 24, no. 4, pp. 2319–2327, Oct. 2009.

[2] Y. Tu *et al.*, "Effect of moisture on temperature rise of composite insulators operating in power system," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 4, pp. 2207–2213, Aug. 2015.

[3] Y. Liu and B. X. Du, "Recurrent plot analysis of leakage current in dynamic drop test for hydrophobicity evaluation of silicone rubber insulator," *IEEE Trans. Power Del.*, vol. 28, no. 4, pp. 1996–2003, Oct. 2013.

[4] M. S. Jadin and S. Taib, "Recent progress in diagnosing the reliability of electrical equipment by using infrared thermography," *Infr. Phys. Tech.*, vol. 55, no. 4, pp. 236–245, Jul. 2012.

[5] S. Bagavathiappan, B. B. Lahiri, T. Saravanan, J. Philip, and T. Jayakumar, "Infrared thermography for condition monitoring—A review," *Infr. Phys. Tech.*, vol. 60, nos. 35–55, Sep. 2013.

[6] M. S. Jadin, S. Taib, and K. H. Ghazali, "Finding region of interest in the infrared image of electrical installation," *Infr. Phys. Tech.*, vol. 71, nos. 329–338, Jul. 2015.

[7] L. Jin and D. Zhang, "Contamination grades recognition of ceramic insulators using fused features of infrared and ultraviolet images," *Energies*, vol. 8, no. 2, pp. 837–858, Jan. 2015.

[8] C. A. L. Almeida *et al.*, "Intelligent thermographic diagnostic applied to surge arresters: A new approach," *IEEE Trans. Power Del.*, vol. 24, no. 2, pp. 751–757, Apr. 2009.

[9] Z. A. Jaffery and A. K. Dubey, "Design of early fault detection technique for electrical assets using infrared thermograms," *Int. J. Elect. Power Energy Syst.*, vol. 63, pp. 753–759, Dec. 2014.

[10] H. Zou and F. Huang, "A novel intelligent fault diagnosis method for electrical equipment using infrared thermography," *Infr. Phys. Technol.*, vol. 73, pp. 29–35, Nov. 2015.

[11] M. S. Jadin, K. H. Ghazali, and S. Taib, "Detecting ROIs in the thermal image of electrical installations," in *Proc. IEEE Int. Conf. Cont. Syst., Comput. Eng. (ICCSCE)*, Nov. 2014, pp. 496–501.

[12] K. Mopuri and R. Babu, "Object level deep feature pooling for compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 62–70.

[13] L. Fei-Fei and P. Perona, "A Bayesian heirarcical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun. 2005, pp. 524–531.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[15] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, May 2006, pp. 404–417.

[16] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 3304–3311.

[17] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2012, pp. 1097–1105.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Sep. 2014, pp. 1717–1724.

[20] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[22] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 21–37.

[23] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2016, pp. 379–387.

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 818–833.

[25] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 392–407.

[26] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[28] A. Ess, B. Leibe, K. Schindler, and L. V. Gool, "A mobile vision system for robust multi-person tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[29] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3119–3127.

[30] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4749–4757.

[31] L. Araaba, Z. Al-Hamouz, and H. Al-Duwaish, "Estimation of high voltage insulator contamination using a combined image processing and artificial neural networks," in *Proc. IEEE 8th Int. Power Eng. Optim. Conf. (PEOCO)*, Langkawi, Malaysia, Mar. 2014, pp. 214–219.

[32] A. K. Chaou, A. Mekhaldi, and M. Teguar, "Elaboration of novel image processing algorithm for arcing discharges recognition on HV polluted insulator model," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 2, pp. 990–999, Apr. 2015.

[33] R. C. Sierra, O. Oviedo-Trespalacios, J. E. Candelo, and J. D. Soto, "Assessment of the risk of failure of high voltage substations due to environmental conditions and pollution on insulators," *Envir. Sci. Pollution Res.*, vol. 22, no. 13, pp. 9749–9758, Jan. 2015.

[34] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 2564–2571.

[35] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 2548–2555.

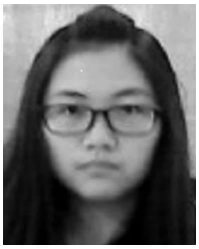
[36] Z. Z. Zhao, G. Z. Xu, and Y. C. Qi, "Representation of binary feature pooling for detection of insulator strings in infrared images," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 5, pp. 2858–2866, Nov. 2016.

[37] C. L. Zitnick and D. Piotr, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 391–405.

ZHENBING ZHAO was born in Jiangsu, China, in 1979. He received the B.S., M.S., and Ph.D. degrees from North China Electric Power University, Baoding, in 2002, 2005, and 2009, respectively. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, North China Electric Power University. His research interests include artificial intelligence, intelligent detection of electrical equipment, and image processing.

XIAOQING FAN was born in 1992. She received the degree in electronic information engineering from the Hebei University of Science and Technology, Shijiazhuang, in 2015. She is currently pursuing the master's degree with North China Electric Power University. Her research interests include infrared image recognition and deep learning.

GUOZHI XU was born in 1992. He received the B.S. and M.S. degrees from North China Electric Power University, Baoding, in 2014 and 2017, respectively. He is currently an Engineer with NetEase. His research interests include infrared image detection, deep learning, artificial intelligence, and image processing.



LEI ZHANG was born in 1992. She received the B.S. degree in communication engineering from the China University of Geosciences, Wuhan, in 2015. She is currently pursuing the master's degree with North China Electric Power University. Her research interests include image detection and deep learning.

YINCHENG QI was born in 1968. He received the B.S., M.S., and Ph.D. degrees from North China Electric Power University, Baoding, in 1990, 1998, and 2009, respectively. He is currently a Professor with North China Electric Power University. His research interests include electric power system communication and information processing.

KE ZHANG was born in 1980. He received the B.S. and M.S. degrees from North China Electric Power University, Baoding, in 2003 and 2006, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, North China Electric Power University. His research interests include deep learning, robot navigation, natural language processing, and spatial relation description.