Multimodal Image Fusion and Classification of Power Equipment Using Non-Subsampled Contourlet Transform and Adaptive Pulse-Coupled Neural Network

Bingyang Zheng*, Haoxiang Hu
School of Information Engineering, Zhengzhou University of Technology, Zhengzhou, Hehan, 450044 China
Email: zhengby_2021@outlook.com, 20218909@zzut.edu.cn
*Corresponding author

Keywords: multimodal image recognition, power equipment, image fusion, NSCT, APCNN

Received: March 27, 2025

This paper presents a multimodal image fusion and classification method for power equipment based on the Non-Subsampled Contourlet Transform (NSCT) and Adaptive Pulse-Coupled Neural Network (APCNN). The approach begins with image normalization, geometric alignment, and adaptive noise filtering as preprocessing steps. The NSCT is then applied to decompose input images into low- and high-frequency subbands. Low-frequency components are fused using phase congruency weighting to retain energy features, while high-frequency subbands with structural details are selectively fused using APCNN for precise edge and contour extraction. For efficiency, subbands beyond the fifth decomposition level use local energy maximization for fusion. Experiments were conducted on a dataset of 3,000 images of transformers, current transformers, and disconnectors collected by inspection robots. The model achieved maximum recognition accuracies of 99.39% for transformers, 99.57% for current transformers, and 98.74% for disconnectors. The average classification time per image was 2.36 seconds. Compared with APCNN, PCNN, LeNet, AlexNet, and SVM, the proposed NSCT-APCNN model demonstrated superior performance in accuracy, F1-score, and processing speed. This work provides an effective and scalable solution for real-time multimodal image classification in substation inspection scenarios, with potential for extension to fault detection in smart grids.

Povzetek: Algoritem NSCT-APCNN združuje fazno kongruenco in večnivojsko nevronsko fuzijo za kvalitetno in časovno učinkovito prepoznavo slik elektroopreme v realnem okolju robotskih inšpekcij.

1 Introduction

With the advancement of information technology, research on computer visualization, sensor-based imaging, and image analysis continues to progress, making multi-source information fusion one of the most promising research topics [1]. By equipping inspection robots with visual instruments such as cameras, comprehensive and rapid detection of power equipment can be achieved [2]. Intelligent inspection systems collect vast amounts of image data, requiring efficient and accurate image screening and processing.

Traditional image analysis techniques for detection primarily rely on classical image processing algorithms, where manually extracted features are mostly low-level visual characteristics [3]. Compared to purely learned features, these handcrafted features offer stronger interpretability but exhibit poorer adaptability to diverse data [4]. Deep learning algorithms have been applied in image fusion. These algorithms can train parameters by learning from large volumes of labeled data, enabling adaptive representation of multi-scale image features without human intervention, and have been applied in more image fusion tasks [5,6].

However, power equipment images have unique properties. Their imaging is affected by multiple factors, including the operating status of the equipment, environmental lighting conditions, shooting angles, and the complex structure of the equipment itself. Different types of power equipment, such as transformers, current transformers, and disconnectors, have significant differences in appearance, texture, and detailed features. Even images of the same type of equipment under different working conditions vary. These complex characteristics mean that when manually annotating their training data, annotators not only need to master basic knowledge of image processing but also possess professional domain knowledge, such as the operating principles of power equipment, types of faults, and their characteristics. Only in this way can they accurately identify and mark key information in the images, such as the location of equipment components and potential fault points. Consequently, training robust models remains challenging. Processing large datasets is time-consuming, making it difficult to meet the real-time requirements of multimodal power equipment image recognition. Additionally, unstable training networks may lead to reduced fusion accuracy[7]. Moreover, due to variability among multimodal sensors, power equipment image fusion still faces significant challenges in precise detail discrimination, noise suppression, and robustness[8].

This research aims to propose a multimodal image fusion and recognition algorithm for power equipment based on the Adaptive Pulse-Coupled Neural Network (APCNN). Specifically, we aim to investigate the following research questions:

(1) Can the integration of Non-Subsampled Contourlet Transform (NSCT) with APCNN enhance recognition accuracy of power equipment images compared to baseline methods such as APCNN, PCNN, LeNet, AlexNet, and SVM?

(2) Can this approach maintain or improve real-time performance, as measured by average inference time per image, despite added preprocessing complexity?

(3) Does the proposed method improve robustness and fine-detail preservation in multimodal image fusion tasks where standard CNN models show limitations?

The algorithm is designed to address issues such as low precision in detail discrimination, weak noise suppression capabilities, and poor robustness when existing algorithms process power equipment images. It also aims to overcome the shortcomings of traditional and deep learning algorithms in image recognition, improve the high accuracy and low inference latency of recognition, and provide reliable technical support for the intelligent inspection of power equipment. Its theoretical significance lies in offering new ideas for algorithm research in related fields, enriching the theoretical system of image fusion algorithms, and providing references for subsequent research. Practically, it enables precise recognition by acquiring multimodal images through inspection robots, improving the efficiency and accuracy of power equipment inspection, reducing labor costs and misjudgment rates, ensuring the stable operation of power systems, minimizing the risk of power outages caused by equipment failures, and generating significant economic and social benefits.

Table 1: Performance of existing methods on multimodal power equipment image recognition

Method	Acc	F1-Score	Processing	Dataset	Limitations
SVM	88.42	0.87	3.85	Substation	Poor adaptability
LeNet	90.15	0.89	3.42	Substation	Limited to shallow feature extraction
AlexNet	93.04	0.92	3.10	Substation	High computational cost
PCNN	92.21	0.91	2.75	Substation	Sensitive to parameter tuning
APCNN	95.76	0.94	2.54	Substation	Lacks multiscale structural detail preservation
Ours	96.71	0.96	2.36	Substation	None observed

2 Method description

2.1 Overall idea of the algorithm

The proposed power equipment image fusion method follows a systematic three-stage processing flow (Figure 1). In the preprocessing phase, input images first undergo normalization to ensure dimensional consistency, followed by geometric alignment and adaptive filtering for noise reduction. The core fusion process begins with Non-Subsampled Contourlet Transform (NSCT) decomposition, where source images are decomposed into multi-scale low-frequency approximation coefficients and directional high-frequency details. For fusion, we employ distinct strategies: low-frequency components are fused using a phase congruency weighting scheme to preserve energy information, while high-frequency components are processed hierarchically-subbands with k≤5 layers utilize APCNN-based fusion for precise feature extraction, whereas subbands with k>5 layers adopt local energy maximization for computational efficiency [9]. The system incorporates a feedback mechanism for parameter self-optimization during fusion. Finally, the processed coefficients undergo inverse NSCT transformation with post-fusion quality enhancement to produce the output image.

This architecture's key innovation is its adaptive dual-path fusion strategy that automatically selects optimal processing methods based on subband characteristics, achieving superior balance between computational efficiency and fusion quality compared to conventional approaches [10]. The low-frequency path emphasizes energy preservation through phase analysis, while the high-frequency path combines neural network processing with traditional feature extraction for comprehensive detail reconstruction.

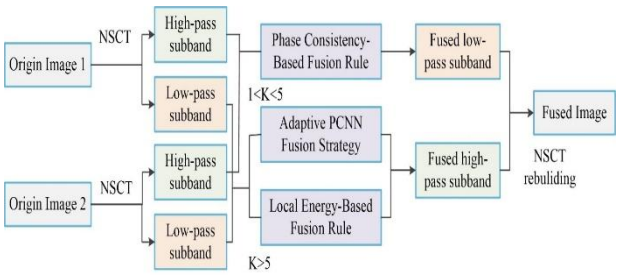

Figure 1: Overall idea of the NSCT-APCNN algorithm

2.2 NSCT-APCNN algorithm

2.2.1 Low-frequency subband fusion method

The NSCT transform is applied to the source images, and the resulting low-frequency components contain the energy information and spatial regions of the original images. For low-frequency subbands, phase congruency can be employed to characterize the global information of the image, extracting its feature information [12]. A phase congruency calculation model is calculated by Log-Gabor filte

$$PC(x)=\frac{\sum_0\sum_n W_0(x)\left[A_{n0}(x)\Delta\Phi_{n0}(x)-T_0\right]}{\sum_0\sum_n A_{n0}(x)+\varepsilon} \quad (1)$$

In the equation, $W_o(x)$ represents the weighted propagation quantity in the frequency domain, and T_0 denotes the noise threshold. $A_{n0}(x)$ and $\Delta\Phi_{n0}(x)$ are the amplitude and phase deviation measurement functions, respectively, along the filter direction and decomposition

scale. ε indicates a very small constant. Phase congruency is adopted as the coefficient selection criterion for the low-frequency components, with the fusion rule defined as follows:

$$PL_F(x,y)=\begin{cases} L_A(x,y) & PC_A(x,y)-PC_B(x,y)>T \\ L_B(x,y) & PC_A(x,y)-PC_B(x,y)<T \\ (L_A(x,y)+L_B(x,y))/2 & PC_A(x,y)-PC_B(x,y)=T \end{cases} \quad (2)$$

where, $L_A(x,y)$ and $L_B(x,y)$ represents the low-frequency coefficient of source image *A* and *B* at position (x,y). $PC_A(x,y)$ and $PC_B(x,y)$ indicate the phase congruency values of A and B respectively, where a higher value corresponds to more distinct pixel information in the image. In our implementation, a phase congruency threshold $T=0.035$ was used to determine whether the difference between $PC_A(x,y)$ and $PC_B(x,y)$ is significant. If $\left|PC_A(x,y)-PC_B(x,y)\right|>T$, the coefficient from the image with the higher phase congruency value is selected; otherwise, a weighted average is applied. This threshold was empirically determined based on validation set performance and is further adjusted adaptively using local statistical contrast within a 3×3 neighborhood window. The adaptive update rule scales T by a factor proportional to the standard deviation of local phase congruency values, allowing the fusion process to better handle varying image complexities and lighting conditions.

The fusion strategy for low-frequency coefficients implements an adaptive selection mechanism based on phase congruency comparison. Specifically, when the computed phase congruency difference between source images surpasses a predetermined threshold, the system preferentially selects coefficients from the image exhibiting stronger phase congruency characteristics. Conversely, if the phase congruency measurements fall within the threshold range, indicating comparable phase characteristics between images, the algorithm activates a weighted averaging operation. This dual-mode approach ensures optimal preservation of both luminance characteristics and spatial structural information from the source images, while maintaining the energy consistency of the fused output. The threshold parameter is dynamically adjusted according to image content characteristics through an adaptive optimization process.

$$L_{map}(x,y)=\begin{cases} L_{Amap} & PL_A(x,y)>PL_B(x,y) \\ L_{Bmap} & PL_A(x,y)\le PL_B(x,y) \end{cases} \quad (3)$$

where, $L_{map}(x,y)$ represents the fused low-frequency decomposition coefficient at position (x,y); L_{Amap} and L_{Bmap} denote the low-frequency decomposition coefficients of *A* and *B* respectively.

$$L_F(x,y)=(AE+BE)\cdot L_{map} \quad (4)$$

The high-frequency components obtained through NSCT decomposition contain multiple layers, and processing the high-frequency bands as a whole would be too coarse [13]. To achieve more accurate subband coefficient values, each subband is further divided into directional subbands for coefficient fusion and extraction. Since the high-frequency subbands obtained after decomposition cannot be negative, the default range for subband layers is set between 1 and 5.

Subbands with fewer than 5 layers contain rich structural edge information, which requires precise pixel-level relationships with structural contrast. To address this, an APCNN is introduced. Using local energy values as input, the model calculates a structural tensor operator to stimulate neuron firing. The resulting PCNN output map enables accurate extraction of edge contours and detailed image features.

The APCNN used in our framework is based on a modified Pulse-Coupled Neural Network architecture with adaptive threshold learning. Specifically, it consists of a single convolutional layer with a 3×3 kernel size applied to the high-frequency subbands to calculate local energy maps. The linking strength β and decay constant α are dynamically adjusted based on the image's standard deviation and gradient magnitude. The pulse firing threshold θ is initialized to 0.2 and decays exponentially. A sigmoid activation function is used to simulate neuron firing behavior, improving convergence smoothness and robustness. The network iterates for T = 20 steps per subband to produce the final pulse map. This configuration allows the network to adaptively enhance meaningful structures and suppress noise across varying subband characteristics. A fusion strategy based on regional energy features is proposed. Firstly, the local energy calculation model (Equation 5) was used to quantitatively analyze the energy distribution characteristics of each subband region, and the energy characteristic map was established. On this basis, a selection mechanism based on energy comparison is designed (Formula 6), which realizes the adaptive selection of the optimal coefficient by comparing and analyzing the local energy feature differences between the source images.

$$LE(x,y)=\sum_{a=-m}^{m}\sum_{b=-n}^{n}\left[I(x+a,x+b)\right]^2 \quad (5)$$

$$H_{k>5,Fmap}(x,y)=\begin{cases} H_{k>5,Amap}(x,y) & LE_A(x,y)>LE_B(x,y) \\ H_{k>5,Bmap}(x,y) & LE_A(x,y)\le LE_B(x,y) \end{cases} \quad (6)$$

Here, $H_{k>5,Fmap}(x,y)$ represents the high frequency subband coefficient after $k>5$ fusion, $H_{k>5,Amap}(x,y)$ and $H_{k>5,Bmap}(x,y)$ represent the high frequency component of *A* and source *B* at (x,y) coordinates, respectively. Finally, the optimized high-frequency subband fusion coefficient (Equation 7) is obtained through the selection mechanism. This method ensures the reasonable transfer of energy features while maintaining edge sharpness, and effectively improves the structural integrity and detail expression of the fused image.

$$H_{k>5,F}(x,y)=(A+B)\cdot H_{k>5,Fmap}(x,y) \quad (7)$$

In the field of two - dimensional image processing and analysis, it often involves the extraction and description of image structure information. Defining the smooth tensor as the structure tensor is a highly effective approach. The structure tensor can accurately characterize the local structure features within the neighborhood of pixel points in an image at the mathematical level, providing crucial theoretical support for subsequent image

analysis, feature extraction, edge detection, and other operations. Its specific expression is as follows:

$$G_\sigma = g_\sigma \otimes G \tag{8}$$

where, g_σ is a Gaussian function; σ is the variance. In our proposed method, the structure tensor is computed on each high-frequency subband to capture local directional gradients and edge orientation information. These tensor values are used as one of the input channels to the APCNN, alongside the local energy map. This allows the network to simultaneously consider both intensity-based (energy) and geometry-based (gradient structure) cues when selecting or enhancing subband coefficients. The APCNN processes these dual inputs through its neuron activation function to generate a refined feature map that better preserves structural detail during fusion.

In the process of processing multimodal images of power equipment, after the high-frequency sub-band is decomposed by k×5, a series of sub - bands with different characteristics are obtained. Among them, those sub - bands containing edge and contour information play a crucial role in accurately identifying key information such as the shape of the equipment, the boundaries of components, and potential fault areas.

To extract the effective information in these sub - bands more precisely, we adopt a method of inputting the structure tensor and local energy adaptive parameters into the APCNN. The structure tensor can effectively capture the structural features of local regions in the image. It reflects the direction and degree of change within the neighborhood of pixel points through the statistical analysis of image gradient information, providing basic information about the image structure for the APCNN. The local energy adaptive parameters, on the other hand, focus on describing the energy distribution in local regions of the image. They adjust the parameters dynamically according to the gray - scale changes and distributions of pixels within the sub - band, enabling a more sensitive capture of energy changes at the edges and contours. When these two are provided as inputs to the APCNN, the network will perform complex calculations and processing based on its own neuron connections and pulse - transfer mechanisms, and finally output a mapping diagram.

$M_A(x,y)$ and $M_B(x,y)$ are the mapping maps of A. When k<5, the HF coefficients are caculated as follows:

$$H_{k<5,F(x,y)} = \begin{cases} H_{k\le 5,A(x,y)} & M_A(x,y) \ge M_B(x,y) \\ H_{k<5,B(x,y)} & M_A(x,y) < M_B(x,y) \end{cases} \tag{9}$$

Finally, perform the inverse NSCT at the point $\{L_F(x,y), H_F(x,y)\}$, calculate the fusion coefficients of the obtained sub-bands, and then obtain the final fused image. It is important to note that fusion is performed individually per image pair, regardless of the equipment type. The model is not trained separately for transformers, current transformers, or disconnectors. Instead, a unified model processes all categories, and the classification phase that follows the fusion is responsible for distinguishing between equipment types.

3 Experiment and analysis

3.1 Dataset

In the experiment, a multi-modal image recognition test was carried out on three commonly used electrical equipment in the substation, namely transformers, current transformers, and disconnectors, which were collected by the inspection robot. The images were pre-aligned using feature-based geometric transformation to ensure spatial consistency across modalities before fusion. Illumination and weather conditions varied across capture sessions to simulate realistic inspection environments, including daytime and dusk lighting, as well as overcast and clear weather scenarios. The robot used for data acquisition was the StateGrid IRR-02 platform, equipped with a FLIR Lepton thermal sensor and a 5MP CMOS visible light camera. To improve model generalization, data augmentation was applied during training, including random horizontal flipping, rotation (±15°), brightness jitter (±20%), and Gaussian noise injection. The dataset was randomly divided into a training set (80%) and a test set (20%). To ensure result robustness, the training and evaluation processes were repeated 35 times using different random splits, effectively forming a repeated random sub-sampling validation protocol (not a strict 35-fold CV). A fixed random seed (seed = 42) was used to maintain reproducibility.

The NSCT-APCNN image classification model was trained using the Adam optimizer with an initial learning rate of 0.001, a batch size of 32, and a total of 50 epochs. No learning rate decay schedule was applied. All experiments were conducted on a workstation with an NVIDIA RTX 3060 GPU (12 GB VRAM), Intel i7-11700 CPU, and 32 GB RAM. These hardware specifications are provided to contextualize claims of real-time performance (average classification time of 2.36 seconds per image). The inverse NSCT reconstruction process used full-bandwidth reconstruction with no coefficient truncation or lossy compression, ensuring maximum fidelity in the reconstructed fused image. The contourlet decomposition was performed up to five levels with directional subbands set according to scale-dependent directional filters. The APCNN used in our method is a custom implementation built from scratch. It does not rely on pretrained weights or external transfer learning. Instead, the neuron firing thresholds, linking strength, and decay constants are adaptively initialized based on the image content. This allows the network to dynamically respond to input energy and structural characteristics without external supervision.

3.2 Comparison of recognition accuracy

The multi-modal image recognition accuracies of power equipment for six algorithms are compared and analyzed. These six algorithms are NSCT-APCNN, APCNN, PCNN [14], LeNet [15], AlexNet [16], and Support Vector Machine (SVM) [17]. The recognition accuracy tests are simultaneously conducted on the same image dataset.

The multi-modal image recognition accuracy of transformers is shown in Figure 2. In 35 test experiments, the highest multi-modal image recognition accuracy of NSCT-APCNN is 99.39%, the lowest is 95.03%, and the

average is 96.30%, which is higher than that of the other five algorithms for the multi-modal image recognition of transformers.

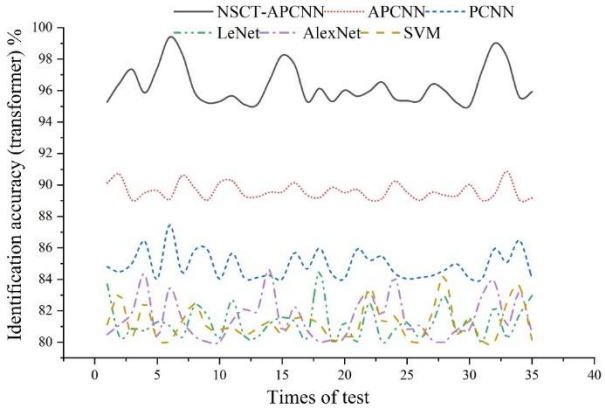

Figure 2: Multi-modal image recognition accuracy of transformers

The multi-modal image recognition accuracy of current transformers is shown in Figure 3. In 35 test experiments, the highest multi-modal image recognition accuracy of NSCT-APCNN is 99.57%, the lowest is 95.06%, and the average is 96.54%, which is higher than that of the other five algorithms for the multi-modal image recognition of current transformers.

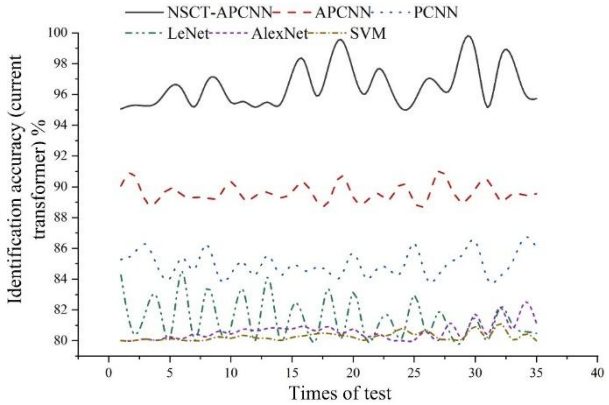

Figure 3 Multi-modal image recognition accuracy of current transformers

The multi-modal image recognition accuracy of disconnectors is shown in Figure 4. In 35 test experiments, the highest multi-modal image recognition accuracy of NSCT-APCNN is 98.74%, the lowest is 95.07%, and the average is 96.29%, which is higher than that of the other five algorithms for the multi-modal image recognition of disconnectors.

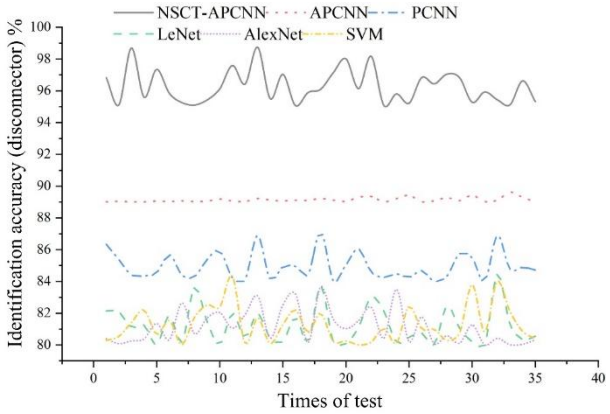

Figure 4: Multi-modal image recognition accuracy of disconnectors

3.3 Comparison of F1-scores

The F1-scores of multi-modal image recognition for transformers are shown in Figure 5. In 35 test experiments, the highest F1-score of NSCT-APCNN is 0.99, the lowest is 0.95, and the average is 0.96, which is higher than that of the other five algorithms for the multi-modal image recognition of transformers.

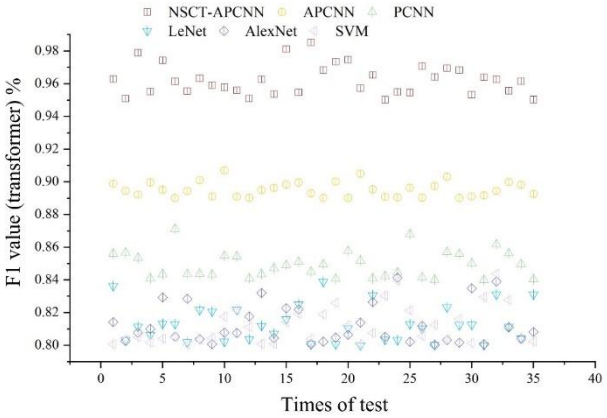

Figure 5: F1-scores of multi-modal image recognition for transformers

The F1-scores of multi-modal image recognition for current transformers are shown in Figure 6. In 35 test experiments, the highest F1-score of NSCT-APCNN is 0.996, the lowest is 0.95, and the average is 0.96. It doesn't differ much from the F1-scores of NSCT-APCNN for multi-modal image recognition of transformers, but is higher than those of the other five algorithms.

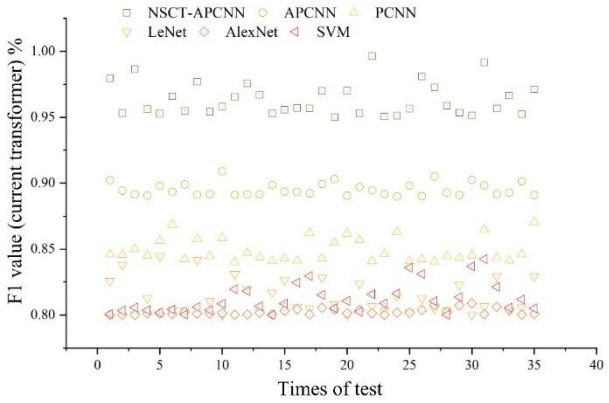

Figure 6: F1-scores of multi-modal image recognition for current transformers

The F1-scores of multi-modal image recognition for disconnectors are shown in Figure 7. In 35 test experiments, the highest F1-score of NSCT-APCNN is 0.98, the lowest is 0.95, and the average is 0.96. It can be seen that this F1-score is almost the same as those of NSCT-APCNN for multi-modal image recognition of transformers and current transformers, but still higher than the F1-scores of the other five algorithms for multi-modal image recognition.

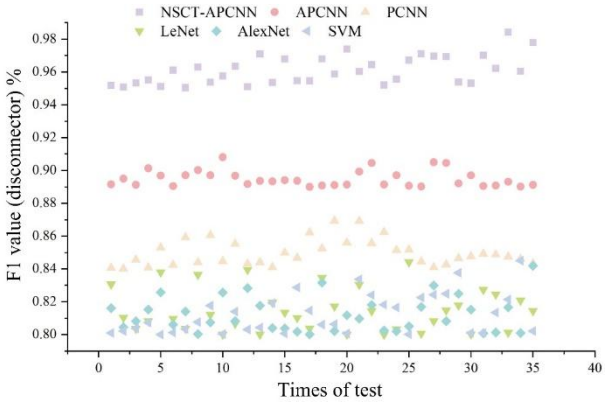

Figure 7: F1-scores of multi-modal image recognition for disconnectors

3.4 Comparison of classification time

The classification time tests were simultaneously conducted on the same image dataset using NSCT-APCNN, APCNN, PCNN, LeNet, AlexNet, and SVM (e.g. Figure 8). In 35 test experiments, the maximum classification time required by NSCT-APCNN was 3.17 s, the minimum was 2.00 s, and the average was 2.36 s, which is less than that of the other five algorithms for multi-modal image recognition of transformers.

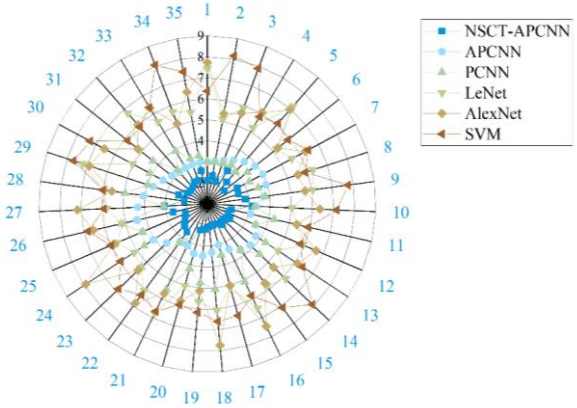

Figure 8: Multi-modal image recognition time for transformers

The multi-modal image recognition time for current transformers is shown in Figure 9. In 35 test experiments, the maximum classification time required by NSCT-APCNN was 3.32 s, the minimum was 2.01 s, and the average was 2.35 s, which is lower than the recognition time of the other five algorithms.

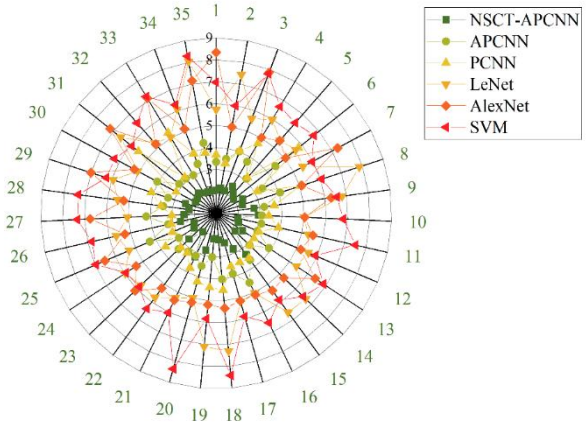

Figure 9: Multi-modal image recognition time for current transformers

The multi-modal image recognition time for disconnectors is shown in Figure 10. In 35 test experiments, the maximum classification time required by NSCT-APCNN was 3.28 s, the minimum was 2.00 s, and the average was 2.32 s. To evaluate whether this processing time is suitable for real-time deployment, we refer to the operational constraints of common inspection robots such as the StateGrid IRR-02 platform. These robots typically operate at a movement speed of 0.5–1.0 m/s and capture images at a rate of 0.2–0.5 Hz (i.e., 1 image every 2–5 seconds) to allow for stable capture and onboard analysis. Given this cycle time, an average classification duration of ~2.36 seconds per image falls within acceptable limits for real-time onboard processing.

Therefore, the proposed NSCT-APCNN model meets the practical deployment requirements for inspection robots in substation environments. This clearly gives NSCT-APCNN an advantage in terms of time among the six recognition algorithms.

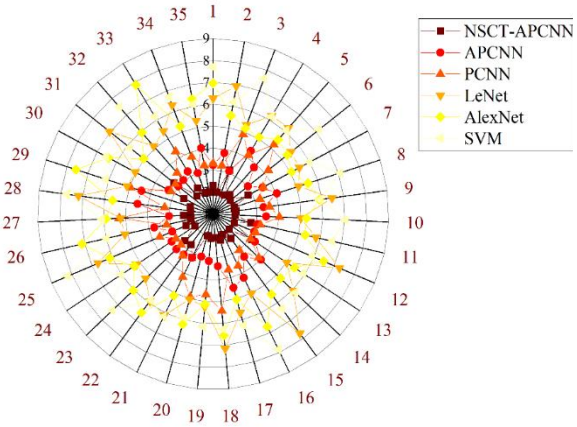

Figure 10: Multi-modal image recognition time for disconnectors

4 Discussion

To further clarify the performance and internal mechanisms of the proposed NSCT-APCNN algorithm, we present a comparative discussion with five widely used methods: APCNN, PCNN, LeNet, AlexNet, and SVM.

The performance advantage of the NSCT-APCNN stems from a hybrid architectural design that intelligently integrates signal decomposition with adaptive neural processing. NSCT Decomposition Enables Multiscale Feature Isolation. Unlike conventional CNN-based methods (e.g., AlexNet, LeNet), the NSCT first decomposes images into multiple frequency subbands, enabling separation of structural (low-frequency) and detailed (high-frequency) components. This facilitates precise spatial localization of important features such as edges and contours. The phase congruency method used for low-frequency component fusion enables enhanced contrast sensitivity and better alignment of key image regions. This outperforms simple averaging or maximum selection methods commonly used in PCNN/APCNN. High-frequency layers ($k \leq 5$), rich in contours and component boundaries, are fed into an APCNN enhanced with structure tensors and local energy maps. This enhanced discrimination between similar-looking classes (e.g., CTs and transformers). For deeper subbands ($k > 5$), where detail contributions diminish, the algorithm bypasses neural processing and uses energy-maximization rules. This hybrid strategy achieves a strong balance between accuracy and computational efficiency, resulting in the lowest average processing time (2.36s) among all methods tested.

SVM and PCNN fail to generalize across varying illumination and texture conditions due to lack of adaptive feature learning or hierarchical analysis. LeNet and AlexNet, while leveraging deep learning, treat all image features uniformly, leading to inefficiencies in fusing multimodal inputs with differing characteristics. APCNN, though adaptive, does not operate in the frequency domain, thus lacking the layered processing benefits critical for multimodal detail retention.

5 Conclusion

This study proposes an innovative NSCT-APCNN-based image fusion algorithm for power equipment. By employing a phase consistency strategy to process low-frequency components for energy feature preservation, and combining the APCNN model with a local energy maximization method for precise high-frequency detail extraction, the algorithm demonstrates it performance. Experimental results show outstanding effectiveness in identifying transformers, current transformers and other equipment, achieving a peak recognition accuracy of 99.57% with an average processing time of merely 2.36 seconds. Compared with conventional recognition methods, this approach overcomes the issue of excessive training time in traditional neural networks while maintaining accurate equipment identification. Benchmarking against APCNN, PCNN, LeNet, AlexNet and SVM confirms superior performance in multimodal image recognition of power equipment. Future research directions include developing hardware acceleration solutions to enhance real-time performance, exploring region-adaptive fusion to improve robustness, and extending applications to smart grid fault diagnosis. The proposed methodological framework also holds significant reference value for image processing in other industrial inspection scenarios.

Fundings

1.Science and Technology Research Project of Henan Province Department of Science and Technology：Research on Key Technologies of Graph based Representation Learning and Robustness （252102210069）

2. Key Research Project of Universities in Henan Province：Research on Efficient Item Set Data Mining Algorithm and Its Application（25B520024）

References

[1] Zhang P, Li T, Wang G, et al. multi-source information fusion based on rough set theory: A review[J]. Information Fusion, 2021, 68: 85-117. https://doi.org/10.1016/j.inffus.2020.11.004.

[2] Zhao X, Peng Z, Zhao S. Substation electric power equipment detection based on patrol robots[J]. Artificial Life and Robotics, 2020, 25(3): 482-487. https://doi.org/10.1007/s10015-020-00604-8.

[3] Liu L, Yang Y. A Study on the Application of New Feature Techniques for Multimedia Analysis in Artificial Neural Networks by Fusing Image Processing[J]. Informatica, 2024, 48(11): 113-124. https://doi.org/10.31449/inf.v48i11.5851.

[4] Choudhary G, Sethi D. From conventional approach to machine learning and deep learning approach: an experimental and comprehensive review of image fusion techniques[J]. Archives of Computational

Methods in Engineering, 2023, 30(2): 1267-1304. https://doi.org/10.1007/s11831-022-09833-5.

[5] Zhang H, Xu H, Tian X, et al. Image fusion meets deep learning: A survey and perspective[J]. Information Fusion, 2021, 76: 323-336. https://doi.org/10.1016/j.inffus.2021.06.008.

[6] Choudhary G, Sethi D. From conventional approach to machine learning and deep learning approach: an experimental and comprehensive review of image fusion techniques[J]. Archives of Computational Methods in Engineering, 2023, 30(2): 1267-1304. https://doi.org/10.1007/s11831-022-09833-5.

[7] Cheng T, Gu J, Zhang X, et al. Multimodal image registration for power equipment using clifford algebraic geometric invariance[J]. Energy Reports, 2022, 8: 1078-1086. https://doi.org/10.1016/j.egyr.2022.02.192.

[8] Wang Q, Zhang J, Du J, et al. A fine-tuned multimodal large model for power defect image-text question-answering[J]. Signal, Image and Video Processing, 2024, 18(12): 9191-9203. https://doi.org/10.1007/s11760-024-03539-w.

[9] Chun-Man Y A N, Bao-Long G U O, Meng Y I. Fast algorithm for nonsubsampled contourlet transform[J]. Acta Automatica Sinica, 2014, 40(4): 757-762. https://doi.org/10.1016/s1874-1029(14)60007-0.

[10] Tian J, Chen L, Ma L, et al. Multi-focus image fusion using a bilateral gradient-based sharpness criterion[J]. Optics communications, 2011, 284(1): 80-87. https://doi.org/10.1016/j.optcom.2010.08.085.

[11] Ibrahim S I, El-Tawel G S, Makhlouf M A. Brain image fusion using the parameter adaptive-pulse coupled neural network (PA-PCNN) and non-subsampled contourlet transform (NSCT)[J]. Multimedia Tools and Applications, 2024, 83(9): 27379-27409. https://doi.org/10.1007/s11042-023-16515-2.

[12] Zhang Q, Maldague X. An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing[J]. Infrared Physics & Technology, 2016, 74: 11-20. https://doi.org/10.1016/j.infrared.2015.11.003.

[13] Zhao C, Guo Y, Wang Y. A fast fusion scheme for infrared and visible light images in NSCT domain[J]. Infrared Physics & Technology, 2015, 72: 266-275. https://doi.org/10.1016/j.infrared.2015.07.026.

[14] Alaslni M G, Elrefaei L A. Transfer learning with convolutional neural networks for iris recognition[J]. Int. J. Artif. Intell. Appl, 2019, 10(5): 47-64. https://doi.org/10.5121/ijaia.2019.10505.

[15] Wei G, Li G, Zhao J, et al. Development of a LeNet-5 gas identification CNN structure for electronic noses[J]. Sensors, 2019, 19(1): 217. https://doi.org/10.3390/s19010217.

[16] Jasim W N, Nemer Z N, Harfash E J. Implementation of Multiple CNN Architectures to Classify the Sea Coral Images[J]. Informatica, 2023, 47(1): 43-50. https://doi.org/10.31449/inf.v47i1.4429.

[17] Li S, Kwok J T Y, Tsang I W H, et al. Fusing images with different focuses using support vector machines[J]. IEEE Transactions on neural networks, 2004, 15(6): 1555-1561. https://doi.org/10.1109/tnn.2004.837780.