Robust and Accurate Recognition of Carriage Linear Array Images for Train Fault Detection

Zhenzhou Fu and Xiao Pan *

Posted Date: 13 September 2024

doi: 10.20944/preprints202409.1075.v1

Keywords: High-resolution image recognition; point set registration; weighted radial basis function; non-linear scale distortion; gaussian mixture model

 doi:10.20944/preprints202409.1075.v1

Article

Robust and Accurate Recognition of Carriage Linear Array Images for Train Fault Detection

Zhenzhou Fu and Xiao Pan *

School of Instrumentation and Optoelectronic Engineering, Beihang University (BUAA), Beijing 100191, China
* Correspondence: panxiao0620541@buaa.edu.cn

Abstract: Train fault detection often relies on comparing collected images with reference images, making accurate image type recognition crucial. Current systems use Automatic Equipment Identification (AEI) devices to recognize carriage numbers while capturing images, but damaged Radio Frequency (RF) tags or blurred characters can hinder this process. Carriage linear array images, with their high resolution, extreme aspect ratios, and local nonlinear distortions, present challenges for recognition algorithms. This paper proposes a method tailored for recognizing such images. We apply an object detection algorithm to locate key components, simplifying image recognition into a sparse point set alignment task. To handle local distortions, we introduce a weighted radial basis function (RBF) and maximize the similarity between Gaussian mixtures of point sets to determine RBF weights. Experiments show 100% recognition accuracy under nonlinear distortions up to 15%. The algorithm also performs robustly with detection errors and identifies categories from 79 image classes in 24 milliseconds on an i7 CPU without GPU support. This method significantly reduces system costs and advances automatic exterior fault detection for trains.

Keywords: high-resolution image recognition; point set registration; weighted radial basis function; non-linear scale distortion; gaussian mixture model

1. Introduction

With the continuous expansion of railway networks and the increase in train speeds, the importance of train body fault detection technology has become increasingly prominent. Failure to accurately identify train faults in a timely manner poses serious safety risks and can increase operating costs. Therefore, researchers have designed systems such as the Trouble of Moving EMU Detection System (TEDS), Trouble of Moving Freight Detection System (TFDS), and Trouble of Moving Vehicle Detection System (TVDS) to monitor train fault conditions [1–4]. These systems use multiple cameras installed around the tracks to capture images of the train car bodies from different angles. Figure 1(a) shows a train carriage line-scan image acquisition system. When the train passes through the camera's field of view, the cameras capture images of the carriage surface from various observation angles. By extracting standard reference images of the same category from a pre-established reference image library and performing comparative analysis, these systems can evaluate the train's fault status. Figure 1(e) illustrates the fault detection process; for more details, please refer to our previous work [3]. The first step in this process is to accurately select the reference images for comparison, which is crucial for the accuracy of subsequent fault analysis. Ensuring the correct selection of reference images depends on accurately identifying the category of the current image. Thus, accurate identification of carriage image categories is essential in the train fault detection system.

 doi:10.20944/preprints202409.1075.v1

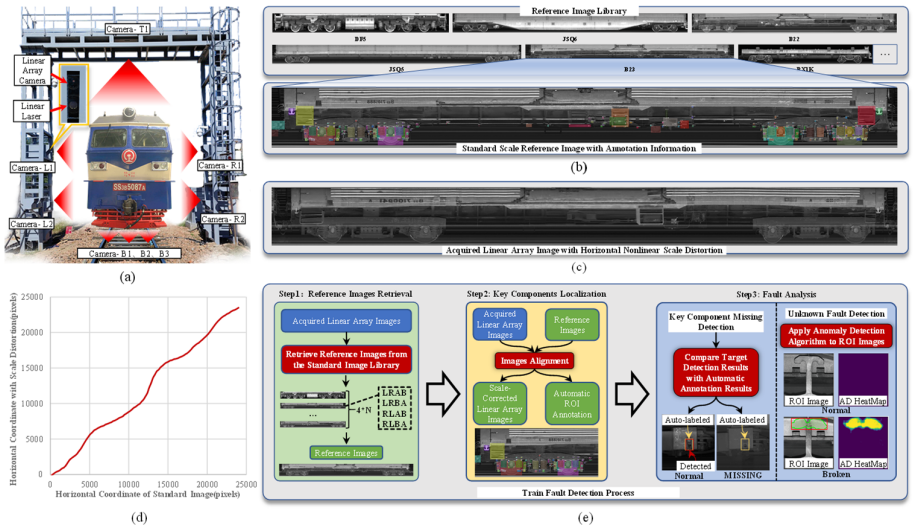


Figure 1. Train Fault Detection System. (a) Train line-scan image acquisition system; (b) The reference image library contains standard images of all carriage types without scale distortion, each annotated with key component information for subsequent fault analysis; (c) Acquired line-scan images may have nonlinear scale distortions in the horizontal direction; (d) Correspondence between horizontal pixel coordinates in (c) and standard reference image coordinates; (e) Main processing steps in train fault detection.

Currently, in train image acquisition systems, image category identification is primarily achieved by recognizing the carriage number information using specialized equipment while capturing the line-scan images of the carriage. This information includes the model and a unique identifier of the carriage, from which the model information can be parsed to label the category of the current image. Existing carriage number recognition equipment can be classified into two types based on their working principles. One type uses microwave communication technology. When the train passes between microwave antennas installed on the tracks, the equipment reads the RFID tags installed at the bottom of the carriages, which store the carriage number information, as shown in Figure 2(a). The other type uses visual character recognition technology to directly recognize the carriage number characters painted on the side of the carriage, as shown in Figure 2(b).

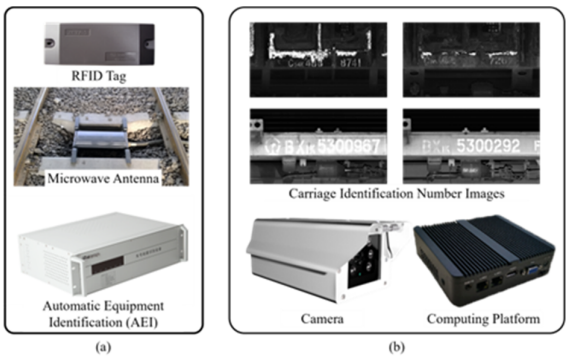


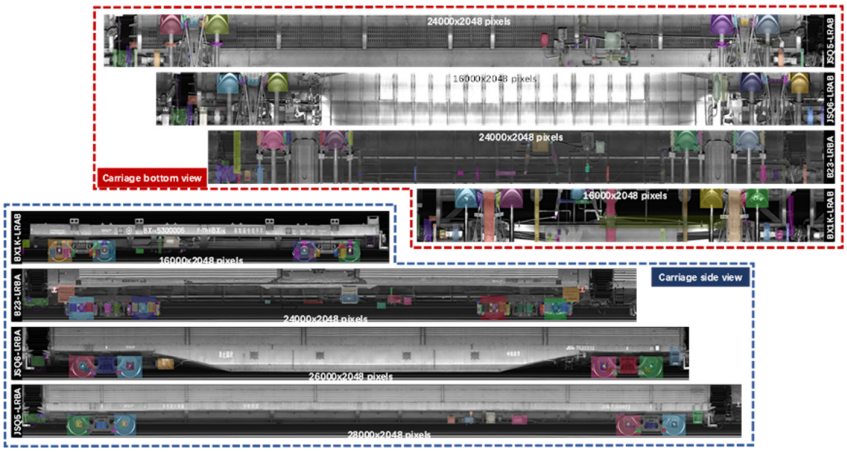
Figure 2. Carriage Identification Equipment. (a) Carriage identification equipment based on microwave communication; (b) Carriage identification equipment based on visual character recognition.

However, the real-world operating conditions of trains are complex, especially for freight trains, where strong vibrations and unpredictable environmental factors can cause the pre-installed RFID tags to be lost or damaged. In such cases, the microwave communication-based carriage number recognition equipment cannot mark the collected image categories, rendering the fault detection

 doi:10.20944/preprints202409.1075.v1

system inoperative. Additionally, the carriage number markings can be worn out or obscured by dust, leading to blurred characters in the images, as shown in Figure 2(b). This compromises the accuracy of character recognition-based carriage number identification. Furthermore, the high cost of AEI equipment significantly increases the cost of the fault detection system. To address these issues, this study focuses on developing a method for identifying train carriage categories directly from line-scan images, without relying on additional specialized equipment. This approach aims to overcome the limitations of existing carriage image identification technologies and improve the stability of train fault detection systems.

Due to the ease of achieving uniform illumination with line-scan cameras, the high consistency of multiple imaging of the same target, and their unique imaging principle which allows capturing high-resolution images of entire carriages without the need for image stitching algorithms, line-scan cameras are widely used in existing train fault detection systems [1–4]. However, line-scan carriage images have unique characteristics such as nonlinear scale distortion, ultra-high image resolution, and extreme aspect ratios, which pose new challenges to image recognition tasks. Firstly, when a train passes a line-scan camera at varying speeds, if the train's speed is not accurately and promptly fed back to the camera's trigger interface, the captured line-scan images of the carriage will exhibit nonlinear scale distortion in the horizontal direction, as shown in Figure 1(c). The relationship between the horizontal coordinates of the distorted image and the standard scale image is illustrated in Figure 1(d). This curve intuitively demonstrates the form of nonlinear distortion in the horizontal direction. This type of scale distortion maintains consistency in the vertical direction but presents unknown nonlinearity in the horizontal direction. It not only alters local image features but also significantly impacts global image features, directly affecting the accuracy of image recognition algorithms. Furthermore, to provide sufficient image detail features for accurate fault identification, carriage line-scan images often have extremely high resolutions. Due to the elongated geometric structure of the carriages, there is a significant difference in the aspect ratio of the line-scan images. Figure 3 shows line-scan images of four different types of carriages, captured by a 2K line-scan camera from both bottom and side perspectives. The resolution in the width direction primarily depends on the length of the carriage. As shown in Figure 3, even the smallest image resolution reaches approximately 32 megapixels, with an aspect ratio of about 1:8. Although existing deep learning-based image recognition algorithms have demonstrated excellent performance across multiple datasets, these models are primarily designed for low-resolution and regular aspect ratio images. Processing high-resolution images for feature extraction and handling consumes substantial computer memory or GPU memory, making it challenging to meet the above requirements due to capacity limitations. If high-resolution images with extreme aspect ratios are forcibly scaled down to conventional proportions and regular resolutions, significant image detail features will be lost, and the extreme scale distortion will cause uncertainty in feature representation, which is crucial for distinguishing different carriage types.



 doi:10.20944/preprints202409.1075.v1

Figure 3. Line-scan images of different types of carriages and their corresponding key component detection results.

To address the challenge of identifying the categories of high-resolution, extreme aspect ratio line-scan images of carriages, this study utilizes the spatial arrangement of key components in the carriage images, which corresponds uniquely to each carriage type. We propose a template matching-based image identification method, with the main contributions summarized as follows:

1. To avoid feature distortion and high memory usage when processing image features during category identification. We leverage the relationship between the layout of key components in the carriage image and the carriage type. By constructing a sparse point set based on the detected key components, we propose a template matching method that registers the sparse point set of the acquired image to the standard template point set of the target category, thereby identifying the carriage image category.

2. To ensure that the coordinate transformation in the template matching process can accommodate the nonlinear scale distortions present in line-scan images of carriages, this study employs a weighted radial basis function to describe the nonlinear transformation relationship between the horizontal coordinates of the two point sets. Furthermore, to address the issues of unknown point correspondences and mismatched point quantities during point set registration, we designed an objective function that maximizes the similarity of the mixed Gaussian distribution between the point sets, thereby solving for the weights of the radial basis function.

3. Extensive experiments demonstrate that our method exhibits excellent performance in terms of recognition accuracy, processing speed, and robustness. In the task of recognizing high-resolution line-scan images of 76 carriage categories, the algorithm achieves 100% recognition accuracy when the local nonlinear scale distortion is less than 15%. Furthermore, it can accurately identify image categories even when the number of false detections increases or decreases by up to 10%. The entire image identification process takes an average of only 24ms on an i7 CPU.

The rest of this paper is organized as follows. Section 2 provides a review of related work. Section 3 introduces the proposed methodology framework in detail. Section 4 discusses experimental results. Finally, Section 5 concludes this article.

2. Related Works

Image classification is a typical task in the field of computer vision, widely applied in object recognition, medical image analysis, and autonomous driving. Based on different image feature descriptions and classification principles, image classification methods are mainly divided into traditional machine learning-based methods [5–9] and deep learning-based methods [11–17]. This article provides a detailed introduction to these methods, focusing on their implementation principles, feature representation methods, and the advantages of each approach.

2.1. Traditional Machine Learning-Based Image Classification Methods

Support Vector Machines (SVM) are widely used in image classification by finding an optimal hyperplane to separate classes. Chandra and Bedi [5] reviewed SVM's principles, kernel function selection, and practical performance. Typically, handcrafted features like SIFT and HOG are used for feature extraction, transforming image information into vectors for classification. SVM is effective in high-dimensional spaces but has long training times and is sensitive to parameter selection, making it less suitable for large-scale datasets. Random Forests, an ensemble method, enhance robustness and generalization through multiple decision trees trained on subsamples and feature subsets. Bosch et al. [6] used Random Forests and Ferns, a variant of decision trees known for computational efficiency, for image classification. While Random Forests are fast and handle high-dimensional data, they can overfit with a small number of trees, and their interpretability is limited. k-Nearest Neighbors (kNN) determines the class of a sample by calculating its distance to k nearest neighbors. Amato and Falchi [7] improved kNN's accuracy using local feature descriptors like SURF and ORB. kNN is simple and adaptive but computationally expensive for large datasets, and performance

 doi:10.20944/preprints202409.1075.v1

depends on the distance metric and k value. The Naive Bayes classifier applies Bayes' theorem to calculate the conditional probability of features. Timofte et al. [8] used statistical principles in their Naive Bayes-based image classification method. While fast and suitable for high-dimensional data, Naive Bayes assumes feature independence, which is often unrealistic. The Bag of Words (BoW) model represents images as unordered collections of feature words. Wang and Huang [9] improved BoW for image classification by refining visual vocabulary generation. BoW is flexible and scalable but lacks spatial information, potentially missing structural details in images.

2.2. Deep Learning-Based Image Classification Methods

EfficientNet is a convolutional neural network that balances depth, width, and resolution through compound scaling, reducing parameters while maintaining performance [10]. Tan and Le [11] improved the architecture with EfficientNetV2, achieving smaller models and faster training. EfficientNet's hierarchical design efficiently extracts high-level features for image classification. ResMLP uses multilayer perceptrons (MLP) and residual connections for image classification, offering advantages in data efficiency and generalization [12]. ResNet, introduced by He et al. [13], solved the gradient vanishing issue with residual connections, improving performance in feature extraction. Liang [14] demonstrated ResNet's superior performance on various benchmarks. Cheng et al. [15] proposed SeNet, enhancing classification by extracting edge information through structured edge detection. ShuffleNet, proposed by Zhang et al. [16], uses depthwise separable convolutions and channel shuffle to optimize performance on mobile devices while reducing complexity. The Vision Transformer (ViT) [17] extends Transformer models from natural language processing to image recognition by treating images as sequences of patches. ViT, pre-trained on large datasets like ImageNet-21k and JFT-300M, has demonstrated superior performance over traditional CNNs on large-scale image tasks.

2.3. High-Resolution Image Detection Related Research

There has been extensive research on high-resolution image detection and recognition in the field of satellite remote sensing. For example, in [18], buildings in remote sensing images captured by drones were successfully identified. In [19,20], ships in Gaofen-3 SAR images were detected and identified. In [21], sparse representation was combined with various characteristics of remote sensing images for the first time, improving the accuracy of high-resolution image target recognition. Representatively, [22] proposed the HRDNet network model designed for small target detection in high-resolution images. To improve detection speed, [23] proposed a multi-GPU distributed computing method, optimizing performance for target detection tasks on 4K and 8K images. However, current research on high-resolution image detection primarily focuses on detecting conventional targets such as faces, pedestrians, buildings, ships, and aircraft [24]. These studies mainly focus on small target detection and typically deal with images of regular size ratios. Unlike target detection tasks, this study focuses on image recognition. The carriage line-scan images in this study are characterized by ultra-high resolution and extreme aspect ratios, with each image corresponding to an entire carriage. Thus, existing research results do not address the image recognition issues in this study.

2.4. Image Retrieval Methods

As shown in Figure 1, searching for target images from a reference library for train fault detection is a typical image retrieval task. Various mature methods are available:

Liu et al. [25] proposed Deep Supervised Hashing (DSH), using CNNs to learn binary codes for efficient image retrieval with low storage and computational costs, though it struggles with high-resolution images and requires extensive labeled data. Gordo et al. [26] introduced an end-to-end learning architecture with R-MAC descriptors and a Siamese network, improving retrieval accuracy but requiring complex training and high-quality labeled datasets. Liu et al. [27] proposed Guided Similarity Separation (GSS) using graph convolutional networks and unsupervised learning to

 doi:10.20944/preprints202409.1075.v1

enhance retrieval accuracy, but the method has high computational complexity. Ramzi et al. [28] introduced a rank loss optimization framework, improving retrieval performance across multiple datasets, although the optimization process is resource-intensive. Smeulders et al. [29] proposed Content-Based Image Retrieval (CBIR), which relies on extracting visual features like color and texture, but the feature extraction process can be slow for large databases. Noh et al. [30] developed DELF, a deep local feature descriptor effective in handling complex images, though initial training requires significant computational resources. Henkel et al. [31] proposed an end-to-end pipeline combining EfficientNet with hybrid Swin-Transformers, achieving state-of-the-art results in large-scale landmark retrieval, though the model's complexity demands substantial computational power and labeled data.

In summary, numerous advanced technologies and application methods have been proposed in the field of image retrieval, demonstrating significant performance improvements in specific application scenarios. In train fault detection systems, the extremely high resolution and extreme aspect ratio of carriage line-scan images present significant challenges in memory and GPU usage for existing image retrieval methods. While traditional machine learning methods perform well on small-scale datasets, they often fall short on large-scale image data compared to deep learning algorithms. Despite deep learning methods showcasing outstanding performance in image recognition, these models are primarily designed for low-resolution and regular aspect ratio images. Forcibly resizing high-resolution carriage line-scan images to regular sizes would result in the loss of critical detailed features essential for vehicle type recognition.

Furthermore, accumulating images of different carriage types requires considerable time, and in training-based methods, algorithm accuracy is positively correlated with the number of images collected. When data categories change, the model needs retraining, which is time-consuming and cannot quickly adjust to new target categories. Due to the unique nature of line-scan carriage images, existing methods do not consider these aspects and fail to meet the specific needs of image recognition in train fault detection.

To address these challenges, this study explores methods for acquiring reference images from the perspective of image classification, aiming to develop a method suitable for quickly obtaining reference images of line-scan carriage images. This approach seeks to reduce dependence on computer hardware configuration and computational resources, improving the efficiency and accuracy of line-scan carriage image recognition.

3. Methodology

To tackle the recognition challenges of carriage line-array images, we propose a template matching-based image identification method. By exploiting the distinct correlation between the spatial arrangement of key carriage components and their categories, the high-resolution image recognition challenge was reframed as a sparse point set template matching task. Considering the non-linear scale distortions in the horizontal dimension of these images, a weighted radial basis function was utilized for coordinate transformations during point set registration. This approach effectively addresses the template matching challenges, enhancing both accuracy and robustness. Before detailing the algorithm, it's essential to categorize the unique features of line-array carriage images.

3.1. Classification of Carriage Line-Array Images

During the acquisition of line-scan images of train carriages, the camera remains stationary, and the movement of the train results in four different categories of line-scan integrated images for the same carriage. As shown in Figure 4(a), L and R represent the left and right sides of the carriage, respectively, while A and B represent the front and rear ends. The red line indicates the scanning line of the line-scan camera. The corresponding image categories are denoted as LRAB, LRBA, RLAB, and RLBA. If there are N types of carriages, the total number of line-scan image categories is 4*N. Images of the same side of the carriage exhibit a left-right flip relationship.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2024 doi:10.20944/preprints202409.1075.v1

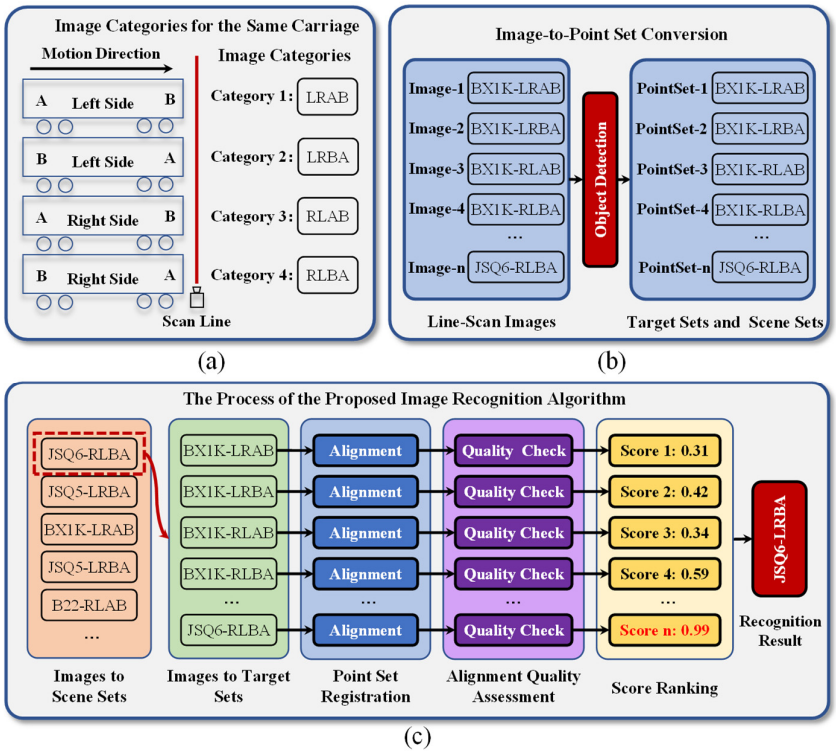


Figure 4. Framework of the Image Recognition Algorithm.

3.2. Template Matching-Based Image Recognition Method

This paper proposes a template matching-based image recognition algorithm, consisting of four main steps: image-to-point conversion, point set registration, matching quality evaluation, and score ranking. As shown in Figure 4(c), the process begins by converting the target detection results of the line-scan carriage images into point sets. In Figure 3, different colored rectangles represent different detection targets. Based on predefined criteria, only the center coordinates of the bounding boxes are retained, and no labels are assigned to the points, forming a two-dimensional point set. This step corresponds to Figure 4(b). The same procedure is applied to both the image to be recognized and the template image, yielding a scene point set and a target point set, respectively. The image to be recognized may exhibit local nonlinear scale distortions, while the template image is a standard scale image for each category.

The algorithm then performs point set registration for the specified scene point set and the template point sets of each category, aiming to align the horizontal coordinates of the point sets. Subsequently, each registration result undergoes a quality evaluation to verify the matching status between point sets and determine the success of the match, scoring each accordingly. Finally, all matching scores are ranked, and the template category with the highest score exceeding a set threshold is selected as the recognition result.

The key to the recognition process is accurately finding the target point set that matches the scene point set. The core challenge is addressing the point set registration problem under local nonlinear scale distortions in the horizontal direction of the line-scan carriage images. Due to the unique nature of line-scan carriage images, the initial vertical coordinates of correctly paired target and scene point sets are generally aligned. To improve point set registration efficiency, only horizontal alignment needs to be addressed, using an appropriate matching quality evaluation method to ensure accurate carriage category recognition. To achieve this, a coordinate transformation equation based on a weighted radial basis function is proposed, along with a numerical solution. Additionally, a specific point set matching quality evaluation method is provided. The following sections will detail these core components.

 doi:10.20944/preprints202409.1075.v1

3.3. Weighed Radial Basis Function for Coordinate Transformation

Radial Basis Function (RBF) neural networks have garnered widespread attention for their excellent capability in approximating nonlinear functions, theoretically able to approximate any nonlinear function [32]. The core of RBF neural networks lies in the linear combination of radial basis functions within the network layer, making them a powerful tool for handling nonlinear problems. Leveraging this characteristic of RBF neural networks, this paper proposes a weighted radial basis function to describe and handle the coordinate transformation problem under local nonlinear scale distortions in line-scan images of train carriages. The horizontal coordinate transformation equation from the scene point set to the target point set to is defined as shown in Equation (1).

$$T(x) = x + \sum_{i=1}^{n} w_i \cdot RBF(x, c_i, \delta) \tag{1}$$

In the equation above, w_i denotes the weighting value of each radial basis function, a parameter to be determined. These weights determine the influence of their corresponding radial basis functions on the overall coordinate transformation. $RBF(\cdot)$ represents a kernel function, with the Gaussian radial basis function selected, as detailed in Equation (2). Here, x designates the horizontal coordinate of a point in the scene point set, c is the center of the radial basis function which corresponds to the horizontal coordinate of a point in the target point set, and δ denotes the standard deviation of the radial basis function. The value of s can either be predefined based on the problem's specifics or adjusted for optimal results.

$$RBF(x, c, s) = \mathrm{e}^{-\frac{(x-c)^2}{2\delta^2}} \tag{2}$$

By employing the weighted radial basis function concept, the strengths of RBF neural networks in approximating non-linear functions are harnessed, providing a robust mathematical model to address the non-linear distortions in carriage line-array images.

3.4. Objective Function Design

When both the scene and target point sets possess an equal number of points with known correspondence, the weight vector $\boldsymbol{w}$ can be directly determined using the least squares method. However, the correspondence between point sets, based solely on object detection results, is often uncertain. Additionally, object detection imperfections, such as false positives or negatives, can cause point count mismatches between the scene and target sets. This makes an analytical determination of the weight variables challenging. To address this, we employ numerical optimization techniques. Iterative optimization, guided by the objective function, yields the weight variables.

The objective function's design aims to measure the alignment between two point sets, particularly when their sizes differ and correspondence is unclear. The Gaussian Mixture Model (GMM) is introduced to represent these point sets. Alignment is gauged by evaluating the congruence between the GMMs of both sets. The GMM's probability density at location x is defined as $gmm(x, U) = \sum_{i=1}^{k} \omega_i \phi(x|u_i, \Sigma_i)$, where u_i represents a point in set U, the gaussian distribution centered at $u_i$is expressed as:

$$\phi(x|u_i, \Sigma_i) = \frac{\exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]}{\sqrt{(2\pi)^d |\det(\Sigma_i)|}} \tag{3}$$

The L2 distance is chosen to determine the similarity between the GMMs of the two sets. Let $\boldsymbol{M}$ represent the target point set, corresponding to centers of components in a standard carriage image, and x a point within $\boldsymbol{M}$. $\boldsymbol{S}$ denotes the scene point set, representing centers of bounding boxes in the image being identified. $\boldsymbol{w}$ is the vector of transformation parameters from $\boldsymbol{S}$ to $\boldsymbol{M}$, comprising weights. The goal is to identify the optimal weight vector $\boldsymbol{w}$ that minimizes the function $F(\boldsymbol{S}, \boldsymbol{M}, \boldsymbol{w})$:

 doi:10.20944/preprints202409.1075.v1

$$\begin{aligned} F(\boldsymbol{S},\boldsymbol{M},\boldsymbol{w}) &= \min_{\boldsymbol{w}} \left(\int G\,dx\right) \\ G &= \Big(gmm\big(x,T(\boldsymbol{S},\boldsymbol{M},\boldsymbol{w})\big) - gmm(x,\boldsymbol{M})\Big)^2 \end{aligned} \tag{4}$$

The squared term in the objective function can be decomposed into three components as shown in Equation (5). Since the third term, $gmm(x,\boldsymbol{M})^2$, is constant during optimization, focus is placed on the remaining two parts.

$$\begin{gathered} \Big(gmm\big(x,T(\boldsymbol{S},\boldsymbol{M},\boldsymbol{w})\big) - gmm(x,\boldsymbol{M})\Big)^2 = A - B + C \\ A = gmm\big(x,T(\boldsymbol{S},\boldsymbol{M},\boldsymbol{w})\big) \\ B = 2gmm\big(x,T(\boldsymbol{S},\boldsymbol{M},\boldsymbol{w})\big)gmm(x,\boldsymbol{M}) \\ C = gmm(x,\boldsymbol{M})^2 \end{gathered} \tag{5}$$

Considering that the scale distortion in line-array images mainly manifests in the scanning direction, only 1D GMM similarity calculations based on horizontal coordinates are considered. If optimal transformation parameters are identified within a specific horizontal range, any point sampled within this range ensures a minimized objective value. Thus, uniform sampling within the model point set's horizontal range is adopted for quick GMM similarity estimation.

For optimization, the Levenberg-Marquardt (L-M) algorithm is utilized, a nonlinear least squares optimization method that merges the Gauss-Newton method with gradient descent [33]. Each iteration requires gradient and Hessian matrix computations for parameter updates. In this context, the objective function's Jacobian matrix is essential. Applying the chain rule, we derive:

$$\frac{\partial F}{\partial w_i} = 2\sum_{i=1}^{N}\frac{\partial \Delta(x_i)}{\partial w_i} \tag{6}$$

where, $\Delta(x) = gmm\big(x,T(M,\boldsymbol{w})\big) - gmm(x,S)$. As $gmm(x,S)$ is a known constant, the partial derivative of $\Delta(x)$ with respect to w_i can further be expressed as:

$$\frac{\partial \Delta(x)}{\partial w_i} = \sum_{j=1}^{N}\frac{\partial gmm\big(x,T(x_j)\big)}{\partial T(x_j)}\cdot\frac{\partial T(x_j)}{\partial w_i} \tag{7}$$

Here, $\frac{\partial T(x_j)}{\partial w_i} = RBF\big(x_j,c_i,\delta\big)$ and N corresponds to the sample count along the horizontal axis of the model point set.

3.5. Alignment Quality Assessment Method

In an ideal alignment between scene and target point sets, their Gaussian mixtures across horizontal coordinates should align closely, as depicted in Figure 5(a). However, even after optimization, notable disparities can remain between point sets of different categories, as highlighted in Figure 5(b). Furthermore, when two carriage images undergo vertical mirroring, their Gaussian mixtures might appear similar despite a mismatch in point set categories, as demonstrated in Figure 5(c). Clearly, relying purely on Gaussian mixture similarity for point set matching isn't sufficient.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2024 doi:10.20944/preprints202409.1075.v1

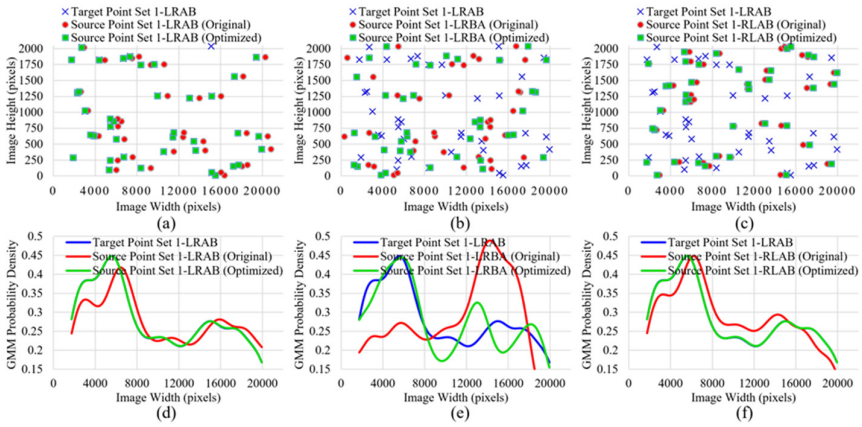


Figure 5. Comparison of point sets and their GMM probability densities before and after RBF transformation with optimized weights. (a) and (d) shows the intra-category transformation within 1-LRAB. (b) and (e) highlights the transformation from 1-LRBA to 1-LRAB. (c) and (f) presents the transformation from 1-RLAB to 1-LRAB.

To tackle these issues, this study proposes an intuitive yet effective method to evaluate the quality of point set alignments, providing a reliable metric to gauge the success of template matching. In detail, following the point set alignment, the scene point set, denoted as $\boldsymbol{S}$, undergoes a transformation into the coordinate space of the target point set $\boldsymbol{M}$. For each point s in $\boldsymbol{S}$, its closest counterpart in $\boldsymbol{M}$ is determined, expressed as:

$$NN(s) = arg\min_{m \in \boldsymbol{M}} distance(T(s), t) \quad (8) \tag{8}$$

Here, $NN(s)$ signifies the closest point to s within $\boldsymbol{M}$, while $distance(a, b)$ quantifies the spatial separation between points a and b. Given τ as a set distance threshold, if $distance(T(s), t) < \tau$, it's inferred that point s has correctly matched. The fraction of accurate matches, denoted as $match\_ratio$, is derived using a subsequent formula. When $match\_ratio$ exceeds a threshold γ, the template matching process is judged successful. The $match\_ratio$ serves as the score for match quality. Finally, among all templates, the category corresponding to the highest-scoring template is recognized as the image's identification result.

$$match\_ratio = \frac{number\ of\ correctly\ matched\ points\ in\ S}{total\ number\ of\ points\ in\ S} \quad (8) \tag{9}$$

4. Experiments

Freight train carriages are designed and manufactured to meet the diverse requirements of different types of cargo. Due to the wide variety of rail freight types, there is also a rich variety of corresponding carriage models. Additionally, the complex outdoor environment on railway lines means that the consistency of line-scan images of freight train carriages is far inferior to that of subway and high-speed train carriages. Therefore, using line-scan images of freight train carriages to evaluate the performance of image recognition algorithms is more convincing. In this experiment, we collected line-scan images of 19 different models of freight carriages. Based on the different movement directions of the vehicles, the images for each model were further subdivided into four subcategories: LRAB, RLBA, RLAB, and RLBA. Each subcategory contains 200 different images, exhibiting varying degrees of local nonlinear scale distortion in the horizontal direction. This results in a dataset comprising 76 categories and a total of 1.52 million high-resolution carriage images. Each image has a height resolution of 2048 pixels, and the width resolution ranges from n×2000 (where n=8 to 13), depending on the length of the carriage. The dataset is evenly divided into a training set and

 doi:10.20944/preprints202409.1075.v1

a testing set. Additionally, to evaluate our method, we constructed a standard reference image library consisting of 76 undistorted standard images, with one reference image for each category. Object detection was performed on both the test set and the template set, and the results were converted into corresponding point sets for subsequent algorithm performance evaluation.

In Section II, we analyzed and discussed existing image classification methods. Among machine learning-based image recognition algorithms, the Bag of Words (BoW) model stands out for its high computational efficiency and scalability, particularly in handling large-scale feature descriptors. Given the high resolution of line-scan carriage images, the extracted local feature quantities are substantial. Considering processing speed and algorithmic complexity, other traditional machine learning methods are unsuitable for recognizing line-scan carriage images. Therefore, in our experiments, we selected the BoW model as a representative machine learning method, combined with different feature extraction methods, to compare recognition performance. Additionally, the unique nature of line-scan carriage images requires substantial memory and GPU resources for existing image retrieval methods. Currently, we do not have the hardware capabilities to conduct comparative experiments under these conditions. Ultimately, we selected three categories of methods for comparison: BoW-based methods, deep learning-based multi-class image classification methods, and the proposed method. Performance evaluation on the carriage dataset was conducted in terms of image recognition accuracy, processing speed, and algorithm robustness.

During the evaluation of deep learning-based methods, high-resolution input images caused GPU memory shortages, hindering model training and inference. Conversely, excessively low resolutions resulted in significant loss of image details, affecting recognition accuracy. To balance GPU memory consumption and accuracy loss, we uniformly scaled all network model input images to a resolution of 448×3200pixels. In other comparative experiments, we downsampled the input images by a factor of 4 to ensure a resolution similar to that of the multi-classification network model, thereby ensuring the fairness and rigor of the comparative experiments.

4.1. Parameter Selection

This method involves four main parameters: τ, γ, δ, and Σ, each influencing the matching process and the final accuracy statistics differently. Below, I will explain the selection rationale for each parameter.

Selection of τ and γ: These two parameters are thresholds set to determine whether template matching is successful. Although they do not directly affect the point set matching process, they determine the accuracy statistics of the matching result:

τ is determined based on the maximum allowable error under scale distortions in the images. Considering the actual detection environment of the train images and the real target detection results, we set τ=200 to account for potential scale variations.

γ represents the proportion of successfully matched points in the template point set. We set γ to 0.75, meaning that if the proportion exceeds this value, the match is considered successful. A higher γ can reduce false positives but may increase the risk of missing matches, while a lower γ may lead to more false positives. Based on experimental results, we selected this value as an optimal balance.

Selection of δ: We chose $\delta = 1$ for the iterative optimization process of point set alignment. The point set coordinates are normalized along the height direction of the image during alignment, but there is local nonlinear distortion in the width direction. The train's speed feedback mechanism somewhat limits this distortion, and the horizontal offset usually does not exceed the height of the image. Therefore, $\delta = 1$ ensures that each point's Gaussian distribution in the template aligns with the height direction of the image, causing most matching points to fall within the high-response region of the Gaussian distribution. This enhances the robustness of the matching process and avoids excessive variance, which could lead to too much overlap of template points and local optima.

Selection of Σ: Σ is a key parameter in the template point set matching process. We tested the impact of Σ values ranging from 0.1 to 1 on the matching success rate. As shown in Table 1, the experimental results demonstrate:

 doi:10.20944/preprints202409.1075.v1

Table 1. Impact of Σ on Mean Recognition Accuracy (%) and Mean Matching Ratio (%).

Σ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Mean Recognition Accuracy (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	87.50
Mean Matching Ratio (%)	99.77	99.84	99.71	99.96	100.00	97.46	96.86	96.66	96.25	95.32

When Σ is less than 0.9, the image recognition accuracy remains close to 100%. It is only when Σ exceeds 0.9 that the recognition accuracy starts to decline. This indicates that the method has good adaptability to parameter selection.

For matching success rate, $\Sigma = 0.5$ produces the best results, showing that this value yields the most stable matching outcome, as more successfully matched points enhance the robustness of the matching process.

Based on the above analysis, we selected the following final parameters for this study: $\delta = 1, \Sigma = 0.5, \tau = 200, \gamma = 0.75$. This combination demonstrated the best matching performance in our experiments.

4.2. Image Recognition Accuracy Evaluation

Due to different movement modes, the same carriage can produce four different types of line-scan integrated images. Since the structure of the carriage is asymmetrical, images captured from the left side differ in features from those captured on the right side. Images captured from the same side but with different movement directions will appear as left-right flipped versions of each other. Spatially, these four subcategories of images have significant differences; however, from an image feature perspective, they exhibit high similarity, which directly affects image recognition accuracy. Although the image category relationships may seem straightforward, they pose significant challenges for existing image recognition algorithms. To fully demonstrate these challenges and highlight the advantages of our proposed method, we reconstructed a dataset with 19 categories based on the original carriage dataset, ignoring distinctions in carriage movement modes. Comparing the accuracy of algorithms on these two datasets allows us to evaluate their ability to recognize different carriage models and their effectiveness in handling images from various movement modes. Given that our method achieves 100% recognition accuracy across all categories, we present the average Top-1 Accuracy of subcategories for each model category to save space.

Table 2 presents the accuracy comparison between our method and the Bag of Words (BoW) image recognition method. The BoW model combines five feature extraction methods to construct visual word representations of images: AKAZE [34], BRISK [35], ORB [36], SIFT [37], and SuperPoint [38], a deep learning-based feature extraction method known for its superior local feature representation capabilities compared to traditional methods. These five feature extraction methods are commonly used in computer vision and are representative of the field. From Table 2, it is evident that the representation ability of feature descriptors directly determines the classification accuracy of the BoW model—the stronger the feature representation, the higher the image recognition accuracy. The deep learning-based feature extraction method achieves the best recognition accuracy in the BoW model. However, the BoW model, which relies on local features, is sensitive to local feature differences in images. For carriage types with inconsistent images due to complex usage environments, the recognition accuracy is poor (e.g., B23, BX1K, BX1K-1, NX17K, X70). For categories with high image consistency, the recognition accuracy approaches 100% (e.g., BDL1, C70, NX70A, C64K).

 doi:10.20944/preprints202409.1075.v1

Table 2. Comparison of Recognition Accuracy: Proposed Method VS. Bag-of-Words.

Class	76-category							19-category					
	AKAZE [34]	BRISK [35]	ORB [36]	SIFT [37]	Super [38]	NCC [39]	Ours	AKAZE [34]	BRISK [35]	ORB [36]	SIFT [37]	Super [38]	Ours
B22	0.087	0.095	0.11	0.13	0.122	0.059	1.0	0.133	0.255	0.225	0.69	0.355	**1.0**
B22-1	0.397	0.225	0.23	0.27	0.32	0.238	1.0	0.488	0.278	0.183	0.71	0.773	**1.0**
B23	0.14	0.167	0.112	0.1275	0.17	0.093	1.0	0.44	0.518	0.365	0.227	0.304	**1.0**
B23-1	0.445	0.502	0.477	0.455	0.477	0.421	1.0	0.688	0.82	0.988	0.30	0.361	**1.0**
BDL1	1.0	1.0	0.975	0.998	0.985	0.942	1.0	1.0	0.715	0.59	0.295	0.613	**1.0**
BH1	0.16	0.09	0.057	0.237	0.45	0.149	1.0	0.308	0.113	0.14	0.41	0.19	**1.0**
BX1K	0.112	0.109	0.108	0.152	0.252	0.097	1.0	0.444	0.438	0.385	0.644	1.0	**1.0**
BX1K-1	0.082	0.129	0.102	0.087	0.112	0.052	1.0	0.179	0.185	0.185	0.35	0.448	**1.0**
C64K	1.0	0.982	1.0	1.0	0.93	0.932	1.0	1.0	0.63	0.565	0.69	0.68	**1.0**
C70E	0.243	0.043	0.058	0.23	0.02	0.069	1.0	0.8	1.0	1.0	0.68	0.57	**1.0**
C70	1.0	0.968	1.0	0.998	0.995	0.942	1.0	1.0	0.578	0.595	0.843	0.503	**1.0**
JSQ5	0.273	0.175	0.168	0.188	0.335	0.178	1.0	0.388	0.315	0.401	0.688	0.995	**1.0**
JSQ6	0.089	0.164	0.196	0.199	0.139	0.107	1.0	0.181	0.544	0.628	0.818	0.74	**1.0**
JSQ6-1	0.163	0.155	0.143	0.155	0.203	0.114	1.0	0.315	0.3325	0.45	0.2525	0.435	**1.0**
NX17K	0.133	0.233	0.165	0.118	0.118	0.103	1.0	0.14	0.388	0.36	0.308	0.463	**1.0**
NX70A	0.998	0.995	0.963	0.998	0.995	0.940	1.0	1.0	1.0	0.963	0.998	0.593	**1.0**
NX70	0.515	0.503	0.548	0.470	0.483	0.454	1.0	0.53	0.683	0.538	0.373	0.123	**1.0**
X68BK	0.5	0.573	0.543	0.455	0.473	0.459	1.0	0.728	0.363	0.18	0.413	0.303	**1.0**
X70	0.1	0.113	0.143	0.280	0.018	0.081	1.0	0.16	0.793	0.668	0.605	1.0	**1.0**
Ave. ACC	0.391	0.379	0.373	0.397	0.399	0.338	1.0	0.515	0.51	0.481	0.538	0.549	**1.0**

To further evaluate the impact of local nonlinear distortions on traditional image recognition methods, we conducted a comparative experiment using the template matching algorithm based on normalized cross-correlation, as described in Yoo and Han [39]. This method is commonly employed in image processing tasks for template matching by calculating the correlation between image regions. In our experiment, we applied this method to the 79-class carriage linear array image classification task. The results demonstrate that the normalized cross-correlation method performs poorly on this task, with a significant decline in recognition accuracy. Specifically, the method fails to effectively handle the local nonlinear scale distortions present in the carriage linear array images, leading to a substantial reduction in recognition accuracy, and in many cases, the method fails to produce valid matches altogether.

The poor performance is attributed to the sensitivity of correlation-based methods to image distortions and inconsistencies. In particular, the height inconsistency between carriage images and the scale distortions cause significant changes in the image features, making it difficult for the normalized cross-correlation to produce reliable results. This further highlights the challenges posed by carriage linear array images and demonstrates the limitations of traditional image recognition methods when dealing with images subject to such distortions.

Comparing the classification accuracy of the two datasets in Table 2, the 19-category dataset performs better than the 76-category one. The total number of images in both datasets is the same, but the reduced number of categories increases the diversity of images within each category. This result indicates that the BoW model has some generalization ability for images under different

 doi:10.20944/preprints202409.1075.v1

movement modes, which is beneficial for general image recognition tasks. However, for line-scan carriage images, this generalization is undesirable because the goal is to distinguish images based on different movement modes. The fundamental reason for this phenomenon is that the BoW model represents images as unordered collections of feature words, losing spatial information among features. The analysis of the experimental results in Table 2 further verifies that the BoW model is unsuitable for recognizing line-scan carriage images.

Table 3 presents the accuracy comparison between our method and multi-class neural network methods. The six multi-class network models used for comparison are representative of current image classification methods, exhibiting high recognition accuracy on multiple datasets as detailed in Section II. Compared with the results in Table 2, the multi-class neural networks significantly outperform the BoW method in overall recognition accuracy. The recognition accuracy distribution is consistent with that of the BoW method, with high accuracy for categories with consistent images (e.g., BDL1, C70, NX70A, C64K). However, for inconsistent image categories, the recognition accuracy still falls short of practical application standards.

Table 3. Comparison of Recognition Accuracy: Proposed Method VS. Multi-Classification Models.

Class	76-category							19-category						
	Efficient NetV2 [11]	Resmlp [12]	Resnet [13]	Senet [15]	Shufflenet [16]	Vit [17]	Ours	Efficient NetV2 [11]	Resmlp [12]	Resnet [13]	Senet [15]	Shufflenet [16]	Vit [17]	Ours
B22	0.123	0.513	0.513	0.600	0.128	0.253	**1.0**	0.165	0.193	0.680	0.787	0.573	0.655	**1.0**
B22-1	0.475	0.380	0.380	0.348	0.438	0.338	**1.0**	0.673	0.440	0.818	0.448	0.328	0.168	**1.0**
B23	0.298	0.488	0.488	0.393	0.158	0.383	**1.0**	0.563	0.406	0.405	0.993	0.660	0.790	**1.0**
B23-1	0.500	0.503	0.503	0.580	0.498	0.535	**1.0**	0.750	0.120	0.635	0.545	0.553	0.693	**1.0**
BDL1	1.0	1.0	1.0	1.0	1.0	1.0	**1.0**	1.0	0.520	0.678	0.750	0.663	0.838	**1.0**
BH1	0.215	0.308	0.308	0.220	0.085	0.120	**1.0**	0.228	0.513	0.713	0.815	0.808	0.011	**1.0**
BX1K	0.149	0.423	0.423	0.460	0.168	0.402	**1.0**	0.149	0.703	0.418	0.753	0.413	0.305	**1.0**
BX1K-1	0.139	0.309	0.309	0.386	0.094	0.215	**1.0**	0.160	0.548	0.680	0.785	0.500	0.453	**1.0**
C64K	1.0	1.0	1.0	1.0	0.985	1.0	**1.0**	1.0	0.730	0.728	0.238	0.185	0.703	**1.0**
C70E	0.033	0.260	0.260	0.160	0.213	0.035	**1.0**	0.033	0.693	1.0	0.873	0.423	0.455	**1.0**
C70	1.0	1.0	1.0	1.0	1.0	1.0	**1.0**	1.0	0.443	1.0	0.765	0.137	0.565	**1.0**
JSQ5	0.788	0.510	0.510	0.623	0.300	0.625	**1.0**	0.908	0.485	0.693	0.560	0.965	0.708	**1.0**

 doi:10.20944/preprints202409.1075.v1

JSQ6	0.366	0.528	0.528	0.558	0.145	0.319	**1.0**	0.468	0.533	0.643	0.778	0.340	0.680	**1.0**
JSQ6-1	0.250	0.268	0.268	0.378	0.158	0.433	**1.0**	0.355	0.573	0.500	1.0	0.513	1.0	**1.0**
NX17K	0.248	0.158	0.158	0.238	0.133	0.118	**1.0**	0.248	0.260	0.608	0.398	0.990	0.398	**1.0**
NX70A	1.0	1.0	1.0	1.0	1.0	1.0	**1.0**	1.0	0.245	0.613	0.440	0.645	0.618	**1.0**
NX70	0.520	0.513	0.513	0.513	0.513	0.513	**1.0**	0.523	0.343	0.659	0.928	0.570	0.280	**1.0**
X68BK	0.500	0.503	0.503	0.500	0.500	0.500	**1.0**	0.750	0.490	0.860	0.270	0.298	0.250	**1.0**
X70	0.555	0.583	0.583	0.553	0.133	0.178	**1.0**	0.630	0.513	0.298	0.680	0.618	0.613	**1.0**
Ave. ACC	0.482	0.539	0.539	0.553	0.402	0.472	**1.0**	0.558	0.460	0.664	0.674	0.536	0.536	**1.0**

Comparing the accuracy performance on the two datasets in Table 3 also shows that the multi-class network models perform significantly better on the 19-category dataset than on the 69-category dataset and outperform the BoW results in Table 1. This indicates that the generalization ability of multi-class network models far exceeds that of the BoW model, primarily due to the enhanced feature representation capabilities of neural networks trained on the carriage dataset. However, the significant accuracy improvement also reveals the limitations of multi-class network models in the 76-category line-scan carriage image recognition task. Despite extensive training, they cannot accurately recognize carriage movement modes for inconsistent image categories. However, accurate identification of reference images for fault analysis is crucial in train fault detection.

Combining the accuracy statistics from Tables 2 and 3, our method achieves 100% Top-1 Accuracy on both datasets, significantly outperforming other methods. Neither the BoW model nor the deep learning-based multi-class network models effectively recognize line-scan carriage images under different movement modes. In contrast, our method can accurately recognize carriage types and further distinguish carriage movement modes, ensuring the accurate acquisition of reference images for fault analysis.

4.3. Comparison of Image Recognition Speeds

To ensure fairness in evaluating the image recognition speeds of different algorithms, all algorithms were run on the same computer. The computer configuration is as follows: CPU—i7-11800, 2.3GHz; GPU—NVIDIA RTX 3050, 8GB VRAM; RAM—16GB. The comparison results of image recognition speeds are shown in Table 4.

Table 4. Comparison of Average Recognition Speeds Among Different Methods.

Method	**AKAZE**	**BRISK**	**ORB**	**SIFT**	**Superpoint**	**EfficientNetV2**	**Resmlp**	**Resnet**	**Senet**	**Shufflenet**	**Vit**	**Ours**
Time(s)	3.263	4.586	5.312	2.274	4.762	0.082	0.031	0.054	0.063	0.139	0.082	**0.024**

As observed, the Bag of Words (BoW) based methods have the slowest recognition speeds. This is primarily due to the large size of the line-scan images of train carriages, which makes the feature

 doi:10.20944/preprints202409.1075.v1

extraction process time-consuming. Additionally, the total number of extracted feature points also impacts the processing time. Considering these factors, the BoW method using SIFT features is the fastest among the BoW methods, with an average recognition time of 2.274 seconds per carriage line-scan image. Compared to the two-stage BoW methods, multi-class neural network methods demonstrate significantly faster recognition speeds, mainly due to their end-to-end structure. Among these, the ResMLP model has the fastest inference speed, with an average detection time of 31ms per image. Our proposed method exhibits the fastest image recognition speed among all methods. For the 76-category recognition task, the average recognition time per image is only 24ms, showcasing a clear advantage over the other two categories of methods. The efficiency of our method can be attributed to the sparse point cloud registration in the template matching process, which inherently involves minimal computation. Additionally, the high efficiency of the C++ programming language contributes to the speed of our algorithm.

4.4. Algorithm Robustness Evaluation

In our proposed method, the point sets used for template matching are derived from the object detection results of images. Due to the complexity of real-world scenarios, certain types of carriage images may exhibit poor consistency. Significant feature differences among these images can lead to false detections or missed detections during the object detection process, resulting in point sets with extra or missing points for template matching. Although object detection algorithms are minimally affected by nonlinear scale distortions in images (data augmentation techniques used during model training effectively mitigate the impact of scale distortions on object detection), the use of weighted radial basis functions to describe the nonlinear scale distortions in images introduces potential challenges. The iterative optimization process for solving the parameters can get trapped in local optima when there are extra or missing points in the point set, or when there are extreme nonlinear scale distortions in the image, leading to incorrect point set registration results and ultimately affecting image recognition accuracy. To evaluate the accuracy and robustness of our proposed method under such conditions, we assess the algorithm's recognition accuracy in terms of resistance to distortion, missed detections, and false detections.

4.4.1. Evaluation of Resistance to Local Nonlinear Distortions

To evaluate our method's ability to resist local nonlinear scale distortions, we generated test point sets with varying levels of local distortion from 0% to 50% in 5% increments for 76 template point sets. The distortion percentage is relative to the horizontal dimension of the image. For each category, 100 sets of test point sets were generated with randomly positioned distortions at each specified distortion level. To realistically simulate the localization error of object detection algorithms, each point in the target point set was randomly offset by up to 20 pixels in both horizontal and vertical directions. Local nonlinear scale distortions were simulated by applying specific scale offsets to random horizontal positions and using cubic polynomial interpolation for local interpolation. Since the point sets of the same carriage type under four movement modes have a unique spatial relationship, their recognition accuracies are essentially the same. Therefore, we calculate the average recognition accuracy by carriage type when reporting accuracy statistics. Figure 6 shows the recognition accuracy of our method under different distortion scales. As expected, the matching accuracy generally decreases with increasing distortion levels. However, the recognition accuracy remains unaffected in the 0% to 10% distortion range. At a 15% distortion level, minor changes in recognition accuracy are observed for some categories. When the distortion level exceeds 25%, the recognition accuracy significantly decreases, but even at a 50% distortion level, the recognition accuracy for all categories remains above 69%. These results demonstrate the excellent robustness of our method in resisting local nonlinear scale distortions in line-scan images of train carriages.

 doi:10.20944/preprints202409.1075.v1

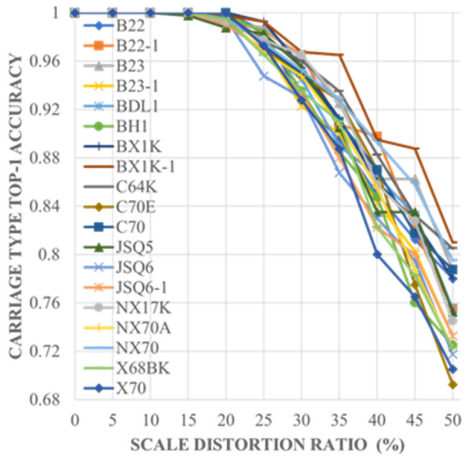


Figure 6. Recognition accuracy under different distortion scales.

4.4.2. Evaluation of Resistance to False Detections

False detections in object detection algorithms mainly fall into two categories: incorrect category identification and extraneous detection results. Since our method does not rely on point categories, the first type of false detection does not affect it. Therefore, this section focuses on evaluating the performance of our method under different levels of extraneous detections. To simulate different false detection levels, we generated new test point sets by adding random points to the original test point sets, increasing the total number of points by 5% increments up to 35%. Figure 7 shows the corresponding accuracy evaluation results using the same statistical method as previously described. The results indicate that although the recognition accuracy generally decreases with increasing false detection rates, recognition accuracy is unaffected when the false detection rate is below 10%. At a 15% false detection rate, some categories experience slight declines in recognition accuracy, but the overall accuracy remains high. When the false detection rate exceeds 20%, overall recognition accuracy starts to decline significantly, but the recognition accuracy for most categories remains relatively high. These experimental results demonstrate the robustness of our method in handling image category recognition problems in the presence of false detections.

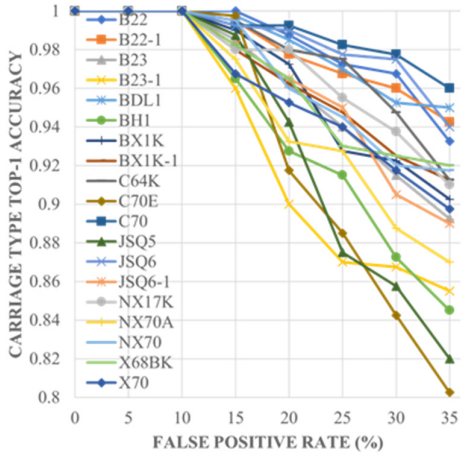


Figure 7. Recognition accuracy under different false detection rates.

 doi:10.20944/preprints202409.1075.v1

4.4.3. Evaluation of Resistance to Missed Detections

To further evaluate our method's performance under different missed detection rates, we generated new test point sets by randomly removing points from the original test point sets, reducing the total number of points by 5% increments up to 35%. Figure 8 shows the corresponding accuracy evaluation results. As the missed detection rate increases, image recognition accuracy shows a decreasing trend. Recognition results are unaffected when the missed detection rate is below 10%. At a 15% missed detection rate, recognition accuracy for some categories slightly decreases. Beyond a 20% missed detection rate, overall recognition accuracy significantly declines, but the accuracy for some categories remains above 90% even at a 35% missed detection rate. These experimental results confirm that our method retains strong image category recognition capabilities in the presence of missed detections.

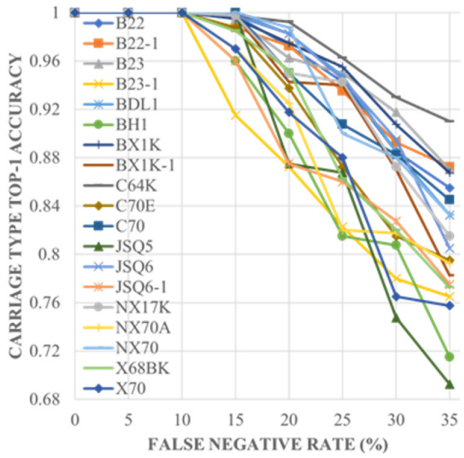


Figure 8. Recognition accuracy under different missed detection rates.

5. Conclusions

This study introduces an image recognition method tailored specifically for carriage linear array images, which are characterized by extreme aspect ratios, ultra-high resolution, and local nonlinear scale distortions. These inherent properties significantly limit the recognition accuracy of many existing algorithms, particularly because they require substantial computational resources and memory to handle such complex image data. Traditional methods, such as feature-based approaches, often struggle with the sheer scale and variability of these images, resulting in inefficient training and poor performance.

To address these challenges, we reformulated the recognition problem into a template matching task focused on sparse point set alignment. A coordinate transformation using weighted radial basis functions was proposed to handle the local nonlinear scale distortions typically present in linear array images. Additionally, we designed a unique objective function to enhance the similarity between mixed Gaussian distributions of point sets, ensuring accurate parameter determination for the transformation equation, even in cases where point set correspondences are ambiguous or the point sets vary in size.

Extensive experimental data validate the effectiveness of our method: under local nonlinear scale distortions of up to 15%, the carriage image recognition accuracy reaches 100%. Even with a 10% increase or decrease in the number of detected targets, the recognition accuracy remains unaffected. Moreover, identifying the target category from 79 templates takes only 24ms. This demonstrates the algorithm's exceptional performance in terms of recognition accuracy, processing speed, and robustness. Our method allows for the direct identification of carriage models from line-scan images, eliminating the need for traditional AEI equipment.

 doi:10.20944/preprints202409.1075.v1

However, we acknowledge that as the number of template categories increases, the image recognition processing time will also increase accordingly. This limitation will be addressed in future research as we explore ways to further improve the algorithm's efficiency.

Author Contributions: Conceptualization, Z.F. and X.P.; methodology, Z.F.; software, Z.F.; validation, Z.F., X.P.; formal analysis, Z.F.; investigation, Z.F.; resources, Z.F.; data curation, Z.F.; writing—original draft preparation, Z.F.; writing—review and editing, Z.F.; visualization, Z.F.; supervision, X.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to restrictions eg privacy or ethical.

Acknowledgments: This work was supported in part by the National Science Fund for Youth Science Foundation of China under Grant 62105015, in part by the Aviation Science Foundation 2023 under Grant 20230046051003, and in part by the Beihang University Frontier Cross Fund 2023.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chang, L.; Liu, Z.; Shen,Y.; Zhang,G. Novel multistate fault diagnosis and location method for key components of high-speed trains. *IEEE Trans. Ind. Electron*. 2020, 68, 3537-3547.
2. Lu, S.; Liu, Z.; Shen, Y. Automatic fault detection of multiple targets in railway maintenance based on time-scale normalization. *IEEE Trans. Instrum. Meas*. 2018, 67, 849-865.
3. Fu, Z.; Pan, X.; Zhang, G. Linear array image alignment under nonlinear scale distortion for train fault detection. *IEEE Sens. J.*, doi: 10.1109/JSEN. 2024, 3404950.
4. Chen, H.; Jiang, B. A review of fault detection and diagnosis for the traction system in high-speed trains. *IEEE Trans. Intell. Transp. Syst*. 2019, 21, 450-465.
5. Chandra, M.A.; Bedi, S.S. Survey on SVM and their application in image classification. *Int. J. Inf. Technol*. 2021, 13, 1-11.
6. Bosch, A.; Zisserman, A.; Munoz, X. Image classification using random forests and ferns. *In Proc. 2007 IEEE 11th Int. Conf. Comput. Vis*. 2007, 1-8.
7. Amato, G.; Falchi, F. kNN based image classification relying on local feature similarity. *In Proc. 3rd Int. Conf. Similarity Search Appl*. 2010, 101-108.
8. Timofte, R.; Tuytelaars, T.; Van Gool, L. Naive Bayes image classification: beyond nearest neighbors. *In Proc. Asian Conf. Comput. Vis., Berlin, Heidelberg: Springer*, 2012, 689-703.
9. Wang, C.; Huang, K. How to use bag-of-words model better for image classification. *Image Vis. Comput.* 2015, 38, 65-74.
10. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *In Proc. Int. Conf. Mach. Learn., PMLR*, 2019, 6105-6114.
11. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proc. Int. Conf. Mach. Learn., PMLR, 2021,10096-10106.
12. Touvron, H.; Bojanowski, P.; Caron, M. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell*. 2022, 45, 5314-5321.
13. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, 770-778.
14. Liang,J.Image classification based on RESNET. J. Phys.: Conf. Ser. 2020, 1634, 012110.
15. Cheng, D.; Meng, G.; Cheng, G. SeNet: Structured edge network for sea–land segmentation. *IEEE Geosci. Remote Sens. Lett.* 2016, 14, 247-251.
16. Zhang, X.; Zhou, X.; Lin, M. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, 6848-6856.
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A. An image is worth 16x16 words: Transformers for image recognition at scale.2020, arXiv preprint, arXiv:2010.11929.
18. Zheng, L.; Ai, P.; Wu, Y. Building recognition of UAV remote sensing images by deep learning. *In Proc. IGARSS 2020 IEEE Int. Geosci. Remote Sens. Symp.*, 2020, 1185-1188.
19. Hou, X.; Ao, W.; Xu, F. End-to-end automatic ship detection and recognition in high-resolution Gaofen-3 spaceborne SAR images. *In Proc. IGARSS 2019 IEEE Int. Geosci. Remote Sens. Symp.*, 2019, 9486-9489.

20. Sun, Z.; Xiong, B.; Lei, Y. Ship classification in high-resolution SAR images based on CNN regional feature fusion. *In Proc. 2021 CIE Int. Conf. Radar*, 2021,1445-1449.
21. Liu, Y.; Chang, M.; Xu, J. High-resolution remote sensing image information extraction and target recognition based on multiple information fusion. *IEEE Access* 2020, 8, 121486-121500.
22. Liu, Z.; Gao, G.; Sun, L.; HRDNet: High-resolution detection network for small objects. *In Proc. 2021 IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, 1-6.
23. Růžička, V.; Franchetti, F. Fast and accurate object detection in high resolution 4K and 8K video using GPUs. *In Proc. 2018 IEEE High Perform. Extreme Comput. Conf. (HPEC)*, 2018, 1-7.
24. Li, Y. Research and application of deep learning in image recognition. *In Proc. 2022 IEEE 2nd Int. Conf. Power Electron. Comput. Appl. (ICPECA)*, 2022, 994-999.
25. Liu, H.; Wang, R.; Shan, S. Deep supervised hashing for fast image retrieval.In Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, 2064-2072.
26. Gordo, A.; Almazan, J. Revaud, J. End-to-end learning of deep visual representations for image retrieval. *Int. J. Comput. Vis.* 2017, 124, 237-254.
27. Liu, C.; Yu, G.; Volkovs, M. Guided similarity separation for image retrieval. *Adv. Neural Inf. Process. Syst*. 2019, 32.
28. Ramzi, E.; Audebert, N.; Rambour, C. Optimization of rank losses for image retrieval. 2023, arXiv preprint, arXiv:2309.08250.
29. Smeulders, A. W. M.; Worring, M.; Santini, S. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 2000, 22, 1349-1380.
30. Noh, H.; Araujo, A.; Sim, J. Large-scale image retrieval with attentive deep local features. *In Proc. IEEE Int. Conf. Comput. Vis.*, 2017, 3456-3465.
31. Henkel, C. Efficient large-scale image retrieval with deep feature orthogonality and hybrid-swin-transformers.2021, arXiv preprint, arXiv:2110.03786.
32. Lee, C. C.; Chung, P.; Tsai, J. Robust radial basis function neural networks. *IEEE Trans. Syst. Man, Cybern. B* 1999, 29, 674-685.
33. Ranganathan, A. The Levenberg-Marquardt algorithm. *Tutoral on LM Algorithm* 2004, 11, 101-110.
34. Alcantarilla, P. F.; Solutions, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Pattern Anal. Mach. Intell*. 2011, 34, 1281-1298.
35. Leutenegger, S.; Chli, M.; Siegwart, R. BRISK: Binary robust invariant scalable keypoints. *In Proc. 2011 Int. Conf. Comput. Vis.*, 2011, 2548-2555.
36. Rublee, E.; Rabaud, V.; Konolige, K. ORB: An efficient alternative to SIFT or SURF. *In Proc. 2011 Int. Conf. Comput. Vis.*, 2011, 2564-2571.
37. Lowe, D.G.; Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 2004, 60, 91-110.
38. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, 224-236.
39. Yoo, J. C.; Han, T. H. Fast normalized cross-correlation. Circuits, Systems and Signal Processing 2009, 28, 819-843.