Deep Local Multi-level Feature Aggregation Based High-speed Train Image Matching

Jun Li[1], Xiang Li[1], Yifei Wei[1,*] and Xiaojun Wang[2]
[1] Beijing Key Laboratory of Work Safety Intelligent Monitoring, School of Electronic Engineering, Beijing University of Posts and Telecommunications
Beijing, 100876, China
[e-mail: lijun2021@bupt.edu.cn, 2020140419@bupt.edu.cn, weiyifei@bupt.edu.cn]
[2] Dublin City University
Dublin, Dublin 9, Ireland
[e-mail: xiaojun.wang@dcu.ie]
*Corresponding author: Yifei Wei

Received February 10, 2022; revised March 31, 2022; accepted April 24, 2022; published May 31, 2022

Abstract

At present, the main method of high-speed train chassis detection is using computer vision technology to extract keypoints from two related chassis images firstly, then matching these keypoints to find the pixel-level correspondence between these two images, finally, detection and other steps are performed. The quality and accuracy of image matching are very important for subsequent defect detection. Current traditional matching methods are difficult to meet the actual requirements for the generalization of complex scenes such as weather, illumination, and seasonal changes. Therefore, it is of great significance to study the high-speed train image matching method based on deep learning. This paper establishes a high-speed train chassis image matching dataset, including random perspective changes and optical distortion, to simulate the changes in the actual working environment of the high-speed rail system as much as possible. This work designs a convolutional neural network to intensively extract keypoints, so as to alleviate the problems of current methods. With multi-level features, on the one hand, the network restores low-level details, thereby improving the localization accuracy of keypoints, on the other hand, the network can generate robust keypoint descriptors. Detailed experiments show the huge improvement of the proposed network over traditional methods.

Keywords: Image matching, High-speed train, Multi-scale features, Artificial intelligence, Joint description and detection of local features

1. Introduction

Image matching can be regarded as a basic task in the detection process of high-speed train chassis. The chassis is an important part connecting the train and the rail, so its safety detection is a key aspect to ensure the safety of high-speed railway train [1].

At present, most of the detection schemes of high-speed railway trains in China are developed on the basis of Trouble of Moving EMU Detection System (TEDS) [2]. The method is to utilize computer vision technology to match the images collected by a line-array camera installed near the rails in different times, then using automatic technology or manpower to identify chassis faults or foreign objects. If any abnormality is detected, it will remind workers to respect. Compared with the traditional purely manual mode, this man-machine interaction saves a lot of time and reduces the demand for workers, thus improves the detection efficiency a lot.

Considering a standard high-speed train images matching process. When a high-speed railway train passes through the image acquisition area of TEDS system, the detection system completes the speed measurement of the train firstly, then the line scan camera takes pictures of the train chassis at an image acquisition rate corresponding to the measured train speed, so that the length of the images in the direction of train travel is approximately equal to the historical images. Inputting the saved previous image I_s and the currently image I_f located near the same chassis position into the computer, it extracts and matches the keypoints of the two associated images. Then, the homography transformation matrix H_{fs} between these two high-speed railway train images is calculated to obtain the transformed I_f. In this way, the image $I_{f'}$ under the coordinate of I_s is obtained. So, the image matching between current and previous high-speed rail train chassis images is realized. In this task, most current matching methods are carried out in pixel space (intensity-based approach) or manual feature space (feature-based approach) in a non-learnable manner. However, due to the speed measurement accuracy of the detection system is usually not so ideal, the imaging speed of the line scan camera does not match the actual running speed of the train, resulting in inconsistencies between the historical and current images of the same train chassis area, which brings challenges to the matching task. Moreover, as a result of the large changes in the appearance of the images caused by weather, light, and seasons at the time of collection, these traditional methods have problems such as low matching accuracy and lack of generalization ability. Since the quality of image matching is very important for subsequent tasks such as abnormal detection of the train chassis, there is an urgent need to improve the matching accuracy, and robustness to more complex scenes, which is of great significance.

With the rapid development of deep learning in the field of computer vision, there are more and more learning image matching methods to solve the drawbacks of traditional image matching methods. However, most of the existing learning methods usually only use the deepest feature map, which is several times smaller than the original image (usually 4 or 8 times) to detect and describe keypoints. The accuracy of locating keypoints in the deep feature map is definitely lower than that in the original image size. In addition, different keypoints in deep feature maps may share the same high level of semantic information. Therefore, their description vectors may be extremely similar, which degrades the performance of feature matching and thus inevitably reduces the accuracy of finding pixel-level correspondences, which is detrimental.

To the best of our knowledge, this work is the first to complete high-speed train image registration using deep learning methods. This work proposes a convolutional neural network (CNN) based high-speed train chassis images matching method, and alleviates the above

limitations from two aspects. Firstly, deep and shallow feature maps extracted by the network are sampled to high resolution, so as to improve the localization accuracy of keypoints. Secondly, multi-scale feature maps are used to collect rich local feature information and screen out the most discriminating descriptors. In short, by extracting dense keypoints with high localization precision, the algorithm is robust to almost all complex situations faced by high-speed railway.

The contributions of this article are as follows:

1. A CNN-based network that enables high-speed rail trains image matching, which improves robustness to complex illumination and weather conditions compared with traditional methods.

2. A dataset Constructed for high-speed rail train chassis matching, including optical distortions such as random illumination changes and perspective changes to simulate the actual detection environment as much as possible.

3. A multi-level feature extraction structure is designed to improve the localization accuracy of keypoints and the richness of the keypoint descriptors.

The rest of this article is as follows. Sec. 2 reviews the related work on image matching. Sec. 3 presents the proposed network architecture and the detailed implementations. Sec. 4 evaluates the performance of the proposed model. Sec. 5 concludes the paper.

2. Related Research

Traditional image matching methods are mainly divided into gray scale matching and feature-based matching. Image matching based on gray scale can be regarded as an iterative optimization process that directly compares all the contents of two images [2-4]. Although the image matching method based on intensity is simple to implement, it is difficult to meet the requirements of dynamic measurement in practice due to the large amount of calculation and time-consuming [5]. The classical feature-based image matching process includes keypoint detection, keypoint description, feature matching and geometric transformation estimation. Most of the early methods focuse on a certain step in keypoint detection [6] or keypoint description [7].

In view of the characteristics of high-speed train images, many researchers have improved popular keypoint detection or description algorithms which are based on classical keypoint detection algorithms such as SIFT [1, 8], SURF [5, 9, 10], ORB [11], etc. Peng [5] propose an abnormal detection algorithm for the bottom parts of the train based on the SURF feature of the rail edge image. Xie [1] designs an improved SIFT keypoint detection algorithm, which can reduce the number of keypoints in the flat area of the train component and increase the number of keypoints where the gradient of train components changes drastically, thereby improving the accuracy of keypoint matching. For feature matching, the nearest neighbor search algorithm is usually used to establish the pixel-level correspondence between two images. For the last step, according to DLT [12] or RANSCA [13], the 3×3 homography transformation matrix H is calculated. Then H is used to warp the source image so that the source image is aligned with the target image. The traditional feature-based image matching method only needs to complete the calculation of keypoint descriptors for keypoints, rather than for the entire image content. Therefore, the amount of calculation is small and thus more efficient, but traditional feature extraction algorithms usually only extract shallow-level features, and it is difficult to obtain deeper and more expressive features for them [14]. Therefore, the match result is poor when the appearance of the images changes too much owing to light, weather, etc. [15].

With the rapid development of deep learning [16-18], more and more image matching methods based on deep learning have been proposed [19-21]. Similar to traditional methods, there are two main categories of methods here, one does not detect and describe keypoint [22, 23], and the other does the opposite [24-26]. Most of the state-of-the-art methods can detect and describe keypoints simultaneously. D2-net [24] highly couples the keypoint detection module and description module. This method has good robustness to illumination changes, but it is slow and has low accuracy. ASLFeat [25] uses DCN (Deformable Convolution Net) and peak detection to extract keypoints, which greatly improves the localization accuracy of keypoints. SuperPoint [26] proposed a self-supervised joint keypoint detection and description architecture, consisting of a common VGG encoder and two independent branches: keypoint detection and keypoint descriptors.

However, most of the above learning feature-based methods do not utilize shallow features. This paper uses the advantages of multi-level feature maps to greatly improve the accuracy of keypoints positioning while generating more expressive and robust keypoint descriptors.

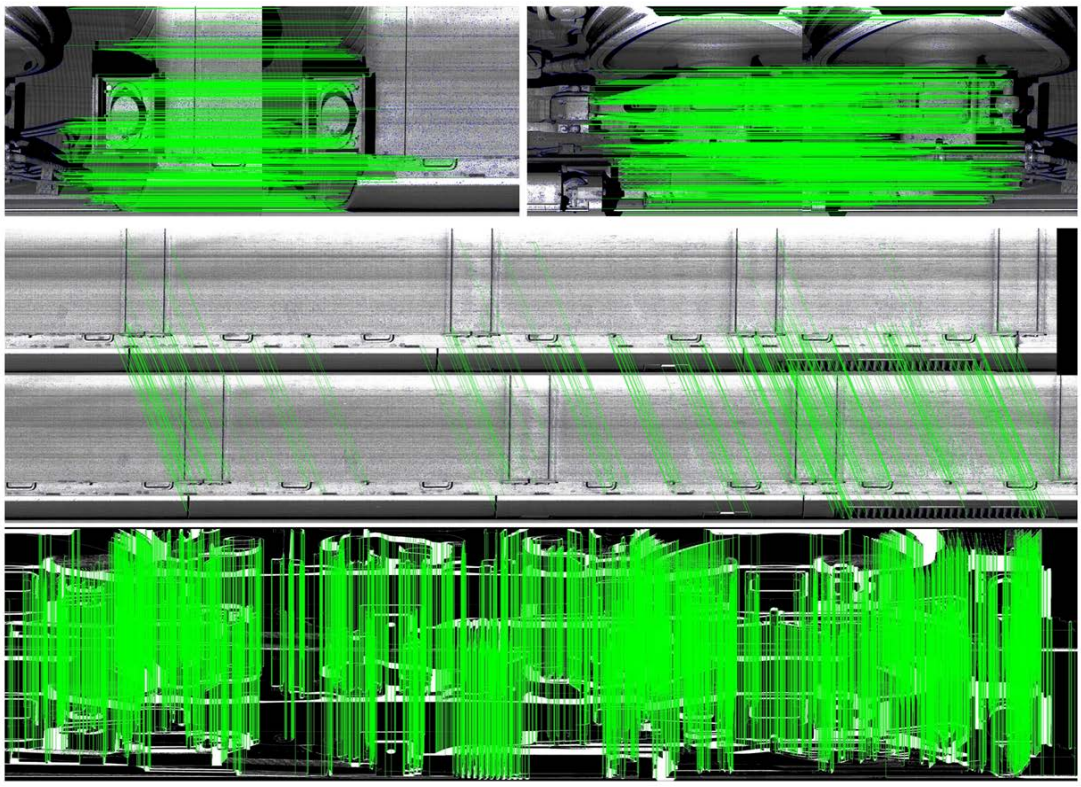


Fig. 1. Visualized results of the proposed method. The proposed method can find high-quality matches even in low-textured regions (the second row). It also works well on edge detection pictures (the third row).

3. Algorithm

Considering a standard high-speed train image matching process. Inputting the previously collected image I_s and the current image I_f of the same train at the same chassis position into a computer. Then, it extracts keypoints and corresponding high-dimensional description vectors, matches these keypoints using certain method, e.g., deep learning method or nearest neighbor

algorithm. Finally, it calculates the perspective transformation matrix H_{fs} between I_s and I_f. Thus, we get the pixel-level correspondence, which means the current and the previous high-speed train chassis images matching is achieved. This work uses a deep learning network for joint keypoint detection and description.

3.1 Keypoint Detection and Description Network

The proposed keypoint detection and description network takes a gray scale high-speed train chassis image as input and outputs the coordinates of detected keypoints and high-dimensional vectors as the corresponding descriptors. As shown in **Fig. 2**, this network consists of three parts: backbone feature encoder, keypoint detector and keypoint descriptor. It up-samples and combines three layers of feature maps. Then, the keypoint detection and description are performed sequentially.

3.1.1 Backbone Feature Encoder

The proposed backbone feature extractor is fully convolutional. The steps size of *Conv4* and *Conv7* layers are set to 2 to complete two times spatial down-sampling, and the steps size of the remaining layers are all set to 1. Considering an input image $I \in \mathbb{R}^{h\times w}$, the multi-scale features $P_i \in \mathbb{R}^{h_i\times w_i\times C_i}$ are the output of Conv3, Conv6 and Conv9 layers. Where $i \in \{1,6,9\}$, $h_i, w_i \in \{H, \frac{H}{2}, \frac{H}{4}\}$ and $C \in \{32,64,128\}$. The backbone multi-scale feature extraction can be expressed as Eq. (1).

$$P_i = Enconer(I), i \in \{1,6,9\} \tag{1}$$

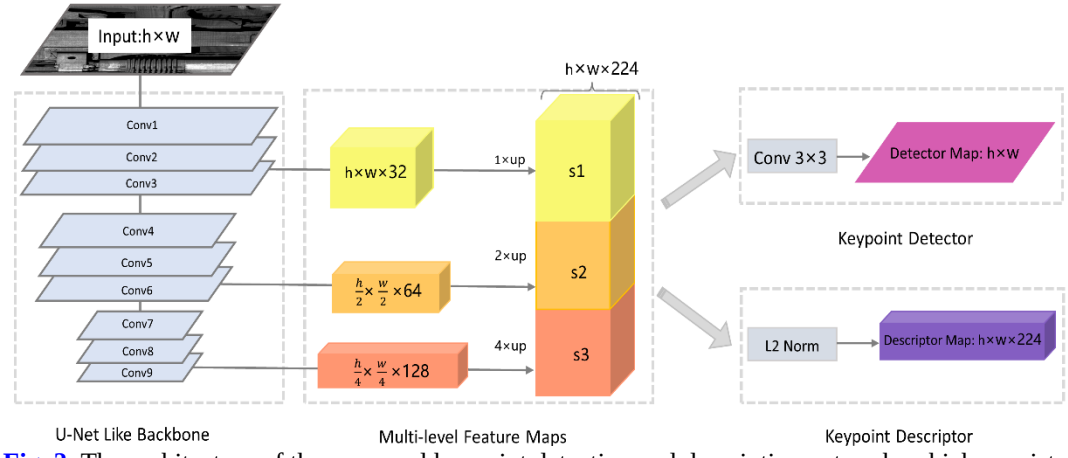


Fig. 2. The architecture of the proposed keypoint detection and description network, which consists of three parts: backbone feature encoder, keypoint detector and keypoint descriptor.

P_i is upsampled to the size of the original image in an unlearnable way, then multiplied by the corresponding weight coefficient and stacked together in channel dimension to obtain the final multi-scale features $M \in \mathbb{R}^{H\times W\times 224}$:

$$M = \frac{1}{\sum w_i}\sum w_i P_i \tag{2}$$

Where $w_i \in \{1,2,3\}$, is the weight coefficient of the multi-scale feature map. In this way, this work obtains the rich local shape details contained in the shallow features with a small amount of computational overhead, which is of great help to improve the localization accuracy of keypoints. At the same time, multiple layers of semantic information and deep

features that are robust to noise caused by illumination changes are also preserved, which is crucial for obtaining keypoint descriptors that are highly distinguishable.

3.1.2 Detector

Feature detector D uses a simple 3×3 convolution layer to process the multi-scale feature graph M, obtains the keypoint map *H* and keypoint set *K*.

$$K, H = D(M) \tag{3}$$

The condition for a point *(i, j)* to be judged as a keypoint is that, the value at the coordinate *(i, j)* in *H* is bigger than the detection threshold α. Following [25], this work sets $\alpha = 0.5$. Considering that overly dense keypoints are detrimental to the performance of the network, a non-maximum suppression (NMS) is used to delete keypoints that are too close in *H*.

3.1.3 Descriptor

The keypoint descriptor *F* constructs the unit length descriptor *F* by L2 regularization of *M* in the channel dimension. Considering a keypoint *k* located at *(i, j)* in the keypoint map *H*, and its descriptor *f* can be constituted by the value located at *(i, j)* in the feature maps of each channel in *M*:

$$f_{i,j} = M_{i,j}, f_{i,j} \in \mathbb{R}^{224} \tag{4}$$

Descriptor *f* contains rich local shape details acquired from shallow features and deep features that are robust to noise caused by illumination and seasonal changes, so as to obtain keypoint descriptors with high discrimination and robustness.

3.1.4 Joint Loss

The total loss consists of two parts: the detection loss L_d and the description loss L_f. Considering two associated input images *I* and *I'*, the keypoint sets are *K* and *K'*. The pseudo-ground truth labels are *Y* and *Y'* respectively. Then, the keypoint detection can be regarded as a binary classification problem, on the predicted keypoint map, whether the point located at the pseudo-truth label point is judged as a keypoint. Obviously, the number of negative samples is much larger than the number of positive samples. Therefore, the loss of detector can be constructed by a biased cross-entropy loss function. The loss of keypoint detection of input image *I* can be described as follows:

$$L_d(K, Y) = \frac{1}{HW}\sum_{i,j}^{H,W} l_d(K_{ij}, Y_{ij}, \lambda) \tag{5}$$

$$l_d = -\lambda Y_{ij}\, log\big(K_{ij}\big) - \big(1 - Y_{ij}\big)\, log\big(1 - K_{ij}\big) \tag{6}$$

Where λ is to deal with the case that positive samples are far fewer than negative samples. Meanwhile, the loss of keypoint detection of the image *I'* is:

$$L_{d'}(K', Y') = \frac{1}{HW}\sum_{i,j}^{H,W} l_d(K'_{ij}, Y'_{ij}, \lambda) \tag{7}$$

For the paired keypoint set *K* and *K'*, the corresponding descriptors set are *S* and *S'*. Then descriptor loss L_f can be described by the ternary loss between *S* and *S'*:

$$L_f(S, S') = \frac{1}{2n}\sum_{i=1}^{n}\big(l_f\big) \tag{8}$$

Where：

$$l_f(S_i) = ma\,x\big(0, p(s_i) - n(s_i)\big),\ s_i \in S \tag{9}$$

Considering two points $k \in K$ and $k' \in K'$, The condition for these two points to be paired is:

$$\big||warp(k, H) - k'|\big| \le \varepsilon \tag{10}$$

For $s_i' \in S'$, $p(s_i)$ and $n(s_i)$ represent the positive sample distance between paired s_i and s_i', the negative sample distance between unpaired s_i and s_k' respectively, s_k' represents points within the image boundary, since the image matching is for image content that exists in both images:

$$p(d_i) = ||s_i - s_i'|| \tag{11}$$

$$n(d_i) = ||s_i - s_k'|| \tag{12}$$

Considering the keypoint detection loss, as well as the description loss, the total loss of the network is shown in Eq. (13):

$$L_b(K, K', Y, Y', S, S') = L_d(K, Y) + L_d(K', Y') + L_f(S, S') \tag{13}$$

3.3 Datasets and Implementation Details

At present, there is no free and open-source image dataset of high-speed rail train chassis for this task. The high-speed train chassis images matching dataset constructed in this paper includes 400 images at 1115×1000 resolution, among which the training set contains 300 images and another 100 images are used to make the test set. The images are captured by linear array cameras installed near the railroad tracks. In order to simulate the actual working environment of railway system as much as possible, this work adds random optical distortion to these 400 pictures, including random brightness, saturation, contrast change and gaussian noise to make the proposed training set and test set. These different photometric distortions are randomly combined to produce images with completely different styles, which will greatly enhance the robustness of the network in complex environments.

It is very difficult to obtain the real corresponding relationship between two high-speed train chassis images, and even the real keypoint coordinate labels for keypoint detection. Following [26], this work constructs a synthetic dataset. Each image contains several basic shapes, such as points, lines, triangles, quadrilaterals, and a corresponding annotation file. The annotation file records the coordinates of the corners and the endpoints, and the coordinate origin is the upper left corner of the image. These coordinates serve as labels for keypoint detector pre-training. Synthetic dataset image size: 1000*1000. We train the detector on the synthetic dataset and save the detector modle. Then, loading the detector model and inputting the images of the high-speed train chassis training set into the network, labelling each image of the training set, we then use the pseudo-ground truth labels and the high-speed train chassis training set image to retrain the keypoint detector. This training, then labeling, and retraining process is repeated several times until the final training loss no longer drops, thus completing the separate training of the keypoint detector. After this, this work uses the high-speed train image training set images and the latest pseudo-keypoint labels to conduct joint keypoint detection and description training. Using Adam optimizer, this work carries out joint training for 50k iterations. $l_r = 10 \times 10^{-3}$, *batch size = 4*.

As shown in **Fig. 3**, a method similar to HPatches [27] is used to produce the test set. Five random projection transformations are applied to each original image in the test set to generate 100 sequences, each of which contains one original image, five transformed images, and their transformation matrix H relative to the original image. At the same time, taking into account the complex working environment of the high-speed rail system, in order to test the robustness of the algorithm proposed in the case of large image appearance changes caused by illumination, season, weather, we randomly select 50 sequences from the 100 test sequences, add some kinds of photometric distortions to these 300 pictures to artificially simulate changes in the working environment of the railway system. Thus, the test set is divided into two parts

randomly and equally: one only contains random viewing angle changes, another has random perspective changes and optical distortions at the same time.

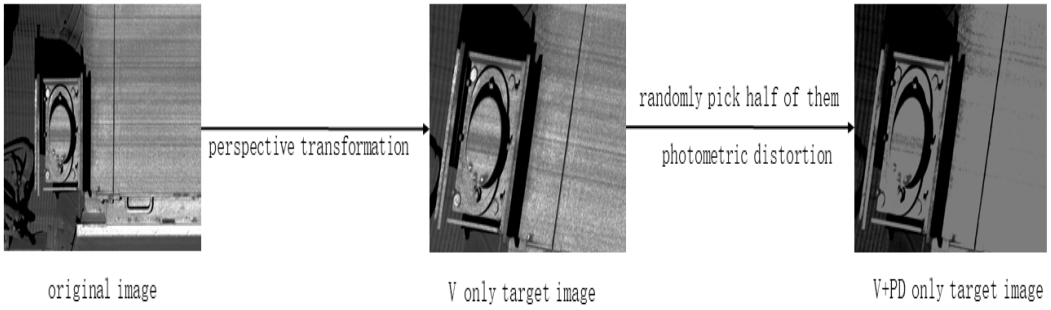


Fig. 3. Dataset generation. A random perspective transformation is performed for every given image resulting in the creation of target image candidate, then we random select half of them to perform random several gaussian noise, contrast, saturation, blur or brightness changes to artificially simulate changes in the working environments.

4. Experiments

This work considers a standard high-speed train image matching process. Inputting the previously collected image I_s and the current image I_f of the same train at the same chassis position into a computer. Then, it extracts keypoints and corresponding high-dimensional description vectors, matches these keypoints using certain method, e.g., deep learning method or nearest neighbor algorithm. Finally, we calculate the perspective transformation matrix H_{fs} between I_s and I_f by the way provided by OpenCV. Thus, we get the pixel-level correspondence, which means the current and the previous high-speed train chassis images matching is achieved. For the purpose of inspecting the localization accuracy of keypoints and the robustness to photometric distortions of the proposed method, the quantitative evaluation is compared with three key and popular metrics: keypoint repetition rate (REP), matching score (M.S.) and mean matching accuracy (MMA) [28]. Their explanations are shown in these following formulas respectively.

1. Mean Matching Accuracy (*MMA%*). *MMA* equals to the average of the ratio of the $N_{correct\ matches}$ to $N_{pre-matches}$ for N pairs of test images.

$$MMA = \frac{1}{N}\sum_{I_f,I_s}\frac{N_{correct\ matches}}{N_{pre-\ matches}} \tag{14}$$

Where I_f and I_s represent a pair of test images. N is the number of test image pairs. $N_{pre-matches}$ represents the number of keypoints that their descriptors are the closest to each other in a pair of test images. $N_{correct\ matches}$ represents the number of keypoint pairs that judged as pre-match whose projection error calculated according to Eq. (10) is less than the error threshold ε. Since the similarity of keypoint descriptors and the coordinate threshold are considered at the same time, the mean matching accuracy measures the comprehensive performance of the proposed network.

2. Keypoint Repetition Rate (*REP%*). *REP* equals to the average of the ratio of the $N_{possible\ matches}$ to $N(shared(I_f, I_s))$ for N pairs of test images.

$$REP = \frac{1}{N}\sum_{shared(I_f,I_s)}\frac{N_{possible\ matches}}{N(shared(I_f,I_s))} \tag{15}$$

Where I_f and I_s represent a pair of test images. $shared(I_f, I_s)$ represents the area where both images are visible in a pair of test images. N is the number of test image pairs. $N_{possible\ matches}$ is the number of pair of keypoints that the projection error calculated according to Eq. (10) is less than the error threshold ε. $N(shared(I_f, I_s))$ represents the number of keypoints that lie within the area visible in both figures. Since the coordinate threshold are considered here, the *REP* measures the performance of the keypoint detector.

3. Matching Score (*M.S.%*): *M.S.* equals to the average of the ratio of the $N_{pre-\ matches}$ to $N(shared(I_f, I_s))$ for N pairs of test images.

$$M.S. = \frac{1}{N}\sum_{shared(I_f,I_s)} \frac{N_{correct\ matches}}{N(shared(I_f,I_s))} \tag{16}$$

Where I_f and I_s represent a pair of test images. $shared(I_f, I_s)$ represents the area where both images are visible in a pair of test images. N is the number of test image pairs. $N_{correct\ matches}$ represents the number of pair of keypoints that their descriptors are the closest to each other.

This work comprehensively compares the proposed network with the traditional hand-designed feature extraction algorithms SIFT, ORB and SURF on the proposed test set. All of them are implemented using default settings in OpenCV, and the results are shown in **Fig. 4**. The proposed approach outperforms traditional methods on all metrics. In terms of *MMA* and *M.S.*, which simultaneously measure keypoint detection and description, the proposed method has obvious advantages compared with the other three methods. Specifically, for *REP*, which only measures keypoint detection, the proposed method does not improve much compared to ORB, but this is because there are not enough keypoints for ORB detection.

It is also worth noting that when both random perspective changes and illumination distortions are included, the proposed network has a greater advantage than when only random perspective changes are included, indicating that it can well deal with the task of matching high-speed rail train chassis images.

The visual comparison between the proposed network and other methods is shown in **Fig. 5**. The green lines in the figure represent the correct keypoint matching pairs, and the blue points represent the unpaired keypoints. The first row is the case that only includes the perspective transformation. Both the proposed method and SIFT, SURF can find enough keypoint pairs, while ORB extracts fewer keypoint pairs. For the second and third row, this work adds random optical distortion on the basis of the first row. The proposed network is the only method that can find dense keypoint pairs in this case, while ORB has obvious mismatches. The last two rows represent the effects in two challenging scenarios. The fourth row is low texture, and we add some noise on the left image, the keypoints extracted by ORB are still very few, while SIFT and SURF extract more keypoints in the left image than in the right image, especially SURF, but there are few keypoint pairs for successful matching. Due to the good design of the proposed network, the number of keypoints in the left figure has not increased. This also explains why the proposed method is significantly higher than the other two methods in the metric of *REP* in **Fig. 4**. The fifth row represents the extreme optical distortion. In this case, the other three methods have a large number of mismatches, while the proposed network is higher than the other three methods in terms of the number of extracted keypoints and the number of correct matches. We can draw the conclusion that, the proposed method is the only one that finds enough correct matches both at the low texture and extreme optical distortion.

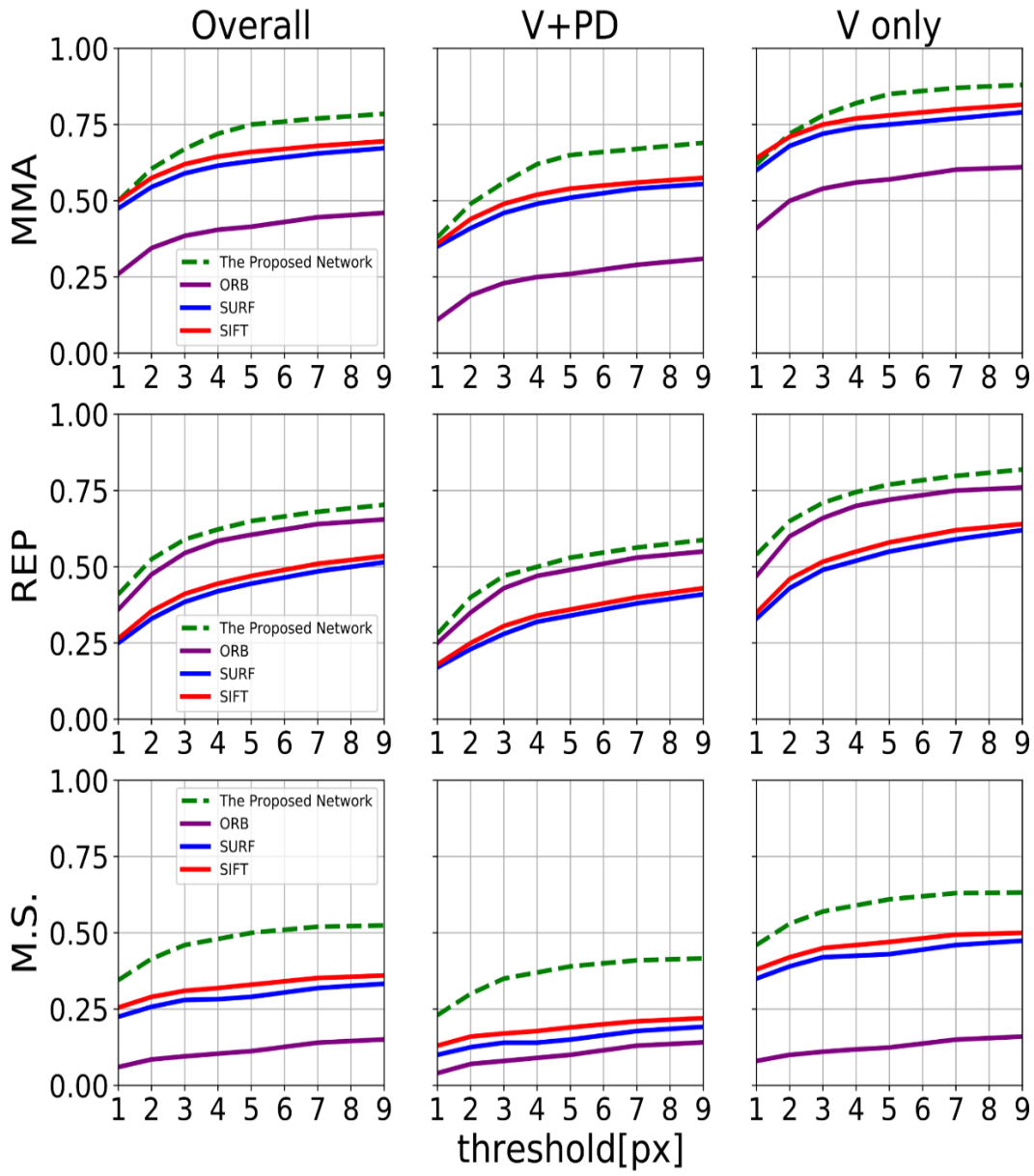


Fig. 4. The comparison between the proposed network and the three methods of SIFT, ORB and SURF with *MMA*, *REP* and *M.S.* on the test set. V+PD indicates the situation with both optical distortion and random perspective changes. V only indicates the situation with random perspective changes. Overall indicates the average error value in these two situations. The proposed method has the best performance on all metrics.

Based on the above analysis, we can conclude that traditional feature extraction methods have good results when facing image matching tasks that only include changes in perspective changes. But when the images to be matched have both the perspective changes and the image appearance and style changes caused by the illumination or weather changes, these methods are somewhat stretched, while the method proposed in this paper can well meet the complex environmental changes faced by the high-speed rail system.

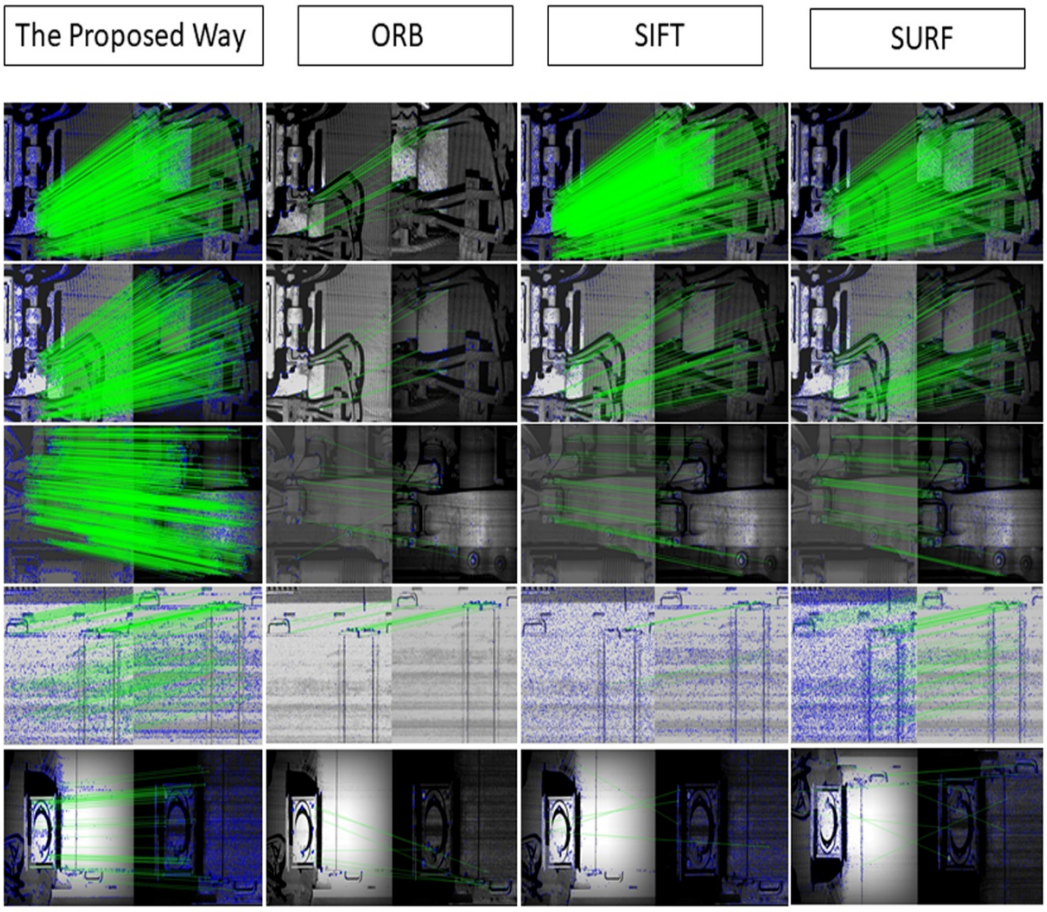


Fig. 5. Visual comparison results of the proposed network with ORB, SIFT and SURF, the green lines indicate the correct keypoint matching pairs, and the blue points indicate the unpaired keypoints. The proposed method is significantly better than the other three methods in terms of both the number of matching pairs and robustness.

5. Conclusion

Image keypoint detection and description is a popular method for image matching. Finding pixel level correspondence precisely in complex work environment of the high-speed rail system is a big challenge, the quality of image registration of high-speed train chassis images has an important impact on the subsequent process such as defect detection and the safety of trains. The current matching methods require a lot of manpower and material resources, but the generalization of complex scenes such as weather, light, and seasonal changes is difficult to meet the actual needs. In this paper, a high-speed rail train chassis image matching dataset is established, including both optical distortions such as illumination and perspective changes to simulate the actual high-speed rail system working environment as much as possible. The high-speed train chassis matching method designed in this paper has achieved good results on the dataset. Detailed experiments show the huge improvement of the proposed network over traditional methods. By robustly extracting dense keypoints, it can be applied to the actual scenes of high-speed train safety detection, improving the matching effect of high-speed train chassis images, thus has a wide range of application prospects.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61871058, WYF, http://www.nsfc.gov.cn/).

References

[1] G. X. Xie, “Research and development of image-based train component integrity detection method and system,” M.S. dissertation, Chang'an University, China, Xi'an, 2019. Article (CrossRef Link)
[2] Z. J. Zhang, “Research on application of dynamic image detection system for EMU vehicle faults (TEDS),” *Railway Locomotive &Car*, vol. 34, no. 4, pp. 82-84, 2014. Article (CrossRef Link)
[3] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proc. of the 7th International Joint Conference on Artificial Intelligence Morgan Kaufmann Publishers Inc*., 674-679, 1981. Article (CrossRef Link)
[4] S. F. Lu and Z. L, “Dynamic image comparison and analysis method for operation faults of EMU,” *Laser & Optoelectronics Progress*, vol. 54, no. 9, pp. 301-307, 2017.
[5] D. Peng, “Anomaly detection algorithm for the bottom parts of high-speed trains based on the SURF feature of rail-side images,” M.S. dissertation, Beijing Jiaotong University, China, Beijing, 2016. Article (CrossRef Link)
[6] E. Rosten and T. Drummond, “Fusing point sand lines for high performance tracking,” in *Proc. of the IEEE International Conference on Computer Vision (ICCV 2005)*, vol. 1, no. 2, pp. 1508–1515, 2005. Article (CrossRef Link)
[7] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005. Article (CrossRef Link)
[8] D. G. Lowe, “Distinctive image features from scale invariant keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004. Article (CrossRef Link)
[9] B. Herbert, T. Tuytelaars and L. V. Gool, “SURF: Speeded up robust features,” in *Proc. of the European Conference on Computer Vision (ECCV 2006)*, pp. 404–417, 2006. Article (CrossRef Link)
[10] Karami E, Prasad S and Shehata M, “Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images,” *arXiv e-prints, arXiv: 1710.02726*, 2017. Article (CrossRef Link)
[11] Rublee. E, “ORB: an efficient alternative to SIFT or SURF,” in *Proc. of IEEE International Conference on Computer Vision*, Barcelona, Spain, pp. 2564-2571, 2011. Article (CrossRef Link)
[12] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Readings in Computer Vision*, pp. 726-740, 1987. Article (CrossRef Link)
[13] E. Rosten and T. Drummond, “Fusing point sand lines for high performance tracking,” in *Proc. of the IEEE International Conference on Computer Vision (ICCV 2005)*, vol. 1, no. 2, pp. 1508–1515, 2005. Article (CrossRef Link)
[14] B, Liu, “Research and thoughts on the application of the image detection system for operational faults of EMUs (TEDS),” *China Railway*, vol. 12, pp. 61-65, 2017.
[15] R. Hartley and A. Zisserman, “Multiple view geometry in computer vision,” *Cambridge University Press*, 2003. Article (CrossRef Link)
[16] Z. G. Qu, Y. M. Huang and M. Zheng, “A novel coherence-based quantum steganalysis protocol,” *Quantum Information Processing*, vol. 19, no. 362, pp. 1-19, 2020. Article (CrossRef Link)
[17] S. Sun, J. Zhou, J. Wen, Y. Wei and X. Wang, "A dqn-based cache strategy for mobile edge networks," *Computers, Materials & Continua*, vol. 71, no.2, pp. 3277–3291, 2022. Article (CrossRef Link)

[18] Y. F. Wei, F. Richard Yu, M. Song and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no.1, pp. 680-692, Jan. 2018. Article (CrossRef Link)

[19] X. Z. Liu, L. Luo and Y. Zhang, "High-speed train image registration algorithm based on deep learning," *Information Technology*, vol. 45, no. 7, pp. 26-30, 2021. Article (CrossRef Link)

[20] L. Xiangchun, C. Zhan, S. Wei, L. Fenglei and Y. Yanxing, "Data matching of solar images super-resolution based on deep learning," *Computers, Materials & Continua*, vol. 68, no.3, pp. 4017–4029, 2021. Article (CrossRef Link)

[21] Z. G. Qu, H. R. Sun and M. Zheng, "An efficient quantum image steganography protocol based on improved EMD algorithm," *Quantum Information Processing*, vol. 20, no. 53, pp. 1-29, 2021. Article (CrossRef Link)

[22] D. Detone, T. Malisiewicz and A. Rabinovich, "Deep image homography estimation," *arXiv preprint, arXiv: 1606.03798*, 2016. Article (CrossRef Link)

[23] J. Zhang, C. Wang and S. Liu, "Content-aware unsupervised deep homography estimation," in *Proc. of European Conference on Computer Vision (ECCV 2020)*, pp. 653–669, 2020. Article (CrossRef Link)

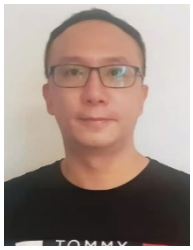
[24] M. Dusmanu, I. Rocco and T. Pajdla, "D2-Net: A trainable CNN for joint detection and description of local features," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 8092-8101, 2019. Article (CrossRef Link)

[25] Z. Luo, L. Zhou, X. Bai, H. Chen and J. Zhang, "ASLFeat: Learning local features of accurate shape and localization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pp. 6589-6589, 2020. Article (CrossRef Link)

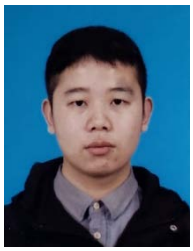
[26] D. Detone, T. Malisiewicz and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 224-236, 2018. Article (CrossRef Link)

[27] V. Balntas, K. Lenc, A. Vedaldi and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pp. 5173–5182, 2017. Article (CrossRef Link)

[28] S. Mutic, J. F. Dempsey and W. R. Bosch, "Multimodality image registration quality assurance for conformal three-dimensional treatment planning," *International Journal of Radiation Oncology, Biology, Physics*, vol.51, no.1, pp. 255-260, 2001. Article (CrossRef Link)



Jun Li received the bachelor's degree and master's degree in electronic engineering from Beijing University of Posts and Telecommunications in 2004 and 2014, respectively. He is currently studying for a doctor's degree in electronic engineering from Beijing University of Posts and telecommunications. He was involved in several innovation projects in his work, and developed a number of rail transit visual detection systems. He has been authorized more than 80 patents, and won the second prize of railway science, and the second prize of Beijing science and technology. His current research interests include photoelectric detection systems, high-speed 3D reconstruction, multidimensional data fusion, and deep learning.



Xiang Li received the B.S. degree in electronic science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2019. He is currently working toward the M.S. degree in electronic information engineering at Beijing University of Posts and Telecommunications (BUPT, China). His current research interests focus on deep learning and computer vision.

Yifei Wei received the B.Sc. and Ph.D. degrees in electronic engineering from Beijing University of Posts and Telecommunications (BUPT, China), in 2004 and 2009, respectively. He was a visiting Ph.D. student in Carleton University (Canada) from 2008 to 2009. He was a postdoctoral research fellow in the Dublin City University (Ireland) in 2013. He was the vice dean of school of science in BUPT from 2014 to 2016. He was a visiting scholar in the University of Houston (USA) from 2016 to 2017. He is currently a Professor in school of electronic engineering at BUPT. His current research interests are in energy-efficient networking, machine learning and deep reinforcement learning.

Xiaojun Wang received the B.Eng. degree in Computer and Communications and the M.Eng. degree in Computer Applications from Beijing University of Posts and Telecommunications (BUPT), China, in 1984 and 1987 respectively. He received the Ph.D. degree in Electronic Engineering from Staffordshire University (then Staffordshire Polytechnic), England, U.K., in 1993. He joined the School of Electronic Engineering, Dublin City University, Ireland, as an Assistant Lecturer in 1992, where he is currently a Senior Lecturer. His current research interests include energy-efficient networking, network security and hardware acceleration of cryptographic algorithms.