Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 7836169, 16 pages
https://doi.org/10.1155/2018/7836169

*Research Article*

# Novel Registration and Fusion Algorithm for Multimodal Railway Images with Different Field of Views

**Baoqing Guo**,[1,2,3] **Xingfang Zhou**,[2] **Yingzi Lin**,[3] **Liqiang Zhu**,[1,2] **and Zujun Yu**[1,2]

[1]*School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China*
[2]*Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology (Beijing Jiaotong University), Ministry of Education, Beijing 100044, China*
[3]*Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115, USA*

Correspondence should be addressed to Baoqing Guo; bqguo@bjtu.edu.cn

Received 16 May 2018; Revised 30 July 2018; Accepted 30 August 2018; Published 6 December 2018

Academic Editor: Krzysztof Okarma



Objects intruding high-speed railway clearance do great threat to running trains. In order to improve accuracy of railway intrusion detection, an automatic multimodal registration and fusion algorithm for infrared and visible images with different field of views is presented. The ratio of the nearest to next nearest distance, geometric, similar triangle, and RANSAC constraints are used to refine the matching SURF feature points successively. Correct matching points are accumulated with multiframe to overcome the insufficient matching points in single image pair. After being registered, an improved Contourlet transform fusion algorithm combined with total variation and local region energy is proposed. Inverse Contourlet transform to low frequency subband coefficient fused with total variation model and high frequency subband coefficients fused with local region energy is used to reconstruct the fused image. The comparison to other 4 popular fusion methods shows that our algorithm has the best comprehensive performance for multimodal railway image fusion.

## 1. Introduction

High-speed railway is a competitive transportation all over the world. As of 2017, the length of commercial high-speed railway lines in China had exceeded 25,000 km. With the increasing of train speed, safety of railway operation has attracted more and more attention. Any object intruding railway clearance will pose a major threat to running trains. Effective intrusion detection methods are needed to prevent accidents.

Image analysis is an efficient and effective non-contact intrusion detection method widely used in target detecting and tracking [1–3]. Most of them use visible images to analyze intrusion activities. Visible images get the detailed information of colors and texture under good illumination. But their qualities are poor under low illumination and disturbance light circumstance. In contrast, infrared (IR) images are sensitive to thermal radiation of objects and tolerant to changing illumination and disturbance light. However, IR image typically has lower spatial resolution and fewer details. Image fusion is widely used in remote sensing [4], military [5, 6], objects tracking and detecting [7, 8], etc. IR and visible image fusion will also bring great benefit to railway clearance intrusion detection. A typical image fusion process is usually divided into two steps: image registration and image fusion. Image registration is the basis of fusion. The quality of fusion is bounded not only by the quality of fusion algorithm, but also by the outcome of prior registration algorithm. Due to this dependency, images are always assumed to be pre-aligned [9]. However, in fact, visible and IR images acquired separately are not pre-aligned and even have different field of views and depth. These bring great difficulties to image fusion. We need both excellent registration and fusion algorithms in railway applications.

Image registration is the process of precisely overlaying two images with the same area through geometrically aligning common features identified in image pairs. It is usually divided into four steps: (1) feature detection and extraction, (2) feature matching, (3) transformation function fitting, and (4) image transformation and resample [10]. For multimodal

images, the first two steps are more difficult challenges since different modalities do not exhibit the same features. Z. Xiong [11] has provided a broad overview of registration methods based on feature extraction and matching. S. A. Raza [12] proposed a registration algorithm for thermal and visible diseased plants images based on silhouette extraction. J. Ma [13] presented a non-rigid IR and visible face images registration method by aligning edge maps with a robust regularized Gaussian fields criterion. G. A. Bilodeau [14] proposed a RANSAC-Based registration method using a novel function of moving objects' trajectories. These researches focused on simple scenes registration where the interested objects occupied most part of source images. For large scene images, M. Gong [15] and J. Woo [16] used mutual information for registration. M. Ghantous [9] presented an area-based registration and object-based fusion combined approach based on Dual-Tree Complex Wavelet Transform to reduce complexity. Z. Song [10] proposed an effective registration approach based on histogram of triangle area representation sample consensus for remote sensing images. J. Han [17] proposed a registration method combined line-based global transform approximation with point-based local transform adaptation. All above methods are registration for multimodal images with the same field of views. However, these methods are not available because railway IR and visible images usually have different field of views and depth.

Image fusion is a technique of combining or integrating complementary information from source images into the fused image. Fusion methods based on saliency detection [18, 19], guided image filters [7], and total variation model [20, 21] play dramatic effect in multimodal image fusion. Recently, multiscale decomposition fusion schemes have been a popular research area. Pyramid, wavelet, and Contourlet based decomposition are included in the schemes. Contourlet transform was proposed based on wavelet transform in 2002. Besides attributes of wavelets, it also has advantages of multiresolution, multidirection, and anisotropy. It can decompose images in arbitrary direction at any scale and is good at describing contours and directional texture of images [22]. Many researches focused on Contourlet transform combined with sparse representation [23, 24], low-level visual features [25], object region detection [26], and other methods [27, 28]. In order to solve the problem of unclear edge and textures, we proposed an improved Contourlet transform method combined with total variation (TV) model and local region energy.

In this paper, we present a novel registration and an improved Contourlet-based fusion algorithm for railway IR and visible images with different field of views. The initial candidate matching points are firstly generated with the ratio of the nearest to next nearest distance of SURF feature points. Then they are refined with consecutive geometric constraint, similar triangle, and RANSAC constraints. In order to overcome the shortcoming of insufficient matching points in single frame pair, matching points are accumulated to calculate the transformation matrix from multiframe sequence. After registration, an improved Contourlet transform combined with total variation and local region energy is proposed for image fusion. Qualitative and quantitative comparison of proposed and other state-of-the-art fusion methods is discussed for fusion effect evaluation. The contributions of this paper are as follows:

(1) A novel multimodal image registration method is presented based on SURF feature points refining with ratio of the nearest to next nearest distance, geometric, similar triangle, and RANSAC constraints consecutively.

(2) An improved multimodal Contourlet transform fusion algorithm combined with total variation model and local region energy is proposed. The total variation model is used for low frequency subband coefficients fusing. The local region energy is used for high frequency subband coefficients fusing.

(3) Qualitative and quantitative comparison of fusion effect on railway IR and visible images is presented to demonstrate the validity of our algorithm.

The remainder of this paper is organized as follows. A novel multimodal image registration method based on SURF feature points matching with a series of refining procedures is outlined in Section 2. Section 3 presents an improved Contourlet transform image fusion algorithm combined with total variation model and local region energy. The fusion evaluation of proposed and other state-of-the-art algorithms is discussed in Section 4. The conclusion is in Section 5.

## 2. Novel Image Registration with SURF Refining

*2.1. Railway Images Registration Overview.* There are some differences between familiar security monitoring and railway clearance intruding detection. In security monitoring application, visible and IR images always have the same field of views and depth. However, railway scene is a long and narrow field. Since the view of one camera is limited, visible cameras are arranged at an interval of 100 meters along rails to obtain seamless overlay images. Because of the much higher cost of IR camera, we usually use IR camera with shorter focal length to monitor a larger scene shown in Figure 1(a). The visible image in Figure 1(b) is a small part of IR image shown as the red rectangular in Figure 1(a). In IR image, the rail and person are clearer than other objects because of their higher thermal radiation. In visible image, the person and laying pole are very clear and easy to identify because of the great illumination at daytime. It contains more details such as color, texture, and edges than IR image. But it is not good enough at night. Image fusion will provide more information than each alone. However, in railway application, the typical multimodal fusion methods cannot be used directly due to their different field of views and depth.

Image registration is the basis of fusion. Our novel multimodal registration process for IR and visible images includes 3 steps:

(1) Feature extraction and description of IR and visible images using SURF descriptor: the initial matching point pairs are generated with ratio of the nearest to next nearest neighbor distance.


(a) IR image of large scene (b) Visible image of small scene

Figure 1: Railway IR and visible images.

(2) Matching point pairs refining with successively geometric constraints: similar triangle and RANSAC constraints in single frame pair.

(3) Transformation matrix calculation with accumulative matching point pairs obtained from multiframe sequences: this can overcome the shortcoming of insufficient matching points and matching deviation introduced by concentrated matching points in single image pair.

*2.2. Feature Extraction and Comparison.* Feature extraction is the first step of image matching and registration. In order to highlight the object, the source IR image is first negated. The commonly used feature descriptors can be divided into two types: histogram-based descriptors and binary descriptors. The *Scale Invariant Feature Transform* (SIFT), *Maximally Stable Extremal Region* (MSER), and *Speed Up Robust Features* (SURF) are typical histogram-based descriptors. Instead of expensive gradients operations, binary descriptors make use of simple pixel comparison, which result in binary strings of typically short length (commonly 256 bits) and lead to a major improvement in both running time and memory footprint [29]. The common binary descriptors include *Oriented Fast and Rotated BRIEF* (ORB), *Binary Robust Invariant Scalable Keypoints* (BRISK), and *Accelerated KAZE* (AKAZE). Each descriptor has its advantages and disadvantages. We will give a comparison of their results and select the best one as feature extraction method.

*2.2.1. Histogram-Based Descriptors. (a) SIFT* is the most renowned feature detection description method introduced by D.G. Lowe in 2004 [30]. It is based on Difference-of-Gaussians (DoG) operator which is an approximation of Laplacian-of-Gaussian (LoG). Feature points are detected by searching local maxim value using DoG at various scales of the subject images. SIFT is robustly invariant to image rotations, scale, and limited affine variations but its main drawback is high computational cost. Equation (1) shows the convolution of difference of two Gaussians (computed at different scales) with image "$I(x, y)$":

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \quad (1)$$

where "$G$" is the Gaussian function and $\sigma$ is the scale.

*(b) SURF* also relies on Gaussian scale space analysis of images. It is based on *determinant of Hessian Matrix* and exploits integral images to improve feature-detection speed [31]. SURF features are also invariant to rotation and scale but they have little affine invariance. The main advantage of SURF over SIFT is low computational cost. Equation (2) represents the Hessian Matrix in point "$\boldsymbol{x} = (x, y)$" at scale "$\sigma$":

$$H(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix} \quad (2)$$

where $L_{xx}(x, \sigma)$ is the convolution of Gaussian second-order derivatives with the image in point $\boldsymbol{x}$ and $L_{xy}$ and $L_{yy}$ are similar to $L_{xx}$.

When nonmaximal suppression is achieved in the $3 \times 3 \times 3$ stereo neighborhood detected by the Hessian Matrix, only the points larger than all 26 nearest neighbor's responses are identified as feature points and then interpolated in the scale space to obtain a stable feature point location and scale.

*(c) MSER* generally attempts to detect a set of connected regions from an image. It is defined as

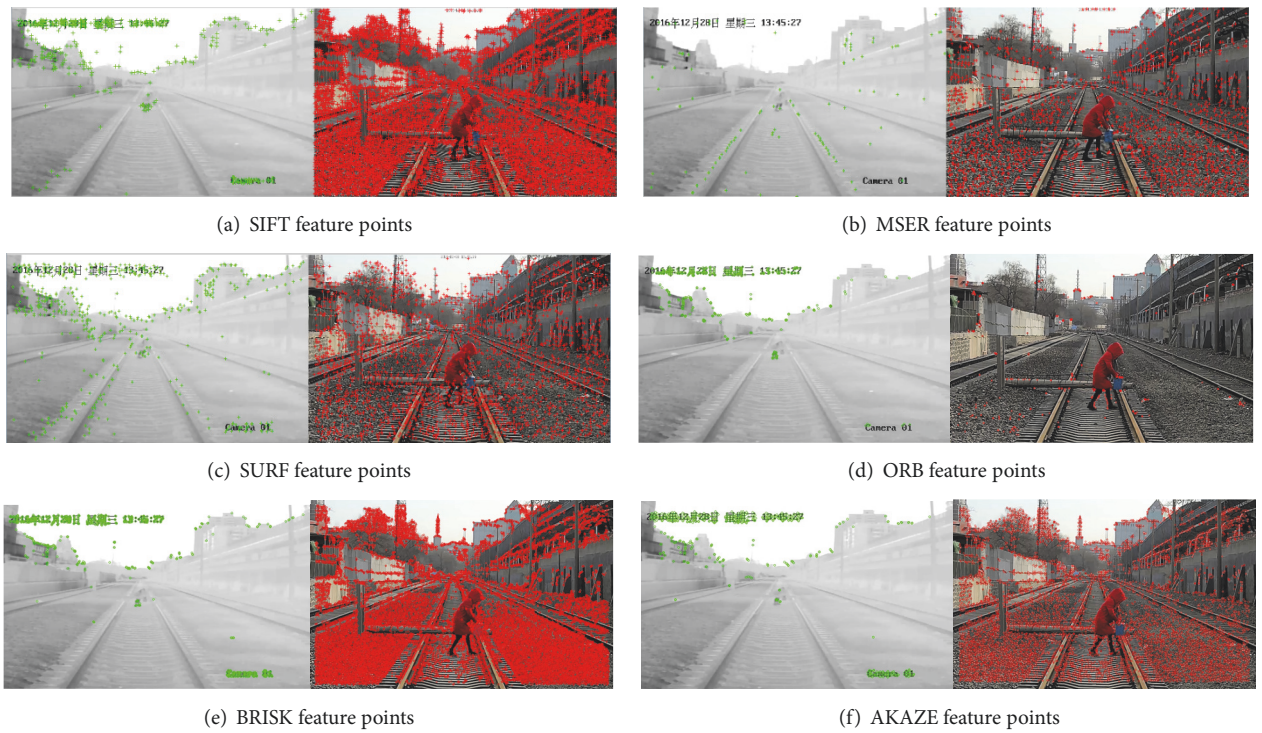
$$g(t) = \frac{(d/dt)\,|Q(t)|}{|Q(t)|} \quad (3)$$

where |Q(t)| indicates the region area. MSER is computationally inexpensive and can be calculated in linear time. It was introduced by extremal to perform scene recognition under arbitrary viewpoints and illumination conditions. MSER performs very well when detected on flat surfaces and changing illumination. The only weakness for MSER is that it is sensitive to image blur. It can perform better in colorful visible image than in blurring IR image.

*2.2.2. Binary Descriptors.* (a) *ORB descriptor* was proposed by Reblee et al. [32]. It builds upon BRIEF by adding a rotation invariance mechanism that is based on measuring the patch orientation using first-order moments. BRIEF compares the same set of smoothed pixel pairs for each local patch that it describes. For each pair, if the first smoothed pixel intensity is larger than that of the second, BRIEF writes 1 in the final descriptor string and 0 otherwise. ORB also uses unsupervised learning to learn the sampling pairs.

(b) *BRISK* descriptor was proposed by Leutenegger in 2011 [33]. It uses a hand-crafted sampling pattern that is composed of set of concentric rings. BRISK uses the long-distance sampling pairs to estimate the patch orientation and the short distance sampling pairs to construct the descriptor itself through pixel intensity comparisons.

(c) *AKAZE* uses the Accelerated-KAZE detector estimation of orientation for rotating the LDB grid to achieve



(a) SIFT feature points

(b) MSER feature points

(c) SURF feature points

(d) ORB feature points

(e) BRISK feature points

(f) AKAZE feature points

Figure 2: Comparison of different descriptors.

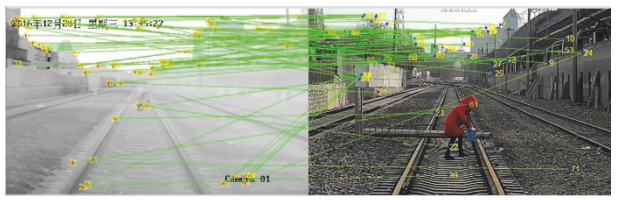


Figure 3: Initial candidate matching points.

rotation invariance [34]. In addition, AKAZE also uses the A-KAZE detector's estimation of patch scale to subsample the grid in steps that are a function of the patch scale.

*2.2.3. Comparison of Different Descriptors.* Figure 2 shows the results of SIFT, MSER, SURF, ORB, BRISK, and AKAZE descriptors. In general, binary descriptors are faster than histogram-based descriptors. ORB obtains the fewest feature points in both visible and IR images. For visible image, SIFT and BRISK obtain much more feature points than MSER, SURF, and AKAZE. For IR image, SIFT and SURF obtain much more feature points than other descriptors. But SURF feature points are more evenly distributed in whole IR image than all other methods. That is conducive to the following registration. We have tried to change the parameters of these detectors but found that parameters made little contribution to the improvement of uneven distribution. So we choose SURF as feature extractors for subsequent matching process.

*2.3. Initial Candidate Matching Point Pairs Generation.* After SURF features extraction, Euclidean distance is used to measure the similarity of feature points in two images. For a feature point in IR image, it looks for two nearest points measured with Euclidean distance in visible image. If the ratio of the neatest distance $d_{ND}$ to next nearest distance $d_{NND}$ is less than threshold $\varepsilon$ in formula (4), the points with the nearest distance are considered as a candidate matching point pair.

$$\frac{d_{ND}}{d_{NND}} < \varepsilon \tag{4}$$

where $\varepsilon$ controls the number of matching points. The number will decrease as $\varepsilon$ decreases. After massive experiments, the default value is 0.8. The initial candidate matching points in IR and visible images are shown in Figure 3. There are many mismatching points that need to be refined.

*2.4. Candidate Matching Points Refining*

*2.4.1. Geometric Constraints Refining.* The centers of IR and visible image can be considered approximately identical when the images are obtained at a long distance. IR image

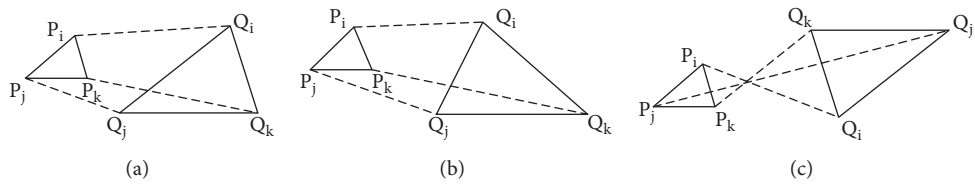



Figure 4: Similar triangle constraints.

center is noted as $(x_{or}, y_{or})$, visible image center is noted as $(x_{ov}, y_{ov})$, and any corresponding matching points $(x_{ir}, y_{ir})$ and $(x_{iv}, y_{iv})$ should obey the following geometric constraints:

(1) The dip angles of corresponding matching points to their own image centers are approximately equal:

$$\left|\arctan\left(\frac{(x_{ir}-x_{or})}{(y_{ir}-y_{or})}\right)-\arctan\left(\frac{(x_{iv}-x_{ov})}{(y_{iv}-y_{ov})}\right)\right|<T \quad (5)$$

where $T$ is a threshold for dip angle difference; the matching points should be deleted when the dip difference is larger than $T$.

(2) The corresponding matching points should be within the same quadrants in both IR and visible images coordinate system. In other words, the vertical and horizontal coordinate differences of corresponding matching points and their respective image centers should have the same sign shown as formula (6):

$$\begin{aligned}(x_{or}-x_{ir})*(x_{ov}-x_{iv})>0\\(y_{or}-y_{ir})*(y_{ov}-y_{iv})>0\end{aligned} \quad (6)$$

(3) Since the IR scene is larger than visible scene, the pixel distance in IR image should be less than that in visible image:

$$\begin{aligned}&\sqrt{(x_{or}-x_{ir})^2+(y_{or}-y_{ir})^2}\\&<\sqrt{(x_{ov}-x_{iv})^2+(y_{ov}-y_{iv})^2}\end{aligned} \quad (7)$$

The candidate matching points should be deleted if any constraints above are not satisfied.

*2.4.2. Similar Triangle Constraint Refining.* After geometric constraints refining, some mismatching points are deleted. But there are still "many-to-one" and crossing mismatching points. We will refine them further with similar triangle constraint.

Assume that P and Q are matching point sets after geometric constraint. Any 3 feature points (noncollinear) in IR image can form a triangle $\Delta P_i P_j P_k$, $(i < j < k, P_i, P_j, P_k \in P)$. The corresponding triangle in visible image is $\Delta Q_i Q_j Q_k$ $(i < j < k, Q_i, Q_j, Q_k \in Q)$. These two triangles should be similar, shown in Figure 4(a).

According to the characters of similar triangles, there exists

$$\frac{P_iP_j}{Q_iQ_j}=\frac{P_jP_k}{Q_jQ_k}=\frac{P_kP_i}{Q_kQ_i} \quad (8)$$

We define $dd_1$ and $dd_2$ as

$$\begin{aligned}dd_1&=\frac{|P_iP_j|/|Q_iQ_j|}{|P_jP_k|/|Q_jQ_k|}\\dd_2&=\frac{|P_iP_k|/|Q_iQ_k|}{|P_jP_k|/|Q_jQ_k|}\end{aligned} \quad (9)$$

From formula (8), we know that $dd_1$ and $dd_2$ in formula (9) should equal 1. But for noise interruption, they should approximately equal 1:

$$\begin{aligned}|dd_1-1|<T_0\\|dd_2-1|<T_0\end{aligned} \quad (10)$$

$T_0$ is a threshold near to zero. In this paper, $T_0$ is 0.1.

In Figure 4(b), $|P_iP_j|/|Q_iQ_j| > |P_iP_k|/|Q_iQ_k|$, $dd_1 > 1$, and $dd_2 < 1$; they are not similar triangles even if their side lengths satisfy formula (10). In this situation, we should firstly sort their side lengths; only the similar triangle with the same sort can be kept. After that, there will also be another situation inversed as Figure 4(c). It can be removed with formula (11):

$$\left|\frac{\overrightarrow{P_iP_j}}{|P_iP_j|}-\frac{\overrightarrow{Q_iQ_j}}{|Q_iQ_j|}\right|<T_1 \quad (11)$$

Formula (11) means that unit vector of any pair of vertices (such as $\overrightarrow{P_iP_j}$ and $\overrightarrow{Q_iQ_j}$) in similar triangle should be approximated. $T_1$ is also a threshold near zero. In this paper, $T_1$ is 0.1.

In order to improve the matching accuracy, a traversal similar triangle test method is adopted. For the matching point set P and Q with $n$ pairs of matching points, we randomly select 3 matching points to consist $C_n^3$ pairs of triangles. Then, we introduce an accumulator for each matching point. When any pairs of triangle satisfy the above similar triangle constraints, the accumulator increases by 1. After traversing all these triangles, the larger the accumulator is, the higher the probability of correct matching is for this point pair. When the accumulator value is larger than threshold $T_2$, the corresponding point pair is considered as a correct matching point pair. The point pairs corresponding to the first k largest accumulators are considered as k correct candidate matching point pairs.

*2.4.3. Fine Matching with RANSAC.* After refining with geometric and similar triangle constraints, the reserved matching

Table 1: Comparison of different matching algorithms.

| Method | feature point numbers in IR/Visible image | Candidate matching points | Fine matching points | Correct points |
|---|---|---|---|---|
| SIFT+RANSAC | 664/11,433 | 664 | 6 | 0 |
| SURF+RANSAC | 523/2,766 | 523 | 4 | 0 |
| Our method | 553/2,766 | 72 | 3 | 3 |

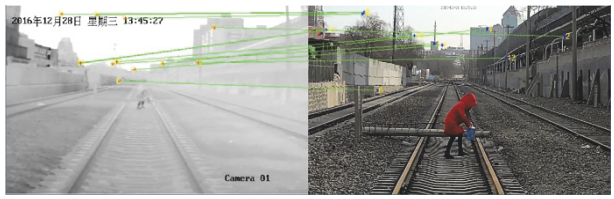



Figure 5: Refining result of geometric constraint.


Figure 6: Refining result of similar triangle constraint.


Figure 7: Result of RANSAC fine matching.

points are almost correct except several separate points. In order to improve the correct matching further, RANSAC fine matching algorithm is finally used for the prior reserved matching points.

The refining process is shown in Figures 5–7. After refining with geometric constraints there are 12 pairs of matching points in Figure 5. After similar triangle refining, 5 pairs are left in Figure 6. After RANSAC matching, only 3 pairs of correct matching points are reserved in Figure 7.

After consecutive procedure of geometric constraints, similar triangle, and RANSAC matching, the reserved point pairs are all correctly matched. In order to evaluate the effect of proposed algorithm, a comparison to SIFT+RANSAC and SURF+RANSAC is shown in Table 1. It is shown that the traditional SIFT+RANSAC and SURF+RANSAC matching methods are not applicable to multimodal images. Our proposed method with successive geometric, similar triangle, and RANSAC constraints can obtain the correct matching point pairs in IR and visible images.

*2.5. Transformation Matrix Calculation.* In our application, both IR and visible cameras are equipped with microdistortion lens. The distortion is so small that it can be ignored in image registration process. The IR and visible cameras have approximate parallel optical axis and follow the perspective projection model. The coordinates between two images are shown as formula:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} m_0 & m_1 & m_2 \\ m_3 & m_4 & m_5 \\ m_6 & m_7 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = M \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{12}$$

where $(x, y), (x', y')$ are coordinates of corresponding matching points in IR and visible images. $M$ is transformation matrix between the two images. Eight parameters require at least 4 pairs of matching points.

As shown in Figure 7, we cannot ensure that there are more than 4 pairs of correct matching points in a single image pair with the prior matching procedures. If the number of matching point pairs is less than 4, it is insufficient to calculate the 8 parameters in matrix $M$. On the other side, even though we got 4 pairs in one single image pair, the transformation will have deviation on another position if the matching pairs are too concentrated. It will still lead to bias if we use this matrix for registration.

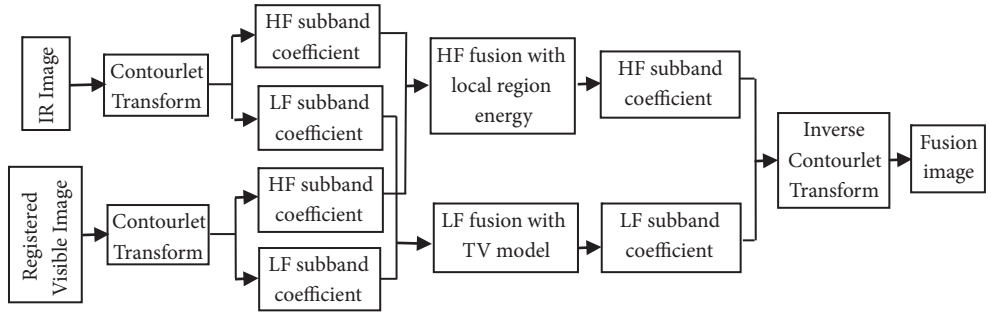



Figure 8: Flow chart of image fusion algorithm.

In order to improve the precision of transformation matrix, we create an accumulative set $Tpts$ from respective matching points sets $pts(i)$ of multiframe images where objects are in different position:

$$Tpts = \{pts(1), \ldots, pts(i), \ldots, pts(n)\} \quad (13)$$

where $n$ is the number of selected frames, $pts(i)$ is the matching points set in the $i$th frame image pairs, and $Tpts$ is the accumulative point sets with the selected $n$ frames of image pairs. In practice, we accumulate at least 12 correct matching point pairs for parameters calculation.

Substitute $Tpts$ coordinates into formula (12); we can get the transformation matrix parameters with the Least Squares method. The transformation matrix obtained with this method has higher accuracy and applicability.

## 3. Improved Contourlet Transform Fusion Combined with Total Variation and Local Region Energy

*3.1. Improved Contourlet Transform.* Contourlet transform is a multiscale geometric image analysis method. It completes multiscale analysis and direction analysis separately. Firstly, an image is multiscale decomposed with Laplacian Pyramid (LP) method to capture the singular points. With LP process, the source signal is decomposed into a low frequency and a band-pass signal of difference between source and predicted signal. Then, the band-pass signals of each pyramid level are direction-filtered. The singular points distributed in the same direction are combined into a coefficient by a *directional filter bank* (DFB). Both LP and DFB can be fully reconstructed. The discrete Contourlet transform of their combination can also be reconstructed perfectly. After registration, the resolution of both visible and IR images is 960×576. There are 3 LP decomposition levels in Contourlet transform. The direction numbers of DFB are 2-3-4.

The procedure of IR and visible images fusion with improved Contourlet transform is shown in Figure 8. There are 3 steps:

(1) Contourlet transform to both IR and visible images, respectively, to get their coefficients of high and low frequency subbands under different scales.

(2) Different fusion rules fused low and high frequency coefficients. For low frequency subband, the coefficients are fused with total variation model. For high frequency subbands, the coefficients are fused with local region energy.

(3) The fusion image is reconstructed by inverse Contourlet transform.

*3.2. Low Frequency Fusion with Total Variation Model.* Due to different principles of IR and visible imaging, different images include their own information. An IR image indicates the thermal distribution of objects. However, a visible image contains detailed information such as color and contour of objects. The low frequency subband keeps the approximate characteristics and contains most energy of source images. The familiar weighted fusion for low frequency subband will reduce the scales of both thermal radiation in IR image and colors in visible image, which will reduce the contrast of fusion. In this paper, the low frequency subband fusion is transformed into the minimization of optimization problem of total variation (TV) model.

Assume that the IR, registered visible, and fused image are all gray images of $m \times n$. Their column vectors are $u, v, x \in R^{mn \times 1}$. Since the targets in IR image are usually much more distinct than in visible image, we expected that the difference between fused and IR image in formula (14) is as small as possible:

$$\varepsilon_1(x) = \|x - u\|_1 \quad (14)$$

On the other hand, we also need details of objects. A simple method is to make the fused image have similar pixel intensity to the visible image. But the detailed appearance of object depends more on gradient than on pixel intensity. So fused image should have similar gradient rather than pixel intensity to visible image. The difference in formula (15) should also be as small as possible.

$$\varepsilon_2(x) = \|\nabla_x - \nabla_v\|_1 \quad (15)$$

where $\nabla$ is gradient operator.

The coefficient solution of low frequency subband fusion can be expressed as a minimization problem of an objective function in formula (16).

$$\varepsilon(x) = \varepsilon_1(x) + \lambda\varepsilon_2(x) = \|x - u\|_1 + \lambda\|\nabla x - \nabla v\|_1 \quad (16)$$

where $\varepsilon_1(x)$ indicates that fused image $x$ has similar pixel intensities to IR image $u$; $\varepsilon_2(x)$ indicates that fused image $x$

has similar gradients to visible image $v$; and $\lambda$ is a positive constant that controls trade-off between the two terms. The objective function (16) aims to transfer the gradients/edges in visible image onto the corresponding positions in IR image. So the low frequency fusion image should look like an IR image but with more appearance details.

Obviously, the objective function (16) is convex and has global optimal solution. The first term $\varepsilon_1(x)$ is smooth, and the second term $\varepsilon_2(x)$ is nonsmooth. Assuming $z = x - v$, then (16) can be rewritten as

$$z^* = \arg\min \left\{ \sum_{i=1}^{mn} \left| z_i - (u_i - v_i) \right| + \lambda J(z) \right\} \tag{17}$$

where

$$J(z) = \sum_{i=1}^{mn} \left| \nabla_i z \right| = \sum_{i=1}^{mn} \sqrt{\left(\nabla_i^h z\right)^2 + \left(\nabla_i^v z\right)^2} \tag{18}$$

The new objective function (17) is a standard total variation (TV) minimization problem. It is the model for low frequency subband fusion. $z^*$ can be calculated with $\lambda$ regularization of the TV minimization problem. Then the global optimal solution of fusion image is $x^* = z^* + v$. The column vector of $x^* \in R^{\mathrm{mn}\times 1}$ is the low frequency fusion coefficients.

*3.3. High Frequency Fusion with Local Region Energy.* High frequency subbands stand for details of images. The local region energy is used for high frequency subband coefficients fusion. Not only the pixel itself but also its local neighbor pixels are involved in the calculation. The procedure is as follows:

(1) For two images $A$ and $B$, calculate their own local region energies $E_{l,A}, E_{l,B}$ on their corresponding pyramid level centered $(n, m)$ separately:

$$E_l(n,m) = \sum_{n' \in J, m' \in K} \omega'\left(n', m'\right) \left[ LP_l\left(n + n', m + m'\right) \right]^2 \tag{19}$$

where $E_l(n,m)$ is local region energy centered $(n,m)$ on the $l$th level of Laplacian Pyramid; $LP_l$ is the $l$th level image; $\omega'(n', m')$ is the weight coefficient corresponding to $LP_l$; $J, K$ define the range of local region in images $A$ and $B$, and $n', m'$ are inner points in $J$ and $K$.

(2) Calculate the matching degree between corresponding local regions in two images:

$$M_{l,AB}(n,m) = \frac{2 \sum_{n' \in J, m' \in K} \left\{ \omega'\left(n', m'\right) LP_{l,A}\left(n + n', m + m'\right) LP_{l,B}\left(n + n', m + m'\right) \right\}}{E_{l,A}(n,m) + E_{l,B}(n,m)} \tag{20}$$

where $E_{l,A}, E_{l,B}$ are calculated with formula (19).

(3) According to the matching degree, different fusion rules are adopted as follows:

If $M_{l,AB}(n,m) < \alpha$ ($\alpha = 0.85$), it indicates that the correlation between two source images is low. The coefficient with larger energy is used as fused coefficients shown as formula (21).

$$\begin{aligned} &if\ E_{l,A}(n,m) \geq E_{l,B}(n,m), \\ &\quad LP_{l,F}(n,m) = LP_{l,A}(n,m) \\ &if\ E_{l,A}(n,m) < E_{l,B}(n,m), \\ &\quad LP_{l,F}(n,m) = LP_{l,B}(n,m) \end{aligned} \tag{21}$$

If $M_{l,AB}(n,m) \geq \alpha$, it indicates that the correlation degree between two source images is very large, and the weighted average fusion method shown in (22) is better.

$$\begin{aligned} &if\ E_{l,A}(n,m) \geq E_{l,B}(n,m), \\ &\quad LP_{l,F}(n,m) = W_{l,\max}(n,m)\, LP_{l,A}(n,m) \\ &\qquad + W_{l,\min}(n,m)\, LP_{l,B}(n,m) \\ &if\ E_{l,A}(n,m) < E_{l,B}(n,m), \\ &\quad LP_{l,F}(n,m) = W_{l,\min}(n,m)\, LP_{l,A}(n,m) \\ &\qquad + W_{l,\max}(n,m)\, LP_{l,B}(n,m) \end{aligned} \tag{22}$$

where

$$W_{l,\min}(n,m) = \frac{1}{2} - \frac{1}{2}\left( \frac{1 - M_{l,AB}(n,m)}{1 - \alpha} \right) \tag{23}$$

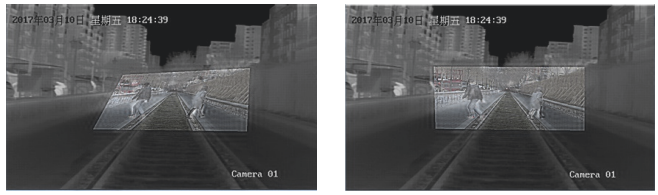
$$W_{l,\max}(n,m) = 1 - W_{l,\min}(n,m) \tag{24}$$

## 4. Experiment Results and Analysis

The performance of proposed registration and fusion algorithm is discussed in this section. The registration results of single frame and multiframe image pairs are introduced in Section 4.1. In Section 4.2, the performance comparison between source images and 5 state-of-the-art fusion results is presented and discussed. In Section 4.3, 15 frames of image pairs sampled from daytime and night videos separately will be used to evaluate the stability of proposed algorithm.

*4.1. Results of Single and Multiframe Registration.* Since single frame pair cannot ensure getting enough matching point pairs for transform matrix calculation, the multiframe

Figure 9: Infrared and visible images.



(a) Single frame registration (b) Multiframe registration

Figure 10: Fusion result of single and multiframe registration.

accumulative features method not only solves the problem of insufficient matching points in single frame pair but also improves the applicability of transform model with evenly distributed feature points. The IR and visible source images of railway scene are shown in Figure 9. Their fused image with single frame registration (using 4 pairs of feature points in single image pair) is shown in Figure 10(a). The left rail in the fused image is not aligned very well because of the concentrated feature points for matrix calculation. The fused image with multiframe registration is shown in Figure 10(b). The coincidence of all objects is much better than that in Figure 10(a).

*4.2. Fusion Performance Evaluation.* Figure 11 shows the source images and fusion results of 5 state-of-the-art methods (including ours) for daytime and night images. From top to bottom, they are source IR image, registered visible image, fusion result with gray weighted average, Contourlet transform, wavelet transform, total variation fusion, and our proposed method. The left column shows daytime images; and the right column shows night images. In Figure 11, the fusion results of gray weighted average, wavelet transform, and total variation are much smoother and more blurred. The results of Contourlet transform and ours are much sharper and closer to source IR image. The result of our method at night is sharpest and of highest gray value. It not only kept the thermal information of IR image, but also greatly weakened the strong disturbance light at night.

Due to the subjective assessment varying from person to person, we also introduce objective evaluation with quantitative indicators. Average gray, standard deviation, average gradient, information entropy, and edge intensity are used for quantitative analysis. Their definitions are as follows:

(a) Average gray is the mean value of all pixels in the image, defined as

$$Ave = \frac{1}{n \times m} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} P(i, j) \tag{25}$$

where $P(i, j)$ is the gray value at pixel $(i, j)$; $n, m$ are horizontal and vertical size of image; and $Ave$ is the average gray of the image.

(b) Standard deviation describes the dispersion between all pixels and average gray, defined as

$$\sigma = \sqrt{\frac{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left[P(i, j) - Ave\right]^2}{m \times n}} \tag{26}$$

(c) Average gradient, also known as sharpness, reflects image sharpness and variation in texture details, defined as

$$Ave_v = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{m-2} \left|F(i+1, j) - F(i, j)\right|}{(n-1) \times m} \tag{27}$$

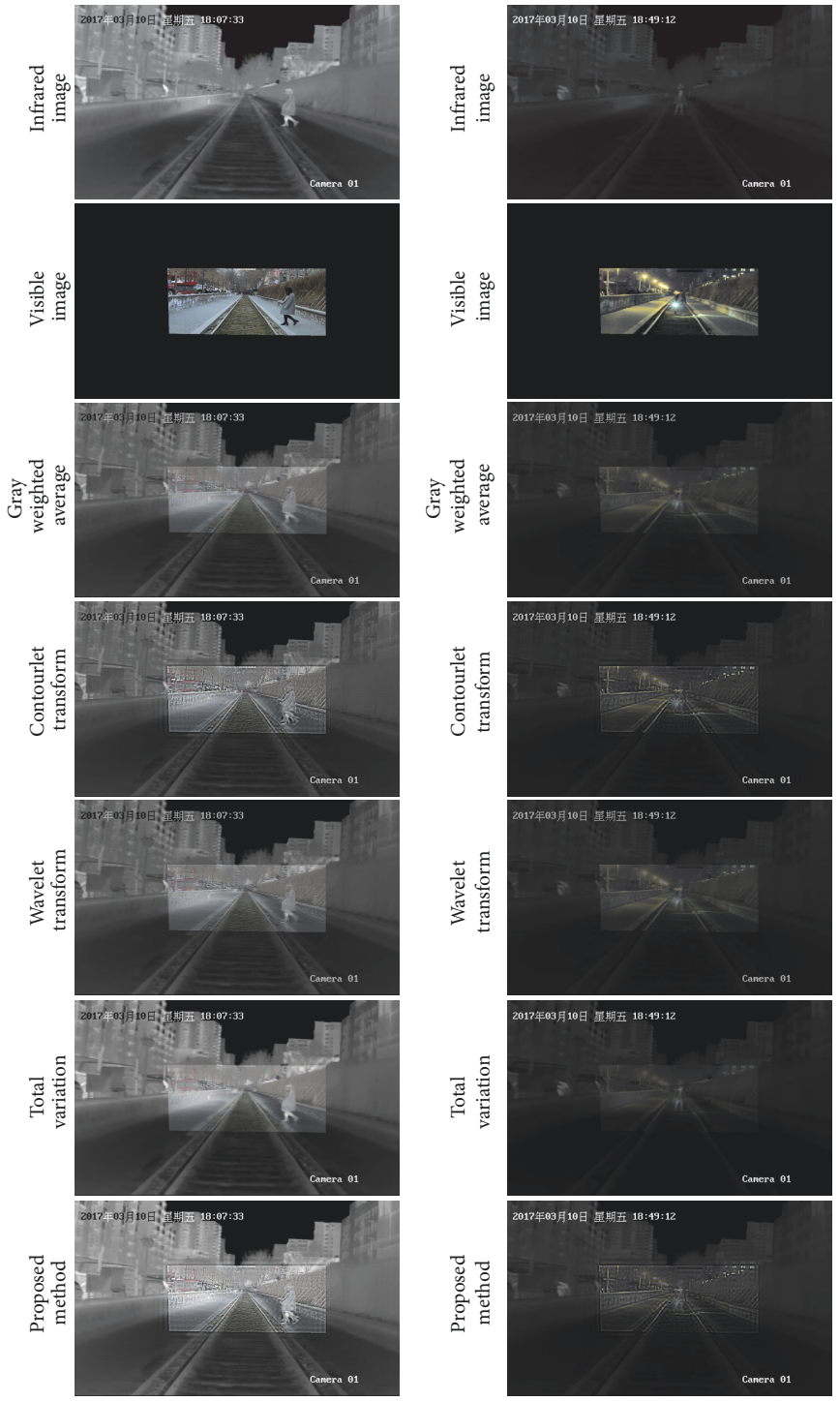
$$Ave_H = \frac{\sum_{i=0}^{n-2} \sum_{j=0}^{m-2} \left|F(i, j+1) - F(i, j)\right|}{(n-1) \times m} \tag{28}$$

$$AG = \sqrt{Ave_v^2 + Ave_H^2} \tag{29}$$

where $Ave_v$, $Ave_H$ are average gradients in vertical and horizontal direction, respectively and $AG$ is the average gradient of the fused image $F$.

(d) Information entropy is used to evaluate the richness of image information. We assume that the gray values of each pixel in an image are independent samples; then the gray distribution is $P = \{P_0, P_1, \ldots, P_i, \ldots, P_n\}$; $P_i$ is the probability of the gray value $i$ in image, that is, the ratio of $N_i$ (the pixels number with gray value $i$) to $N$ (the total pixels number in the image). $L$ is the total number of gray levels in the image. According to Shannon's theorem, information entropy of an image is defined as

$$H = -\sum_{i=0}^{L-1} P_i \log P_i \tag{30}$$




(a) Images at daytime (b) Images at night

Figure 11: Comparison of visual quality of source and fused images of various methods at daytime and night.

(e) Edge intensity is essentially the amplitude of edge gradient. The first-order difference in directions $x$ and $y$ is defined as

$$\nabla P_x(i,j) = P(i,j) - P(i-1,j) \tag{31}$$

$$\nabla P_y(i,j) = P(i,j) - P(i,j-1) \tag{32}$$

The amplitude of gradients is

$$G(i,j) = \sqrt{\nabla P_x(i,j)^2 + \nabla P_y(i,j)^2} \tag{33}$$

Table 2: Quantitative evaluation of source image and fusion results at daytime.

| evaluation indicators | Source image | | Fusion results | | | | |
|---|---|---|---|---|---|---|---|
| | IR | Registered visible | gray weighted average | Contourlet Transform | Wavelet transform | Total variation | proposed |
| average gray | 88.37 | 17.28 | 67.09 | 67.37 | 67.07 | 76.66 | **91.60** |
| standard deviation | 47.73 | 43.20 | 37.07 | 40.13 | 37.85 | 46.46 | **51.63** |
| average gradient | 2.60 | 2.55 | 2.41 | **4.90** | 3.37 | 2.64 | 4.89 |
| Information entropy | 6.70 | 1.87 | 6.46 | 6.87 | 6.61 | 6.95 | **7.07** |
| edge intensity | 27.74 | 22.15 | 24.21 | **47.04** | 30.33 | 27.18 | 46.98 |

Table 3: Quantitative evaluation of source image and fusion results at night.

| evaluation indicators | Source image | | Fusion results | | | | |
|---|---|---|---|---|---|---|---|
| | IR | Registered visible | gray weighted average | Contourlet Transform | Wavelet transform | Total variation | proposed |
| average gray | 32.40 | 13.97 | 26.89 | 27.10 | 26.90 | 21.22 | **33.06** |
| standard deviation | 27.49 | **35.89** | 22.74 | 26.06 | 22.91 | 26.48 | 29.27 |
| average gradient | 1.67 | 1.07 | 1.44 | 2.61 | 1.65 | 1.55 | **2.65** |
| Information entropy | 5.82 | 1.83 | 5.76 | 5.93 | 5.82 | 5.09 | **6.06** |
| edge intensity | 17.70 | 11.53 | 15.23 | 27.69 | 17.07 | 16.56 | **28.05** |

For all the five indicators, the higher the value is, the better its performance is. The quantitative evaluation results for Figure 11 are shown in Tables 2 and 3.

For daytime image fusion results in Table 2, relative to the high value of source images, our proposed fusion method improved 3.66% at average gray, 8.17% at standard deviation, and 5.52% at information entropy. Because of the differences between IR and visible images, the improvement at average gradient and edge intensity is even more, 88.08% and 69.36%, respectively. Compared to other 4 methods, our method performs the best in average gray, standard deviation, and information entropy and the second best in average gradient and edge intensity. In general, compared to source images and other 4 fused results, our proposed method has the best performance.

For night image fusion results in Table 3, relative to the high value of source images, our proposed fusion method improved 2.03% at average gray, 4.12% at information entropy, and 56.68% and 58.47% at average gradient and edge intensity, respectively. But it has lower standard deviation than visible source image because of great illumination variation of lamps and shining of torch in dark environment. The illumination variety of large standard deviation leads to mass misdetection in visible images analysis. That is what we want to eliminate at night. In conclusion, our proposed fusion algorithm improved all other evaluation indicators except standard deviation. It has the best performance among the 5 fusion methods.

*4.3. Stability Verification of Fusion Algorithm.* Section 4.2 only shows the effectiveness of our algorithm on one typical frame pair at daytime and night separately. In order to verify the stability and reliability of our algorithm, 15 frames of image pairs sampled from different daytime and night videos, respectively, are used for quantitative evaluation. In this section, besides three indicators of standard deviation, information entropy, and sharpness defined in Section 4.2, we also introduced another two indicators of mutual information and cross entropy defined as follows:

(a) Mutual information represents a measure of correlation between multiple variables. The mutual information $MI((A,B):F)$ between images $A$, $B$, and $F$ is defined as

$$MI((A,B):F) = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\sum_{k=0}^{L-1} P_{abf}(i,j,k)\log\frac{P_{abf}(i,j,k)}{P_{ab}(i,j)\,P_f(k)} \tag{34}$$

Similarly, the relation chart between $i$ and $P_{\mathrm{ab}}(i,j)$ can be thought as a normalized joint grayscale histogram of images $A$ and $B$; the relation chart between $i$ and $P_{\mathrm{abf}}(i,j,k)$ is the normalized joint grayscale histogram of images $A$, $B$, and $F$.

The larger the mutual information is, the more information the fusion image gets from source images.

(b) Cross entropy is used to measure the information difference of gray distribution between two images. Assuming reference image $R$ and fusion image $F$, the cross entropy of $R$ and $F$ is defined as

$$CE_{R,F} = \sum_{i=0}^{L-1} P_{R_i}\log\frac{P_{R_i}}{P_{F_i}} \tag{35}$$

where $P_{Ri}$ and $P_{Fi}$ are gray distribution of reference image $R$ and fusion image $F$, respectively.

The smaller the cross entropy value is, the more information the fusion image gets from source images and the better performance the fusion algorithm has. Because of the different quality of daytime and night images, we evaluate them separately as follows:

*(1) Fusion Evaluation of Daytime Images.* We sampled 15 frames of image pairs from daytime videos and compared 5 evaluation indicators of 4 state-of-the-art methods and our

Table 4: Average evaluation value of different fusion methods at daytime.

| | Standard deviation | Mutual information | Information entropy | Cross entropy | Sharpness |
|---|---|---|---|---|---|
| Gray weighted average fusion | 36.686 | 1.571 | 6.454 | 1.084 | 3.099 |
| Contourlet transform fusion | 39.767 | **1.771** | 6.860 | 0.802 | **7.202** |
| Wavelet transform fusion | 37.463 | 1.644 | 6.602 | 0.860 | 5.224 |
| Total variation fusion | 44.223 | 1.687 | 6.889 | 0.590 | 3.099 |
| Proposed | **51.078** | 1.717 | **7.061** | **0.279** | 6.923 |

Table 5: Average evaluation value of different fusion methods at night.

| | Standard deviation | Mutual information | Information entropy | Cross entropy | Sharpness |
|---|---|---|---|---|---|
| Gray weighted average fusion | 22.270 | 1.555 | 5.674 | 0.623 | 1.540 |
| Contourlet transform fusion | 25.425 | **1.722** | 5.841 | 0.628 | **2.852** |
| Wavelet transform fusion | 22.413 | 1.597 | 5.729 | 0.587 | 1.936 |
| Total variation fusion | 26.322 | 1.681 | 5.066 | 0.849 | 1.618 |
| Proposed | **28.846** | 1.690 | **6.001** | **0.134** | 2.828 |

proposed method. The 4 compared methods include gray weighted average, Contourlet transform, wavelet transform, and total variation methods. The 5 indicators including standard deviation, mutual information, information entropy, cross entropy, and sharpness of 5 methods are shown in Figures 12(a)–12(e). In Figure 12, except the large deviation in standard deviation, information entropy, and cross entropy of total variation method, all the other indicators of 5 methods have little deviation in the 15 frames. So, we can draw the conclusion that, except total variation method, the other 4 fusion methods have stability for different day images.

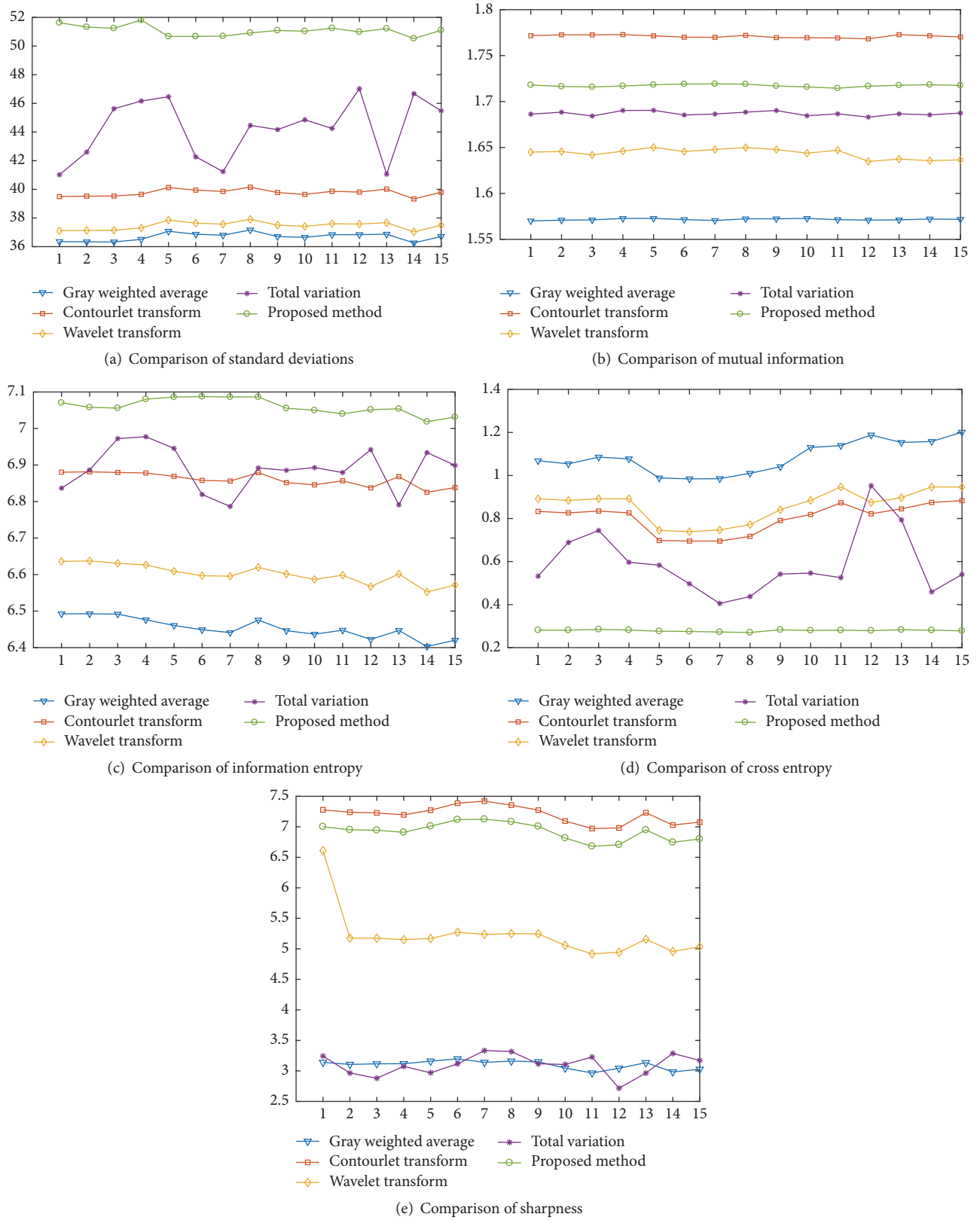
The average values of 15 daytime frame pairs are shown in Table 4. In the 5 evaluation indicators, for cross entropy, the less its value is, the more information of source images is translated into the fused image and the better the fusion effect is. For the other 4 indicators, the larger they are, the better the fusion effect is. For the 5 algorithms, our proposed method is the best in standard deviation, information entropy, and cross entropy and the second best in mutual information and sharpness. It improved 15.5% at standard deviation, 2.5% at information entropy, and 52.7% at cross entropy compared to the second best method. And as the second best one in mutual information and sharpness, the difference to the best one is very little. In summary, our proposed method is much better than other methods and can get best performance for day images fusion.

*(2) Fusion Evaluation of Night Images*. The visible images usually do not have good quality at night because of bad illumination and disturbance light. So we should also evaluate the fusion effect for night images. We still sampled 15 frames of railway night image pairs from different night videos. The 5 evaluation indicators of different methods are shown in Figures 13(a)–13(e). In Figure 13, different methods have relatively stable performance. The deviations of all indicators are much greater than daytime because of the poor quality of night visible images. The deviations of 5 methods on the same frame have similar trends. This phenomenon indicates that deviations depended more on image qualities rather than on different fusion methods. In fact, frame 6 has the biggest deviation because of strong disturbance lights.

The average values of 15 night frames pairs are shown in Table 5. In Table 5, our proposed method has the best performance in standard deviation, information entropy, and cross entropy and improved 9.6%, 18.5%, and 77.2%, respectively, compared to the second best one. It is also the second best method in mutual information and sharpness. In summary, our proposed method has the most outstanding performance for night images fusion. It reduced the strong disturbance of torchlight and improved the contour and thermal information of objects at night.

## 5. Conclusion

In this paper, we propose a novel registration and fusion algorithm for multimodal railway images with different field of views. One of the main novelties is the multimodal matching SURF features refining procedure with geometric, similar triangle, and RANSAC constraints in registration process. Another novelty is the improved Contourlet transform image fusion algorithm combined with total variation model and local region energy. Experiment results of railway images validate the effectiveness of our proposed registration and fusion approach. Compared to other 4 state-of-the-art methods, our method performs best and improves 15.5%, 2.5%, and 52.7% for day images and 9.6%, 18.5%, and 77.2% for night images compared to the second best one in standard deviation, information entropy, and cross entropy. And it is the second best one in mutual information and sharpness. Our method greatly reduced the strong disturbance light of torch at night. This paper gives impetus to the research on objects detection in complex circumstance. As for future work, we plan to use the fusion results to study the railway clearance intrusion detection methods.




(a) Comparison of standard deviations

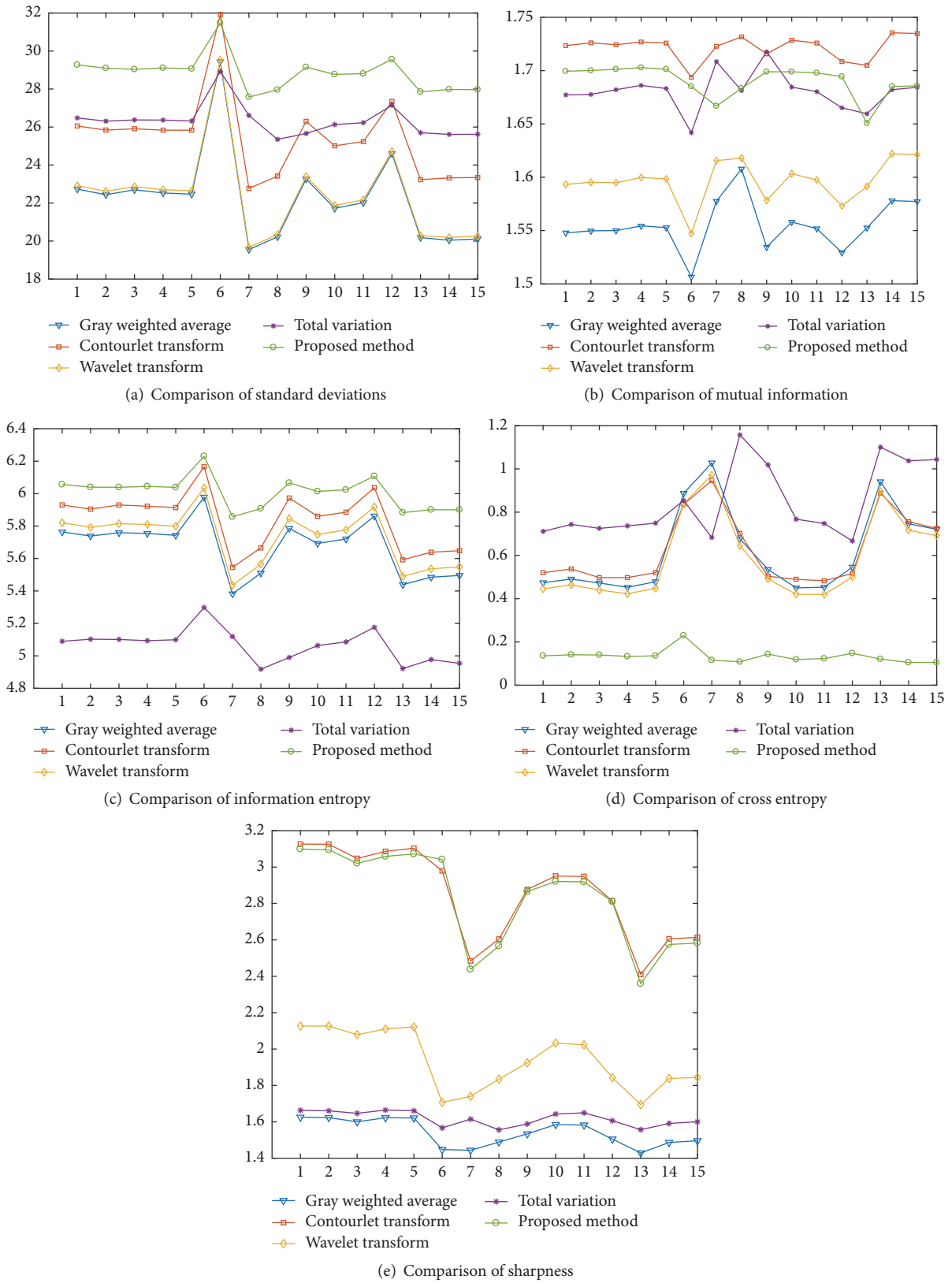
(b) Comparison of mutual information

(c) Comparison of information entropy

(d) Comparison of cross entropy

(e) Comparison of sharpness

Figure 12: Quantitative evaluation of day images fusion.




(a) Comparison of standard deviations

(b) Comparison of mutual information

(c) Comparison of information entropy

(d) Comparison of cross entropy

(e) Comparison of sharpness

Figure 13: Quantitative evaluation of night images fusion.

## Data Availability

The data (including the images data) used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work is partially supported by the National Key Research and Development Program of China (2016YFB1200402), the Research and Development Plan of Chinese Railway Company (2017T001-B), and Chinese Scholarship Council (20170709507).

## References

[1] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparse collaborative appearance model," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2356–2368, 2014.

[2] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "Car Detection from Low-Altitude UAV Imagery with the Faster R-CNN," *Journal of Advanced Transportation*, vol. 2017, Article ID 2823617, 10 pages, 2017.

[3] S. Dugad, V. Puliyadi, H. Palod, N. Johnson, S. Rajput, and S. Johnny, "Ship Intrusion Detection Security System Using HoG and SVM," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 5, no. 10, pp. 2504–2507, 2016.

[4] L. M. Dong, Q. X. Yang, H. Y. Wu, H. Xiao, and M. Xu, "High quality multi-spectral and panchromatic image fusion technologies based on Curvelet transform," *Neurocomputing*, vol. 159, pp. 268–274, 2015.

[5] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised Multilayer Feature Learning for Satellite Image Scene Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 2, pp. 157–161, 2016.

[6] J. Zhao, Q. Zhou, Y. Chen, H. Feng, Z. Xu, and Q. Li, "Fusion of visible and infrared images using saliency analysis and detail preserving based image decomposition," *Infrared Physics & Technology*, vol. 56, pp. 93–99, 2013.

[7] X. Qian, Y. Wang, and L. Han, "An object tracking method based on guided filter for night fusion image," *Infrared Physics & Technology*, vol. 74, pp. 38–43, 2016.

[8] Z. Yu, L. Yan, N. Han, and J. Liu, "Image fusion algorithm based on contourlet transform and PCNN for detecting obstacles in forests," *Cybernetics and Information Technologies*, vol. 15, no. 1, pp. 116–125, 2015.

[9] M. Ghantous and M. Bayoumi, "MIRF: A multimodal image registration and fusion module based on DT-CWT," *Journal of Signal Processing Systems*, vol. 71, no. 1, pp. 41–55, 2013.

[10] Z. Song, S. Zhou, and J. Guan, "A novel image registration algorithm for remote sensing under affine transformation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4895–4912, 2014.

[11] Z. Xiong and Y. Zhang, "A critical review of image registration methods," *International Journal of Image and Data Fusion*, vol. 1, no. 2, pp. 137–158, 2010.

[12] S.-E. Raza, V. Sanchez, G. Prince, J. P. Clarkson, and N. M. Rajpoot, "Registration of thermal and visible light images of diseased plants using silhouette extraction in the wavelet domain," *Pattern Recognition*, vol. 48, no. 7, pp. 2119–2128, 2015.

[13] J. Y. Ma, J. Zhao, Y. Ma, and J. W. Tian, "Non-rigid visible and infrared face registration via regularized gaussian fields criterion," *Pattern Recognition*, vol. 48, no. 3, pp. 772–784, 2015.

[14] G. A. Bilodeau, A. Torabi, and F. Morin, "Visible and infrared image registration using trajectories and composite foreground images," *Image and Vision Computing*, vol. 29, no. 1, pp. 41–50, 2011.

[15] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 4328–4338, 2014.

[16] J. Woo, M. Stone, and J. L. Prince, "Multimodal registration via mutual information incorporating geometric and spatial context," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 757–759, 2015.

[17] J. Han, E. J. Pauwels, and P. De Zeeuw, "Visible and infrared image registration in man-made environments employing hybrid visual features," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 42–51, 2013.

[18] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[19] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

[20] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, pp. 100–109, 2016.

[21] Y. Ma, J. Chen, C. Chen, F. Fan, and J. Ma, "Infrared and visible image fusion using total variation model," *Neurocomputing*, vol. 202, pp. 12–19, 2016.

[22] H. Li, L. Liu, W. Huang, and C. Yue, "An improved fusion algorithm for infrared and visible images based on multi-scale transform," *Infrared Physics & Technology*, vol. 74, pp. 28–37, 2016.

[23] G. He, D. Dong, Z. Xia, S. Xing, and Y. Wei, "An Infrared and Visible Image Fusion Method Based on Non-Subsampled Contourlet Transform and Joint Sparse Representation," in *Proceedings of the 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 492–497, Chengdu, China, December 2016.

[24] J. Cai, Q. Cheng, M. Peng, and Y. Song, "Fusion of infrared and visible images based on nonsubsampled contourlet transform and sparse K-SVD dictionary learning," *Infrared Physics & Technology*, vol. 82, pp. 85–95, 2017.

[25] H. Li, H. Qiu, Z. Yu, and Y. Zhang, "Infrared and visible image fusion scheme based on NSCT and low-level visual features," *Infrared Physics & Technology*, vol. 76, pp. 174–184, 2016.

[26] F. Meng, M. Song, B. Guo, R. Shi, and D. Shan, "Image fusion based on object region detection and Non-Subsampled Contourlet Transform," *Computers and Electrical Engineering*, vol. 62, pp. 375–383, 2017.

[27] Z. Fu, X. Wang, J. Xu, N. Zhou, and Y. Zhao, "Infrared and visible images fusion based on RPCA and NSCT," *Infrared Physics & Technology*, vol. 77, pp. 114–123, 2016.

[28] K. He, D. Zhou, X. Zhang, R. Nie, Q. Wang, and X. Jin, "Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain," *Journal of Applied Remote Sensing*, vol. 11, no. 1, 2017.

[29] X. Yang and K.-T. T. Cheng, "Local difference binary for ultrafast and distinctive feature description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 188–194, 2014.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[31] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2564–2571, Barcelona, Spain, November 2011.

[33] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: binary robust invariant scalable keypoints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2548–2555, Barcelona, Spain, November 2011.

[34] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE Features," in *Computer Vision – ECCV 2012*, vol. 7577 of *Lecture Notes in Computer Science*, pp. 214–227, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.