IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2018, accepted March 23, 2018, date of publication March 29, 2018, date of current version April 23, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2820680*

# Cross-Domain Co-Occurring Feature for Visible-Infrared Image Matching

**JING LI[1], (Member, IEEE), CONGCONG LI[1], (Student Member, IEEE), TAO YANG[2], (Member, IEEE), AND ZHAOYANG LU[1], (Senior Member, IEEE)**

[1]School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China

Corresponding author: Jing Li (jinglixd@mail.xidian.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61502364, Grant 61672429, and Grant 61272288.

**ABSTRACT** As the two most commonly used imaging devices, infrared sensor, and visible sensor play a vital and essential role in the field of heterogeneous image matching. Therefore, visible-infrared image matching which aims to search images across them has important application and theoretical significance. However, due to the vastly different imaging principles, how to accurately match between visible and infrared image remains a challenge. In fact, the two images describe one scene from different aspects. There is a symbiotic relationship between their features, which we named as cross-domain co-occurring feature. In this paper, based on cross-domain co-occurring feature, we present a novel visible-infrared image matching algorithm. Concretely, co-occurring feature is first constructed by cross-domain image database and feature extraction approach. Then three visual vocabulary trees can be built by visible feature, infrared feature, and co-occurring feature. Thus, the symbiotic relationship between the two domains is established by co-occurring feature and vocabulary trees. With this relationship, each image is represented by a list of leaf node of co-occurring vocabulary tree. Finally, we measure the image similarity and the highest scoring image is the matching result. As a bi-directional method, we evaluate the proposed algorithm on two tasks: visible-to-infrared matching and infrared-to-visible matching. Experiments on the Korea Advanced Institute of Science and Technology all-day place recognition database captured from 42-km sequences demonstrate that co-occurring feature is effectiveness and efficiency to link different domains. And the matching approach also achieves superior performance.

**INDEX TERMS** Image matching, cross-domain co-occurring feature, visible-infrared image matching.

## I. INTRODUCTION

Multi-sensor image matching [1]–[3] has become a hot research topic in the field of computer vision with the rapid development of imaging sensor performance and abundance of sensor types. It aims to solve the matching problem across different imaging sensors and enables several sensors cooperate for a matching task. Especially, multi-sensor matching avoids the single sensor limitation, and provides more comprehensive, more accurate information for users. As the two most common imaging sensors, visible and infrared imaging devices play a fundamental role in many important applications, including visual data fusion [4]–[6], scene match location [7], [8], visual navigation [9], face recognition [10], [11] and so on. Fig. 1 provides an illustration of visible-infrared image matching.

However, due to the different imaging principles, visible image and infrared image differ greatly in many characteristics shown in Fig. 1). That is the basic reason why cross-domain image matching is so difficult. To be more specific, the main challenges cover the following points: (1) Different imaging mechanism. Infrared images and visible images reflect the properties of objects in different light wave bands. Infrared images are based on the emissivity of the object, whereas visible imaging sensor is based on its reflectivity. (2) Different imaging conditions. Image gray scale distortion and geometric deformation are easily affected by shooting time, season, light intensity, etc. That causes infrared image and visible image distinctly different in many characterises. The factors above bring certain difficulties to visible-infrared image matching.

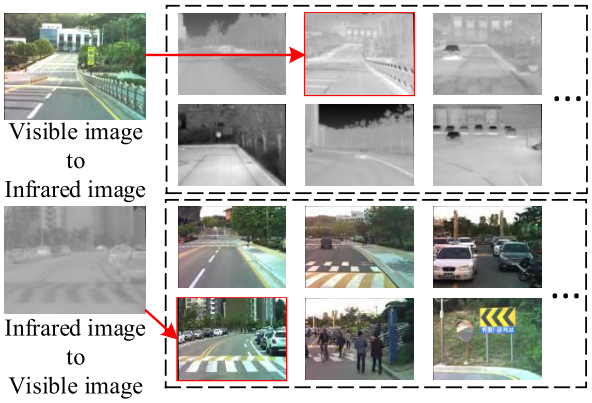



**FIGURE 1.** An illustration of visible-infrared image matching. Queries can search for corresponding images in another domain.(Above) visible-to-infrared matching. (Below) infrared-to-visible matching.

To address this problem, some researchers have made their contributions to match optical image and infrared image. Back in 1990s, Dana and Anandan [12] are the first to put forward a visible-infrared matching method, which uses multi-scale edge detection algorithm to obtain surface boundaries and implements a matching system with the hierarchical estimation process. Then in 2000, Coiras *et al.* [13] purpose a segmentation-based method for cross-domain image image matching. Their method is based on the relation between segmented triangles and it requires no pre-knowledge. Subsequently, the use of mutual information provides an effective approach for image matching and many experts develop mutual information to connect thermal and visual-light visual domain [14], [15]. Jing and Zhang [16] employ wavelet transform and maximization of mutual information to achieve high matching performance between infrared and optical airborne images. Zhuang *et al.* [17] propose a novel hybrid algorithm which combines mutual information and two optimization methods (particle swarm optimization and Powell search method) to obtain better matching performance. Moreover, in the recent years, there have appeared several new methods to solve this problem. [18], [19]. Argulewar and Jain [20] review the approaches based on LBP (local binary pattern). With LBP and relevance machine classification, they design a visible-infrared matching system and it shows great results on face image. Through a key point selection approach, Ghosh *et al.* [11] introduce a cross-domain matching algorithm which enables us to compute a fast approximation based filter [21]. Cunjian and Ross [22] present a Heterogeneous Face Recognition (HFR) framework, which uses multiple sets of subspaces generated by sampling patches from visible and thermal face images and subjects them to a sequence of transformations. With the wide application of deep learning, Liong *et al.* [23] propose a new deep coupled metric learning (DCML) method for cross-modal matching. They design two feed-forward neural networks which learn two sets of hierarchical nonlinear transformations to nonlinearly map samples from visible and infrared modality into a shared latent feature subspace. In a word, with the unceasingly thorough study, the researches in optical-infrared matching have made considerable progress and achieved great matching performance.

But despite all that, these methods have their own drawbacks and limitations: (1) Most existing methods are independent researches which focus on the cross-domain images in certain fields, such as face images, remote sensing images, airbornes images, etc. Although the visible-infrared work have emerged many, they are not universal for different matching tasks. (2) Previous researches can only accomplish unidirectional matching. That means that users can only use visible image to search for relevant infrared image or use infrared query image to find correctly visible image. These work cannot solve both ways at the same time. (3) Infrared image describes the radiation information of the image, whereas visible image reflects its reflection information. This has created the weak correlation between the two visual fields. And in the existing work, there is a lack of consistent features between infrared and visible image. To sum up, in the current stage, how to build a general bridge that connects infrared domain and visible domain is the primary consideration and foremost problem.

Therefore, in order to achieve visible-infrared matching, we firstly explore the inherent relationship between optical and thermal visual fields. In a certain view, different imaging sensors observe an object from different perspectives and display different kinds of image features. This principle is the same as the truth in the story of the blind men and elephant which is originated in ancient India. The imaging sensor which can only describe one aspect of a scene just like the blind which can only obtain a part of object features by touch. Therefore, we start with this story to find its inner relationship and then extend it to visible-light and infrared visual domain. As Fig. 2 shown, supposing that two blind men touch an elephant at the same time, the first blind begins to touch from the front and the other one from the back. At the first time, blind 1 feels a big ear and the other touches the thin tail. Then blind 1 touches the long nose and blind 2 gets the thick leg. After many times such a process, many pairs of characteristics of elephant are obtained. Although these features are quite different, they are all belong to elephant. Next time, if an animal has both big ear and thin tail, or long nose and thick leg, we guess it maybe an elephant. In other words, they coexist for the animal: elephant. And there is a symbiotic relationship between these pairs of features. So we name such a pair of features as co-occurring feature.

As for visible-light and infrared visual domain, the two imaging sensors can be seen as the blind men and there is also a symbiotic relationship between the two kinds of image features. To put it more specifically, infrared sensor describes the approximate contour feature of one important object and optical sensor reflects the detailed information such as color. Visible feature and infrared feature which are from one scene vary widely and they are the cross-domain features. However, there is still a certain internal

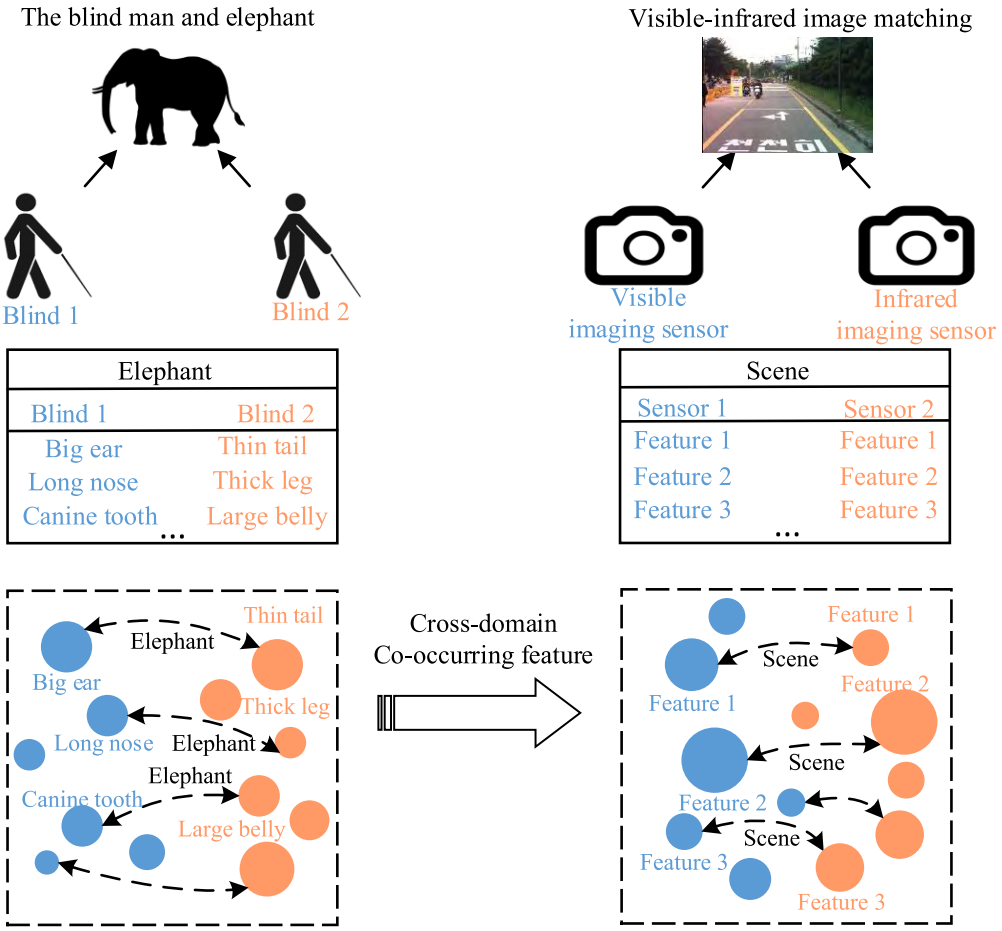



**FIGURE 2.** The story of the blinds and the elephant and visible-infrared image matching. We extend the principle in this story to visible-infrared image matching. The blue balls and the orange balls represent the features from different blinds/imaging sensors. The features connected with a dotted line are cross-domain co-occurring features.

relationship between the two kinds of features because they describe the same scene. If visible visual domain exists an image feature, its corresponding infrared feature which describes the same scene must can be obtained by infrared imaging sensor. That is the cross-domain co-occurring feature in visible-infrared image matching. With this, a bridge across different visual domain is built up. In practical image matching, if we train some cross-domain co-occurring features in advance, the inherent relation of the two visual domains is established and visible-infrared matching can be implemented.

In this paper, we systematically investigate the key idea above and purpose a novel visible-infrared image matching approach based on cross-domain co-occurring feature. Specifically, our algorithm includes two main parts: one is cross-domain co-occurring feature construction, whereas the other is visible-infrared image matching process. The first part employs a visible-infrared image database which is composed of a number of one-to-one cross-domain image pairs. With this database, we extract feature from optical image and infrared image respectively. Cross-domain co-occurring feature can be constructed by connecting the two kinds of features in series. Then, based on these features, a visible vocabulary tree, an infrared vocabulary tree and a co-occurring vocabulary tree are built. In the second portion, through these vocabulary trees, the query image and database images are quantized into the leaf nodes of their own visual domain. With the correspondence relationship in co-occurring features, the query and database images are represented as a vector by the leaf node IDs of co-occurring vocabulary tree. Finally, we compare the similar distance between these vectors and the image with the highest similarity score is the matching result.

The main contributions of our work can be summarized as follows:

- We propose a novel concept of cross-domain co-occurring feature to explore the relationship between visible and infrared images. Cross-domain co-occurring feature is constructed by a pair of cross-domain feature. The visible feature and infrared feature in one co-occurring feature are corresponding one by one. By co-occurring feature, the interrelations among different visual domain are well mined.
- We introduce an image representation approach which combines vocabulary trees and cross-domain co-occurring features. We convert each visible/infrared

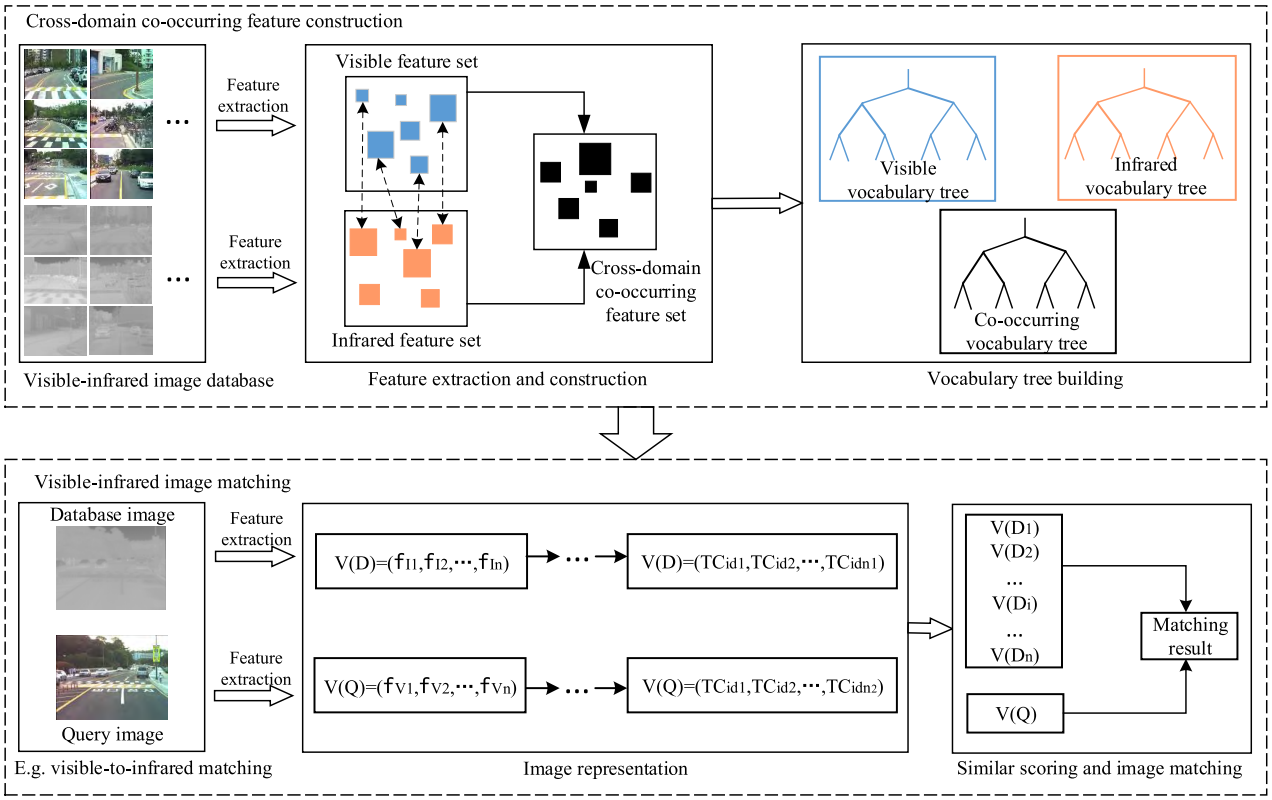



**FIGURE 3.** An illustration of our visible-infrared image matching method based on cross-domain co-occurring feature. Our approach contains two main parts: cross-domain co-occurring feature construction and visible-infrared image matching. In the first part, we employ a visible-infrared image database as the training database. At the beginning, through feature extraction, we obtain two kinds of image features and construct the cross-domain co-occurring feature. The vocabulary trees are built by hierarchical clustering these features. On the basis of the first part, we utilize the cross-domain co-occurring feature and vocabulary trees to represent each image as a vector which is a list of leaf node IDs. Here TCid is the leaf node ID of co-occurring feature vocabulary tree. Finally, we measure the similarity of these vectors and the matching result is the highest scoring database image.

feature into a co-occurring feature by utilizing the relationship among visible vocabulary tree, infrared vocabulary tree and co-occurring vocabulary tree. After co-occurring feature mapping, images from different visual domains can be represented by a list of co-occurring vocabulary leaf nodes. Our method overcomes the gap between different imaging sensors and it has been proved that this approach enjoys strong robustness.

- Based on cross-domain co-occurring feature and image representation, we present a visible-infrared image matching algorithm. What's more, our approach is a two-way work which can achieve visible-to-infrared matching and infrared-to-visible matching simultaneously. To evaluate the proposed algorithm, we apply it on the KAIST All-day Place Recognition Database [24]. Since the raw data is a video, we frame it and set up a new visible-infrared database with these images. Experimental results demonstrate that the proposed method achieves encouraging result.

The remainder of this paper is organized as follows. In Section II, we propose a visible-infrared image matching algorithm base on cross-domain co-occurring feature. The experiment result is presented in Section III. Finally, we conclude the paper in Section IV.

## II. CROSS-DOMAIN CO-OCCURRING FEATURE FOR VISIBLE-INFRARED IMAGE MATCHING

This section elaborates our proposed approach for visible-infrared image matching based on cross-domain co-occurring feature. An overview of this algorithm is shown in Fig. 3. The co-occurring feature construction method is firstly presented which is the fundamental technique in this paper. In this step, with the visible-infrared image database, we detail the feature extraction and construction process to lay a solid foundation for the follow-up work. And the vocabulary tree training method is discussed by using these features. Next, on the basis of co-occurring feature and vocabulary tree, a bridge connected the two domains is established. With this relationship, we introduce the procedure of image representation as shown. After this process, each image is converted into a list of leaf node IDs. Finally, we implement a visible-infrared image matching system by similarity measurement.

### A. CROSS-DOMAIN CO-OCCURRING FEATURE CONSTRUCTION

The proposed visible-infrared image matching approach is based on the idea of "cross-domain co-occurring feature". In our opinion, although visible image and infrared image show completely different in many kinds of features, there is

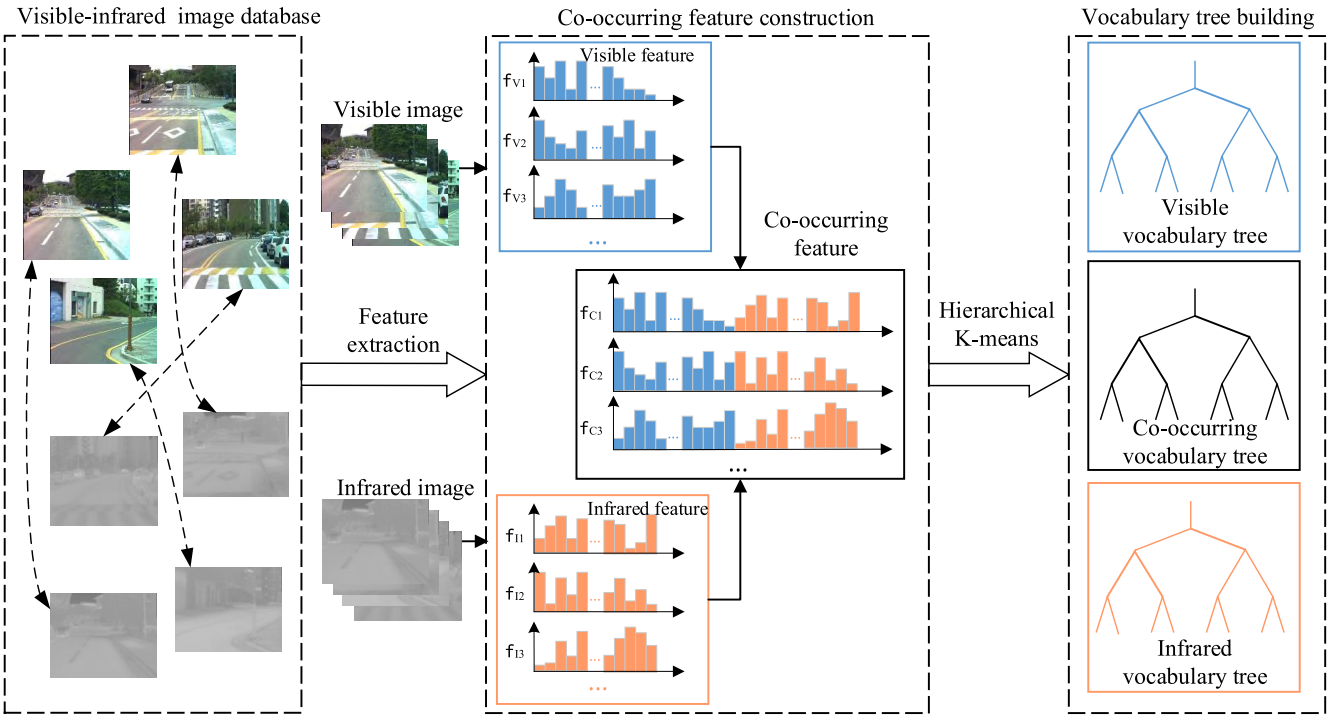



**FIGURE 4.** Cross-domain co-occurring feature construction and vocabulary trees building. With visible-infrared image database, we firstly extract their feature separately. Co-occurring feature is generated by connecting visible feature with its corresponding infrared feature in series. On the basis of these features, a visible vocabulary tree, an infrared vocabulary tree, and a co-occurring vocabulary tree will be built.

still a symbiotic relationship between these features. Therefore, in this part, we firstly extract image feature from the training database and construct cross-domain co-occurring feature. Then, on the basis of these features, visual vocabulary trees are built by feature clustering algorithm.

### 1) FEATURE EXTRACTION AND CONSTRUCTION

Before cross-domain co-occurring feature construction, we firstly introduce the training database: visible-infrared image database. This database contains a number of one-to-one visible-infrared image pairs and images from the same pair are matched one by one (shown in Fig. 4). Why should we use such an image database? This is because we must ensure that there is one-to-one correspondence between visible feature and infrared feature in the process of co-occurring feature construction. In this way, the symbiotic relationship between visible domain and infrared domain can be established accurately, which is the key to the success of visible-infrared matching. Moreover, this database is also employed for visible and infrared vocabulary tree training.

Let image sets $DC = (V_1, I_1, \ldots, V_i, I_i, \ldots, V_m, I_m)$ is a visible-infrared image database which contains $m$ pairs of cross-domain image, how do we use it to establish co-occurring feature? Fig. 5 describes the procedure intuitively. Firstly, we extract feature of each training database image. Here we adopt Scale Invariant Feature Transform (SIFT) descriptor [25] due to its good performance for a variety of image matching tasks, its matching speed, robustness of rotation, adaptability to infrared image and its wide application. SIFT algorithm includes four steps: scale-space extrema detection, keypoint localization, orientation assignment and keypoint descriptor. In the first step, difference of gaussian (DoG) simulates the multi-scale images and establishes a gaussian pyramid. Once the DoG is built, they search for the local extreme point over scale and space. The results are the potential keypoint locations. In order to get more accurate results, they use Taylor series expansion of scale space to determine the feature points. Finally, according to the gradient direction, the major direction to each feature point is assigned, and the keypoint descriptor is constructed by its orientation information. However, since visible light image and infrared image vary greatly in feature, it cannot guarantee the one-to-one corresponding relation between visible feature and infrared feature if detect feature point directly. Without this correspondence relationship, the cross-domain co-occurring feature cannot be constructed.

To tackle this problem, we use the method shown in Fig. 5 to extract and construct feature. At the beginning, the original image is parted into disjoint sub-blocks with same size and we take the center of each block as the feature point. The methods of uniform image segmentation and specified feature location can guarantee the one to one correspondence of cross-domain features. For example, if the image size is 320 $\times$ 240 and each block size is 16 $\times$ 16, an image will be divided into 300 patches and the patch enters are the keypoints. After such a procedure, features from the same keypoint of one visible-infrared image pair are mutually matched. Next, with these accurately matched feature pairs, we connect a visible

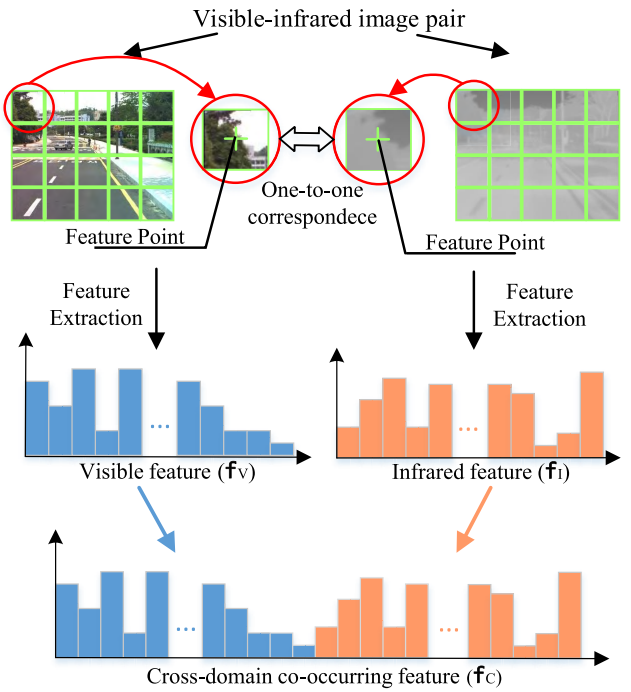



**FIGURE 5.** The illustration of feature extraction and construction. Take a visible-infrared image pair as an example, we first block the images and the center of each block is the feature point. Then through feature extraction, the visible feature and the infrared feature are generated. As the figure shown, the cross-domain co-occurring feature is constructed by connecting and in series.

feature and its corresponding infrared feature in series to form a cross-domain co-occurring feature. Suppose that a visible feature vector $\mathbf{f}_V$ and its corresponding infrared feature vector $\mathbf{f}_I$, the cross-domain co-occurring feature vector $\mathbf{f}_C$ is constructed as following;

$$\mathbf{f}_C = \left\{ \underbrace{P_1, P_2 \ldots, P_{128}}_{\mathbf{f}\,\mathrm{V}}, \underbrace{W_1, W_2 \ldots, W_{128}}_{\mathbf{f}\,\mathrm{I}} \right\} \tag{1}$$

where $P_i$ and $W_i$ represent an integer which is computed by direction amplitude of SIFT feature, 128 is the dimension of feature descriptor. And in our method, $\mathbf{f}_V$ and $\mathbf{f}_I$ are 128-dimensional, the dimension of $\mathbf{f}_C$ is 256. In this feature pair, the infrared feature and optical feature is a pair of symbiotic feature and they construct a cross-domain co-occurring feature vector $\mathbf{f}_C$. In this way, cross-domain co-occurring feature records this one-to-one relationship. The visible feature and infrared feature in one co-occurring feature is symbiotic. In the practical matching, when a feature appears in one visual domain, its symbiotic feature is likely to occur in another visual domain. At this point, co-occurring feature is employed to help us match the two kinds of features across different visual domain.

After feature extraction, we can obtain three feature collections by collating the three kinds of features above. They are visible feature set $FV$, infrared feature set $FI$ and co-occurring feature set $FC$. If an image is represented by 100 SIFT features, feature collections are denoted as:

$$FV = \{\underbrace{\mathbf{f}_{V_1}, \ldots, \mathbf{f}_{V_{100}}}_{\text{VIS image V1}}, \underbrace{\mathbf{f}_{V_{101}}, \ldots, \mathbf{f}_{V_{200}}}_{\text{VIS image V2}}, \ldots, \mathbf{f}_{V_{100\times m}}\} \tag{2}$$

$$FI = \{\underbrace{\mathbf{f}_{I_1}, \ldots, \mathbf{f}_{I_{100}}}_{\text{IR image I1}}, \underbrace{\mathbf{f}_{I_{101}}, \ldots, \mathbf{f}_{I_{200}}}_{\text{IR image I2}}, \ldots, \mathbf{f}_{I_{100\times m}}\} \tag{3}$$

$$FC = \{\underbrace{\mathbf{f}_{C_1}, \ldots, \mathbf{f}_{C_{100}}}_{\text{VIS-IR pair } C1}, \underbrace{\mathbf{f}_{C_{101}}, \ldots, \mathbf{f}_{C_{200}}}_{\text{VIS-IR pair } C2}, \ldots, \mathbf{f}_{C_{100\times m}}\} \tag{4}$$

### 2) VOCABULARY TREE BUILDING

In order to integrate these image features for image matching, we employ vocabulary tree [26] to index features. Although this method has been put forward for nearly a decade, vocabulary tree shows good performance on feature index and has great potential in image matching. Vocabulary tree combines the bag of words model [27] and tree framework. Each leaf node of the vocabulary tree can be seen as a visual word. Image matching is realized by quantizing features into leaf nodes and measuring similarity. In this section, we build three vocabulary trees: a visible vocabulary tree, an infrared vocabulary tree and a co-occurring feature vocabulary tree with the feature collections $FV$, $FI$ and $FC$. The building process mainly includes two steps: (1) Vocabulary tree establishment. (2) Creation of leaf node index file.

After determining the tree parameter and feature collection, the first step of vocabulary tree building is feature clustering. We utilize hierarchical K-means (HKM) as the clustering approach. HKM is usually used for speeding up the large-scale vocabulary tree construction which just an approximate method of K-means. In our system, HKM cluster SIFT feature which efficiently groups visually similar patches into one cluster. Supposing that the vocabulary tree depth $L$ and the branch factor $K$, HKM splits all features in feature collection into $K$ clusters firstly and then divides the data in the same cluster recursively. Feature vector is denoted as $f_j$, we divide all data into $K$ classes as $S = \{S_1, \ldots, S_i, \ldots, S_K\}$. And we compute the minimize value to establish each layer of the vocabulary tree, as in (5):.

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{f_j \in S_i} \|f_i - c_i\|^2 \tag{5}$$

where $c_i$ is the centroid of $i$th classes $S_i$. After $L$ iterations, a vocabulary tree with K-branch and L-depth is built. The total number of cluster centers is calculated by (6). In the vocabulary tree, each leaf node (that is the cluster center) is labeled by an integer of $0 \sim K^L - 1$.

$$\sum_{i=1}^{L} K^i = \frac{K^{L+1} - K}{K - 1} \approx K^L \tag{6}$$

Through this vocabulary tree building process and three feature collections (visible feature collection FV, infrared feature collection FI, co-occurring feature collection FC), we obtain three vocabulary trees: visible vocabulary tree TV, infrared vocabulary tree TI and co-occurring vocabulary

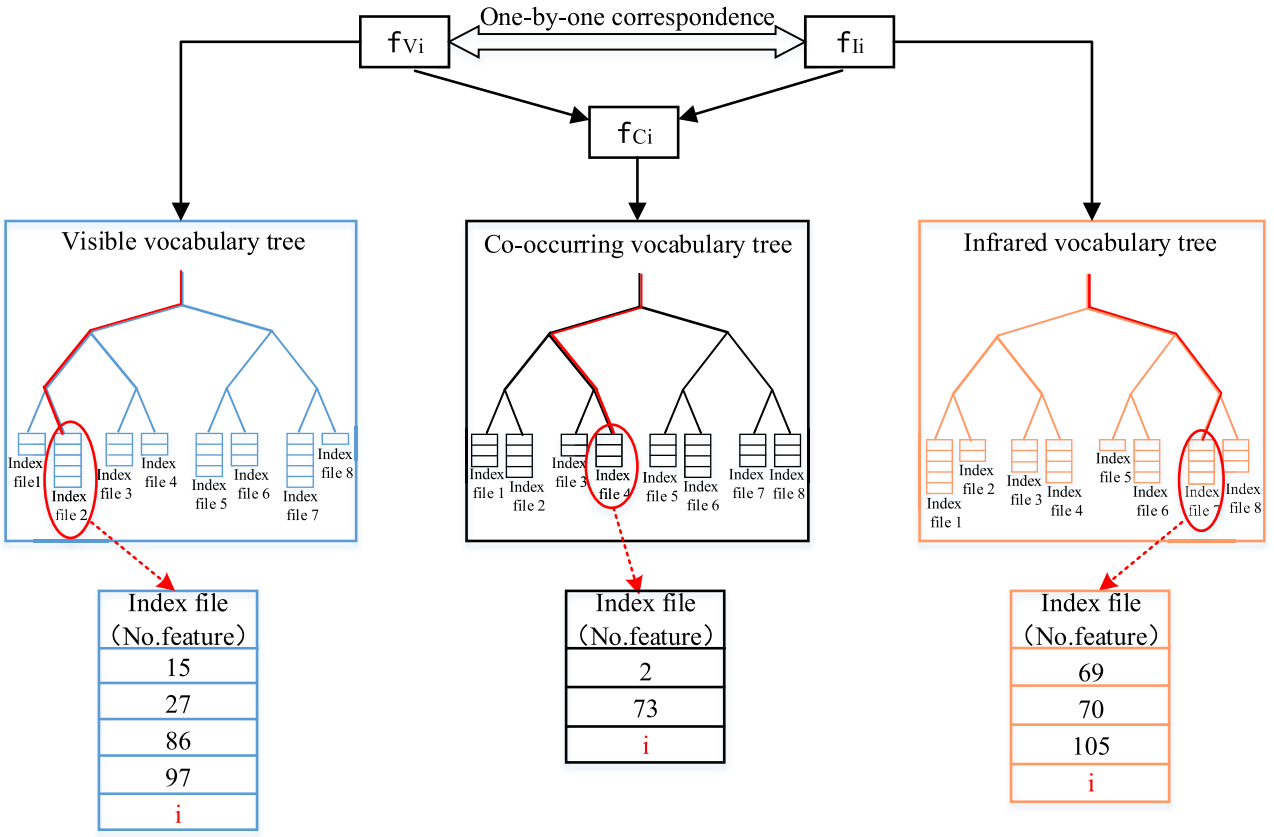



**FIGURE 6.** The process of index file creation. Suppose that the ith visible feature vector is fVi and its corresponded infrared feature vector fIi, the cross-domain co-occurring feature vector constructed by them is fCi. These features are mapped to a path of visual word from root to a leaf node in their own trees and we add its feature tag i at the end of the index file which is attached with leaf node.

tree TC. These three vocabulary trees can be regarded as the dictionaries in visible visual domain, infrared visual domain and co-occurring feature domain. The tree leaf nodes are the visual words in dictionary. Through vocabulary tree, features are represented by a vector which is a list of leaf nodes in their own field.

After vocabulary tree establishment, index file attached with leaf node is needed to create for image matching. We add each feature to a leaf node and the structure of index file is shown in Fig. 6. Take $i$th visible feature vector $f_{Vi}$ in $FV$ and the $i$th infrared feature vector $f_{Ii}$ in $FI$ as an example, the $i$th feature vector $f_{Ci}$ in $FC$ is the co-occurring feature vector which is constructed by connect a visible feature vector and its corresponding infrared feature vector in series. Feature vector $f_{Vi}$, $f_{Ii}$ and $f_{Ci}$ will be mapped to a path of visual word from root to a leaf node in their own vocabulary tree $TV$, $TI$ and $TC$. And we add its feature tag at the end of the index file which is attached to leaf node. In addition, to represent the symbiotic relationship between features, the cross-domain co-occurring features are labeled by the same number as follows:

$$FV = \{\mathbf{f}_{V_1}, \quad \mathbf{f}_{V_2}, \ldots, \mathbf{f}_{V_i}, \quad \ldots\} \tag{7}$$

$$FI = \{\mathbf{f}_{I_1}, \quad \mathbf{f}_{I_2}, \ldots, \mathbf{f}_{I_i}, \quad \ldots\} \tag{8}$$

$$\begin{aligned} FC &= \{\mathbf{f}_{C_1}, \quad \mathbf{f}_{C_2}, \ldots, \mathbf{f}_{C_i}, \quad \ldots\} \\ No.feature: &\quad 1 \quad\quad 2 \quad \ldots \quad i \quad \ldots \end{aligned} \tag{9}$$

### B. VISIBLE-INFRARED IMAGE MATCHING

So far, cross-domain co-occurring feature and three vocabulary trees are all set up. But how do we apply it to visible-infrared image matching? In this section, inspired by the image representation method [28], we propose a novel matching algorithm to deal with this problem. As illustrated in Fig. 7, we take visible-to-infrared image matching as an example to express our approach conveniently and infrared-to-visible image matching is the same principle as it. Let the image database to be matched denoted as $D = \{D_1, D_2, \ldots, D_i, \ldots, D_L\}$, which contains $L$ infrared images. $Q$ is the query image from visible visual domain.

Before facilitate data matching, each image should be represented by a vector. First of all, we extract SIFT feature from each database image $D_i$ and query image $Q$, as in (11) and (16)). Then, the features will be mapped to the vocabulary tree in their own visual domain. Each feature $\mathbf{f}_i$ is traversed in their vocabulary tree from root to a leaf node to find the nearest node. Here we use Euclidean distance to find which leaf node $\mathbf{W}_i$ is most similar with the image feature vector $\mathbf{f}_i$.

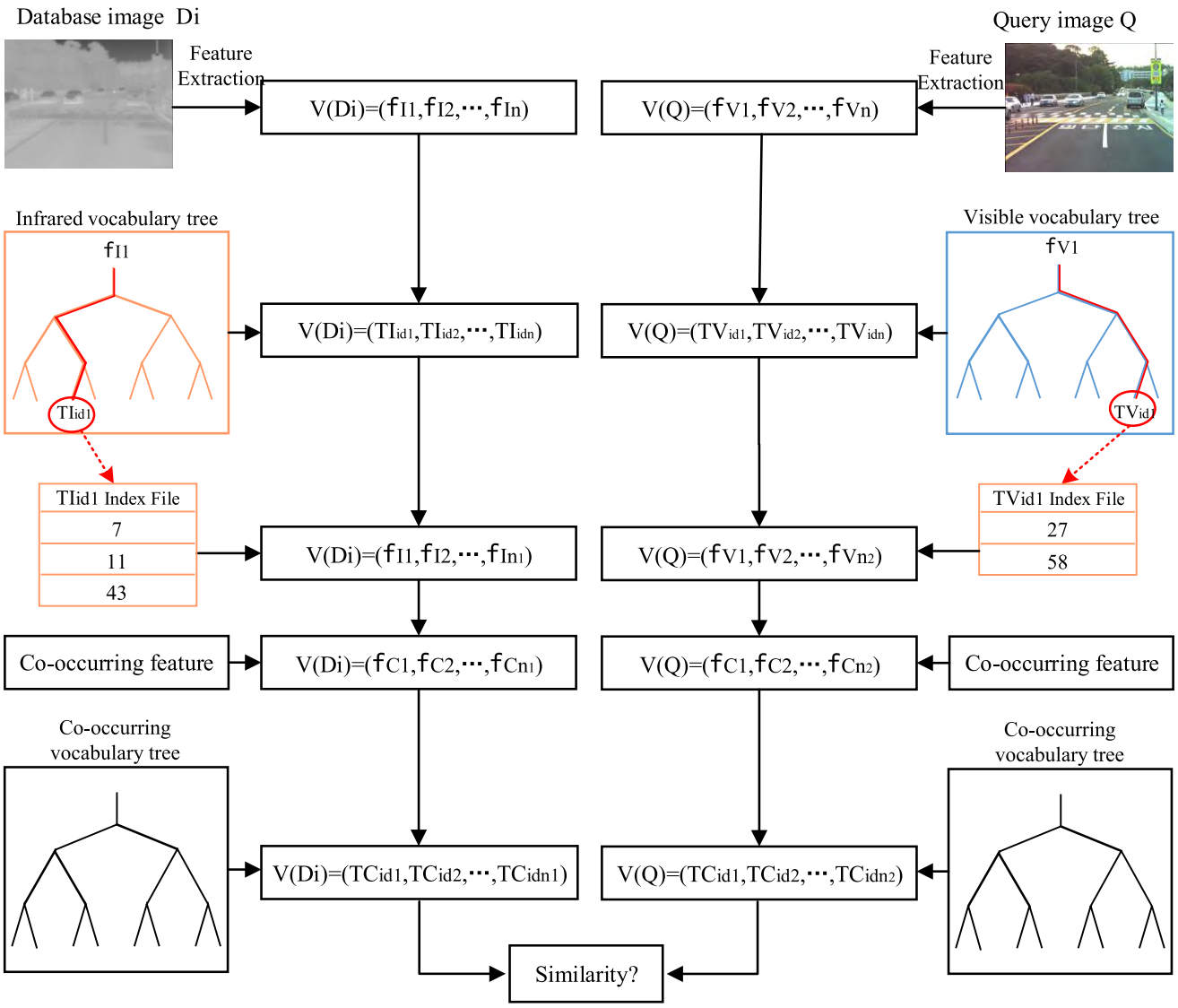



**FIGURE 7.** Image representation. On the basis of cross-domain co-occurring feature and vocabulary tree, database image and query image are represented by a list of leaf node IDs in co-occurring vocabulary tree from feature vector step by step.

The concrete calculating method is as follows:

$$D(\mathbf{f}_i, \mathbf{W}_i) = \|\mathbf{f}_i - \mathbf{W}_i\|^2 \tag{10}$$

According to feature extraction approach, SIFT feature vector is 128-dimensional and the high dimensional feature will increase the complexity of subsequent computation. In order to simplify the operation and improve the matching efficiency, we replace each 128-dimensional feature vector by an integer. And in this paper, on the basis of vocabulary tree, feature vectors are replaced by its nearest leaf node IDs as in (12) and (17). Meanwhile, the index files attached with these leaf nodes can be obtained.

However, we cannot match them directly since the query image and database image belong to different visual domain and they are represented by the leaf node of different vocabulary trees. To overcome this gap, it time to employ cross-domain co-occurring feature which contains the relationship between visible and infrared visual domain. Inspired by the matching method in [28] which is also applied to cross-domain matching, we propose a matching approach on the basis of co-occurring feature. The specific process is as follows. First of all, according to the leaf node in (12) and (17), we can know which feature is persevered in its index file during the training process. The features from the same index file can be seen as similar features. Then, by the co-occurring feature with the same label which is marked at the last section, we represent visible features and infrared features with cross-domain co-occurring feature as (14) and (19). In other words, with the symbiotic relationship between infrared feature and visible feature,

each infrared or visible is replaced by its corresponding co-occurring
feature.

Thus, database image $D_i$ and query image $Q$ are both expressed by a bag of co-occurring features. And next we use co-occurring feature vocabulary tree $TC$ to quantize each co-occurring feature into its nearest leaf node. Finally, to simplify representation, each feature descriptor is replaced by the leaf node IDs in co-occurring vocabulary tree as (15) and (20). Therefore, the image representation process is finished.

$$V(D_i) \Longrightarrow \{f_{I1}, f_{I2}, \ldots, f_{In}\} \quad (11)$$
$$\Longrightarrow \{TI_{id1}, TI_{id2}, \ldots, TI_{idn}\} \quad (12)$$
$$\Longrightarrow \{f_{I1}, f_{I2}, \ldots, f_{In1}\} \quad (13)$$
$$\Longrightarrow \{f_{C1}, f_{C2}, \ldots, f_{Cn1}\} \quad (14)$$
$$\Longrightarrow \{TC_{id1}, TC_{id2}, \ldots, TC_{idn1}\} \quad (15)$$
$$V(Q) \Longrightarrow \{f_{V1}, f_{V2}, \ldots, f_{Vn}\} \quad (16)$$
$$\Longrightarrow \{TV_{id1}, TV_{id2}, \ldots, TV_{idn}\} \quad (17)$$
$$\Longrightarrow \{f_{V1}, f_{V2}, \ldots, f_{Vn2}\} \quad (18)$$
$$\Longrightarrow \{f_{C1}, f_{C2}, \ldots, f_{Cn2}\} \quad (19)$$
$$\Longrightarrow \{TC_{id1}, TC_{id2}, \ldots, TC_{idn2}\} \quad (20)$$

After the above procedure, we are ready to match visible image and infrared image. Assume that there is a query image **Q** which is a visible light image, how to find its corresponding infrared image? The method is detailed following. We firstly measure the similarity of the query image vector $V(Q)$ and candidate vector $V(D_i)$. The highest scoring image in database is the matching result. For similarity scoring method, we count how many the same leaf nodes they have and the number is the similar score. This is because $V(D_i)$ and $V(Q)$ are made up of leaf node IDs instead of feature vectors. The more same leaf nodes they have, the more similar they are.

## III. EXPERIMENTS

Extensive experiments are conducted to evaluate the performance of the visible-infrared image matching approach based on cross-domain co-occurring feature. In this section, we employ the KAIST All-day Place Recognition Datasetas the evaluated database and it is described in subsection *A*. As for visible-infrared image matching, we perform two kinds of cross-domain matching tasks. One is visible-to-infrared matching, i.e., using a visible image to find relevant infrared images. The other is infrared-to-visible matching, i.e. using an infrared image to match relevant visible images. In addition, to validate our approach, we comprehensively compare it with some other cross-domain matching work.

The common configurations for all experiments are summarized here. All the images are resized to 320 × 240. The program is implemented in C++ and all results are based on an Intel(R) Core(TM) i7-4700MQ (2.40 GHz CPU, 8 GB RAM).

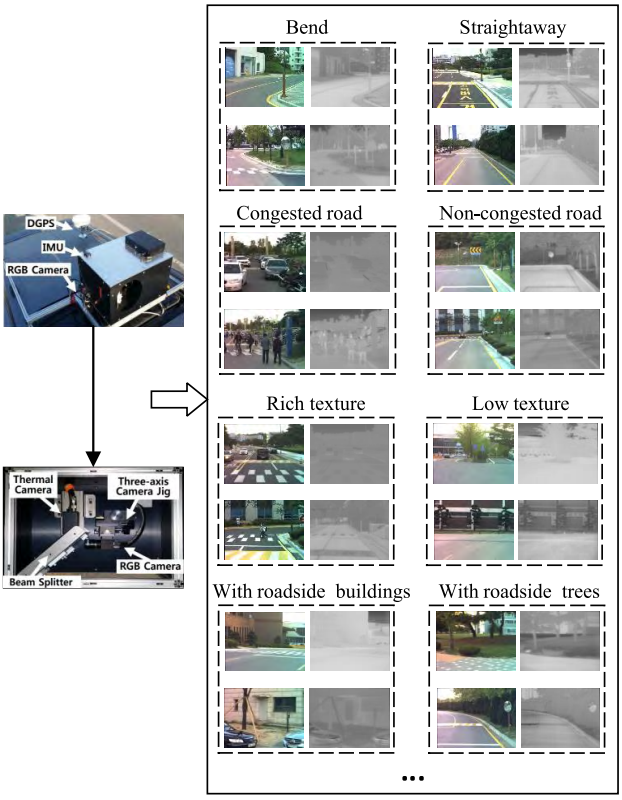



**FIGURE 8.** The general view of the KAIST All-day Place Recognition Dataset [24]. (Left) Experimental installation, The top left figure is the Beam-splitter and the side view of sensor setup is under it. (Right) Some typical scenarios of the KAIST all-day place recognition database, including bend or straightway, congested road or not, road with rich texture or not, road with roadside buildings or trees.

### A. EXPERIMENTAL SETUP

#### 1) DATASET

To manifest the advantages and generalization of the propose framework, experiments are performed on the KAIST All-day Place Recognition Dataset. This database is initially released by Yukung *et al.* [24] as a benchmark for testing computer vision and robotics algorithms. They utilize multiple imaging sensors to collect a lot of data around the Korea Advanced Institute of Science and Technology (KAIST) campus of 42km sequences at 15-100Hz. The sensors onto a standard sport utility vehicle are illustrated on the left of Fig. **??**. Note that the visible imaging sensor and thermal imaging sensor are combined with beam-splitter, which is made of zinc-oxide and silicon materials. This optical device can reflect visible wavelengths and transmit long-wavelengths infrared lights (LWIR).

The reasons why we utilize this database are summarized as follows: (1) Unlike most visible-infrared databases which are monitoring scene, this database is a novel database including dynamic objects and diverse illumination changes. It is captured in six fixed time : 4:00, 6:00, 11:00, 14:00, 18:00 and 24:00. The richness of scene and diversity of imaging condition are just what we need to prove the

**TABLE 1. The performance of different matching tasks.**

| Task | Precision | | | | Time |
|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-5 | Top-10 | |
| Visible-to-infrared | 70.6% | 79.0% | 90.6% | 95.0% | 632ms |
| Infrared-to-visible | 58.0% | 70.2% | 85.4% | 93.0% | 547ms |

robustness of the proposed method. (2) With the beam-splitter setup, the alignment of visible image and thermal image is completely parallax-free. Thus, without additional use of image rotating or straightening algorithms, the accurately matching between thermal image and visible image is easily done. It meets the demands of visible-infrared image database which contains a number of corresponding cross-domain image pairs.

To be specific, we use three subsets of this database which is captured at 11:00, 14:00, and 18:00. In order to apply it to image matching, we resolve these video into frames and obtain 3500 visible light images and 3500 infrared images. We divided these images into two parts: one is utilized to test our approach which contains 500 images in total and the remaining of them are employed as the training database. The training database and test images are captured from different roads to better evaluate the effectiveness of co-occurring feature. Moreover, in order to guarantee the completeness of our experiments, these images contain variety road conditions ( shown on the right of Fig. 8), including bend, straightway, congested traffic, non-congested traffic, road with rich texture, road with low texture, road with roadside buildings, and road with roadside trees.

#### 2) EVALUATION METRICS

We employ the Top-K precision as the performance measure. As for relevant images of each query, we define the database images which are spaced within 5 frames as the ground truth. Although these images are not exactly registered with the query, they share at least 70% same scene. Suppose N is the number of queries, we compute the Top-K precision of each query:

$$Top - K = \frac{1}{N}\sum_{i=1}^{N} P(q_i) \tag{21}$$

where $q_i$ is the $i$th query image and $P(\cdot)$ is defined as an indicator function whose value is 1 if one of the first K returns is relevant to query or 0 otherwise.

### B. EXPERIMENTAL RESULTS

As a bi-directional matching approach, we explore the performance of the proposed method on two kinds of visible-infrared matching tasks: visible-to-infrared matching and infrared-to-visible matching. Fig. 9 and Fig. 10 display some matching results and Table 1 summaries the matching data of quantitative analysis.

We first analyze the system performance qualitatively (shown in Fig. 9 and Fig. 10). We divided the query images into four traffic situations: bend or straightway, congested traffic or not, road with rich texture or not, road with many roadside buildings or trees. As we can see from these figures, the queries which are presented in the first column are great different in variety image characterises and the richness of scene help us to better evaluate performance. we firstly analyze the matching data as a whole. Most of the returns are highly related to the query image, even some of them and its queries are from the same scene. The first returns which have the highest similarity scoring are accurately registered with query images in many cases. What's more, according to statistics, the Top-1 precision is more than 70 percent for visible-to-infrared matching and approximately 60 percent for infrared-to-visible matching. For scene matching,we only need to a correct result to help us on scene location and recognition. The first return precision of the proposed approach is enough to meet the need of practical applications. As for the other returns, we find that these images also show the target scene in a similar viewpoint. This provides user with more information about the query scene to analyze comprehensively. However, as these figures shown, the matching system still returns some error images. Although these images are not from the corresponding scene, they have some extent coincided with the query scene. In another sense, these images are not the exactly results for image matching task, but some of them is the right returns for image retrieval task which is not such stringent on the degree of image similarity. This also indicates that our approach has potential on image retrieval and we will study the performance of this field in the future work.

We then carry out a quantitative analysis of the proposed method. For quantitative evaluation, we calculate the Top-K precision on the two matching directions respectively, where K is 1, 2, 5, 10. As shown in Table 1, both visible-to-infrared matching and infrared-to-visible matching show encouraging results in efficiency and precision. In accuracy, as the number of results increases, the rate of correctness rises and more than 90 percent queries can get its correct matches within the top 10 results on both matching directions. Thus, if we return the first ten returns, almost all images can obtain the correct result. That means visible-infrared image matching is well realized with our approach. The great performance is not just in terms of precision, but also in terms of efficiency. The proposed method only takes about 600ms per query which verifies the better time-consuming performance of the proposed method.

In addition, from Table 1, our approach achieves unbalanced performance in the two matching directions on time and accuracy. Visible-to-infrared matching shows better Top-K precision and infrared-to-visible matching has lower response time. The reasons of this phenomenon are analyzed as follows. Firstly, infrared image tends to be less significant in the area of image feature than visible image. That makes some local features extracted from infrared query image difficult for matching. Then, visible imaging sensor has the higher resolution because of its imaging principle. For matching,

**TABLE 2.** The performance comparison of previous cross-domain matching methods.

| Task | Methods | Precision | | | | Time |
|---|---|---|---|---|---|---|
| | | Top-1 | Top-2 | Top-5 | Top-10 | |
| Visible-to-infrared | Vocabulary tree [26] | 2.6% | 3.0% | 5.6% | 11.2% | 1108ms |
| | Visual Translator [28] | 2.0% | 2.0% | 2.6% | 3.4% | 766ms |
| | SIFT [25] | 1.8% | 3.0% | 6.4% | 9.2% | −− |
| | SURF [29] | 2.2% | 3.8% | 6.6% | 8.8% | −− |
| | ORB [30] | 1.0% | 1.8% | 3.0% | 4.4% | −− |
| | Our proposed method | **70.6%** | **79.0%** | **90.6%** | **95.0%** | **632ms** |
| Infrared-to-visible | Vocabulary tree [26] | 1.8% | 2.4% | 2.8% | 6.6% | 1053ms |
| | Visual translator [28] | 2.8% | 3.4% | 5.6% | 9.2% | 744ms |
| | SIFT [25] | 1.2% | 2.8% | 5.0% | 7.4% | −− |
| | SURF [29] | 1.0% | 2.4% | 4.8% | 7.2% | −− |
| | ORB [30] | 1.2% | 1.6% | 3.2% | 4.2% | −− |
| | Our proposed method | **58.0%** | **70.2%** | **85.4%** | **93.0%** | **547ms** |

the higher resolution the image is, the better performance achieves. The two reasons mainly affect the matching precision. However, due to the lower resolution, infrared image is faster in image reading and feature extraction, so infrared-to-visible image matching is more efficient.

All in all, extensive experimental results demonstrate the effectiveness and efficiency of the proposed approach in both matching directions and the great performance can satisfy users demands well.

### C. DISCUSSION

#### 1) THE PERFORMANCE COMPARISON

The following matching algorithms are chosen as the contrast experiments.

##### *a: VOCABULARY TREE [26]*

Vocabulary tree is an index scheme which is a significantly powerful tool for many image matching tasks. This approach first hierarchically quantizes the feature descriptors extracted from local regions. Then it employs L1-norm as image similarity definition and the matching result is the highest scoring database image. Vocabulary tree is robust to background clutter and it help us process large data sets more efficiently.

##### *b: VISUAL TRANSLATOR [28]*

Visual vocabulary translator is proposed to establish a bridge between different visual domains. Visual translator consists of two main modules: one is a pair of vocabulary trees which can be regarded as the codebooks in their respective fields, whereas the other is the index file based on cross-domain image pair. Through such a translator, a feature from one visual domain is translated into another and cross-domain image matching system is implemented. After extensive experiments, the proposed algorithm shows great results on different cross-domain matching tasks and visual vocabulary translator is effectiveness and efficiency for cross-domain image.

##### *c: SIFT [25], SURF [29], AND ORB [30]*

SIFT is one of the most widely used feature extraction algorithms. Speeded Up Robust Features (SURF) and Oriented FAST and Rotated BRIEF (ORB) are the improved methods for SIFT. Since they are proposed, there have appeared several visible and infrared researches [31]–[34] on the basis of these three methods. Therefore, we chose them as the comparison methods. For visible-infrared image matching, we firstly extract feature from visible and infrared image, respectively. Then, these features are matched each other by Euclidean distance and the number of good matches is the image similarity. The image with the highest score of similarity is the matching result. In addition, since these methods do not exploit any index algorithm which has a great impact on the matching time, we only compare the precision with our proposed method.

Table 2 shows the performance of visible-infrared image matching in terms of Top-K precision over the KAIST database. The proposed method outperforms the other methods in both precision and matching time. (1) For precision, the proposed method based on cross-domain co-occurring feature achieves the relative improvement of more than 65 percent for visible-to-infrared image matching and approximately 55 percent for infrared-to-visible image matching. Fig. 11 and Fig. 12 intuitively shows the advantages of our approach. We can see from these figures that the vocabulary tree basically impossible to match between visible image and infrared image. Visual translator also returns incorrect database images, but some of these results have a certain similarity with the query image in structure. Almost all the matching results with SIFT, SURF and ORB methods are incorrect. That means the classical matching algorithms which are effective on intra-domain matching cannot solve the problem of cross-domain matching. But with the proposed method, most of query images can obtain its right matches. (2) For matching time, the approach using vocabulary tree has the lowest efficiency and the average matching time of the proposed algorithm is the shortest which takes 600ms per query execution. In other words, the proposed method can matches two times in a second. That indicates that our algorithm based on ordinary hardware configuration is significant to meet the demands of real-time image processing.

Moreover, as figures display, the comparing methods only can correctly match for a few times. This has not been of statistical significance to visible-infrared image matching.

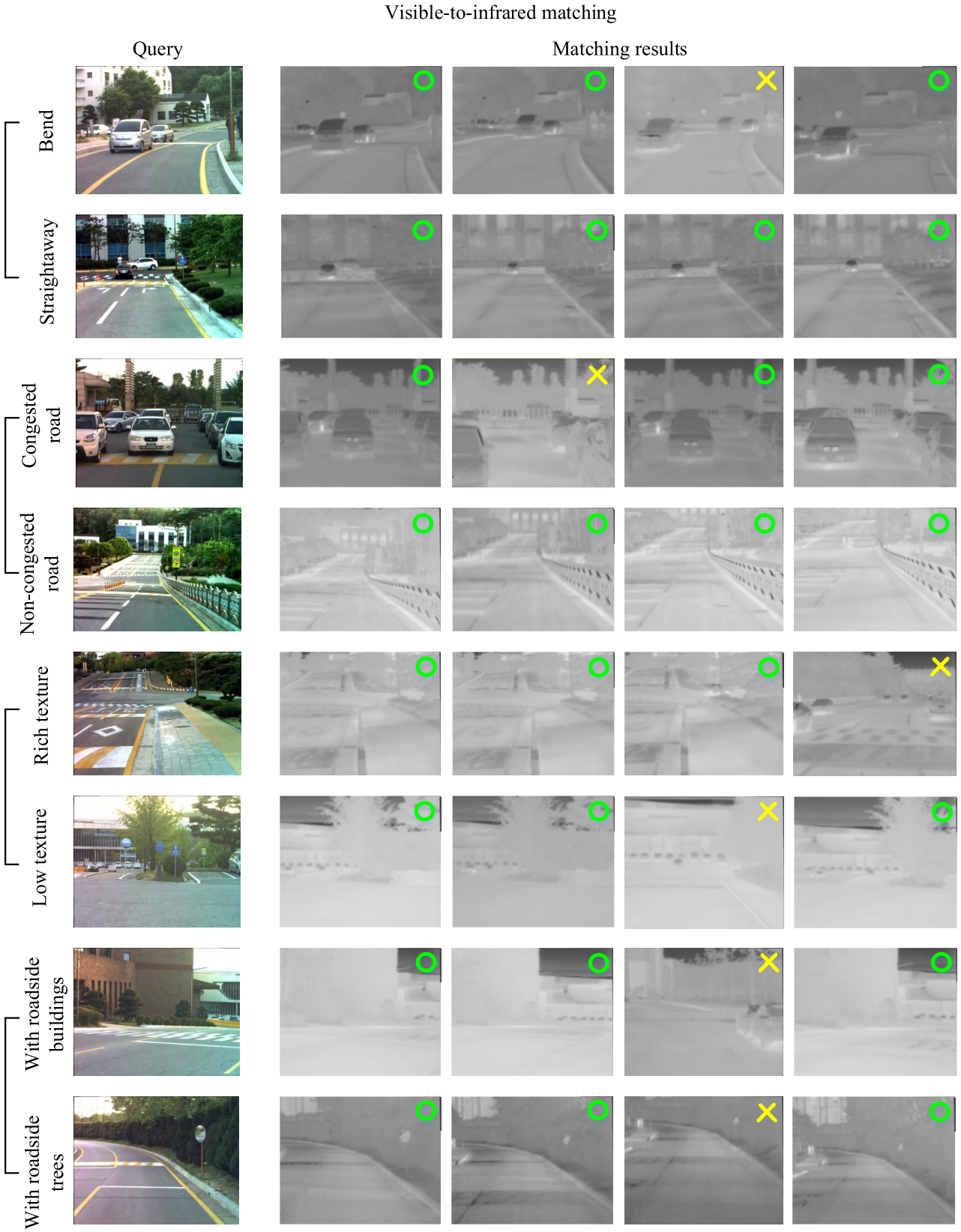



**FIGURE 9.** Some examples of visible-to-infrared matching. The query visible images are in the first column and the following images are matching results. The similarity score of each result is reduced in turn from left to right. Infrared images marked by a yellow cross mean that they are irrelevant with its corresponding query, and the results with a green circle at the top right corner have the same scene with the query.

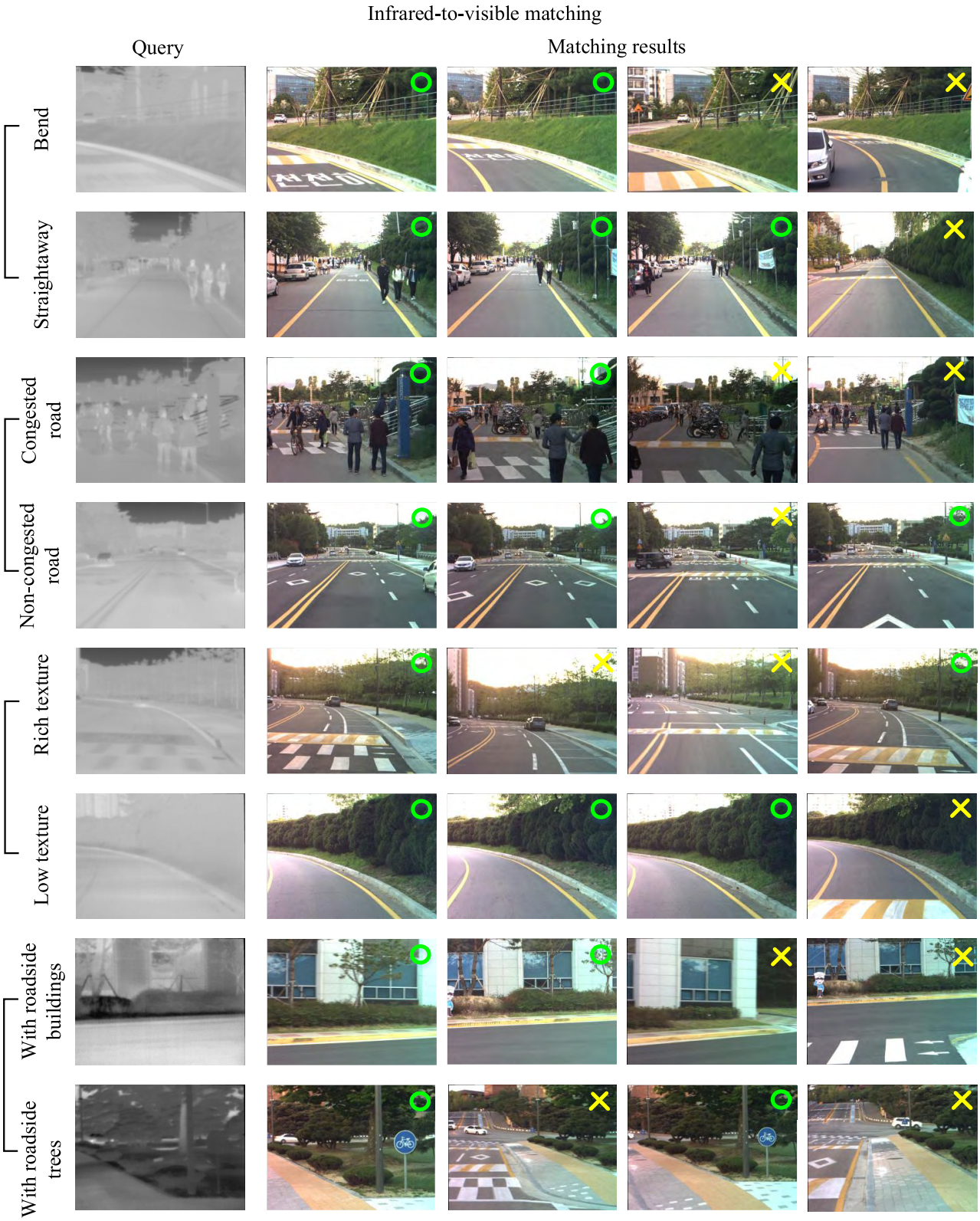



**FIGURE 10.** Some examples of infrared-to-visible matching. The infrared images in the first column are query images and the matching results are shown in the other columns. From left to right, the similarity score of each result is reduced in turn. The results with a green circle in the top right are the relevant images, whereas images marked by cross indicate that they are not from the same scene as the query.

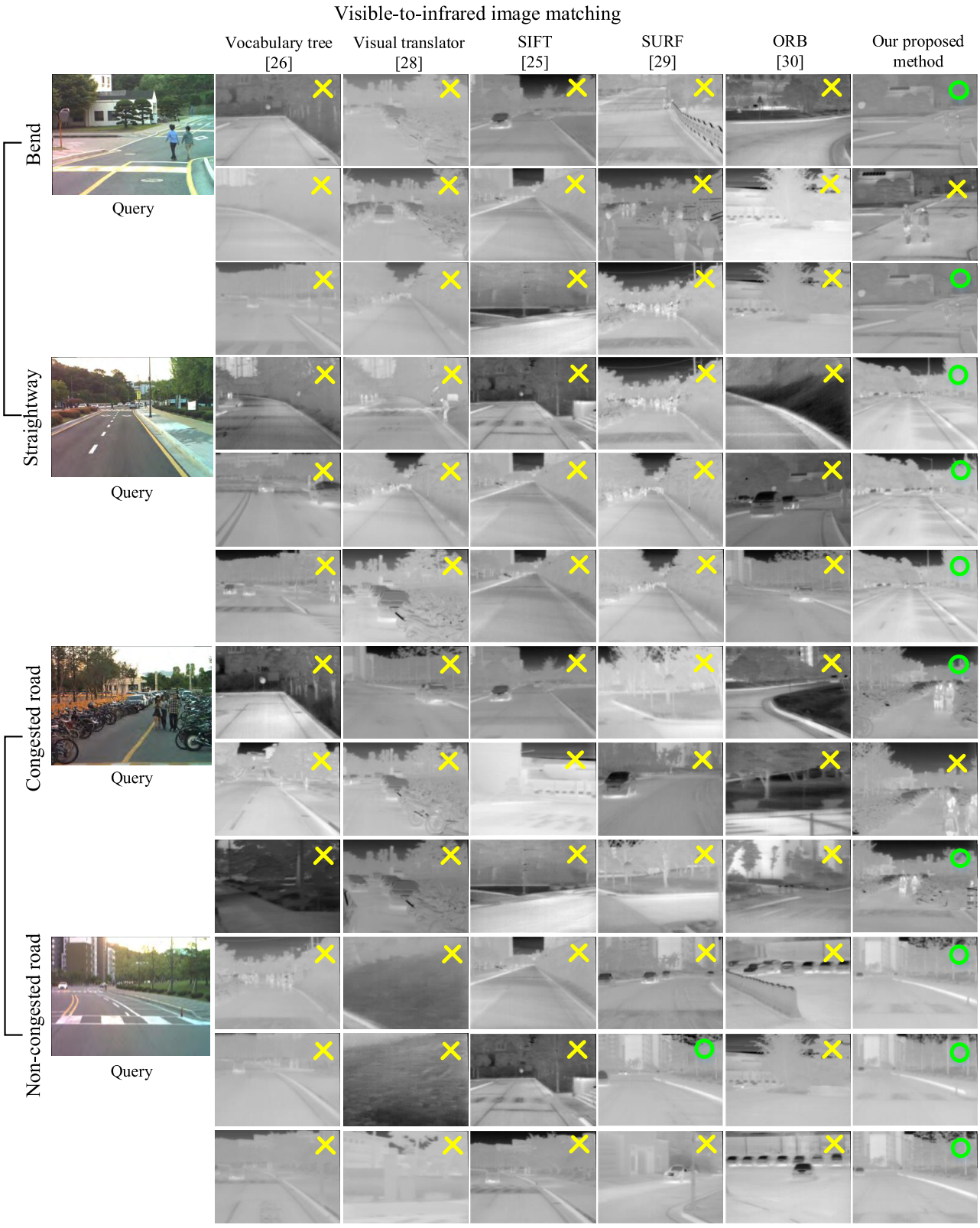



**FIGURE 11.** Qualitative comparison of our approach against the other contrast methods for visible-to-infrared matching. We evaluate the performance on two traffic situations (bend and straightway, congested road and non-congested road). Yellow cross: irrelevant image with the query. Green circle: the correct matches which are from the same scene as the query.


**FIGURE 12.** Qualitative comparison of our approach against the other contrast methods for infrared-to-visible matching. We evaluate the performance on two traffic situations (road with rich texture or not, road with roadside buildings or trees). Yellow cross: irrelevant image with the query. Green circle: the correct matches which are from the same scene as the query.

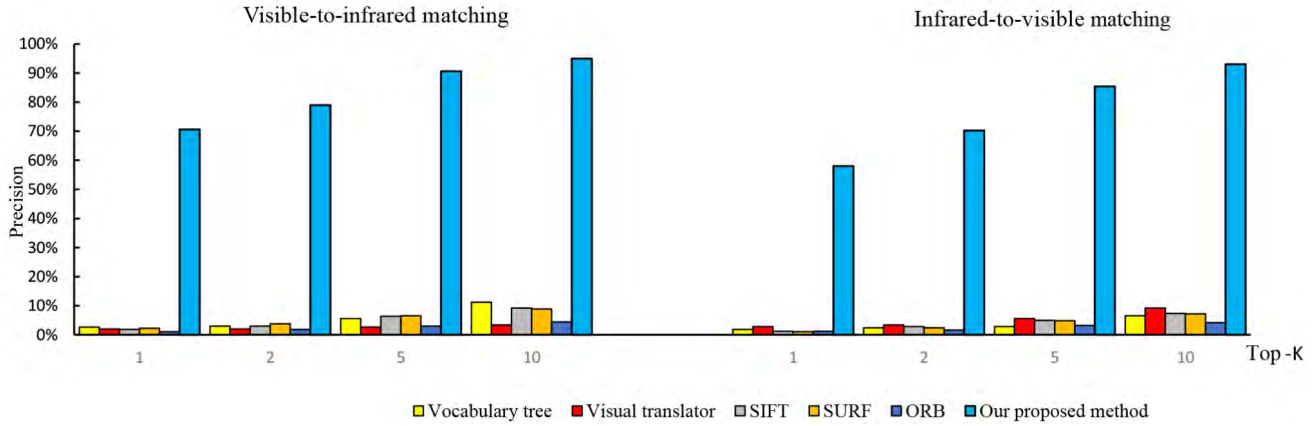



**FIGURE 13.** Quantity comparison of our approach against the other contrast methods for visible-to-infrared matching and infrared-to-visible matching.The performance of our approach which is shown with blue rectangle far better than the other methods. And the Top-10 precision of the proposed approach is more than 90 percent on both matching directions.

Due to the poor performance, we think that they do not have ability to solve the matching between visible image and infrared image. On the other hand, our approach achieves great precision on both matching directions. That verifies the proposed method not only able to realize visible-infrared matching but also shows satisfying result.

The high performance of our approach is due to cross-domain co-occurring feature, which well links up the visible domain and infrared domain. Firstly, the vocabulary tree method directly treats visible image and infrared image as one kind of image. SIFT, SURF and ORB immediately match these cross-domain features. These approaches do not consider the huge difference between visible feature and infrared feature. Its precision of cross-domain image matching is naturally low. But the proposed method respects the difference of the two kinds of features and employs co-occurring features to establish the symbiotic relationship between visible domain and infrared domain. Then, as for visual translator, though it is proposed for cross-domain image matching tasks, it still exists its limitation. This method requires a certain repeatability on keypoint detection and it is suitable for cross-domain image with certain structural similarity. Visible image and infrared image appear quite different in image structure. Visual translator directly detects feature point which may lead to a wrong visual translator. So this translator cannot build a correct bridge between the two visual domains and unable to match. However, the proposed method takes more account of these differences and using the strategy of image blocking to build a connection of the two features. So it achieves far better performance than the comparing method. Finally, for the matching time, vocabulary tree uses high-dimensional feature vector to represent the image and it will cost more time on the matching process. This leads to low efficiency. Visual translator only takes a little longer matching time than our approach, which is because its SIFT feature detection. The keypoint localization of our approach is the center of each image block that is much time-saving than visual translator. To sum up, for visible-infrared image matching, our approach based on cross-domain co-occurring feature is more accurate and more efficient.

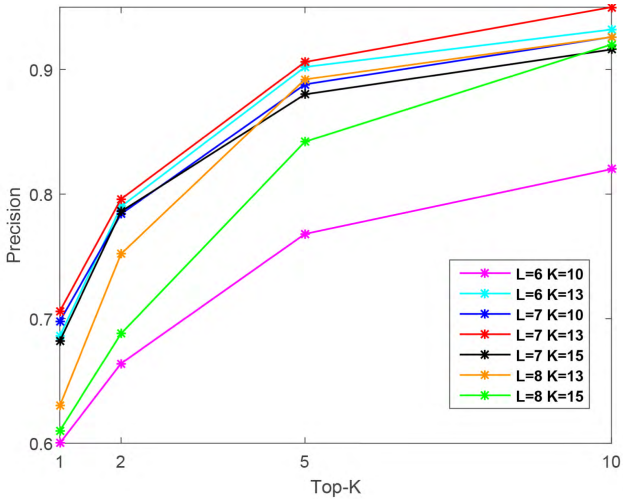



**FIGURE 14.** Precision curves under different parameters of vocabulary tree: tree depth L and branch number K. When the tree depth L is 7 and branch number K is 13 (as shown with a red line), matching performance is best.

### 2) THE EFFECT OF VOCABULARY TREE PARAMETER

Since the performance of visible-to-infrared matching and infrared-to-visible matching is similar, we take visible-to-infrared matching as an example to discuss the effect of vocabulary tree parameter on matching performance. There are two parameters in vocabulary tree: tree depth $L$ and branch number $K$. The precision curves by varying the two parameters are given in Fig. 14.

From this figure, we find that when the parameters are increased, the performance is increased at the beginning. But when the parameters are set too big, the precision is dropped. That is caused by the index file which is attached with leaf node. When the parameters are small, the number of leaf node is small. The number of image feature is same, so one

index file will contains too many features. That increases the effect of unrelated features and the matching precision is naturally influenced. And if the number of leaf node is too large, some of the index file maybe empty and the similar feature will be missed which makes the accuracy lower. After a great deal of experiments and analysis, the best parameter is $L = 7, K = 13$ with 3000 visible-infrared image pairs in training database. Moreover, for other training database scale, the performance will be better when each index file contains 5-10 features.

## IV. CONCLUSION

This paper proposes a novel visible-infrared image matching approach based on cross-domain co-occurring feature. In the proposed approach, we exploit cross-domain co-occurring feature to explore the inherent relationship between visible domain and infrared domain. For co-occurring feature construction, we firstly establish a cross-domain image database which is composed of some one-to-one visible-infrared image pairs. Then through feature extraction, cross-domain co-occurring feature is constructed by connecting a visible feature and its corresponding infrared feature in series. With these features, a visible vocabulary tree, an infrared vocabulary tree and a co-occurring vocabulary tree can be built, respectively. Finally, on the basis of co-occurring features and three vocabulary trees, a visible-infrared image matching system is successfully executed. Extensive experimental results on a public database (the KAIST All-day Place Recognition Database) confirm that co-occurring feature is effective and efficient for visible-infrared image matching. In addition, it can solve the matching problem bidirectionally. With the same co-occurring features, users can match visible images with a query infrared image or use a visible image to match infrared images. Furthermore, the current work shows great potential on various visual domains. In the future work, we will consider to extend our work to multiple fields and develop it to general.

## ACKNOWLEDGMENT

The authors give thanks to Erkang Li and Nianzeng Zhang from Xidian University for data collection and analysis. All authors also appreciate the careful and valuable comments of the reviewers to improve the quality of this paper.

## REFERENCES

[1] H. Zheng, S. Li, Y. Shao, and S. Yang, "Typical building of multi-sensor image feature extraction and recognition," in *Proc. Int. Conf. Artif. Intell. Sci. Technol.*, 2017, pp. 259–272.

[2] J. Son, S. Kim, and K. Sohn, "A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8830–8839, 2015.

[3] Y. Xu, J. Zhou, and L. Zhuang, "Binary auto encoding feature for multi-sensor image matching," in *Proc. Int. Conf. Ubiquitous Positioning, Indoor Navigat. Location Based Services*, 2016, pp. 278–282.

[4] W. Sun, X. Zhang, Z. Zhang, and R. Zhu, "Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb," *Spectrochimica Acta, A Mole. Biomole. Spectrosc.*, vol. 171, pp. 72–79, Jan. 2017.

[5] C. Wei, B. Zhou, and W. Guo, "A three scale image transformation for infrared and visible image fusion," in *Proc. Int. Conf. Inf. Fusion*, 2017, pp. 1–6.

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

[7] A. Apatean, A. Rogozan, and A. Bensrhair, "Objects recognition in visible and infrared images from the road scene," in *Proc. IEEE Int. Conf. Autom., Quality Testing, Robot.*, May 2008, pp. 327–332.

[8] M. Liu *et al.*, "Scene recognition for indoor localization using a multi-sensor fusion approach," *Sensors*, vol. 17, no. 12, pp. 2847–2867, 2017.

[9] F. Andert and S. Krause, "Optical aircraft navigation with multi-sensor SLAM and infinite depth features," in *Proc. Int. Conf. Unmanned Aircraft Syst.*, 2017, pp. 1030–1036.

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

[11] S. Ghosh, T. I. Dhamecha, R. Keshari, and R. Singh, "Feature and keypoint selection for visible to near-infrared face matching," in *Proc. IEEE Int. Conf. Biometrics Theory, Appl. Syst.*, Sep. 2015, pp. 1–7.

[12] K. J. Dana and P. Anandan, "Registration of visible and infrared images," in *Proc. Opt. Eng. Photon. Aerosp. Sens.*, 1993, pp. 2–13.

[13] E. Coiras, J. Santamaria, and C. Miravet, "A segment-based registration technique for visual-IR images," *Opt. Eng.*, vol. 39, no. 1, pp. 282–289, 2000.

[14] L. Zang and J. Wang, "Infrared and visible light image fast registration based on mutual information," *Infr. Laser Eng.*, vol. 37, no. 1, pp. 164–168, 2008.

[15] L. Bai, J. Han, Y. Zhang, and Q. Chen, "Registration algorithm of infrared and visible images based on improved gradient normalized mutual information and particle swarm optimization," *Infr. Laser Eng.*, vol. 41, no. 1, pp. 248–254, 2012.

[16] J. Jing and X. Zhang, "Multi-resolution registration of visible and infrared imagery by maximization of mutual information," in *Proc. Int. Conf. Adv. Comput. Control*, 2010, pp. 120–123.

[17] Y. Zhuang, K. Gao, X. Miu, L. Han, and X. Gong, "Infrared and visual image registration based on mutual information with a combined particle swarm optimization—Powell search algorithm," *Optik Int. J. Light Electron Opt.*, vol. 127, no. 1, pp. 188–191, 2016.

[18] J. Ma, J. Jiang, C. Liu, and Y. Li, "Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration," *Inf. Sci.*, vol. 417, pp. 128–147, Nov. 2017.

[19] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[20] K. Arguelwar and S. V. Jain, "Review on matching infrared face images to optical face images using LBP," *Int. J. Adv. Comput. Res.*, vol. 4, no. 17, pp. 950–955, 2014.

[21] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Int. Conf. Mach. Learn.*, 2003, pp. 856–863.

[22] C. Cunjian and A. Ross, "Matching thermal to visible face images using hidden factor analysis in a cascaded subspace learning framework," *Pattern Recognit. Lett.*, vol. 72, pp. 25–32, Mar. 2016.

[23] V. E. Liong, J. Lu, Y. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[24] Y. Choi, N. Kim, K. Park, S. Hwang, J. S. Yoon, and I. S. Kweon, "All-day visual place recognition: Benchmark dataset and baselines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 1–9.

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

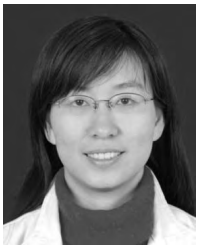
[26] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2161–2168.

[27] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 1470.

[28] J. Li, C. Li, T. Yang, and Z. Lu, "A novel visual vocabulary translator based cross-domain image matching," *IEEE Access*, vol. 5, pp. 23190–23203, 2017.

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

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 2564–2571.
[31] A. A. Sima and S. J. Buckley, "Optimizing SIFT for matching of short wave infrared and visible wavelength images," *Remote Sens.*, vol. 5, no. 5, pp. 2037–2056, 2013.
[32] X. Li and N. Aouf, "SIFT and SURF feature analysis in visible and infrared imaging for UAVs," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst.*, Aug. 2014, pp. 46–51.
[33] B. Besbes, A. Rogozan, A. M. Rus, A. Bensrhair, and A. Broggi, "Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF," *Sensors*, vol. 15, no. 4, pp. 8570–8594, 2015.
[34] Y. Chen, X. Zhang, F. Li, and Y. Zhang, "Multi-modal image registration based on modified-SURF and consensus inliers recovery," in *Proc. Image Graph.*, 2017, pp. 612–622.



**JING LI** (M'14) received the Ph.D. degree in control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 2008. She was a Visiting Scholar with the University of Delaware, USA, from 2013 to 2014. Before that, she was a Visiting Scholar with the National Laboratory of Pattern Recognition, Beijing, China, from 2004 to 2005, and also a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, in 2008. She is currently an Associate Professor with the School of Telecommunications Engineering, Xidian University, Xi'an, China. She is also serving as the Leader of the Intelligent Signal Processing and Pattern Recognition Laboratory. She has authored over 50 research papers in international journals and conference proceedings in the areas of computer vision and pattern recognition. Her research interests include image registration, matching and retrieval, and video content analysis and understanding.

**CONGCONG LI** received the B.S. degree from the School of Electronic Information Engineering, Hebei University, China, in 2015. She is currently pursuing the M.S. degree with the School of Telecommunications Engineering, Xidian University, Xi'an, China. Her research interests include cross-domain image matching and visual searching.

**TAO YANG** (M'13) received the Ph.D. degree in control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 2008. He was a Post-Doctoral Fellow with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University, from 2008 to 2010. Before that, he was a Visiting Scholar with the Intelligent Video Surveillance Group, National Laboratory of Pattern Recognition, Beijing, China, from 2004 to 2005, also a Research Intern with the FX Palo Alto Laboratory, Inc., Palo Alto, CA, USA, from 2006 to 2007, and also a Microsoft Research Asia. He was a Visiting Scholar with the University of Delaware, USA, from 2013 to 2014. He is currently a Full Professor with the School of Computer Science, Northwestern Polytechnical University. His research interests include video content analysis and understanding and image and video registration. He has published over 50 research papers in these fields. He is serving as a reviewer to numerous international journals, conferences, and funding agencies.

**ZHAOYANG LU** (SM'14) received the bachelor's, master's, and Ph.D. degrees in communication and information systems from Xidian University, in 1982, 1985, and 1990, respectively. He is currently a Full Professor with the School of Telecommunications Engineering, Xidian University, Xi'an, China. He has written and co-authored over 100 papers. He also holds over 10 patents in the field of pattern recognition and image processing. His current research interests include image matching and recognition and video content analysis and understanding.