





Article

# Multi-Level Features Extraction for Discontinuous Target Tracking in Remote Sensing Image Monitoring

Bin Zhou <sup>1,2,3,\*</sup> , Xuemei Duan <sup>1</sup>, Dongjun Ye <sup>1</sup>, Wei Wei <sup>4,\*</sup>, Marcin Woźniak <sup>5</sup> , Dawid Połap <sup>5</sup>  and Robertas Damaševičius <sup>5</sup> 

<sup>1</sup> School of Sciences, Southwest Petroleum University, Chengdu 610500, China; duan\_xuem@163.com (X.D.); dongjunye@yeah.net (D.Y.)

<sup>2</sup> Institute of Artificial Intelligence, Southwest Petroleum University, Chengdu 610500, China

<sup>3</sup> Research Center of Mathematical Mechanics, Southwest Petroleum University, Chengdu 610500, China

<sup>4</sup> School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

<sup>5</sup> Institute of Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland; marcin.wozniak@polsl.pl (M.W.); Dawid.Polap@polsl.pl (D.P.); Robertas.Damasevicius@polsl.pl (R.D.)

\* Correspondence: binzhou@swpu.edu.cn (B.Z.); weiwei@xaut.edu.cn (W.W.)

Received: 16 October 2019; Accepted: 4 November 2019; Published: 7 November 2019



**Abstract:** Many techniques have been developed for computer vision in the past years. Features extraction and matching are the basis of many high-level applications. In this paper, we propose a multi-level features extraction for discontinuous target tracking in remote sensing image monitoring. The features of the reference image are pre-extracted at different levels. The first-level features are used to roughly check the candidate targets and other levels are used for refined matching. With Gaussian weight function introduced, the support of matching features is accumulated to make a final decision. Adaptive neighborhood and principal component analysis are used to improve the description of the feature. Experimental results verify the efficiency and accuracy of the proposed method.

**Keywords:** feature; tracking; WMSNs; matching; weight; multi-level

## 1. Introduction

Remote sensing technology and wireless multimedia sensor networks (WMSNs) have been widely used in various fields of the national economy and are able to collect lots of data such as video and audio streams, still images, and scalar sensor data from the environment. It has been one of the most interesting research fields in the past few years [1–3].

Remote sensing monitoring is the remote observation of the characteristics or phenomena of the target through monitoring devices such as infrared detectors, multimedia sensors, and some other electronic or optical instruments. It means monitoring and analyzing a target/phenomenon without directly contacting the target/phenomenon when collecting information. Remote sensing technology can be used to quickly locate the ecological environmental pollution sources or other interested targets [4,5].

WMSNs are often composed of many wirelessly interconnected devices such as low-cost hardware CMOS cameras, microphones, and other sensor nodes with computational and wireless sensing capabilities; they can help to complete various tasks in remote sensing [2,6,7].

With the advances in wireless and electronic technologies, a variety of intelligent systems based on WMSNs have been developed for target tracking, behavioral analysis, identification, traffic surveillance, healthcare monitoring, environment monitoring, and so on [8,9]. Many techniques are presented to address these tasks based on video sequence analysis [10–12], and they are advantageous in ideal urban scenes with abundant computational ability.

A supervised learning framework is presented to generate compact and bit-scalable hashing codes directly from raw images [13]. Then, the deep convolutional neural network is utilized to train the model with the image features and hash functions simultaneously optimized. [10] proposes a novel network named part-based convolutional baseline (PCB) based on a convolutional descriptor consisting of several part-level features and a refined part pooling method for person retrieval. [14] presents a new object tracking approach for surveillance applications developed using a big data model based on graphs and a multilevel fusion. With enough quantity of pose-rich samples generated from the original image and skeleton samples, a novel unsupervised pose augmentation cross-view scheme is proposed for person re-identification [12]. In [15], an improved method is developed to detect and track multiple heads by considering them as rigid body parts for real-time video surveillance. The appearance model of human heads is updated according to the fusion of color histogram and oriented gradients. Acoustic and image hybrid wireless multimedia sensors' networks are introduced to trajectory prediction for target tracking [16].

However, there are still some tasks must be completed in special scenes with limited computational ability or power support, such as wildlife monitoring and tracking [1,17]. Remote sensing image monitoring based on an optimized convolutional neural network model is introduced to the conservation of rare wild animals [1]. A hierarchical wireless sensor network is installed in Doñana National Park to collect information about animals' behaviors. The placed intelligent devices contain a neural network implementation to classify the animals' behavior [17]. A novel energy-efficient object detection based on the image transmission approach is proposed for wireless multimedia sensor networks [18].

Similar contents in different images can be found by analyzing the pixel values and their potential features, named features' detection and image matching [19–21]. Traditional features' detection approaches are started with Harris's corner detection and Forstner's work on the fast operator for precise location of distinct points [22,23]. In addition, various methods and algorithms are then developed with different comprehensive understanding of the features [21,24,25].

Distinctive invariant features can be extracted to perform reliable matching between different views of an object or scene [20], and this named scale-invariant feature transform technology has been widely applied in lots of computer vision tasks. The principal component analysis is introduced to describe the feature points and the dimension-reduced descriptor can hold some advantageous in robustness or computation [26]. By relying on integral images for image convolutions and using a Hessian matrix-based measure, a novel scale-and rotation-invariant interest point detector and descriptor (namely Speeded Up Robust Features, SURF) is presented and it approximates or even outperforms previously proposed methods respected to robustness and computation [27]. With the maximum similarity measure defined in terms of geometric and photometric properties of regions, a hierarchical image matching based on a tree matching problem is presented to identify the largest similar part [28]. A practical method is proposed to establish dense correspondences between two images with similar content, but possibly different 3D scenes [29].

In the traditional scenes, with the sufficient support of computation ability and power, enough image sequences are easy to be captured and convenient to be analyzed. Many previous methods are presented based on this ideal situation. However, sometimes, the targets move from one camera to another, and very short video or a few images can be captured for the limited environment condition or device condition. In addition, the relation among the videos or images may be loose or not so continuous. In this case, feature detection should be implemented more efficiently to finish the task with less computation and power.

In recent years, convolution neural networks and machine learning have been rapidly developed in many fields. Large scale databases are often applied to training various deep and structural models. Though a considerable efficiency achieved in the scenes covered by these data, it may be also mean that it is more difficult to adapt some scenes with unexpected data. Furthermore, it is not easy to collect enough data in any scenes. Thus, it is still necessary to address some tasks via interpretable methods

with some mathematical base. In this paper, a novel multi-level features extraction is approached for discontinuous target tracking in remote sensing image monitoring. Multi-level features of the reference image are pre-extracted. The rough features are used to exclude the obvious error targets. In addition, the refined features are used to compare with the rest candidate target. Adaptive neighborhood and the principal component analysis (PCA) are used to describe the feature. The weighted support of matching features will be accumulated based on a Gaussian function to make the final decision.

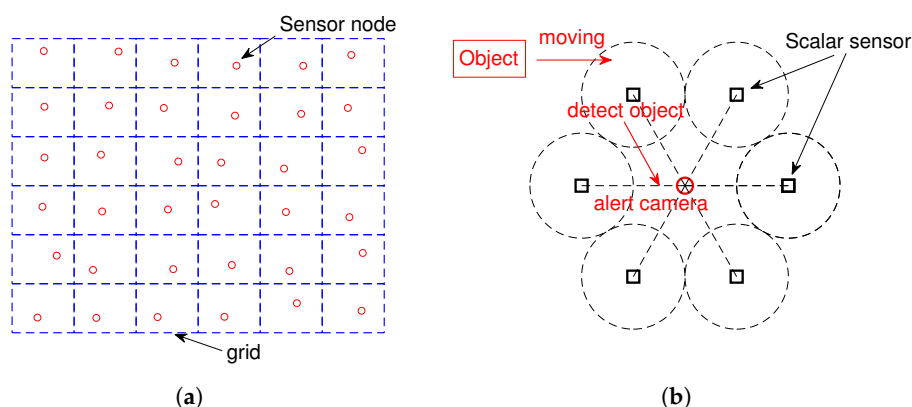
The rest of this paper is organized as follows. The related fundamentals of remote sensing monitoring, target tracking, and features extraction are prepared in Section 2. The multi-level features extraction and discontinuous targets tracking are proposed in Section 3. Several experiments are implemented in Section 4 to verify the accuracy and efficiency of the proposed method.

## 2. Fundamentals

### 2.1. Remote Sensing and Wireless Multimedia Sensor Networks

There are many techniques, such as photography, infrared scanning, correlation spectroscopy, lidar detection, and unmanned aerial vehicles, that can be used to achieve remote sensing monitoring [30,31]. Remote sensing cameras can be remotely monitored by installing them on a flying device or on a satellite to capture targets on the ground, vegetation, and plant emissions. The principle of remote sensing technology is that the reflection characteristics of electromagnetic waves are often not the same due to different objects or phenomena, and photographs of different colors or tones can be obtained by photosensitive recording of the photosensitive film. In some cases, the surveillance area can be deployed with a wireless distributed sensor network consisting of a set of multimedia sensor nodes, so-called wireless multimedia sensor networks. These nodes are connected or connected to the main gateways using a wireless communication protocol.

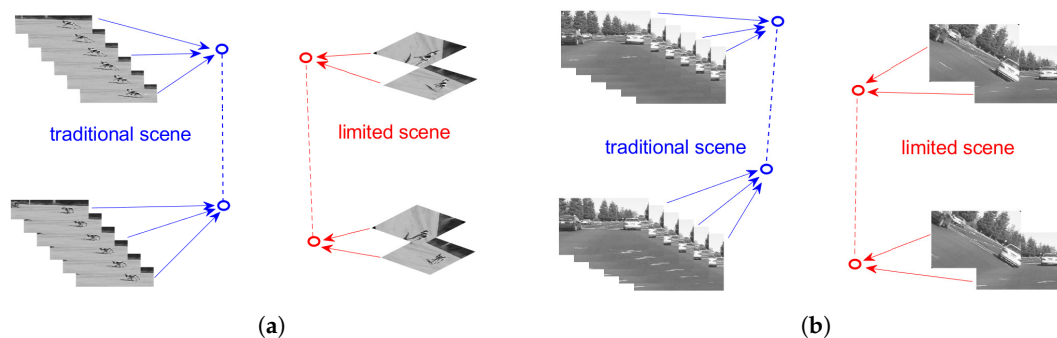
Suppose there are  $N$  sensor nodes deployed in a square surveillance zone that is divided into  $n \times n$  grids. The grid approach is commonly used to monitor the entire area without leaving gaps between the sensor nodes [32]. Each sensor node is fixed on a position  $(x_s, y_s)$ , and there are several scalar sensors (such as seismic and acoustic) that are deployed around for detecting moving targets and awaken the camera sensor. As an object enters the grid area, it will be first detected by a scalar sensor. Then, the camera sensor will be awakened and try to take a short video or some images according to the position of alerting the scalar sensor. Figure 1 shows the topology and workflow of wireless multimedia sensor networks.



**Figure 1.** Topology and workflow of wireless multimedia sensor networks: (a) topology of the networks; (b) workflow of the networks.

Our goal is to track a moving target using such sensor networks composed of camera sensor nodes and scalar sensors.

Different from traditional urban scenes, the computation ability and power support are limited. The cameras only work as they are awakened and very short videos or very few images can be captured for recognition and tracking. The multimedia data may be linked to several objects moving, and it means that the videos or images are not continuous in the spatial or time. It will be more difficult to recognize and track the target with such a discontinuous data. Figure 2 shows the difference between traditional scenes and limited scenes.



**Figure 2.** Illustration of a traditional and limited scenes: (a) Dog case; (b) Panda case. In both cases, a number of obtained frames can be seen.

In the traditional scene, with the powerful support of energy and computation ability, enough long video and lots of image sequences can be easily collected for the latter computing. Many popular deep learning methods can be applied to complete the tasks. However, in a limited scene, there are only a few short videos and images available.

## 2.2. Multi-Cam Tracking and Re-Identification

The research about disjoint cameras is started with Huang and Russell's work on Bayesian formulation. They use the formulation to estimate the posterior of predicting the appearance of objects in one camera given evidence observed in other camera views. Multiple spatial-temporal features such as color, vehicle length, height and width, velocity, and time of observation are all included in the appearance model [33,34].

The term "person re-identification" is first proposed by Zajdel, Zivkovic, and Krose [35] in the research about multi-camera tracking.

They aim to recognize a person when it leaves the field of view and re-enters later. A dynamic Bayesian network is defined to encode the probabilistic relationship between the labels and features (color and spatial-temporal cues).

After then, many technologies are developed to address this problem such as independence of re-ID (image-based), video-based re-ID, deep learning for re-ID, end-to-end image-based re-ID, and so on [34,36,37].

## 2.3. Feature Extraction

Feature extraction is one of the most important techniques in computer vision and many high-level applications must be implemented on it [21,26]. Earlier approaches are started with the detection of corner points or distinct points [23]. Based on the local auto-correlation function, Harris proposed a combined corner and edge detector to cater for image regions containing texture and isolated features [22].

An important milestone of feature extraction is the presentation of scale-invariant feature transform (SIFT) [20]. This local image features is proposed to develop an object recognition system, and it is invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection.

To reduce the 128-dimensional feature descriptor, principle component analysis is introduced to normalized gradient patch instead of using smoothed weighted histograms [26].

In addition, various methods and algorithms are then developed with different views to features [24,25,38]. Speeded Up Robust Features (SURF) is a novel interest point descriptor with scale-and rotation-invariance based on image convolution and Hessian matrix-based measure [27]. Some other methods are also presented by different principles such as geometric and photometric properties, dense correspondence, human visual system, dedicated sampling, and so on [29,39].

### 3. Proposed Method

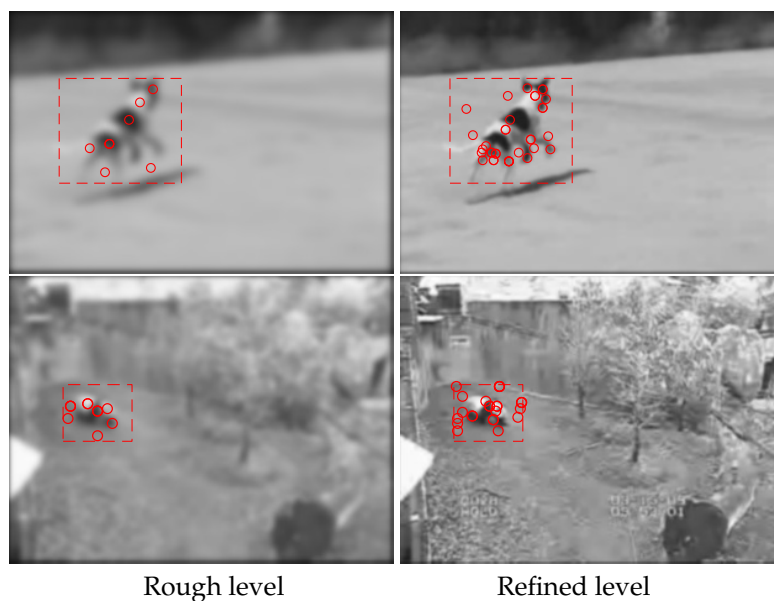
#### 3.1. Optimal Selection to the Principle Components

By optimizing a closed-world toy model, Gheissari et al. [40] addresses person re-identification based on single-image.  $G$  is assumed to be a gallery composed of  $m$  images, denoted as  $\{g_i\}_{i=1}^m$ . It means there are  $m$  different identities,  $1, 2, \dots, m$ . Given a probe image collected by WMSNs, its identity can be determined by

$$i^* = \operatorname{argmax}_i \operatorname{sim}(q, g_i), \quad (1)$$

where  $i^*$  means the decision and  $\operatorname{sim}(\cdot, \cdot)$  means the similarity function. In general, the similarity can be computed based on the image features such as SIFT, SURF, etc.

Multi-level features points of each image in the reference gallery are pre-extracted based on the classical SIFT as shown in Figure 3—the first column shows the rough feature points and the second one shows refined feature points.



**Figure 3.** Multi-level features points of reference images, where features are described as a red circle, and a detected object is surrounded by a rectangle. In the first column, the rough features are marked, and, in the second one, refined ones.

Assume that  $p_{i,j}^{(k)}, j = 1, 2, \dots, n_i$  denotes the  $k$ th level features points of image  $g_i$ . In this paper,  $k = 1$  means rough level and  $k = 2$  means refined level. Then, each  $b \times b$  local area centered a rough level point is reshaping to a row vector after necessary rotated normalization. All the vectors are arranged to a matrix denoted by  $H^{(k)}$  and apply principle component analysis on it. The main advantages of using principal component analysis are that the method is performed without supervision, so there is no need to have any information about classes during size reduction. As a result, the method indicates the dominant patterns in the analyzed sets.

Let  $H^{(k)}[i, j]$  denote the reshaped row vector of the normalized local area centered  $p_{i,j}$ . To well separate the multi-targets, a set of components to describe the feature points should be determined by

$$\begin{aligned} & \max_{\delta} \sum_{i_1 \neq i_2} \sum_{j_1, j_2} \|H^{(k)}[i_1, j_1] \text{diag}(\delta)V - H^{(k)}[i_2, j_2] \text{diag}(\delta)V\|^2 \\ & \text{s.t.} \\ & \delta = [\delta_1, \delta_2, \dots, \delta_{b^2}], \\ & \delta_s \in \{0, 1\}, s = 1, 2, \dots, b^2. \end{aligned} \quad (2)$$

Here,  $V$  denotes the components' matrix. The binary vector  $\delta$  means selection to the components. Maximizing the objective function means to maximize the distance between two different targets.

### 3.2. Image Matching via Refined Feature Describing

A probe image can be checked by model (1) with a proper similarity threshold set. If passed, it will be further matching via refined features.

Scale-invariant feature transform (SIFT) is a common descriptor widely applied in many computer vision problems such as image matching, object recognition, and so on. Though some classical improvements have been approached in recent years, it still is one of the most representative techniques to well describe the image features. Based on the feature points computed previously, the traditional SIFT can be implemented as three main steps:

- Step 1. Determine candidate key-points via peak selection in the difference of Gaussian space;
- Step 2. key-point checking and orientation assignment;
- Step 3. Eight direction statistics and key-point describing.

To make it work well in these limited WMSN scenes, we introduce adaptive neighborhoods to key-point checking and PCA to orientation assignment. This paper presents a novel frame for multi-targets tracking in limited WMSNs, and the feature extraction method can be directly replaced if necessary. For better performance of feature describing, some other techniques can also be introduced to these steps.

### 3.3. Evaluation of the Matching Results

Suppose that there are some feature points  $pos_{i,j}^{(k)}$  ( $j = 1, 2, \dots, m_i$ ) in the probe image  $q$  found to be matching the feature points  $p_{i,j}^{(k)}$  in a reference image  $g_i$ . The aggregation of  $pos_{i,j}^{(k)}$  and  $p_{i,j}^{(k)}$  can be measured independently to make a binary decision on the target tracking.

It is natural to introduce the Gaussian weight function to measure the aggregation of  $p_{i,j}^{(k)}$  in the reference image and regard it as similarity:

$$\text{sim}(q, g_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \exp\left(-\frac{r_j^2}{2\sigma^2}\right), \quad r_j = p_{i,j}^{(k)} - \frac{1}{m_i} \sum_{j=1}^{m_i} p_{i,j}^{(k)}, \quad (3)$$

where  $\sigma$  denotes a distance scale factor and  $r_j$  means the distance from  $p_{i,j}^{(k)}$  to their center.

However, the aggregation of  $pos_{i,j}^{(k)}$  in the probe image  $q$  is ignored and the above measurement can be improved as

$$\begin{aligned} \text{sim}(q, g_i) &= \frac{1}{m_i} \sum_{j=1}^{m_i} \exp\left(-\frac{r_j^2}{2\sigma^2}\right) + \frac{1}{m_i} \sum_{j=1}^{m_i} \exp\left(-\frac{s_j^2}{2\sigma^2}\right), \\ r_j &= p_{i,j}^{(k)} - \frac{1}{m_i} \sum_{j=1}^{m_i} p_{i,j}^{(k)}, \quad s_j = pos_{i,j}^{(k)} - \frac{1}{m_i} \sum_{j=1}^{m_i} pos_{i,j}^{(k)}, \end{aligned} \quad (4)$$



Then, the similarity can be easily applied to make a binary decision on target tracking.

$$\text{IsIdentifying}(q, g_i) = \begin{cases} 1, & \text{sim}(q, g_i) > \rho_0, \\ 0, & \text{sim}(q, g_i) \leq \rho_0. \end{cases} \quad (5)$$

Though in a discontinuous tracking scene, it is still assumed that a few images could help to determine the moving area. Then, there is a probability of whether a feature belongs to the target. It is similar to a matching. Several intensive matching means more probability of target identification. Equation (5) is introduced to determine the identification based on the concentration of the matching.

For more accurate computing, some other techniques can be introduced to improve the similarity such as adaptive weight function, distance metric learning, and so on.

#### 4. Experiments

In limited scenes, very few images can be collected to match test and target tracking. However, we still assume that there are two images of each moving object that can be captured by a camera each time. The object in the probe image can be located by the difference of the convolution with the Gaussian kernel. Figure 4 shows the object location computed from the two probe images. For well matching results, the local areas have been extended to include most feature points of the target and nearby surroundings.



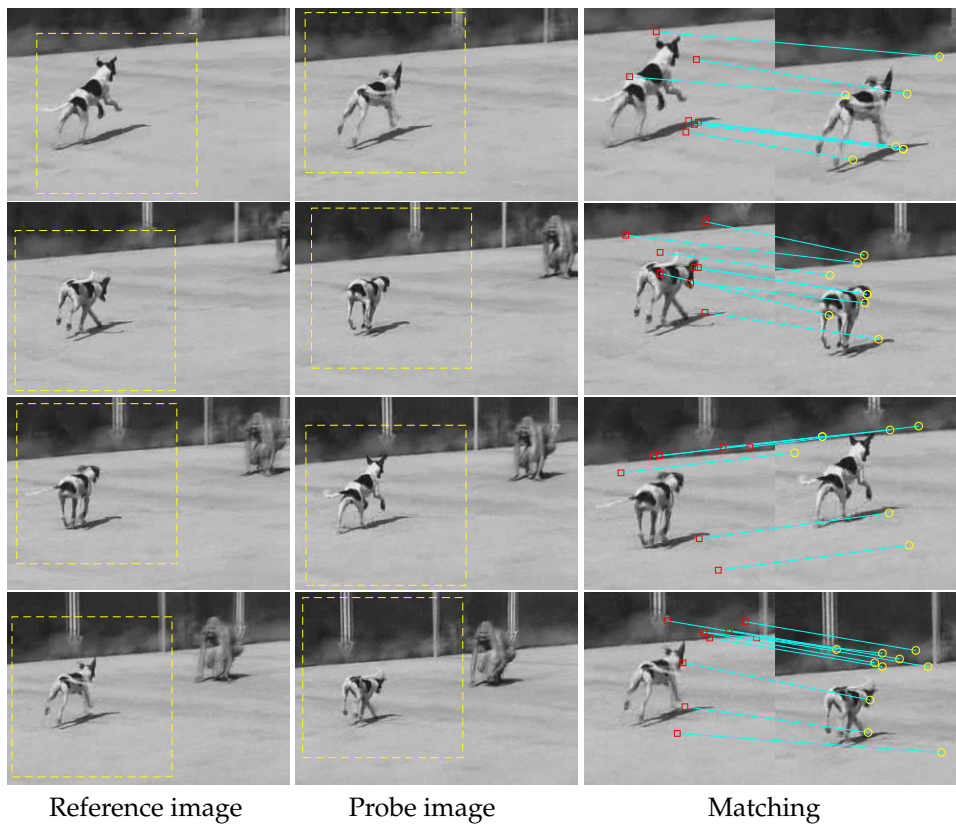
Figure 4. Sample images with marked moving objects in yellow rectangles.

We present several experiments to evaluate the performance of the proposed method. The first and second are implemented to explore the recognition ability of our method to discontinuous probe images. Each pair of probe images are selected from a continuous sequence of images and the interval between them is more than five frames. The third and last are used to explore the recognition ability of the proposed method to multi-targets tracking. The experiment data are downloaded from a Visual Tracker Benchmark (v1.0) [41]. All the experiments are completed under Windows 7 system with Matlab R2017b. The related parameters are set as follows. The features matching was determined traditionally by Euclidean distance and the ratio between the shortest one and the second shortest one. The ratio threshold is set to be 0.6 in this paper. The distance scale factor  $\sigma$  in Equation (4) is set to be  $\frac{d}{2}$  and  $d$  means the local image area size. The similarity threshold  $\rho_0$  is set to be 0.6.

Features matching test 1. Four pairs of images of a dog are selected from the Dog image sequence in OTB-100. Frames 1, 11, 18, and 31 are assumed to be reference images and frames 6, 16, 23, and 36 are regarded as probe images.

As shown in Figure 5, the left column means reference images with moving detection, the middle means probe images with moving detection, and the right column is the matching results. The rows

mean different image pairs and the detected moving areas are  $200 \times 200$ . It can be found that there are 7–9 matches in each pair. Some are related to the dog self and others are related to the surroundings. It is interesting that several means the correspondence of shadows.



**Figure 5.** First feature matching test. In the first two columns, moving objects are marked, and, in the last column, there are marked key-points (as red squares and yellow circles for better visualization) with an indication of their position on two different frames with a blue line.

Features matching test 2. Four pair images of a panda (shown in Figure 6) are selected from the Panda image sequence in OTB-50. Frames 1, 11, 23, and 31 are assumed to be reference images and frames 10, 22, 34, and 42 are regarded as probe images.

The columns from left to right mean reference images with moving detection, probe images with moving detection, and the matching results. The rows correspond to different image pairs.

It can be found that there are a few matches in each pair compared to experiment 1 because of the smaller local image area ( $61 \times 61$ ). Most of the matches are related to the panda self and very few related to the surroundings.

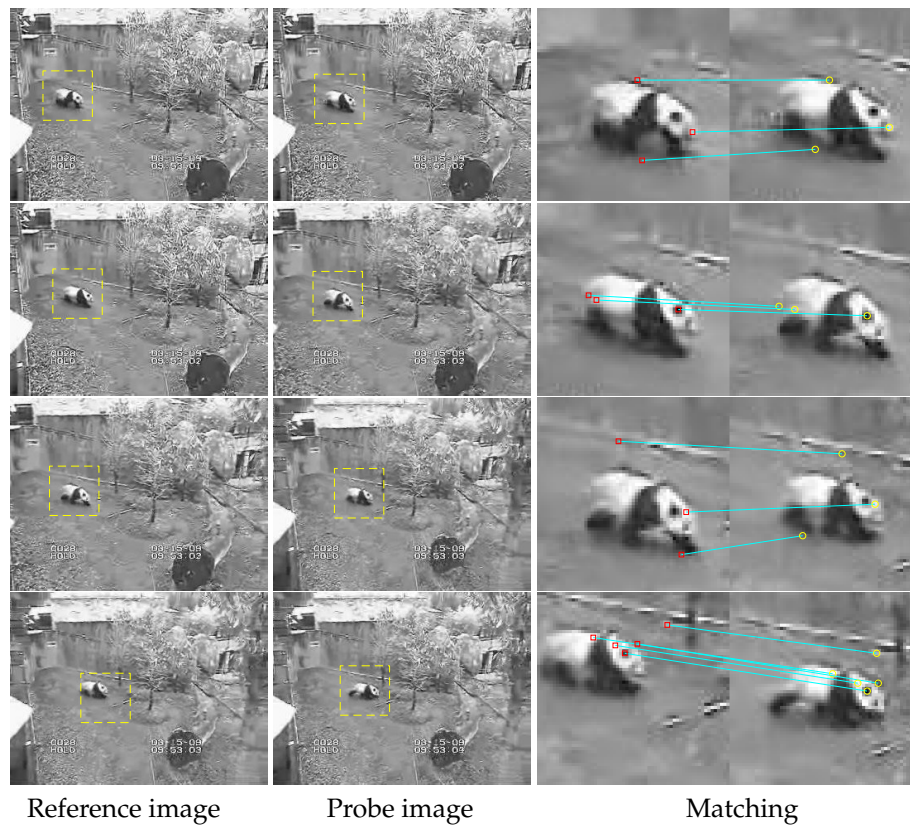
Target tracking test 1. Frame 1 is selected from the Dog image sequence and supposed to be a reference image. Frames 9, 16, 23, and 30 are regarded as probe images captured by different cameras.

As shown in Figure 7, the images in the first column are the same as frame 1, which is regarded as a reference image. Probe images with moving detection are shown in the second column. Matching results are shown in the last column.

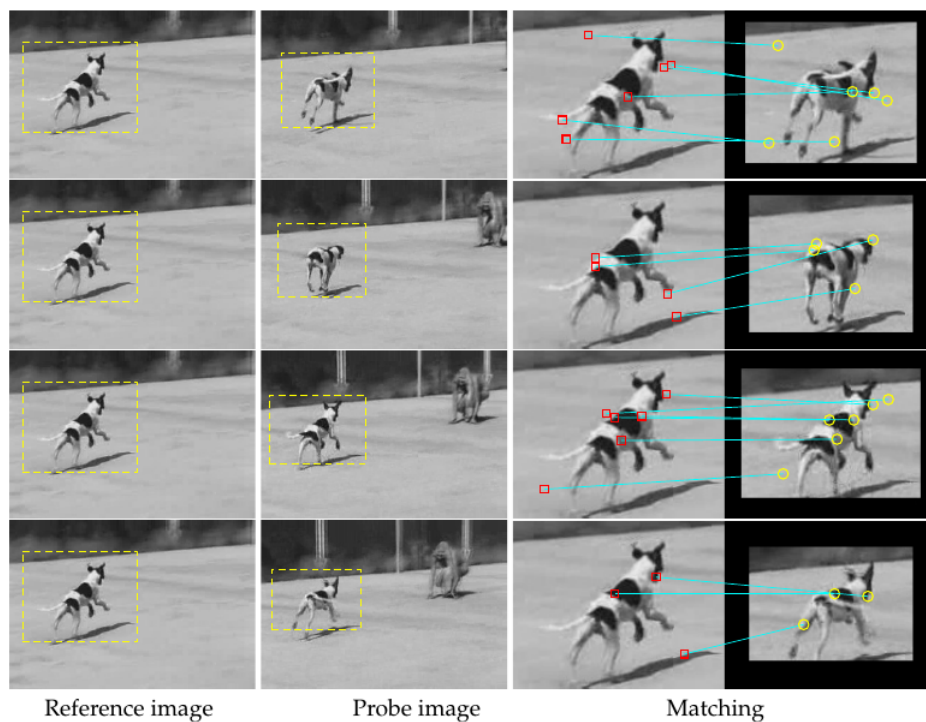
The detected moving areas are often in different size and they are adaptively computed in the tracking. It can be found that there are about five matches in each pair.

Target tracking test 2. Frame 1 is selected from the Panda image sequence and supposed to be a reference image. Similar to the above test, we select frames 9, 16, 23, and 30 to be probe images for tests, as shown in Figure 8. Different probe images with moving detection are shown in the middle column. Matching results are shown in the last column.

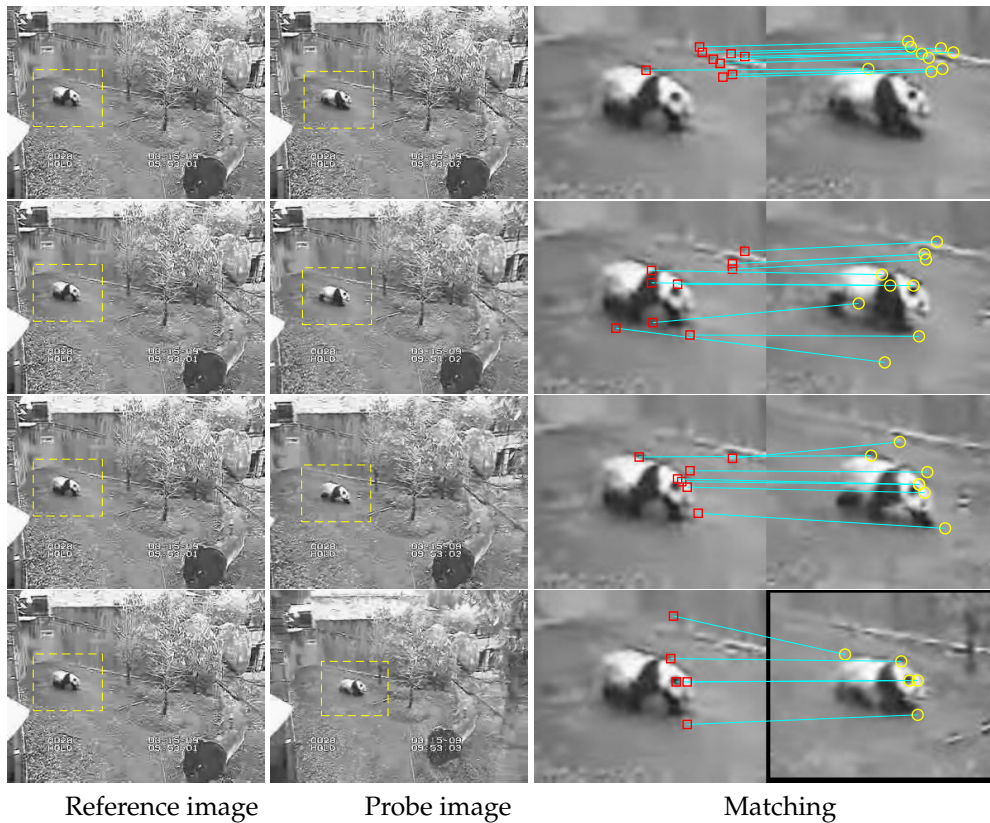




**Figure 6.** Second feature matching test. In the first two columns, moving objects are marked, and, in the last column, there are marked key-points (as red squares and yellow circles) with an indication of their position on two different frames with a blue line.

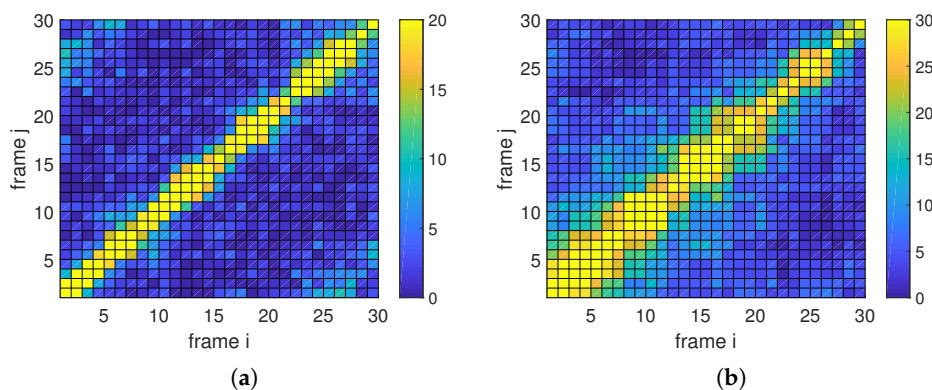


**Figure 7.** First target tracking test. In the first two columns, moving objects are marked, and, in the last column, there are marked key-points (as red squares and yellow circles) with an indication of their position on two different frames with a blue line.



**Figure 8.** Second target tracking test. In the first two columns, moving objects are marked, and, in the last column, there are marked key-points (as red squares and yellow circles) with an indication of their position on two different frames with a blue line.

More details about the matching relation of frame 1 to frame 30 are shown in Figure 9. The color value at position  $(i, j)$  means the matching number of frame  $i$  and frame  $j$ . It is found that more matching can be captured between a pair of close images, and there is a lot of matching that can be found in most of the image pairs. However, there is not yet matching that can be found in a few image pairs. Thus, the features describing and matching still should be improved for serious discontinuous targets tracking, although the proposed method has provided a solution to this problem in some sense.



**Figure 9.** Matching relations of discontinuous frames: (a) matching relation of Dog frames; (b) matching relation of Panda frames.

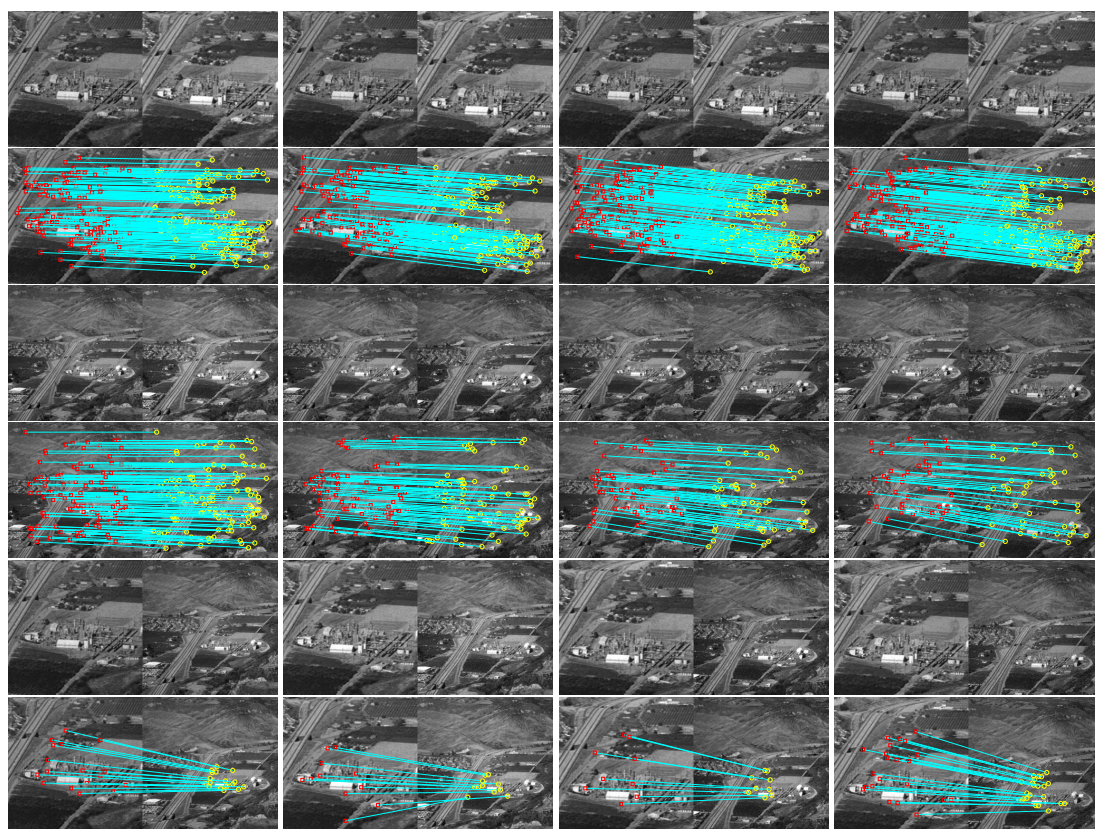


**Remote sensing image matching test.** There are two image sequences (sequence A and sequence B) from a scene. The camera views and scales are much different from each other. We try to explore the tracking ability of the proposed method from three points.

(1) Suppose frame 1 of sequence A to be a reference image and then frames 7, 13, 19, and 25 are selected to test the tracking ability. The moving area is set to be the whole image. Original image pairs are shown in the first row of Figure 10 and matching results are shown in the second row.

(2) Frame 1 of sequence B is supposed to be a reference image and then frames 3, 5, 7, and 9 are selected to test the tracking ability. Original image pairs and matching results are shown in the third and fourth row.

(3) Two images are selected independently from sequence A and sequence B to generate an image pair for discontinuous targets tracking ability test. Original image pairs and matching results are shown in the fifth and sixth row.



**Figure 10.** Remote sensing image matching test—the original frames are shown in the odd rows, and the even ones, frames with the found key-points (marked as yellow and red points) and their locations as a blue line.

It can be found from the results that more matching can be captured in (1) than (2) because of the short shooting distance. Though a significant scale difference between sequence A and sequence B can be found, there are still considerable matching that can be captured in each image pair crossed the wide-scale gap.

In these discontinuous targets tracking scenes, sequence analysis-based methods or learning-based methods are difficult to get to work well because of insufficient data. The methods based on single-image re-identification require considerable computation for features detection on each probe image and matching them to the reference images. However, sometimes, it is not necessary to introduce refined computing at first. Furthermore, little guarantee can be achieved for distinguishing different targets because of the independent features describing. Compared to the traditional methods, a two-stage procedure is introduced to reduce some unnecessary computation. In addition, the optimal

set of components based on PCA is applied to well distinguish the features from different targets. These contribute to the effectiveness of the proposed method.

## 5. Conclusions

In this paper, a multi-level features extraction is presented for discontinuous target tracking in remote sensing image monitoring. The features of reference images are extracted at different levels in advance. The rough-level features are used to discard the error target and refined-levels are used to target matching. Proper neighborhood can be set adaptively and principal component analysis is used to improve the descriptor. The weighted support of matching features can be accumulated to make the final decision. Experimental results verify the efficiency and accuracy of the proposed method.

**Author Contributions:** Conceptualization, B.Z., X.D. and M.W.; resources, W.W.; data curation, D.Y.; writing—original draft preparation, B.Z. and X.D.; writing—review and editing, D.P., M.W., and R.D.; visualization, D.Y.; supervision, B.Z.

**Funding:** This work is supported in part by the NSF of China (11226173, 11301414).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chen, G. Remote Sensing Image Monitoring and Recognition Technology for the Conservation of Rare Wild Animals. *Rev. Cient.* **2019**, *29*, 301–311.
2. Roopa, D.; Chaudhari, S. A survey on Geographic Multipath Routing Techniques in Wireless Sensor Networks. In Proceedings of the IEEE 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 15–16 March 2019; pp. 257–262.
3. Küçükkeçeci, C.; Yazıcı, A. Big data model simulation on a graph database for surveillance in wireless multimedia sensor networks. *Big Data Res.* **2018**, *11*, 33–43. [[CrossRef](#)]
4. Aguirre-Gutiérrez, J.; Seijmonsbergen, A.C.; Duivenvoorden, J.F. Optimizing land cover classification accuracy for change detection, a combined pixel-based and object-based approach in a mountainous area in Mexico. *Appl. Geogr.* **2012**, *34*, 29–37. [[CrossRef](#)]
5. Chen, G.; Weng, Q.; Hay, G.J.; He, Y. Geographic Object-based Image Analysis (GEOBIA): Emerging trends and future opportunities. *GISci. Remote Sens.* **2018**, *55*, 159–182. [[CrossRef](#)]
6. Akyildiz, I.F.; Melodia, T.; Chowdhury, K.R. A survey on wireless multimedia sensor networks. *Comput. Netw.* **2007**, *51*, 921–960. [[CrossRef](#)]
7. Abbas, N.; Yu, F.; Fan, Y. Intelligent Video Surveillance Platform for Wireless Multimedia Sensor Networks. *Appl. Sci.* **2018**, *8*, 348. [[CrossRef](#)]
8. Amjad, M.; Rehmani, M.H.; Mao, S. Wireless multimedia cognitive radio networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 1056–1103. [[CrossRef](#)]
9. Usman, M.; Jan, M.A.; He, X.; Chen, J. A mobile multimedia data collection scheme for secured wireless multimedia sensor networks. *IEEE Trans. Netw. Sci. Eng.* **2018**. [[CrossRef](#)]
10. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
11. Yu, H.X.; Zheng, W.S.; Wu, A.; Guo, X.; Gong, S.; Lai, J.H. Unsupervised Person Re-identification by Soft Multilabel Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2148–2157.
12. Li, M.; Zhu, X.; Gong, S. Unsupervised Tracklet Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [[CrossRef](#)]
13. Zhang, R.; Lin, L.; Zhang, R.; Zuo, W.; Zhang, L. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Trans. Image Process.* **2015**, *24*, 4766–4779. [[CrossRef](#)]
14. Küçükkeçeci, C.; Yazıcı, A. Multilevel Object Tracking in Wireless Multimedia Sensor Networks for Surveillance Applications Using Graph-based Big Data. *IEEE Access* **2019**. [[CrossRef](#)]

15. Xu, R.; Guan, Y.; Huang, Y. Multiple human detection and tracking based on head detection for real-time video surveillance. *Multimed. Tools Appl.* **2015**, *74*, 729–742. [[CrossRef](#)]
16. Xiao, S.; Li, W.; Jiang, H.; Xu, Z.; Hu, Z. Trajectory prediction for target tracking using acoustic and image hybrid wireless multimedia sensors networks. *Multimed. Tools Appl.* **2018**, *77*, 12003–12022. [[CrossRef](#)]
17. Dominguez-Morales, J.P.; Rios-Navarro, A.; Dominguez-Morales, M.; Tapiador-Morales, R.; Gutierrez-Galan, D.; Cascado-Caballero, D.; Jimenez-Fernandez, A.; Linares-Barranco, A. Wireless sensor network for wildlife tracking and behavior classification of animals in Doñana. *IEEE Commun. Lett.* **2016**, *20*, 2534–2537. [[CrossRef](#)]
18. Rehman, Y.A.U.; Tariq, M.; Sato, T. A novel energy efficient object detection and image transmission approach for wireless multimedia sensor networks. *IEEE Sens. J.* **2016**, *16*, 5942–5949. [[CrossRef](#)]
19. Lowe, D.G. Object recognition from local scale-invariant features. In *Proceedings of the 1999 7th IEEE International Conference on Computer Vision*; IEEE Computer Society: Los Alamitos, CA, USA, 1999; Volume 2, pp. 1150–1157. [[CrossRef](#)]
20. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
21. Guo, Y.; Bennamoun, M.; Sohel, F.; Lu, M.; Wan, J.; Kwok, N.M. A comprehensive performance evaluation of 3D local feature descriptors. *Int. J. Comput. Vis.* **2016**, *116*, 66–89. [[CrossRef](#)]
22. Harris, C.; Stephens, M. A combined corner and edge detector. In *Alvey Vision Conference*; Citeseer: Princeton, NJ, USA, 1988; Volume 15, pp. 10–5244.
23. Förstner, W.; Gülch, E. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Proceedings ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, Interlaken, Switzerland, 2–4 June 1987; pp. 281–305.
24. Li, Y.; Li, Q.; Liu, Y.; Xie, W. A spatial-spectral SIFT for hyperspectral image matching and classification. In *Pattern Recognition Letters*; Elsevier: Amsterdam, The Netherlands, 2018.
25. Rodríguez, M.; Delon, J.; Morel, J.M. Fast Affine Invariant Image Matching. *Image Process. Line* **2018**, *8*, 251–281, doi:10.5201/ipol.2018.225. [[CrossRef](#)]
26. Ke, Y.; Sukthankar, R. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—CVPR 2004*, Washington, DC, USA, 27 June–2 July 2004; Volume 2, p. II.
27. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2006; pp. 404–417.
28. Todorovic, S.; Ahuja, N. Region-based hierarchical image matching. *Int. J. Comput. Vis.* **2008**, *78*, 47–66. [[CrossRef](#)]
29. Tau, M.; Hassner, T. Dense correspondences across scenes and scales. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 875–888. [[CrossRef](#)]
30. Shao, Z.; Fu, H.; Li, D.; Altan, O.; Cheng, T. Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation. *Remote Sens. Environ.* **2019**, *232*, 111338. [[CrossRef](#)]
31. Long, N.; Millescamp, B.; Guillot, B.; Pouget, F.; Bertin, X. Monitoring the topography of a dynamic tidal inlet using UAV imagery. *Remote Sens.* **2016**, *8*, 387. [[CrossRef](#)]
32. Masazade, E.; Niu, R.; Varshney, P.K. Dynamic bit allocation for object tracking in wireless sensor networks. *IEEE Trans. Signal Process.* **2012**, *60*, 5048–5063. [[CrossRef](#)]
33. Wang, X. Intelligent multi-camera video surveillance: A review. *Pattern Recognit. Lett.* **2013**, *34*, 3–19. [[CrossRef](#)]
34. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
35. Zajdel, W.; Zivkovic, Z.; Krose, B. Keeping track of humans: Have I seen this person before? In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Barcelona, Spain, 18–22 April 2005; pp. 2081–2086.
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.



37. Xu, Y.; Ma, B.; Huang, R.; Lin, L. Person search in a scene by jointly modeling people commonness and person uniqueness. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 937–940.
38. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput.* **2004**, *22*, 761–767. [[CrossRef](#)]
39. Alahi, A.; Ortiz, R.; Vanderghenst, P. Freak: Fast retina keypoint. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 510–517.
40. Gheissari, N.; Sebastian, T.B.; Hartley, R. Person reidentification using spatiotemporal appearance. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1528–1535.
41. Visual Tracker Benchmark. Available online: <http://www.visual-tracking.net> (accessed on 30 July 2019).



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).