Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8313471, 11 pages
https://doi.org/10.1155/2022/8313471

Research Article

Real-Time Multitarget Tracking for Panoramic Video Based on Dual Neural Networks for Multisensor Information Fusion

Qing Lin

Intelligent Science and Information Engineering College, Xi'an Peihua University, Xi'an, Shaanxi 710125, China

Correspondence should be addressed to Qing Lin; 530610232@qq.com

Received 8 March 2022; Revised 5 May 2022; Accepted 9 May 2022; Published 2 June 2022

Academic Editor: Zhihan Lv

A multitarget real-time tracking system for panoramic video with multisensor information fusion dual neural networks is studied and implemented by combining dual neural networks, fused geometric features, and deep learning-based target real-time tracking algorithms. The motion model of multisensor information fusion is analyzed, the dual neural network perturbation model is introduced, and the state variables of the system are determined. Combined with the data structure model of multisensor information fusion, a least-squares-based optimization function is constructed. In this optimization function, the image sensor is constructed as an observationally constrained residual term based on the pixel plane, the inertial measurement unit is constructed as an observationally constrained residual term based on the motion model, and the motion relationship between sensors is constructed as a motion-constrained residual term to derive a mathematical model for real-time multitarget tracking of panoramic video. The simulation experimental results show that the integrated position estimation accuracy of the multisensor information fusion dual neural network-based panoramic video multitarget real-time tracking fusion algorithm is 10.5% higher than that of the optimal distributed panoramic video multitarget fusion algorithm without feedback. The results of the simulation experiments show that the integrated position estimation accuracy of the multitarget real-time tracking fusion algorithm for panoramic video improves by 10.96%, and the integrated speed estimation accuracy improves by 6.32% compared with the optimal distributed multitarget fusion algorithm without feedback. The proposed panoramic video multitarget real-time tracking algorithm based on the dual neural network can effectively improve the target tracking accuracy of the model on degraded frames (motion blur, target occlusion, out-of-focus, etc.), and the stability of the algorithm for target location and category detection is effectively improved by multiframe feature fusion. The research in this paper provides better technical support and research theory for panoramic video multitarget real-time tracking.

1. Introduction

Panoramic video multitarget tracking, as one of the basic elements of video analysis in the field of computer vision, is the basis for various subsequent high-level vision processing. In general, before tracking, it is necessary to first detect the panoramic video multitargets using a target detection algorithm [1]. Panoramic video multitarget detection refers to extracting the panoramic video multitarget from the background image based on the feature that the target pixels change while the background pixels remain unchanged in the video sequence; target tracking refers to finding the location of the target and determining the path or trajectory of its motion on each frame of the video. And emerging deep learning-like single-target tracking methods advocate unifying the above methods into an end-to-end framework, where the tracking results are output directly from the input frames into the network. Similar to single-target tracking, multitarget tracking has evolved from the traditional two parts of target detection and data association toward an end-to-end tracking approach. In turn, the measurement information of the sensors on the target has different degrees of noise, so it is difficult to accurately estimate the real motion of the target [2]. After using sensors to detect the target and obtaining the measurement information of the target, combined with state filters to reduce the effect of sensor noise, more accurate estimates of the target's position and velocity can be obtained as data input to the multisource

information fusion center to obtain more accurate track estimation information. Therefore, the direction of multisensor data fusion has become a mainstream direction to solve the scale ambiguity and cumulative drift. With the acquisition data from different sensors, it is possible to repeat sampling in multiple dimensions to correct the offset data due to noise [3]. Although the data acquisition aspect is too redundant, it exists only at the data acquisition stage and does not affect the subsequent system solution, and most importantly, it can ensure the validity of the data while providing a reference basis for the accuracy of the system motion estimation and can establish a corresponding feedback mechanism [4].

With the continuous development of information intelligence and network technology, video data in practical applications are increasing day by day, which demands higher quality for video target tracking at the academic level. Currently, most tracking algorithms are trained, tested, and evaluated online with parameters based on publicly available datasets, and the target tracking models trained in this way show good tracking performance on certain datasets [5]. However, the ultimate task of target tracking algorithms is to deal with target tracking problems in real application scenarios, and the actual tracking environment is subject to unpredictable interference factors at any time, which leads to complex and variable tracking scenarios and eventually affects the tracking results of the algorithms [6]. Currently, target tracking in complex scenarios such as target occlusion, illumination change, deformation, low resolution, low illumination, and rotation may still cause tracking drift or loss due to interference in the target appearance representation model, weak model discrimination, or incorrect model update, and tracking accuracy and robustness still need to be further improved to continuously meet the needs of practical applications [7]. Through the data collected by different sensors, repeated sampling can be performed in multiple dimensions to correct the offset data caused by noise. Although the data collection is too redundant, it only exists in the data collection stage and does not affect the subsequent system solution. The most important thing is to ensure the validity of the data and at the same time provide a reference for the accuracy of the system motion estimation and can establish the corresponding feedback mechanism.

Compared with panoramic video multitarget real-time tracking, the research related to panoramic video multitarget real-time tracking is still relatively small, but because it can reflect the real demand of autonomous driving environment perception, it is still in a booming stage of development, and a large number of related new algorithms are constantly proposed [8]. From the technical point of view, panoramic video multitarget real-time tracking needs to consider more frequent interactions and occlusions, and the interaction process is mostly nonregular and random, so it is suitable for the research of multisensor information fusion dual neural network approach. The first section is the introduction, which introduces the background and significance of the research and the research progress of multisensor information fusion dual neural network and panoramic video multitarget real-time tracking and also explains the research framework of this paper. The second section mainly analyzes the research status of multisensor information fusion dual neural network technology and panoramic video multitarget real-time tracking technology, from which the research ideas and technical theories of this paper are obtained. Section 3 investigates the real-time tracking of panoramic video multitarget based on multisensor information fusion dual neural network, introduces the process of constructing multisensor information fusion dual neural network tracking model, improves the real-time tracking algorithm of panoramic video multitarget, and designs and implements the real-time tracking system of panoramic video multitarget using the technology studied in this paper. Section 4 provides an experimental analysis of the research in this paper and builds a simulation platform for vertical and horizontal comparison experiments on two analysis systems of absolute and relative positional errors, respectively, and the experimental results show the feasibility, robustness, and accuracy of the multitarget real-time tracking system in this paper. Section 5 summarizes the target tracking algorithm proposed in this paper and provides an outlook on the improvement of the algorithm in the future, taking into account the future trend of panoramic video multitarget real-time tracking algorithm.

2. Key Technology

As target detection algorithms continue to advance, more and more traditional multitarget tracking algorithms have started to adopt a detection-based and tracking framework, and a large number of excellent algorithms have been proposed for the data association part [9]. And with the development of deep learning, researchers have proposed end-to-end multitarget tracking networks that use convolutional neural networks or recurrent neural networks for target tracking. The weak classifier is combined into a strong classifier, the classifier is used to track the location of gray blocks in the image, and the classifier is automatically updated online, which has strong robustness to changes such as illumination. The multitarget real-time tracking algorithm of panoramic video builds histograms of various features of the tracked target. Experiments show that the algorithm can avoid the interference of complex backgrounds, has obvious advantages over single-feature tracking, and can adapt to more complex background scenes.

2.1. Dual Neural Network Technology for Multisensor Information Fusion. Klein et al. first proposed PTAM, a visual SLAM scheme that parallelizes the tracking and map building process, and in doing so distinguished between a front-end, which allocates a large number of computational resources to feature tracking to achieve real-time response to sensor data and ensure system real time and a back-end, which uses offline optimized computation [10]. Blott G et al. proposed a direct monocular vision for a large-scale range. The LSD-SLAM scheme, which directly estimates similar transformations between keyframes and scale-aware image matching by feature points for loopback detection, enables

map construction of semidense scenes on the CPU. However, it is sensitive to the initial value of the image sensor's internal reference and exposure level, and it is easy to cause tracking loss when it moves fast [11]. Ni T et al. theoretically analyzed the effect of system error on target measurement trajectory and proposed a phase-correlation-based target trajectory alignment correlation algorithm for the relationship between trajectory correlation and system error in radar networking [12]. The algorithm uses one- and two-dimensional phase correlation to estimate rotation and motion based on Radon transform and Fourier transform properties to achieve an accurate correlation of target trajectories provided by different radars in the same coordinate system without a priori estimation and alignment of radar system errors [13]. Multisensor data fusion can carry out multifaceted and multilevel measurements on the description and status of a single feature, which can further ensure measurement accuracy. It can greatly improve the stability of system positioning and adaptability to the environment.

Using offline training, the tracking algorithm based on deep convolutional networks can learn the common feature model that can represent the robustness of the target and dynamically update the coefficients of the classifier by online learning to achieve the purpose of improving the tracking performance [14]. However, the tracking process involves the adjustment and update of huge network parameters, which consumes a lot of computing time and does not fully meet industrial-grade standards in terms of real-time performance [15]. The model of the tracking algorithm is motivated by continuous optimization, which can be done with the help of multifeature fusion on the one hand and from the optimization of the filter mathematical modeling on the other hand. Make full use of multiple information sources, and combine the redundant or complementary information of multiple information sources in space or time according to specific standards to obtain a consistent interpretation or description of the measured object, so that the information system is relatively [16]. The system formed by the subsets it contains has better performance.

2.2. Panoramic Video Multitarget Real-Time Tracking Technology. To reduce the effects of illumination changes and appearance deformation on the model, the mean-shift algorithm uses a kernel density function to represent the target appearance model and locate the local optimal position by iterative means. Mean-shift algorithm is widely used for its computational simplicity and high real-time performance [17]. However, mean-shift is insensitive to rotation, scale, background motion, and so on and lacks real-time update of the target model; thus, it is easy to cause tracking failure due to target scale change in engineering applications [18]. Li B et al., to solve the problem that the mean-shift algorithm cannot update the target model in real-time, improved the mean-shift algorithm and proposed the Camshift algorithm, which is a tracking algorithm based on color probability distribution. This algorithm is suitable for tracking scenarios where the target color is single and has a large color difference with the background, but not for target tracking scenarios where the target and background colors are similar or the background is complex and the target texture is rich [19]. Luo L et al. proposed a continuous adaptive mean drift tracking algorithm with a background suppression histogram model, which improves the tracking accuracy and stability by suppressing the hues belonging to the background in the original color model [20]. Lu W et al. extended the Camshaft algorithm by adding a fast panoramic video multitarget state prediction algorithm with adaptive kernel bandwidth and state estimation, which reduces the mean-shift [21]. Saha P et al. used the Camshaft tracker to detect hand motion trajectories from video images, combining Markov model sequence classification methods to improve the success rate of target detection [22].

The tracking targets are chunked and then matched in grayscale, and the matching coefficients of grayscale are used to update the template, which reduces the computing time and increases the real-time and reliability of tracking [23]. The algorithm for panoramic video multitarget tracking uses the RGB color space model, this algorithm applies a learning technique, the RGB component of the image generates alternative targets, combines weak classifiers into strong classifiers, and uses classifiers to locate grayscale blocks in the image to achieve tracking, and the classifiers are automatically updated online with strong robustness to changes such as illumination [24]. The panoramic video multitarget tracking algorithm with multifeature fusion creates histograms of multiple features of the tracked target, and experiments show that the algorithm can avoid the interference of complex backgrounds and has obvious advantages over single-feature tracking, which can adapt to more complex background scenes [25].

3. Research on Real-Time Multitarget Tracking of Panoramic Video Based on Dual Neural Networks for Multisensor Information Fusion

3.1. Dual Neural Network Tracking Model for Multisensor Information Fusion. We can use multiple groups of sensors to achieve a variety of navigation and positioning panoramic video multitarget real-time tracking methods, and then the collected information is processed and fused to improve the accuracy, stability, and reliability of the panoramic video multitarget operation. Information fusion then requires specific research on the fusion method strategy and fusion algorithm, which can be realized in data layer fusion, feature layer fusion, and decision layer fusion. The three fusion methods have their advantages and disadvantages, and it is necessary to choose the appropriate fusion method according to the specific actual situation. Data layer fusion has a large amount of information, small loss, and high accuracy, which leads to an increase in the amount of computation and a decrease in the speed of operation; feature layer fusion has a reduced amount of information, but the accuracy is also reduced and the speed of operation is improved; decision layer fusion has the smallest amount of information, the loss of information is relatively large, the

accuracy of control is reduced, and the speed of operation is the fastest. The structure diagram of the dual neural network to achieve multisensor fusion feature layer fusion is shown in Figure 1.

Target detection requires better extraction of image features on a baseline network. The current research trend in feature extraction mainly features fusion and learning of high-resolution features under larger perceptual fields. Target detection consists of target classification, which requires position-insensitive feature invariant representations. And target localization requires covariant representation of features sensitive to position and scale. Feature fusion can enhance the detection results from these two aspects. Current feature fusion methods generally consider top-down or bottom-up flow processing feature layer fusion and various element-by-element operations. Learning higher resolution and larger perceptual field enables us to obtain richer scale content information and local details.

After the input images of the two branches pass through the feature extraction network, the middle result of $8 \times 8 \times 128$ and $24 \times 24 \times 128$ will be obtained, respectively, then the response map of size $19 \times 19 \times 1$ will be obtained by the related operation to obtain the corresponding result of the template in the search area, and the position of the maximum value of the response map will be estimated as the position of the target in the current frame. Because there is a fully connected layer in the network structure, the calculation process can be expressed by (1), where ϖ, the convolutional embedding, is a function and $\beta * \xi \subseteq G$ represents the compensation signal value.

$$g(m,n) = \varpi(m) + \varpi(n) + \beta * \xi. \tag{1}$$

The overall architecture is based on the twin network CFNet, where the input dimensions of the network template branch and search branch in CFNet are the same as Siamese-FC, but the correlation filter block f is added between the input on the template branch and the final correlation operator, which can be expressed in the form of (2), where $\varpi = (m)$ denotes the standard correlation filtering operation and w denotes $m = g_\beta(m)$ from the trained feature map.

$$f_{\beta,\xi,\psi}(m,n) = \varphi * \varpi\left(g_\beta(n)\right) \odot \left(g_\beta(m)\right) + \varphi. \tag{2}$$

As a typical geometrical feature, the flow feature can help the algorithm to adapt to the above scenario better and can make the algorithm more robust. By adding stream shape analysis on top of the original framework, the formula can be written in (3), where $\chi(m)$ represents the stream form template function and ξ is the fusion operation to combine the semantic features $\varpi(m)$ and geometric features $\chi(m)$.

$$g(m,n) = (\chi(m) \otimes \chi(n)) * \varpi(m) + \beta * \xi. \tag{3}$$

Each target in the video frame is considered as a point on the Grassmann manifold and represented by its corresponding basis matrix. The basis matrix U can be represented by the principal feature vector of the sample, which can represent the geometric structure of the target. For a given video sequence, its observation matrix H is a sequence of M frames, where $h_1, h_2 \ldots h_M$ are denoted as the i-th mean extracted observation of the target, h_i denotes the vector corresponding to the target image, and T is the mean vector. The fundamental matrix can be represented by the principal eigenvectors of the sample, which can represent the geometry of the target. For a given video sequence, its observation matrix is a sequence of frame images.

$$H = \left[h_1, h_2 \ldots h_M\right]^T + \left[h_1, h_2 \ldots h_M\right]^{-T}. \tag{4}$$

In the streamlined sample pool, different samples are set to have different weights to reflect their degree of contribution to the overall understanding of the geometric structure of the continuous image sequence. The Gaussian mixture model is used to estimate the sample weights, which can be expressed by (5), where M represents the total number of dual neural network components $N(\varpi; \vartheta(k); W)$, $\sigma(k)$ is the weight of the corresponding component W among them, and $\vartheta(k)$ represents the component mean. By analyzing the kinematics of the inertial measurement unit and combining its inherent cumulative drift property, the parameters of the inertial measurement unit are calibrated. The preintegration theory of the inertial measurement unit is introduced, the data structure model of the multisensor odometer is deduced, and the multisensor odometer is jointly calibrated to obtain the optimal parameters.

$$\chi(\varpi) = \sum_{k=1}^{M} \sigma(k) * N(\varpi; \vartheta(k); W). \tag{5}$$

As shown in Figure 2, the Lemming sequence from the dataset OTB2021 is used as an example to visualize the effect of occlusion on the response map. At frame 50, the target is very clear and unobstructed, and the response graph tends to have a large value of the ideal single-peaked Gaussian distribution indicator. At frame 350, the target is heavily occluded, and the response map shows multiple low-peaked Gaussian distributions with significant noise and significantly lower indicator values.

3.2. Panoramic Video Multitarget Real-Time Tracking Algorithm. To address the target tracking loss phenomenon, a proposed region solution is proposed and combined with a robust target model update criterion to achieve continuous target tracking. The algorithm model is shown in Figure 3. In Figure 3, the blue box indicates the search box, the yellow box indicates the candidate suggestion region, and the red box is the target location.

In solving the filter $L(x)$, it is described as a regularized least-squares objective function in the form of the tangent function as in (6), where $l(x)$ is the filter associated with the x-th frame, $l(x)^n$ denotes the filter corresponding to each feature dimension, $g(x)^d$ is the feature map corresponding to each feature in the input candidate frame, and $n = 1, 2, \ldots, N$, N is the number of feature dimensions and takes the value of 10.

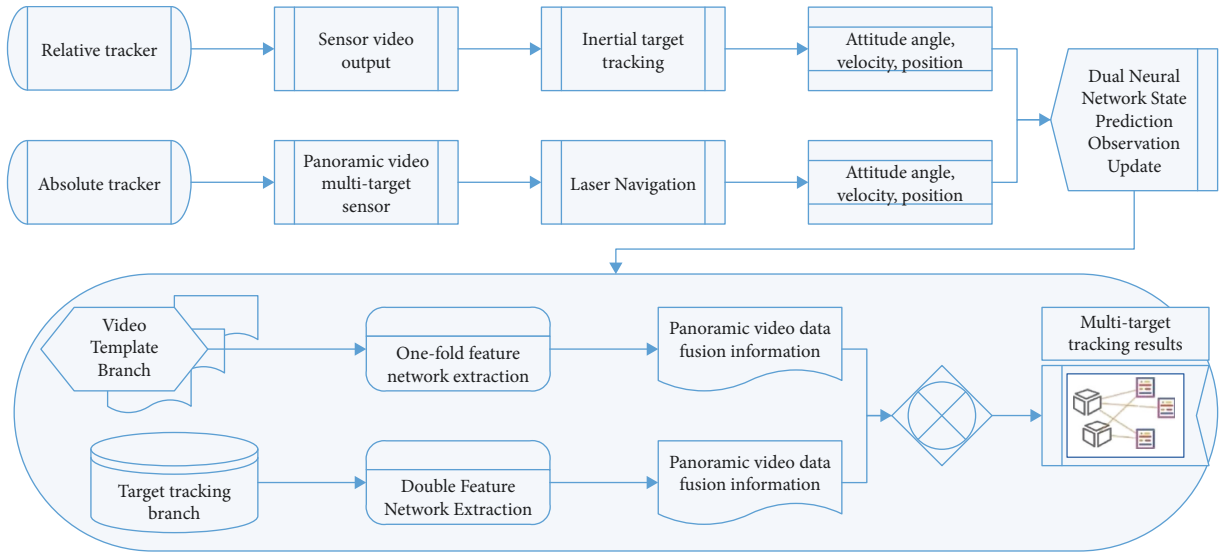

Figure 1: Feature layer fusion structure diagram.

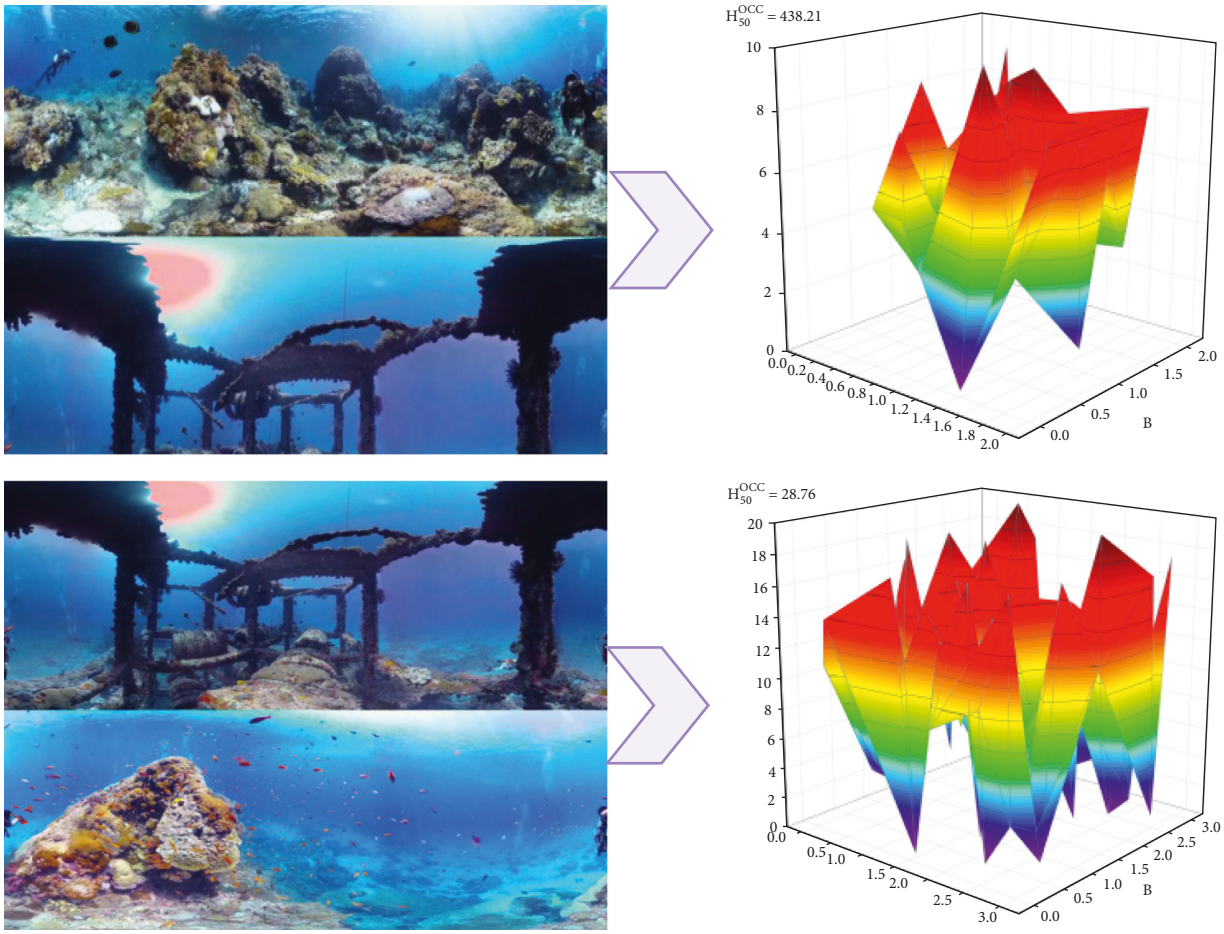

Figure 2: Schematic diagram of the tracking result evaluation method.

$$L(x) = \tan\max_{l(x)} \left\| \sum_{n=1}^{N} l(x)^n * g(x)^d - k * \psi \right\|^2. \quad (6)$$

In the target tracking process of the actual scene, the target and background of the two adjacent frames do not change much or are similar. Then, the target model obtained according to the target is also basically the same, and the filter obtained by learning in the $x+1$ frame and the filter obtained in the x frame are the same, which can be represented by the mathematical model of (7), where $f(w)$ and $f(w)^x$ are trade-off coefficients to control the strength of the regular term formed by the difference between the filters of the two adjacent frames to prevent it from becoming larger in the process of the optimal solution. Otherwise, it will lead to tracking drift or tracking failure when the target is obscured.

$$\tan\max_{f(w)} \left\| f(w+1)^{x+1} - f(w)^x \right\|^2. \quad (7)$$

Assuming that the scale of the target at frame x is $K(x) * H(x)$, the target scale pyramid for scale estimation is constructed near the target location with the number of layers B. The scale $Q(x, n)$ for any image slice in the target scale pyramid is expressed as (8). $\xi(x)$ and $\zeta(x)$ denote the scale factor of different scale layers.

$$\begin{cases} Q(x,n) = \xi(x) * H(x) + \zeta(x) * K(x), \\ x \subseteq \left[-\dfrac{X+Y+Z-1}{3}, \dfrac{X+Y+Z-1}{3} \right], \\ B = X + Y + Z. \end{cases} \quad (8)$$

The position of the current frame is found in the position, and the size of the previous frame is known. The current frame is assumed to be the x frame, and the target position of the x frame 1 is

$$Q(x-1) = (m(x-1), n(x+1)). \quad (9)$$

The scales are

$$B(x, L(x-1)) = \left(\xi(x)^n * H(x-1), \xi(x)^{n+1} * K(x+1) \right). \quad (10)$$

The scale of each candidate sample is $B(x, L(x-1))$, and the region samples are represented $F(\text{pad}, x)$ as in (11).

$$F(\text{pad}, x) = \frac{H(x) * K(x) \| Q(x) - Q(x-1)}{\text{pad}}. \quad (11)$$

When the target is disturbed, blurred, or obscured by the background, the associated filtered response map will oscillate substantially, and the location with the largest response value is not necessarily the target. According to the maximum filtered response strategy, it will be determined that the tracking is correct and the target model is updated, while the target model has drifted, thus causing the subsequent tracking to fail. For this reason, in addition to the target model update based on the correlation filtered maximum response value, the average peak correlation energy is introduced as the basis for judging the target model update, and APCE can be calculated using (12), where $\max(x)$ denotes the maximum value of the response map, $\min(x)$ denotes the minimum value of the response map, and $F(h, k)$ denotes the response value at the (h, k) position.

$$S(h,k) = \frac{|\max(x) - \min(x)|^2}{\text{sum}\left(\sum_{h,k=1}^{N} (F(h,k) - \min(x)) \right)}. \quad (12)$$

Table 1 shows the maximum response values and $S(h, k)$ before and after the target is occluded in the girls2 video sequence. The target is occluded from frame 205 onwards, and at first, the occlusion area is small, and the maximum response values $S(h, k)$ do not decrease significantly (maximum response value of 0.749 and $S(h, k)$ 48.617). As the occlusion area becomes larger, the maximum response value $S(h, k)$ decreases significantly, and the target is completely occluded at frame 110 until the target appears completely at frame 236 (maximum response value of 0.274 and $S(h, k)$ 25.801). The target model was not updated between frames 209 and 222, and the maximum response value and APCE returned to normal from frame 226, reflecting the robustness of the designed model update strategy. In order to enhance the generalization ability of the algorithm, a double detection module needs to be added when the maximum filter response value fluctuates. That is, the maximum response position of the current frame and the filter model of the first frame are subjected to a correlation filtering operation to calculate the maximum response value. If the threshold condition is met, the model has not drifted; if the threshold condition is not met, the candidate region proposal scheme is started, and the target position is reacquired.

3.3. Panoramic Video Multitarget Real-Time Tracking System Design and Implementation. The design of this system is based on the study of deep learning target detection and data fusion, aiming at the detection and tracking of panoramic video multiple targets, as well as the function of giving already detected panoramic video multiple targets. The design of the system is divided into hardware design and software design. The main function is to acquire video information through multisensors, transmit it to the panoramic video multitarget detection and tracking server, process it, and then pass it to the display terminal and display it on the system display through the transmission equipment. The system realizes the functions of video acquisition and storage, digital image processing, panoramic video multitarget detection, panoramic video multitarget tracking, browsing, and so on. The software system is designed in modules and divided into five modules. The module distribution of the system is shown in Figure 4. The pooling operation of the neural network reduces the spatial resolution of the image, increases the receptive field, and makes the high-level features scale and rotation invariant. In order to improve the ability of semantic description and precise positioning, some advanced tracking algorithms integrate

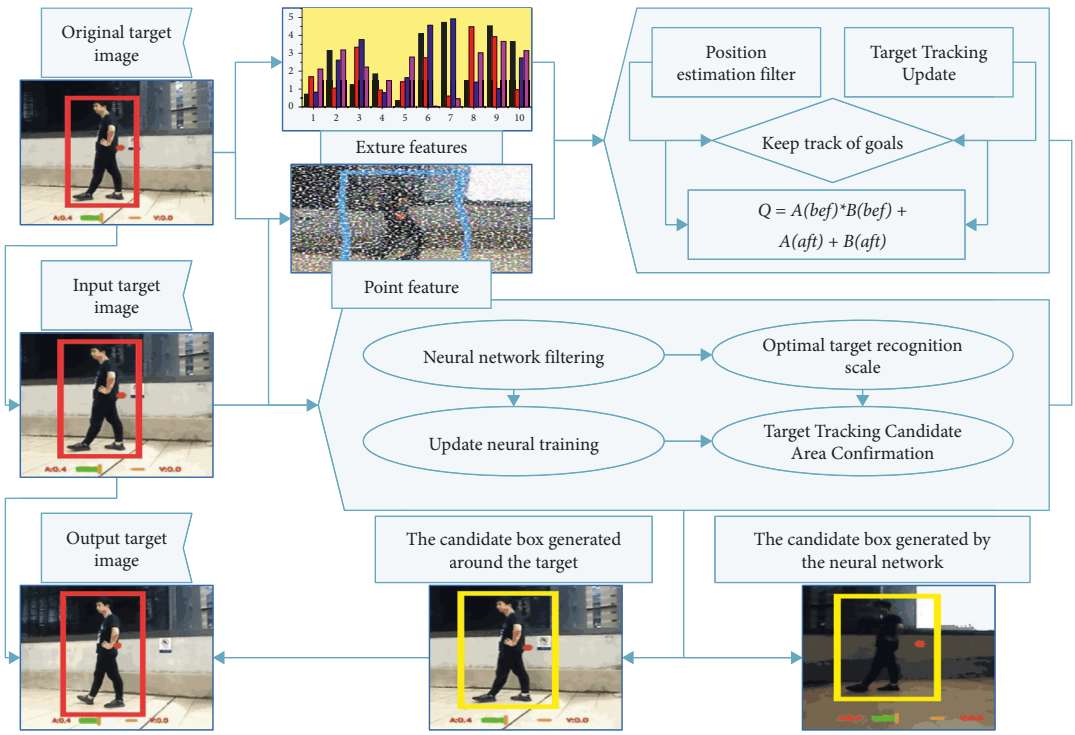

Figure 3: Algorithm model.

Table 1: Maximum response values and $S(h,k)$ before and after target masking.

Video frame	200	205	209	214	218	222	226	231	236
Maximum response value	0.749	0.231	0.682	0.505	0.569	0.740	0.033	0.449	0.274
$S(h,k)$	48.617	49.405	28.492	26.963	24.117	46.469	40.580	23.899	25.801

convolutional neural networks and handcrafted features to improve the tracking accuracy. However, the manual features often contain a large amount of background noise, which affects the tracking performance. So consider extracting complete edge information from the convolutional neural network for precise target location.

The algorithm module reads video data from the image read/write module, does algorithmic processing, and then returns the processed video data to the image read/write module. Open is to open a channel to access a file or device. If successful, it will return a file descriptor that can be used for Write and Read operations or system calls. This file descriptor is unique and cannot be shared with other running processes. If two programs open a file at the same time, they will each hold different file descriptors. If a program writes to a file, it will be written to the location where the program last left the file (not the end of the file). The two programs do not write data interleaved, but one program's write may overwrite the other program's write. Each program only remembers its position in the file and only starts reading and writing from this position. We can prevent such conflicts from happening by locking the file.

Two storage spaces are used for reading video data. When the video stream writes data to space 1, the application reads data from space 2 and does the processing, and when the processing is completed and the video stream finishes writing a frame to space 1, space 1 and space 2 are exchanged, at which point the video stream writes data to space 2 while the application reads data from space 1. The above operation is achieved by controlling the multisensor information fusion, setting Buffer as the starting address of space 1 and Back Buffer as the starting address of space 2. When the Energy Harvesting Aided Massive Multiple Access Networks is used to implement the function calculation, the cycle required is very long. For the lookup table implementation method, no matter how complex the function is, only one clock cycle is required for memory reading; the second case is that the calculation complexity is very high. When the resource is too high and consumes too much Energy Harvesting Aided Massive Multiple Access Networks resources, use the lookup table to reduce the resource consumption. Of course, if the precision is very high, more memory is required. In practice, it depends on the resource usage and determines whether to use lookup tables to implement Source-Aware Packet Management functions.

4. Experimental Analysis

4.1. Multisensor Information Fusion Simulation Analysis. The instantaneous position and velocity error curves of the two sensors and the instantaneous position curves after the fusion of each fusion algorithm are shown in Figure 5. It can

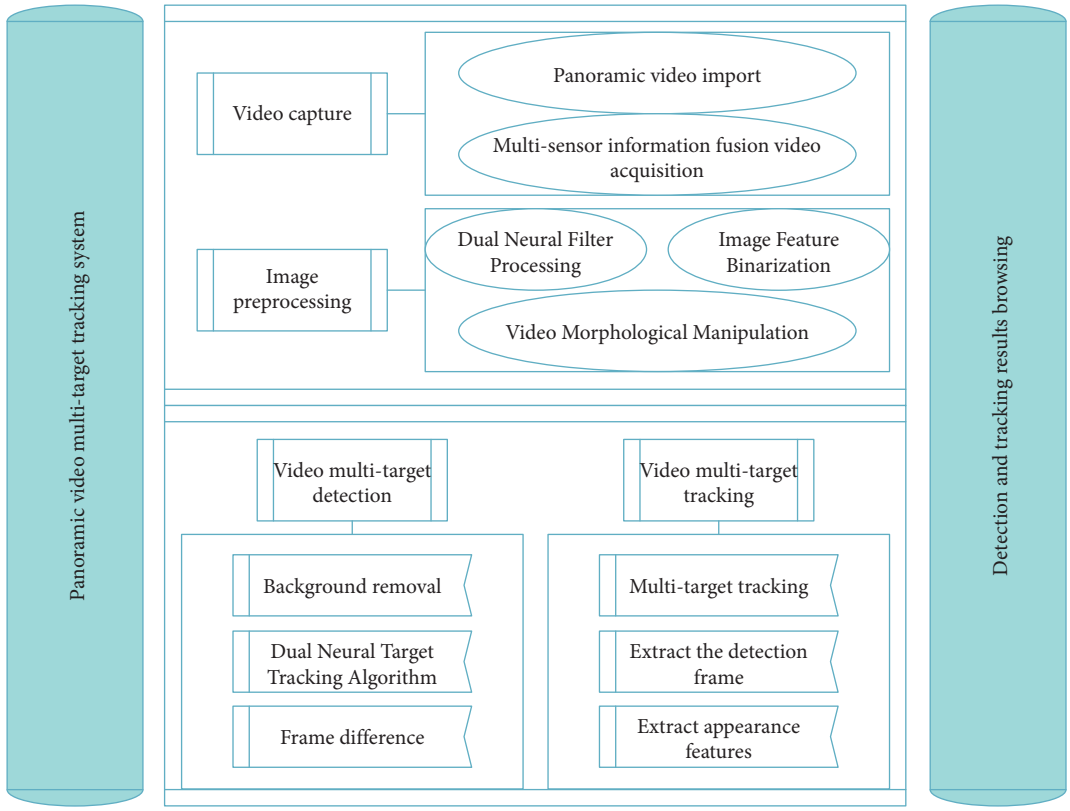

Figure 4: Module diagram of panoramic video multitarget tracking system.

be seen that the position measurement error of sensor 1 multifunctional radar fluctuates around 14.6 mm, and the position measurement error of sensor 2 infrared photoelectric radar fluctuates around 6.7 mm. After fusion with the time-convolution network-based track fusion algorithm, the instantaneous position error can be stabilized at around 13.2 mm. Compared with the optimal distributed fusion algorithm without feedback, the fusion algorithm based on the temporal convolutional network has improved the comprehensive position estimation accuracy by 12.96% and the comprehensive speed estimation accuracy by 6.52%, which verifies the proposed method in this paper.

The velocity error curves are shown in Figure 6. Figure 6 reflects the instantaneous velocity estimation error, the velocity measurement error of sensor 1 multifunctional radar fluctuates around 6 mm/s, and the measurement error of sensor 2 infrared photoelectric radar fluctuates around 8 m/s. After fusion, the error is stabilized at around 4 mm/s. It can be seen that the fusion algorithm can greatly reduce the error, although it cannot eliminate it.

The Bar-Shalom–Campo algorithm takes into account the correlation between sensor measurement errors and can reach the optimum in the sense of great likelihood but cannot reach the optimum in the sense of minimum mean square error. The estimation accuracy is higher than that of the simple convex combinatorial track fusion algorithm but lower than that of the feedback-free optimal distributed track fusion algorithm and the time-convolutional network-based track fusion algorithm. The optimal distributed track fusion algorithm without feedback avoids the calculation of mutual covariance, which is computationally larger than the simple convex combined track fusion algorithm and smaller than the Bar-Shalom–Campo algorithm and achieves higher accuracy in position and velocity estimation. The time-convolutional network-based track fusion algorithm has the highest accuracy of integrated position and integrated velocity estimation compared with the other three fusion algorithms, which verifies the effectiveness of the time-convolutional network-based track fusion algorithm.

4.2. Performance Analysis of Multiobjective Real-Time Tracking Algorithms. Real time is one of the requirements of panoramic video multitarget tracking, and the time evaluation index can accurately reflect the real-time performance of the algorithm. The first 1000 frames of the video are counted, the real coordinates of the center of the tracked target and the tracking coordinates of the algorithm are recorded every 50 frames to calculate the absolute error of the coordinates, and the absolute error statistics are obtained. The absolute error is calculated and shown in Figure 7. The traditional multitarget real-time tracking algorithm will experience tracking drift or even failure when encountering occlusion and illumination changes. The multitarget real-time tracking algorithm in this paper can effectively solve the problem of tracking failure caused by the occlusion of moving objects, but it is robust to illumination changes. The stickiness is poor, and the multitarget real-time

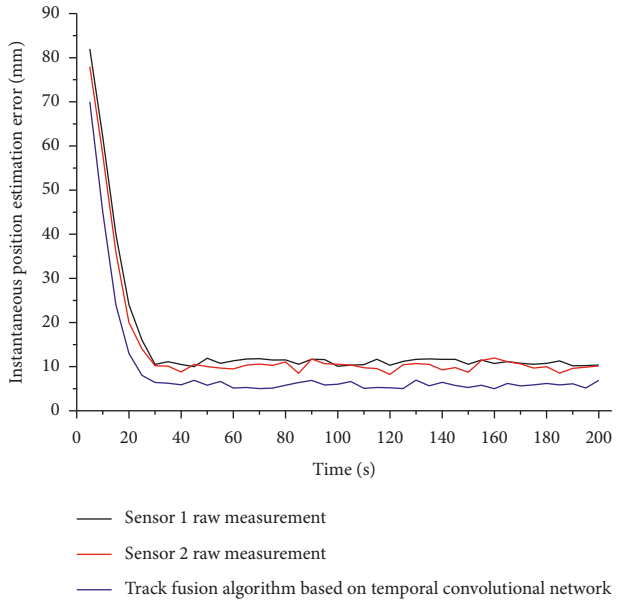

Figure 5: Instantaneous position estimation error.

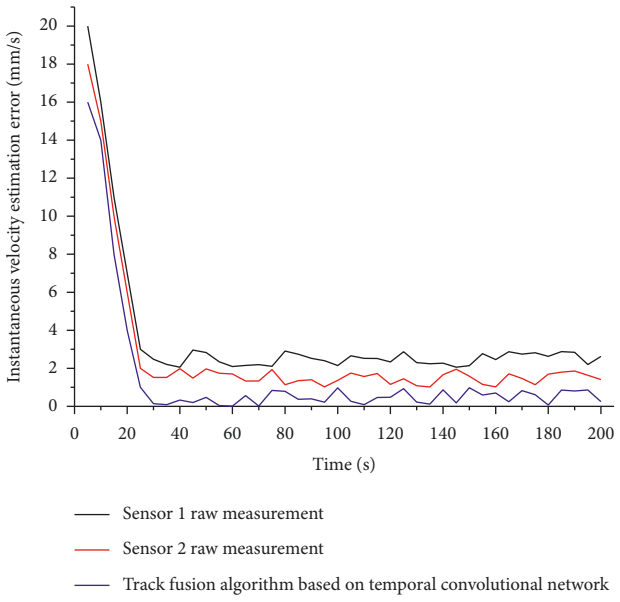

Figure 6: Instantaneous velocity estimation error.

tracking algorithm in this paper has great advantages in tracking effects in occlusion and lighting environments.

Through comparison experiments, the Camshaft algorithm has high computing efficiency and can track accurately under simple background and no lighting changes but loses the target when there are occlusion or other interference factors, and the robustness of the algorithm is poor. The improved algorithm proposed in this paper can perform accurate tracking with high robustness in both scenes, ORB-LBP feature point matching is added to the Camshift + Kalman algorithm, the tracking is corrected with the feature point matching method, which increases the complexity of the algorithm, and it can be concluded that the

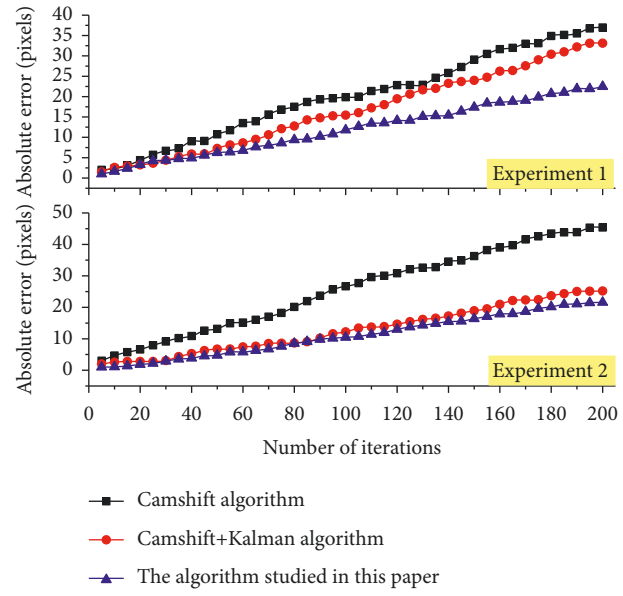

Figure 7: Tracking error graph.

improved algorithm has better performance. The algorithm increases a small amount of processing time, but the reliability and accuracy of tracking are improved, and the improved algorithm has more advantages in panoramic video multitarget tracking.

The distance accuracy achieved by the algorithms in each chapter on different datasets shows an increasing trend chapter by chapter. The TC-256 dataset is more challenging than some video sequences in the OTB2021 dataset, and the algorithms are slightly less effective in tracking in the TC-256 dataset. Both Algorithm 2 and Algorithm 3 use traditional manual features to model the target appearance model. The target-guided saliency-based detection algorithm of Algorithm 2 is more effective than the candidate region suggestion scheme of Algorithm 3, but the target scale estimation does not have the advantage of Algorithm 3, and Algorithm 3 performs filter optimization in solving the multifeature coupled model, so the distance accuracy of Algorithm 3 is improved compared to that of Algorithm 2. The algorithm in this paper achieves the highest distance accuracy because the algorithm in this paper uses hierarchical depth features to model the target appearance model, which is robust to scenes with severe target deformation and rotation, and also improves the tracking effectiveness of the algorithm by improving the filter modeling. The comparison of distance accuracy and overlap success rate of the algorithms in each chapter on OTB2021 and TC-256 datasets is shown in Figure 8.

4.3. Multitarget Real-Time Tracking System Test Analysis. Figure 9 depicts the distribution of relative positional errors in time, it can be seen that this multitarget real-time tracking system outperforms VINS and OpenVINS under the RPE evaluation system, and it can be seen that this multitarget real-time tracking system outperforms VINS-Mono and OpenVINS in six statistical indicators, such as standard

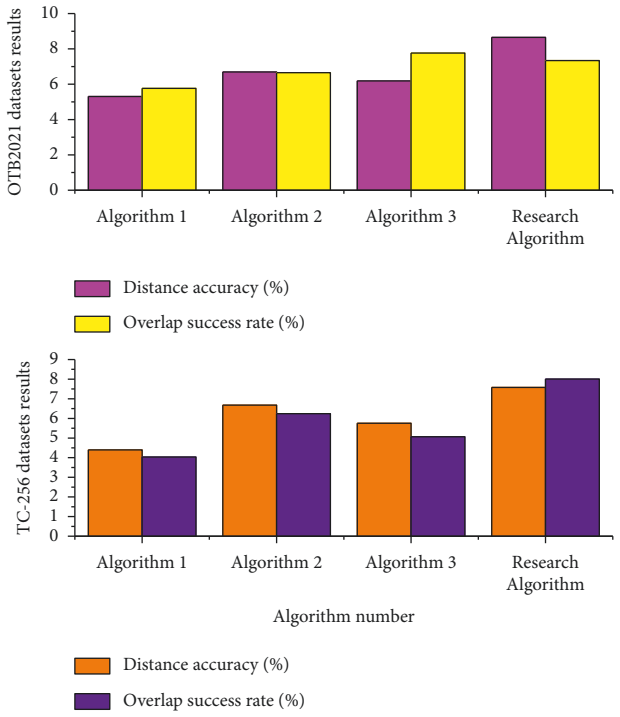

Figure 8: Distance accuracy and overlap success rate of the algorithm.

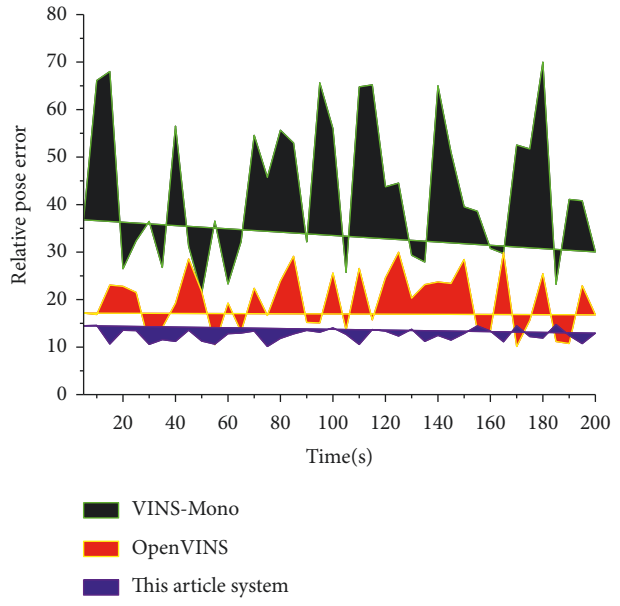

Figure 9: The relative attitude error curve.

deviation, root mean square, minimum, median, mean, and maximum OpenVINS.

The experimental results show that the motion estimation of this multitarget real-time tracking system is consistent with the real trajectory, which confirms the feasibility of this multitarget real-time tracking system. In the APE analysis dimension, there is a certain error with the real trajectory; the maximum value is 0.444 m, which mainly occurs in the initial position pair. The analysis suggests that the main reason is the lack of a fixed world coordinate system for the sensor and provides a direction for subsequent improvement. The cross-sectional comparison experiment compares the motion estimation implemented in this paper with the current mainstream OpenVINS and VINS-Mono motion estimation. The experimental results show that the motion estimation of this multitarget real-time tracking system is lower than OpenVINS slightly higher than VINS-Mono in the APE analysis system and higher than VINS-Mono and OpenVINS in the RPE analysis system, which achieves the research objective.

5. Conclusion

In this paper, we analyze the current research status of target tracking algorithms at the level of tracking model modeling and summarize the difficulties and challenges encountered by current target tracking algorithms. Based on this, the target tracking algorithm is discussed and designed based on two models, generative and discriminative, in terms of mathematical modeling of the model, modeling of target appearance using manual and depth features, relocation-tracking strategy, and target model update, aiming to make the algorithm achieve a balance between accuracy, robustness, and real-time performance in complex scenarios and meet the needs of practical applications. The algorithm is designed to achieve a balance between accuracy, robustness, and real-time performance in complex scenarios and meet the needs of practical applications. By analyzing the multisensor data fusion model, a multisensor odometer design scheme is proposed. When the application environment in which the mobile device is located is complex, the single sensor has problems of scale ambiguity, positioning failure, and cumulative drift in positioning, which cannot meet the application requirements of mobile devices in special environments. And using multisensor joint detection and performing fused positional estimation is an effective method. By constructing a mathematical model of the bit-pose estimator with a least-squares problem locally, an error term is constructed for each sensor device measurement and solved optimally to achieve the localization function. The tracking algorithm based on deep convolutional networks learns a robust common feature model of the target by offline training and dynamically updates the coefficients of the classifier by online learning, which greatly improves the accuracy and robustness of the tracking algorithm. However, the tracking process will involve the adjustment and update of huge network parameters, which will consume a large amount of computing time and cannot fully reach industrial-grade standards in terms of real-time performance. How to effectively integrate the offline network model and online correlation filtering algorithm is a key task in the next step, to improve the accuracy and robustness of the tracking algorithm but also to improve the processing speed of the algorithm to meet the needs of real-time tasks. This can be started by designing parameter optimization algorithms and combining them with large-scale public datasets for training to achieve algorithm effect equivalence on different datasets.

Data Availability

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

Conflicts of Interest

There are no conflicts of interest in this paper.

Acknowledgments

This work of this paper was supported by Xian Peihua University.

References

[1] S. Wang, F. Jiang, B. Zhang, R. Ma, and Q. Hao, "Development of UAV-based target tracking and recognition systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3409–3422, 2020.

[2] Z. Yan, T. Duckett, and N. Bellotto, "Online learning for 3D LiDAR-based human detection: Experimental analysis of point cloud clustering and classification methods," *Autonomous Robots*, vol. 44, no. 2, pp. 147–164, 2020.

[3] X.-W. Liu, Q. Zhang, Y. Luo, X. Lu, and C. Dong, "Radar network time scheduling for multi-target ISAR task with game theory and multiagent reinforcement learning," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4462–4473, 2021.

[4] W. Li, L. Cao, L. Yan, J. Liao, and Z. Wang, "Vacant parking slot detection and tracking during driving and parking with a standalone around view monitor," *Proceedings of the Institution of Mechanical Engineers - Part D: Journal of Automobile Engineering*, vol. 235, no. 6, pp. 1539–1551, 2021.

[5] E. Gärtner, A. Pirinen, and C. Sminchisescu, "Deep reinforcement learning for active human pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10835–10844, New York, NY, USA, February 2020.

[6] A. Delforouzi, S. A. H. Tabatabaei, K. Shirahama, and M. Grzegorzek, "A polar model for fast object tracking in 360-degree camera images," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 9275–9297, 2019.

[7] Y. Choo, J. Jang, and J. Paik, "Scene mapping-based video registration using frame similarity measurement and feature tracking," *IEIE Transactions on Smart Processing & Computing*, vol. 8, no. 6, pp. 456–464, 2019.

[8] M. Simic, M. Peric, I. Popadic et al., "Big Data and development of Smart City: System architecture and practical public safety example," *Serbian Journal of Electrical Engineering*, vol. 17, no. 3, pp. 337–355, 2020.

[9] A. A. Micheal, K. Vani, S. Sanjeevi, and C.-H. Lin, "Object detection and tracking with UAV data using deep learning," *Journal of the Indian Society of Remote Sensing*, vol. 49, no. 3, pp. 463–469, 2021.

[10] A. Rangesh and M. M. Trivedi, "No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 4, pp. 588–599, 2019.

[11] G. Blott, J. Yu, and C. Heipke, "Multi-view person Re-identification in a fisheye camera network with different viewing directions," *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 87, no. 5, pp. 263–274, 2019.

[12] T. Ni, Y. Chen, S. Liu, and J. Wu, "Detection of real-time augmented reality scene light sources and construction of photorealis tic rendering framework," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 271–281, 2021.

[13] W. Sun and C. Mo, "High-speed real-time augmented reality tracking algorithm model of camera based on mixed feature points," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 249–259, 2021.

[14] L. Yang, Y. Liu, H. Yu et al., "Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: A review," *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2785–2816, 2021.

[15] S. Hu, K. Shimasaki, M. Jiang, T. Senoo, and I. Ishii, "A simultaneous multi-object zooming system using an ultrafast pan-tilt camera," *IEEE Sensors Journal*, vol. 21, no. 7, pp. 9436–9448, 2021.

[16] J. P. Amezquita-Sancheza, M. Valtierra-Rodriguez, and H. Adeli, "Machine learning in structural engineering," *Scientia Iranica*, vol. 27, no. 6, pp. 2645–2656, 2020.

[17] Y. Zhou, L. Chang, and B. Qian, "A belief-rule-based model for information fusion with insufficient multi-sensor data and domain knowledge using evolutionary algorithms with operator recommendations," *Soft Computing*, vol. 23, no. 13, pp. 5129–5142, 2019.

[18] N. Shaukat, A. Ali, M. Javed Iqbal, M. Moinuddin, and P. Otero, "Multi-sensor fusion for underwater vehicle localization by augmentation of rbf neural network and error-state kalman filter," *Sensors*, vol. 21, no. 4, pp. 1149–1168, 2021.

[19] B. Li, Y. Xian, D. Zhang, J. Su, X. Hu, and W. Guo, "Multi-sensor image fusion: A survey of the state of the art," *Journal of Computer and Communications*, vol. 9, no. 6, pp. 73–108, 2021.

[20] L.-X. Luo, "Information fusion for wireless sensor network based on mass deep auto-encoder learning and adaptive weighted D-S evidence synthesis," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 2, pp. 519–526, 2020.

[21] W. Lu, M. Zeng, and H. Qin, "Intelligent navigation algorithm of plant phenotype detection robot based on dynamic credibility evaluation," *International Journal of Agricultural and Biological Engineering*, vol. 14, no. 6, pp. 195–206, 2021.

[22] K. Wang, C. Cao, S. Ma, and F. Ren, "An optimization-based multi-sensor fusion approach towards global drift-free motion estimation," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 12228–12235, 2021.

[23] R. R. Nair and T. Singh, "A multi-resolution approach," *IET Image Processing*, vol. 13, no. 9, pp. 1447–1459, 2019.

[24] M. S. Safizadeh and A. Golmohammadi, "Ball bearing fault detection via multi-sensor data fusion with accelerometer and microphone," *Insight - Non-Destructive Testing and Condition Monitoring*, vol. 63, no. 3, pp. 168–175, 2021.

[25] S. Ma, Y. Yuan, J. Wu, Y. Jiang, B. Jia, and W. Li, "Multisensor decision approach for HVCB fault detection based on the vibration information," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 985–994, 2021.