symmetry

MDPI

*Article*

# An Intelligent Vision-Based Tracking Method for Underground Human Using Infrared Videos

**Xiaoyu Li [1,2], Shuai Wang [1,3,*], Wei Chen [1,4,*], Zhi Weng [2], Weiqiang Fan [2] and Zijian Tian [1]**

[1] School of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), Beijing 100083, China
[2] School of Electronic Information Engineering, Inner Mongolia University, Hohhot 010021, China
[3] Inner Mongolia Bureau of the National Mine Safety Administration, Hohhot 010010, China
[4] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221100, China
* Correspondence: zqt2000407167@student.cumtb.edu.cn (S.W.); chenw@cumt.edu.cn (W.C.)

**Abstract:** The underground mine environment is dangerous and harsh, tracking and detecting humans based on computer vision is of great significance for mine safety monitoring, which will also greatly facilitate identification of humans using the symmetrical image features of human organs. However, existing methods have difficulty solving the problems of accurate identification of humans and background, unstable human appearance characteristics, and humans occluded or lost. For these reasons, an improved aberrance repressed correlation filter (IARCF) tracker for human tracking in underground mines based on infrared videos is proposed. Firstly, the preprocess operations of edge sharpening, contrast adjustment, and denoising are used to enhance the image features of original videos. Secondly, the response map characteristics of peak shape and peak to side lobe ratio (PSLR) are analyzed to identify abnormal human locations in each frame, and the method of calculating the image similarity by generating virtual tracking boxes is used to accurately relocate the human. Finally, using the value of PSLR and the highest peak point of the response map, the appearance model is adaptively updated to further improve the robustness of the tracker. Experimental results show that the average *precision* and *success* rate of the IARCF tracker in the five underground scenarios reach 0.8985 and 0.7183, respectively, and the improvement of human tracking in difficult scenes is excellent. The IARCF tracker can effectively track underground human targets, especially occluded humans in complex scenes.

**Keywords:** computer vision; human tracking; infrared videos; response map; appearance model

check for updates

**Citation:** Li, X.; Wang, S.; Chen, W.; Weng, Z.; Fan, W.; Tian, Z. An Intelligent Vision-Based Tracking Method for Underground Human Using Infrared Videos. *Symmetry* **2022**, *14*, 1750. https://doi.org/10.3390/sym14081750

Academic Editors: João Ruivo Paulo, Cristina P. Santos and Gabriel Pires

Received: 28 July 2022
Accepted: 19 August 2022
Published: 22 August 2022

## 1. Introduction

Human safety monitoring based on computer vision technology is of great significance to the development of informatization and intelligence in modern industries. Human tracking detection is a necessary prerequisite for realizing personnel identification based on symmetrical image features of human organs, such as face, iris, and gait [1–4], and also lays the foundation for the research work in related fields such as personnel indoor positioning, artificial intelligence, and human activity recognition [5,6]. Human tracking aims to predict and locate the human's position in subsequent frames based on the initial human position in the first frame of the video, so as to estimate the movement track of the tracked human. After decades of development, human tracking technology has been widely used in the fields of security surveillance, human-robot interactions, sporting events, traffic monitoring [7–10], and has made great progress.

Human tracking is one of the important means of locating and detecting personnel targets in underground mines continuously in video sequences. By tracking the human whereabouts in real time, it can help to monitor where the person is walking and the safety of personnel. If a dangerous situation occurs, the safety monitoring system can be coordinated in the fastest and most efficient way to issue early warning signals, which is very important for ensuring personnel's safety. In addition, human identification using symmetric biometrics is more direct and convenient than other asymmetric human biometrics, so it has rapidly developed into a popular personnel identification method. More importantly, human tracking and detection is a fundamental premise for human identification based on symmetrical biometrics.

However, environmental characteristics such as complex backgrounds, loud noise, dust diffusion and poor lighting conditions in underground mines [11–14] make it difficult for existing methods to obtain robust human tracking results. Specifically: there is much electrical equipment in underground mines, the space for human activities is narrow, making it easy for personnel to be blocked or even lost; the interference factors such as dust particles, fine water mist, and thermal noise of imaging equipment in the environment often make the underground video images not clear enough, the feature similarity between human and background is high, and it is difficult for the tracker to accurately distinguish humans from backgrounds; the underground lighting conditions are limited and the overall light is dim, which makes the chromaticity of the personnel target insufficient, although the miner's lamp on the helmet can play a good role in supplementing the light, due to its strong local illumination, the global illumination of the human will be uneven and low, and it is difficult for the tracker to accurately capture the human. The above-mentioned situations are the main difficulties faced by human tracking in underground mines, which makes it difficult for the traditional trackers to successfully track underground humans using visible light videos.

In order to overcome the difficulties caused by the above problems to underground human tracking, this paper aims to develop a human tracker that is not affected by special light conditions and physical environment underground, and has high robustness to complex situations, such as crowding occlusion among humans and occlusion interference of production equipment. In this respect, this paper proposes an improved aberrance repressed correlation filter (IARCF) tracker for human tracking in underground mines based on thermal infrared videos, which has the advantages of simple implementation, high computational efficiency and timely correction of tracking errors.

Firstly, image preprocessing technologies are used to enhance the image quality of the infrared human video, so as to better extract the image features of the video frames. Secondly, according to the morphological characteristics and peak to side lobe ratio (PSLR) change of the ARCF's response map, the methods for human abnormal location discrimination and human relocation are designed to improve the tracking accuracy in each frame. Finally, in order to optimize the tracker performance, the appearance model is adaptively updated by the PSLR and the highest peak point value of the response map. The main contributions of this paper are as follows:

(1) Considering the poor imaging effect caused by dim lighting conditions in underground mines, the infrared video dataset is collected to track the underground humans, which effectively alleviates the difficulty of human tracking based on visible light images.

(2) To enhance the image features of infrared human video and improve the tracking quality of human targets, the preprocessing operations of edge sharpening, contrast adjustment and noise removal were performed with the infrared video dataset.

(3) To improve the accuracy of human tracking, the morphological characteristics of the response map are analyzed to determine whether the human positioning is abnormal, and if an abnormal location occurs, the human is relocated by shifting the tracking boxes to calculate the image similarity.

## 2. Related Work

Expanding in chronological order, the development process of target tracking can be summarized as from the early classical algorithm to the later kernel correlation filtering algorithm, and then to the deep learning algorithm.

### *2.1. Classical Tracking Algorithm*

Classical tracking methods are mainly represented by the Kalman filter, particle filter, edge detection; background subtraction and their improved forms, the researchers have proposed a variety of related tracking algorithms. For example, Srilekha et al. introduced a technique for detecting, tracking as well as counting the vehicles, in which background is updated using the Kalman filter [15]. Chu et al. proposed an innovative method that uses projected gradient to facilitate multiple-kernel tracking, which is combined with Kalman filter to form a complete automatic tracking system [16]. Obolensky et al. described a time varying extension to the Kalman filter for tracking moving objects in images, the proposed Kalman filter adapts its model at each step to better estimate the movement of the maneuvering target [17]. Majumdar et al. attempted to track the human in a cluttered environment by identifying the human body based on an RGB color model, particle filter tracking, frame difference, and also detect collisions between multiple humans [18]. Additionally, Majumdar et al. also proposed an automatic detection of moving objects and initiating the particle filter algorithm to track the objects [19]. Further to this, Majumdar et al. proposed a robust human tracking framework combined optical flow and with the particle filter [20]. Beaugendre et al. presented a robust object tracking method based on a particle filter, and the main feature in which is multi-candidate object detection based on a background subtraction algorithm combined with color and interaction features [21]. Kaur et al. first performed background modelling by taking the mean of the first $n$ frames, and then performed human detection using a background subtraction algorithm, finally completing the tracking using a Kalman filter [22]. Huang et al. presented a computational framework for robust detection, tracking, and pose estimation of faces captured by video arrays, and discussed the development of a multi-primitive skin-tone and edge-based detection module embedded in a tracking module for efficient and robust face detection and tracking [23].

### *2.2. Correlation Filter Tracking Algorithm*

The characteristic of the correlation filtering tracking algorithm is that the fast Fourier transform method is used in the filtering process, and then the time domain operation is converted to the frequency domain, which greatly improves the processing speed of the tracking algorithm. In 2010, Bolme et al. first proposed the minimum mean square error (minimum output sum of squared error, MOSSE) filter algorithm [24], and opened the era of correlation filter-based tracking algorithms. In 2012, Henriques et al. proposed a circulant structure with kernels (CSK) filter algorithm based on MOSSE [25], CSK mainly solved the problem of sample redundancy caused by sparse sampling in traditional algorithms. In 2014, Henriques et al. proposed the kernelized correlation filters (KCF) tracking algorithm after improving the CSK [26]. The kernel function is also introduced, and the unique Fourier space diagonalization property of the circulant matrix can simplify the calculation, thus greatly improving the processing efficiency and robustness of the algorithm. In the same year, Danelljan et al. proposed a discriminative scale space tracker (DSST) to overcome the change problem of target scale during motion [27]. In 2015, in order to solve the problem of boundary effects, Danelljan et al. proposed the spatially regularized discriminative correlation filter (SRDCF) [28], the SRDCF algorithm uses spatial regularization to increase the weight constraint on the filter boundary function, which optimizes the tracking accuracy. In 2018, Li et al. introduced temporal regularization into the single-sample SRDCF, and further proposed the spatial-temporal regularized correlation filters (STRCF) [29], which effectively solves the problem of boundary effects in the cyclic sampling process of correlation filtering. In 2019, in order to deal with the problem of increasing background interference information, due to the expansion of the search area, Huang et al. introduced a variation suppression in regular terms into the tracking model, and proposed a tracker named aberrance repressed correlation filter (ARCF) [30]. In 2020, Wang et al. proposed a novel ARCF tracker by introducing a temporal regularization into ARCF tracker [31], which can not only make the filter template retain the historical information for filter learning, but also achieve a long-time and high precision tracking. Besides, according to different application requirements, researchers have also carried out different forms of improvement around these mainstream correlation filter trackers.

### *2.3. Deep Learning Tracking Algorithm*

With the rise of deep learning methods in the field of video image processing, various target tracking algorithms related to deep networks have also sprung up. Representative research results include Nam et al. took the lead in proposing a multi-domain learning network model (MDNet) based on part of the VGG network structure, MDNet convert the tracking task into a two-classification problem of target and background. Meanwhile, it also causes the network to be easily interfered by objects with high similarity to the target, and the processing effect of occlusion is not good enough [32]. In response to this drawback, Fan et al. proposed a structure-aware network (Sanet) algorithm, Sanet uses a recurrent neural network (RNN) to model the target structure to improve its robustness [33]. In addition, Nam et al. also leveraged multiple convolutional neural networks (CNN) to build a tree-structured CNN (TCNN) tracking algorithm, which helps to achieve multi-modality and reliability of object appearance [34]. Subsequently, the emergence of fully convolutional Siamese networks (SiamFC) [35] and Siamese instance search network for tracking (SINT) [36] introduced Siamese neural networks into the field of object tracking. With the in-depth development of the Siamese network, the target tracking has been greatly improved, compared with the previous methods. Further, for the problem of insufficient target samples and unbalanced positive and negative samples, Song et al. proposed an adversarial learning target tracking algorithm of visual tracking via adversarial learning (VITAL) based on MDNet [37]. After that, as new techniques associated with deep learning continue to emerge, researchers have also developed other forms of object tracking algorithms.

### *2.4. Mine Personnel Tracking Algorithms*

In terms of human tracking in underground mines, Zhou et al. proposed a tracking method for scene information fusion to improve the accuracy and robustness of human tracking in complex lighting environments [38], that is, a new particle filter framework was designed. Wang et al. proposed the SI-Camshift algorithm based on surf features and improved Camshift model, the surf feature plays the role of relocating the search window when the object is lost, the improved Camshift model can obtain better tracking results without the interference of non-uniform illumination and similar color [39]. Jiang et al. proposed an improved algorithm based on principal component analysis-scale invariant feature transform (PCA-SIFT) and Meanshift for the complex and changeable environment in coal mines [40], it uses the scale invariance of the PCA-SIFT to establish a target tracking model, and then uses the Meanshift method to track the moving human.

## 3. ARCF Tracker

The basic principle of the correlation filter is to calculate the similarity of two target signals through the convolution operation. The higher the response output of the correlation filter, the higher the similarity of the two target signals. So, the tracking algorithm takes advantage of this point to track objects. First, the correlation filter is trained according to the target samples given in the previous frame, and then the response output of the candidate sub-image region in the subsequent frames is calculated. The candidate sub-image region with the highest response output is the new position of the tracking target.

Usually, the correlation filter tracking method uses the circulant matrix method to collect a large number of positive and negative samples to train the correlation filter, and according to the diagonalizability of circulant matrices, the calculation is converted from the time domain to the frequency domain to further improve the operation speed. The mathematical idea of the traditional correlation filter tracker is to find a suitable filter $w$ by solving the minimum mean square error between the expected output $y$ and the actual response output. So, the idea of a traditional correlation filter tracker is as follows:

Given a $D$-dimensional training sample $x$ and an ideal response map $y \in \mathbb{R}^N$,$x^d \in \mathbb{R}^N (d = 1, 2, \cdots, D)$, $D$ represents the number of sample $x$ collected for correlation filter training, $y$ is represented by a Gaussian vector. By introducing the regularized least squares and the kernel function, the traditional tracker model can be trained by

$$\varepsilon(w_k) = \left\| y - \sum_{d=1}^{D} f(x_k^d) \right\|_2^2 + \lambda \sum_{d=1}^{D} \left\| w_k^d \right\|_2^2 \tag{1}$$

where $k$ represents the $k$th frame in the video, $w^d \in \mathbb{R}^M$ represents the correlation filters trained in the $d$th channel, $\lambda$ represents the regularization coefficient to prevent sample training from overfitting.

However, the traditional correlation filter tracker often suffers from boundary effects due to the limited search region originating from its periodic shifting of the area near original object [30]. This can easily cause aberrance tracking, which directly affects the accuracy of the tracking result. To suppress aberrance tracking during training, the training objective of the ARCF tracker is optimized according to Equation (2) to minimize the loss function. The specific form is that the aberrance penalty term is added on the basis of Equation (1).

$$\varepsilon(w_k) = \frac{1}{2} \left\| y - \sum_{d=1}^{D} \mathbf{B} x_k^d * w_k^d \right\|_2^2 + \frac{\lambda}{2} \sum_{d=1}^{D} \left\| w_k^d \right\|_2^2 + \frac{\gamma}{2} \left\| \sum_{d=1}^{D} (\mathbf{B} x_{k-1}^d * w_{k-1}^d) [\psi_{p,q}] - \sum_{d=1}^{D} \mathbf{B} x_k^d * w_k^d \right\|_2^2 \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{M \times N}$ represents the cropping matrix used to select the center of each channel $x^d$ of the input vectorized sample, $k$ and $(k-1)$ represent the $k$th frame and the $(k-1)$th frame in the video respectively, $\mathbf{B} x_k^d * w_k^d$ represents the response map of the $k$th frame, $\mathbf{B} x_{k-1}^d * w_{k-1}^d$ represents the response map of the $(k-1)$th frame, $\gamma$ denotes the aberrance penalty parameter to control the suppression of aberrance tracking, $p$ and $q$ represent the peak positions of the two response maps in the two-dimensional space respectively, $[\psi_{p,q}]$ represents the shift operation performed to make the two peaks match each other. The third term on the right side of the Equation (2) is the regularization term used to restrict the aberrance.

First, as described above, to further improve the computational efficiency, Equation (2) is converted to the frequency domain space according to

$$\begin{cases} \hat{\varepsilon}(w_k, \hat{g}_k) = \frac{1}{2} \left\| \hat{y} - \hat{\mathbf{X}}_k \hat{g}_k \right\|_2^2 + \frac{\lambda}{2} \| w_k \|_2^2 + \frac{\gamma}{2} \left\| \hat{M}_{k-1}^s - \hat{\mathbf{X}}_k \hat{g}_k \right\|_2^2, \\ \quad s.t. \quad \hat{g}_k = \sqrt{N} (\mathbf{I}_D \otimes \mathrm{F} \mathbf{B}^H) w_k \end{cases} \tag{3}$$

where the variable with the superscript ˆ represents the signal that has been Fourier transformed, $\hat{g}_k \in \mathbb{C}^{DN \times 1}$ represents the parameters introduced for further optimization, $\hat{M}_{k-1}^s$ represents the discrete Fourier transform of the shifted signal that can be regarded as a constant signal.

Next, the augmented Lagrangian equation of Equation (3) is iteratively solved with the alternative direction method of multipliers (ADMM), that is, by converting it into the form of solving two sub-problems of $w_{k+1}^*$ and $\hat{g}_{k+1}^*$, the correlation filter of ($k$ + 1)th video frame is calculated. The solution results of $w_{k+1}^*$ and $\hat{g}_{k+1}^*$ are

$$w_{k+1}^* = (\lambda + \mu N)^{-1} (\sqrt{N} (\mathbf{I}_D \otimes \mathbf{B} \mathrm{F}^H) \hat{\xi} + \mu \sqrt{N} (\mathbf{I}_D \otimes \mathbf{B} \mathrm{F}^H) \hat{g}_k) = (\frac{\lambda}{N} + \mu)^{-1} (\xi + \mu g_k), \tag{4}$$

$$\begin{aligned} \hat{g}_{k+1}(n)^* \qquad &= \gamma^* (\hat{x}_k(n) \hat{y}(n) + \gamma \hat{x}_k(n) \hat{\mathrm{M}}_{k-1}^s - \hat{\xi}(n) + \mu \hat{w}_k(n)) \\ &- \gamma \tfrac{\hat{x}_k(n)}{b} (\hat{\mathbf{S}}_{xk}(n) \hat{y}(n) + \gamma \hat{\mathbf{S}}_{xk}(n) \hat{\mathrm{M}}_{k-1}^s \hat{\mathbf{S}}_{\xi}(n) + \mu \hat{\mathbf{S}}_{wk}(n)) \end{aligned} \tag{5}$$

Subsequently, the Lagrangian parameters are updated by

$$\zeta_{k+1}^{(i+1)} = \zeta_{k+1}^{i} + \mu(\hat{g}_{k+1}^{*(i+1)} - \hat{w}_{k+1}^{*(i+1)}), \tag{6}$$

where the superscript $i$ and ($i$ + 1) represent the $i$th and ($i$ + 1)th ADMM optimization iterations respectively, $\hat{g}_{k+1}^{*(i+1)}$ represents the solution of $\hat{g}_{k+1}^{*}$ in the ($i$ + 1)th iteration, $\hat{w}_{k+1}^{*(i+1)}$ represents the solution of $w_{k+1}^{*}$ in the ($i$ + 1)th iteration, $\hat{w}_{k+1}^{*(i+1)} = (\mathbf{I}_D \otimes \mathrm{F}\mathbf{B}^H) w_{k+1}^{*(i+1)}$.

Finally, ARCF updates the appearance model by

$$\hat{x}_{k+1}^{\mathrm{M}} = \alpha \hat{x}_{k}^{\mathrm{M}} + \beta \hat{x}_{k+1}, \tag{7}$$

where $k$ and $k$ + 1 represent the $k$th frame and the ($k$ + 1)th frame respectively, $\alpha$ and $\beta$ represent the learning rate of the appearance model.

## 4. Improved ARCF (IARCF) Tracker

### *4.1. Preprocessing of Infrared Personnel Video*

Infrared video mainly relies on the thermal radiation effect of objects for imaging. Tracking humans based on infrared videos can effectively overcome the influence of non-uniform lighting conditions in underground mines. However, the infrared video has shortcomings such as obvious noise, low contrast, blurred edges, etc., which directly affect the accuracy of human tracking without the complement of color features. Therefore, in order to improve the visual effect of the infrared human videos in underground mines, this paper uses preprocessing methods to enhance the image features of infrared video, which includes un-sharp masking for edge sharpening, edge sensing for local contrast adjustment, and NL-means for denoising.

As shown in Figure 1, after the original infrared human image is sharpened by the un-sharp filter operation, the blurred boundary pixels in the image are sharpened. Then, the local contrast adjustment operator focusing on edge perception further improves the brightness of the boundary contour in the infrared human image, which helps to identify the human target and background in the image more clearly. In addition, since the infrared human video will bring degradation factors during the imaging process, mainly including thermal noise generated by imaging equipment and environmental noise generated by underground dust particles, so the NL-means operator is used to denoise at the end, which can effectively improve the quality of infrared human video images.

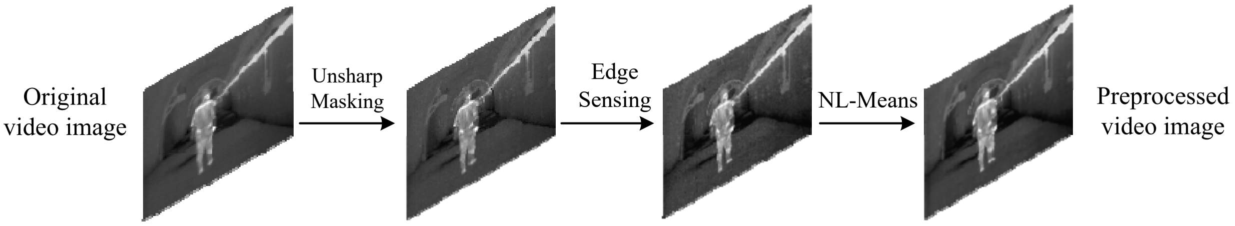
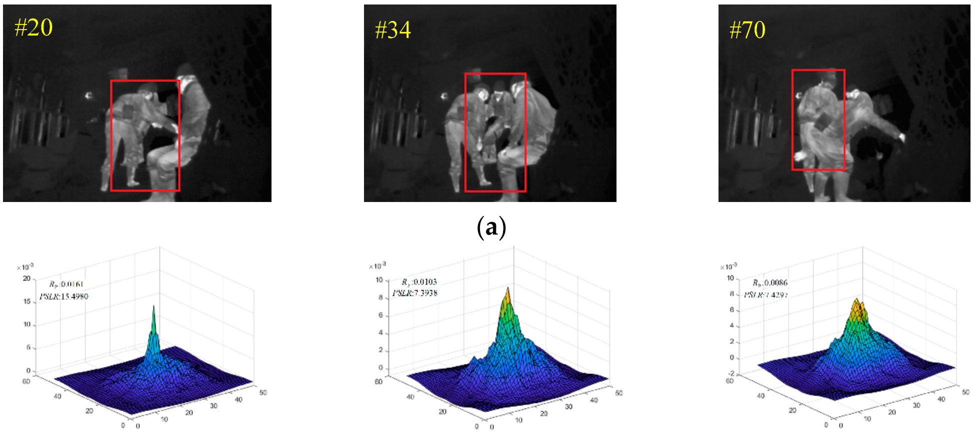


**Figure 1.** Preprocessing of infrared human video.

### *4.2. Abnormal Location Identification*

For the traditional discriminant correlation filter (DCF) trackers, the important basis for locating the target in each frame is the maximum confidence of the response map, that is, the highest peak point $R_p$ of the response map. Ideally, the DCF tracker considers the position box in the video frame corresponding to $R_p$ as the positioning result of the tracked target. However, the underground environment is special, the space available for human activities is limited, and the production equipment is widely distributed. And so, the human is easily disturbed by the occlusion problem during tracking, including the occlusion between human and the occlusion between human and other objects.

As shown in the first column of Figure 2a, if the human is not disturbed by other similar thermal objects or occlusion during movement, an obvious sharp peak will appear in the response map, and the confidence at the peak is relatively large. In this case, the red positioning box at frame #20 corresponding to $R_p$ is basically consistent with the actual human location. On the contrary, as shown in the second column and third column of Figure 2a, if the human is disturbed by other similar thermal objects or occlusion during the movement, multiple peaks will appear in the response map, the confidence at the main peak is relatively small and the shape of the main peak is blunt. In this case, the main peak at frame #34 is a pseudo peak, and the red positioning box at frame #34 and frame #70 corresponding to their $R_p$ is basically not consistent with the actual human location.



(**a**)

(**b**)

**Figure 2.** Human location and response map. (**a**) Human location result; (**b**) Peak shape of the response map.

From the above analysis, it can be seen that the traditional method of locating the target based on $R_p$ is not completely suitable for tracking underground humans. So, after testing and analyzing a large number of infrared mine video sequences, we find that once the human is disturbed by environmental factors during the tracking, the response map of the current frame will be significantly different in *PSLR* and the peak shape *Rs* compared with the previous historical frames. The calculation method of *PSLR* [24] is defined as

$$PSLR = \frac{m - a}{d}, \tag{8}$$

where $m$ represents the maximum peak of the response map, $a$ and $d$ represent the mean and standard deviation of all response values, respectively.

It can be seen from Figure 2a that the tracker shows obvious drift at frame #34, correspondingly, the *PSLR* value decreased at frame #34 compared with frame #20 in Figure 2b, besides, the *Rs* appears multi-peaked and the main peak becomes obviously thicker at frame #34 and frame #70 compared with frame #20 in Figure 2.

Figure 3 shows the change curve of the *PSLR* value and its historical average multiplied by a fixed factor with human tracking sequences corresponding to Figure 2, it can be seen that *PSLR* almost drops to the lowest compared with the previous historical frames and is below the historical average at frame #34. Figure 4 shows the two-dimensional forward plan of the response map, where the red point is the highest peak $R_p$, the yellow point is the second highest peak $R_s$. Draw horizontal lines $l1$ and $l2$ through $R_p$ and $R_s$ respectively, and record the height difference between them as $h$. Draw vertical lines $l3$ and $l4$ through the intersection points $p1$ and $p2$ obtained from the outermost contour of the main peak $R_p$ where $l2$ is located, and record the width difference between them as $w$. In this paper, $R_s$ is defined as the ratio of $h$ to $w$, that is, $Rs = h/w$. Based on the above discussion, it can be seen that the larger $Rs$ is, the better the peak shape of the response map is, and the higher the positioning accuracy of human tracking is. On the contrary, the smaller *Rs* is, the worse the peak shape of the response map is, and the lower the positioning accuracy of human tracking is. Figure 5 shows the change curve of *Rs* value with human tracking results in the video sequence corresponding to Figure 2, similar to Figure 3, the *Rs* drops to the lowest value at frame #34 compared with the previous historical frames.

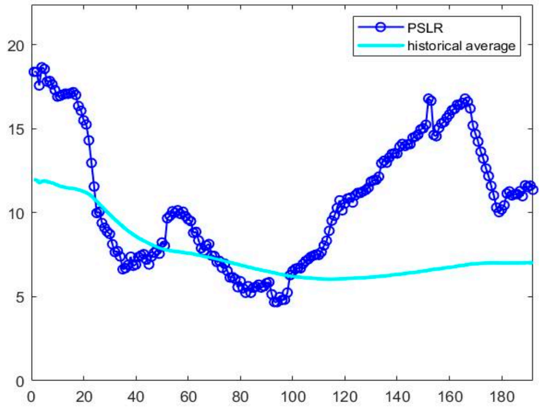


**Figure 3.** PSLR change curve of the response map.

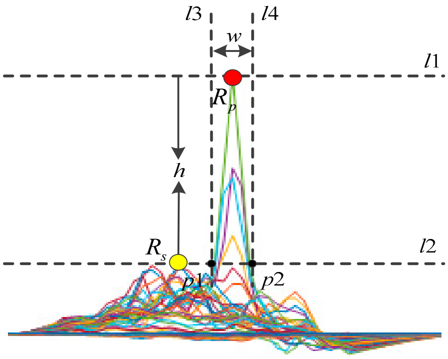


**Figure 4.** 2D peak shape model for response map.

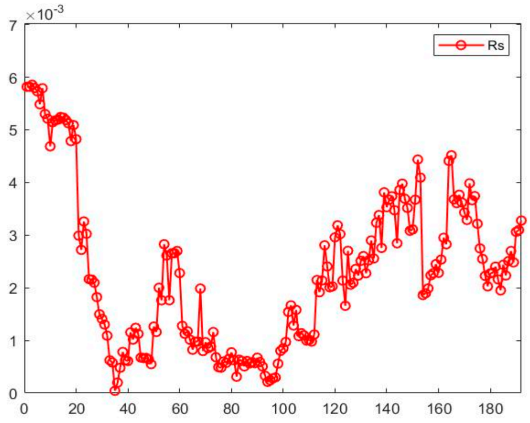


**Figure 5.** *Rs* change curve of the response map.

In summary, the conditions for judging the abnormal human positioning in this paper are defined as

$$\begin{cases} Rp^k < \chi_1 \text{mean}(Rp^h) \\ Rs^k < T \end{cases}, \tag{9}$$

where $Rp^k$ represents the $PSLR$ value of the current frame, $\chi_1$ represents the factor, mean$(\cdot)$ represents the averaging function, $Rp^h$ represents the historical average of $PSLR$ value before $k$th frame, $Rs^k$ represents the $Rs$ of the current frame, $T$ represents a fixed threshold.

### *4.3. Human Relocation*

If the two indicators of $Rp^t$ and $Rs^t$ in current frame meets the judgment conditions for abnormal positioning, it means that the human box $P_c$ in the frame corresponding to $R_p$ of the current response map is not the actual human location, and a more accurate human positioning box $R_a$ needs to be relocated in the current frame. Human relocation is completed based on the $P_c$ through two links of positioning box translation and positioning box filtering, as shown in Figure 6.

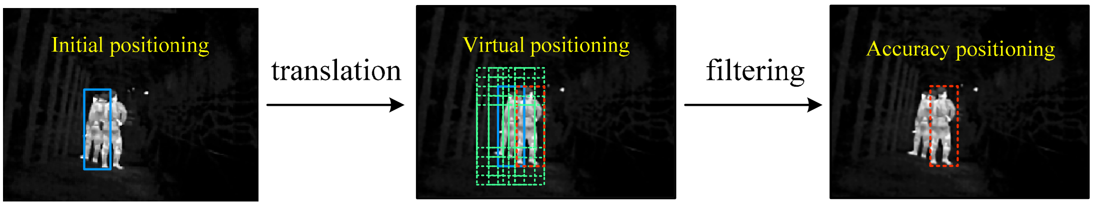


**Figure 6.** Process of human relocation.

Combining with the human positioning at frame #34 in Figure 2a, it can be seen that even if there appears to be abnormal positioning, the location of the positioning box will not deviate too far from the actual human position, and it is basically located in the vicinity of the actual human position. Therefore, positioning box translation is to move $P_c$ along the horizontal and vertical directions with a certain step size, where the horizontal step size and the vertical step size are consistent with the aspect ratio of $P_c$, so that a series of virtual candidate boxes $P_v$ can be generated. In addition, due to the cramped space for human activities, the dangerous environment, and the complex road conditions in underground mines, human movement speed is relatively slow, and the changes of human characteristics between two consecutive frames are very tiny. So, positioning box filtering is to calculate the similarity between the sub-image $I_p$ of $R_a$ in the previous frame and the sub-image $I_c$ of $P_v$ in the current frame, and filter out the $I_c$ with the highest similarity with $I_p$, then $R_a$ in the current frame can be obtained.

The filtering basis of $R_a$ in this paper is composed of three parts: the histogram cosine operator $Hc$, the hash operator $Ha$, and the $Hu$ moment. Among them, the feature similarity $S_c$ and $D_a$ of $Hc$ and $Ha$ is represented by the similarity value of $I_p$ and $I_c$ respectively, and the feature similarity $S_m$ of $Hu$ moment is represented by the absolute difference between the $Hu$ moment features of $I_p$ and $I_c$.

First, the calculation process of $Hc$ is as follows:

1. Get the grayscale image histograms $h_p$ and $h_c$ of $I_p$ and $I_c$;
2. Divide $h_p$ and $h_c$ into 64 intervals $h_p^i$ and $h_c^i$ $(i = 1, 2, \cdots, 64)$, where $h_p^i$ and $h_c^i$ contain four consecutive gray levels;
3. Sum up the four gray levels in $h_p^i$ and $h_c^i$, and get the $1 \times 64$ fingerprint vectors $I_p$ and $I_c$ of $V_p$ and $V_c$;
4. Calculate the cosine similarity $S_c$ of $V_p$ and $V_c$ (the smaller $S_c$, the greater the image similarity).

Then, the calculation process of $Ha$ is as follows:

1. Resize $I_p$ and $I_c$ into $8 \times 8$ or $32 \times 32$, and convert them to grayscale images;
2. Calculate $I_p$ and $I_c$ through discrete cosine transform (DCT) compression to obtain $I_p^t$ and $I_c^t$;
3. Compare the grayscale values of $I_p^t$ and $I_c^t$ with the average DCT value, and obtain the $1 \times 64$ hash codes $C_p$ and $C_c$ corresponding to $I_p$ and $I_c$, respectively;

4. Calculate the Hamming distance $D_a$ of $C_p$ and $C_c$ (the smaller the $D_a$, the greater the image similarity).

Finally, the computational principle of *Hu* moment [41] is as follows:

For the image *f*, whose pixel coordinates can be defined as two-dimensional random variables $(x, y)$, and *f* can be further expressed as the two-dimensional density function $f(x, y)$, its ($p$ + $q$) order geometric moment is defined as

$$m_{pq} = \sum_{x=1}^{A} \sum_{y=1}^{B} x^p y^q f(x, y), \tag{10}$$

$$\mu_{pq} = \sum_{x=1}^{A} \sum_{y=1}^{B} (x - x_0)^p (y - y_0)^q f(x, y), \tag{11}$$

where the $A$ and $B$ represent the width and height of the grayscale image respectively, and $x_0$ and $y_0$ represent the centroid coordinates of the grayscale image respectively. The normalized center distance is defined as

$$\eta_{pq} = \mu_{pq}/(\mu_{00}^{\rho}), \;\; \rho = (p + q)/2 + 1. \tag{12}$$

Seven invariant moments $M_1$~$M_7$ are formed by the second-order and third-order normalized central moments to describe the contour of *f*, which are defined as

$$\begin{aligned}
&M_1 = \eta_{20} + \eta_{02} \\
&M_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
&M_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
&M_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
&M_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right) + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right) \\
&M_6 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})(\eta_{21} + \eta_{03}) \\
&M_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + (3\eta_{21} - \eta_{30})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]
\end{aligned} \tag{13}$$

In this paper, the $S_m$ of $I_p$ and $I_c$ is defined as

$$S_m = \text{abs}(\sum_{i=1}^{7} (M_i^p - M_i^c)), \tag{14}$$

where $M_i^p$ and $M_i^c$ represent the *i*-order moment of $I_p$ and $I_c$, and $\text{abs}(\cdot)$ represents the function to get the absolute value.

To sum up, for each $P_v$ participated in the similarity calculation of the current frame, a set of feature vectors $Q^j (j = 1, 2, \cdots, N)$ composed of three elements of $S_c$, $D_a$, and $S_m$ will be obtained, that is, $Q^j = \left\langle S_c^j, D_a^j, S_m^j \right\rangle$, $N$ is the number of $P_v$. In this paper, the $I_c$ with the smallest sum of $Q_j$ is selected as the optimal similar human sub-image, and the corresponding $P_v$ is the accurate result of human relocation.

### 4.4. Adaptive Appearance Model Update

For the visual tracking, the establishment and update of the target appearance model also plays an important role for the tracker performance. Currently, the strategy of linearly updating the target appearance model via Equation (7) is commonly used. It can be considered as the superposition result of the original appearance model and the new appearance model, so that the DCF filters can be updated in real time with the change in the target appearance characteristics, so as to improve the tracking effect.

Nevertheless, updating the human appearance with Equation (7) is not sufficient for application in the underground mine. Specifically, Equation (7) updates the appearance model of each frame, in fact, for different appearance changes, the update requirement is not the same. For example, when the tracked human is occluded by equipment or there is interference by other humans, the tracker does not need to update the appearance model, otherwise the filter will learn more background-independent information, which will eventually lead to tracking failure. On the contrary, when the tracked human has undergone significant morphological changes in the two frames before and after, the tracker needs to update the appearance model in time to locate the human more accurately.

Figure 7 shows the change curve of $R_p$ and its historical average multiplied by a fixed factor, with human tracking sequences corresponding to Figure 2. Similarly, it can be seen that $R_p$ almost drops to its lowest value compared with the previous historical frames and is below the historical average at frame #34.

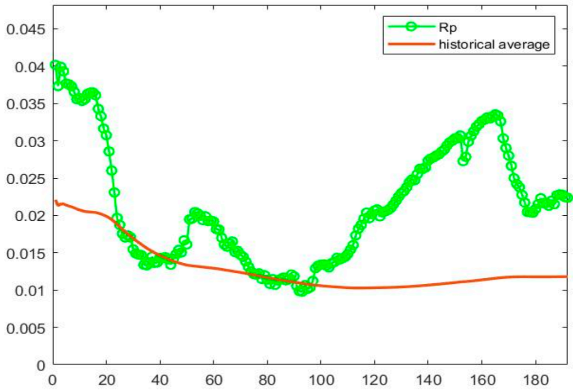


**Figure 7.** Change curve of the response map.

Combined with the above discussion, for the improved ARCF tracker in this paper, the *PSLR* and $R_p$ of the response map are used as the reference indicators to update the appearance model [42]. The calculation method of model update is

$$\begin{cases} Rp^k < \chi_1 \mathrm{mean}(Rp^h) \\ Rf^k < \chi_2 \mathrm{mean}(Rf^h)' \end{cases} \tag{15}$$

where $Rf^k$ represents the $R_p$ value of the current frame, $\chi_2$ represents a coefficient, $Rf^h$ represents the historical average of $R_p$ value before $k$th frame.

Only when the values of $Rp^k$ and $Rf^k$ satisfy Equation (15) at the same time, is it considered that the response map conforms to the unimodal characteristic and the model can be updated by Equation (7), otherwise it will not be updated. Figure 8 shows the update schematic of the appearance model in this paper.

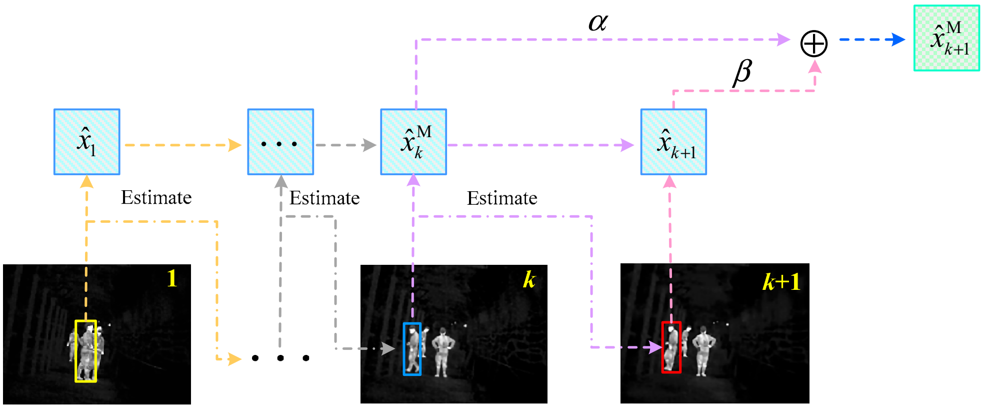


**Figure 8.** The update principle of the appearance model.

In Figure 8, the initial human positioning box represented by $\hat{x}_1(\hat{x}^{M}_{1})$ is known, $\hat{x}^{M}_{2}$ consists of $\hat{x}_1$ and the estimated sample $\hat{x}_2$ where the human location in the first frame is located in the second frame, $\hat{x}^{M}_{3}$ consists of $\hat{x}_1$, $\hat{x}^{M}_{2}$ and $\hat{x}_3$ where the human location in the second frame is located in the third frame. By analogy, it can be inferred that the template $\hat{x}_{k+1}^{\mathrm{M}}$ of the ($k$ + 1)th frame consists of $\hat{x}_1$, accumulated positioning sample $\hat{x}_k^{\mathrm{M}}$ of the $k$th frame, and estimated sample $\hat{x}_{k+1}$ where the human location in the $k$th frame is located in the ($k$ + 1)th frame.

## 5. Results

### *5.1. Dataset and Evaluation Indicators*

The dataset of infrared human video at underground mine (IHV-UM) used in this paper was collected from the Shuangma No. 1 Mine of Ningxia Coal Group. It mainly includes a total of 50 groups of video sequences in multiple scenes such as underground alleys, fully mechanized faces, working face ends, substations, refuge chambers and electromechanical rooms, which involve common tracking challenges such as human scale changes, posture change and occlusion, etc. The specific situation is as follows: IHV-UM contains eight groups of videos in the alley, eleven groups of videos in the fully mechanized mining face, nine groups of videos in the substations, twelve groups of videos in the refuge chamber, and ten groups of videos in the electromechanical chamber. The alley videos are accompanied by a change of human scale and slight human occlusion, the fully mechanized mining face videos are accompanied by a change of human posture and severe occlusion, the substation videos are accompanied by a change of human scale, the refuge chamber videos are accompanied by a change of human scale and angle changes, the electromechanical chamber videos are accompanied by human occlusion and angle changes. To verify and analyze the tracking performance of the IARCF tracker with the humans in underground mines, this paper compares the IARCF with KCF [26], DSST [27], ECO [43], BIT [44] and ARCF on the IHV-UM dataset.

For the preprocessing of IHV-UM dataset, the three indicators of peak signal-to-noise ratio (PSNR), mean square error (MSE) and energy of gradient (EOG) were used to evaluate the dataset quality in terms of noise, contrast and sharpness, whose calculation methods are

$$PSNR = 10 \times \log_{10}\left(\frac{\sum\limits_{x=1}^{L}\sum\limits_{y=1}^{W}(255)^2}{MSE}\right), \tag{16}$$

$$MSE = \frac{1}{L \times W}\sum_{x=1}^{A}\sum_{y=1}^{B}(f(x,y) - g(x,y))^2, \tag{17}$$

$$EOG = \sum_{x=1}^{L}\sum_{y=1}^{W}\left\{[f(x+1,y) - f(x,y)]^2 + [f(x,y+1) - f(x,y)]^2\right\}. \tag{18}$$

where $L$ and $W$ represent the length and width of the image, $f(x,y)$ and $g(x,y)$ represent the gray value at $(x,y)$ of the original image and the preprocessed image, respectively.

In this paper, two common indicators of *precision* and *success* rate are used to evaluate the human tracking performance of all trackers.

The *precision* means that the ratio of the frames number in which the difference between the human position predicted and the actual human position does not exceed a fixed threshold to the total number of frames. Different thresholds are set to obtain different ratios, so the *precision* of the trackers can be described by the change curve. The calculation method of *precision* is

$$\begin{cases} precision = \sum\limits_{i=1}^{N_t}\frac{f_p}{N_t} \\ s.t.\ \ f_p = \begin{cases} 0, CLE \leq T_p \\ 1, CLE \geq T_p \end{cases} \\ CLE = \mathrm{Eucl}(p_1, p_2) \end{cases}, \tag{19}$$

where $f_p$ represents the video frame, $N_t$ represents the total number of frames in the video, $T_p$ represents the fixed threshold, $\mathrm{Eucl}(\cdot)$ represents the Euclidean distance function between two points of $p_1$ and $p_2$, $p_1$ and $p_2$ represent the human center position predicted by the tracker and the actual human center position, respectively.

The *success* rate means that the ratio of frames number in which the overlapping pixel area of the human position box predicted, and the actual human position box does not exceed a fixed threshold to the total number of frames. In the same way, different ratios can be obtained by setting different thresholds, so the *success* rate of the tracker can be described by the change curve. The calculation method of success is

$$\begin{cases} success = \frac{\sum\limits_{i=1}^{N_t} f_s}{N_t} \\ s.t.\ \ f_s = \begin{cases} 0, S \le T_s \\ 1, S \ge T_s \end{cases} \\ S = \frac{\mathrm{Area}(a \cap b)}{\mathrm{Area}(a \cup b)} \end{cases}, \tag{20}$$

where $f_s$ represents the video frame, $N_t$ represents the total number of frames in the video, $T_s$ represents the fixed threshold, $\mathrm{Area}(\cdot)$ represents the pixel area function about personnel position box, $S$ represents the proportion of $\mathrm{Area}(\cdot)$ about the intersection of $a$ and $b$ and the union of $a$ and $b$, $a$ and $b$ represent the human position box predicted by the tracker and the actual human position box, respectively.

### *5.2. Experimental Parameters*

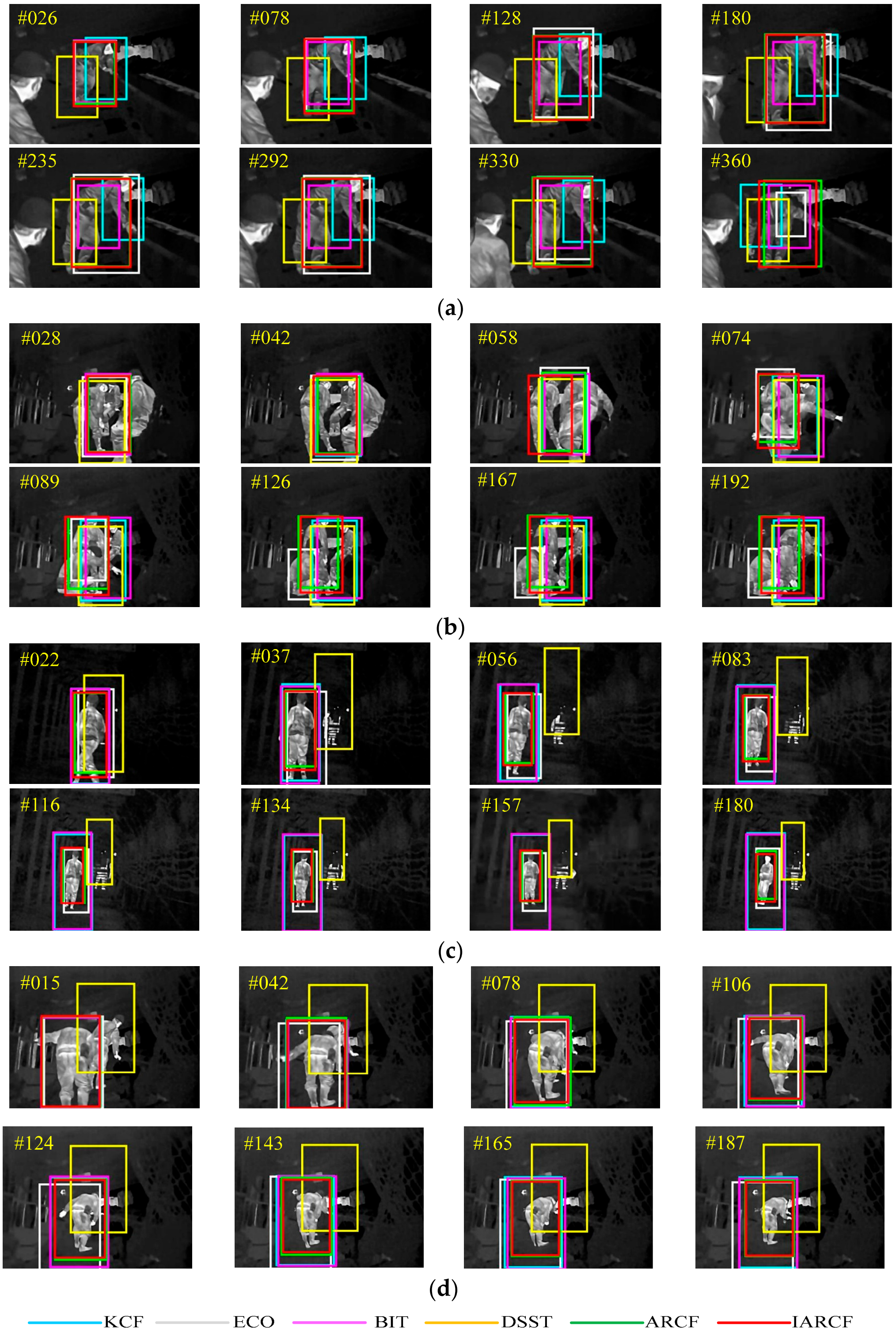
The IARCF tracker in this paper is implemented with Matlab2021a, and the PC equipped with Intel (R) Core (TM) i5-1035G1 CPU, 4 GB RAM, 64-bit Windows10. In the experiment, the regularization coefficient $\lambda$ is set to 0.0001, the aberrance penalty parameter $\gamma$ is set to 0.71, the number of ADMM iterations is set to 5, the learning rate $\alpha$ is set to 0.03, and the learning rate $\beta$ is set to 0.07, $\chi_1$ is set to the range between 0.65–0.87, $\chi_2$ is set to 0.55.

The average EOG of the original IHV-CM dataset is 1.6495607 × 107, after being preprocessed, the average EOG rise to 2.574393 × 107, and the average PSNR and MSE is 27.7328 and 109.5971, respectively.

### *5.3. Qualitative Experiments and Discussion*

Figure 9 shows the qualitative experimental results of different trackers on the algorithm performance of four video sequences in the IHV-UM dataset. The first video sequence in Figure 9a contains a total of 360 frames, the human target is interfered with by a thermal radiation from an object. The second video sequence in Figure 9b contains a total of 192 frames, the human target is interfered with by human occlusion. The third video sequence in Figure 9c contains a total of 180 frames, the human target is lightly interfered with by human occlusion. The fourth video sequence in Figure 9d contains a total of 187 frames, the human target is interfered with by thermal radiation from an object and human occlusion.

For the first video sequence, there is a thermal interference source from the shearer cantilever, the scale of the human target changes from small to large and is accompanied by posture changes in the latter. ARCF and IARCF can locate the human target accurately by introducing a regularization term that suppresses aberrance tracking, and IARCF is little better than ARCF when using the human relocation strategy and adaptive appearance model update. Although the BIT tracker can more accurately locate the human position in each frame, the tracking area does not change with the scale of human target. The overall tracking effect of ECO is good, but when the human has an obvious posture change at the 360th frame, the tracking result of ECO appears to drift. The tracking effect of KCF and DSST on the human target is relatively poor; they have shown a certain deviation in tracking the human target at the 26th frame, and in the subsequent frame sequence, as the scale of the human target increases. The tracking area has not been extended accordingly, so the tracking error is more obvious. In addition, it can be seen from the tracking results of the 360th frame that when the human pose changes significantly, the localization and coverage of the human target by the IARCF tracking box is the best among all trackers, it shows that the IARCF tracker has stronger adaptability than the other trackers in dealing with sudden abnormal localization of the human.



(**a**)

(**b**)

(**c**)

(**d**)

**Figure 9.** Comparison of qualitative results with different trackers. (**a**) The first video sequence; (**b**) The second video sequence; (**c**) The third video sequence; (**d**) The fourth video sequence.

For the second video sequence, there is a serious occlusion problem, and the human posture changes relatively quickly. The six trackers can basically track the human accurately at the 28th frame, but from the 42nd frame to the 192th frame, the tracking effects of the six trackers gradually show differences. ARCF and IARCF are relatively stable and accurate, and the IARCF is better than ARCF. The other four trackers of KCF, ECO, BIT and DSST show different degrees of tracking errors from the 89th frame because of the difficulty in identifying the tracked human and other humans, and even mistaking other humans as the tracked human. In terms of the overall tracking effect, the IARCF tracking box performs better than other trackers in both the positioning accuracy and the area coverage accuracy of the tracked human target. This is because the human relocation function of the IARCF tracker can correct abnormal human positioning in time, and at the same time, the IARCF tracker can update the appearance model according to the actual demand by using the adaptive model update strategy, which effectively avoids the accumulation of tracking errors.

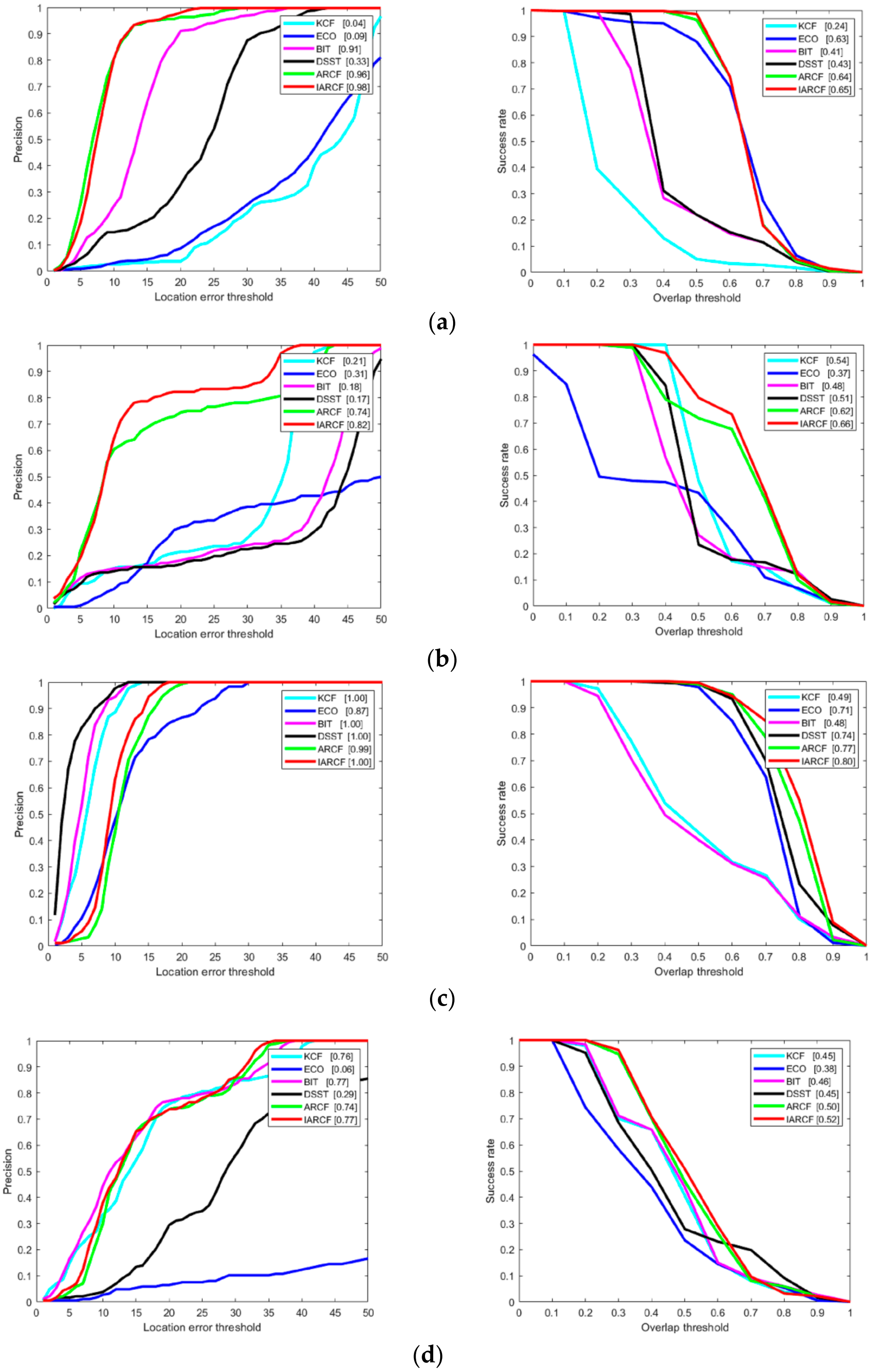
For the third video sequence, the human walks from near to far, and the target scale changes from large to small, accompanied by angular rotation changes. Since the scene conditions where the human is located are ideal, the ARCF, IARCF and ECO trackers show uniform stability and accuracy during the whole process, and the tracking scales of ARCF and IARCF are more suitable than other trackers. For the DSST, it shows a significant tracking drift phenomenon during the whole process. Although the other two trackers of KCF and BIT can locate the human target more accurately, their tracking area is much larger than the human scale. Since there are few environmental disturbance factors in this scene, the trackers of ARCF and IARCF that with better tracking performance than other trackers don't show obvious difference in visual effect. However, it can be seen from the details that the IARCF tracking box has a more accurate positioning than ARCF, which reflects the feasibility of improving ARCF for underground human tracking in this paper, and thereby, effectively improves the quality of human tracking.

For the fourth video sequence, the human also moves from near to far, and the target scale changes from large to small correspondingly. The five trackers of ARCF, IARCF, ECO, BIT and KCF can basically track the human accurately all the time, and in contrast, the tracking scales of ARCF and IARCF are still more accurate than the trackers of ECO, BIT and KCF. Similar to the third video sequence, DSST shows a significant tracking drift phenomenon during the whole process, and its tracking area basically does not change with the change of human scale. Similarly, the tracking results of IARCF and ARCF are basically the same as that of the third video sequence. Overall, the tracking effects of IARCF and ARCF are the best among all trackers, this is because the scene conditions are ideal and there is no obvious human pose change. But careful observation shows that in terms of the tracking box, positioning accuracy and coverage accuracy, IARCF still has a tracking effect that is closer to the actual position of humans than ARCF, which further illustrates that the location discrimination strategy of IARCF can play a powerful role in enhancing the human tracking performance.

On the whole, from the above analysis and discussion, it can be seen that for the four groups of human video sequences in different scenarios, the tracking effect of IARCF is always better than the other five trackers of KCF, DSST, ECO, BIT and ARCF, which shows that IARCF has the most excellent tracking performance. Meanwhile, it also confirms the necessity of considering the human anomaly localization in each frame and adaptive updating of the human appearance model according to actual needs. Additionally, it can be seen from Figure 9 that the tracking effect of IARCF in the first two complex scenarios is better than that in the last two simple scenarios, which also proves that IARCF has stronger robustness for human tracking in difficult underground scenes.

### *5.4. Quantitative Experiments and Discussion*

Figure 10 shows the quantitative experimental results of different trackers on the algorithm performance of four video sequences in the IHV-UM dataset, and the two indicators of *precision* and *success* rate are selected to evaluate first. In order to compare the tracking performance of different trackers more intuitively, the score when the location error threshold is 20 pixels is selected as the representative *precision* value, and it is presented in the legend of the *precision* figure, the score of the area under the *success* rate curve is selected as the representative *success* value; it is presented in the legend of the *success* figure.



(**a**)

(**b**)

(**c**)

(**d**)

**Figure 10.** Comparison of quantitative results with different trackers. (**a**) The first video sequence; (**b**) The second video sequence; (**c**) The third video sequence; (**d**) The fourth video sequence.

For the first video sequence shown in Figure 10a, the IARCF tracker achieves a *precision* of 98% and a *success* rate of 65%, respectively, which is 0.02% and 0.01% higher than that of ARCF tracker, and reaches the highest value of all trackers. For the second video sequence shown in Figure 10b, the IARCF tracker achieves a *precision* of 82% and a *success* rate of 66%, respectively, which is 0.08% and 0.04% higher than that of the ARCF tracker, and also reaches the highest value of all trackers. For the third video sequence shown in Figure 10c, the IARCF tracker achieves a *precision* of 100% and a *success* rate of 80%, respectively, which is 0.01% and 0.03% higher than that of the ARCF tracker. Since the scene conditions where the human target is located in the video sequence are relatively ideal, the three comparison trackers of KCF, BIT, and DSST also reach a *precision* of 100%, and only the IARCF reaches the highest *success* rate. For the fourth video sequence shown in Figure 10d, the IARCF tracker achieves a *precision* of 77% and a *success* rate of 52%, respectively, which is 0.03% and 0.02% higher than that of the ARCF tracker, BIT tracker also reaches the highest *precision* of 77%, and still only the IARCF reaches the highest *success* rate.

For the comprehensive indicator values of *precision* and *success* rate for the above four video sequences, the IARCF achieves the highest among all trackers. It means that IARCF's tracking results of human targets in underground mines are closer to the actual location of human targets, which again proves the effectiveness and feasibility of correcting abnormal human location based on the morphological characteristics of the response map. Apart from this, it is shown that the *precision* and *success* rate of the IARCF tracker are the highest in the third video sequence. This is because the human in the third video sequence is less disturbed by other humans and environmental objects, and the human appearance is basically stable in terms of angle and posture. In addition, as shown in Section 5.1, the human tracking in the second video sequence is the most complicated and difficult, and there is serious human occlusion interference. Despite this, IARCF's improvements in the *precision* and *success* rate perform best in the second video sequence compared to the ARCF tracker, this fully demonstrates that the IARCF tracker has strong robustness and adaptability in dealing with the human tracking problems in complex underground scenes.

Furthermore, to test and evaluate the tracking performance of all trackers on the whole IHV-UM dataset, Table 1 shows the global index values for the six trackers under five scenarios of alley, fully mechanized face, substation, refuge chamber and electromechanical room. The penultimate column in the table is the average index values under each scenario, and the last column is the improvement of IARCF compared to ARCF.

**Table 1.** Performance comparison of different tracking algorithms.

| Indicators | Scenes | KCF | ECO | BIT | DSST | ARCF | IARCF | Average | Improved |
|---|---|---|---|---|---|---|---|---|---|
| precision | alley | 0.9453 | 0.8515 | 0.9802 | 0.4594 | 0.9124 | 0.9647 | 0.8523 | 0.0523 |
| | fully mechanized face | 0.3625 | 0.3178 | 0.6464 | 0.6297 | 0.7028 | 0.7918 | 0.5752 | 0.0890 |
| | substation | 0.6452 | 0.7235 | 0.8800 | 0.3851 | 0.9255 | 0.9436 | 0.7505 | 0.0181 |
| | refuge chamber | 0.7984 | 0.4710 | 0.5227 | 0.7838 | 0.8819 | 0.9094 | 0.7279 | 0.0275 |
| | electromechanical room | 0.8554 | 0.4211 | 0.7339 | 0.4316 | 0.8358 | 0.8829 | 0.6934 | 0.0471 |
| success | alley | 0.6124 | 0.7538 | 0.6242 | 0.4778 | 0.8057 | 0.8278 | 0.6836 | 0.0221 |
| | fully mechanized face | 0.5878 | 0.5696 | 0.5719 | 0.6255 | 0.5634 | 0.6116 | 0.5883 | 0.0582 |
| | substation | 0.4616 | 0.5041 | 0.5321 | 0.5439 | 0.6994 | 0.7325 | 0.5789 | 0.0331 |
| | refuge chamber | 0.4449 | 0.5805 | 0.4818 | 0.7412 | 0.7298 | 0.7514 | 0.6216 | 0.0216 |
| | electromechanical room | 0.6284 | 0.6628 | 0.5262 | 0.4202 | 0.6518 | 0.6684 | 0.5930 | 0.0166 |

Horizontal observation of the data in the columns of KCF, ECO, BIT, DSST, ARCF, and IARCF shows that the IARCF tracker has the highest *precision* and *success* rate for the five scenarios, which illustrates that IARCF successfully improves the tracking quality by introducing a human relocation strategy and adaptive model update mechanism, and its tracking performance is not affected by the characteristics of underground scenes.

Longitudinal observation of the average data show that for the five tracking scenarios in the IHV-UM dataset, the tracking performance of alley, substation, refuge chamber and electromechanical room is better than that of fully mechanized face. This is because the spatial environmental conditions of the previous four scenarios is relatively good, and the human is less disturbed by the background objects. However, what is worth mentioning more is that from the data in the improved column, it can be seen that the improvement of the *precision* and *success* rates of the fully mechanized face are significantly higher than that of other scenarios, which indicates that the IARCF tracker in this paper can not only track humans in complex underground scenes, but is also more suitable for dealing with situations where humans are heavily occluded.

## 6. Conclusions

In this paper, an improved ARCF tracker based on thermal infrared videos is proposed to track the humans in underground mines, which can be embedded in machine vision systems for human security monitoring and personnel identification based on human symmetric biometrics. First, the preprocessing method for infrared videos is used to improve the accuracy of human location. Next, by analyzing the shape of the response map and the change characteristics of the PSLR, the positioning accuracy of the human is measured, and the way of relocating humans by calculating the image similarity is specified. Then, according to the numerical relationship of the highest peak point and the PSLR between the previous frames and current frame, the adaptive update strategy of the appearance model is formulated. Finally, the experiments are performed on the IHV-UM dataset, the results show that the average *precision* and *success* rate of the IARCF tracker in the five underground scenarios reach to 0.8985 and 0.7183 respectively, which is 0.1771, 0.3415, 0.1459, 0.3606, 0.0468 higher than the average *precision* and 0.1713, 0.1041, 0.1711, 0.1566, 0.0283 higher than the average *success* rate of the KCF, DSST, ECO, BIT and ARCF trackers. This proves the effectiveness and feasibility of the IARCF applied to underground human tracking. In future, the algorithm calculation amount of the relocation link can be compressed and refined, to further improve the tracking speed without sacrificing the tracking performance.

**Author Contributions:** Conceptualization, X.L. and W.C.; methodology, X.L.; software, W.F.; validation, X.L., S.W. and Z.T.; investigation, X.L., Z.W. and S.W.; resources, S.W.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L., S.W., W.C., Z.W. and Z.T.; funding acquisition, W.C. and Z.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China (Grant no. 52074305 and 51874300), the National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon (Grant no. U1510115), and the Open Research Fund of Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences (Grant no. 20190902, 20190913).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

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

## References

1. Abdul-azeez, L.; Aibinu, A.M.; Akanmu, S.O.; Folorunso, T.A.; Salami, M.E. Intelligence security check system using face recognition algorithm: A review. In Proceedings of the 5th International Conference on Electronics, Computer and Computation (ICECCO), Abuja, Nigeria, 10–12 December 2019.
2. Borkar, K.; Salankar, S. IRIS Recognition System. In Proceedings of the International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, India, 3–4 December 2021.
3. Liu, X.Y.; Liu, J.Q. Gait recognition method of underground coal mine personnel based on densely connected convolution network and stacked convolutional autoencoder. *Entropy* **2020**, *22*, 695. [CrossRef] [PubMed]

4. Liu, S.H.; Song, Y.; Zhang, M.Y.; Zhao, J.W.; Yang, S.H.; Hou, K. An identity authentication method combining liveness detection and face recognition. *Sensors* **2019**, *19*, 4733. [CrossRef] [PubMed]
5. Thakur, N.; Han, C.Y. Indoor Localization for Personalized Ambient Assisted Living of Multiple Users in Multi-Floor Smart Environments. *Big Data Cogn. Comput.* **2021**, *5*, 42. [CrossRef]
6. Ranieri, C.M.; MacLeod, S.; Dragone, M.; Vargas, P.A.; Romero, R.A.F. Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors. *Sensors* **2021**, *21*, 768. [CrossRef]
7. Kumar, M.; Ray, S.; Yadav, D.K. Moving human detection and tracking from thermal video through intelligent surveillance system for smart applications. *Multimed. Tools Appl.* **2022**, 1–20. [CrossRef]
8. Li, Z.Z.; Xu, J.C. Target Adaptive Tracking Based on GOTURN Algorithm with Convolutional Neural Network and Data Fusion. *Comput. Intell. Neurosci.* **2021**, *2021*, 4276860. [CrossRef]
9. Zhang, Y.F.; Zhang, M.; Cui, Y.X.; Zhang, D.Y. Detection and tracking of human track and field motion targets based on deep learning. *Multimed. Tools Appl.* **2020**, *79*, 9543–9563. [CrossRef]
10. Liu, C.H.; Huynh, D.; Sun, Y.C.; Reynolds, M.; Atkinston, S. A Vision-Based Pipeline for Vehicle Counting, Speed Estimation, and Classification. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 7547–7560. [CrossRef]
11. Huo, Y.H.; Fan, W.Q. Face recognition method under complex light conditions in coal mine. *Laser Optoelectron. Prog.* **2019**, *56*, 11003. [CrossRef]
12. Chai, Y.; Gao, R.; Deng, L.J. Study of image enhancement algorithms in coal mine. In Proceedings of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 27–28 August 2016.
13. Fan, W.Q.; Huo, Y.H.; Li, X.Y. Degraded image enhancement using dual-domain-adaptive wavelet and improved fuzzy transform. *Math. Probl. Eng.* **2021**, *2021*, 5578289. [CrossRef]
14. Li, C.L.; Liu, J.H.; Zhu, J.J.; Zhang, W.Z.; Bi, L.H. Mine image enhancement using adaptive bilateral gamma adjustment and double plateaus histogram equalization. *Multimed. Tools Appl.* **2022**, *81*, 12643–12660. [CrossRef]
15. Srilekha, S.; Swamy, G.N.; Anudeep Krishna, A. A Novel Approach for Detection and Tracking of Vehicles using Kalman Filter. In Proceedings of the 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 12–14 December 2015.
16. Chu, C.T.; Hwang, J.N.; Pai, H.I.; Lan, K.M. Tracking Human Under Occlusion Based on Adaptive Multiple Kernels with Projected Gradients. *IEEE Trans. Multimed.* **2013**, *15*, 1602–1615. [CrossRef]
17. Obolensky, N.; Erdogmus, D.; Principe, J.C. A Time-varying Kalman Filter Applied to Moving Target Tracking. In Proceedings of the ONTROLO'02, Aveiro, Portugal, September 2002.
18. Abhinava, B.N.; Majumdar, J. Automatic Detection of Human in Video and Human Tracking. *Int. J. Eng. Res. Technol. (IJERT)* **2017**, *6*, 265–273. [CrossRef]
19. Dinesh, H.R.; Majumdar, J.; Kiran, S. Automatic Object Tracking with Particle Filter Coupled to Edge Detectors. *Int. J. Sci. Res. (IJSR)* **2014**, *3*, 262–267.
20. Majumdar, J.; Bhattaral, A.; Adhikari, S. Optical Flow-Initiated Particle Filter Framework for Human-Tracking and Body-Component Detection. *Adv. Sci. Lett.* **2017**, *23*, 11217–11222. [CrossRef]
21. Beaugendre, A.; Miyano, H.; Ishidera, E.; Goto, S. Human Tracking System for Automatic Video Surveillance with Particle Filters. In Proceedings of the 2010 IEEE Asia Pacific Conference on Circuit and System (APCCAS), Kuala Lumpur, Malaysia, 6–9 December 2010.
22. Kaur, R.; Singh, S. Background Modelling, Detection and Tracking of Human in Video Surveillance System. In Proceedings of the 2014 Innovative Applications of Computational Intelligence on Power Energy and Controls with their Impact on Humanity (CIPECH), Ghaziabad, India, 28–29 November 2014.
23. Huang, K.S.; Trivedi, M.M. Robust Real-Time Detection, Tracking, and Pose Estimation of Faces in Video Streams. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR), British Machine Vis Assoc, Cambridge, UK, 23–26 August 2004.
24. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010.
25. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the circulant structure of tracking-by-detection with kernels. In Proceedings of the 12th European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012.
26. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal.* **2015**, *37*, 583–596. [CrossRef]
27. Danelljan, M.; Khan, F.S.; Felsberg, M.; Felsberg, M.; Weijer, J.V.D. Adaptive color attributes for real-time visual tracking. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
28. Danelljan, M.; Häger, G.; Khan, F.; Felsberg, M. Accurate scale estimation for robust visual tracking. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014.
29. Li, F.; Tian, C.; Zuo, W.M.; Zhang, L.; Yang, M.H. Learning spatial-temporal regularized correlation filters for visual tracking. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
30. Huang, Z.Y.; Fu, C.H.; Li, Y.M.; Lin, F.L.; Lu, P. Learning aberrance repressed correlation filters for real-time UAV tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.

31. Wang, X.Y.; Fan, B.J. Learning Aberrance Repressed and Temporal Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020.
32. Nam, H.; Han, B. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016.
33. Fan, H.; Ling, H.B. SANet: Structure-Aware Network for Visual Tracking. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 June 2017.
34. Nam, H.; Baek, M.; Han, B. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking. *arXiv* **2016**, arXiv:1608.07242v1.
35. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.
36. Tao, R.; Gavves, E.; Smeulders, A.W.M. Siamese Instance Search for Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016.
37. Song, Y.B.; Ma, C.; Wu, X.H.; Gong, L.J.; Bao, L.C.; Zuo, W.M.; Shen, C.H.; Lau, R.W.H.; Yang, M.H. VITAL: VIsual Tracking via Adversarial Learning. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
38. Zhou, X.; Chen, K.X.; Zhou, Q.D. Human tracking by employing the scene information in underground coal mines. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017.
39. Wang, J.; Ye, G.Q.; Li, J.H.; Kou, Q.Q. Improved object-tracking algorithm for an underground mining environment. In Proceedings of the 2019 6th International Conference on Soft Computing and Machine Intelligence (ISCMI), Johannesburg, South Africa, 19–20 November 2019.
40. Jiang, D.H.; Dai, L.; Li, D.; Zhang, S.Y. Moving-Object tracking algorithm based on PCA-SIFT and optimization for underground coal mines. *IEEE Access* **2019**, *7*, 35556–35563.
41. Hu, M.K. Visual pattern recognition by moment invariants. *Ire Trans. Inf. Theory* **1962**, *8*, 179–187.
42. Pan, D.F.; Li, Y.T.; Han, K. A target tracking method based on multi-correlation filter combination. *J. Hunan Univ.* **2019**, *46*, 112–122.
43. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient convolution operators for tracking. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
44. Cai, B.L.; Xu, X.M.; Xing, X.F.; Jia, K.; Miao, J.; Tao, D.C. BIT: Biologically inspired tracker. *IEEE Trans. Image Process.* **2016**, *25*, 1327–1339. [CrossRef]