Multi-target Tracking Algorithm for Film and Video under Multi-Layer Convolution Feature

Lyuqi Li[1], Shaoyang Zeng[2]
[1]School of Humanities and Arts, Xi'an International University, Xi'an 710077, Shaanxi, China
[2]MSE Development Department, Packet Core Network, Huawei Technologies Co., Ltd., Xi'an 710000, Shaanxi, China
Corresponding author: Lyuqi Li (e-mail: llq_xaiu@163.com)

ABSTRACT Accurate multi-target tracking in film and video has become increasingly important for intelligent visual perception systems and electromagnetic-enabled sensing applications, where robust signal representation and feature propagation are essential for reliable scene understanding. This study investigates the accuracy and precision of multi-target character tracking in film and video by employing a multi-layer convolution feature framework integrated with convolutional neural networks (CNNs). In the proposed neural network architecture, video data are processed through multiple convolution and pooling layers followed by fully connected layers, while the Rectified Linear Unit (ReLU) is adopted as the activation function to enhance network performance. Experimental results demonstrate that the proposed method achieves a multiple object tracking accuracy (MOTA) of 56.8%, a multiple object tracking precision (MOTP) of 79.8%, and an Identification F-score (IDF1) of 60.1%. On the MOT16 benchmark, the proposed approach improves MOTA, MOTP, and IDF1 by 1.8%, 4.0%, and 2.7%, respectively, compared with the Dynamic Memory based Attention Network (DMAN), indicating superior tracking consistency and feature discrimination across different video sequences. The results confirm that the CNN-based multi-target tracking framework significantly enhances MOTA and IDF1 while providing a reliable computational foundation for intelligent visual sensing and signal processing systems in advanced communication and electromagnetic perception environments.

INDEX TERMS multi-target tracking, convolutional neural network, signal processing, electromagnetic sensing, visual perception

I. INTRODUCTION

WITH the continuous development of modern science and technology, movies and videos based on Internet technology constantly impact human vision, much like the sophisticated visual surface inspection and material characterization technologies evolving within engineering. This digital progression facilitates a deeper integration of high-resolution imaging and automated analysis, which are now foundational to the advanced manufacturing standards of the industry [1-3]. After work, most people spend their free time watching videos. Nowadays, movies, videos and sports events are all spread through the Internet. With the increasing requirements for experience of watching movies and videos, researchers related to film and television production have integrated various high technologies to realize target tracking [4-6]. This technology enables the precise localization and persistent identification of key characters within complex cinematic sequences. By providing robust temporal-spatial consistency, target tracking serves as a foundational component for advanced post-production tasks, such as automated character-centric editing and dynamic visual effects, thereby enhancing the overall technical quality of film production. To evaluate the proposed multi-target tracking framework, widely recognized public benchmarks are utilized to simulate diverse real-world shooting scenarios captured by cinematic cameras [7-9]. According to the given situation, research demonstrates that multi-target tracking constitutes a fundamental technical pillar in the evolution of digital film and video processing. In the actual complicated and changeable video shooting environment, there are uncontrollable changes in the shape, movement and scale of the shooting target, which has high requirements for multi-target tracking technology.

In order to improve the accuracy of multi-target tracking algorithm, researchers all over the world have done a lot of research, and have obtained many effective methods. Multi-target tracking technology includes many categories in advanced science and technology. Such as image processing, optimization methods and signal processing. Li et al. (2021) [10] pointed out in medical diagnosis and multi-modal medical image research that deep learning algorithm can pre-process the data of pattern recognition

and image processing, automatically learn and constantly optimize itself. Besides, this algorithm can also deal with image fusion problems well, and has the advantage of good stability. Sun et al. (2018) [11] proposed in the "Multi-Target Pig Tracking Loss Correction Algorithm Based on Faster Region-Convolution Neural Network (Fast R-CNN)" that combining Background Difference Method with Speed Up Robust Feature (SURF) can produce better robustness and adaptability, and this method can well solve the problem of multi-target tracking loss. Concurrently, multi-targets can still be tracked accurately of the occluded target. He et al. (2018) [12] introduced a new joint multi-target tracking and tracking maintenance algorithm based on sensor networks, and the research results confirmed the authenticity and effectiveness of the proposed multi-target tracking algorithm based on multi-sensors [13-15]. According to the above-mentioned multi-target tracking algorithms and the related research status of deep learning model all over the world, further research is made on the existing problems and research methods that is not involved. In addition, no relevant researchers have combined convolution features videos with multi-target tracking algorithm to research on film and video. On this basis, multi-target tracking algorithm and convolution features are used to study on the video. The innovation lies in the integration of a specialized spatial attention mechanism with multi-layer convolution features to dynamically weight character appearance in occluded video scenes, a refinement that enhances identity preservation compared to standard CNN-based tracking frameworks. The research topic lays a theoretical foundation for the follow-up related research work, and concurrently, it has a referencing significance for the relevant personnel engaged in film and video research.

II. METHOD

A. CONVOLUTIONAL FEATURES

The CNN algorithm in deep learning is employed to extract features from standard multi-object tracking datasets (e.g., MOT15, MOT16, and MOT17), which serve as the primary data source for validating the model's performance. CNN [16-18] algorithm can minimize preprocessing, and besides, it can directly extract the most expressive features from the original data input by specifying features manually. In the CNN model, video data goes through multiple convolutional layers and pooling layers. Then it is connected with each other by multiple fully connected layers. In order to improve the network performance, the activation function used by all neurons is usually ReLU function. Meanwhile, in order to balance computational efficiency and real-time tracking requirements, AlexNet is selected as the lightweight backbone for feature extraction. Although modern architectures like ResNet-152 provide higher peak accuracy, the streamlined structure of AlexNet significantly reduces the latency in processing high-frame-rate film data, which is crucial for maintaining consistent tracklets. Therefore, the functional layer of convolutional layer of AlexNet model is further improved. Primarily, the local part is normalized, then the priority of the pooling operation is put to the next step. Ultimately, the local part is normalized again. Two advantages of this improved method are: first, it can further enhance the generalization ability of AlexNet network, simultaneously, it can also weaken the phenomenon of over-fitting, which greatly reduces the training time; Second, overlapping pooling operation before local normalization can not only keep more data information and weaken redundant information in the pooling process, but also accelerate the convergence rate in the training process of video prediction model. It can also highlight the advantages of overlapping maximum pooling over previous maximum pooling methods. Figure 1 concludes the framework of CNN algorithm.

In the algorithm flow of AlexNet [19-21], initialize $\lambda_3 = \lambda_4$ and s as shown in Equations (1), (2) and (3).

$$\lambda_3 = \lambda_4 = \frac{1}{n}\sum_{i=2}^{n}\left(kd_{i,k+1}^{x} - \frac{1}{2}\sum_{j=1}^{k} d_{ij}^{x}\right) \tag{1}$$

$$s_{ij} = \begin{cases} 0, & x^j \notin N_k(x^i), \\ \dfrac{d_{i,k+1}^{x} - d_{ij}^{x}}{kd_{i,k+1}^{x} - \sum_{j=1}^{k} d_{ij}^{x}}, & x^j \in N_k(x^i). \end{cases} \tag{2}$$

$$d_{ij}^{x} = \|x^i - x^j\|_2^2 \tag{3}$$

Equation (3) is the supplement of Equations (1) and (2). Index d_{ij}^{x} is sorted in ascending order to find out the smallest k d_{ij}^{x} among them. Besides, $N_k(x^i)$ stands for the k samples nearest to x^i. Update F^t as displayed in Equation (4).

$$\min_{F} \operatorname{tr}(F^T L_s F) \quad \text{s.t.} \quad F \in R^{n\times c},\ F^T F = I \tag{4}$$

According to Equations (1-4), the eigenvalues corresponding to the smallest c eigenvectors of Laplace matrix L_s can be used to form F. Parameter $s(t)$ can be obtained by merely solving Equation (5).

$$\begin{aligned} \min_{s_{i1}=1,\ s_i\geq 0} \sum_{i=1}^{n} \left(\lambda_2 \|w^T x_i - w^T x_j\|_2^2 s_{ij} + 2\lambda_3 s_{ij}^2\right) \\ + \sum_{i=1}^{n} \lambda_4 \|f_i - f_j\|_2^2 s_{ij} \end{aligned} \tag{5}$$

The prevalence of deep learning models causes that different CNN models gradually appear in the public's field of vision. CNN model has become the basic model to solve various problems, on which background, the CNN model is adopted to deconvolute the film and video images and visualize the characters in the videos. Then, Figure 2 presents what is obtained and signifies the visualization process of video image features.

Further, the features of video and TV are extracted, and the tth feature map of the lth convolution layer is analysed in the way of overlapping pooling to sample $y_t^l(i,j)$.

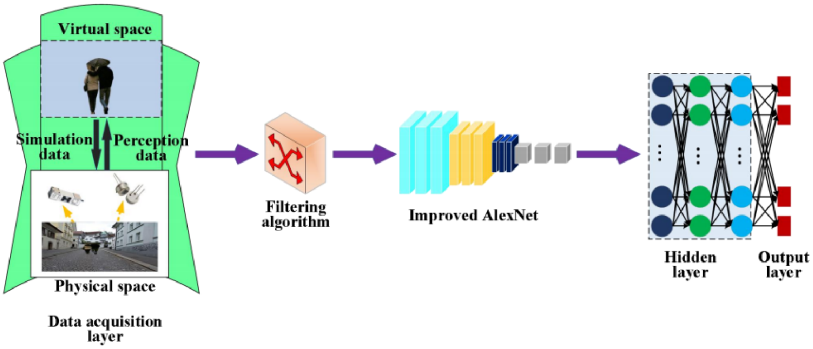

FIGURE 1. Framework of CNN algorithm.

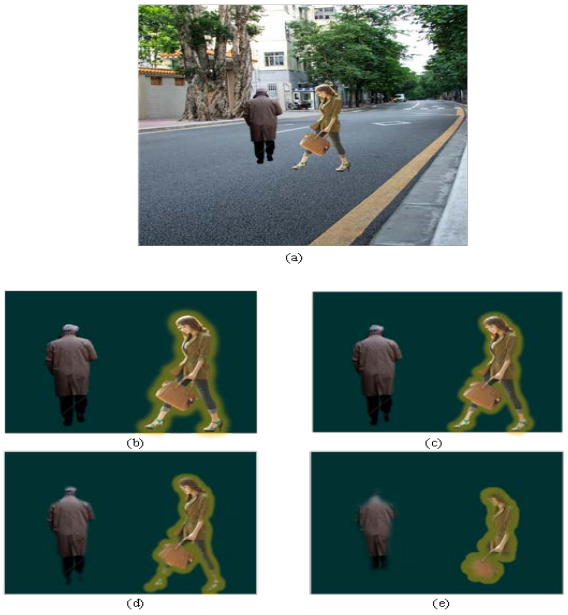

FIGURE 2. Visualization of video image features: (a) the original image in the video; (b) the image obtained by the first part of convolution; (c) the image obtained by convolution in the second part; (d) the image obtained by convolution in the third part; (e) images obtained by convolution of the fourth and fifth parts

$$a_t^l(i,j) = \max\{y_t^l(i,j),\ is \le i \le is + w_c - 1,\ js \le j \le js + w_c - 1\} \tag{6}$$

In Equation (6), s means the pool moving step. w_c equals to the width of pool area, meanwhile. $w_c > s$. After the first and second pooling layers of AlexNet model, local normalization layers are added to standardize the feature map $c_t^l(i,j)$:

$$c_t^l(i,j) = a_t^l(i,j) / \left(k + \alpha \sum_{\max(0,t-m/2)}^{\min(N-1,t+m/2)} (a_t^l(i,j))^2 \right)^{\beta} \tag{7}$$

In Equation (7), k, α, β, m refer to super-parameters, which are valued in sequence as 3, 0.76, 10^{-5} and 8, respectively. Parameter n means the total number of convolution kernels in the lth convolution layer. According to the target detection model, the specific position of the video target is determined, and then the characteristics of the target characters are further obtained through the detection results of the target objects, then the targets are matched, and finally the corresponding association is made clear. Figure 3 denotes the feature extraction process of video images.

In view of the gradient dispersion problem in the network model, ReLU function is taken as the activation function to activate the output convolution $S_t^l(i,j)$. Equation (8) expresses the process.

$$y_t^l(i,j) = f(S_t^l(i,j)) = \max\{0, S_t^l(i,j)\} \tag{8}$$

In Equation (8), $S_t^l(i,j)$ represents the activation function of ReLU. If the weight $\theta = [K_t^l, W_i^l, b_t^l, b_i^l]$, the gradient g is expressed as Equation (9).

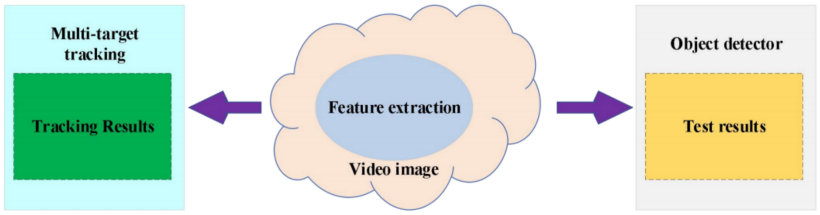

FIGURE 3. Feature extraction process of video images

$$g = \left[\frac{\partial L}{\partial K_t^l}, \frac{\partial L}{\partial W_i^l}, \frac{\partial L}{\partial b_t^l}, \frac{\partial L}{\partial b_i^l}\right] \tag{9}$$

Then, Equations (10) and (11) illustrate the bi-first-order moment estimation s and second-order moment estimation r of the updated gradient, respectively.

$$s = \rho_1 s + (1-\rho_1)g \tag{10}$$

$$r = \rho_2 r + (1-\rho_2)g^2 \tag{11}$$

In Equations (10) and (11), ρ_1, ρ_2 define the exponential decay rate for the moment estimation. Deviation $\hat{s}$ of the corrected first moment s and deviation $\hat{r}$ of the second moment r are calculated as shown in Equations (12) and (13), respectively.

$$\hat{s} = \frac{s}{1-\rho_1^t} \tag{12}$$

$$\hat{r} = \frac{r}{1-\rho_2^t} \tag{13}$$

Finally, Equation (14) illustrates the updated weights θ.

$$\Delta\theta = -\epsilon\frac{\hat{s}}{\sqrt{\hat{r}}+\delta}, \quad \theta = \theta + \Delta\theta \tag{14}$$

In Equation (14), ϵ means learning step, and δ represents a small constant for numerical stability. In order to improve the accuracy of data association, the characteristic of high response of CNN is fully utilized, and then reinforce the network with spatial attention mechanism. Figure 4 presents a video image spatial attention module.

Generally, in order to prevent over-fitting, the parameter of dropout operation is usually set as 0.4 in the fully connected layer. When l equals 5 in Equation (4), all feature images are reconstructed into a high-dimensional single-layer neuron structure C_5^6, then the input Z_i^6 of the ith neuron in the 6th fully connected layer can be obtained through Equation (15).

$$Z_i^6 = W_i^6 C_5 + b_i^6 \tag{15}$$

In Equation (15), W_i^6 and b_i^6 refer to the weight and offset of the ith neuron in the 6th layer of the fully connected layer, respectively. In the process of improving generalization ability, a dropout strategy is implemented on the neurons in the 6th and 7th fully connected layers. Equation (16) expresses the principle of the process.

$$r_j^l \sim \text{bernoulli}(dp), \quad \widetilde{C}^l = r^l C^l \tag{16}$$

Then, the input of the ith neuron Z_i^{l+1} in the 7th and 8th fully connected layers can be represented by $W_i^{l+1}\widetilde{C}^l + b_i^{l+1}$. The output C_i^l of the i-th neuron in the 6th and 7th fully connected layers is defined as $f(Z_i^l)$, namely $\max\{0, Z_i^l\}$. Finally, the input of neurons q_i in the 8th fully connected layer can be solved out by Equation (17).

$$q_i = \text{softmax}(Z_i^8) = \frac{e^{Z_i^8}}{\sum_{j=1}^{12} e^{Z_j^8}} \tag{17}$$

Additionally, the cross entropy loss function is suitable for the classification task as the error function of the model. Equations (18) and (19) signify how it works.

$$Loss = \sum_{i=1}^{K} y_i \cdot \log(p_i) \tag{18}$$

$$p_i = \frac{\exp(\widetilde{y}_i)}{\sum_{j=1}^{K}\exp(\widetilde{y}_j)} \tag{19}$$

In Equations (18) and (19), parameter K stands for the number of types, parameter y_i means the true category distribution of samples, $\widetilde{y}_i$ refers to the output result of neural network, and p_i equals the classification result after passing through Softmax classifier [22-24]. The input of the Softmax function is an n-dimensional real number vector, which is set as x. Equation (20) presents the specific expression.

$$\xi(x)_i = \frac{e^{x_i}}{\sum_{n=1}^{N} e^{x_n}}, \quad i = 1, 2, \ldots, N \tag{20}$$

Fundamentally speaking, Softmax function can map any n-dimensional real number vector into an n-dimensional vector whose values of all elements are in the range of (0,1), thus realizing the normalization of the vector. In order to reduce the computational complexity of the model system, compandor is generally adopted to transform the data so as to improve the prediction efficiency of the model.

$$f(x_t) = \frac{\text{sign}(x_t)\ln(1+\mu|x_t|)}{\ln(1+\mu)}, \quad |x_t| < 1 \tag{21}$$

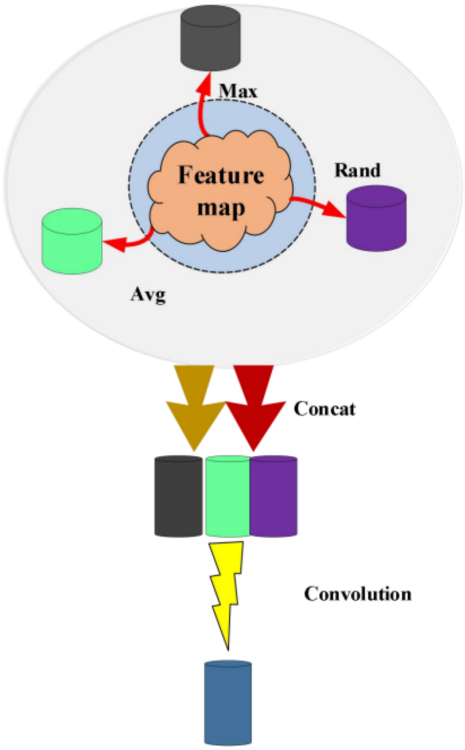

FIGURE 4. Spatial attention module of video image.

B. MULTI-TARGET TRACKING ALGORITHM

In real life, people often encounter all kinds of complicated scenes, such as the shooting and making process of film and videos. Because of the actors' movements being blocked by various obstacles, the multi-target tracking [25-27] is greatly interfered. Time complexity and tracking accuracy are very important for multi-target tracking technology. Figure 5 displays a scene of moving people tracking.

Another method is to use Kalman filter to model various trajectories of target objects. Firstly, prediction is made on the possible position of the target object in the next second. Secondly, a coordinate origin is established for the predicted position of the target object. Thirdly, the filter function is used for preliminary processing. Equations (22) and (23) are the calculation of radius. Equation (24) is used to get the prediction probability of the target trajectory.

$$r_1 = \frac{h}{2} \tag{22}$$

$$r_2 = \frac{w}{2} \tag{23}$$

$$p = \frac{1}{2\pi r_1 r_2} \exp\left(-\left(\frac{x^2}{2r_2^2} + \frac{y^2}{2r_1^2}\right)\right) \tag{24}$$

In Equations (22-24). r_1 and r_2 refer to the radius, h and w represent the heights and width of the target. The target characters are divided into multiple grids, track and locate them concurrently to predict the probability of their categories, and calculate the offset. Figure 6 presents the result of target tracking output.

In order to improve the accuracy and efficiency of tracking, usually, some negative samples need to be removed when dealing with distant objects.

$$p_{ij} = \alpha_{t_1}^i * (X_{t_1}^i \cdot X_{t_2}^j) \tag{25}$$

Generally, the target is described according to the motion characteristics and appearance characteristics. In Equation (25), $X_{t_1}^i$ means the appearance description vector of the ith object at t_1; $X_{t_2}^j$ stands for the appearance description vector of the jth object at t_2; $\alpha_{t_1}^i$ refers to the confidence score obtained by Gaussian model. p_{ij} denotes the similarity between any two objects. Equation (26) illustrates the position of the same target at different times.

$$T_i = (O_{t-K}^i, \cdots, O_t^i) \tag{26}$$

In Equation (26), T_i stands for the current track, and $i = 1, 2, \ldots, m$. Equation (27) reveals the similarity between the moving track of the target and the target by:

$$S_{ij} = \alpha_t^i * \sum (X_t^j \cdot X_v^i) \tag{27}$$

In Equation (27), X_v^i means the vector expression of targets' feature of target track T_i, meanwhile, $X_v^i \in T_i$. If the similarity matrix S is set as a negative number, the cost matrix $-S$ can be obtained, and then correlate the

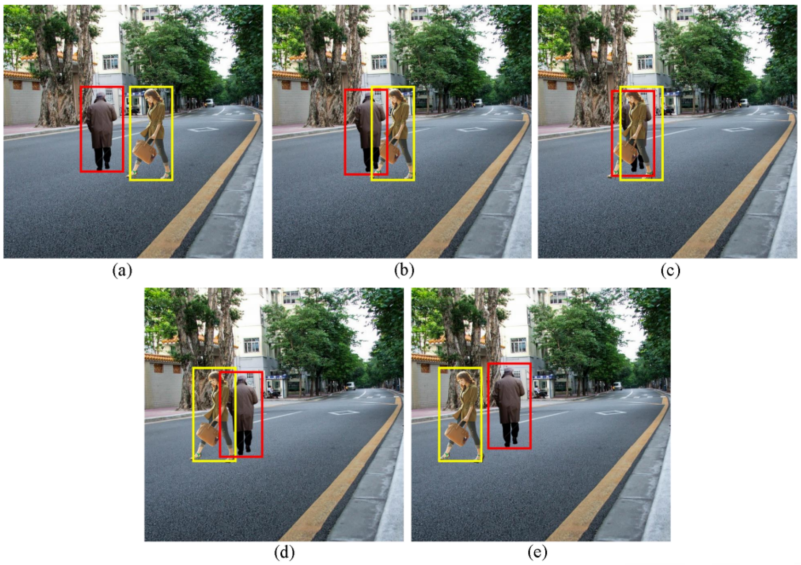

FIGURE 5. Tracking movement of characters: (a) the movement at the 5 second in the video; (b) the motion at the 10 second in the video; (c) the motion at the 15 second in the video; (d) the motion at the 20 second in the video; (e) the motion at the 25 second in the video

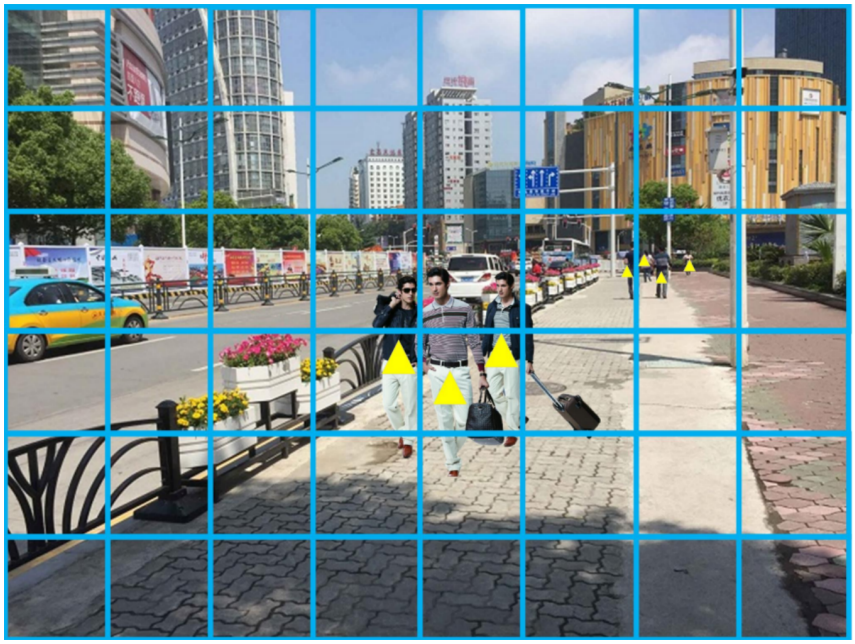

FIGURE 6. Result of target tracking output

current video data with the tracking track data. Equation (28) denotes the expression.

$$\text{assignment} = \text{Hungarian}(-S) \tag{28}$$

The loss functions of the multi-objective model studied are cross entropy loss function and multi-objective tracking loss function. Equation (29) is a cross entropy loss function.

$$L_C = -\sum p_i * \log(q_i) \tag{29}$$

In Equation (29), p_i can be calculated as shown in Equation (30), which represents the meaning of predicting different types of probabilities. Parameter q_i refers to the true type of the target object.

$$p_i = \frac{\exp(x_i)}{\sum \exp(x_j)} \tag{30}$$

In Equation (31), smooth L1 denotes the loss function of the specific position regression of the target object.

$$L_R = \sum \sum \text{smoothL1}(l_i - \hat{g}_j) \tag{31}$$

In Equation (31), l_i means the specific position of the target output by the model, $\hat{g}_j$ represents the true position of the target object. By changing it nonlinearly, the expression

of the specific position on the characteristic map after the change can be obtained as shown in Equation (32).

$$\begin{cases} \hat{g}_{x_i} = \dfrac{g_{x_i} - a_{x_i}}{a_{w_i}}, \\ \hat{g}_{y_i} = \dfrac{g_{y_i} - a_{y_i}}{a_{h_i}}, \\ \hat{g}_{w_i} = \log\left(\dfrac{g_{w_i}}{a_{w_i}}\right), \\ \hat{g}_{h_i} = \log\left(\dfrac{g_{h_i}}{a_{h_i}}\right). \end{cases} \tag{32}$$

In Equation (32), $(a_{x_i}, a_{y_i}, a_{w_i}, a_{h_i})$ denotes the position and size of the Anchor, $(g_{x_i}, g_{y_i}, g_{w_i}, g_{h_i})$ signifies the specific position of the target mark. Equation (33) is the loss function of target tracking. It is specifically expressed as a target matrix.

$$G_{ij} = \begin{cases} 1, & \text{if } id(O_i) = id(O_j), \\ 0, & \text{if } id(O_i) \neq id(O_j). \end{cases} \tag{33}$$

In Equation (33), O_i indicates the object in the film and video, O_j stands for the target in different video frames. For matching two images of different video frames, Equation (34) can describe the process.

$$\sum_i G_{ij} \in \{0, 1\} \quad \sum_j G_{ij} \in \{0, 1\} \tag{34}$$

For the loss of target tracking, it is divided into forward matching and backward matching. Equation (35) displays its detailed expression.

$$\begin{aligned} Loss = & \sum_{i \in \Omega_1^m} \sum_j h(S_{ij}^1) \\ & + \sum_{i \in \Omega_1^{um}} \sum_j (S_{ij}^1 > \sigma) h(1 - S_{ij}^1) \\ & + \sum_i \sum_{j \in \Omega_2^m} h(S_{ij}^2) \\ & + \sum_i \sum_{j \in \Omega_2^{um}} (S_{ij}^2 > \sigma) h(1 - S_{ij}^2) \end{aligned} \tag{35}$$

In Equation (35), Ω_1^m means the matching relationship of different objects in video, σ equals threshold value of the confidence of prediction.

$$h(x) = -\alpha(1 - x)^{\gamma} \log(x) \tag{36}$$

In Equation (36) h(x) denotes the probability of constraint prediction. In Figure 7, the red cross represents the specific position predicted of the target person, and the probability of the target's appearance can be obtained through calculation. On this basis, further judgement can be made on the similarity between the tracked target and the detected target on the moving track. Figure 7 shows the probability of the target position.

Equation (27) displays multiple object tracking accuracy (MOTA) [28-30].

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{37}$$

In Equation (37), m_t means the number of lost objects in the video, fp_t stands for the number of false samples, mme_t indicates the number of objects that are incorrectly matched. Equation (38) illustrates multiple object tracking precision (MOTP).

$$MOTP = \frac{\sum_i^t d_i^t}{(\sum_t c_t)} \tag{38}$$

d_i^t denotes the prediction error of the target in the film and video and c_t refers to the number of objects in the video. According to the multi-target tracking algorithm, any target in video can be effectively tracked, no matter whether these targets are still or in motion. MOT15, MOT16 and MOT17 are different data sets. Figure 8 shows the tracking of multi-targets in the video.

According to the description vector of the target object, it is convoluted to obtain the eigenvalue. Equations (39) and (40) express the calculation of output results.

$$M_C(\Phi) = \sigma(F_C) \tag{39}$$

$$F_C = f_{1\times1}(F_{\max}^C + F_{avg}^C) \tag{40}$$

In Equation (39), σ is sigmoid function. In Equation (40), $f_{1\times1}$ stands for point convolution. On this basis, after the feature map of the target object is weighted and combined, it can be solved out by Equation (41).

$$\Phi_C = M_C(\Phi) \otimes \Phi \tag{41}$$

In Equation (41), $\otimes$ stands for multiplication of numerical values at one-to-one positions. After obtaining the specific position of the target object, the sigmoid function is used to normalize the output data, and it is expressed as Equations (42) and (43).

$$M_S(\Phi_C) = \sigma(F_S) \tag{42}$$

$$F_S = f_{7\times7}(F_{\max}^S + F_{avg}^S) \tag{43}$$

Φ_C is the input characteristic. In Equation (43). $f_{7\times7}$ represents a convolution function. In Equation (42), F_{avg}^S and $F_{\max}^S$ are the internal conditions of the feature.

$$\Phi_S = M_S(\Phi_C) \otimes \Phi_C \tag{44}$$

The Equation (44) is the expression of the feature map after the feature points in the three-dimensional space are strengthened. In order to reduce the occlusion of the target object, it is weighted and redistributed and specifically expressed in Equation (45).

$$S_{ij} = \sum_t \alpha^t sim(X_i, X_j^t) \tag{45}$$

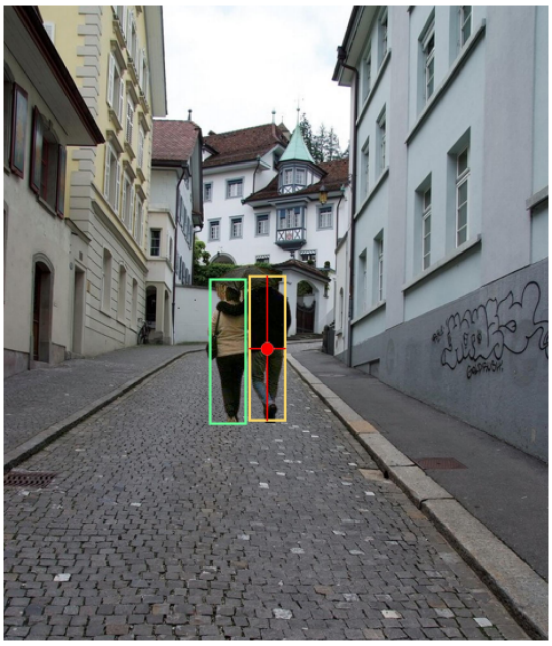

FIGURE 7. Probability of target position.

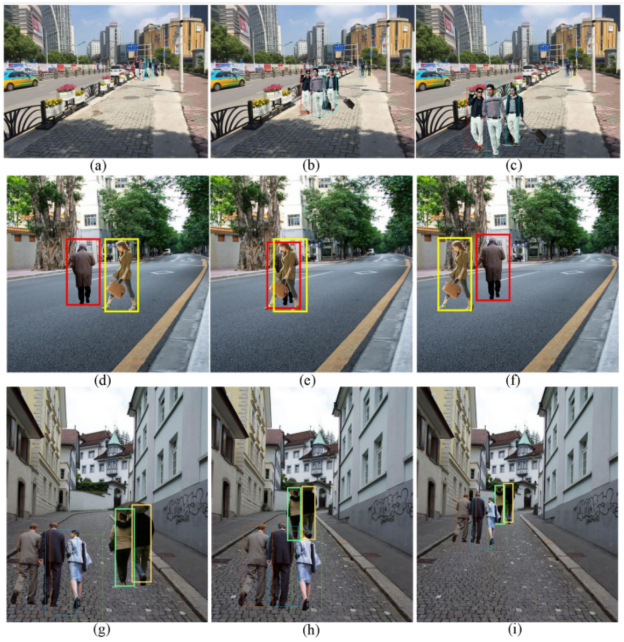

FIGURE 8. Video multi-target tracking situation: (a, b and c display the motion situations when testing the video at the 5th, 10th and 15th second on MOT15, respectively; d, e, and f show the movements of the test video at the 5th, 10th, and 15th seconds respectively on MOT16; g, h, and i present the movements of the test video at the 5th, 10th, and 15th seconds on MOT17, separately).

In Equation (45), X_i denotes the feature vector of the target object, $sim()$ means the similarity between different objects, α^t stands for the weight value of the target object.

III. RESULTS AND DISCUSSION

A. VIDEO IMAGE ANALYSIS FOR CNN

Figure 9 indicates that the accuracy of ImageNet Top1 and ImageNet Top2 is 76.9% and 92.9% respectively for ResNet 101. For Darknet 53, the accuracy of ImageNet Top1 is 77.3%, and that of ImageNet Top2 is 93.5%. With respect to ResNet152, the accuracy of ImageNet top1 is 77.4%, and that of ImageNet top2 is 93.7%. Figure 9 presents the comparison of network classification accuracy.

Improvement is made on the value of MOTA and the value of IDF1 of calculated detection number, effective identification detection number and real number in different videos. Figures 10 and 11 display the variation range. Figure 10 illustrates that for MOT17-1, Baseline+C+A (C stands for candidate target; A stands for association algorithm) has a value of 45 and Baseline has a value of 43; With regard

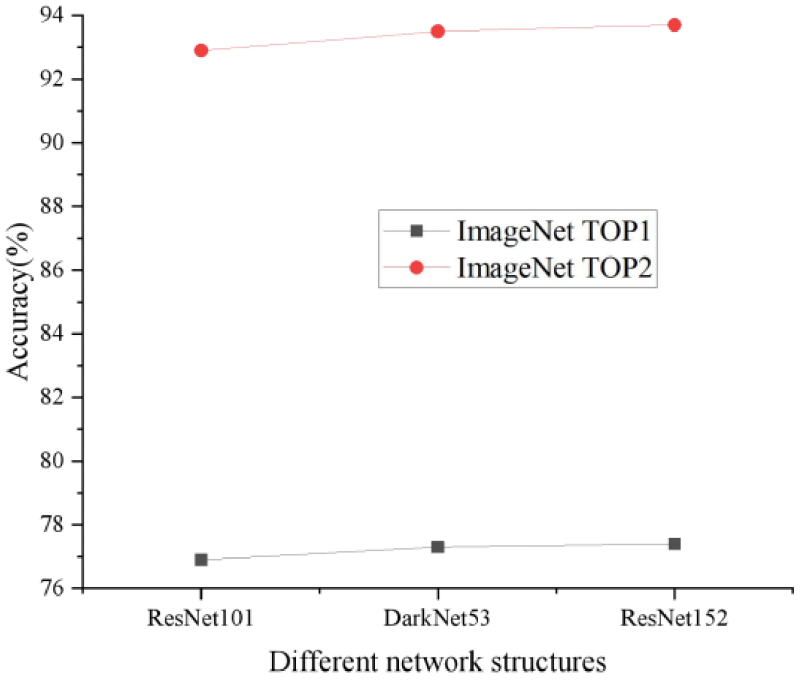

FIGURE 9. Comparison of network classification accuracy

to MOT17-3, the value of Baseline+C+A is 73 and the Baseline value is 68; With regard to MOT17-6, the value of Baseline+C+A is 57, and the baseline value is 58. Figure 10 shows the comparison of MOTA values of video sequences before and after improvement.

Figure 11 indicates that, as for MOT17-1, the Baseline+C+A value is 46, and the Baseline value is 39; With regard to MOT17-3, the value of Baseline+C+A is 69, and the baseline value is 62; With regard to MOT17-6, the value of Baseline+C+A is 60, and the baseline value is 59. Figure 11 illustrates the comparison of IDF1 values of video sequences before and after improvement.

Data in Figures 10 and 11 suggest that the algorithm after improvement greatly improves the MOTA value and IDF1 value of different sequences of video.

B. VIDEO IMAGE RESULT ANALYSIS BASED ON MULTI-TARGET TRACKING

This section is classified according to the collection process of video image sample data, and is specifically divided into three parts, the first part is the tracking result in data set MOT15, the second part is the tracking result in data set MOT16, and the third part is the tracking result in data set MOT17. Figure 12 (a) is the data set MOT15, the MOTA, multiple object tracking precision (MOTP) and IDF1 obtained by traditional method KCF (Kernel Correlation Filter) are 39.1%,71.1% and 46.1%, respectively. MOTA, MOTP and IDF1 obtained by CDA (Certified Data Analyst) are 50.9%,73.9% and 53.9% respectively. The MOTA, MOTP and IDF1 obtained by this method are 48.9%,75.8% and 56.2%, respectively. And Figure 12 (b) stands for the data set MOT16, the MOTA, MOTP and IDF1 obtained by the traditional method single object tracker with spatial-temporal attention mechanism (STAM) are 46.5%, 74.6% and 49.9%, respectively. The MOTA, MOTP and IDF1 obtained by using DMAN are 47.1%, 74.3% and 54.6%, respectively. The MOTA, MOTP and IDF1 obtained by this method are 58.9%, 78.3% and 57.3%, respectively. Figure 12 (c) refers the data set MOT17, in which MOTA, MOTP and IDF1 obtained by traditional method DeepMOT are 54.5% and 77.6% and 54.9%; MOTA, MOTP and IDF1 obtained by SST are 52.6%, 78.3% and 49.6% respectively. The MOTA, MOTP and IDF1 obtained by this method are 56.8%, 79.8% and 60.1%, respectively. Figure 12 displays the results of multi-target tracking in video.

According to the above results, comparison is made between the proposed method with the traditional method. Results indicate that in MOT15, the MOTP and IDF1 of the proposed method are improved by 1.9% and 2.3% respectively compared with CDA. In MOT16, the MOTA, MOTP, and IDF1 of this method are improved by 1.8%, 4% and 2.7%, respectively compared with those of DMAN. In MOT17, the MOTA, MOTP and IDF1 of this method are improved by 2.3%, 2.2% and 5.2% respectively compared with DeepMOT.

IV. CONCLUSION

The rapid development of various advanced technologies has brought innovations to film production and video image processing, mirroring the digital transformation of automation and structural monitoring within engineering. These cross-disciplinary advancements facilitate more precise vi-

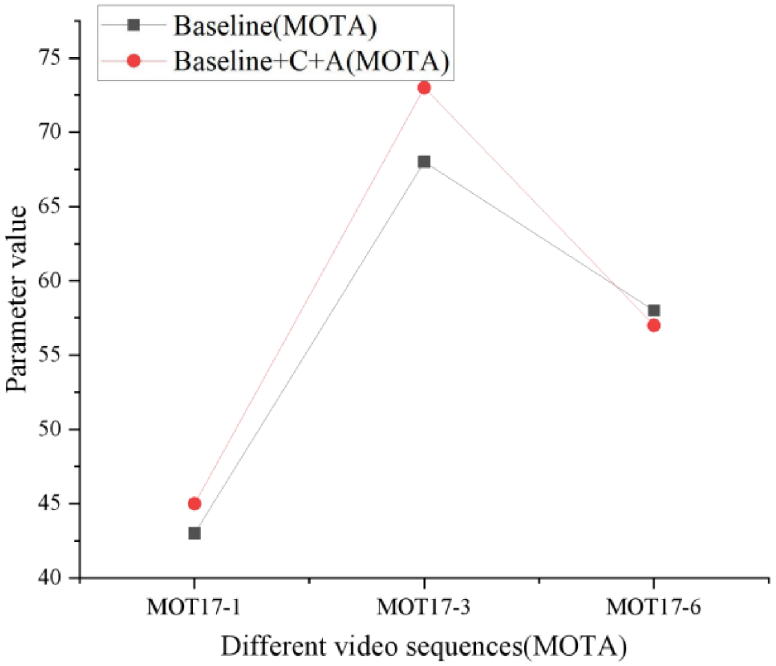

FIGURE 10. Comparison of MOTA values of video sequences before and after improvement.

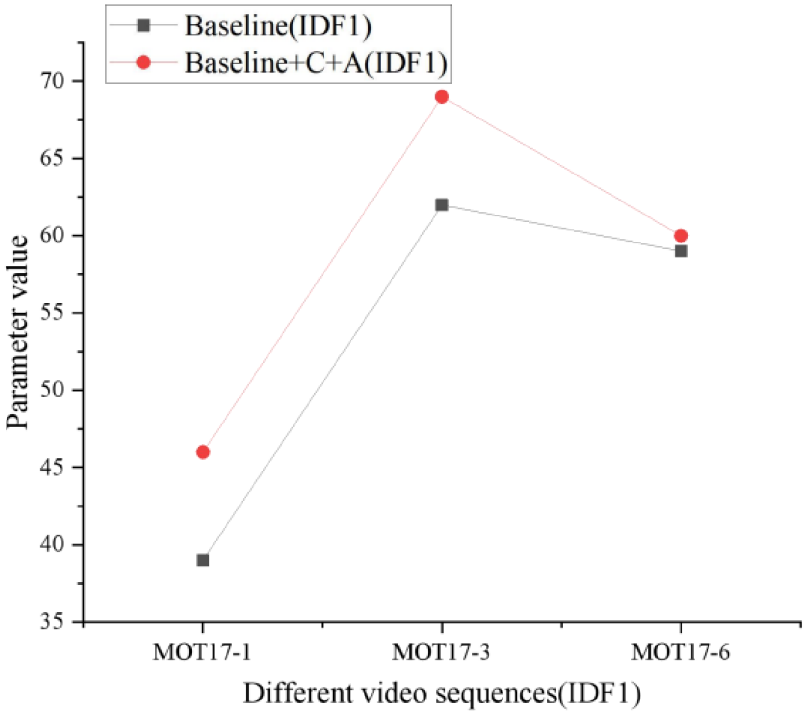

FIGURE 11. comparison of IDF1 value of video sequence before and after improvement.

sual analytics and material simulation techniques that are becoming integral to the modern industry. With the wide popularization of short videos and movie videos, discussion is made on how to improve the multi-target tracking in videos. On this basis, the CNN and MTTA are used to deeply study the problem of people tracking in movies and videos. The result shows that the MOTA, MOTP and IDF1 obtained by this method are 56.8%, 79.8% and 60.1%, respectively. In MOT16, the MOTA, MOTP and IDF1 of this method are improved by 1.8%, 4% and 2.7% respectively

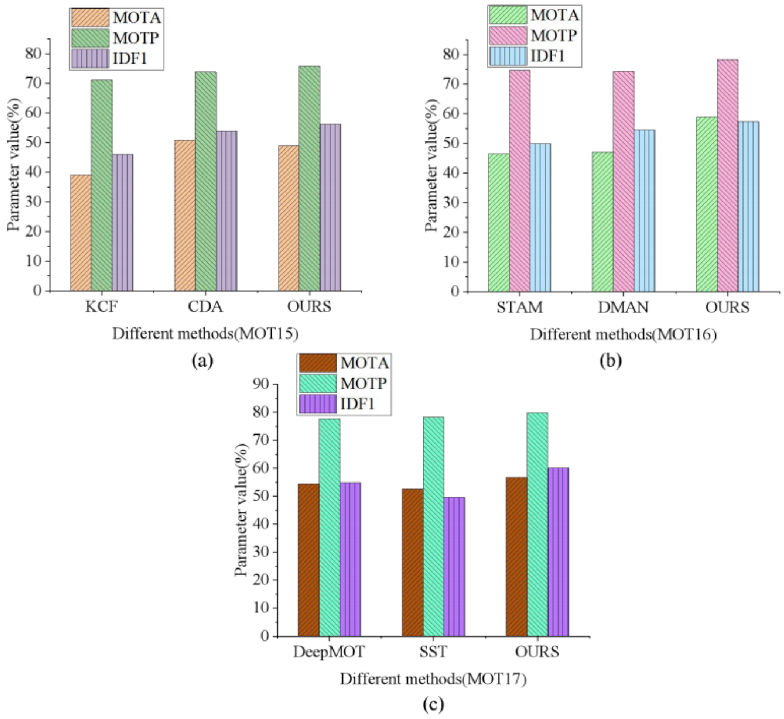

FIGURE 12. multi-target tracking results in video: (a) tracking results in data set MOT15; (b) the tracking result in the data set MOT16; (c) the tracking result in the data set MOT17

compared with those of DMAN. Therefore, the algorithm proposed here improves the MOTA value and IDF1 value of different video sequences. At last, the following conclusions are drawn: based on the CNN, the multi-target tracking algorithm has greatly improved the MOTA value, MOTP value, and IDF1 value of different sequences of videos. The shortcomings lie in: the proposed algorithm occasionally suffers from ID switching during prolonged periods of heavy occlusion and extreme lighting variations in cinematic scenes. Future research will focus on incorporating temporal consistency constraints and motion re-identification (Re-ID) modules to further refine the tracking precision under these complex conditions. In the next stage, further improvement will be made on the accuracy of tracking multi-targets in movies and videos based on the research results.

AUTHOR CONTRIBUTIONS

Conceptualization –Lyuqi Li and Shaoyang Zeng; methodology – Lyuqi Li and Shaoyang Zeng; investigation – Lyuqi Li and Shaoyang Zeng; writing-original draft preparation – Lyuqi Li and Shaoyang Zeng. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by,

1. The 2025 annual project of the "14th Five-Year Plan" Educational Science Planning of Shaanxi Province, titled "Research on the Application of AIGC Imaging Technology in Film and Television Creation Course Teaching for Journalism and Communication Major" (Project Approval Number: SGH25Q634)

2. The 2025 annual educational and teaching reform research project of Xi'an International University, "Research on the Teaching Reform of Film and Television Special Effects Courses under the 'Human- Machine Collaboration' Creation Paradigm Based on AIGC" (Project Approval Number: 2025Z07).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] H. Wang, L. Ding, R. Guan, et al., "Effects of advancing internet technology on Chinese employment: a spatial study of inter-industry spillovers," Technological Forecasting and Social Change, vol. 161, 120259, 2020, doi: 10.1016/j.techfore.2020.120259.

[2] B. Johnston A N, J. Barton M, A. Williams-Pritchard G, et al., "Youtube for millennial nursing students; using internet technology to support student engagement with bioscience," Nurse education in practice, vol. 31, pp. 151-155, 2018, doi: 10.1016/j.nepr.2018.06.002.

[3] J. Chiu C, C. Tasi W, L. Yang W, et al., "How to help older adults learn new technology? Results from a multiple case research interviewing the internet technology instructors at the senior learning center," Computers & Education, vol. 129, pp. 61-70, 2019, doi: 10.1016/j.compedu.2018.10.020.

[4] J. Yan, W. Pu, S. Zhou, et al., "Collaborative detection and power allocation framework for target tracking in multiple radar system," Information Fusion, vol. 55, pp. 173-183, 2020, doi: 10.1016/j.inffus.2019.08.010.

[5] A. Liu and S. Zhao, "High-performance target tracking scheme with low prediction precision requirement in WSNs," International Journal of Ad Hoc and Ubiquitous Computing, vol. 29, no. 4, pp. 270-289, 2018, doi: 10.1504/IJAHUC.2018.096081.

[6] A. Altan and R. Hacıoğlu, "Model predictive control of three-axis gimbal system mounted on UAV for real-time target tracking under external disturbances," Mechanical Systems and Signal Processing, vol. 138, 106548, 2020, doi: 10.1016/j.ymssp.2019.106548.
[7] L. Sun and Y. Li, "Multi-target pig tracking algorithm based on joint probability data association and particle filter," International Journal of Agricultural and Biological Engineering, vol. 14, no. 4, pp. 199-207, 2021, doi: 10.25165/IJABE.V14I4.6105.
[8] M. Ullah, H. Ullah, and A. Cheikh F, "Single shot appearance model (ssam) for multi-target tracking," Electronic Imaging, vol. 2019, no. 7, pp. 466-1-466-6, 2019, doi: 10.2352/ISSN.2470-1173.2019.7.IRIACV-466.
[9] H. Chen, Z. He, and B. Liu, "Sensor control method based on information entropy measure for multi-target tracking," Control and Decision, vol. 33, no. 2, pp. 337-344, 2018, doi: 10.13195/j.kzyjc.2016.1424.
[10] Y. Li, J. Zhao, Z. Lv, et al., "Medical image fusion method by deep learning," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 21-29, 2021, doi: 10.1016/j.ijcce.2020.12.004.
[11] L. Sun, Y. Zou, Y. Li, et al., "Multi target pigs tracking loss correction algorithm based on Faster R-CNN," International Journal of Agricultural and Biological Engineering, vol. 11, no. 5, pp. 192-197, 2018, doi: 10.25165/j.ijabe.20181105.4232.
[12] S. He, S. Shin H, and A. Tsourdos, "Multi-sensor multi-target tracking using domain knowledge and clustering," IEEE Sensors Journal, vol. 18, no. 19, pp. 8074-8084, 2018, doi: 10.1109/JSEN.2018.2863105.
[13] E. Delande, J. Houssineau, J. Franco, et al., "A new multi-target tracking algorithm for a large number of orbiting objects," Advances in Space Research, vol. 64, no. 3, pp. 645-667, 2019, doi: 10.1016/j.asr.2019.04.012.
[14] Y. Tian, A. Dehghan, and M. Shah, "On detection, data association and segmentation for multi-target tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9, pp. 2146-2160, 2018, doi: 10.1109/TPAMI.2018.2849374.
[15] W. Yi, Y. Yuan, R. Hoseinnezhad, et al., "Resource scheduling for distributed multi-target tracking in netted colocated MIMO radar systems," IEEE Transactions on Signal Processing, vol. 68, pp. 1602-1617, 2020, doi: 10.1109/TSP.2020.2976587.
[16] B. Traore B, B. Kamsu-Foguem, and F. Tangara, "Deep convolution neural network for image recognition," Ecological Informatics, vol. 48, pp. 257-268, 2018, doi: 10.1016/j.ecoinf.2018.10.002.
[17] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network," EURASIP Journal on Image and Video Processing, vol. 2019, no. 1, pp. 1-11, 2019, doi: 10.1186/s13640-019-0417-8.
[18] A. Sharma, E. Vans, D. Shigemizu, et al., "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," Scientific reports, vol. 9, no. 1, pp. 1-7, 2019, doi: 10.1038/s41598-019-47765-6.
[19] T. Shanthi and S. Sabeenian R, "Modified Alexnet architecture for classification of diabetic retinopathy images," Computers & Electrical Engineering, vol. 76, pp. 56-64, 2019, doi: 10.1016/j.compeleceng.2019.03.004.
[20] S. Lu, Z. Lu, and D. Zhang Y, "Pathological brain detection based on AlexNet and transfer learning," Journal of computational science, vol. 30, pp. 41-47, 2019, doi: 10.1016/j.jocs.2018.11.008.
[21] M. Hosny K, A. Kassem M, and M. Fouad M, "Classification of skin lesions into seven classes using transfer learning with AlexNet," Journal of digital imaging, vol. 33, no. 5, pp. 1325-1334, 2020, doi: 10.1007/s10278-020-00371-9.
[22] J. Kadam V, M. Jadhav S, and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," Journal of medical systems, vol. 43, no. 8, pp. 1-11, 2019, doi: 10.1007/s10916-019-1397-z.
[23] J. Yang, Y. Bai, F. Lin, et al., "A novel electrocardiogram arrhythmia classification method based on stacked sparse auto-encoders and softmax regression," International Journal of Machine Learning and Cybernetics, vol. 9, no. 10, pp. 1733-1740, 2018, doi: 10.1007/s13042-017-0677-5.
[24] L. Chen, M. Zhou, W. Su, et al., "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," Information Sciences, vol. 428, pp. 49-61, 2018, doi: 10.1016/j.ins.2017.10.044.
[25] D. Liang C, L. Wang, Y. Yao X, et al., "Multi-target tracking of networked heterogeneous collaborative robots in task space," Nonlinear Dynamics, vol. 97, no. 2, pp. 1159-1173, 2019, doi: 10.1007/s11071-019-05038-x.
[26] H. Kim S and L. Choi H, "Convolutional neural network for monocular vision-based multi-target tracking," International Journal of Control, Automation and Systems, vol. 17, no. 9, pp. 2284-2296, 2019, doi: 10.1007/s12555-018-0134-6.
[27] T. Tesfaye Y, E. Zemene, A. Prati, et al., "Multi-target tracking in multiple non-overlapping cameras using fastconstrained dominant sets," International Journal of Computer Vision, vol. 127, no. 9, pp. 1303-1320, 2019, doi: 10.1007/s11263-019-01180-6.
[28] I. Ahmed, S. Din, G. Jeon, et al., "Towards collaborative robotics in top view surveillance: A framework for multiple object tracking by detection using deep learning," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 7, pp. 1253-1270, 2020, doi: 10.1109/JAS.2020.1003453.
[29] L. Li, X. Wang, Z. Liu, et al., "A novel intuitionistic fuzzy clustering algorithm based on feature selection for multiple object tracking," International Journal of Fuzzy Systems, vol. 21, no. 5, pp. 1613-1628, 2019, doi: 10.1007/s40815-019-00645-7.
[30] W. Gan, S. Wang, X. Lei, et al., "Online CNN-based multiple object tracking with enhanced model updates and identity association," Signal Processing: Image Communication, vol. 66, pp. 95-102, 2018, doi: 10.1016/j.image.2018.05.008.