# Pedestrian Target Recognition Algorithm in Public Places Based on Representation Learning and Similarity Learning

Xiaowen Li
School of Computer and Artificial Intelligence, Henan Finance University, China
kittylxw@163.com

**Abstract:** *To ensure public safety, the government has placed a lot of cameras in public places and used them to monitor key targets. In target detection, pedestrian target detection is undoubtedly a research hotspot. How to realize the high efficiency of pedestrian detection is the focus of this field. As a result, this research suggests an algorithm for pedestrian target detection in public spaces that is based on representation learning and similarity learning. The algorithm uses representation learning to extract pedestrian features and Singular Value Decomposition (SVD) to build a more trustworthy Feature Extraction Network (FEN). In addition, the improved softmax function is used for similarity learning, and the K-Nearest Neighbor algorithm (KNN is applied for image retrieval to fgreatly increase the identification accuracy of pedestrians. The algorithm proposed in this study only needs two rounds of constraint training to achieve the best state. The mean absolute error and mean square error are 0.31 and 9.38, respectively. Its Relative Robustness (RR), Relative Generalization (RG) and Relative Scalability (RS) are excellent. In the final practical test, the model finally achieved 98.8% accuracy and 2.1% false positive rate. The proposed algorithm in this study has good application value in pedestrian target detection, and can better promote the development of social public safety.*

**Keywords:** *DenseNet, singular value decomposition, KNN algorithm, AM-softmax function, pedestrian target recognition*.

*Received October 30, 2023; accepted March 21, 2024*
*https://doi.org/10.34028/iajit/21/3/3*

## 1. Introduction

For social and economic development, public safety has always been a very important issue. Currently, various new crimes and modus operandi are also increasing [4]. In order to effectively detect criminal behavior, the government has installed a large number of camera devices in various public places [22]. Video surveillance systems can be used to obtain a huge quantity of video data. From these video data, the information of interested target personnel can be extracted to realize the trajectory tracking and behavior of key personnel and ensure public safety [7]. However, the quantity of video data gathered is rather substantial, and if only the traditional manual recognition method is used, it is not only difficult to extract the key information in these video data, but also the efficiency is very low [18]. Recently, using computer vision effect to optimize the intelligence level of video surveillance has become a research hotspot. It not only meets the actual needs, but also has great significance for the monitoring of mass behavior in reality [23]. From a technical point of view, the pedestrian recognition system currently used in monitoring devices can generally be divided into three functional units, namely pedestrian detection, pedestrian tracking and pedestrian retrieval [21]. The first two parts generally belong to separate topics (target detection and target tracking), so this study will focus on pedestrian retrieval for pedestrian recognition research. Thus, this study suggests a pedestrian recognition model based on representation learning and similarity learning. This method uses representation learning to extract pedestrian features through the classification effectiveness of the basic network to improve the difficulty of the same pedestrian across scenes recognition in traditional pedestrian recognition algorithms. In addition, similarity learning is used to extract the correlation between pedestrian features to optimize the recognition accuracy of traditional pedestrian recognition algorithms. There are three innovations in this research. The first is that the characteristics of pedestrians are extracted using Convolutional Neural Network (CNN), which solves the matching problem when the appearance of the same pedestrian changes dramatically across scenes in multi-camera video surveillance. The second point is the use of Singular Value Decomposition (SVD) to remove the correlation between the CNN output all connected layers to construct a reliable Feature Extraction Network (FEN). The third point is that the improved Softmax function is used for similarity learning, and the K-Nearest Neighbor algorithm (KNN) is used for data reordering and image retrieval. This study is split into four parts: the first part is a summary of related study areas; the second part is the construction of the method proposed in this paper; the third part is the verification of the constructed method. The last part is a

comprehensive summary of the content of this study.

## 2. Related Works

The research topic of pedestrian target recognition is derived from multi-camera target tracking. The standard of pedestrian target recognition is "to recognize a person when they return to the field of view after leaving the field of view." Saho *et al*. [15] suggested a pedestrian recognition method using Doppler radar and CNN to solve the issue of visual interference existing in traditional visual pedestrian recognition systems in the field of contact-free pedestrian recognition. It can accurately identify pedestrians from sitting to standing or from standing to sitting, and can maximize the impact of occluders. In the conventional pedestrian target recognition system, Sumari *et al*. [16] introduced a full-frame person recognition device to solve the problem of fuzzy person recognition in some cases. After evaluating the specific indicators of its implementation, the hybrid human-machine collaboration framework is standardized. This framework performs well in experiments and plays a driving role in the research of pedestrian target recognition. In order to minimize the occurrence of pedestrian traffic accidents at night, Ogura *et al*. [14] studied the method of pedestrian recognition at night. A method of continuous night image conversion using vehicle-mounted camera is proposed. This method can effectively improve the personal safety of pedestrians in poor lighting conditions at night. To enhance the convenience of pedestrian target recognition and the performance of the method, Li *et al*. [10] proposed a multi-task learning method combining pedestrian attributes and identity in the traditional pedestrian target recognition system. Through simulation experiments and actual detection, this method is in a leading position in the field of pedestrian recognition, and it reduces the false alarm rate and has low power consumption performance. To enhance the detection error rate of pedestrian recognition in intelligent transportation system, Dow *et al*. [5] combined with neural network and traditional visual pedestrian recognition system to construct a real-time pedestrian recognition system. This system ensures high precision and reduces the error rate. Its accuracy is far higher than that of similar models, and it has high efficiency, effectively promoting the development of pedestrian target recognition in traffic. To address the issue of poor recognition accuracy in conventional pedestrian target identification, An *et al*. [2] suggested a novel hierarchical reasoning network for pedestrian attribute recognition. This method has excellent performance and reduces the training cost to a certain extent.

Representation learning and similarity learning process the collected pedestrian video data by feature extraction, through the similarity measurement to obtain the similarity of each feature in the world, and then judge whether these features belong to the same pedestrian. To handle the issue that it is tough for ordinary learning prediction systems to predict learners' performance in new problems, Gan *et al*. [8] constructed a model for learning graph representations with additional enhancements. This method successfully predicts learners' performance in new questions. Xue *et al*. [20] constructed a sub supervised learning method based on mutual information representation learning in the traditional field of visual learning to handle the issue that most video learning still requires manual annotation. This method can eliminate human interference in the process of video learning, and its generalization is good and its efficiency is good. It can learn representations from videos without manual annotation, effectively saving manpower and material resources, and improving accuracy. In the field of electron observation in materials, Na *et al*. [13] focus on the classical bandgap problem. In order to observe crystal compounds as comprehensively as possible and solve the complicated calculation of bandgap problem for crystal compounds at present, a new representation learning method using a newly developed meta-graph neural network is suggested. The band gap of crystalline substances could be predicted with high precision using this approach, which relied on machine learning. The model's precision is superior to the standard density functional theory calculation, and the cumbersome calculation process is eliminated, the work efficiency is greatly improved, and the development of crystal chemistry is effectively promoted. To realize the graph-based clustering method, Zhong and Pun [24] designed a similarity learning algorithm based on augmented Lagrange multipliers. This method effectively obtains the optimal solution and maintains excellent performance. Luo *et al*. [12] suggested an effective unsupervised hashing method based on similarity learning under the background of popular hashing algorithms to solve the shortcomings of the popular deep unsupervised hashing methods in recent years. This method is superior to many current benchmark methods and effectively promotes the development of hash algorithms. Ali *et al*. [1] proposed to apply transfer learning technology to the target task of insufficient annotation data, and proved that this method can enrich prior knowledge of data in related fields, significantly improve model performance, and effectively avoid overfitting.

To sum up, there are still some problems in the current pedestrian recognition algorithm. For example, it is difficult to identify a group of people when their appearance changes dramatically across scenes. As a result, this paper uses a Deep Learning (DL)-based system for pedestrian detection. This model uses the representation learning method to obtain the image features of pedestrians from the collected video data and learn the similarity measure. To simplify the training and use of the algorithm as much as possible, CNN-

based representation learning is used to extract features. To making the most of the relationship between the images in the dataset, a recognition algorithm based on similarity learning is adopted.

## 3. Construction of Pedestrian Target Recognition Algorithm in Public Places Based on Representation Learning and Similarity Learning

Traditional pedestrian recognition algorithms are often complicated in computation and training, and are prone to difficulties in recognizing the same pedestrian when the appearance changes dramatically across scenes. Thus, this paper constructs a pedestrian recognition algorithm based on representation learning and similarity learning. CNN is used to extract pedestrian features, which simplifies the training and learning process, and solves the difficulty of recognizing the same pedestrian when the appearance changes dramatically across scenes. SVD algorithm is applied to construct the FEN, which makes the extracted pedestrian features robust and reliable. The Additive Margin Softmax (AM-Softmax) function is used for similarity learning. The KNN is used for image retrieval.

### 3.1. Construction of Pedestrian Target Recognition Algorithm in Public Places Based on Representation Learning

In this study, representation learning is used to solve the problem of pedestrian recognition. A pair of pedestrian images are input, CNN is adopted as the FEN, and similarity learning is adopted to obtain the similarity between the features extracted in each pedestrian image, so as to determine whether they belong to the same pedestrian [6]. CNN usually consists of verification model and recognition model. In the verification model, if two images belong to the same pedestrian, they are mapped to adjacent points in the feature space; for images belonging to different pedestrians, the mapping points in the special space will be far away. Figure 1 is a schematic of the validation model.

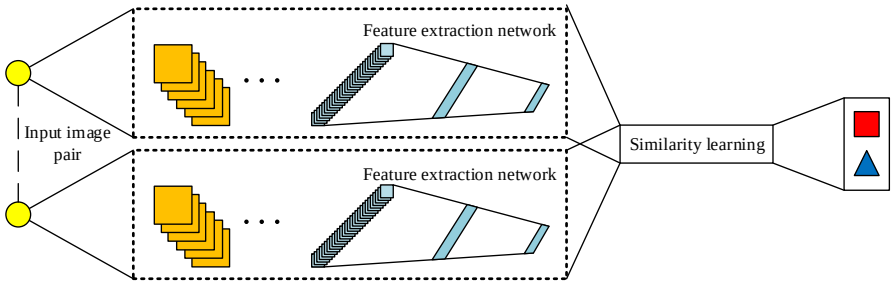



Figure 1. Validation model diagram.

In the field of pedestrian recognition, performance learning is a common method of pedestrian re-recognition. Through the processing of video data, a robust and stable pedestrian image appearance feature is extracted as a descriptive feature. In the field of pedestrian recognition, the training set of pedestrian recognition can be regarded as the process of classifying pictures according to different identity characteristics. Therefore, pedestrian recognition and image classification often have a certain correlation. Therefore, pedestrian recognition often uses a pre-trained general image classification network as the basic network for image extraction. Figure 2 shows the general process of pedestrian feature extraction using CNN.

In the feature extraction part of pedestrian recognition algorithm, the last step of feature extraction for pedestrian image is usually pedestrian feature matching. For image information, it is necessary to find an effective image feature similarity measurement criterion, so that similarity measurement can be carried out according to the extracted pedestrian feature vector, and then it can be determined whether the test samples belong to the same pedestrian. Then, according to the similarity between them, the order is sorted, so as to realize the recognition of pedestrian targets. The classical similarity measurement method adopted in this study is Euclidean distance, and its calculation process is shown in Equation (1).

$$d_{12}(a,b) = \sqrt{(a-b)+(a-b)^T} \tag{1}$$

In Equation (1), $a$ represents an eigenvector $a(x_{11}, x_{12}, \ldots, x_{1n})$ of dimension $n$, and $b$ represents an eigenvector $b(x_{11}, x_{12}, \ldots, x_{1n})$ of dimension $n$. In this study, the representation learning approach seeks to maximize the degree of resemblance across many images of the same pedestrian, and at the same time to make the similarity between different pictures of different pedestrians as far as possible. It is reflected in the loss function that the distance between the same pedestrian and the camera is closer than that between various pedestrians. For the convenience of expression, use $I_1$ and $I_2$ to represent the two input pictures respectively, then the calculation process of Euclidean distance between them is shown in Equation (2).

$$d_{I_1,I_2}\left\|f_{I_1} - f_{I_2}\right\|_2 \tag{2}$$

In Equation (2), vector $f_{I_1}$ and vector $f_{I_2}$ respectively represent the normalized feature vector obtained by the forward propagation of the image through CNN. The input to A twin network is usually A pair of images,

namely $I_1$ and $I_2$ in Figure 2. Each pair of input images may not have the same identity. Data labels are used to label images, which are denoted as $y$. If when $y$=1, the two pictures belong to the same pedestrian; Otherwise, it means that they belong to different pedestrians. Contrast loss is often used for training twin networks, and its calculation process is shown in Equation (3).

$$L_C = yd_{I_1,I_2}^2 + (1-y)\left(a - d_{I_1,I_2}\right)_+^2 \quad (3)$$

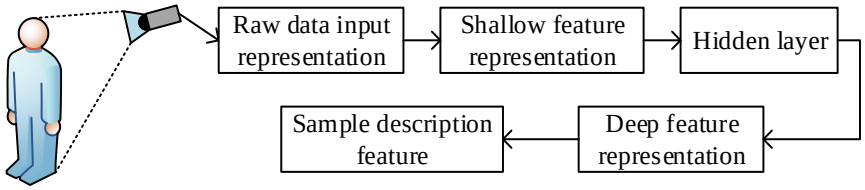



Figure 2. The process of CNN extracting pedestrian features.

In Equation (3), $(\bullet)_+$ represents max(●, 0). $a$ is the threshold parameter considered in the design according to the actual situation. In the pedestrian recognition work, the goal is to make the sample distance of the same identity closer. In a certain sense, the optimization process of neural network is the process of continuously minimizing the loss function $L_C$. In the actual training process, when the network input a positive sample pair, $d_{I_1,I_2}$ will gradually decrease; When a negative sample pair is entered into the network, it will cause $d_{I_1,I_2}$ to gradually grow larger than the preset threshold $a$. The classification performance and generalization ability of CNN model are excellent, and the features extracted by CNN model are more robust and reliable, so CNN is selected as a classification network. In this study, DenseNet model is selected, and its basic idea is similar to ResNet. However, the "short-circuit connection" in DenseNet is a Dense connection established between all the front layers and the back layers, so it is called DenseNet [3]. Figure 3 shows the structure diagram of DenseNet.

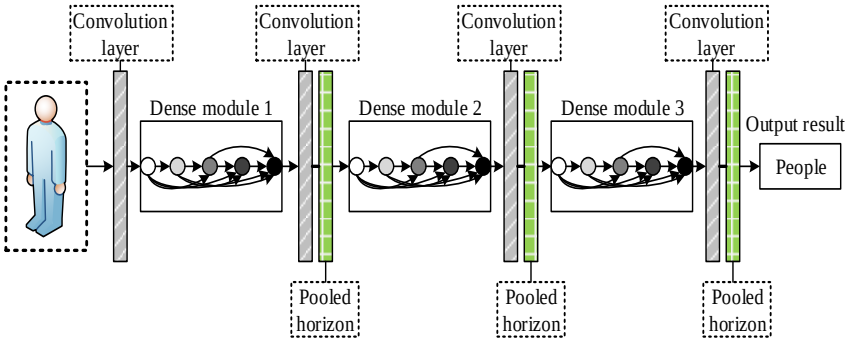



Figure 3. DenseNet structure diagram.

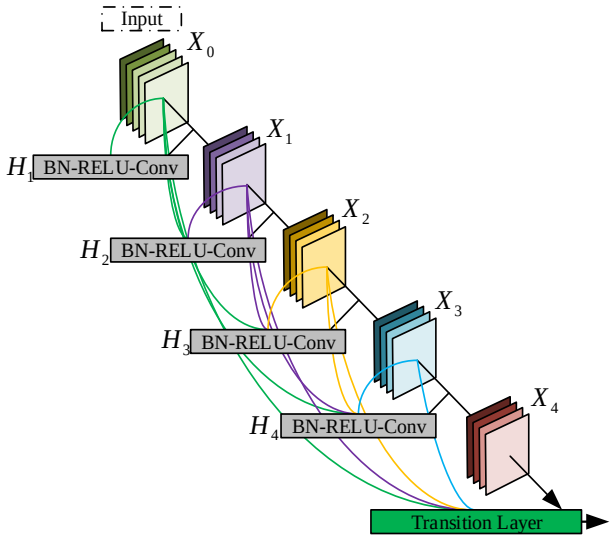



Figure 4. Dense module internal construction.

The basic structure of DenseNet mainly consists of dense modules and transition layers. The dense module is a module containing its own unique dense connection. The transition layer is the middle of the two dense modules. Dense connections exist only within a single dense module, not between different dense modules. This special network structure has better advantages. It resolves the gradient disappearance issue, improves the propagation of features, maximizes the usage of features, and, to a certain degree, minimizes the number of parameters. Dense module is the core structure of DenseNet. Figure 4 illustrates the specific internal structure of a dense module.

In Figure 4, it is a dense module with 5 layers, that is, it has a nonlinear transformation of 5 layers Batch Normalization (BN) layer +Relu layer+3*3 convolution layer. Its network growth rate is 4, indicating that the feature graph dimension of each layer is 4. In a traditional convolutional network, if the network has a

common layer $L$, the number of connections in the network is $L$. In A dense module containing layers $L$, a dense connection exists because each layer makes use of information from all previous layers in the module. That is, each of the front layers is connected to the back layer by a short circuit, so there is a dense $\frac{L*(L+1)}{2}$ connection. As shown in Figure 4, $X_0$ is the input of the entire dense module, so the input of the first set of nonlinear transformations $H_1$ is also $X_0$; the input of $H_2$ is sequentially deduced as $X_0$ and $X_1$; and so on for the rest. Usually in the ResNet model structure, the output $X_l$ of layer $l$ is the output $X_{l-1}$ of layer $l$-1 plus the nonlinear transformation to $X_{l-1}$, and its calculation process is shown in Equation (4).

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{4}$$

In Equation (4), $H_l$ represents the nonlinear transformation corresponding to layer $l$. ResNet is a short-circuit connection between each layer and a previous layer, which uses the addition of the element level as the connection method. In DenseNet, however, a more direct method of dense connection is used, which is to connect all the layers directly to each other in a more radical way. Each layer specifically takes all the layers that came before it as supplementary input. Therefore, for DenseNet, the output $X_l$ of $l$ is shown in Equation (5) [17].

$$x_l = H_l\left[\left(x_0, x_1, \ldots, x_{l-1}\right)\right] \tag{5}$$

Compared with ResNet, DenseNet uses dense concatenation to directly concatenate feature maps of different layers, thus realizing feature reuse and improving model efficiency. And the dense connection does not lead to an increase in computation and parameter number. Because the output from all preceding levels is included in each layer, this feature reuse requires very few feature maps, which decreases the number of parameters significantly. Each layer of this connection connects the input to the loss function, so the gradient loss problem is also mitigated, thus deepening the depth of the network.

## 3.2. Construction of Pedestrian Target Recognition Algorithm in Public Places Based on Representation Learning and Similarity Learning

Considering that the features have high correlation, it will affect the distance variable of Euclidean distance. Therefore, to further make the extracted pedestrian feature robust and reliable, this study uses SVD to remove the correlation between the weight vectors of the fully connected layer. For example, A real matrix $W=UV^T$ of size $n\times m$ can be decomposed using SVD as shown in Equation (6).

$$W = USV^T \tag{6}$$

In Equation (6), $U$ represents an orthonormal matrix of magnitude $n\times n$; $V$ represents an orthonormal matrix of magnitude $m\times m$; $S$ represents A diagonal matrix of magnitude $n\times m$; Matrix $W$ has a unique value equal to the value on the diagonal. $US$ is further used instead of $W$ to make it orthogonality without losing the ability to judge the feature representation in the entire sample space. Assuming that the input images are $x_i$ and $x_j$, the Euclidean distance can be expressed as shown in Equation (7).

$$\begin{aligned} d_{i,j} &= \left\|g_i - g_j\right\|_2 \\ &= \sqrt{\left(g_i - g_j\right)^T \left(g_i - g_j\right)} = \sqrt{\left(f_i - f_j\right)^T USV^T VS^T U^T \left(f_i - f_j\right)} \end{aligned} \tag{7}$$

In Equation (7), because $V$ is the identity orthogonal matrix, it can be further simplified as by Equation (8).

$$d_{i,j} = \sqrt{(f_i - f_j)^T USS^T U^T (f_i - f_j)} \tag{8}$$

In Equation (8), when $US$ is used instead of $W$, the Euclidean distance $d_{i,j}$ of the picture $x_i$ and $x_j$ does not change [9]. Therefore, the network performance is not affected. SVD is applied to the FEN based on DenseNet, and its main network structure is shown in Figure 5.

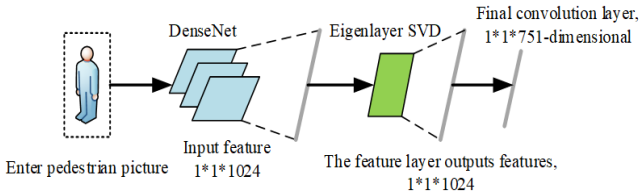



Figure 5. SVD network structure.

In Figure 5, the last fully connected layer of the initial network receives a convolutional layer. For more convenient expression, it is named the feature layer. In the training process, SVD decomposition is used to remove the correlation of the feature layer. In this study, the AM-Softmax loss function will be used for similarity learning. AM-Softmax function is improved by adding the Angle boundary to the traditional Softmax function, which is easier to calculate and has better performance. It is often used in multi-classification processes and can map the output of multiple neurons into the (0, 1) interval; the target sample obtained by using network prediction belongs to each classification. Equation (9) depicts the Softmax computation procedure.

$$L_S = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{W_{y_i}^T f_i}}{\sum_{j=1}^{c} e^{W_j^T f_i}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{\left\|W_{y_i}\right\|\left\|f_i\right\|\cos\left(\theta_{y_i}\right)}}{\sum_{j=1}^{c} e^{W_j^T f_i}} \tag{9}$$

In Equation (9), $n$ denotes the quantity of categories of the total; $f$ denotes the input to the last fully connected layer; $f_i$ is sample $i$; $Wj$ is the weight vector of column $j$ of the last fully connected layer. The calculation process of A-Softmax is shown in Equation (10).

$$L_S = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{\|f_i\|\psi(\theta_{y_i})}}{e^{\|f_i\|\psi(\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{c}e^{\|f_i\|\cos(\theta_{y_i})}} \quad (10)$$

In Equation (10), $\psi(\theta)$ represents the piecewise function, which can be specifically expressed as Equation (11).

$$\psi(\theta)=\frac{(-1)^k\cos(m\theta)-2k+\lambda\cos(\theta)}{1+\lambda},\theta\in\left[\frac{k\pi}{m},\frac{(k+1)\pi}{m}\right] \quad (11)$$

In Equation (11), $m$ usually represents an integer with a value greater than 1 and is used to adjust the distance between features; $\lambda$ represents a hyperparameter used to control classification boundaries to avoid network divergence. In this study, a specific $\psi(\theta)$ function is further used to introduce an additional boundary to the Softmax loss function.

$$\psi(\theta)=\cos\theta-m \quad (12)$$

In Equation (12), $\psi(\theta)$ is a monotonically decreasing function compared to the traditional definition, and is simpler and more intuitive. In Equation (11), $m$ is used to multiply $\theta$, but in Equation (12), cos $\theta$-$m$ is used. The angular distance is calculated in Equation (11), while the cosine distance is calculated in Equation (12). Since the inner product of $W$ and $f$ is obtained in the network, if cos ($\theta$-$m$) needs to be optimized, it will involve the inverse trigonometric function, which increases the calculation amount and makes the process slightly cumbersome. Therefore, cos $\theta$-$m$ is chosen instead of cos ($\theta$-$m$). The feature weight vector and the fully connected layer are normalized, and cosine similarity is employed to obtain the distance measure, and a hyperparameter $s$ is added as a scaling factor, then Equation (13) indicates the loss function.

$$\begin{aligned}L_{AMS}&=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot(\cos\theta_{y_i}-m)}}{e^{s\cdot(\cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}e^{s\cdot\cos\theta_j}}\\&=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot(W_{y_i}^{T}f_i-m)}}{e^{s\cdot(W_{y_i}^{T}f_i-m)}+\sum_{j=1,j\neq y_i}^{c}e^{sW_j^{T}f_i}}\end{aligned} \quad (13)$$

Figure 6 shows the boundary comparison between the traditional Softmax loss function and the AM-Softmax loss function in two dimensions. Among them, the distribution of the eigenvector after normalization is a circle. The vector $P_0$ represents the boundary of the traditional Softmax loss function. The AM-Softmax loss function is different in that its boundary is a marginal region rather than a single variable. AM-Softmax loss function makes the intra-class variance decrease and the inter-class distance increase.

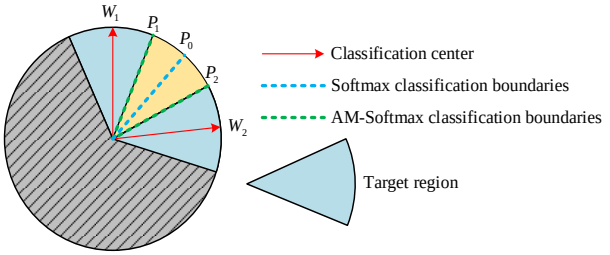



Figure 6. Comparison of the boundary between Softmax and AM-Softmax in two dimensions.

Due to the structural similarity between the training set for pedestrian recognition and the training set for image classification, the training set for pedestrian recognition does not match the pedestrian ID in the test set, so the traditional image classification method cannot be used to solve the problem of pedestrian recognition. Therefore, pedestrian recognition can be regarded as an image retrieval problem, which is processed by KNN algorithm in this study. Figure 7 is a schematic diagram of KNN algorithm.

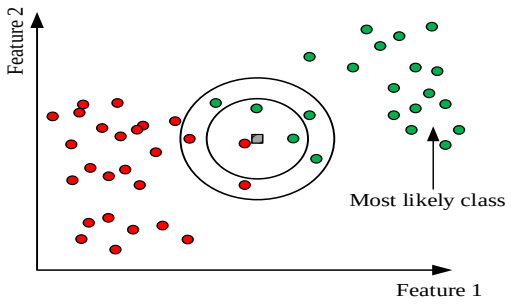



Figure 7. KNN algorithm diagram.

Firstly, the target pedestrian sample to be identified is set as $p$, and the candidate pedestrian sample set contains $N$ candidate pedestrian samples, which are denoted as $G=\{g_i|i=1, 2, \ldots, N\}$. Firstly, the feature expressions $p$ and $G$ of all samples in $f_p$ and $f_{g_i}$ are obtained by using the deep network proposed above, and the Euclidean distance $d(p, g_i)$ between $p$ and $g_i$ is further calculated. By sorting this distance, you get a sorting table $L=(p,G)=\{g_1^0,g_2^0,\ldots,g_N^0\}, d(p,g_i)<d(p,g_{i+1})$. However, due to the different changes of light, Angle, obstacles and posture in the actual image data, the correct matching result may be in a relatively backward position, so it is necessary to reorder it to further improve the accuracy of pedestrian recognition. Therefore, this study adopts KNN algorithm to reorder them, and defines $N(p, k)$ as the nearest $k$ neighbor of $p$.

$$N(p,k)=\{g_1^0g_2^0\ldots g_k^0\},|N(p,k)|=k \quad (14)$$

In Equation (14), if $A$ is similar to $B$, then $B$ must also be similar to $A$; therefore, the $k$ nearest neighbor $R(p, k)$ of the target $p$ is further defined as shown in Equation (15).

$$R(p,k)=\left\{g_i \left| \left(g_i \in N(p,k)\right) \cap \left(p \in N(g_i,k)\right)\right.\right\} \tag{15}$$

In Equation (15), $p$ and $g_i$ are neighbors. It is obvious that the $k$ nearest neighbor is closer to $p$ than the $k$ nearest neighbor. Therefore, the $\frac{1}{2}k$ mutual nearest neighbor $R\left(q,\frac{1}{2}k\right)$ of each candidate sample $q$ in $p$'s mutual nearest neighbor $R(p, k)$ is further used to extend $R(p, k)$ to make it a more robust $R^*(p, k)$. The specific conditions are shown in Equation (16).

$$\begin{cases} R^*(p,k)=R(p,k)\cup r\left(q,\frac{1}{2}k\right) \\ s.t\left|R(p,k)\cup R\left(q,\frac{1}{2}k\right)\right| \geq \frac{2}{3}\left|R\left(q,\frac{1}{2}k\right)\right|, \forall q \in R(p,k) \end{cases} \tag{16}$$

After the above processing, $R^*(p, k)$ contains more samples similar to the candidate samples than $R(p, k)$. This expansion process can add more positive samples to, $R^*(p, k)$ which helps to enhance the precision of pedestrian recognition. To determine whether a particular pedestrian is present in a picture, pedestrian recognition employs computer vision techniques. Firstly, the target detection technology is used to detect the pedestrian target that appears in the camera monitoring field for the first time. The second step is feature learning, that is, feature extraction of the pedestrian image information obtained through the deep network. The third step is similarity learning, mainly by learning the distance measurement function between pedestrian features, and using it to obtain the similarity between pedestrian features, so as to realize the target recognition of pedestrians. Its system framework is shown in Figure 8.

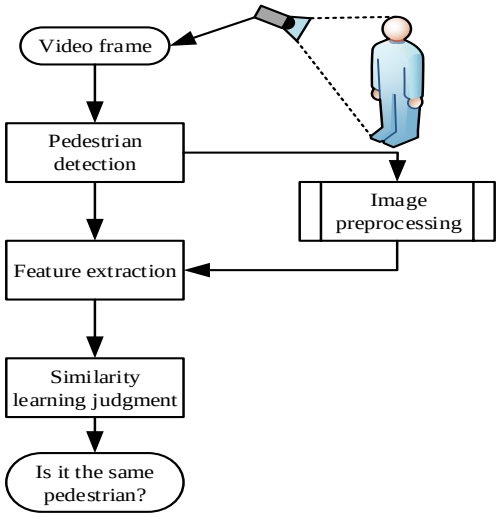



Figure 8. Pedestrian identification system framework.

In the field of pedestrian target recognition in public places, an algorithm framework based on representation learning and similarity learning is adopted. The algorithm uses CNN to extract features from pedestrian images. The multi-layer structure of CNN can automatically learn the representative features of pedestrians from large data sets, which not only improves the feature extraction ability of the model, but also simplifies the complexity of the learning process. In order to enhance the robustness of the extracted features, the SVD algorithm is introduced to construct the FEN. By removing the redundant information of the data and strengthening the key features, the obtained features are guaranteed to have high recognition accuracy and stability in different environments. In addition, the AM-Softmax function was adopted to optimize similarity learning, increasing the spacing between classes by introducing a boundary at the decision margin, thereby improving the discriminant efficiency of the model. Finally, the pedestrian image retrieval task is realized by the K-NN algorithm, which retrieves the target pedestrian image by direct comparison of feature similarity. This method provides fast and efficient re-recognition capability without explicit learning of discriminant function.

For the model parameter optimization problem, the CNN structures involved, such as the convolutional layer, the fully connected layer and the parameters of their associated activation functions, need to be fine-tuned in order to achieve the best performance in multi-level feature learning. In addition, the SVD method is used to further construct and refine the FEN, in which additional parameters such as dimensionality reduction need to be adjusted according to the characteristics of the data set to improve the strength and robustness of feature expression. The AM-Softmax function is used in similarity learning. Adding marginal parameters can enlarge the classification interval and improve the discriminating power of learning. The optimization of this parameter requires a series of experiments to evaluate its impact on classification performance, and consider combining cross-validation and other techniques to refine its value. As another important parameter that affects the efficiency and effect of the retrieval, the choice of K value of the K-NN algorithm must also be determined by experimental design to adapt to the data sets with different sizes and characteristics.

For the model training problem, the core training goal is to minimize the image feature distance of the same pedestrian and maximize the image feature distance of different pedestrians. The loss function used is designed to reflect this and improve the recognition performance of the model by ensuring that the distance between the feature vectors of the same pedestrian images is smaller than the distance between the feature vectors of different pedestrian images. The algorithm uses the architecture of twin network to evaluate the similarity of features by calculating the Euclidean distance between a pair of images. Each pair of input images may represent the same pedestrian or different pedestrians, annotated with data labels. The contrast loss function is used to train the twin network, which is designed based on minimizing

the distance between identical pedestrian pictures while making the distance between different pedestrian pictures exceed a preset threshold.

In the training process, CNN is first used to learn features from pedestrian images, and these features are optimized by contrast loss function, so that the algorithm can distinguish different pedestrians. The twin network architecture is used to process pairs of images and train the model by minimizing the distance between the same pedestrian images and maximizing the distance between different pedestrian images. SVD is used to improve the feature representation of fully connected layer weights and reduce overfitting. The DenseNet network architecture was chosen to enhance feature passing and reduce network parameters. The whole training is an iterative process until the algorithm shows good performance on the verification set. Throughout the process, techniques such as early stopping, regularization, and data augmentation may also be employed to improve model generalization.

## 4. Evaluation of Pedestrian Target Recognition Algorithms in Public Places Based on Representation Learning and Similarity Learning

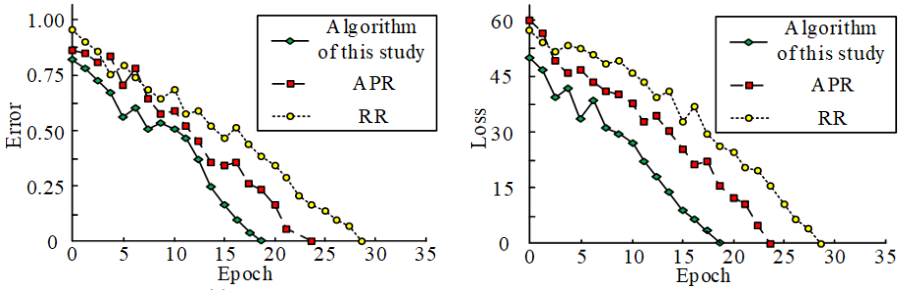
This study optimizes the pedestrian recognition algorithm based on DL from two perspectives of representation learning and similarity learning. To test the actual effectiveness of the constructed model and verify its effectiveness in practical application, two public mainstream pedestrian recognition data sets were selected to train and test the algorithm. Market1501 and Chinese University of Hong Kong datasets (CUHK03) were selected for the experiment. There is no overlap between the training set and the rows in the test set in the two datasets. The photos in these data sets reflect the challenges faced in the actual pedestrian recognition scene, such as the impact of uncertain factors such as light, Angle, obstacles, distance and posture on the visual images of pedestrians, taking into account inter-class differences and intra-class changes, so that the algorithm can be evaluated more reasonably and objectively. The quality of the training set directly affects whether the algorithm can accurately identify the target. High-quality data sets should cover a wide range of situations, allowing algorithms to learn more generalized feature representations. If the quality of the data set is poor, such as a single sample, inaccurate labeling, or insufficient sample size, the algorithm may only perform well on the training set and cannot adapt to new or unseen data, which is the phenomenon of overfitting. Overfitting means that the algorithm is over-sensitive to the training data, captures the noise and contingency in the training data, and fails to grasp the essential characteristics of pedestrian recognition, which will lead to the performance of the algorithm in practical applications. Therefore, ensuring the high quality of the data set can not only improve the overall performance of the algorithm, but also reduce the risk of overfitting, making the algorithm more robust and reliable in the face of diverse real-world data.

In this study, Attribute-Person Recognition network (APR) and Re-Ranking (RR), which are also DL-based pedestrian recognition algorithms in recent years, are further selected for analogy. Experiments of all algorithms proposed in this study are completed by calling python and MATLAB interfaces in the DL framework Caffe, and the specific software and Table 1 shoes the hardware environments of the experimental platform.

Table 1. Experimental software and hardware environmental parameters.

| Software and hardware environment name | Parameter specification |
|---|---|
| Device type | DL server |
| Processor | Intel(R) Xeon(R) CPU E5-2640 v4@ 2.40Ghz |
| Graphics card | Nvidia GTX 1080 Ti 11G *2 |
| Internal memory | 80GB |
| Operating system | Ubuntu 14.04 |
| Experimental platform | Caffe, Python 2.7.12(Anaconda 4.2.0), MATLAB R2015b |

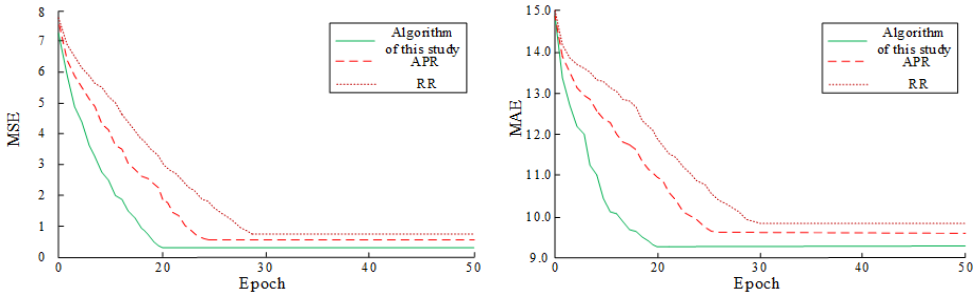
In Figure 9, firstly, the convergence of several algorithms is tested and compared. In the training process, the improved algorithm can converge faster to reach the target precision value and the best Loss value faster; It only needs to be iterated 19 times, which is 6 times and 11 times ahead of the APR algorithm and RR algorithm respectively. The improved model has a simpler training process and its convergence is better than the APR algorithm and RR algorithm.




a) Error. b) Loss.

Figure 9. Convergence of three algorithms.

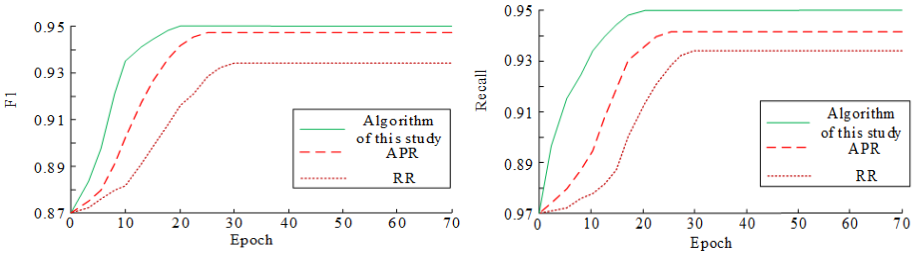
The Mean Squared Error (MSE) and Mean Absolute Error (MAE) values of the three algorithms are compared to evaluate their error performance in practice. In Figure 10, the MSE value of the proposed algorithm reached 0.31, which was 0.12 and 0.37 lower than that of the APR algorithm and the RR algorithm, respectively. The MAE value reached 9.38, which was 0.38 and 0.81 lower than the APR algorithm and RR algorithm, respectively. This indicates that the algorithm proposed in this study has a lower error value.




a) MSE. b) MAE.

Figure 10. MSE and MAE of three algorithms.

In Figure 11, F1 and Recall values of the three algorithms are evaluated using data sets. Among them, F1 of the algorithm proposed in this study has the fastest growth, with the highest value reaching 0.949, which is 0.006 and 0.019 ahead of the APR algorithm and RR algorithm, respectively. The recall value reached 0.948, which was 0.011 and 0.018 higher than the APR algorithm and RR, respectively.




a) F1. b) Recall.

Figure 11. F1 and recall of three algorithms.

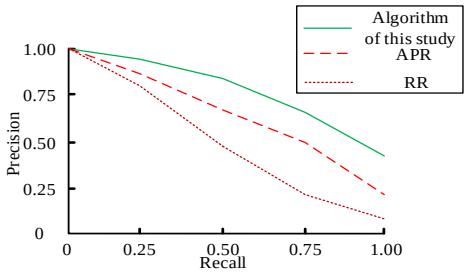
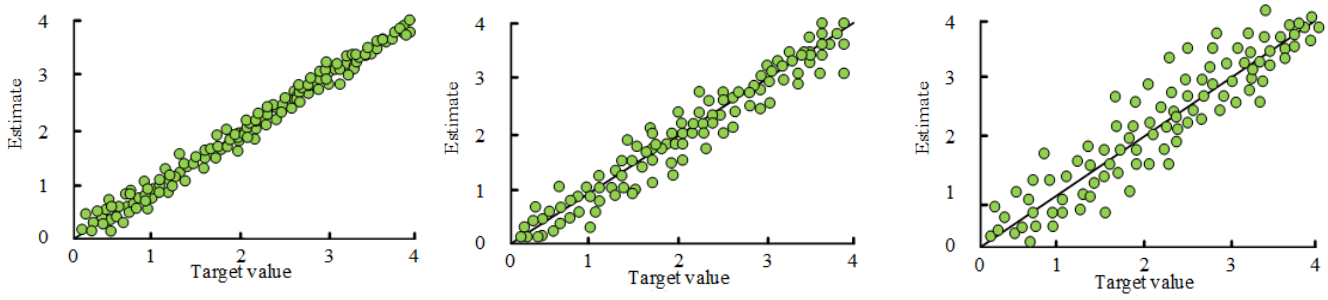



Figure 12. Precision/recall of three algorithms.

In Figure 12, to measure the efficiency of the model, the accuracy rate/recall rate curve was established. With the increase of recall rate, the accuracy rate of the proposed algorithm declines slowly, while the accuracy rate of the APR algorithm and the RR algorithm declines more obviously. The curve of the extraction algorithm in this study is always the highest, so it can be concluded that the comprehensive performance of the extraction model is superior to APR and RR.

In Figure 13, to further evaluate the fit degree of the three algorithms, simulation experiments are conducted using data sets. The fit degree of the improved algorithm in this study reaches 0.982, which is higher than that of the APR and RR algorithms by 0.048 and 0.101 respectively. This study's model has better practical significance in practice.




a) Algorithm of this study. b) APR. c) RR.

Figure 13. Fit of three algorithms.

In Figure 14, several algorithms are trained by constraints to compare their rank-1 accuracy with the variation trend of mAP value. After the first round of constraint training, the rank-1 accuracy and mAP accuracy of the proposed algorithm increase rapidly. However, it can reach the highest value after two rounds of Progressive Reinforcement Iteration (PRI) training, and then fluctuate slightly. The APR algorithm and RR algorithm need to reach the highest value in the third round of PRI training, indicating that better convergence performance is achieved by the approach suggested in this work. Finally, the rank-1 accuracy of the suggested model achieves 85.90%, which is 0.2% and 0.4% higher than APR algorithm and RR algorithm, respectively. The mAP value of the suggested model achieves 65.4%, which is 0.3% and 0.7% higher than APR algorithm and RR algorithm, respectively.

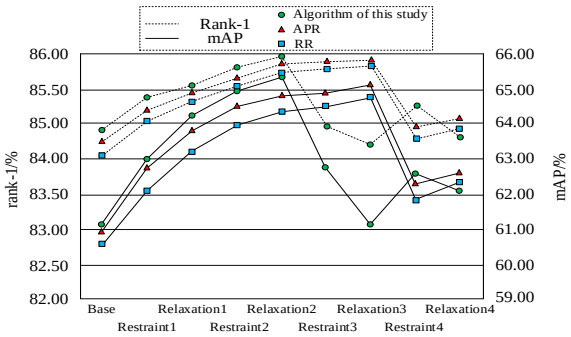



Figure 14. Rank-1 and mAP of three algorithms.

To further test the superiority of the model, this research selects other popular pedestrian recognition algorithms such as Pose Sensitive Embedding (PSE), Pose algorithm, Singular Value Decomposition Network algorithm (SVDNet), MultiLoss algorithm, You Only Look Once version 4 (YOLOv4) [11] and Mask Region Convolutional Neural Network Model Based on Attention Mechanism (MRCNNAM) [19]. Test data sets are used to compare rank-1 accuracy values, mAP accuracy values, Relative Robustness (RR), Relative Generalization (RG), and Relative Scalability (RS). In Table 2, the rank-1 accuracy of the algorithm proposed in this study reaches 89.58%; its mAP accuracy also reached 72.48%, which was the highest value. At the same time, the proposed model has the best relative robustness, generalization and expansibility. This means that the study's model possesses a higher precision value and can be more accurate in pedestrian recognition.

Table 2. The comparison results of nine algorithms on the test set.

| Algorithms | Rank-1(%) | mAP (%) | RR(%) | RG(%) | RS(%) |
|---|---|---|---|---|---|
| **APR** | 84.97 | 64.44 | 71.27 | 65.27 | 77.14 |
| **RR** | 82.75 | 63.84 | 73.47 | 66.94 | 78.24 |
| **PSE** | 85.42 | 61.67 | 76.28 | 69.62 | 79.51 |
| **POSE** | 81.95 | 65.34 | 78.39 | 75.14 | 80.36 |
| **SVDNet** | 83.45 | 66.84 | 79.97 | 76.84 | 81.95 |
| **MultiLoss** | 84.28 | 67.51 | 82.14 | 77.28 | 82.84 |
| **YOLOv4** | 85.12 | 68.71 | 84.65 | 79.34 | 83.49 |
| **MRCNNAM** | 87.25 | 69.63 | 86.25 | 81.41 | 84.62 |
| **Algorithm of this study** | **89.58** | **72.48** | **91.24** | **85.29** | **86.97** |

In Table 3, the average accuracy rate, average false positive rate and evaluation false negative rate of several algorithms are tested using the actual scenario. The average accuracy of the suggested model achieves 98.8%. The average false alarm rate is 2.1%. On the other hand, the average failure rate is 0.2%, which is the best performance in several values. In summary, suggested model is advanced to a certain extent, and it is also verified that the proposed algorithm can work together with popular target detection algorithms and target tracking algorithms in the actual constructed scene, and has better performance than other algorithms.

Table 3. Average accuracy, false alarm rate and missing report rate of nine algorithms.

| Algorithm | Average accuracy/% | Average false alarm rate/% | Average missing report rate/% |
|---|---|---|---|
| **APR** | 95.8 | 4.3 | 6.1 |
| **RR** | 93.4 | 5.8 | 7.9 |
| **PSE** | 92.4 | 6.2 | 7.1 |
| **POSE** | 91.8 | 7.4 | 8.1 |
| **SVDNet** | 94.6 | 5.9 | 6.9 |
| **MultiLoss** | 95.5 | 4.8 | 6.2 |
| **YOLOv4** | 96.7 | 3.5 | 4.9 |
| **MRCNNAM** | 97.2 | 3.1 | 2.7 |
| **Algorithm of this study** | **98.8** | **2.1** | **0.2** |

## 5. Discussion

The representation learning adopted in the study is a common way to automatically discover data representations in order to better identify or classify data. Similarity learning is concerned with learning a task so that the algorithm can judge the similarity between different data points. Saho *et al.*'s [15] method of combining Doppler radar with CNNs, which is actually an application of representation learning. They use radar data to enhance visual information to provide a more comprehensive representation of the data, which facilitates pedestrian identification in complex environments. Sumari *et al.* [16] used a full-frame person recognition device, which may rely on the extracted full-frame features for recognition. This is also an application scenario of representation learning, in which the performance is improved by learning to recognize the features of the entire screen. The night pedestrian recognition technology highlighted by Ogura *et al.* [14] Adapts to the low-light environment through image conversion, which also relies on representation learning, because it is necessary to extract feature representations that can work effectively at night through the network. The multi-task learning method of Li *et al.* [10] combines pedestrian attributes and identities. In fact, this method uses additional label information to enhance the differentiation ability of representations under the framework of representation learning. The real-time pedestrian recognition system of Dow *et al.* [5] combines DL classifier and zebra crossing recognition, which is also a manifestation of representation learning, and improves the accuracy and speed of real-time recognition by jointly learning

multiple features and scene information.

In the field of public safety, the research algorithm can extract the characteristics of pedestrians through cameras or other sensors and identify their identity or behavior pattern, which is used for real-time monitoring of crowded areas such as airports, stations, shopping malls, etc., so as to timely detect suspicious behavior or manage emergency situations. For autonomous driving, research algorithms can help vehicles more accurately identify and track pedestrians, even in the complex and changeable urban traffic environment, can also improve the safety of driving and reduce the risk of accidents. In general, the research content has important value in ensuring personnel safety and improving traffic management efficiency.

Although the human object recognition algorithm performs well in laboratory conditions, its performance in real-world variable environments still needs to be evaluated. Specifically, in real-world applications, pedestrian target recognition faces a variety of challenges, including but not limited to lighting changes in different environments, various types of occlusion, and changes in human posture. These factors are likely to affect the performance of the algorithm, especially when the pedestrian target is in a crowded or complex background. Future research will focus on improving the adaptability and robustness of the algorithm in various environments. Further work will include optimizing the algorithm to solve problems such as occlusion and complex lighting variations in real-world scenarios, and testing it on larger data sets to confirm the algorithm's performance in real-world applications. Possible directions include introducing techniques such as color space conversion and dynamic range compression to deal with image recognition under extreme lighting conditions.

## 6. Conclusions

Crime in today's society cannot be ignored and is always difficult to contain. Therefore, based on cameras placed in public places, this paper constructs a pedestrian recognition algorithm based on representation learning and similarity learning. In this method, the FEN based on CNN is adopted, and the pedestrian feature parameters are extracted by representation learning. At the same time, similarity learning is used to calculate the similarity measure of features, so as to realize the recognition of pedestrian identity. In this study, the algorithm can reach the best state only 19 times of conventional training; F1 value and Recall value increased rapidly, with the highest values reaching 0.949 and 0.948, respectively. The fit degree reached 0.982. The MAE and MSE values were 0.31 and 9.38, respectively. It also performed well in restraint training, only two rounds to achieve the best. The final rank-1 accuracy and mAP accuracy reached 89.58% and 72.48%, respectively. The accuracy rate/recall rate curve of the suggested model in this study also performs well, and it is the best curve overall. In the final simulation experiment, the proposed algorithm can be well combined with the current mainstream pedestrian target detection algorithm and pedestrian tracking algorithm, indicating that its practicability is good. It achieved 98.8 percent accuracy, 2.1 percent false positives and 0.2 percent false positives. The relative robustness reaches 91.2%, the RG reaches 85.29, and the RS reaches 86.97%. The algorithm in this study has made remarkable progress in processing speed and recognition accuracy, which is of great significance for the real-time monitoring system in the field of public security. However, despite its excellent performance in a laboratory setting, the algorithm's adaptability in diverse and dynamically changing real-world scenarios remains to be further validated. Going forward, research will continue to focus on the adaptability and robustness of algorithms in a wider range of environments. Further work will include optimizing the algorithm to handle issues such as occlusion, lighting changes in more complex scenes, and testing it on larger data sets to verify its actual performance in real-world applications.

## References

[1] Ali A., Yaseen M., Aljanabi M., and Abed S., "Transfer Learning: A New Promising Techniques," *Mesopotamian Journal of Big Data*, vol. 2023, pp. 29-30, 2023. https://doi.org/10.58496/MJBD/2023/004

[2] An H., Hu H., Guo Y., Zhou Q., and Li B., "Hierarchical Reasoning Network for Pedestrian Attribute Recognition," *IEEE Transactions on Multimedia*, vol. 23, no.1, pp. 268-280, 2021. DOI:10.1109/TMM.2020.2975417

[3] Bagyaraj S., Tamilselvi R., Gani P., and Sabarinathan D., "Brain Tumour Cell Segmentation and Detection Using Deep Learning Networks," *IET Image Processing*, vol. 15, no. 10, pp. 2363-2371, 2021. https://doi.org/10.1049/ipr2.12219

[4] Bi D., Kadry S., and Kumar P., "Internet of Things Assisted Public Security Management Platform for Urban Transportation Using Hybridised Cryptographic-Integrated Steganography," *IET Intelligent Transport Systems*, vol. 14, no. 11, pp. 1497-1506, 2020. DOI:10.1049/iet-its.2019.0833

[5] Dow C., Ngo H., Lee L., Lai P., Wang K., and Bui V., "A Crosswalk Pedestrian Recognition System by Using Deep Learning and Zebra-Crossing Recognition Techniques," *Software: Practice and Experience*, vol. 50, no. 5, pp. 630-644, 2020. https://doi.org/10.1002/spe.2742

[6] Fang B., Jiang M., Shen J., and Stenger B., "Deep Generative Inpainting with Comparative Sample Augmentation," *Journal of Computational and Cognitive Engineering*, vol. 1, no. 4, pp. 174-180,

2022. DOI: 10.47852/bonviewJCCE2202319

[7] Fetene D., Higgs P., Nielsen S., Djordjevic F., and Dietze P., "The impact of Victoria's Real Time Prescription Monitoring System (SafeScript) in a Cohort of People Who Inject Drugs," *The Medical Journal of Australia*, vol. 214, no. 5, pp. 234-235, 2021. DOI: 10.5694/mja2.50958

[8] Gan W., Sun Y., and Sun Y., "Knowledge Structure Enhanced Graph Representation Learning Model for Attentive Knowledge Tracing," *International Journal of Intelligent Systems*, vol. 37, no. 3, pp. 2012-2045, 2021. https://doi.org/10.1002/int.22763

[9] Kapoor S., Sharma A., Verma A., Dhull V., and Goyal C., "A Comparative Study on Deep Learning and Machine Learning Models for Human Action Recognition in Aerial Videos," *The International Arab Journal of Information Technology*, vol. 20, no. 4, pp. 567-574, 2023. https://doi.org/10.34028/iajit/20/4/2

[10] Li C., Yang X., Yin K., Chang Y., Wang Z., and Yin G., "Pedestrian Re-Identification Based on Attribute Mining and Reasoning," *IET Image Processing*, vol. 15, no. 11, pp. 2399-2411, 2021. https://doi.org/10.1049/ipr2.12225

[11] Li K., Zhuang Y., Lai J., and Zeng Y., "PFYOLOv4: An Improved Small Object Pedestrian Detection Algorithm," *IEEE Access*, vol. 11, pp. 17197-17206, 2023. DOI:10.1109/ACCESS.2023.3244981

[12] Luo X., Ma Z., Cheng W., and Deng M., "Improve Deep Unsupervised Hashing via Structural and Intrinsic Similarity Learning," *IEEE Signal Processing Letters*, vol. 29, no. 1, pp. 602-606, 2022. DOI: 10.1109/LSP.2022.3148674

[13] Na G., Jang S., Lee Y., and Chang H., "Tuplewise Material Representation Based Machine Learning for Accurate Band Gap Prediction," *The Journal of Physical Chemistry A*, vol. 124, no. 50, pp. 10616-10623, 2020. https://doi.org/10.1021/acs.jpca.0c07802

[14] Ogura R., Nagasaki T., and Matsubara H., "Improving the Visibility of Nighttime Images for Pedestrian Recognition Using in-Vehicle Camera," *Electronics and Communications in Japan*, vol. 103, no. 10, pp. 35-43, 2020. https://doi.org/10.1002/ecj.12268

[15] Saho K., Shioiri K., and Inuzuka K., "Accurate Person Identification Based on Combined Sit-to-Stand and Stand-to-Sit Movements Measured Using Doppler Radars," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4563-4570, 2020. DOI:10.1109/JSEN.2020.3032960

[16] Sumari F., Machaca L., Huaman J., Clua E., and Guérin J., "Towards Practical Implementations of Person Re-Identification from Full Video Frames," *Pattern Recognition Letters*, vol. 138, no. 10, pp. 513-519, 2020. https://doi.org/10.1016/j.patrec.2020.08.023

[17] Totaro S., Hussain A., and Scardapane S., "A Non-Parametric Softmax for Improving Neural Attention in Time-Series Forecasting," *Neurocomputing*, vol. 381, pp. 177-185, 2020. https://doi.org/10.1016/j.neucom.2019.10.084

[18] Uras M., Cossu R., Ferrara E., Liotta A., and Atzori L., "PmA: A Real-World System for People Mobility Monitoring and Analysis Based on Wi-Fi Probes," *Journal of Cleaner Production*, vol. 270, pp. 1-14, 2020. https://doi.org/10.1016/j.jclepro.2020.122084

[19] Wang Y., "MRCNNAM: Mask Region Convolutional Neural Network Model Based on Attention Mechanism and Gabor Feature for Pedestrian Detection," *Journal of Applied Science and Engineering*, vol. 26, no. 11, pp. 1555-1561. http://jase.tku.edu.tw/articles/jase-202311-26-11-0005

[20] Xue F., Ji H., and Zhang W., "Mutual Information Guided 3D ResNet for Self-Supervised Video Representation Learning," *IET Image Processing*, vol. 14, no. 13, pp. 3066-3075, 2020. DOI:10.1049/iet-ipr.2020.0019

[21] Yang F., Wang X., Zhu X., Liang B., and Li W., "Relation-Based Global-Partial Feature Learning Network for Video-Based Person Re-Identification," *Neurocomputing*, vol. 488, pp. 424-435, 2022. https://doi.org/10.1016/j.neucom.2022.03.032

[22] Zhang H., Li P., Du Z., and Dou W., "Risk Entropy Modeling of Surveillance Camera for Public Security Application," *IEEE Access*, vol. 8, no. 1, pp. 45343-45355, 2020. DOI:10.1109/ACCESS.2020.2978247

[23] Zhao Z., Sun R., Yang Z., and Gao J., "Visible-Infrared Person Re-Identification Based on Frequency-Domain Simulated Multispectral Modality for Dual-Mode Cameras," *IEEE Sensors Journal*, vol. 22, no. 1, pp. 989-1002, 2021. DOI:10.1109/JSEN.2021.3130181

[24] Zhong G. and Pun C., "Subspace Clustering by Simultaneously Feature Selection and Similarity Learning," *Knowledge-Based Systems*, vol. 193, no. 1, pp. 105512, 2020. https://doi.org/10.1016/j.knosys.2020.105512

**Xiaowen Li**, born in December 1984, female, Hui ethnicity, Zhoukou City, Henan Province. She obtained a Bachelor's degree in Computer Science and Technology from Henan University in 2007 and a Master's degree in Computer Application Technology from Henan University in 2010, majoring in Computer Networks and Artificial Intelligence. June 2010 to December 2014, worked as a teaching assistant at the former Defense Air Force Command College; from December 2014 to March 2018, lecturer at the former Defense Air Force Command College. From March 2018 to present, lecturer at Henan Finance University. Published 5 papers in domestic and foreign academic journals. Her areas of interest are Computer Networks, Intelligent Networks, and Artificial Intelligence.