Received March 29, 2020, accepted April 4, 2020, date of publication April 7, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986394

Pedestrian Re-Identification Monitoring System Based on Deep Convolutional Neural Network

WENZHENG QU, ZHIMING XU, BEI LUO, HAIHUA FENG, AND ZHIPING WAN
Department of Information Science, Xinhua College of Sun Yat-sen University, Guangzhou 510520, China

Corresponding author: Zhiping Wan (wzp888@xhsysu.edu.cn)

This work was supported in part by National Training Program of Innovation and Entrepreneurship for Undergraduates under Grant 201813902001, and in part by the Guangdong University Student Science and Technology Innovation Special Funding Project under Grant pdjh2019b0626.

ABSTRACT The gradual establishment of large-scale distributed camera networks and the rapid development of "Internet +" have resulted in the recent popularization of massive video surveillance systems. As pedestrians are the key monitoring targets in video surveillance systems, many studies are focusing on pedestrian re-identification monitoring algorithms across cameras. At present, the pedestrian re-identification model is not only faced with the difficulty of training the network model due to the huge quantity difference between different types of training samples, but also needs to reduce the impact of the large difference in visual performance on the model identification accuracy. To solve these difficulties, this paper proposed a deep learning model and designed a system based on a deep convolutional neural network for pedestrian re-identification. In particular, we determined the difference between the system input neighborhoods in order to derive the local relationship between the two input images, thus reducing the effects of illumination and perspective. Furthermore, we employed focal loss to solve the phenomenon of sample imbalance in the pedestrian re-identification process in order to enhance the actual application potential of the model. The proposed method was implemented in our developed end-to-end monitoring system for pedestrian re-identification. The hardware component of the system design framework was composed of a digital matrix, streaming media storage server and a network high-speed dome, with the ability to extend to additional tasks in the future. Our approach reduces the effects of data imbalances and visual performance differences, with a score of 76.0% for rank-1 and 99.5% for rank-20 on large data sets (CUHK03), which is not only a significant improvement over the previous IDLA, but also superior to other existing approaches.

INDEX TERMS Deep convolution, monitoring system, neural network, pedestrian re-identification.

I. INTRODUCTION

The establishment of large-scale distributed camera networks has put much attention on massive video surveillance systems. Surveillance cameras produce large numbers of videos with multiple data types at a low value density and a strong real-time performance. There are many challenges to security inspection [1], [28], [29], [37], [49]. However, the shooting range of current monitoring systems cannot cover all key monitoring areas, while in general, the areas monitored by multiple cameras do not overlap. This makes it extremely difficult for relevant departments to track suspect trajectories through monitoring systems. In current practical applications, the target pedestrians are matched via manual real-time monitoring, consuming a large amount of human resources and yielding a low information utilization rate. Furthermore, current methods are associated with unstable accuracies and low efficiency levels [2], [30], [35]. The bottlenecks of manual camera network monitoring methods are even more pronounced in big data environments [31], [32], [37]. Thus, in order to keep up with the rapid development of computer vision technology, a video analysis system that can simultaneously process multiple surveillance videos and accurately identify effective target pedestrians is urgently required [4].

Pedestrian re-identification refers to the similarity matching of pedestrian targets under the surveillance of multiple cameras without the overlapping of viewing angles. Its application in monitoring systems is a hot topic in the field

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

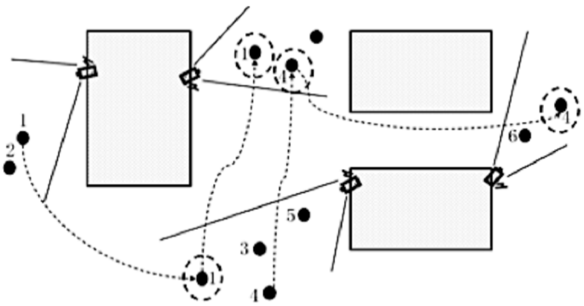


FIGURE 1. Pedestrian re-identification display.

of computer vision [5], [36]. As shown in Figure 1 [6], the target (the pedestrian, point 1), can be identified and locked in multiple cameras in different locations or in the same video (point 1, dotted circles). As the camera resolution in such systems is generally low, it is difficult to obtain identifiable features such as faces. Furthermore, the process is affected by additional factors (e.g. illumination and viewing angles), resulting in discrepancies within the same pedestrians in different cameras. It is often difficult to capture the invariant factors of the sample due to the extreme variability and high-dimensionality of the visual characteristics [7]. The simultaneous identification of multiple pedestrians proves to be a difficult task for traditional pedestrian re-identification methods. Such methods are also associated with long operation times and low recognition efficiency. Therefore, it is necessary to further improve the pedestrian re-identification method.

Deep learning has become a popular topic in the field of machine learning over the recent years. A convolutional neural network (CNN) is a representative and widely adaptable model in the field of deep learning. In the current paper, we propose a deep learning model suitable for pedestrian re-identification. In addition, a pedestrian re-identification system based on a deep convolutional neural network was designed, with outputs having significant social benefits for accident prevention and emergency treatment. Moreover, our proposed model can potentially generate economic benefits by saving human resources and improving the information usage, identification accuracy and work efficiency of camera monitoring systems.

More specifically, we propose a deep convolutional neural network structure that employs deep learning to solve the pedestrian re-identification problem in monitoring systems [38], [39]. The convolutional neural network extracts the feature vector of each pedestrian image, while the Euclidean distance between the feature vectors is then calculated to measure the similarity between pedestrian images. Following this, the loss function is determined, with the focal loss used to solve the sample imbalance, thus reducing the model image feature storage space, accelerating the image feature matching speed and improving the recognition accuracy [40]–[46]. Moreover, a surveillance system for pedestrian re-identification is implemented by a host computer and several high-speed network dome cameras to acquire images of the target. Images of the same person are then selected from similar image groups. The feature extraction and automatic learning of targets in complex environments are the core modules of the system, and can be applied to video surveillance and intelligence analysis in the live broadcast, education, film and television industries. In addition, our proposed model has key applications in human body recognition for complex environments such as machine vision human-computer interactions and automated robot tracking.

Recent studies have applied Siamese networks for re-identification [47], [48], demonstrating that novel deep CNN and integrated focal loss are the two key directions in the field of person re-identification. In the current paper, we explore these two fields to solve the problem in a unified framework. Our study has the following contributions: (1) An end-to-end pedestrian feature extraction architecture was proposed that learns hard examples using a specific fully connected layer; (2) the algorithm architecture was developed and implemented in our designed monitoring prototype system using digital matrix hardware and high-speed domes, which can be further extended to additional tasks for future applications; and (3) experiments on a public benchmark dataset were conducted and their competitive performance was observed to be similar to existing methods.

The rest of this paper is organized as follows. Related work is described in Section 2, and the framework of pedestrian feature extraction is explained in Section 3. The system design and implementation is introduced in detail in Section 4. The experimental results and analysis are presented in Section 5, while the conclusions and suggestions for future work are proposed in Section 6.

II. RELATED WORKS

Complications in pedestrian re-identification originate from multi-camera tracking problems. Early research on pedestrian re-identification was closely related to multi-camera tracking problems. Since pedestrian re-identification was originally employed for tracking pedestrians in videos, most studies focused on the matching of pedestrian images [8], [12]. For example, Huang and Rueesl *et al.* (1997) proposed the use of the Bayesian method to estimate the posterior probability of a target appearing under another camera by observing the appearance of the target under the field of view of a known camera [9]. Wojciech and Zajdel *et al.* (2005) first proposed the term "pedestrian re-identification" [10], [2], [33], [34]. Gheissari and Niloofar *et al.* (2006) followed this by investigating pedestrian re-identification as an independent visual problem and proposed the use of the space-time segmentation algorithm to detect the foreground area of pedestrians [8].

The complexity of pedestrian re-identification subsequently increased. Farenzena *et al.* (2010) extended pedestrian re-identification research from images to videos by proposing a feature extraction scheme. This was an important step for the application of pedestrian re-identification technology [11]. Following the success of convolutional neural networks in pattern recognition and classification, Ouyang Wanli *et al.* (2013) proposed joint deep learning for pedestrian re-identification, achieving promising results [13]. Anelia *et al.* (2015) combined the soft-cascade with CNN networks of different complexities to obtain a highly accurate and real-time pedestrian detection system [14]. Weihua Chen *et al.* (2016) proposed a multi-task deep network (MTDnet) and an across-domain architecture, with the employment of the auxiliary set to assist in the training of small target sets [26]. Song Bai *et al.* (2017) proposed an unconventional manifold-preserving algorithm that was able to take full advantage of the supervision of training data as well as demonstrating applications as a post-processing program for most existing algorithms, thus improving recognition accuracy [4].

After much research, pedestrian re-identification is currently in the application stage. With further studies focusing on deep learning, the performance stability of pedestrian re-identification has gradually improved for small datasets. It is an irresistible trend to apply deep learning to solve the problem of pedestrian re-identification and monitoring. Recently, a number of methods have applied Siamese networks for re-identification [47], [48]. In particular, Lin Wu *et al.* (2016) proposed PersonNet, integrating an adaptive root-mean-square gradient decent algorithm to seek robust features within images [47], [48]. In addition, Dangwei Li *et al.* (2017) used dilated conv to extract multi-scale features and reduce the information loss of traditional CNN feature extraction [25]. Furthermore, Yeong-Jun Cho *et al.* (2017) did pedestrian reidentification by estimating human posture and establishing multi-posture matching model [52]. As for feature selection, Luo Hao *et al.* (2019) proposed BagTricks in which global feature learning directly learns the representation of the whole image without local reduction [51].

III. PEDESTRIAN FEATURE EXTRACTION

The techniques used for searching in pedestrian re-identification generally include feature extraction and feature similarity measurements. Traditional feature extraction methods include color histograms, LBP, Gabor and Local Patch. Feature similarity measurements are often performed using M-distance, LFDA, MFA, etc. [6]. In the current paper, we propose a deep convolutional neural network structure that employs deep learning to solve the pedestrian re-identification problems within the monitoring system. The monitoring system acquires a target image and a set of similar images, with the subsequent task of matching different images of the same person from a similar group of images.

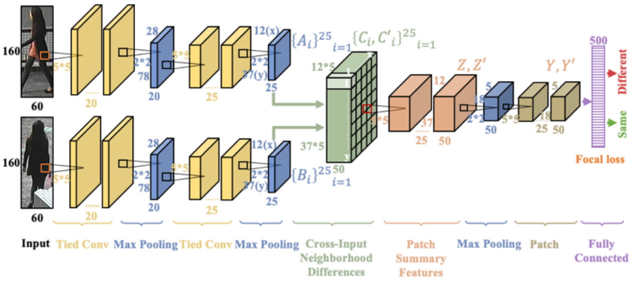


FIGURE 2. Visualization of the proposed neural network architecture.

A. NEURAL NETWORK ARCHITECTURE

The neural network architecture of the proposed model is presented in Figure 2. It is a six-component scalable network with the generalization ability to extract image features. Two-view 60*160*3 RGB images are input into the network where context information of the human body is extracted. First, the first two-layer-tied convolution structure with 20 5*5*3 filters is deployed to calculate the high-order features of the two input images. Note that the network weights are shared by the parameter sharing mechanism to ensure that two-view features are calculated with the same filters. The length and width of the feature maps are then halved using the max pooling layer. Next, the second tied convolution structure is employed to determine the feature maps with 25 5*5*20 filters, followed by the deployment of the second max pooling to obtain 25 feature maps with a size of 12*37.

In the second component, a cross-input neighborhood difference structure is implemented to compare the feature maps in the adjacent neighborhood positions. These 25 feature maps are recorded as a_i and b_i $(0 < i \leq 25)$. The final output feature maps are the differences of the features in the five neighborhoods around the feature values of the corresponding two feature maps in A_i and B_i. The neighborhood difference map C_i is generated from A_i and B_i with a grid size of 12*37 5*5, as follows:

$$C_i\,(x, y) = A_i\,(x, y)\,\alpha\,(5, 5) - \omega\,[B_i\,(x, y)] \qquad (1)$$

where $\alpha\,(5, 5) \in \mathbb{R}^{5\times5}$ is a 5×5 matrix and $\omega\,[B_i\,(x, y)] \in \mathbb{R}^{5\times5}$ is a neighborhood matrix centered on (x, y). C'_i is determined by exchanging A_i, B_i in Eq. (1). A total of 50 neighborhood difference maps are calculated, with the ReLu activation function used to obtain the final outputs.

The third component is a patch summary feature that uses C and 25 5*5*25 filters with a step size of 5 to convolve and sum each 5*5 matrix in C_i to obtain the overall difference. The output is recorded as Z. C_i denotes the same calculation excluding the weight sharing, with the output denoted as Z'.

The fourth component involves the learning of the neighborhood difference spatial relationships. In particular, Z is convolved with 25 3*3*25 filters and a step size of 1. The pool then halves the feature maps to 25 12*37. Following this, Y and Y′ are calculated using 25 5*18 feature maps by max pooling.

The fifth component of the network obtains high-order connections through the fully connected layer, combining the information of the matrix with a long distance. The information from Y and Y' is combined to generate a 500-dimensional vector. The ReLu function is first used to activate the features, which are then classified with the focal loss layer of two nodes.

B. NETWORK TRAINING BASED ON FOCAL LOSS

In the vast majority of existing pedestrian re-identification datasets, the number of potentially generated negative samples is much larger than the number of positive samples [15]. In this case, if all the samples are indiscriminately concentrated and used in the training of the network model (after scrambling the order) to minimize the risk of prediction errors, the trained network model will predict that “the two images do not belong to the same person”, regardless of the specific content of the input image. If the number of negative samples in the training sample is limited, only a limited amount of negative samples are retained from those generated. In addition, the information from the large set of negative samples that is helpful for the training of the network model will be seriously wasted [7]. This imbalance between positive and negative samples increases the difficulty in training network models due to the large number of differences across training samples.

In order to reduce the experimental error caused by this imbalance, our proposed model attempts to use focal loss to replace the traditional SoftMax loss calculation method. Focal loss is actually a normal cross-entropy loss plus a $(1-p_t)^\gamma$ factor that is used to reduce the loss function of high-quality categorical samples [15]. Intuitively, dynamic scaling can automatically reduce the contribution parameters of simple samples, such that more attention is focused on difficult samples and the impact of data imbalance can be solved.

In our model, we define focal loss as a binary classification test as follows. Let CE denote cross entropy, while y $\in \{-1,+1\}$ represents the real label of the sample, with -1 indicating a negative sample and $+1$ a positive. $p \in (0,1)$ represents the sample prediction value produced by the classifier. For $y = 1$, the value of p is large, and the confidence of the classifier predicting the sample as a positive example is higher. This indicates that the classifier performs well with a low value loss function. Similarly, when $y = -1$, if p is small ($1-p$ is large), the classification performance is strong. In particular, the cross-entropy loss is defined as follows:

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p), & y = 1 \\ -\log(1-p), & y \neq 1 \end{cases} \tag{2}$$

where p_t represents the degree of matching between the predicted value produced by the classifier and the true value of the sample, namely,

$$p_t = \begin{cases} p, & y = 1 \\ 1-p, & y \neq 1. \end{cases} \tag{3}$$

p_t is used to calculate the focal loss as follows:

$$\mathrm{FL}(p_t) = -(1-p_t)^\gamma \log(p_t). \tag{4}$$

For large p_t, the value of $(1-p_t)^\gamma$ is small ($\gamma > 0$), and the degree of matching between the predicted value and the true value is high. As an example, let us assume that the predicted value of a positive sample is 0.99, then the sample belongs to the “simple sample” in the training of the learner. That is, the learner is able to judge the true category of the sample well. Focal loss attenuates the loss value of this sample by $(0.001)^g$ times. Assuming that $g = 2$, the result will be attenuated 10,000 times, namely, the loss values of 10000 $p_t = 0.99$ and $p_t = 0.99$ will be equal before and after attenuation. This property reduces the influence of a large number of simple samples throughout the training process.

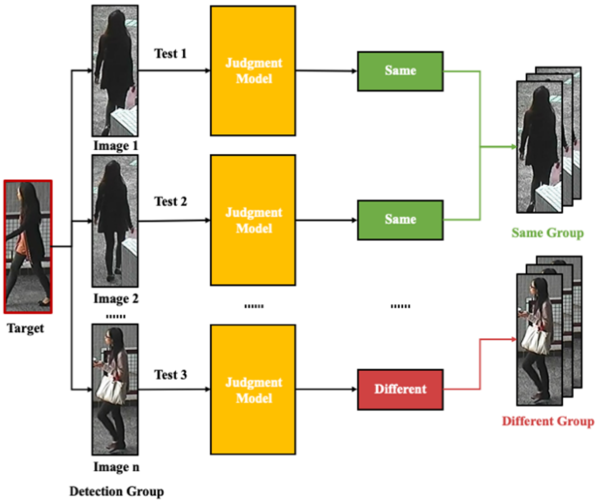


FIGURE 3. Visualization of the neural network architecture.

In the actual experiment, the model will also add weights to the positive and negative samples, further reducing the impact of the sample size on the experimental results. For example, for high negative sample frequencies, the weight of the negative samples is reduced, while if the number of positive samples is small, the weight of the positive samples increases. The formula is as follows:

$$\mathrm{FL}(p_t) = -\alpha (1-p_t)^\gamma \log(p_t) \tag{5}$$

Assuming a balanced sample size, for the focal loss-based network training, the loss function value of the “simple sample” model is small, and the influence of the reduction in the loss function value during the training process is also little. This method can effectively train the pedestrian re-identification model by increasing the weight of positive samples and reducing the weight of negative samples, greatly reducing the experimental error. Thus, the problem of sample imbalance in pedestrian re-identification is minimized.

Following the completion of the network model training, image traversal retrieval is executed, with the results presented in Figure 3. This process begins with the determination of a target and a detection group. The traversal method is then used to compare the images in the target and detection groups using the trained network model. Lastly, the results are divided into two categories, namely “same” and “different”. Hence, a similar image of the same pedestrian from different cameras is obtained, and the entire pedestrian re-identification is thus completed.

IV. SYSTEM DESIGN AND IMPLEMENTATION

The proposed pedestrian re-identification and monitoring system consists of five components: digital matrix, network switch, storage server, master computer and camera matrix.

Figure 4 depicts the monitoring system framework. The high-speed network dome camera and the network switch form the camera network. The master computer and storage server are connected by the network switch, for system control and dataset storage, respectively. The digital matrix is connected to the network switch in order to make the monitoring process visible in real time. The PC component is equipped the ‘‘Person re-identification-collection’’ and ‘‘Person re-identification-handle’’ security system software, whereby searching commences when the target person and dataset are provided.

Figure 5 presents the monitoring system identification process. The user is able to perform the video monitoring and data collection operations through the ‘‘Person re-identification-collection’’ step.

Note that in practice, the input of the system should be a complete surveillance video. An in-depth algorithm is involved here to transform the video collected by the ‘‘Person re-identification-collection’’ into a database that can be processed by the ‘‘Person re-identification-handle’’. This means that the system will automatically extract a single image of a human from the video and set a unique label for each image. By doing so, it can determine which position of the recognized human body is from the original video. Due to the space, this will be investigated further in future research.

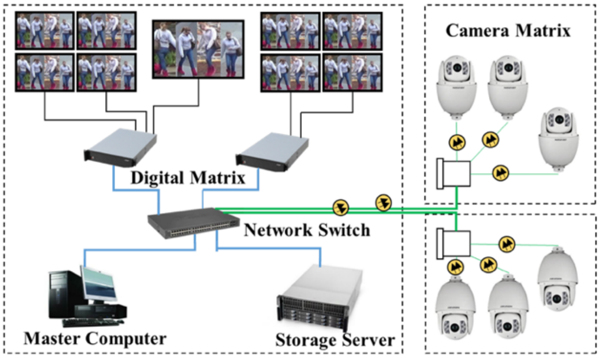


FIGURE 4. Monitoring system framework.

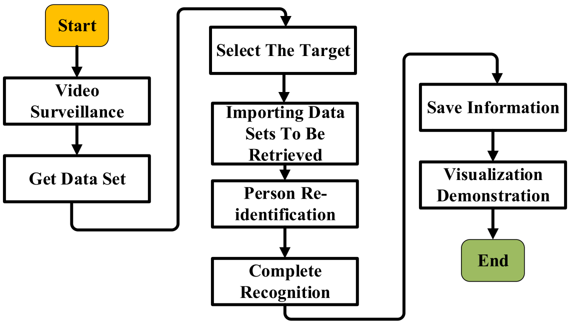


FIGURE 5. Monitoring system re-identification flow chart.

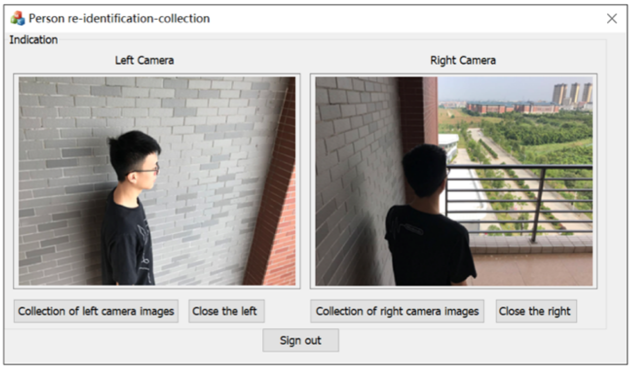

Person re-identification-collection
Indication
Left Camera
Right Camera
Collection of left camera images
Close the left
Collection of right camera images
Close the right
Sign out

FIGURE 6. Person re-identification- collection.

For our experiments, we accomplished the data segmentation by default, obtaining the dataset of single body images.

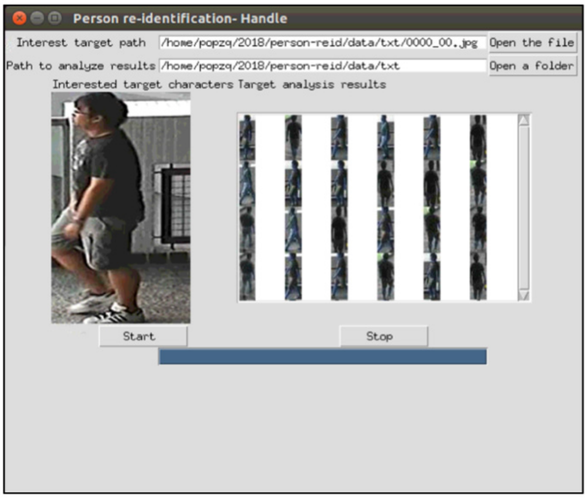
The left of Figure 6 is the pedestrian image captured by the left camera, while the right is the same pedestrian captured by the right camera at different angles in the same position. Note that the image patches in our pedestrian re-identification experiment and monitoring system were obtained from public datasets.

During the ‘‘Person re-identification-handle’’ step, the path of the target person is selected for retrieval using the ‘‘Open the file’’ and ‘‘Open a folder’’ buttons above. The imported dataset is then retrieved to start the pedestrian re-identification search. Once the searching ends, the information is stored and the search result can be visually displayed. Figure 7 presents an example search result, with the left showing the target character, and the right the result images with high similarity to the target.

V. EXPERIMENTAL VERIFICATION

A. TRAINING NETWORK

In the experimental phase, network model training was first performed on the host computer. The main control computer hardware had the following configurations: A NVIDIA GTX


Interest target path /home/popzq/2018/person-reid/data/txt/0000_00.jpg

Path to analyze results /home/popzq/2018/person-reid/data/txt

Person re-identification- Handle
Open the file
Open a folder
Interested target characters
Target analysis results
Start
Stop

FIGURE 7. Person re-identification- Handle.

1080ti (1) GPU; a graphics card single-precision floating-point performance of approximately 10.7 TFlops; a 11 G memory; 3584 stream processor cores; a i7 920 CPU with 2.66 GHz and quad-core eight threads; and a RAM of 18 G. The network parameter settings are reported in Table 1:

TABLE 1. Network parameter settings.

Parameter	Numerical value
Epoch	200000
Learning Rate	0.01
Batch Size	150
Gamma	2.0
Alpha	8

The initial learning rate during training was set to 0.01, a network model was generated and saved every 10,000 iterations, and the training process of the entire network was iterated 200,000 times. Gamma in Table 1 denotes the focal loss hyperparameter used to control the ‘‘simple sample’’ loss rate. Moreover, Alpha is the weight applied to balance the positive and negative samples in order to reduce the influence of sample size on the model.

B. DATASET AND EVALUATION FRAMEWORK

We employed the CUHK03-Labeled public dataset to train our proposed network. As shown in Figure 8, the CUHK03-Labeled dataset had a total of 1316 for 1360 pedestrians. The images of the dataset came from 6 (3 pairs) different cameras on the campus of the Chinese University of Hong Kong, and each pedestrian appears in just one of the cameras. Since the images were taken from real scenes, images of the same pedestrian under different cameras exhibited large variations in posture and mutual occlusion, amongst other issues.

As pedestrian re-identification is a similarity ordering problem, we adopted the Cumulative Matching Characteristic (CMC) curve to evaluate the experimental results [50]. The images corresponding to 1260 and 1360 pedestrian IDs were randomly selected as the training set. The images corresponding to the remaining 100 pedestrian IDs were used as the test sets, and the first camera image under each pair of cameras was used as the target image. In the subtest process, an image of the second camera was randomly selected for each pedestrian ID in the test set, thereby forming an image library in which the ID of each image is known. The cumulative matching characteristic of each retrieved image in the image library was calculated and averaged to form a cumulative matching characteristic curve for the sub-process. The random sampling was repeated 100 times. The average of the 100 cumulative matching characteristic curves was then taken, resulting in the cumulative matching characteristic curve of the entire testing process.

FIGURE 8. Subset of the dataset used in the experiment.

C. DATASET AND EVALUATION FRAMEWORK

The proposed network model was tested using the single-shot setting method with the CUHK03-Labeled dataset. The Rank-1, Rank-10 and Rank-20 accuracies were recorded and compared with current methods. The results are reported in Table 2:

TABLE 2. Rank scores of different methods using the CUHK03 dataset.

Method	CUHK03（p=100）		
	Rank-1	Rank-10	Rank-20
FPNN[17]	20.65	66.50	80.00
IDLA[7]	54.74	93.88	98.10
LOMO+XQDA[18]	52.20	92.14	96.25
LOMO+MLAPG[19]	57.96	94.74	98.00
SSSL[20]	57.00	-	-
EDM[21]	61.32	-	-
Ensembles[22]	62.10	94.30	97.80
LDNS[23]	62.55	90.05	98.10
PersonNet[49]	64.8	-	-
GOG+XQDA[24]	67.3	96.0	-
MSCAN[25]	74.21	**97.54**	99.25
Latent Parts [50]	74.21	-	-
MTDNet[26]	74.68	97.47	-
DNNIM[27]	72.43	95.51	98.40
SSM[17]	**76.63**	-	97.95
Our	76.0	95.6	**99.5**

“-” denotes no corresponding data in the open literature.

Table 2 gives a comparison of the CUHK03 data sets. Our network is up to 76 percent Rank-1; In contrast to PersonNet, which is also improved on IDAL, it tries to replace the shallow convolution layer with multiple 3 × 3 convolutional layers, which deeptens the number of network layers and makes the effect significantly improved. However, it is difficult to train the model due to data imbalance, and the deeper network layers will aggravate this phenomenon. In Rank-1, the score of our model is about 11% higher than that of PersonNet. Therefore, compared with deepening the number of network layers, how to reduce the experimental error caused by negative samples is more effective to improve the accuracy of the model. Furthermore, SSM surpassed all of the other pedestrian re-identification methods in terms of judging the similarity of two pedestrian images by comparing their

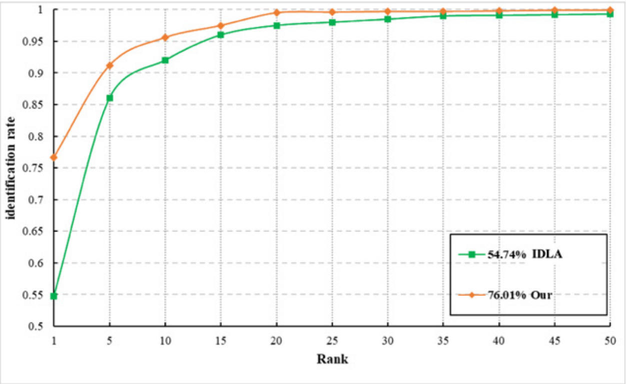


FIGURE 9. Comparison of traditional loss function and our proposed method.

feature maps. The SSM method uses the CNN feature as the input of the algorithm, while the supervised embedded manifold is employed to reduce the dimensionality of the pedestrian feature vector. At the same time, it guaranteed the constraint on intra-class distance and inter-class distance in dimension reduction, which belonged to subspace learning [27]. Different from SSM, the network model proposed in this paper uses a deep convolutional neural network to learn features and corresponding similarity metrics simultaneously, introducing a cross-input neighborhood difference layer. The local correlation was calculated according to the convolutional feature map of the image pair. The neighborhood of the output feature map was added using the additive feature, and the correlation of the distant pixel points was then calculated. The traditional SoftMax Loss function calculation was replaced with the focal loss, allowing for the data imbalance of pedestrian recognition to be efficiently solved.

Figure 9 is the comparison between our model and IDLA's scores on CMC. It is not difficult to find that we use the Focal Loss, to reduce the Loss function quality classification samples, the contribution of narrowing the simple sample parameters automatically, so as to faster the attention of the difficult samples, solve the unbalanced data, can effectively improve the scoring model at Rank-1, which means that the model can be faster from the pedestrian database retrieval to the pedestrian targets accurately, but also illustrate the Focal Loss contrast than traditional SoftMax Loss is more suitable for the pedestrian recognition model.

Due to practical constraints such as funding, privacy, etc., our experiments were conducted on the open source CUHK03 dataset. The lack of further verification using real data may lead to a sharp decline in the identification accuracy of the model. In a real environment, pedestrians may be affected by factors such as illumination, angle of view and attitude. This leads to larger visual performance differences of the same pedestrian from different cameras. To solve this problem, existing datasets can be generated from different perspectives via human posture estimation and antagonistic network combinations. This will standardize the posture of pedestrians in real data, thus helping the model to better acquire pedestrian characteristics and improve its recognition accuracy for real data.

VI. CONCLUSION

Despite the gradual improvement in performance stability for small pedestrian re-identification datasets, massive video datasets and human analysis prove to be severe bottlenecks. The application of deep learning to solve pedestrian recognition has become a current trend. In this paper, a convolutional neural network was employed to extract the feature vector of each pedestrian image. The Euclidean distance between the feature vectors was then calculated to measure the similarity between pedestrian images, followed by the determination of the loss function. The focal loss was used to solve the sample imbalance phenomenon, reduce the model image feature storage space, speed up the image feature matching and improve the recognition accuracy. In addition, we implemented an end-to-end monitoring system for pedestrian re-identification based on the deep convolution neural network and focal loss. The real-time, simple and rapid characteristics of the proposed method are in line with practical application requirements, which will be expanded to other tasks in the future.

REFERENCES

[1] S. Ma, X. Chen, Z. Li, and Y. Yang, "A retrieval optimized surveillance video storage system for campus application scenarios," *J. Electr. Comput. Eng.*, vol. 2018, Aug. 2018, Art. no. 3839104, doi: 10.1155/2018/3839104.

[2] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, Apr. 2014.

[3] P. H. Tu, G. Doretto, and N. O. Krahnstoever, "An intelligent video framework for homeland Protection," *Proc. SPIE Int. Soc. Opt. Photon.*, vol. 6562, May 2007, Art. no. 65620C.

[4] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2530–2539.

[5] S. Huo, T. Hu, and C. Li, "Improved collaborative representation classifier based on l2-regularized for human action recognition," *J. Electr. Comput. Eng.*, vol. 2017, Jan. 2017, Art. no. 8191537, doi: 10.1155/2017/8191537.

[6] S. Gong, M. Cristani, and S. Yan, *Person Re-Identification*. London, U.K.: Springer, 2014.

[7] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[8] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1528–1535.

[9] T. Huang and S. Russell, "Object identification in a Bayesian context," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 97, 1997, pp. 1276–1282.

[10] W. Zajdel, Z. Zivkovic, and B. J. A. Krose, "Keeping track of humans: Have i seen this person before?" in *Proc. IEEE Int. Conf. Robot. Autom.*, Jun. 2005, pp. 2081–2086.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.

[12] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

[13] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063.

[14] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proc. Brit. Mach. Vis. Conf.*, 2015, p. 32.

[15] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.
[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*. [Online]. Available: http://arxiv.org/abs/1708.02002
[17] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[19] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3685–3693.
[20] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1278–1287.
[21] H. Shi, Y. Yang, and X. Zhu, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 732–748.
[22] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1846–1855.
[23] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.
[24] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1363–1372.
[25] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 384–393.
[26] W. Chen, X. Chen, and J. Zhang, "A multi-task deep network for person re-identification," In *Proc. Assoc. Adv. Artif. Intell.*, 2016, pp. 3988–3994.
[27] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2667–2675.
[28] Y. Sun, L. Zhu, G. Wang, and F. Zhao, "Multi-input convolutional neural network for flower grading," *J. Electr. Comput. Eng.*, vol. 2017, Jun. 2017, Art. no. 9240407, doi: 10.1155/2017/9240407.
[29] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1992–2000.
[30] A. Barman and S. K. Shah, "SHaPE: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1124–1133.
[31] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3239–3248.
[32] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3200–3208.
[33] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[34] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2528–2535.
[35] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[36] K. Han, W. Wan, G. Chen, and L. Hou, "Distance aggregation for person re-identification using simulated annealing algorithm," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2016, pp. 399–403.
[37] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person re-identification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, p. 13, 2018.
[38] S. Tan, F. Zheng, L. Liu, J. Han, and L. Shao, "Dense invariant feature-based support vector ranking for cross-camera person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 356–363, Feb. 2018.
[39] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.
[40] X. Zhao, N. Wang, Y. Zhang, S. Du, Y. Gao, and J. Sun, "Beyond pairwise matching: Person reidentification via high-order relevance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3701–3714, Aug. 2018.
[41] X. Li, "Multishot person reidentification using joint group sparse representation," *J. Electron. Imag.*, vol. 27, no. 6, Dec. 2018. Art. no. 063012.
[42] S. Zhou, J. Wang, and R. Shi, "Large margin learning in set-to-set similarity comparison for person re-identification," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 593–604, Dec. 2018.
[43] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1470–1478.
[44] H. Shi, S. Liao, and D. Yi, "Learning deep metrics for person re-identification," in *Deep Learning Biometrics*. Boca Raton, FL, USA: CRC Press, Mar. 2018, pp. 109–125.
[45] H. Li, M. Yang, Z. Lai, W. Zheng, and Z. Yu, "Pedestrian re-identification based on tree branch network with local and global learning," 2019, *arXiv:1904.00355*. [Online]. Available: http://arxiv.org/abs/1904.00355
[46] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.
[47] L. Wu, C. Shen, and A. van den Hengel, "PersonNet: Person re-identification with deep convolutional neural networks," 2016, *arXiv:1601.07255*. [Online]. Available: http://arxiv.org/abs/1601.07255
[48] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, "Person re-identification with triplet focal loss," *IEEE Access*, vol. 6, pp. 78092–78099, 2018.
[49] Z. Liu and C. Wang, "Design of traffic emergency response system based on Internet of Things and data mining in emergencies," *IEEE Access*, vol. 7, pp. 113950–113962, 2019, doi: 10.1109/ACCESS.2019.2934979.
[50] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," 2017, *arXiv:1703.07220*. [Online]. Available: http://arxiv.org/abs/1703.07220
[51] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," 2019, *arXiv:1906.08332*. [Online]. Available: http://arxiv.org/abs/1906.08332
[52] Y.-J. Cho and K.-J. Yoon, "Distance-based camera network topology inference for person re-identification," *Pattern Recognit. Lett.*, vol. 125, pp. 220–227, Jul. 2019, doi: 10.1016/j.patrec.2019.04.009.

WENZHENG QU was born in Shandong, China, in 1997. He received the bachelor's degree from the Xinhua College of Sun Yat-sen University, in 2019. He was with Tencent Youtu Lab, Shenzhen, as an AI Engineer. He has published five articles. His research interests include computer vision and deep learning.

ZHIMING XU was born in Guangdong, China, in 1996. He received the bachelor's degree from the Xinhua College of Sun Yat-sen University, in 2019. He was with the School of Information Science, Xinhua College of Sun Yat-sen University, as a Lab Manager. He holds over three patents and two inventions. His research interests include embedded systems and wireless sensor networks.

BEI LUO was born in Guangdong, China, in 1996. He received the bachelor's degree from the Xinhua College of Sun Yat-sen University, in 2019. His research interest includes embedded systems.

HAIHUA FENG was born in Guangdong, China, in 1996. He received the bachelor's degree from the Xinhua College of Sun Yat-sen University, in 2019.His research interest includes automatic systems.

ZHIPING WAN was born in Hubei, China, in 1980. He received the bachelor's degree from the Huanggang Normal College, in 2003, and the master's degree from the Guangdong University of Technology, in 2008. He was a Visiting Scholar with Sun Yat-sen University. He was a Teacher with the Huali College Guangdong University of Technology, from 2004 to 2008. Since 2010, he has been with the School of Information Science, Xinhua College, Sun Yat-sen University, and an Associate Professor, in 2018. He has published 60 articles. He holds over two patents and one invention. His research interests include wireless sensor networks, cognitive radio networks, and network security.