Multimedia Tools and Applications (2022) 81:39169–39184
https://doi.org/10.1007/s11042-022-13170-x

A feature extraction method for person re-identification based on a two-branch CNN

Bo Yang[1] · Yao Shan[1] · Rui Peng[1] · Jian Li[1] · Shaohui Chen[2] · Linlin Li[3]

Received: 23 December 2020 / Revised: 15 April 2021 / Accepted: 11 April 2022 /
Published online: 28 April 2022

Abstract
A two-branch convolutional neural network (CNN) architecture for feature extraction in person re-identification (re-ID) based on video surveillance is proposed. Highly discriminative person features are obtained by extracting both global and local features. Moreover, an adaptive triplet loss function based on the original triplet loss function is proposed and is used in the network training process, resulting in a significantly improved learning efficiency. The experimental results on open datasets demonstrate the effectiveness of the proposed method.

Keywords Person re-identification · Two-branch convolutional network · Triplet loss function

✉ Linlin Li
linlinliits@163.com

Bo Yang
13910700045@139.com

Yao Shan
ShanYao@ncist.edu.cn

Rui Peng
pengrui232001@163.com

Jian Li
641985217@qq.com

Shaohui Chen
chensh01@ehualu.com

[1] School of Emergency Technology and Management, North China Institute of Science & Technology, No. 467 academy Street, Sanhe Yanjiao Development Zone, 065201 Langfang, China

[2] Beijing Gaocheng Technology Development Co. LTD, Beijing 100043, China

[3] Transport planning and Research Institute Ministry of Transport, Building 2, Time International Building, No.a6, Shuguang Xili, Chaoyang District, 100028 Beijing, China

1 Introduction

Recent years have witnessed a rapid development of the global economy and continuous progress of society, and thus, higher requirements have been put forward in various fields for video surveillance systems, the widespread use of which has brought much convenience for people's lives and has also made important contributions to public safety [13].

The targets with the highest frequency of occurrence in surveillance video are persons. Person re-identification (re-ID) refers to the analysis and judgement of the trajectory and range of motions of the targets of interest (such as criminal suspects and terrorists) by retrieving the obtained images of these targets in videos captured by other surveillance cameras in a timely manner to provide technical support and decision-making assistance to the government or the security sector. Common person re-ID tasks are generally composed of a query library and a search library. The query library contains the targets of interest. The search library consists of person images, which are generally obtained from videos using target detection algorithms. Specifically, the image in the query library is compared with each image in the search library, and the person image with the highest similarity is returned as the final recognition result. In the era of video data explosion, it is impossible to meet practical needs by relying only on manual data processing methods. Therefore, it is of important theoretical and practical significance to research and develop the corresponding person re-ID technology.

As early as in 1996, person re-ID began to draw the attention of researchers [2]. In 2006, Gheissari et al. [6] proposed for the first time the concept of person re-ID in an academic conference, followed by a surge of relevant studies. The VIPeR dataset is the first dataset specifically designed for person re-ID research [8] and greatly promoted the development of person re-ID field. The first related monograph authored by Gong et al. [7] was published in 2013, which discusses in detail the cutting-edge technologies and major challenges in the field of person re-ID. In recent years, researchers have begun to use deep learning to solve problems related to the person re-ID and have made great breakthroughs [11, 23, 31, 33]. Many research outcomes have been published in major computer vision conferences and journals, and the recognition results on multiple datasets have also been significantly improved.

As the topic of person re-ID has attracted increasing attention, many solutions have been proposed, and good results have been achieved on different test datasets. However, person re-ID still faces a number of challenges stemming from, for example, illumination changes, different person postures, varying viewing angles, non-aligned person images, complex image background, and scale changes.

Therefore, person re-ID in real situations is still one of hot topics in surveillance video research. In this study, focusing on the challenges including illumination, different person postures, varying viewing angles, non-aligned person images, complex image background and scale changes, we propose a two-branch convolutional neural network (CNN) architecture to extract global and local features of persons from surveillance videos to obtain person features with strong discriminative ability. Moreover, to address the problems in the network learning process, we propose an adaptive triplet loss function based on the conventional triplet loss function to greatly improve the learning efficiency. The experimental results on multiple open datasets verify the proposed method is effective for person re-ID from practical surveillance videos.

2 Related works

One of the key issues in the person re-ID method based on feature extraction is to design a robust, reliable representation of person features, which should be able to not only identify different persons but also overcome the influence of complex environment on the identification. Current research methods in this category are mainly divided into manual feature extraction and deep learning-based feature extraction.

2.1 Manual feature extraction methods

In early research on person re-ID, shallow visual features, including color, shape, and trajectory features, were mainly used to solve the person re-ID problem. In 2012, Farenzena et al. proposed the symmetry-driven accumulation of local features (SDALF) [4]. According to the physiological structure of the human body, they divided a human into different parts, from which various color histogram features based on HSV color space were extracted and then combined into a whole for person matching. Later, Mignon et al. horizontally divided a person image into several blocks [22], extracted for each block the color features based on the YUV, HSV, and RGB color space in addition to the LBP texture features, and finally fused these different features into a whole to describe a specific person.

With the deepening of research, it was found that the use of shallow visual features alone was unable to well solve the person re-ID problem. Shallow visual features can partially represent the exterior intrinsic attributes of a person but are poorly adaptable to viewing angles and illumination. Therefore, researchers began to explore more effective representations of visual features, mainly including the semantic attributes and advanced visual features of a person.

In ECCV 2014, Yang et al. proposed the salient color names-based color descriptor (SCNCD) [29]. They argued that the clothing color of a person is very crucial for recognition; after extracting a variety of basic colors from a person, they extracted the corresponding color histograms from different areas of the person image and fused these color histograms as the final feature description. Liao et al. proposed in 2015 a feature descriptor named LOMO (local maximal occurrence) [15], described as follows. First, a person image is divided into six horizontal long stripes, and then a window of a certain size is used to move across each horizontal stripe to extract the HSV color histogram and SILTP histogram; the feature with the maximum value among these features is taken as the feature of a horizontal stripe, and finally the features of all horizontal stripes are combined as the LOMO feature descriptor. The LOMO feature is strongly invariant with respect to angle and illumination, making it widely applicable and often used for comparison with other algorithms. Recently, Matsukawa et al. proposed a feature descriptor named GOG (Gaussian of Gaussian) [21], in which the local area of an image is modeled by Gaussian distribution to simulate the appearance information of different local areas; very good results were obtained on multiple different datasets. In addition, some researchers also studied the use of semantic information and advanced visual features to represent a person. Because of the stability of the two, good recognition results can be achieved even if the posture of a person has changed considerably.

Although the abovementioned manual feature extraction methods lead to good results, they are mostly designed for certain specific situations and unable to achieve satisfactory test results in other scenarios. In other words, manually extracted features have weak robustness and poor universality. Moreover, it is difficult to define the validity or applicable situation of each

feature. In addition, most of the features are obtained using the multifeature fusion method, for which the fusion strategy of these features cannot ensure that the fused feature is optimal. Therefore, it is especially important to design more effective feature fusion strategies.

2.2 Feature extraction methods based on deep learning

The person feature extraction methods based on deep learning mainly use CNNs to extract person features. Compared to those extracted via the traditional manual feature extraction methods, the features extracted by the CNN model are relatively expressive, and thus the performance of recognition algorithms established with the CNN model will be substantially improved.

A common practice is to use the loss function as a constraint to train the parameters of the model to achieve the goal of "small intraclass distance and large interclass distance." In 2016, Geng et al. proposed combining the classification loss and verification loss to train a network [5]. The main network has a two-stream CNN architecture, which is connected to the classification subnet and verification subnet. The classification subnet is used to predict the identity of the image, and the classification error loss is calculated based on the prediction results. The verification subnet fuses the features of the two images to determine whether the two images belong to the same person. During the test, the trained network is directly used to extract person features for re-ID. In fact, there are many significant researches on the study of multi-branch neural network. Focusing on object segmentation from video, for example, the paper [18] proposed a novel network, called CO-attention Siamse Network to adderss the unsupervised object segmentation. Based on graph theory, Lu et.al. described novel and effective graph networks, and realized object segmentation perfectly [19, 27]. In 2017, Lin et al. noted that person identity information alone is not sufficient to learn a model with a high generalization ability. Therefore, they introduced person attribute labels through the labeled attribute information of a person so that the model needs to predict not only the person identity but also each person attribute correctly; the combined constraints of multiple features not only enhance the generalization ability of the model but also effectively improve its recognition results [16].

The rapid development of deep learning has promoted significant improvement in the person re-ID performance. At present, research on target feature extraction based on deep networks has the following deficiencies. First, the training datasets for person re-ID are generally small, and thus the trained network model tends to be overfitted, leading to insufficient generalization ability of person re-ID in real surveillance scenarios. Second, the deep features extracted based on deep learning networks are unable to effectively distinguish fine-grained target recognition; therefore, it is necessary to construct a new type of network to extract more essential target features.

3 Person re-ID based on CNN feature extraction

Due to the complexity and diversity of real surveillance scenarios, the effectiveness of traditional person re-ID methods based on manual features is far from satisfactory. Therefore, an increasing number of studies have focused on personal re-ID based on CNN. In 1989, LeCun et al. proposed for the first time a network capable of multilayer training named LeNet

network [12], which was subsequently thoroughly studied by many researchers. ResNet is a more frequently used CNN for person re-ID.

3.1 Residual neural network (ResNet) architecture

Researchers have already realized that as the number of network layers is increased, the network can be expressed more effectively. Therefore, theoretically better results can be obtained with a deeper model. However, experimental findings reveal that a deep network is prone to the degradation problem, that is, as the network depth increases, the accuracy of the network becomes saturated or even decreases. This is because when the depth of CNN exceeds a certain number of layers, it will be very difficult to train the CNN due to the problems of gradient disappearance and gradient explosion, thus affecting the final recognition accuracy.

ResNet was proposed by He et al. of Microsoft Research in 2016 [9]. By using the residual unit, ResNet successfully trained a CNN with a depth of 152 layers. ResNet has a small number of parameters, and the residual unit designed by it can very quickly accelerate the training of neural networks. Moreover, the accuracy of the model is also greatly improved. ResNet learns the difference between the input and output of the network, i.e., the residual $H(x) - x$. Such a residual hopping structure enables some information at the front end of the network to be directly transmitted to the back end of the network without the need for calculation in the middle layer. Therefore, the problem of gradient disappearance can be avoided during training so that the network can be trained very deeply.

ResNet is referenced to the VGG19 network, and on this basis, downsampling is performed using convolution with a step size of 2, the full connection layer is replaced by the pooling layer, and the residual units are added through a short-circuit mechanism. These operations not only significantly reduce the number of parameters in ResNet but also improve the expressiveness of the network. Commonly used residual networks include ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152.

ReseNet50 is used as the backbone network in most of the deep learning-based person re-ID algorithms. To facilitate comparison with relevant algorithms, we also used ResNet50 as a backbone network to carry out relevant research.

3.2 Person feature extraction based on a two-branch CNN

Good results were achieved via the use of CNNs for global person feature extraction during the early stage of person re-ID. With the development of research, it was found that use of global features only is not sufficient for person re-ID. Global features are effective for identifying persons with large differences in shape and color and poor otherwise. Therefore, researchers gradually paid attention to person re-ID based on local features. Attempts were made to extract local features of a person and then combine them with global features for person re-ID. Such a design can often achieve better results for person re-ID compared with use of only global features.

Currently, there are many methods for person re-ID based on combinations of global and local features. These methods focus on how to extract effective local features. Typical algorithms include Spindle Net [32], PDC [25], PL-Net [30], and GLAD [28]. Some of these algorithms directly divide a person image horizontally, extract the features separately from the divided image blocks, and finally combine them as a local feature. Using the key points of the human body for estimation, some other algorithms first obtain the key points of persons in the

image. Then, the image is divided into blocks for different parts of the human body based on the key points, features of different blocks of the human body image are extracted, and the features of these different parts of the body are combined to yield the final representation of the features. There are also algorithms that, according to the physiological structure of the human body and combining with the algorithm for estimation of human body key points, divide the person image into horizontal blocks for different parts, separately extract features from the divided image blocks, and finally fuse them into a local feature.

The above methods for the extraction of local features from a person image are simple, straightforward, easy to understand, and consistent with the human recognition process, thus also achieving good recognition results. However, it can be observed that these methods rely on the algorithm for estimation of human body key points, which is very time consuming and hence very inefficient for both the training process and practical applications. Moreover, in the process of feature extraction using the above algorithms, the local features and global features do not constrain and promote each other, and hence the learning efficiency must be improved.

Based on the above considerations, a CNN that can simultaneously extract the global and local features of a person is designed in this section. ResNet50 is used as the backbone network, and the first three layers of the network are used to extract the basic features of the image. Two branches are designed on the high-level semantic-level features to extract the global features and local features, respectively, of a person. The two parts work both collaboratively and separately, with weight sharing for the first three layers and independent weights for the subsequent advanced layers. In this manner, like the principle of human cognition of things, not only the overall information of a person can be seen but also the local information of different scales can be taken into account. Details of the network architecture are shown in Fig. 1.

For building the positive and negative samples. The positive samples include that the Re-ID persons with different person postures, illumination changes, varying viewing angles, non-aligned person images, complex background, and scale changes. The negative samples include that the person (different the Re-ID person) images with the different cases as same as positive samples.

In Fig. 2, GMP stands for the global maximal pooling used for feature dimension reduction; 1 × 1 conv represents connection with a convolutional layer of size 1 × 1; $L_{triplet}$ denotes a triplet loss function; and $L_{softmax}$ is a cross entropy loss function.

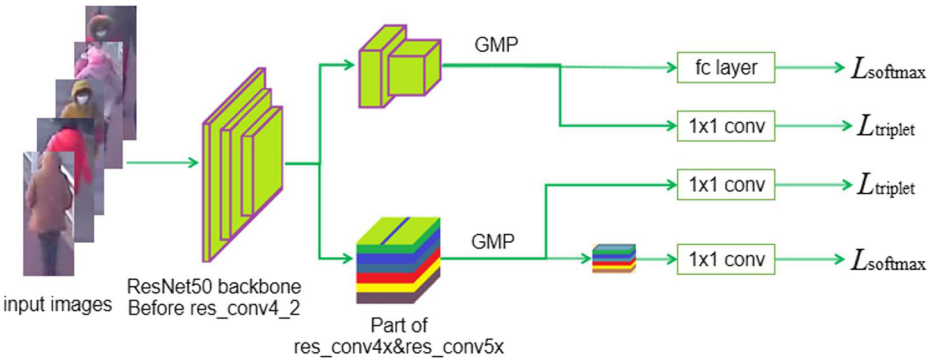

Fig. 1 Schematic diagram of the feature extraction network architecture

Fig. 2 The heat-map of features

The overall feature is the overall description of the pedestrian. In this paper, the pedestrian overall feature extraction method first goes through Resnet50, then goes through Global Max pooling (GMP) to reduce the dimension of the feature, and finally enters two channels. One channel is operated through a 1 × 1 CONV and trained with *Ltriplet*(Triplet Loss). The other channel is to directly enter the features extracted from Resnet50 into the full link layer. The function used in this channel is cross entropy loss function (*Lsoftmax*). In this way, the overall feature extraction of two dimensions is realized.

The local feature map extracted from the first three layers of RESNET 50 is divided into several horizontal strips. Each strip passes through the fourth and fifth layers of RESNET 50, and the output results are reduced by GMP feature dimension, and then enter two channels. In the first channel, 1 × 1 Conv operation is performed, and use triplet loss function to trained the local features. The second channel converges each horizontal strips, then 1 × 1 Conv operation is performed, and uses cross entropy loss as the training function to obtain the local features. After the above operation, two global features and two local features are obtained. The approach for horizontal division of the feature map simulates the division of the human body structure. Each stripe area of the feature map represents a different part of the human body. The full learning of these stripe areas is actually learning of the individualized areas of different parts of the human body, which is conducive to the learning of a discriminative feature of a person. Moreover, it can be seen that this design is in line with the human approach of judging a person. This is because in real life, it is often only necessary to determine who a person is based on parts of him or her rather than the appearance information. This recognition strategy of inferring the whole from the local is simple and effective, and the local branches in the designed network play exactly such a role. From another perspective, if the information based on each small piece can be recognized correctly, then the accuracy of recognition by combining all small pieces of information will naturally increase significantly.

3.3 Design of network loss functions

The loss function of a network is used to evaluate the degree of difference between the predicted value of the model and the real value. It is often used as the objective function of a neural network and represents the optimization direction of an algorithm. In person re-ID, commonly used loss functions include the cross entropy loss function, comparison loss function [26], triplet loss function [24], and optimized triplet loss function [10]. Besides these conventional loss function, there are many novel loss functions were proposed. For the task of

regression tracking, to balance training data, a shrinkage loss was suggested to penalize the importance of easy training data, its dramatic effectiveness was demonstrated by several benchmark datasets [17, 20].

The cross entropy loss function describes the distance between two probability distributions, and the smaller the cross entropy, the closer the two are. This loss function can lead to relatively good results for coarse-grained target classes. For fine-grained recognition tasks such as person re-ID, it can neither guarantee that the intraclass distance of samples is small enough nor that the interclass distance of samples is large enough. Therefore, the learning outcomes are very limited. Consequently, it is necessary to include the constraints on the spatial distribution of samples in the design of the loss function. Only by including the spatial constraints can the same class of samples be clustered in the feature space while different classes of samples be far apart from each other, which is conducive to the subsequent recognition. The triplet loss function based on the distance constraint between positive and negative samples is often used.

The triplet loss function [24] is a commonly used loss function for tasks such as retrieval and fine-grained recognition. Unlike the cross entropy loss function introduced above, the triplet loss function constraints the pairwise distance between samples; through constant iterative learning, it can decrease the intraclass distance and increase the interclass distance of samples in the feature space, thereby distinguishing different classes of persons.

The triplet loss function can decrease the distance between pairs of positive samples and increase the distance between pairs of negative samples. Finally, the person images of the same label form clusters in the feature space to distinguish different persons. However, we also observed that the triplet loss randomly sampled three images from the training data. Although this approach was simple, most of the samples were sample pairs that were simple and easy to distinguish. Thus, there were problems such as low training efficiency and nonideal convergence results. Therefore, it has been found that use of harder samples to train the network can make full use of the complex distribution characteristics of the training data, thereby improving the generalization ability of the network. Therefore, there appears a triplet loss function with batch hard mining.

The triplet loss function with batch hard mining (TriHard Loss function) is an optimization of the above basic triplet loss function [10]. By optimizing the network input triplets, each triplet involved in each training is optimal, thus improving the training efficiency and convergence speed of the network.

The core idea of TriHard Loss is as follows. To form a training batch, P persons with IDs are randomly selected, and K different images are randomly selected for each person, that is, one batch contains $P \times K$ images. Then, for every image a in the batch, we can select the hardest positive sample and the hardest negative sample, which together with a form a triplet. First, the image set with the same ID as a is defined as A, and the remaining sets (with different ID) are B. Then, the TriHard Loss is expressed as

$$L_{trihard} = \frac{1}{P \times K} \sum_{a \in batch} \left(\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + m \right)_+ \tag{1}$$

where m is the manually set interval between pairs of positive and negative samples, $d_{a,\,p}$ represents the distance between the sample a and its positive samples, $d_{a,\,n}$ represents the distance between the sample a and its negative samples, and $()_+$ represents the greater value of the calculation result and zero. TriHard Loss fully utilizes the idea of batch hard mining in the

loss calculation. Specifically, it calculates the distance between a and each image in feature space and then selects the positive sample p with the longest distance from a and the negative sample n with the smallest distance from a to calculate the loss. Thus, TriHard Loss usually leads to better results than the traditional triplet loss.

We find that the interval between the positive and negative sample pairs in the triplet loss function is always fixed during the training. However, this fixed interval setting has an inevitable defect. When the interval is relatively large, the network needs to learn a relatively large interval in the feature space at the beginning of training, which is very difficult for the network. When the interval is relatively small, the network can be easily trained at the beginning and the learning outcomes are good; however, with the continuous deepening of learning, the setting of small intervals can greatly reduce the learning effect of the network, and at this time, a relatively large interval should be set instead to increase the learning difficulty of the network so that the network can continue to learn.

Second, we also find that during each training, TriHard Loss only selects the hardest samples of a sample a to participate in the training while ignore other samples, resulting in low training efficiency and insufficient learning. For the sample a, although other samples are not the hardest, they are the second hardest. With the continuous iteration of training, these second-hardest samples will gradually become the hardest and thus participate in the training. Therefore, for the sample a, during each training, we should consider that not only the hardest samples need to be fully learned but also other second-hardest samples need to participate in learning. Only in this manner can the learning efficiency of the network be improved and the convergence speed be accelerated.

Considering the above two points, we propose an adaptive TriHard Loss based on the original TriHard Loss.

First, we let the interval between pairs of positive and negative samples increase with the increasing number of iterations. The dynamic interval is expressed as follows:

$$m_{adaptive} = k \times epoch + b \tag{2}$$

where the left side of the equation represents the dynamic interval, k represents the degree of increase in interval, epoch represents the number of times that all data are trained in a round, and b represents an initial interval. With this design, the interval is small at the beginning of the training, and as the training continues, the interval gradually increases. Thus, during the whole training process, the feature intervals can all be fully learned, and the training efficiency will also be greatly improved.

Second, for every sample in the batch, regardless of whether it is a positive or negative sample, we assign an appropriate weight according to the distance between the two, so that every sample in the batch participates in training each time, greatly accelerating the efficiency of network learning. We refer to the improved TriHard Loss as the adaptive TriHard Loss, which is expressed as follows:

$$Loss = \frac{1}{P \times K} \sum_{a \in batch} \left(\sum_{p \in A} wpd_{a,p} - \sum_{n \in B} wnd_{a,n} + m_{adaptive} \right)_+ \tag{3}$$

where A represents the set of samples in a batch belonging to the same class as that of the sample a and B represents the set of samples in a batch belonging to different classes from that of the sample a. To calculate the specific loss for the sample a, we calculate the sum of weighted distances between the positive and negative sample pairs, respectively, formed with

the sample a and then combine with the adaptive interval between sample pairs to obtain the loss of the sample a. The corresponding loss is calculated for each sample in the batch to obtain the overall loss. *wp* and *wn* are the weights assigned to the positive and negative samples, respectively, of sample a. They are expressed as follows:

$$wp = \frac{e^{d_{a,p}}}{\sum_{pi \in A} e^{d_{a,pi}}}, wn = \frac{e^{-d_{a,n}}}{\sum_{ni \in B} e^{-d_{a,ni}}} \tag{4}$$

The adaptive TriHard Loss can not only dynamically adjust the interval between positive and negative sample pairs but also set the adaptive weight. Therefore, the efficiency and stability of training can both be greatly improved, and the convergence speed will also be faster, effectively reducing the risk of overfitting.

4 Experimental results and analysis

In this section, we first explore the outcomes and effects of different forms of loss function and verify the effectiveness of the loss function proposed in this paper. Second, for the branches of local feature extraction, we explore the influence of dividing the feature map into different number of stripes on recognition. Finally, we also test the effectiveness and performance of different branches in the network.

4.1 Experimental configuration and parameter settings

The system environment of this study includes an Ubuntu 16.04 operating system, Intel Core i7-7700 K CPU, 8 GB memory, NVIDIA GeForce GTX 1080 graphics card, PyTorch 0.4 open source deep learning framework, CUDA 8.0, and cuDNN 5.0.

The parameters during model training are as follows: the size of the input image =384 × 128; use of ADAM optimization algorithm for training, with the exponential decay rates of the first and second moment estimation, beta1 and beta2, being 0.9 and 0.999, respectively, and epsilon = 10^{-8}; initial learning rate =2×10^{-4}, epoch = 200, and delay of learning rate at the 140th and 180th epochs, respectively, at a rate of 5×10^{-4}; batch_size = 48, K = 4, and P = 12.

For the parameters in Eq.(2), we conduct statistical analysis on the distance of eigenvectors of inner-class samples in the experimental data set. The average intra-class distance is about 0.6. In order to increase the separability between classes, the inter-class distance after network training is set according to the principle of 3 times variance, and the expected inter-class distance after iteration is 6 times the intra-class distance, i.e., 3.6. According to this principle, under the constraint condition of epoch = 200, we set the initial B as 1.6 and K as 0.01.

In terms of experimental data, we mainly use the Market1501 dataset, and the subsequent experimental environment and parameters are basically consistent with the above settings.

4.2 Evaluation indices for person re-ID algorithm

The commonly used indices for evaluating person re-ID algorithms include rank-k and mAP.

(1) rank-k

The rank-1 matching rate (rank-1) [1] is a commonly used index, which refers to the probability that the image to be queried and the image that ranks the first in similarity in the search library belong to the same target, i.e.,

$$rank{-}1 = \frac{\sum_{i\in 1,2,\ldots,m} Si}{m} \tag{5}$$

where m is the total number of images and si is a flag variable representing whether the ith image to be queried and the image ranking the first in similarity belong to the same target (si=1 if so, and si=0 if not). In general, a larger rank-1 means a better performance of the model. Therefore, rank-1 is the most direct and the most important index, and it has been used extensively to evaluate the performance of the model.

The k-matching rate (rank-k) denotes the probability that the image in the top k-position of similarity ranking in the retrieval database belongs to the same pedestrian as the image to be retrieved. Commonly used are rank-1, rank-3, rank-5, rank-10, and rank-20. As with rank-1, a larger rank-k means a better performance of the model. The rank-k index represents the judgment of whether there exists at least one image in the first k images belonging to the same person as the image to be queried. Hence, rank-k is a reflection of the comprehensive search ability of the model and can more comprehensively measure the performance of the model than rank-1.

(2) mAP

Because the rank-k index cannot well measure the recall rate of the model, Zheng et al. in 2015 introduced for the first time the mean average precision (mAP) index into the evaluation system for person re-ID [3]. The mAP index is a trade-off between the precision rate and the recall rate and can more objectively and comprehensively evaluate the performance of the model.

Based on the combinations of real categories and model prediction categories, Common classification problems can be divided into four cases, namely, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The precision rate P and the recall rate R are as follows:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

Then, a two-dimensional curve (called P-R curve) can be drawn by taking the precision rate P as the vertical coordinate and the recall rate R as the horizontal coordinate. The area enclosed under the P-R curve is called the average prevision (AP), and the mean of AP values of all classes is called mAP.

4.3 Experimental comparison of loss functions

In this section, the effects of different loss functions are tested. ResNet50 is used as the backbone network. The outputs of the last convolutional layer of ResNet50 are used as the

person features. Then, the Euclidean distance is used to calculate the similarity for recognition. The parameters of different loss functions remain consistent during the training. Market1501 [3] is used as test dataset. The test results of different loss functions are reported in Table 1.

It can be observed that because the triplet loss constraints the distance between sample pairs, it has better test results than the softmax loss, which also shows that it is not enough to learn different classes only; rather, an in-depth learning of the distance relationship between classes is also required. Second, comparison of the triplet loss and the TriHard Loss finds that both mAP and rank1 indices are greatly improved, indicating that batch hard mining can be of great hep to the triplet loss in that improving the quality of triplet sampling ensures the availability of high-quality triplets every time to train the network to improve the training results. Finally, a comparison of our proposed adaptive TriHard Loss and the original TriHard Loss reveals that mAP and rank-1 are increased by 1.7% and 1.4%, respectively, indicating the effectiveness of our proposed adaptive TriHard Loss.

4.4 Experimental results of person feature extraction

As is known from the introduction of the network architecture in earlier sections, local feature extraction requires horizontal division of the feature map. Therefore, experiments are carried out on the number of horizontal stripes resulting from the division of the feature map to obtain the optimal design of the network architecture. The detailed experimental results are reported in Table 2.

It can be found from observation of the above experimental results that with increasing number of divided stripes, both the mAP and rank-1 values of the model first increase and then decrease, a variation pattern that is consistent with common sense. The smaller the number of divided stripes in the feature map, the closer the results to those of the global feature. As a coarse-grained person descriptor, the global feature can only be used to compare the overall appearance and outline of different persons, but it is not sufficient to distinguish persons that require local fine comparison, resulting in a low recognition rate at the beginning. With increasing number of divided stripes, the model will pay more attention to the individualized area in each of the different stripes to learn more discriminative local features, and therefore the results will be increasingly better. However, as the number of divided stripes increases, the information contained in each stripe area decreases, leading to disordered network learning and lowered learning outcomes. Therefore, d = 6 is generally used as the default setting.

To illustrate the effectiveness of the network designed in this paper, we conduct experiments on the global feature branch, local feature branch, and the combination of the two, respectively, in the network to clarify the function of each branch. The loss function is set as 1*CrossEntropy Loss +2*Adaptive TriHard Loss, where 1* CrossEntropy Loss represents the use of one times the cross entropy loss function, and 2*Adaptive TriHard Loss represents the use of two times the adaptive TriHard Loss. Hence, the final total loss function contains both

Table 1 Test results of different loss functions

Loss function	mAP (%)	rank1(%)
Softmax	41.3	65.8
Triplet loss	54.8	75.9
TriHard Loss	68.0	83.8
$L_{adaptive}$	69.7	85.2

Table 2 Effect of the number of divided stripes in the feature map on the network performance

Number of local regions divided, d	mAP (%)	Rank-1(%)
1	75.4	88.5
2	79.7	90.3
4	81.8	92.7
6	83.4	93.2
8	81.6	92.4
12	80.5	90.8

the cross entropy loss function and the adaptive TriHard Loss, making the overall learning outcome of the network better. The detailed experimental results are reported in Table 3.

In Table 3, "Global" denotes the global features and "Local" represents the local features. From Table 3, it can be found that local features have a significant effect on the values of mAP and rank-1, which means that the local feature branch can provide the discriminative part of a person. Second, it can be observed that the combination of the global feature branch and the local feature extraction branch achieves the best results, indicating that although the local feature branch is highly efficient, there is still some overall information that is not learned, and the global feature branch can exactly make up for this defect. Therefore, it is illustrated that the two are highly complementary and can promote each other to jointly improve the recognition outcomes.

Moreover, we compare the multibranch network proposed in this paper with a similar algorithm published recently (see Tables 4 and 5) using the test datasets of CUHK03 [14] and Market1501 [3].

As can be observed from the above tables, compared with other algorithms that are based on combinations of global and local features, the network proposed in this paper achieves the best results, and values of each index far exceed those of the other algorithms, indicating the effectiveness of the multibranch network.

To show the performance of feature extraction proposed in this paper intuitively, we give the corresponding heat-map for several images as shown in Fig. 2.

In the heat map, the redder the color of the region means that the features of the region play a greater role in recognition. From Fig. 2, it can be founded that the color on the human body is

Table 3 Test results of different branches of the network

Networks with different branches	mAP (%)	rank1(%)
Global	75.4	88.5
Local	83.4	93.2
Global + Local	84.6	93.8

Table 4 Test results of different algorithms on the CUHK03 dataset

Algorithm	Rank-1(%)	Rank-5(%)	Rank-10(%)
Spindle Net [32]	88.5	97.8	98.6
PDC [25]	88.7	98.6	99.2
PL-Net [30]	82.8	96.6	98.6
GLAD [28]	85.0	97.9	99.1
This paper	90.6	98.7	99.4

Table 5 Test results of different algorithms on Market1501 dataset

Algorithm	mAP (%)	rank1(%)
Spindle Net [32]	–	76.9
PDC [25]	63.4	84.1
PL-Net [30]	69.3	88.2
GLAD [28]	73.9	89.9
This paper	84.6	93.8

more red relative to the background. It shows that the feature extraction network proposed in this paper can pay more attention to the human body, which is conducive to person re-identification.

5 Conclusion

This paper first introduces a CNN-based person feature extraction method; second, a multibranch network architecture combining a global feature branch and a local feature branch is constructed, and the construction of each branch of the network is described in detail; third, the loss functions commonly used in person re-ID are discussed, and an adaptive triplet loss function is proposed. In the experiments, the effects of various types of loss functions are validated, the rationality of the design of the multibranch network architecture is also demonstrated, and the effectiveness of the setting of global feature branches and local feature branches is verified.

Acknowledgments We are grateful for the assistances of the reviewers and editors, and this research was supported in part by the Langfang secience and technology research and development project(project number: 2021013071, 2021011066).

Funding This research was funded by the National Natural Science Foundation of China (Grant Nos. 41671441, 41531177, and U1764262). This research was funded by the Langfang science and technology research and development project(project number: 2021013071, 2021011066).

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Bolle RM, Connell JH, Pankanti S et al (2005) The relation between the ROC curve and the CMC [C]. Fourth IEEE workshop on automatic identification advanced technologies (AutoID'05). IEEE, pp 15–20
2. Cai Q, Aggarwal JK (1996) Tracking human motion using multiple cameras [C]. International Conference on Pattern Recognition. Vienna, Austria, pp 68–72
3. Chen YC, Zheng WS, Lai JH, Yuen PC (2017) An asymmetric distance model for cross-view feature mapping in person reidentification [J]. IEEE Trans Circuits Syst Video Technol 27(8):1661–1675
4. Farenzena M, Bazzani L, Perina A et al (2010) Person re-identification by symmetry-driven accumulation of local features [C]. Computer vision and pattern recognition(CVPR), 2010 IEEE Conference on 2010, pp 2360–2367
5. Geng M, Wang Y, Xiang T et al (2016) Deep transfer learning for person re-identification [J]
6. Gheissari N, Sebastian TB, Hartley R (2006) Person re-identification using spatiotemporal appearance [C]. IEEE Conference on Computer Vision and Pattern Recognition. New York, USA, pp 1528–1535
7. Gong SG, Cristani M, Yan SC et al (2013) Person re-identification [J]. Adv Comput Vis Pattern Recognit 42(7):301–313
8. Gray D, Brennan S, Tao H et al (2007) Int J Comput Vis 89(2):56–68
9. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
10. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification [J]
11. Jungling K, Bodensteiner C, Arens M (2010) Person re-identification in multi-camera networks [C]. Computer Vision and Pattern Recognition Workshops. Colorado, USA, pp 709–716
12. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition [J]. Neural Comput 1(4):541–551
13. Li Y, Wang X (2018) Investigation of intelligent video surveillance system based on artificial intelligence technology [J]. Technol Innov Appl 34:76–77
14. Li W, Zhao R, Xiao T et al (2014) DeepReID: deep filter paring neural network for person re-identification [C]. Proceeding of the IEEE Computer Society Conference on Computer Vision and Person Recognition. Columbus. Ohio, pp 152–159
15. Liao S, Hu Y, Zhu X et al (2015) Person re-identification by local maximal occurrence representation and metric learning [C]. The IEEE Conference on Proceedings of Computer Vision and Pattern Recognition, pp 2197–2206
16. Lin Y, Zheng L, Zheng Z et al (2017) Improving person re-identification by attribute and identity learning [J]
17. Lu X, Ma C, Ni B, et al (2018) Deep regression tracking with shrinkage loss[C]. Proceedings of the European conference on computer vision (ECCV), pp 353–369
18. Lu X, Wang W, Ma C et al (2019) See more, know more: Unsupervised video object segmentation with co-attention siamese networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3623–3632
19. Lu X, Wang W, Danelljan M et al (2020) Video object segmentation with episodic graph memory networks[J]. arXiv preprint arXiv:2007.07020
20. Lu X, Ma C, Shen J et al (2020) Deep object tracking with shrinkage loss[J]. IEEE Trans Pattern Anal Mach Intell
21. Matsukawa T, Okabe T, Suzuki E et al (2016) Hierarchical gaussian descriptor for person re-identification[C]. IEEE Conference on Computer Vision and Proceedings of the Pattern Recognition, pp 1363–1372
22. Mignon A, Jurie F (2012) PCCA: a new approach for distance learning from sparse pairwise constraints [C]. 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 2666–2672
23. Oreifej O, Mehran R, Shah M (2010) Human identity recognition in aerial images [C]. IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, pp 709–716.
24. Schroff F, Kalenichenko D, Philbin J (2015). Facenet: A unified embedding for face recognition and clustering [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
25. Su C, Li J, Zhang S et al (2017) Pose-driven deep convolutional model for person re-identification [C]. 2017 IEEE International Conference on Computer Vision (ICCV), pp 3980–3989
26. Varior RR, Shuai B, Lu J et al (2016) A siamese long short-term memory architecture for human re-identification [C]. European conference on computer vision. Springer, Cham, pp 135–153
27. Wang W, Lu X, Shen J et al (2019) Zero-shot video object segmentation via attentive graph neural networks[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9236–9245
28. Wei L, Zhang S, Yao H et al (2017) Glad: global-local-alignment descriptor for person retrieval [C]. Proceedings of the 25th ACM international conference on multimedia. ACM, pp 420–428
29. Yang Y, Yang J, Yan J et al (2014) Salient color names for person re-identification [C]. European conference on computer vision, pp 536–551

30. Yao H, Zhang S, Hong R, Zhang Y, Xu C, Tian Q (2019) Deep representation learning with part loss for person re-identification [J]. IEEE Trans Image Process 28:2860–2871
31. Yi D, Lei Z, Li SZ (2014) Deep metric learning for practical person re-identification [C]. International Conference on Pattern Recognition. Stockholm Waterfront, Sweden
32. Zhao H, Tian M, Sun S et al (2017) Spindle net: Person re-identification with human body region guided feature decomposition and fusion [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1077–1085
33. Zheng WS, Li X, Xiang T et al (2015) Partial person re-identification. In: ICCV