Pattern Recognition Letters 117 (2019) 161–168

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Multilevel triplet deep learning model for person re-identification

Cairong Zhao[a,b,1,*], Kang Chen[a,*], Zhihua Wei[a,1], Yipeng Chen[a], Duoqian Miao[a], Wei Wang[a]

[a] Department of Computer Science and Technology, Tongji University, Shanghai, China
[b] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou 350121, China

ARTICLE INFO

Article history:
Available online 20 April 2018

Keyword:
Multilevel feature extraction
Triplet architecture
Person re-identification

ABSTRACT

Person re-identification (Re-ID) is a typical computer vision problem which matches pedestrians from different cameras. It remains challenging to cope with the variation in light, the change of human pose and view point difference. Many existing person re-identification methods may have difficulty in matching pedestrians when their pictures are similar in appearance or there is object occlusion in pictures. The main problem with these existing methods is that the detail and global features of the images are not well combined. In this paper, we improved the performance of deep CNN network with the proposed Multilevel feature extraction strategy and built a novel Multilevel triplet deep learning model corresponding to our method. The Multilevel feature extraction strategy focuses on combining fine, shallow layer information with coarse, deeper layer information by extracting fusion feature maps from different layers for a better representation of pedestrians. The Multilevel triplet deep learning model (MT-net) provides an end-to-end training and testing plain for our feature extraction strategy. The experiment on the benchmark datasets validated that our multilevel triplet deep learning model had better performance comparing with many state-of-the-art person re-identification methods.

1. Introduction

Person Re-Identification can be easily introduced as given one single shot or multiple shots of a pedestrian and match the same individual among a set of gallery pictures photographed by different set of cameras. As a computer vision task, it remains challenging to cope with the variation in light, the change of human pose and view point which may increase the difficulty in matching different individuals. Earlier studies [11,18–20,27–29,47,49] mainly focus on two components: extracting discriminative and robust features of input images as descriptors and learning an effective metric to measure the similarity between input images. These methods mostly based on man-coded features. They are restricted by the uncontrollable visual appearance of pedestrians in different viewpoints, especially when the cameras are not overlapped. Thus the extracted feature descriptors are not robust enough for metric calculation.

Recently, with the development of deep learning and the availability of large-scale person Re-ID datasets like CUHK03 [24], Market1501 [25] and DukeMTMC-reID [41], a bunch of effective CNN frameworks [1–7] were proposed for person re-identification and they have made significant improvement on improving the accuracy of re-identification.

There are several effective deep learning models for person re-identification: The verification model, classification model and triplet model. The verification models [2,3,31] treat re-identification task as a binary verification problem. The input pictures of pedestrians are organized in pairs. Learned features are mapped directly to fully connected layer to measure the similarity between two input pictures and then determine whether they were from the same or different individuals. The verification models are generally supervised by Softmax loss. These kind of models can't make full use of the label of each image and the procedure of training and testing is different. The classification models [14,32,33] set every individual as a class and every pedestrian has their labels. During the training process, discriminative features are learned to represent the reflection of pictures. The training procedure is supervised by cross-entropy loss. The testing process was independent from training and the square Euclidean distance is used to measure the similarity between pictures. The disadvantages of these models are the overmuch number of classes and each of them has only a few instances for training and testing. The triplet models [9,12,38] focus on the ranking task. Input pictures are organized in triplets, containing positive pairs and negative pairs. The models consider samples of positive/negative pairs simultaneously and Euclidean distances of embedding feature map

* Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, China.
E-mail addresses: zhaocairong@tongji.edu.cn (C. Zhao), zhihua_wei@tongji.edu.cn (Z. Wei).
[1] The authors contribute equally to this work.

https://doi.org/10.1016/j.patrec.2018.04.029
0167-8655/

Fig. 1. Examples of pairs of pictures that are overall approximate in appearance but different in details and a mismatch caused by only focusing on the overall appearance.

are directly compared. Triplet loss and its variants are used for supervising the training process.

Most of the methods mentioned above make operations on the global features extracted by deep CNN network. To get these global features, an efficient way is transfer learning, that is initializing weights and biases with the model (ResNet [34], GoogleNet [21]) pretrained on ImageNet [23] and finetune the parameters with datasets for person re-identification. This is an effective way to extract robust and discriminative global features from input pictures for person re-identification.

However, descriptors extracted by ResNet or GoogleNet are mostly global, semantic features after deep convolution operations. The small-size differences and precise details are often ignored after Pooling or Relu layers. These may cause confusion when two individuals are overall approximate in appearance but different in details. The left part of Fig. 1 shows two instances of this circumstance. We can see intuitively that the appearance and posture of pedestrians in each group are very close. The difference lies in several pixels and the global, semantic features may lose these details. Besides, because of object occlusion, sometimes there is an obvious difference in some areas of the images of the same person. This may lead a part of the global feature cannot describe pedestrians and mislead the judgment while matching. At the same time, the detail features are often ignored in the deep CNN network. If we only consider the global feature, there will be a mismatch as shown in the right half of Fig. 1.

In some works [13,15–17], they combine fine, shallow layer information with coarse, deeper layer information to make a more comprehensive representation of input objects. They did feature fusion from different layers or regions. These works achieved good results in the field of object detection and semantic segmentation. Although there are some commonalities in computer visual aspects, person re-identification is another problem and we cannot use these methods directly. Affected by the works mentioned above, we applied the idea of feature fusion to person re-identification task. We proposed a more reasonable feature extraction strategy that based on triplet model. We combine information of each layer and concern both detail and global feature through a standardized method. The new triplet deep learning model shows advantages in the following aspects: Comparing with previous deep CNN networks, the information distributed over different layers are collected and better organized. The detail feature and global feature are both taken into account and the proportion is optimized with training. This strategy took full advantages of the pretrained model and accelerated the training process. Besides, our model jointly composes multilevel feature extraction and representation operation for end-to-end training. The experiment on the benchmark datasets showed that our model has an improved performance on person re-identification task by using the new feature extraction strategy.

The contributions of this paper can be summarized as follows: (1) We proposed a novel Multilevel feature extract strategy to combine coarse and fine information from different layers, which contributes to a more robust and discriminative representation of input pictures of individuals. (2) We designed an end-to-end Multilevel triplet deep learning model (MT-net) to compute extracted Multilevel features efficiently for person re-identification task. (3) Experiments on two popular large scale datasets, CUHK03 [24] and Market1501 [25]. The result showed that our model was efficient and significantly improved the performance comparing with many existing state-of-the-art methods.

2. Related work

In this part, we review existing person re-identification methods most related to our work, especially deep leaning based methods.

Typical person re-id systems are consist of two parts: getting robust and discriminative descriptors for the query image and the gallery images; finding an effective way to measure the distance between different images and constructing a distance metric for comparing. A lot of works [10,42–45] have been done for digging better man-coded features from raw pictures invariant to pose, lighting and view variations. HSV, SILTP [35] and Gabor features [36] were able to describe identities to a certain extent. Several effective metric learning and ranking algorithms [20,26,28,29] were proposed to measure the feature map distance and distinguish between pedestrians.

Recently, inspired by the implement of CNN network in many computer vision problems, people use deep CNN models for person re-identification. Achievements have been made with the availability of large scale datasets such as CUHK03 [2] and Market-1501 [3].

Typical deep learning architecture for person re-identification can be divided into two subnetworks. The first subnetwork aims to learn a person's representation of input images. It develops from shallow convolutional network into today's more popular deep convolutional network (ResNet, GoogleNet), which were proved to be able to extract deep semantic features. They have good performance in learning discriminative and robust features and can be easily transferred and fine-tuned for person re-identification task. Ahmed et al. [3] proposed a method which simultaneously learns features and a corresponding similarity metric for person re-identification. The feature extraction structure included a layer that computes cross-input neighborhood differences. Zhao et al. [38] divided human images into different parts and get receptive field in the stage of feature extraction. The second subnetwork aims to compare the feature representation and penalize the misalignment between learnt similarities and ground-truth similarities. Usually we use cross entropy loss for classification, Siamese contrastive loss for verification and triplet loss for ranking. The Euclidean distances are utilized for comparison. Zheng et al. [2] proposed a siamese network that simultaneously computes the identification loss and verification loss. Hermans et al. [9] showed that, for models trained from scratch as well as pretrained ones, using triplet loss to perform end-to-end deep metric learning outperforms any other published method by a large margin. Chen et al. [1] proposed a multi-task deep network that combines verification model and ranking model. It contains a cross-domain architecture that is capable of using an auxiliary set to assist training on small target sets.

To improve the accuracy of object detection and classification, there is a trend that to study the fusion of features between different layers. A very intuitive idea is that the lower layer of the neural network usually retains much more fine-grained features;

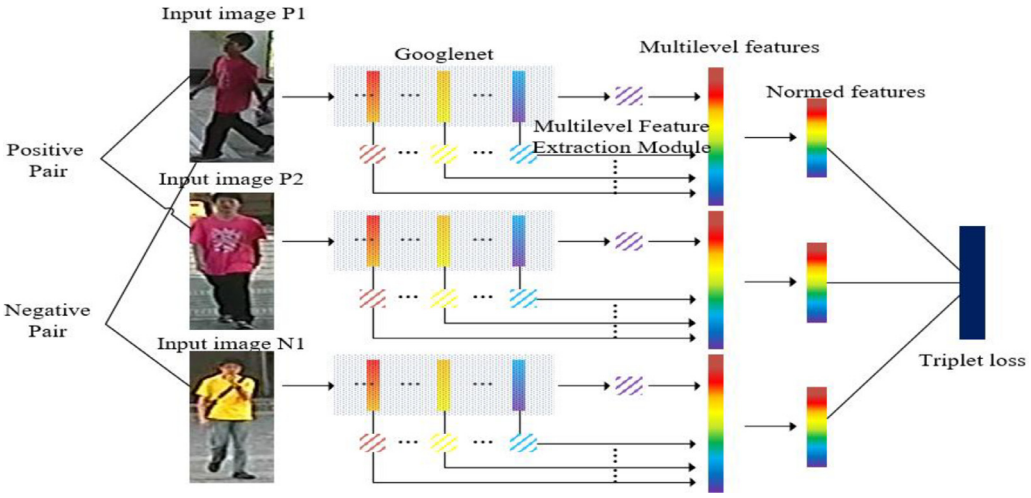

Fig. 2. Proposed multilevel triplet deep learning model (MT-net).

and the deeper layer usually has better semantic features. So a lot of work has been done to combine the features of different layers. Cai et al. [15] made proposal on different layers for object detection. Kong et al. [13] improved accuracy of object classification by cascaded features of different layers. Lin et al. [16] proposed feature image pyramid and Abhinav et al. used Top-down modulation extracted different layers' features. These works showed their advantages in object representation and computing accelerate.

Inspired by these works, we proposed a novel Multilevel feature extract strategy and designed an end-to-end Multilevel triplet deep learning model. Our framework was also related to transfer learning [7], data augmentation and step-wise learning [46,48]. These methods improved the performance of our network.

3. Proposed multilevel deep learning model (MT-net)

Person re-identification aims to find the images of the same identity with the probe image from a set of gallery images. When two individuals are overall approximate in appearance but different in details or there is object occlusion in the picture, there will be confusion in matching individuals. To deal with this problem, we proposed a novel feature extraction strategy that consider global and detail feature simultaneously and built a Multilevel triplet model for end-to-end training and testing. The proposed Multilevel triplet model took three pictures as input. The raw pictures were fed into the triplet deep convolutional network. We first estimate the activate area from different layers then extract features from the detected area. Concatenate the features extracted from layers to form fusion feature map. The extracted descriptors considered both fine, shallow layer information and coarse, deeper layer information. To make the presentation of images from the same identity are closer to those from different identities, the network was optimized by triplet loss.

3.1. The overall framework

Fig. 2 illustrates the overall framework of our methods. The training data was given as a set of triplets,

$$\text{Input} = \{\text{Tr}_1, \text{Tr}_2, \text{Tr}_3, \ldots, \text{Tr}_n\}\ ,\ \text{Tr}_i = \langle \text{p1}_i\ \text{p2}_i\ \text{n1}_i \rangle,$$

where i means the i_{th} triplet, $(\text{p1}_i\ \text{p2}_i)$ is a positive pair of pictures from the same individual and $(\text{p1}_i\ \text{n1}_i)$ is a negative pair of pictures from different individuals.

Triplet training images were fed into three deep CNN models with shared parameter for Multilevel feature extraction. The output of each model is the extracted feature map for each picture. The raw images were mapped into a learned feature space, denoted by $\omega(\text{Tr}_i) = \langle \omega(\text{p1}_i)\ \omega(\text{p2}_i)\ \omega(\text{n1}_i) \rangle$. This process will be described in detail in Section 3.2. During training process, three network models share parameter set. Weights and biases of the network were optimized by triplet loss.

To make sure $d_{l2}(\omega(\text{p1}_i),\ \omega(\text{p2}_i))$ is less than $d_{l2}(\omega(\text{p1}_i),\ \omega(\text{n1}_i))$, where d_{l2} means the L2-norm distance, that is

$$d_{l2}(\omega(\text{p1}_i),\ \omega(\text{p2}_i)) - d_{l2}(\omega(\text{p1}_i),\ \omega(\text{n1}_i)) \le \gamma \tag{1}$$

In this equation γ is negative to enforce the requirement by a predefined margin. We set γ as 0.2 empirically.

So we use triplet loss as the loss function, which can be represented by the following equation.

$$L_{tr} = \frac{1}{N}\sum_{i=1}^{N}[d_{l2}(\omega(\text{p1}_i),\ \omega(\text{p2}_i)) - d_{l2}(\omega(\text{p1}_i),\ \omega(\text{n1}_i)) + \alpha]_{-} \tag{2}$$

where

$$d_{l2}(\omega(\text{p1}_i),\ \omega(\text{p2}_i)) = ||\omega(\text{p1}_i) - \omega(\text{p2}_i)||_2^2\ d_{l2}(\omega(\text{p1}_i),\ \omega(\text{n1}_i)) = ||\omega(\text{p1}_i) - \omega(\text{n1}_i)||_2^2 \tag{3}$$

When combine (1, 2, 3)

$$L_{tr} = \frac{1}{N}\sum_{i=1}^{N}\left[||\omega(\text{p1}_i) - \omega(\text{p2}_i)||_2^2 - ||\omega(\text{p1}_i) - \omega(\text{n1}_i)||_2^2 + \alpha\right]_{-} \tag{4}$$

where N is the number of triplets, it changes with the number of pictures in training set. α represents the margin of positive and negative pairs $\omega(\text{p1}_i)\ \omega(\text{p2}_i)\ \omega(\text{n1}_i) \in \mathbb{R}^{512}$ denotes the features of each input picture.

Assume that we have 1000 training pictures from 100 different pedestrians. The number of triplets $N_0 \approx C_{10}^2 * (99 * 10) * 100 = 4{,}950{,}000$, after preprocessing such as cutting and inversion for expanding training sets, N_0 will grow geometrically, which ensures the number of triplets is large enough for optimizing our MT-network. We randomly select N triplets from N_0 in training. The size of N depends on the batchsize and epoch. When we set batchsize as 400 and epoch as 20, then N will be their product 8000.

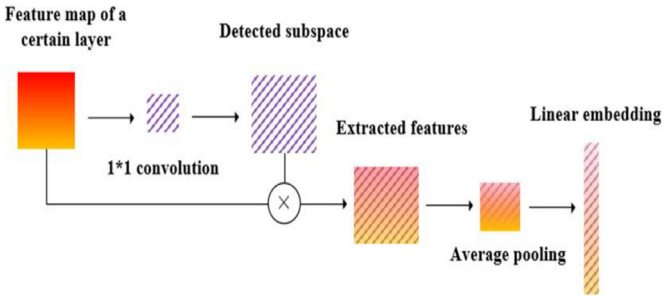

Fig. 3. Multilevel feature extraction module.

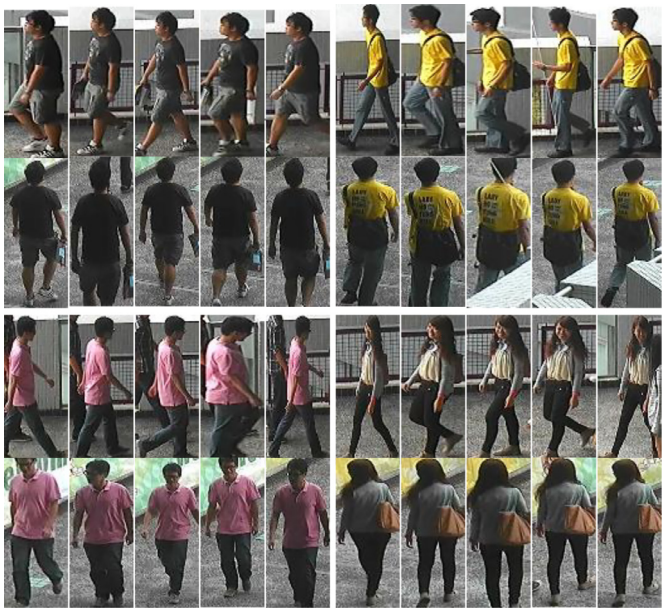

Fig. 4. Example images from CUHK03 dataset.

There are two aspects of our network structure need to be specifically explained. The next two sections will introduce the feature extraction strategy and network optimization.

3.2. Multilevel feature extract strategy

In our work, we chose GoogleNet as the base CNN network for it has the considerable capability for feature extraction to classify general images. We transferred the Googlenet model which pre-trained on ImageNet and initialized weights and biases with it. During the training process, the parameters were fine-tuned to fit the demands of person re-identification task.

To study the fusion of features from different layers, we concatenated multiple feature maps from different layers. To deal with the different output dimensions of convolution layers in Googlenet and combine Multilevel features at the same solution, we carried out a standardized sampling scheme, which is illustrated in Fig. 3.

For each feature extraction module, First we created a detected subspace M(x,y) from an 1*1 convolution layer and a sigmoid layer. The convolution of 1*1 can better combine the information of each channel and reduce the dimension, through a sigmoid layer we got an detected subspace and then extended the dimension of subspace to 512 to form M(x,y). Then we found the discriminative region of a certain layer and extract features from M(x, y). We extracted features $F_{ext}(x, y, c)$ over these subspaces and then let the tensor went through an average pooling layer. The output of the feature extraction module can be represented as f_k. The input of the feature extraction module can be represented as a tensor F(x, y, c).

Fig. 5. Example images from Market1501 dataset.

$$\begin{aligned} M_{sub}(x, y) &= Conv\ (F) \\ F_k(x,\ y,\ c) &= F(x, y, c) \times M_{sub}(x, y) \end{aligned} \tag{5}$$

The feature F(x,y,c) means the c_{th} channel over the location (x, y) and k means the k_{th} feature extraction module.

Before concatenated all f_k as the final feature map, we implement a linear dimension-reduction layer to guarantee the consistency of each feature extraction module.

After concatenated multilevel features were extracted, we use local response normalization (LRN) to normalize our feature maps before compute losses.

Multilevel features carried both coarse and fine-granted features that can represent pedestrians better. The features of lower layers are considered making computation more efficient without overmuch redundant parameter.

3.3. Optimization

The parameters of our network are donated by θ. We minimize the triplet loss function over triplets formulated above.

$$L_{tr} = \frac{1}{N}\sum_{i=1}^{N}\left[||\omega(p1_i) - \omega(p2_i)||_2^2 - ||\omega(p1_i) - \omega(n1_i)||_2^2 + \alpha\right]_+ \tag{6}$$

The gradient can be formulated as:

$$\frac{\partial L_{tr}(\theta, p1_i,\ p2_i,\ n1_i)}{\partial\theta} = 2\sum_{i=1}^{N}\left[\frac{\partial\omega(p1_i)}{\partial\theta}(\omega(n1_i) - \omega(p2_i)) + \frac{\partial\omega(p2_i)}{\partial\theta}(\omega(p2_i) - \omega(p1_i)) + \frac{\partial\omega(n1_i)}{\partial\theta}(\omega(p1_i) - \omega(n1_i))\right] \tag{7}$$

When

$$d_{l2}(\omega(p1_i),\ \omega(p2_i)) - d_{l2}(\omega(p1_i),\ \omega(n1_i)) \le \gamma \tag{8}$$

Algorithm1:

Input
Training samples: $\{p1_i\ p2_i\ n1_i\}$
Output
Parameter of network: θ
while t < T or not convergent do:
  $t \leftarrow t + 1$
  $\frac{\partial L_{tr}(\theta, p1_i,\ p2_i,\ n1_i)}{\partial\theta} = 0$
  for all input image samples $p1_i\ p2_i\ n1_i$,
  Compute feature embedding $\omega(p1_i),\ \omega(p2_i)\ ,\ \omega(n1_i)$ by forward propagation;
  Calculate $\frac{\partial\omega(p1_i)}{\partial\theta}$, $\frac{\partial\omega(p2_i)}{\partial\theta}$, $\frac{\partial\omega(n1_i)}{\partial\theta}$ by back propagation;
  Calculate $\frac{\partial L_{tr}(\theta, p1_i,\ p2_i,\ n1_i)}{\partial\theta}$ according to Eq. (2);
Update the parameter: $\theta^t = \theta^{t-1} - \epsilon_t \frac{\partial L_{tr}(\theta, p1_i,\ p2_i,\ n1_i)}{\partial\theta}$
end while

The formulation shows that the triplet loss can be easily compute by one-dimensional operation. So we draw a mini-batch of samples instead of down sampling triplets. Directly handle batch of samples was more efficient in computation.

We use the Goolenet model trained on the Imagenet to initialize the network parameters and use person re-identification datasets for fine-tune. By training our MT-network with triplets and optimize hyper parameters with triplet loss for about 50,000 iterations, the network has the ability to make the distance of pictures from the same individual closer to different ones. This can help us to distinguish different pedestrians.

3.4. Comparison to prior works

Here, we compare our Multilevel triplet model with state-of-the-art person re-identification frameworks and point out similarities and differences between them.

Verification&Identification model for person re-identification: A typical architecture of these models is shared Siamese CNN architecture and compute verification loss and identification loss simultaneously when training. It used dropout strategy to prevent overfitting. Our Multilevel triplet model shared some similarities with these architectures in sharing weights and using pretrained models. Besides, our model allows us to perform end-to-end learning that directly optimizes the network for final task. We can match individuals by simply compute the Euclidean distance of descriptor vectors.

Triplet Re-identification: Triplet model is a popular ranking model for person re-identification. In [9], Alexander et al., discussed the triplet model with various triplet loss functions. Our Multilevel triplet model share some similarities that the triplet CNN deep learning structure with triplet loss can efficiently complete end-to-end training and testing system for person re-identification. The triplet re-identification faces poor converge outcomes without hard-mining. The embedding feature map may cause overmuch clustering center. Our Multilevel triplet model improved the accuracy by extracted fusion feature from different layers rather than simply transfer exist models (VGG,ResNet,GoogleNet) pretrained over ImageNet. The experiment in CUHK03 and MARKET1501 in Section 4 shows the difference.

Attribute learning Re-identification: Attribute learning is a new trend in computer vision. Lin et al at [8] proposed an attribute learning method for person re-identification. The core idea of their work is to get a more comprehensive representation of pedestrians. The attributes are extracted from both coarse grained and fine-grained features. Our work involves the same thought. The attribute learning took a lot of time and effort to make proper datasets and it may face difficulties in overfitting. Our Multilevel triplet model directly trained over CUHK03 and MARKET1501 and can be easily transferred to other computer vision tasks.

To sum up, our method extends the advantages of the triplet model. It can make more use of the global and local information with the multilevel feature extraction strategy. Besides, it calculates more efficiently and can be better expanded with the end-to-end training plain.

4. Experiment

4.1. Datasets

CUHK03: This dataset consists of 13, 164 images of 1,360 identities. Images are collected from 6 different pairs of camera views. It contains both human-labeled and detected sets. We randomly select the provided training/test set. We randomly set two images as the probe and gallery respectively for testing.

Market-1501: Market-1501 contains 32,668 detected-person bounding boxes of 1501 identities. The images came from 6 different cameras, one of which was low pixel. At the same time, the data set provides training set and test set. The training set contains 12,936 images, and the test set contains 19,732 images. The image is automatically detected and cut by the detector, which contains some detection errors (close to the actual use). There are 751 identities in the training data and 750 identities in the test set. So in the training set, there are 17.2 training data per class (each person).

4.2. Evaluation metrics

We calculated the similarities between the query images with all the gallery images and rank them into a list according to the similarities. We used the Cumulative Matching Characteristic (CMC) curve and the mean Average Precision (mAP) to estimate the expectation of finding the correct match in the top n most similar matches. The mean average precision (mAP) score over Market1501 dataset is reported in Table 4.

4.3. Implementation details

We used Caffe [22] framework to implement our Multilevel triplet model. The settings and other details are introduced in this part.

Data preparation: Before training, we resized the pictures of pedestrians to 160×80. During the training process, the batchsize was set as 400. There were 40 individuals on average and each has 10 pictures of them at most. We shuffled the dataset and randomly ordered the images. We sampled two images from the same class and another image from a different class as a triplet.

Network architecture: We use a subnet of GoogleNet from the data layer to inception_4e layer. For each convolution layer, it followed by an extra 1×1 convolutional layer and then a nonlinear sigmoid layer. We extracted a part of feature map by conduct matrix multiplication between the output of sigmoid layer and the convolution layer. There are eight convolution layers and we extracted features from all of them. Then we concatenated eight extracted feature maps and obtain a 512-dim pedestrian descriptor f. Given a query image, its descriptor was extracted online. We sorted the cosine distance between the query and all the gallery features to obtain the final ranking result. Note that the cosine distance is equivalent to Euclidean distance when the feature is L2-normalized.

Training settings: We initialized the Multilevel triplet model with the GoogleNet model. The mini-batch size was 400. The learning rate was set as 0.001 at the beginning and then divided by 5 for every 20 K iterations. The weights of all new layers were initialized with “Xavier” . The weight decay is 0.0002 and the momentum for gradient update is 0.9. We trained our Multilevel triplet deep learning model for 50 K iterations within 6 hours on a NVIDIA 1080Ti GPU.

Data augmentation: To counter overfitting we performed data augmentation before training our network. We cropped images to 160×80 to generate 5 augmented images around the center by performing random 2D transformations for each training image.

Dropout: Experiment shows that dropout strategy is not applicable for our method. So we give up dropout layers. The result of the comparative experiments will be delivered in Section 4.4.

4.4. Experimental results

We carried out detailed experiments on two Re-ID datasets, CUHK03 and Market1501. The result showed the superiority of our proposed method.

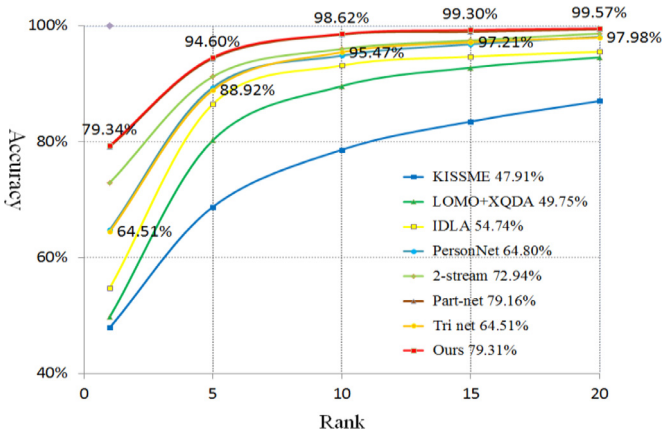

Fig. 6. CMC curves on CUHK03 dataset.

Table 1
Performance comparison on CUHK03 dataset.

Methods	Rank1	Rank5	Rank10	Rank20
KISSME[28]	47.91%	68.75%	78.62%	87.05%
XQDA[11]	49.75%	80.26%	89.61%	94.58%
IDLA[3]	54.74%	86.50%	93.17%	95.54%
PersonNet[40]	64.80%	89.40%	94.89%	98.12%
2-stream[2]	72.94%	91.26%	95.03%	98.69%
Part-net[38]	79.16%	94.39%	97.61%	99.41%
Pose[37]	78.22%	93.45%	96.32%	98.84%
T-net	64.51%	88.92%	95.47%	97.98%
MT-net	79.34%	94.60%	98.62%	99.57%

Table 2
Different numbers of feature extraction modules on CUHK03.

Num of modules	Rank1	Rank5	Rank10	Rank15	Rank20
2	68.33%	89.07%	92.56%	94.89%	96.55%
4	70.92%	92.85%	96.28%	97.50%	98.14%
6	75.40%	92.91%	96.89%	98.02%	98.70%
8	79.34%	94.60%	98.62%	99.36%	99.57%
10	75.15%	94.07%	97.46%	99.00%	99.21%

4.4.1. Results on CUHK03 dataset

For CUHK03 dataset, we used the detected image set for training and testing. The training or testing sets were randomly selected from the 'exp_set' folder. Fig. 6 and Table 1 illustrated the identification accuracy and CMC performance of our method comparing with other state-of-the-art person re-identification methods.

Our method outperformed in the accuracy by a large margin when compared with man-coded feature extraction and metric learning methods like KISSME [28] or XQDA [11]. When Comparing with popular deep learning methods like 2-stream [2] and Pose [37], our MT-net is different in feature extraction strategy and optimize methods. We have a small increase in result. When comparing with the competitive method Part-net [38] that also use triplet loss and based on Googlenet, the rank-1 accuracy is slightly better by 0.2%.

In order to reflect the superiority of our method, we introduced the T-net baseline as contrast in CUHK03 dataset. The T-net is the triplet deep learning architecture without multilevel feature extraction module. As is shown in Fig. 6 and Table 1, the accuracy of matching individuals was greatly enhanced when implementing the multilevel feature extraction module in the triplet architecture.

We set different numbers of feature extraction modules and did experiments on CUHK03 dataset to obtain the optimal parameter. These modules are evenly distributed in different convolution layers of GoogleNet. The results were listed in Table 2 and Fig. 7. We found that empirically the number of modules should be set as 8.

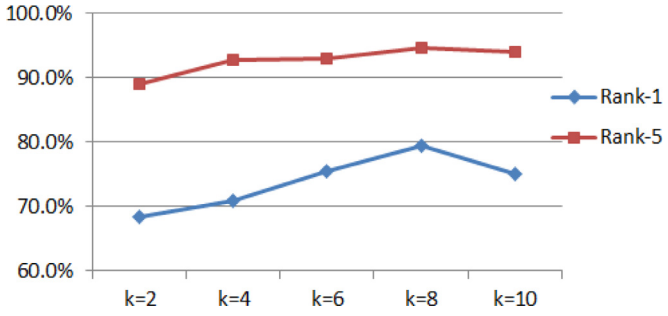

Fig. 7. Different numbers of feature extraction modules on CUHK03 dataset.

Table 3
With/without dropout layer on CUHK03 dataset.

Dropout/no dropout	Rank1	Rank5	Rank10	Rank15	Rank20
CUHK03(dropout)	20.56%	56.21%	72.64%	82.28%	87.92%
CUHK03(no dropout)	79.34%	94.60%	98.62%	99.36%	99.57%

Table 4
Performance comparison on Market1501 dataset.

Methods	Rank1	Rank5	Rank10	Rank20	mAP
KISSME[28]	46.72%	64.87%	73.18%	–	12.27
XQDA[11]	26.12%	–	–	–	20.04
LSTM[39]	62.44%	–	–	–	39.21
S-CNN[30]	66.76%	–	–	–	40.77
Pose[37]	78.06%	90.76%	94.17%	96.02%	58.33
2-stream[2]	78.15%	89.84%	93.27%	95.33%	57.92
Part-net[38]	79.44%	90.96%	93.52%	95.70%	60.40
T-net	65.02%	84.26%	88.21%	92.70%	50.71
MT-net	81.95%	92.53%	94.30%	96.20%	62.98

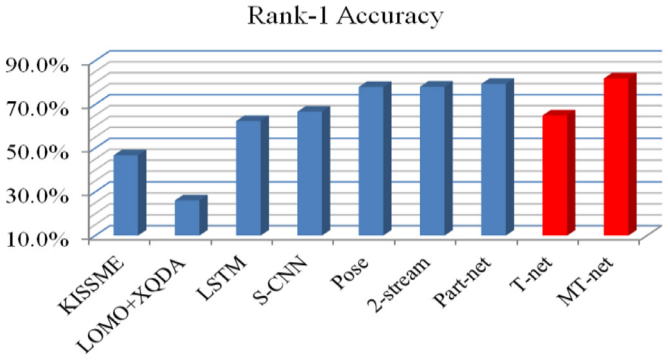

Fig. 8. Rank-1 accuracy on market1501 dataset.

When we added dropout layer after the normed feature layer, the accuracy decreased. It can be seen in Table 3. This is because our method took advantage of the depth of the neural network and the features are distributed in many layers. So the dropout strategy may cause the loss of information and affect the final result.

4.4.2. Results on Market1501 dataset

For Market1501 dataset, it's one of the largest datasets and our method has made success on this dataset. We compared our method with other state-of-the-art methods, the result are shown in Table 4. Fig. 8 is the intuitionistic rank-1 accuracy. We also introduced T-net as the baseline for comparing.

We can see that the same as CUHK03 dataset, the result of our method on Market1501 dataset is also good. Our method has the highest rank-1 accuracy and mAP. It is validated that our approach can be applied to multiple large-scale datasets.

Besides, we checkout the effects of module numbers and dropout layer may have. The comparative experiment was set on

Table 5
Different numbers of feature extraction modules on Market1501 dataset.

Num of modules	Rank1	Rank5	Rank10	Rank15	Rank20
2	74.78%	87.28%	90.63%	92.29%	93.51%
4	78.10%	89.71%	92.44%	93.87%	94.58%
6	80.37%	91.32%	93.81%	94.64%.	95.87%
8	81.95%	92.53%	94.30%	95.51%	96.20%
10	80.02%	90.91%	93.27%	94.08%	95.79%

Table 6
With/without dropout layer on Market1501 dataset.

Dropout/no dropout	Rank1	Rank5	Rank10	Rank15	Rank20
Market15 01(dropout)	25.64%	64.77%	77.46%	84.47%	89.21%
Market1501 (no dropout)	81.95%	92.53%	94.30%	95.51%	96.20%

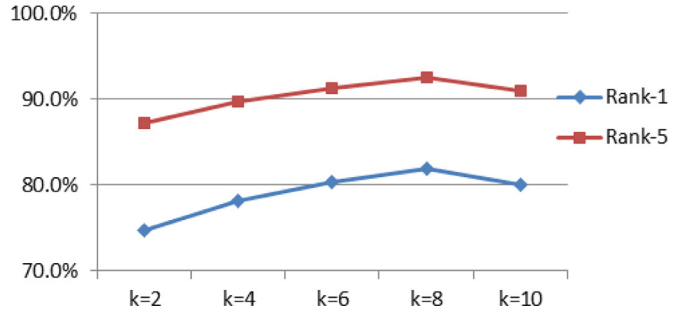

Fig. 9. Different numbers of feature extraction modules on Market1501 dataset.

Market1501 dataset. We got the best results in the case of 8 modules with no dropout layer. This is consistent with the result on CUHK03 dataset. The results are shown in Tables 5 and 6 and Fig. 9.

5. Conclusion

In this work, we proposed a novel Multilevel feature extract strategy to combine coarse and fine information from different layers, which contributes to a more robust and discriminative representation of pedestrian pictures. Further we designed an end-to-end Multilevel triplet deep learning model to compute extracted Multilevel features efficiently for person re-identification task. The experiments on two popular large-scale datasets showed that our model was efficient and significantly improved the performance compared with many existing state-of-the-art methods. The result validated that our feature extract strategy learns more useful and discriminative features and the Multilevel triplet model is effective and reliable for person re-identification task.

Acknowledgments

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the China National Natural Science Foundation under Grant No. 61673299, 61203247, 61673301, 61573259, 61573255. It was also supported by the Fundamental Research Funds for the Central Universities(Grant No. 0800219327). It was also partially supported by the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) under Grant No. MJUKF201721.

References

[1] W Chen, X Chen, J Zhang, K Huang, A multi-task deep network for person re-identification, AAAI Conference on Artificial Intelligence 1 (2) (2017) 3.
[2] Z Zheng, L Zheng, Y Yang , A discriminatively learned CNN embedding for person re-identification, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14 (1) (2017) 13.
[3] E Ahmed, M Jones, K. Marks T, An improved deep learning architecture for person re-identification, Comput. Vis. Pattern Recognit. IEEE (2015) 3908–3916.
[4] W Chen, X Chen, J Zhang, K Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, Comput. Vis. Pattern Recognit, IEEE (2017) 1320–1329.
[5] I Filkovic, Z Kalafatic, T. Hrkac, Deep metric learning for person re-identification and de-identification, Int. Convent. Inf. Commun. Technol. Electron. Microelectron. (2016) 1360–1364.
[6] H Jin, X Wang, S Liao, SZ Li, Deep person re-identification with improved embedding and efficient training. arXiv preprint arXiv:1705.03332, 2017.
[7] M Geng, Y Wang, T Xiang, Y Tian, Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244, 2016.
[8] Y Lin, L Zheng, Z Zheng, Y Wu, Y Yang, Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220, 2017.
[9] A Hermans, L Beyer, B Leibe, In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.
[10] YJ Cho, KJ Yoon, PaMM: pose-aware multi-shot matching for improving person re-identification. arXiv preprint arXiv:1705.06011, 2017.
[11] S Liao, Y Hu, X Zhu, SZ Li, Person re-identification by local maximal occurrence representation and metric learning, Comput. Vis. Pattern Recognit. IEEE (2015) 2197–2206.
[12] D Cheng, Y Gong, S Zhou, J Wang, N Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, Comput. Vis. Pattern Recognit. IEEE (2016) 1335–1344.
[13] T Kong, A Yao, Y Chen, F Sun, HyperNet: towards accurate region proposal generation and joint object detection, Comput. Vis. Pattern Recognit. IEEE (2016) 845–853.
[14] C Su, J Li, S Zhang, J Xing, W Gao, Q Tian, Pose-driven deep convolutional model for person re-identification, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 3980–3989.
[15] Z Cai, Q Fan, R S Feris, N Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: European Conference on Computer Vision, Springer, Cham, 2017, pp. 354–370.
[16] TY Lin, P Dollár, R Girshick, K He, B Hariharan, S Belongie, Feature pyramid networks for object detection, Comput. Vis. Pattern Recognit. IEEE 1 (2) (2016) 4.
[17] A Shrivastava, R Sukthankar, J Malik, A Gupta, Beyond skip connections: top-down modulation for object detection. arXiv preprint arXiv:1612.06851, 2016.
[18] D Gray, H Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision (ECCV), Marseille, France, October 12-18, 2008, Proceedings. DBLP, 2008, pp. 262–275.
[19] B Ma, Y Su, F Jurie, Local descriptors encoded by fisher vectors for person re-identification, in: International Conference on Computer Vision (ICCV), Springer-Verlag, 2012, pp. 413–422.
[20] T Matsukawa, T Okabe, E Suzuki, Y Sato, Hierarchical gaussian descriptor for person re-identification, Comput. Vis. Pattern Recognit. IEEE (2016) 1363–1372.
[21] C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, A Rabinovich, Going deeper with convolutions, Comput. Vis. Pattern Recognit. IEEE (2015) 1–9.
[22] Y Jia, E Shelhamer, J Donahue, S Karayev, J Long, R Girshick, S Guardarrama, T Darrell, Caffe: convolutional architecture for fast feature embedding, in: ACM international conference on Multimedia, 2014, pp. 675–678.
[23] J Deng, W Dong, R Socher, R Socher, J Li, K Li, F Li, ImageNet: a large-scale hierarchical image database, in: Comput. Vis. Pattern Recognit. IEEE, 2009, pp. 248–255.
[24] W Li, R Zhao, T Xiao, X Wang, DeepReID: deep filter pairing neural network for person re-identification, Comput. Vis. Pattern Recognit. IEEE (2014) 152–159.
[25] L Zheng, L Shen, L Tian, S Wang, J Wang, Q Tian, Scalable person re-identification: a benchmark, in: IEEE International Conference on Computer Vision. IEEE, 2016, pp. 1116–1124.
[26] D Gray, H Tao, Evaluating appearance models for recognition, reacquisition, and tracking, IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS). Citeseer 3 (5) (2007) 1–7.
[27] JV Davis, B Kulis, P Jain, S Sra, IS Dhillon, Information-theoretic metric learning, in: International Conference on Machine Learning, ACM, 2007, pp. 209–216.
[28] PM Roth, P Wohlhart, M Hirzer, M Kostinger, H Bischof, Large scale metric learning from equivalence constraints, in: Comput. Vis. Pattern Recognit., IEEE Computer Society, 2012, pp. 2288–2295.
[29] F Xiong, M Gou, O Camps, M Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision (ECCV), Springer International Publishing, 2014, pp. 1–16.
[30] RR Varior, M Haloi, G Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision(ECCV), Springer, Cham, 2016, pp. 791–808.
[31] W Li, R Zhao, T Xiao, X Wang, DeepReID: deep filter pairing neural network for person re-identification, Comput. Vis. Pattern Recognit. IEEE (2014) 152–159.
[32] S Chen , C Guo, J Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.
[33] Y Zhang, X Li, L Zhao, Zhang, Semantics-aware deep correspondence structure learning for robust person re-identification, in: International Joint Conference On Artificial Intelligence, AAAI Press, 2016, pp. 3545–3551.
[34] K He, X Zhang, S Ren, J Sun, Deep residual learning for image recognition, in: Comput. Vis. Pattern Recognit., IEEE, 2016, pp. 770–778.

[35] S Liao, G Zhao, V Kellokumpu, M Pietikainen, SZ Li, Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes, Comput. Vis. Pattern Recognit. IEEE (2010) 1301–1306.
[36] C Liu, H Wechsler, Independent Component Analysis of Gabor Features for Face Recognition, IEEE transactions on Neural Networks 14 (4) (2003) 919–928.
[37] L Zheng, Y Huang, H Lu, Y Yang, Pose invariant embedding for deep person re-identification. arXiv preprint arXiv:1701.07732, 2017.
[38] LM Zhao, X Li, YT Zhuang, J Wang, Deeply-learned part-aligned representations for person re-identification. arXiv preprint arXiv:1707.07256, 2017.
[39] RR Varior, B Shuai, J Lu, D Xu, G Wang, A siamese long short-term memory architecture for human re-identification, in: European Conference on Computer Vision (ECCV), Springer, Cham, 2016, pp. 135–153.
[40] L Wu, C Shen, AVD Hengel, PersonNet: Person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255, 2016.
[41] M Gou, S Karanam, W Liu, O Camps, RJ Radke, in: DukeMTMC4ReID: a large-scale multi-camera person re-identification dataset. *Comput. Vis. Pattern Recognit Workshops. IEEE* , 2017, pp. 1425–1434.
[42] Z Lai, D Mo, WK Wong, Y Xu, D Miao, D Zhang, Robust discriminant regression for feature extraction, IEEE Transactions on Cybernetics (2017) 1–13 PP(99).
[43] WK Wong, Z Lai, J Wen, X Fang Lu Y, Y Lu, Low rank embedding for robust image feature extraction, IEEE Transactions on Image Processing (2017) PP(99)1-1.
[44] J Wen, Z Lai, Y Zhan, J Cui, The L 2,1 -norm-based unsupervised optimal feature selection with applications to action recognition, Pattern Recognit. 60 (C) (2016) 515–530.
[45] WK Wong, Z Lai, Y Xu, J Wen, CP Ho, Joint tensor feature analysis for visual object recognition, IEEE Transactions on Cybernetics 45 (11) (2015) 2425.
[46] F Shen, Y Xu, L Liu, Y Yang, Z Huang, H Shen, Unsupervised deep hashing with similarity-adaptive and discrete optimization, IEEE Transactions on Pattern Analysis & Machine Intelligence (2018) PP(99)1-1.
[47] F Shen, Y Yang, L Liu, W Liu, D Tao, H Shen, Asymmetric binary coding for image search, IEEE Transactions on Multimedia (2017) PP(99)1-1.
[48] F Shen, X Zhou, Y Yang, J Song, H Shen, D Tao, A fast optimization method for general binary code learning, IEEE Transactions on Image Processing 25 (12) (2016) 5610–5621.
[49] F Shen, C Shen, Q Shi, AVD Hengel, Z Tang, H Shen, Hashing on nonlinear manifolds, in: IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 24, 2015, p. 1839.