Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

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

Partial person re-identification with two-stream network and reconstruction

Suguo Zhu[a,*], Xiaowei Gong[a], Zhenzhong Kuang[a], Junping Du[b]

[a] Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hang Zhou 310018, Zhejiang, China
[b] Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Article history:
Received 2 January 2019
Revised 10 April 2019
Accepted 15 April 2019

Keywords:
Partial person re-identification
Two-stream network
Reconstruction
Deep learning

ABSTRACT

Partial person re-identification is a challenging issue at present. However, affected by occlusions, features in person re-identification cannot be detected and the traditional person re-identification methods can not accurately deal with it. In order to solve this problem, we propose to match query and gallery by combining different modes from two-stream network with sparse reconstruction to realize partial person re-identification. For acquiring features, bilinear pooling is applied to fuse the two different modes from the appearance network and pose network aiming at better performance. For matching query and galley, the robust sparse representation reconstructs the features extracted by the network for flexible solution, using the parameters learned from galley. The reconstruction process achieves arbitrary size images in partial person re-identification. In addition, we extract mid-level feature and fuse it with the high-level feature for more accuracy. Experiments demonstrate the performance of the proposed method better compared with the methods of state-of-the-art person re-identification methods on dataset Market1501, CUHK03, DukeMTMC-reID and partial person dataset Partial-REID, Partial-iLIDS.

1. Introduction

Partial person re-identification (Partial person re-id) is an important area of cross-mode research in the field of computer vision. It is also an important application in fine-grained classification and retrieval. Person re-identification is a sub-problem of fine-grained Image and video search and retrieval [1,2], which combines fine-grained problem with retrieval problem and is studied as a new problem.

Although existing researches have achieved satisfactory results in person re-identification, there are still many problems and challenges in fine-grained classification of cross-mode problem [3,4]. For example, occlusion. It causes some parts of person body dropped. And this is very common in the wild scene in person re-identification research. Zheng et al. [5] addressed partial person re-identification problem, and proposed a matching framework. They demonstrated it on partial iLIDS [6]. Some examples of the datasets are shown in Fig. 1. The first row is images with non-partial characteristic, and the second row is images with partial characteristic. From the perspective of fine-grained visual understanding, the research of person re-id needs to employ the fine-grained characteristics of persons. And it is important to capture different modes of features and each mode can be achieved by one task. Person re-id approaches include strip-based re-id and pose-based re-id. Fig. 2 shows the traditional frameworks of the two kinds of methods.

Strip-based re-id. The common scheme was to slice the image horizontally into several horizontal strips. Since the human body is often similar in stature, a simple horizontal bar was used to make one-to-one comparisons [7–9]. For example, Varior R et al. [10] slice frames into several horizontal parts, and input them into LSTM network(Long short term memory network). The final feature fused the local features of all image blocks. However, the disadvantage of this method is that the requirement of image alignment is relatively high. If the two images were not aligned up and down, head-to-body comparison would be likely to occur, which made the model judgement not accuracy or even wrong.

Pose-based re-id. To solve the problem of invalidation of manual image slices when images were not aligned, another relatively new method was to detect parts (hands, legs, trunk, etc.) on the human body before matching them, named as pose-based person re-id, aiming at reducing the position errors, which occurs in strop-based re-id. Zhang et al. [11] employed pose invariant embedding(PIE) to handle pedestrian misalignment. To solve the same problem, Sarfraz et al. [12] incorporated both the fine and coarse pose information of the person to learn a discriminative embedding. These

* Corresponding author.
E-mail address: zsg2016@hdu.edu.cn (S. Zhu).

https://doi.org/10.1016/j.neucom.2019.04.098
0925-2312/

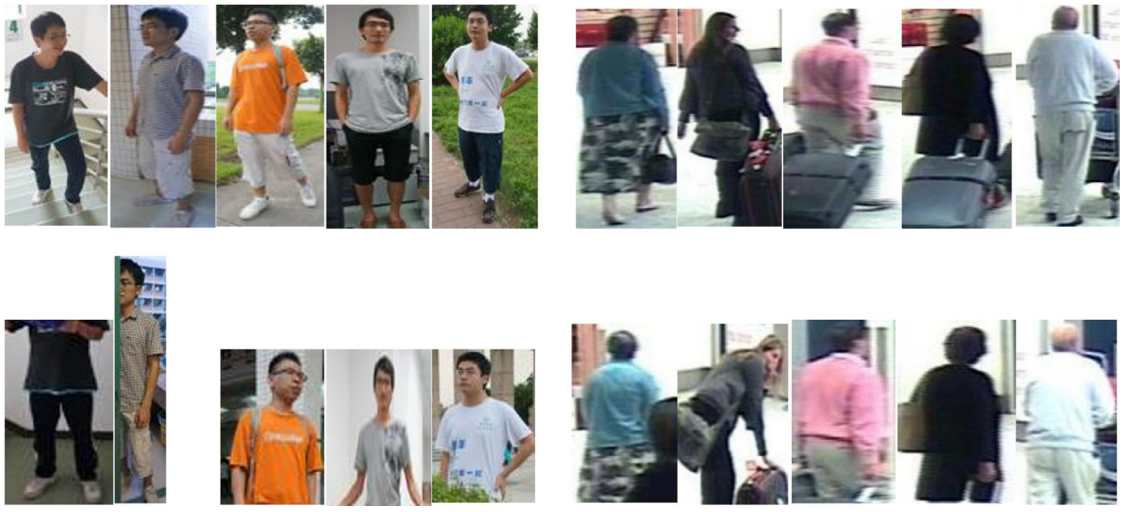


Fig. 1. Examples of two non-partial person images(first row) and the corresponding partial images (the second row).

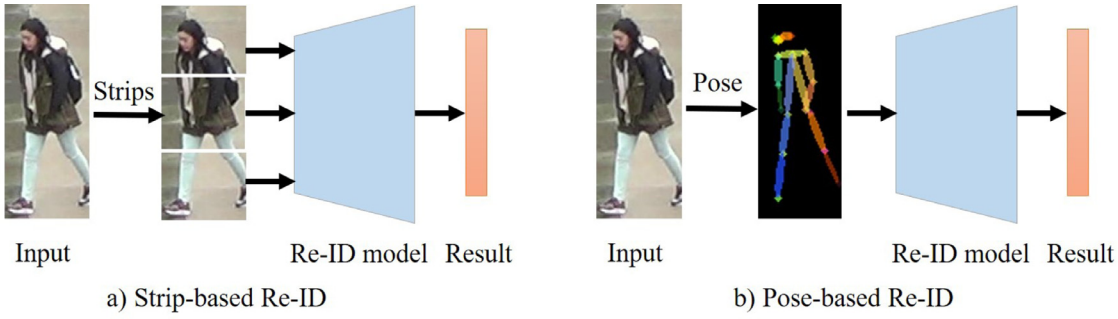


Fig. 2. The frameworks of strip-based re-id and pose-based re-id.

approaches both used the currently popular datasets, however, current standard datasets(MARKET1501 [9], CUHK03 [13], DukeMTMC-reID [14]) did not have sufficient data for pedestrian pose transformation. Liu et al. [15] proposed a method of data enhancement for the re-id model and generated much data for training by changing various postures. That was effective for data deficiencies, but the complexity of the model was very high. Although pose-based re-id approaches can be used as the fine-grained characteristics to achieve a certain degree of performance improvement, the errors of detection components may be introduced.

From the perspective of image retrieval and video retrieval [16,17], the research of person re-id approach needs to employ the idea of retrieval to find a given pedestrian from a large dataset. Multi-mode based approaches [18,19] are popularly introduced into partial person re-id, such as [20]. Following the partial-based method, researchers proposed different methods [5,21] to solve partial person re-identification problem. However, few studies have provided flexible solutions to identify a person in an image containing arbitrary part of the body. He et al. [21] employed the deep spatial feature reconstruction (DSR) method, matching a pair of person images of different sizes. It is effective for partial person re-identification. However, for the wild world, the current methods still cannot achieve the partial person re-id problem well enough on speed and in real scene.

Inspired by the above problems, we propose a novel framework in this paper, which aims at solving partial person re-identification more accurate and more effective. As illustrated in Fig. 3, the proposed framework contains the network part and the reconstruction part. The network part applies two stream networks to generate features of the queries, including appearance branch and pose branch; multi-scale pooling and bilinear method are used for better performance. The features are reconstructed using sparse representation method for searching the final image.

The main contributions of this paper are as follows. We propose a novel framework for partial person re-id with fine-grained multi-scale based feature. It is designed to combine the two-stream network with sparse reconstruction. The appearance information and pose information are extracted from two-stream network and fused by bilinear approach for better describing person. In the proposed method, not only different modes are considered, sparse reconstruction is introduced for flexible solution.

2. Our approach

As illustrated in Fig. 3, the proposed approach contains two-mode feature extraction module and feature reconstruction module. In the two-mode feature extraction part, for focussing on the fine-grained features of persons, one of the modes is the appearance feature extracted through the fusion module which takes the 2 dimension adaptive pooling with the fourth layer(middle-high level) feature of Resnet-50. The other mode is the pose features by OpenPose [22]. And bilinear pooling is employed to joint the two modes. In the feature reconstruction module, we take advantage of the robustness of sparse representation to deal with the person matching task. The core of this module is feature reconstruction with the output of the two-mode feature extraction module.

2.1. Two-mode feature extraction module

Two-mode feature extraction module is implemented through two branches: appearance branch and pose branch. The output fea-

Please cite this article as: S. Zhu, X. Gong and Z. Kuang et al., Partial person re-identification with two-stream network and reconstruction, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.04.098

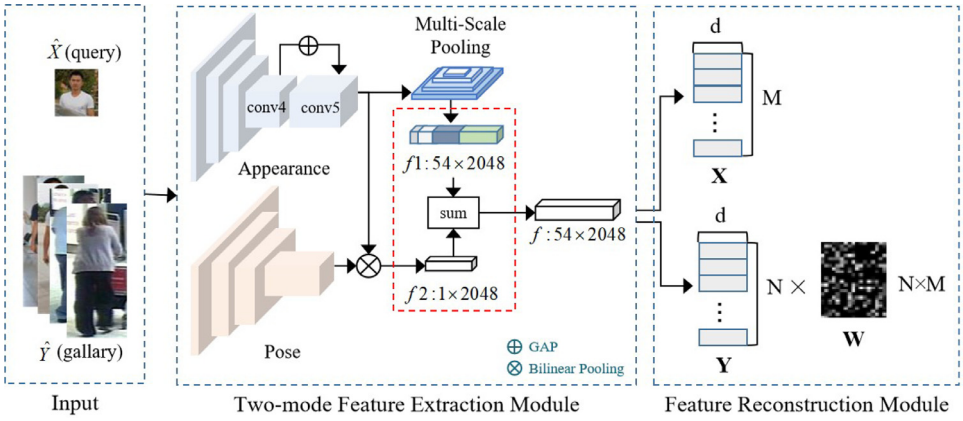


Fig. 3. Overview of the proposed framework.

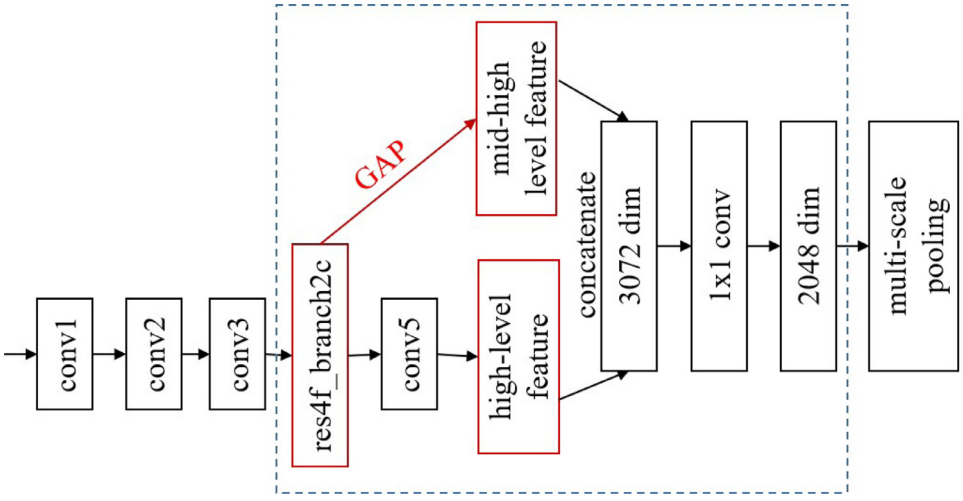


Fig. 4. Fusion of conv4 and conv5 in ResNet-50.

tures of the two branches are fused. It is important to capture the fine-grained features of one person compared with the others.

Appearance branch. In the appearance branch, we concatenate middle-high level features and high level features to extract the appearance features on the basic of ResNet-50 [23], as shown in Fig. 4 in detail. It is inspired by the fact that: the features learned in the low level are mainly the color, edge and other low-level features; the middle layer becomes more complex, learning texture features (such as mesh patterns); middle-high level features learn more distinctive features (such as backpack); the high level features are complete and have the key distinguishing features. We mainly fuse the features of middle-high level (the output of res4f_branch2c layer in ResNet-50) into the features of high level. The purpose is to use the backpack, shoes or other real information to assistant models concerning on middle-high level to better identify the right pedestrians, when the input images is affected by blur or illumination, etc.

As described in Fig. 4, the output of res4f_branch2c layer(or conv4) in ResNet-50 is used to extract the middle level feature(including more detailed information of one person) and high level feature(including more semantic information of one person) for better performance. GAP operation is the 2D adaptive average pooling, and it is set to capture the features. The feature size of the res4f_branch2c layer output is $16\times8\times1024$, and after GAP, the size of mid-high level feature is $16\times8\times1024$. The feature size of the fifth layer(conv5) is $16\times8\times2048$. We take the concoction and convolution operation to get the fine-grained feature of the appearance branch.

For the purpose of different size of the input images, we obtain multi-scale features through multi-scale pooling while the extracted image features have better scale invariance. We set four different scales, including $2\times2/stride=2$, $6\times6/stride=2$, $4\times4/stride=4$ and $8\times8/stride=8$. The appearance feature 16×8 times different scales of multi-scale pooling, separately and the spatial features of 8×4, 6×2, 4×2 and 2×1 large span are obtained. Subsequently, we concatenate all the spatial features together, and the new feature is described as $f1$ with the size of 54×2048. At the same time, feature fusion with different scales is more likely to capture all the features. This is more consistent with the logic of person watching the world. When viewing an image, it is possible to resize different size of object for recognition. Since only one forward convolution operation is required for an image, it improves the efficiency of the proposed model.

Pose branch. As one of the mode of two-mode feature extraction module, pose of one person describe the attention feature which is a different scale to distinguish different persons. In terms of the speed of the model, we use OpenPose [22] to train the parameter matrix in the initialization stage, by extracting pose features in Pose branch, with the purpose of assisting the appearance features to better define the body parts. However, it is not enough for pose

Please cite this article as: S. Zhu, X. Gong and Z. Kuang et al., Partial person re-identification with two-stream network and reconstruction, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.04.098

information to identify one person. We aggregate the output features of appearance branch and pose branch to obtain $f2$ feature by bilinear pooling, and the size of $f2$ is 1×2048.

2.2. Feature reconstruction module

Seeking for better performance, sparse representation is taken advantage of to deal with the person matching task. Sparse coding is robust to match images [24,25]. It calculates the distance between the query and gallery images. Through the comparation of similarities, the most similar image is found.

Let $\hat{X}$ and $\hat{Y}$ denote the query image and one image of the gallery respectively. We input $\hat{X}$ and $\hat{Y}$ respectively to the two-mode feature extraction module, and the output features X and Y are obtained. If $X = \{x_1, \ldots, x_M\} \in R^{d \times M}$, $Y = \{y_1, \ldots, y_N\} \in R^{d \times N}$, $x_i (i = 1, 2, \ldots, N)$ can be represented by linear combination of Y. Therefore, we want to find the minimum linear representation coefficients $w_i (i = 1, 2, \ldots, M)$ of $x_i (i = 1, 2, \ldots, N)$ with respect to Y and we constrain w_i using $l2 - norm$. Then, the objective function can be written as follows:

$$\min_{w_i} \|Yw_i - x_i\|_2^2 + \gamma \|w_i\|_1$$

where $W = \{w_1, \ldots, w_M\} \in R^{N \times M}$ is the dictionary and is trained on dataset Market1501 [9]. The distances between query image and gallery can be described as $\|Yw_i - x_i\|_2^2$. Followed with He et al. [21], we also fix the parameter γ as 0.4, and the size of dictionary W after being trained is 54×54. We take control of the sparsity of the task by the parameter γ, and $\|w_i\|_1$ means l_1. Although sparse reconstruction is an effective method to calculate two images of different sizes, its disadvantage is obvious that it is time-consuming. Therefore, our next work intends to find a time-saving method to replace sparse reconstruction.

2.3. Implementation details

We implement the proposed model using PyTorch framework. Our model is trained and tested on Linux with NVIDIA TITAN Xp. The backbone network in appearance branch is the ResNet-50 [23] model pre-trained on ImageNet [26]; in pose branch, the network is OpenPose [22] pre-trained parameter matrix for initialization. In the actual training process, only the network of appearance branch is gradient propagated, so the human key point datasets is not required. The input image is re-scaled to 256×128, and the stride of the last layer is set to 1 in appearance branch. Thus spatial features with size of 16×8 are generated by appearance branch. At the same time, after the global average pooling(GAP), features of the fourth layer(middle-high level) is with size of $16 \times 8 \times 1024$ and concatenated with features of the fifth layer(high level) the size of which is $16 \times 8 \times 2048$, the size of the new feature is $16 \times 8 \times 3072$. Finally, by the 1×1 convolution layer, the size of $f1$ becomes $16 \times 8 \times 2048$. The pose feature is $16 \times 8 \times 128$ extracted by pose branch through OpenPose network. After bilinear pooling [27], the features of appearance branch and pose feature are aggregated into $f2$ with the size of $16 \times 8 \times 2048$. Then, $f2$ is added by $f1$ and the final feature f with size of 54×2048 is generated.

What's more, the mini-batch size is set to 33, in which each identity has 4 images. In addition, we use the Adam [28] optimizer with the default hyper-parameters to minimize the network; and triplet hard loss [29] margin is set to 0.3 based on experience. The initial learning rate is $1.5 \times 10 - 4$, and the learning rate decreases at epoch 151 to $1.5 \times 10 - 5$ (weight decay is set to 0.1), and the total training takes 400 epochs when the model outputs the best results.

Algorithm 1 Framework of the proposed algorithm.

Input:
A probe person image p of an arbitrary-size; a gallery person image q;
Output:
Parameter θ;
1: **for** each $t \in [1, T]$ **do**
2: Extract appearance feature a through the appearance branch and pose feature b through the pose branch.
3: Through multi-scale pooling, feature $f1$ is obtained from a; and feature $f2$ is obtained from b by bilinear pooling.
4: Extract probe feature maps X and gallery feature maps Y.
5: Divide X and Y into multiple blocks: $X = \{x_1, \ldots, x_n\} \in R^{d \times M}$ and $Y = \{y_1, \ldots, y_n\} \in R^{d \times N}$.
6: Obtain the sparse reconstruction coefficient matrix $W = \{w_1, \ldots, w_n\} \in R^{N \times M}$.
7: Obtain the sparse reconstruction score.
8: Compute the sparse reconstruction error.
9: Update the sparse reconstruction coefficient matrix W.
10: Update the gradients of $L - th$ with respect to X and Y.
11: Update the parameters θ.
12: **end for**

Table 1
Experimental datasets introduction.

Datasets	Publisher	Cameras	Identities	Images
Market1501(holistic)	Zheng et al. [9]	6	1501	32,668
CUHK03(holistic)	Li et al. [13]	6	1467	13,164
DukeMTMC-reID(holistic)	Zheng et al. [14]	8	1812	36,441
Partial-REID(partial)	Zheng et al. [5]	–	60	600
Partial-iLIDS(partial)	Zheng et al. [6]	–	119	476

3. Experiments and analysis

3.1. Datasets

We train a two-stream network end-to-end on two partial ReID datasets, including Partial-REID [5] and Partial-iLIDS [6] with pre-trained model in the Market1501 [9] dataset. Furthermore, in order to verify the universality of the algorithm proposed in this paper, we also test the proposed model on three holistic person datasets, including Market1501, CUHK03 [13] and DukeMTMC-reID [14]. Table 1 shows the details of the datasets used in this paper.

Partial-REID contains 600 images of 60 people, with 5 partial images and 5 full-body images per person. In this dataset, we use the single-shot, where there is only one instance of the gallery images in each identity. Partial-iLIDS is a simulated partial person datasets based on i-LIDS [6]. There are 119 people with total 476 person images captured by multiple non-overlapping cameras. Different from most other datasets, the original images have fair amount of occlusion, sometimes rather severe, caused by people and luggage. Market1501 contains 32,668 annotated bounding boxes of 1501 person identities captured by six cameras(including 5 high-resolution cameras, and one low-resolution camera) in front of a supermarket in Tsinghua University. The training set contains 12,936 images from 751 identities and testing set contains the other 19,732 images from 750 identities. CUHK03 contains 14,096 images from 1467 person identities captured by two different cameras in the CUHK campus. The average person has 9.6 pieces of training data. In this paper, we use the bounding boxes detected from deformable part models (DPM) version. DukeMTMC-reID contains 36,411 images from 1812 person identities captured by eight different high-resolution cameras. There are 1404 identities appear in more than two cameras and the other 408 identities are

Please cite this article as: S. Zhu, X. Gong and Z. Kuang et al., Partial person re-identification with two-stream network and reconstruction, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.04.098

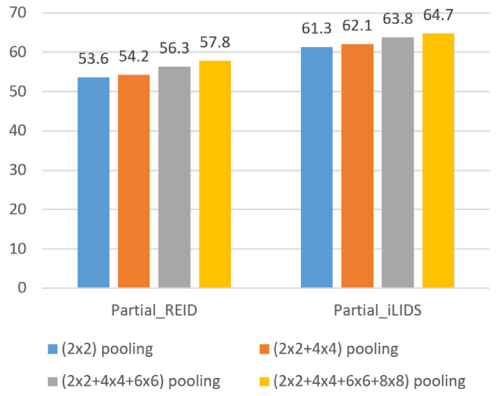


Fig. 5. Rank-1 accuracy of our approach with multi-scale pooling representation.

Table 2
Accuracy comparison on Partial-REID and Partial-iLIDS with single query.

Method	Partial-REID			Partial-iLIDS		
	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$
Resizing model	19.3	40.0	51.3	24.3	52.3	61.3
SWM [5]	33.3	52.0	62.0	36.0	60.0	70.7
AMC [5]	43.0	75.0	76.7	21.9	43.7	55.5
AMC+SWM [5]	33.6	53.8	63.3	46.7	69.6	81.9
DSR [21]	49.5	72.7	84.7	54.5	73.1	85.7
Ours(bilinear)	56.1	78.8	85.8	57.1	80.2	87.8
Ours(bilinear+middle-high)	61.3	75.6	84.8	63.0	78.9	87.3

regarded as distractors. Training set contains 702 identities and testing set contains the rest 702 identities, with an average of 23.5 pieces of training data for each person.

3.2. Evaluation metrics

We use both the cumulative matching characteristics(CMC) and mean average precision(mAP) to evaluate the proposed method. They are the most popular performance evaluation methods in the field of person re-id. The CMC score measures the quality of identifying the correct match at each rank. For multiple ground truth matches, CMC cannot measure how well all the images are ranked. Therefore, we report the mAP scores for Market-1501, CUHK03 and DukeMTMC-reID, where more than one ground truth images are in the gallery.

3.3. Performance on partial_REID and partial_ iLIDS

We employ different scales pooling method to evaluate the performance of the proposed approach with regard to the multi-scale pooling. As shown in Fig. 5. Four different scales are considered for the representation fusion. Four different fusion ways are adopted: (2×2) pooling, $(2\times2+4\times4)$ pooling, $(2\times2+4\times4+6\times6)$ pooling and $(2\times2+4\times4+6\times6+8\times8)$ pooling. The experimental results show that the performance of multi-scale pooling achieves best when probe representation is pooled by $(2\times2+4\times4+6\times6+8\times8)$ pooling.

3.4. Experimental results and analysis

Experiments on Partial-REID and Partial-iLIDS. In Table 2, we compare the current state-of-the-art partial ReID methods Resizing model(The size of the input image is resized as 320×120), SWM [5], AMC [5], AMC and SWM [5] and DSR [21] for Partial-REID and Partial-iLIDS datasets in single query. Compared to the DSR method, our method achieves the improvement of 14.1% and 8.5% (rank=1 accuracy) in Partial-REID and Partial-iLIDS datasets, respectively. The models with middle-high level features were improve 1% and 1.7% (rank=1 accuracy) than that without concatenating middle-high level features in Partial-REID and Partial-iLIDS datasets, respectively. Therefore, the model accuracy has a lot of room to improve, and the performance improvement brought by the fusion of middle-high level is obvious. It is shown that the method of concatenating middle-high level plays a certain role in Partial-REID and Partial-iLIDS datasets.

Table 3
Accuracy comparison on Market1501 with single query and multiple query (r=1: rank=1 and mAP: mean average precision).

Method	Single query				multiple query			
	mAP	$r=1$	$r=5$	$r=10$	mAP	$r=1$	$r=5$	$r=10$
Spindle [31]	–	76.9	91.5	94.6	–	–	–	–
PIE [11]	56.0	79.3	90.8	94.4	–	–	–	–
SVDNet [16]	62.1	82.3	92.3	95.2	–	–	–	–
DSR [21]	64.2	83.5	–	–	–	–	–	–
Transfer [32]	65.5	83.7	–	–	73.8	89.6	–	–
APR [33]	64.7	84.3	93.2	95.2	–	–	–	–
XQDA+SSM [34]	68.8	82.2	–	–	76.2	88.2	–	–
TriHard [29]	69.1	84.9	94.2	–	76.4	90.5	96.3	–
CamStyle [30]	68.8	87.7	–	–	77.1	91.7	–	–
Ours(bilinear)	75.6	90.2	96.4	97.9	82.5	93.0	97.5	98.5
Ours(bilinear+middle-high)	77.0	90.5	96.6	97.7	83.2	93.3	97.9	98.6

Table 4
Accuracy comparison on CUHK03 with detected dataset.

Method	mAP	**$r=1$**	**$r=5$**	**$r=10$**
Pose-transfer [15]	28.2	30.1	–	–
PSE [12]	30.2	27.3	–	–
HA-CNN [36]	38.6	41.7	–	–
MGCAM [37]	46.8	46.7	–	–
DML [8]	–	52.1	84.0	92.0
LSTM Simaese [10]	–	57.3	80.1	88.3
PCB [17]	57.5	61.3	–	–
LDN [35]	–	62.6	90.0	94.8
Ours(bilinear)	57.5	61.7	80.1	86.4
Ours(bilinear+middle-high)	59.6	64.2	81.4	88.5

Experiments on Market1501. Table 3 shows the comparison over two query schemes, single query and multiple query. Single query means that only one image per person exists in the probe set, whereas multiple query means that multiple images per person exist in the probe set. For the multiple-query setting, one descriptor is obtained from multiple images by averaging the feature from each image. Compared to the CamStyle [30] method, our method achieves the improvement of 8.2% mAP and 2.8% rank=1 accuracy in single query. The models with middle-high level features were improve 1.4% mAP and 0.3% rank=1 accuracy than without concatenate middle-high level features in single query, improve 0.7% mAP and 0.3% rank=1 accuracy in multiple query. Except the rank=10 of ours(bilinear) approach, our concatenate middle-high level method has achieved a competitive result in both single query and multiple query. Therefore, it is shown that the method of concatenate middle-high level it helps in Market1501 dataset.

Experiments on CUHK03. Table 4 shows the results of our method on the CUHK03 dataset in detail. Compared with the methods in the table, our method achieves 59.6% and 64.2% performance on mAP and rank=1 accuracy, but there is still a certain gap between LDNS [35] methods on rank=5 and rank=10. The models with middle-high level features were improve 2.1% mAP and 2.5% rank=1 accuracy than without concatenate middle-high level features in single query. Therefore, it is shown that the method of concatenate middle-high level plays a certain role in CUHK03 dataset.

Please cite this article as: S. Zhu, X. Gong and Z. Kuang et al., Partial person re-identification with two-stream network and reconstruction, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.04.098

Fig. 6. Examples of non-partial person images and the corresponding partial images and the visualizations of their feature maps from appearance branch and pose branch of our framework. (first row shows the non-partial persons and the second row shows the corresponding partial images).

Table 5
Accuracy comparison on DukeMTMC-reID.

Method	mAP	$r=1$	$r=5$	$r=10$
DCGAN [14]	47.1	67.7	–	–
APR [33]	51.9	70.7	–	–
ACRN [38]	52.0	72.6	–	–
SVDNet [16]	56.8	76.7	86.4	89.9
PN-GAN [39]	60.6	79.2	–	–
Random Erasing [40]	62.4	79.3	–	–
Ours(bilinear)	66.4	82.2	92.0	94.3
Ours(bilinear+middle-high)	67.1	83.1	92.2	94.7

Experiments on DukeMTMC-reID. Table 5 shows the results of our method on the DukeMTMC-reID dataset in detail. Compared with the other methods in the table, our method achieves the best performance 67.1%, 83.1%,92.2% and 94.7% on mAP, rank=1, rank=5 and rank=10. The models with middle-high level features were improve 0.7% mAP and 0.9% rank=1 accuracy than without concatenate middle-high level features in single query, but it is not obvious in rank=5 and rank=10. We find that the same problem exists in the Market1501 dataset. Now we analyze this reason: the features learned in the low level are mainly the color, edge and other low-level features; The middle layer becomes more complex, learning texture features (such as mesh patterns); Middle-high level features learn more distinctive features (such as backpack); The high level features are complete and have the key distinguishing features. We mainly fusion the features of middle-high level into the features of high level. The purpose is to use the backpacks, shoes or other information assistant models concerned by middle-high level to better identify the right pedestrians when the input images is affected by blur or illumination, etc. Whereas, when the model accuracy is very high, the performance improvement brought by the fusion of middle-high level is not particularly obvious.

4. Conclusions

In this paper, we propose a novel framework aiming at partial person re-identification more accurate, including two-mode feature extraction module and feature reconstruction module. Two-stream network and sparse reconstruction are combined for the searching. Firstly, in two-mode feature extraction module, we employ the two-stream network to generate features of images. For better representation of the features, we fuse middle-high layer feature to high layer and apply multi-scale pooling to capture different scale features. The pose feature is fused with the feature extracted from the appearance branch. Through fusion of the two mode information (appearance feature and pose feature), the more accurate and invariance feature is obtained. For searching the final image, spares representation is employed to reconstructed the images. The experimental results show better performance compared with other methods on partial person re-identification problem.

Declaration of interest

None.

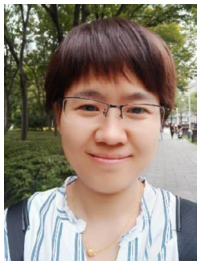
Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61472110 and Grant 61772161.

References

[1] X.S. Wei, J.H. Luo, J. Wu, Z. Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, IEEE Trans. Image Process. 26 (6) (2017) 2868–2881, doi:10.1109/TIP.2017.2688133.
[2] J. Yu, Z. Kuang, B. Zhang, Z. Wei, J. Fan, Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing, IEEE Trans. Inf. Forensics Secur. (2018) 1317–1332, doi:10.1109/TIFS.2017.2787986.
[3] X. Yang, W. Liu, D. Tao, J. Cheng, Canonical correlation analysis networks for two-view image recognition, Inf. Sci. Int. J. (2017) 338–352, doi:10.1016/j.ins.2017.01.011.
[4] W. Liu, X. Yang, D. Tao, J. Cheng, Y. Tang, Multiview dimension reduction via hessian multiset canonical correlations, Inf. Fusion 41 (2017) S1566253517300519, doi:10.1016/j.inffus.2017.09.001.
[5] W. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, S. Gong, in: Partial person re-identification, Proceedings of the IEEE International Conference on Computer Vision (2015) 4678–4686, doi:10.1109/ICCV.2015.531.
[6] W. Zheng, S. Gong, T. Xiang, in: Person re-identification by probabilistic relative distance comparison, CVPR 2011 IEEE (2011) 649–656, doi:10.1109/CVPR.2011.5995598.
[7] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, Comput. Vis. Pattern Recognit. (2015) 2197–2206, doi:10.1109/CVPR.2015.7298832.
[8] D. Yi, Z. Lei, S. Liao, S.Z. Li, in: Deep metric learning for person re-identification, 2014 22nd International Conference on Pattern Recognition. IEEE (2014) 34–39, doi:10.1109/ICPR.2014.16.
[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, in: Scalable person re-identification: A benchmark, Proceedings of the IEEE International Conference on Computer Vision (2015) 1116–1124.
[10] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, in: A siamese long short-term memory architecture for human re-identification, European Conference on Computer Vision (2016) 135–153, doi:10.1007/978-3-319-46478-7_9.
[11] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person re-identification, IEEE Trans. Image Process PP (99) (2017), doi:10.1109/TIP.2019.2910414.
[12] M.S. Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, Comput. Vis. Pattern Recognit. (2018) 420–429, doi:10.1109/CVPR.2018.00051.
[13] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2018) 1, doi:10.1109/TCSVT.2018.2873599.
[14] Z. Zheng, L. Zheng, Y. Yang, in: Unlabeled samples generated by gan improve the person re-identification baseline in vitro, International Conference on Computer Vision (2017) 3774–3782, doi:10.1109/ICCV.2017.405.
[15] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, in: Pose transferrable person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 4099–4108.
[16] Y. Sun, L. Zheng, W. Deng, S. Wang, in: Svdnet for pedestrian retrieval, International Conference on Computer Vision (2017) 3820–3828, doi:10.1109/ICCV.2017.410.

[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling, Comput. Vis. Pattern Recognit. (2017), doi:10.1007/978-3-030-01225-0_30.
[18] J. Yu, B. Zhang, Z. Kuang, L. Dan, J. Fan, Iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, IEEE Trans. Inf. Forensics Secur. (2017) 1005–1016, doi:10.1109/TIFS.2016.2636090.
[19] J. Yu, C. Hong, Y. Rui, D. Tao, Multi-task autoencoder model for recovering human poses, IEEE Trans. Ind. Electron. (2018) 5060–5068, doi:10.1109/tie.2017.2739691.
[20] W. Chen, X. Chen, J. Zhang, K. Huang, in: A multi-task deep network for person re-identification, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) (2017) 3988–3994.
[21] L. He, J. Liang, H. Li, Z. Sun, Deep spatial feature reconstruction for partial person re-identification: alignment-free approach, Comput. Vis. Pattern Recognit. (2018) 7073–7082, doi:10.1109/CVPR.2018.00739.
[22] Z. Cao, T. Simon, S. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, Comput. Vis. Pattern Recognit. (2017) 1302–1310, doi:10.1109/CVPR.2017.143.
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Comput. Vis. Pattern Recognit. (2016) 770–778, doi:10.1109/CVPR.2016.90.
[24] W. Liu, Z.J. Zha, Y. Wang, L. Ke, D. Tao, P-laplacian regularized sparse coding for human activity recognition, IEEE Trans. Ind. Electron. (2016) 5120–5129, doi:10.1109/TIE.2016.2552147.
[25] W. Liu, X. Ma, Y. Zhou, D. Tao, J. Cheng, P-laplacian regularization for scene recognition, IEEE Trans. Cybernet. (2018) 1–14, doi:10.1109/TCYB.2018.2833843.
[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, doi:10.1007/s11263-015-0816-y.
[27] Y. Gao, O. Beijbom, N. Zhang, T. Darrell, Compact bilinear pooling, Comput. Vis. Pattern Recognit. (2016) 317–326, doi:10.1109/CVPR.2016.41.
[28] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, 2015.
[29] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, IEEE Trans. Image Process 28 (9) (2019) 4500–4509, doi:10.1109/TIP.2019.2910414.
[30] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, in: Camera style adaptation for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 5157–5166.
[31] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion (2017) 907–915. doi:10.1109/CVPR.2017.103.
[32] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, T. Xiang, Deep transfer learning for person re-identification, in: IEEE International Conference on Multimedia Big Data, 2018.
[33] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161, doi:10.1016/j.patcog.2019.06.006.
[34] S. Bai, X. Bai, Q. Tian, in: Scalable person re-identification on supervised smoothed manifold, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 2530–2539.
[35] L. Zhang, T. Xiang, S. Gong, in: Learning a discriminative null space for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 1239–1248.
[36] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, Comput. Vis. Pattern Recognit. (2018) 2285–2294, doi:10.1109/CVPR.2018.00243.
[37] C. Song, Y. Huang, W. Ouyang, L. Wang, in: Mask-guided contrastive attention model for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 1179–1188.
[38] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information (2017) 1435–1443. doi:10.1109/CVPRW.2017.186.
[39] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations (2017) 2590–2600. doi:10.1109/ICCVW.2017.304.
[40] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation (2017) arXiv preprint arXiv:http://1708.04896.



Suguo Zhu received the B.S. degree in School of Computer Science and Technology from Henan Normal University, Master's Degree from School of Computer from Gugangzhou University of Technology and Ph.D. Degree in School of Compute Science and Technology from Beijing University of Posts and Telecommunications. Now she is working at School of Computer Science and Technology, Hangzhou Dianzi University. Her research interests include computer vision and machine learning, including person re-identification, action detection and recognition and object tracking.

Xiaowei Gong now he is a graduate student of the Media Intelligence Lab in Hangzhou Dianzi University, Hangzhou, China. His research interests are in Computer Vision, focusing on applications of deep learning to person re-identification and action recognition/detection in video.

Zhenzhong Kuang is currently with School of Computer Science and Technology, in Hangzhou Dianzi University (HDU). He received his master degree and Ph.D degree in computer science from China University of Petroleum. He was a visiting Ph.D Candidate in University of North Carolina at Charlotte. He mainly applies machine learning techniques to computer vision problems. His research interests include, deep learning, large-scale image classification/retrieval, image privacy protection and 3D shape retrieval.

Junping Du received the Ph.D. degree in computer science from the University of Science and Technology Beijing. She held a Post-Doctoral Fellowship with the Department of Computer Science, Tsinghua University. She is currently a Professor with the School of Computer Science, Beijing University of Posts and Telecommunications. Her current research interests include artificial intelligence, data mining, intelligent management system development, and computer applications.

Please cite this article as: S. Zhu, X. Gong and Z. Kuang et al., Partial person re-identification with two-stream network and reconstruction, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.04.098