Fusing Local and Global Features for Person Re-identification Using Multi-stream Deep Neural Networks

Mahmoud Ghorbel[1,3(✉)], Sourour Ammar[1,2(✉)], Yousri Kessentini[1,2(✉)], Mohamed Jmaiel[1,3(✉)], and Ahmed Chaari[4(✉)]

[1] Digital Research Center of Sfax, 3021 Sfax, Tunisia
[2] MIRACL Laboratory, Sfax University, Sfax, Tunisia
{sourour.ammar,yousri.kessentini}@crns.rnrt.tn
[3] ReDCAD Laboratory, Sfax University, Sfax, Tunisia
mohamed.jmaiel@redcad.org
[4] Anavid, Paris, France
ahmed.chaari@anavid.co

Abstract. The field of person re-identification remains a challenging topic in video surveillance and public security because it is facing many problems related to the variations of the position, background and brightness scenes. In order to minimize the impact of those variations, we introduce in this work a multi-stream re-identification system based on the fusion of local and global features. The proposed system uses first a body partition segmentation network (SEG-CNN) to segment three different body regions (the whole body part, the middle and the down body parts) that will represent local features. While the original image will be used to extract global features. Second, a multi-stream fusion framework is performed to fuse the outputs of the individual streams and generate the final predictions. We experimentally prove that the multi-stream combination method improves the recognition rates and provides better results than classic fusion methods. In the rank-1/mAP, the improvement is of $7,24\%/9,5$ for the Market-1501 benchmark dataset.

Keywords: Person re-identification · Semantic segmentation · Multi-steam fusion · CNN

1 Introduction

In recent years, person re-identification (re-ID) has become increasingly popular due to its important applications in many real scenarios such as video surveillance [2], robotics [23] and automated driving. Person re-ID achieves constant improvements in recent years thanks to the considerable progress of the deep learning techniques. It can be seen as an image matching problem [5]. The challenge is to match two images of the same person coming from non-overlapping camera views [7]. Despite the advances in this field, many challenges still remain

C. Djeddi et al. (Eds.): MedPRAI 2020, CCIS 1322, pp. 73–85, 2021.
https://doi.org/10.1007/978-3-030-71804-6_6

such as complex environment, various body poses [4], occlusion [10], background variations [6,20], illumination variations [11] and different camera views [13]. Two images of the same person can present many differences related to the variation of the background, or two different persons can be captured under the same background.

Recent re-identification approaches are generally based on deep learning techniques. They learn Convolutional Neural Networks (CNN) to extract global features from the whole image, without differentiating the different parts of the persons to be identified. Some recent works in the literature has shown that it is critically important to examine multiple highly discriminating local regions of the person's images to deal with large variances in appearance. Authors in [24] proposed a network that generate saliency maps focusing on one part and suppressing the activations on other parts of the image. In [21], the authors introduced a gating function to selectively underline such fine common local patterns by comparing the mid-level features across pairs of images. The network proposed in [3] is a multi-channel convolutional network model composed by one global convolutional layer which is divided into four equal local convolutional layers. Authors in [18] proposed a part-aware model to extract features from four body parts and generate four part vectors. [15] presented a multi-scale context-aware network to learn features from the full body and body parts in order to capture the local context knowledge. The work [12] proposed a multi-stream network that fuses feature distance extracted from the whole image, three body regions with efficient alignment and one segmented image. Besides the use of global and local features proposed by the papers cited above, we have showed in [6] that we can improve the person re-identification performance by background subtraction. Authors in [20] proposed a person-region guided pooling deep neural network based on human parsing maps in order to avoid the background bias problem. In addition to the way of extracting the characteristics from the image, some work has been devoted to the re-ranking process by considering the person re-identification as a retrieval process. The work presented in [17,27] demonstrated that re-ranking process can significantly improve the re-ID accuracy of multiple baseline methods.

In this work, we propose a multi-stream person re-ID method that combines multiple body-part streams to capture complementary information. We propose first to extract person body parts using a semantic segmentation network based on CNN. Second, we propose to fuse the predicted confidence scores of multiple body-part stream re-ID CNNs. In order to extract visual similarities from the different body parts, four streams are combined: the first one deals with the original image to extract global features while the second stream exploit a background subtracted image to avoid the background bias problem. Stream three and stream four takes as input respectively a segmented image that contains the middle and the down part of the body in order to focus on local features. To evaluate our proposed method, we conducted experiments on the large benchmark dataset Market-1501 [25].

2 Overview of the Proposed Method

In this section, we provide the details of the proposed method. Figure 1 shows a flowchart of the whole framework. Our approach can be divided into three separated stages. The first stage is the semantic segmentation, which consists in extracting three images of the person body parts and an image without background from the original image. The second stage is about the feature extraction and re-ID on different streams witch takes as input four images of the same person (original image and three segmented images) and yields as output four similarity measures vectors. In the end, the third stage is the combination of the four stream outputs.

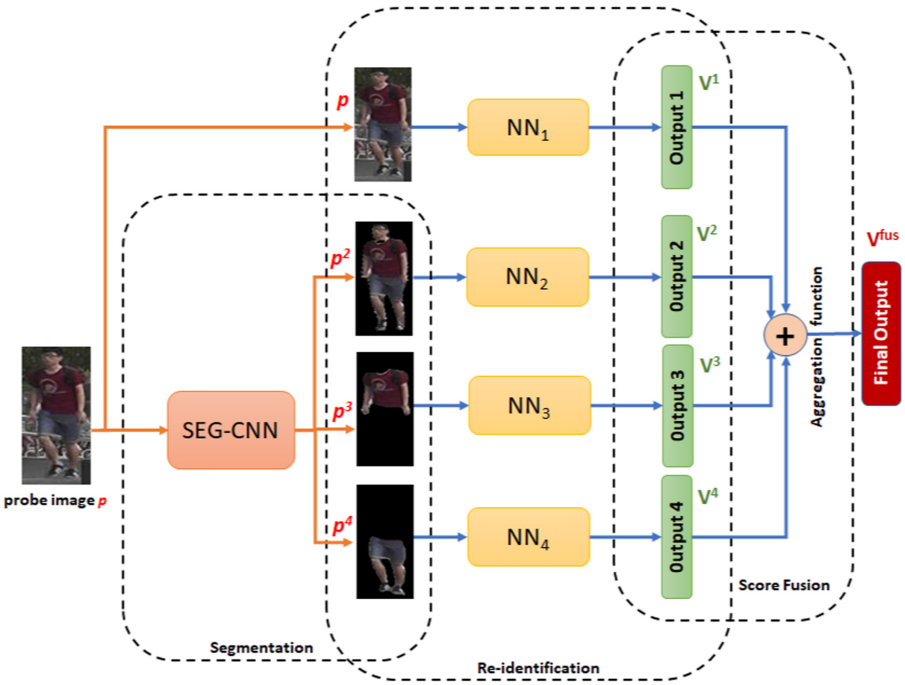

Fig. 1. Overview of the proposed multi-stream feature similarity fusion method. (Images used are from the dataset Market-1501)

2.1 Segmentation Step

The use of the segmentation in this work has two advantages. First, it reduces the effect of the background variation due to the person posture and the multitude of cameras. In some cases, two images of the same person taken by two different cameras may present some difficulties caused by the background variation. The

first image can be with a light background while the second can be with a dark background, even the texture can be different. We propose to deal with this problem by the background subtraction. Second, segmentation allows us to extract information from local parts of the image that is complementary to the global information extracted from the original image. Rather than using boxes to detect the body parts [12], we use the semantic segmentation to remove the background and to create specific body parts. The goal is to extract information only from the person body and not from the background which still present inside the box and can have an influence on the re-ID decision [20].

Our segmentation network (SEG-CNN) is inspired by the work proposed in [16]. SEG-CNN is a deep residual network based on ResNet-101 with atrous spatial pyramid pooling (ASPP) in the output layer so that it improve the segmentation and make it stronger. Moreover, in the attention of generating the context utilized in the refinement phase, two convolutions Res-5 are following. A structure-sensitive loss and a segmentation loss are also utilized to compute the similarity between the ground truth and the output image. Unlike [16] which aims to segment the person into 19 body parts and to predict the pose, we have changed the network architecture to obtain only 3 body parts: top, middle and down (See Fig. 3). From one person image, we get three different images, each image containing a part of the person body. In addition, we use those 3 parts to create a binary mask which is used to remove the background from the original image.

Since the Market-1501 dataset is not annotated for the segmentation field, SEG-CNN trained on a dataset named Look Into Person (LIP) [16] which is made for human parsing segmentation, was used to segment the Market-1501 dataset. Results on Market-1501 dataset are not satisfactory either as multiple images are over-segmented (see Fig. 2 on the right). To overcome this problem, we have visually chosen the well segmented images to create a Market-1501 sub-dataset which is then used to fine-tune our SEG-CNN.

Thanks to the fine-tuning, the segmentation results with SEG-CNN were considerably improved compared to the segmentation results obtained when the network was trained on only the LIP dataset. Figure 2 shows the improvement in the segmentation of SEG-CNN after it was fine-tuned on the Market-1501 sub-dataset.

2.2 Multi-stream Feature Extraction Method

We propose in this work to combine global and local features to enhance the re-ID performances. Our multi-stream feature extraction module is composed of four branches as shown in Fig. 1. The first branch is used to process the whole image to obtain global features (called *Full* stream). The second branch processes the background subtracted image (called *No_bk* stream) to focus only on the body part. This allows us to deal with the background variation problem. For local features, we use the last two branches that focus each on a segmented image (called *Mid* and *Dwn* streams respectively for middle and lower body parts). We empirically prove that the top body part don't contribute on the

Fig. 2. Qualitative comparison between the segmentation of SEG-CNN when it was trained on only the LIP dataset (image on the right) and SEG-CNN when it was fine-tuned on the Market-1501 sub-dataset (image on the left).

Fig. 3. Sample image of the semantic segmentation of SEG-CNN. (from the right to the left: original image, down part image, middle part image and top part image)

improvement of the final fusion score since the top part only contains the person face which is not clear because the resolution of the images is low.

Given a probe image p and a gallery set G with N images, with $G = \{g_1, g_2, ..., g_N\}$, the output of the first features extractor network is a vector $V^1 = \{S_1^1, S_2^1, ..., S_N^1\}$ of N similarity measures S_i^1 calculated between the probe image p and each image g_i of the gallery set. By the same way, the output of the three other feature extractor networks are three vectors (V^2, V^3 and V^4) that contain the similarity between the segmented probe image p^* and each segmented image g_i^* in the gallery set.

2.3 Similarity Scores Fusion

There are several score fusion strategies that have been introduced in the literature [1]. Two types of fusion can be distinguished: early fusion and late fusion. Early fusion is a fusion of feature levels in which the output of a unimodal analysis is fused before training. For the late fusion, the results of the unimodal analysis are utilized to find out distinct scores for every modality. After the fusion, a final score is computed through the combination of the outputs of

every classifier. Compared with early fusion, late fusion is easier to implement and often shows effective in practice.

We propose in this work to adopt a late fusion model. It is used to aggregate the outputs of the four feature extractor networks. To compute the final similarity measure $V^{fus} = \{S_1^{fus}, S_2^{fus}, ..., S_N^{fus}\}$, the weighted sum is used in order to combine the outputs of each stream feature extractor (see Eq. 1).

$$V^{fus} = V^1 \oplus V^2 \oplus V^3 \oplus V^4 \tag{1}$$

where $S_i^{fus} = \alpha \cdot S_i^1 + \beta \cdot S_i^2 + \sigma \cdot S_i^3 + \gamma \cdot S_i^4$;
α, β, σ, $\gamma \in [0, 1]$
and $\alpha + \beta + \sigma + \gamma = 1$.

In order to fix the weighted sum parameters (α, β, σ and γ), we consider a greedy search algorithm over the search space. At each iteration, we fix three weights and we vary the fourth weight according to a step, so as to always have the sum equal to 1. In this way, we went through all the possible combinations. A step of 0.05 has been chosen.

3 Experiments

We empirically evaluate the proposed method in this section. First, we introduce the used datasets, then we present the implementation details. Finally we present the experimental results of the proposed approach.

3.1 Datasets

LIP dataset [16] is a dataset focusing on semantic understanding of person. It gives a more detailed comprehension of image contents thanks to the segmentation of a human image into multiple semantics parts. It is composed by over than 50,000 annotated images with 19 semantic body parts indicating the hair, the head, the left foot, the right foot, etc. semantic part labels and 16 body joints, taken from many viewpoints, occlusions and background complexities.

Market-1501 [25] is a publicly available large-scale person re-ID dataset which is used in this work to evaluate our proposed method. The Market-1501 dataset is made up of 32,667 images of 1501 persons partitioned into 751 identities for training stage and 750 for testing stage. Images are taken by one low-resolution and five high-resolution cameras.

Three segmented databases (Images without background, images of the middle body parts and images of the down body parts) have been created from the Market-1501 database which are organized and composed exactly by the same number of images as Market-1501.

3.2 Implementation

Two neural networks architectures were used for the feature extraction and re-ID step. First, we used a ResNet-50 [8] (baseline) pre-trained on the ImageNet dataset [19] and then trained separately on different datasets (original Market-1501 dataset, Market-1501 dataset without background, middle and down part datasets). The categorical-cross-entropy was used to output a probability over all identities for each image. The stochastic Gradient descent (SGD) is utilized to minimize the loss function and then to refresh the network parameters. Moreover, we applied data augmentation to make the training dataset bigger by the use of transformations like shear-range, width-shift-range, height-shift-range and horizontal-flip. We used 90% of the images for training and the remaining 10% for validation.

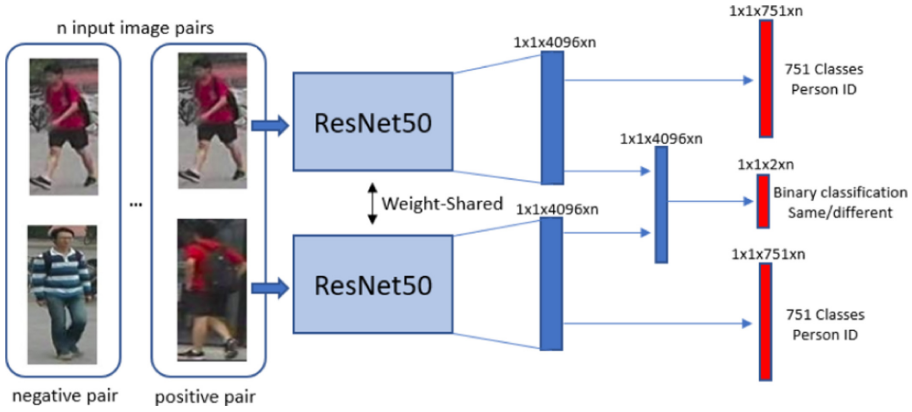

Fig. 4. In our S-CNN, two ResNet-50 are utilized to extract individually two features vectors from two images, which are utilized to predict the identity of the two input images and to predict the binary classification

The second network is a siamese network (S-CNN) [26] that combines the verification and identification losses. This network is trained to minimize three cross-entropy losses jointly, one identification loss for each siamese branch and one verification loss. We show in Fig. 4 the architecture of our siamese network. S-CNN is a two-input network formed by two parallel ResNet-50 CNNs with shared weights and bias. The final fully connected layer of the ResNet-50 network architecture is removed and three additional convolutional layers and one square layer are added [8]. Each Resnet-50 CNN is already pre-trained on different datasets. S-CNN is then trained separately on the four datasets using alternated positive and negative pairs (details of S-CNN training can be found in our previous work [6]).

To evaluate the performance of our re-identification method, we use two common evaluation metrics: the cumulative matching characteristics (CMC) at rank-1, rank-5 and rank-10 and mean average precision (mAP).

3.3 Results

We provide in this section the empirical results of the proposed method and we show the activation maps to demonstrate that using segmented images push the network to extract local features.

Extracting features from only the whole body can skip some significant information (See Fig. 5). The active region in the *No_bk stream* is larger than the active region in the *Full stream* what can cover more discriminative features. The activation map of the *dwn stream* show clearly that the use of segmented images allows us to exploit some regions of the image whose are not activated in the global stream.

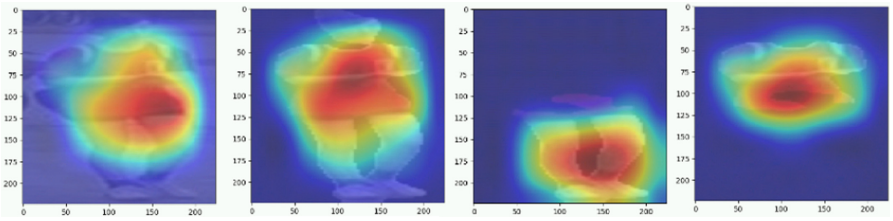

Fig. 5. Samples of activation maps from our four streams for the same image. From left to right: the activation map of the *Full stream*, the *No_bk stream*, the *dwn stream* and the *Mid stream*

Table 1. Fusion results obtained for different fusion methods on Market-1501 dataset.

Fusion method	S-CNN				ResNet50			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Max	74.04	89.42	93.11	56.83	68.34	87.02	91.86	50.97
Sum	86.34	94.12	96.02	71.22	82.77	92.63	95.21	65.56
Accuracy weighted sum	86.54	94.23	**96.37**	71.70	82.95	92.72	95.19	65.57
Product	86.10	93.82	95.99	70.76	82.60	**92.96**	95.33	65.28
Greedy weighted sum	**87.02**	**94.26**	96.31	**71.72**	**83.87**	92.81	**95.63**	**66.15**

In the literature, there are several methods of score fusion that can be used in the context of our work [14]. In order to choose the best method, a comparative study was made. The fusion methods that have been tested are: max rule, sum rule, product rule, accuracy weighted sum and greedy weighted sum. We display in Table 1 the result of this study. Table 1 shows that the weighted sum using the greedy search technique gave the highest accuracy for both S-CNN and Resnet-50 networks. Since we have four streams with fairly different recognition rates, the weighted sum makes it possible to penalize the stream with low precision

and favor the one with high precision. With greedy search technique, we can found suitable weights for our recognition model.

We show in the top of Table 2 the result of our two re-ID networks for the four streams (*Full*, *No_bk*, *Mid* and *Dwn*). We notice that S-CNN provides better results except for the stream *Mid* where S-CNN is slightly worse than the baseline. For both of S-CNN and ResNet-50, the best performance is obtained for the stream *Full* which extract information from the whole image.

Table 2. Fusion results of the multi-stream outputs on Market-1501 dataset.

Streams combination	S-CNN				ResNet50			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Full	79.78	91.30	94.32	62.22	73.60	88.21	92.39	54.89
No_bk	73.93	88.62	92.10	58.06	67.33	84.88	90.26	50.75
Mid	47.47	68.61	75.32	29.03	47.47	68.52	75.83	28.81
Dwn	47.77	68.73	76.72	34.18	46.49	69.86	77.25	32.65
Full + No_bk + Mid + Dwn	**87.02**	**94.26**	**96.31**	**71.72**	**83.87**	**92.81**	**95.63**	**66.15**
Full + Mid + Dwn	85.71	94.00	96.22	70.70	83.16	92.78	95.21	65.16
Full + No_bk + Mid	84.88	93.40	95.51	69.01	80.10	91.83	94.56	61.66
Full + No_bk + Dwn	84.82	93.82	95.72	69.74	81.23	91.98	94.95	63.98
No_bk + Mid + Dwn	83.64	92.99	95.54	67.98	80.78	91.83	94.32	62.45
Full + Dwn	83.07	93.08	95.36	67.46	79.39	91.24	94.06	61.71
Full + Mid	83.01	92.63	94.89	66.85	77.90	90.40	93.76	58.62
Full + No_bk	82.57	92.69	95.30	67.49	77.22	90.17	93.61	59.76
No_bk + Mid	79.89	91.35	94.26	62.41	74.55	88.86	93.02	55.89
No_bk + Dwn	79.78	91.38	94.26	63.57	75.00	88.77	93.11	57.76
Mid + Dwn	78.97	90.29	93.20	60.09	76.90	89.75	92.63	56.42

In the bottom of Table 2, we present the combination results of the four stream re-ID model outputs using a weighted sum method. A greedy search method is used to get the best results on the validation dataset. The weights setting consists in choosing the best parameters: α, β, σ and γ (combination coefficients of the four streams respectively *Full*, *No_bk*, *Mid* and *Dwn*), where $\alpha + \beta + \sigma + \gamma = 1$.

Table 2 shows that the best combination is when the four streams are activated and S-CNN always provide better results than the baseline for the different combinations. With S-CNN, we get 87.02% in the rank-1 accuracy and with Resnet-50 we get 83.87%.

The obtained results show that the combination of multiple streams significantly improves the re-ID performances. Comparing to *Full* stream results, the improvement in rank-1/mAP is 7.24%/9.5% with S-CNN and 10.27%/11.26% with Resnet-50 when the four streams are combined. Less improvement values

are shown when only three or two streams are combined. This improvement confirms that the features extracted from the body parts (local features) give complementary information to the whole person image features (global features).

We notice that for a re-ID model based on only three streams, the best combination is when the two streams Mid and Dwn are combined with the $Full$ stream. The rank-1 accuracy respectively for S-CNN and Resnet-50 are, 85.71 and 83.16. This result confirms that local features provide additional complementary information that can enhance the re-ID decision.

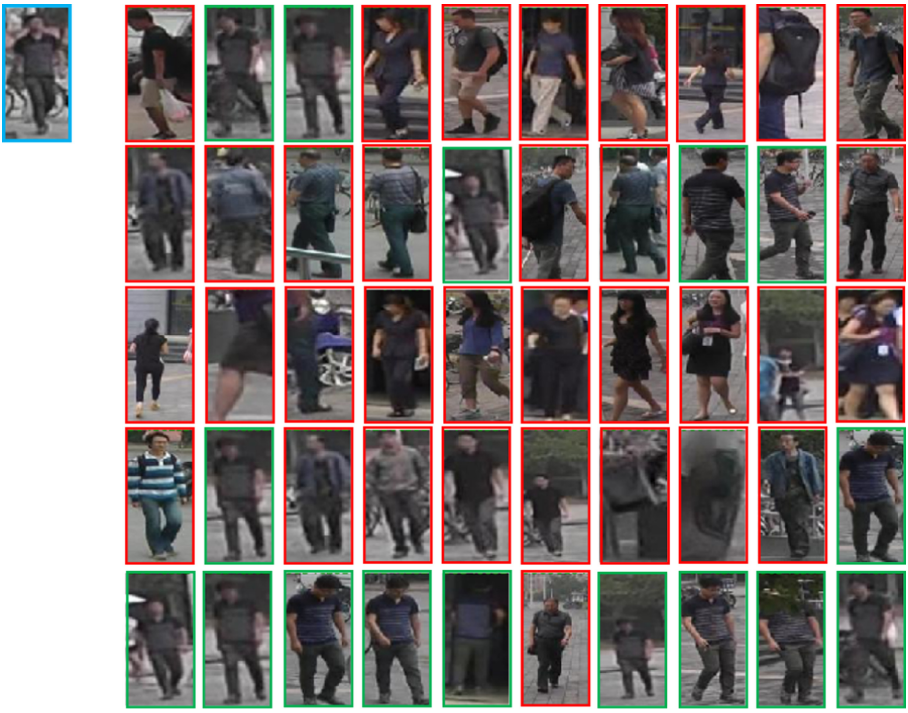

Fig. 6. Top 10 predictions for each stream and the fusion for the same query image. Person surrounded by blue box corresponds to the probe image. The first four rows correspond to the results produced by $S-CNN$ on the $Full$, No_bk, Mid, and Dwn streams respectively. The fifth row corresponds to the results produced by the fusion method. Person surrounded by green box denotes the same person as the probe and person surrounded by red box denotes the negative predictions. (Color figure online)

We display in Fig. 6 one qualitative result showing how the fusion of multi-stream outputs can significantly improve the prediction results. Even when all of the four streams provide a negative prediction in rank-1, the fusion method success to provide the same identity as the probe image. Some images with the same identity as the probe image can move up in the top 10 ranking list despite their absence in the top 10 of all the four streams. This result can be explained

by the slight difference between the scores in some cases where the images of different people are visually difficult to distinguish. The fusion method therefore makes it possible to favor images corresponding to the probe identity at the expense of false identities.

Our method is also compared to several state-of-the-art methods for person re-ID based on multi-stream strategy. This comparative study is presented in Table 3. The obtained results highlight the performances of the proposed method and show that our method outperforms many previous works by a large margin in rank-1 accuracy and mAP. After embedding the re-ranking method [27], we obtain 89.22% in rank-1, an improvement of 2.2%.

Table 3. Comparison with the state-of-the-art approaches on Market-1501 dataset.

Method	Rank-1	mAP
Gated S-CNN [21]	65.88	39.55
PL-Net [24]	69.3	88.2
MSCAN [15]	80.31	57.53
BSTS S-CNN [6]	81.79	66.78
GPN [9]	81.94	87.07
MSCF_RK [12]	85.7	–
GLAD [22]	89.9	73.9
Ours	87.02	71.72
Ours + re-ranking	89.22	83.72

4 Conclusion

In this paper, we proposed a multi-stream method for person re-ID that aims to exploit global and local features and to avoid problems related to background variations. We propose to first perform a semantic segmentation to extract body parts, then we combine local and global feature extractor outputs to make final decision. We showed that combining multiple highly discriminative local region features of the person images leads to an accurate person re-identification system. Experiments on Market-1501 dataset clearly demonstrate that our proposed approach considerably improves the performance as compared with mono-stream methods and to the state-of-the-art approaches.

Acknowledgement. This project is carried out under the MOBIDOC scheme, funded by the EU through the EMORI program and managed by the ANPR. We thank Anavid for assistance. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Syst. **16**(6), 345–379 (2010). https://doi.org/10.1007/s00530-010-0182-0
2. Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P.: A database for person re-identification in multi-camera surveillance networks. In: International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 1–8. IEEE (2012)
3. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Computer Vision and Pattern Recognition (CVPR). pp. 1335–1344 (2016)
4. Cho, Y.J., Yoon, K.J.: Improving person re-identification via pose-aware multi-shot matching. In: Computer Vision and Pattern Recognition (CVPR), pp. 1354–1362 (2016)
5. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1528–1535. IEEE (2006)
6. Ghorbel, M., Ammar, S., Kessentini, Y., Jmaiel, M.: Improving person re-identification by background subtraction using two-stream convolutional networks. In: Karray, F., Campilho, A., Yu, A. (eds.) ICIAR 2019. LNCS, vol. 11662, pp. 345–356. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27202-9_31
7. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-Identification, 1st edn., p. 445. springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
9. Hu, X., Jiang, Z., Guo, X., Zhou, Y.: Person re-identification by deep learning muti-part information complementary. In: IEEE International Conference on Image Processing (ICIP), pp. 848–852 (2018)
10. Huang, H., Li, D., Zhang, Z., Chen, X., Huang, K.: Adversarially occluded samples for person re-identification. In: Computer Vision and Pattern Recognition (CVPR), pp. 5098–5107 (2018)
11. Huang, Y., Zha, Z.J., Fu, X., Zhang, W.: Illumination-invariant person re-identification. In: ACM International Conference on Multimedia, pp. 365–373 (2019)
12. Huang, Z., et al.: Contribution-based multi-stream feature distance fusion method with k-distribution re-ranking for person re-identification. IEEE Access **7**, 35631–35644 (2019)
13. Karanam, S., Li, Y., Radke, R.J.: Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: IEEE International Conference on Computer Vision, pp. 4516–4524 (2015)
14. Kittler, J.: Combining classifiers: a theoretical framework. Pattern Anal. Appl. **1**(1), 18–27 (1998). https://doi.org/10.1007/BF01238023
15. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Computer Vision and Pattern Recognition (CVPR), pp. 384–393 (2017)
16. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 871–885 (2018)

17. Mansouri, N., Ammar, S., Kessentini, Y.: Improving person re-identification by combining Siamese convolutional neural network and re-ranking process. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE (2019)
18. Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y.: Auto-ReID: searching for a part-aware convnet for person re-identification. arXiv preprint arXiv:1903.09776 (2019)
19. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
20. Tian, M., et al.: Eliminating background-bias for robust person re-identification. In: Computer Vision and Pattern Recognition (CVPR), pp. 5794–5803 (2018)
21. Varior, R.R., Haloi, M., Wang, G.: Gated Siamese convolutional neural network architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 791–808. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_48
22. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: GLAD: global-local-alignment descriptor for scalable person re-identification. IEEE Trans. Multimedia **21**(4), 986–999 (2018)
23. Weinrich, C., Volkhardt, M., Gross, H.M.: Appearance-based 3D upper-body pose estimation and person re-identification on mobile robots. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 4384–4390. IEEE (2013)
24. Yao, H., Zhang, S., Hong, R., Zhang, Y., Xu, C., Tian, Q.: Deep representation learning with part loss for person re-identification. IEEE Trans. Image Process. **28**(6), 2860–2871 (2019)
25. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124 (2015)
26. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person reidentification. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **14**(1), 13 (2018)
27. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Computer Vision and Pattern Recognition (CVPR), pp. 1318–1327. IEEE (2017)