Cross-Scenario Person Re-identification

Ruoran Jia*, Shuguang Liu

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences

*liushuguang@cigit.ac.cn, jiaruoran@foxmail.com

Abstract—Person Reid is a challenging task for two factors, first one is background interference, such as changes in light, weather, posture, and camera position. Second is domain adaptive capacity, such as model train by market1501 achieve the same performance on Duke dataset. To solve the above problems, we come up with adopt human semantic to remove clutter from unwanted background information, is naturally a better alternative compare with bounding box, we adopt Local Regions Representation to extra the image features, which can preeminently improve the representation of local feature and global feature. Our proposed CSReID integrates human semantic and Local Regions Representation in person re-identification and not need to train on the evaluation dataset can achieve state of the art cross-modal performance.

1. INTRODUCTION

Person Reid is retrieve a query image from a given large person image gallery to get all have same id person images, which query images and gallery images should come from different cameras.

Person ReID has two challenge. First, when the same person is captured in different cameras, the image will be affected by weather, light, body posture, camera angle and other factors and show various differences. Second, even under the same camera, the colors and contents of the images are greatly different due to the changes of people's posture and the surrounding environment. Third, models that are trained in one data set have a significant reduction in performance in another data set. Consider all of the above, an effective person ReID model’s target is to learn representations that are identity-specific, context-invariant and domain adaptive.

In recent years, replacing global features with local features(part-level) to represent pedestrian characteristics has become the mainstream research direction of Pedestrian Reid feature representation research. While the feature expression of the whole image is easily interfered by the background information, while the local feature information can effectively reduce this influence and have higher robustness. And this is where some of the classic papers tend to be simply extract representations from multiple image patches, most horizontal strips, that is closer to the physical relationship of the human body. But all previous method have two questions. First, are the best method to extract human characteristics with human body boundary image? Second, all the ReID train on one dataset can achieve the same effect in another scenario dataset without Extra training?

To solve the first problem, we have put forward using semantic segmentation, to be more specific human semantic, as an alternative method of boundary box to extract local features of human body. Compared with the bounding box, the semantic segmentation of human body is more accurate than the bounding box, which can effectively remove the extra class gap caused by the background information
in the bounding box at different positions to the same person. We first train a semantic model that learns to segment the human body and then use them to extract local features for person Re-ID task.

To address the second question, we propose to use local region representation instead of multi-branch full connection feature to achieve good retrieval results without extra training. The contributions of this paper are as follows:

• Through a series of experiments, we proved that, our local regions representation can significantly outperform current local features presentation method.

• We put forward CSReID, by combining the semantic segmentation of human body with the expression of local features for person re-identification. Through a series of experiments we show it helps improving effectiveness and robustness, The validity of the method is proved.

2. RELATED WORK
In recent years, remarkable progress has been made in different areas of computer vision, including the re-identification of people. This is due to the emergence of deep learning, particularly deep convolutional neural networks. In the problem of human re-identification, due to the challenge of attitude and light change, occlusion and background clutter, the researchers has been concerned with two main sub-problems, that is feature representation and model generalization. The improvement of feature expression ability has been greatly improved in the following literatures, but how to improve the robustness and generalization ability of the model is still a problem worthy of continuous research.

In order to improve the feature expression ability of the model, authors in [8] propose extract the features of each part of the limb to form the overall features. Li et al. [8] take body parts roughly classified as head-shoulder, upper-body and lower-body using a human key point detection network [6]. And then there's another branch, with shared weights, through a multi-scale CNN structure process these parts, The final global representation vector is obtained by combining the characteristics of each part. Zhao et al.[17] use a similar method also achieve better results. Their model is very complex and has been trained in several non-trivial stages. In order to avoid the network training and calculation cost of detecting the key parts of human body parts, authors in [22] attempts were made to extract multiple image blocks closely related to human body parts from the image, but such a frame could not properly handle the dislocation of human body parts. Some work attempts to solve the problem of asymmetry by explicitly integrating attitude estimation into human re-identification, in which the existing attitude estimation model is used to initialize part of the position to quadrilateral, and then alignment is carried out through affine transformation or spatial transformation network [6].

Of course, many researchers have tried [21] use people attributes to improve people re-identification performance. However, the premise is that the attribute detection network can be prepared to detect these attributes under different conditions, so a large number of attribute data sets under various conditions are needed to train the attribute identification network with excellent performance and high robustness.

Unlike the previous methods, Zhao et al. [17], adopt The original global feature representation, but the classification loss function is modified to be a multi-class loss function. Compared to the above works, we propose to use local regions representation to get more distinctive feature and combine with human semantic get better re-identification performance.

3. METHODOLOGY
In our work, we adopt Inception-V3 network[14] as the CNN backbone for person ReID feature learning.

3.1 Human Semantic Network
In order to improve the ability of network feature expression for person ReID, we propose to combine human semantic and local regions representation. We believe that semantic segmentation is better than boundary box detection due to its accuracy of pixel level and robustness of attitude and background
environment change[2], and compared with extract representations from multiple image patches with full connection, local regions representation can extract the features of each position of the image.

We adopt Inception-V3 [14] as the CNN backbone network in our re-identification model. However, we make two modifications to fit our model. We have changed the step of the last grid reduction module in the Inception-V3 network[14] from 2 to 1, which resulting output step size is 16, compared to 32 in the original architecture. To solve the problem of adding extra computation of the last Inception block, we use the dilated convolution[16] replace the corresponding convolution filters. With the above adjustments, we can perform multi-class classification at the pixel level and image classification after backbone.

3.2 Local Regions Representation
To extra the local visual features which representative and differentiated, we use the local regions representation. In this section, we will describe how to derive the representation of the image regions by activating the convolutional layer of the backbone CNN, each region will generate a description vector, vectors generated by these regions are aggregated together to produce a description vector of 2048 dimensions called \( \mathbf{f}_\Omega \). The feature vector \( \mathbf{f}_\Omega \) is a representation for the whole image, \( \Omega \) representation whole feature map. Let's define a rectangular region \( R \subseteq \Omega = [1; W] \times [1; H] \), and define the regional feature vector formally as the following:

\[
\mathbf{f}_R = [\mathbf{f}_{R;1} \ldots \mathbf{f}_{R;i} \ldots \mathbf{f}_{R;K}]
\]

We define \( \mathbf{f}_{R;i} = \max_{p \in R} x_i(p) \) is the maximum activation value of the \( i \)-th channel on the specified region. The regions \( R \) are defined on the space \( \Omega \) of all valid positions and areas for the specified feature map (not on the origin input image). A feature vector of a region with size 1 corresponds to consisting of a single activation value at a particular pixel on the feature map. We can now construct a representation for multiple regions in specified feature map without need to re-enter additional information to network.

We now define a batch of \( R \) regions with different positions and sizes, sampled these square regions on \( S \) scales. At the largest scale \( (s = 1) \), the size of region should be as large as possible, i.e., its height and width are both equal to \( \min(W; H) \) (\( W, H \) is width and height of feature map). In order to reduce unnecessary calculation, the sampling area is uniform, so that the overlap between continuous areas is less than 30%. According to the above definition, we sampled \( s \times (s + m - 1) \) regions of width \( 2 \min(W; H) = (s + 1) \) on each scale \( s \).

Once we have these regions, we compute the feature vector for each region, and process them with L2-normalization, PCA-whitening and L2-normalization. We add the collected regional feature vectors and finally conduct summing and L2-normalization to obtain a single image vector, which keeps the dimension low, equal to the number of feature channels.

The resulting vector can be regarded as the aggregation of local features of different scales and locations, with excellent representativeness and differentiation.

3.3 Person ReID Framework
Our network structure, illustrated in Figure 1 consists of a human semantic branch, a classification branch and a aggregation branch. We call it as CSReID: Cross-Scenario Person Re-identification. The classification backbone in CSReID has same structure with the backbone of human semantic branch but has difficult parameters. Hence, it can perform different functions under same receptive field.

We pass the classification branch to a multiclass classification objective with softmax cross-entropy loss to fine tuning the parameters pre-trained from ImageNet to train the network, pass the human semantic branch to a human and background segmentation with mix softmax cross-entropy loss to get common human segmentation parameters. At the predict time, even if the scene is changed, there is no need to train again. Final representations vector is aggregate the feature maps before the classifier layer of classification branch and probability maps of human semantic and cross local regions representation get it, which is used to retrieve correct matches of a given query person image from the gallery images.
In CSReID, we use a pixel level matrix multiplication between the output of classification backbone and human semantic probability maps where their corresponding spatial domain have same size, the value in human semantic branch’s feature maps can represent the weight of the human body in classification branch feature map with some position. After the matrix multiplication we use the local regions representation features to derive representations vector for image.

Figure 1: CSReID framework: our network first convert the input RGB image into a tensor of activations via a modified Inception-V3 backbone, the human semantic branch generating probability maps. CSReID combine the probability maps to aggregate the convolutional activations from classification backbone, finally by local regions representation generation 2048-dim vector.

4. EXPERIMENTS

4.1 Datasets and Evaluation

In order to evaluate our proposed methodology, we used three publicly available large-scale data sets, CUHK03 [11], Market-1501[18] and DukeMTMC [19].

In addition to the data sets listed above, we also adopted 3DPeS [1], VIPeR [3], PRID [5], Shinpuhkan [7], CUHK02 [9], CUHK01 [10] and PSDB [15] datasets as train data, which to train our model and current state-of-the-art models. We evaluate the quality of different person ReID models using topK accuracy and mean average precision (mAP). All the evaluate experiments are conducted without re-rank.

4.2 Person Re-identification Performance

In this section, we analyzing the performance of our person ReID model with current state-of-the-art models with the scene of cross data set.

4.3 Comparison with the state-of-the-art

| Model          | Market-1501 |
|----------------|-------------|
| Model          | mAP(%)      | rank1 | rank10 |
| Resnet[4]      | 41.1        | 72.3  | 85.2   |
| SOMAnet[23]    | 43.9        | 73.9  | 85.2   |
| SVDNet[24]     | 52.1        | 76.3  | 82.3   |
| Triplet Loss[25]| 55.1        | 72.9  | 78.5   |
| Cam-GAN[26]    | 59.7        | 72.3  | 76.8   |
|                | mAP(%) | rank1 | rank10 |
|----------------|--------|-------|--------|
| MultiRegion[27]| 66.96  | 74.63 | 87.65  |
| PAR[28]        | 63.5   | 75.0  | 84.7   |
| MultiLoss[29]  | 64.4   | 76.9  | 84.3   |
| MultiScale[30] | 63.1   | 75.9  | 85.3   |
| HA-CNN[31]     | 65.7   | 76.2  | 86.9   |
| PCB[32]        | 70.4   | 82.3  | 88.2   |
| CSReID         | 75.5   | 89.5  | 90.5   |

CUHK03

| model          | mAP(%) | rank1 | rank10 |
|----------------|--------|-------|--------|
| Resnet[4]      | 37.2   | 67.5  | 73.1   |
| SOMAne[23]     | -      | 65.4  | 68.6   |
| SVDNet[24]     | -      | 67.4  | 69.4   |
| Triplet Loss[25]| -      | 65.7  | 71.4   |
| Cam-GAN[26]    | -      | 67.6  | 69.8   |
| MultiRegion[27]| -      | 70.54 | 74.6   |
| PAR[28]        | -      | 71.0  | 74.6   |
| MultiLoss[29]  | -      | 72.0  | 75.7   |
| MultiScale[30] | -      | 70.4  | 76.4   |
| HA-CNN[31]     | -      | 69.4  | 73.3   |
| PCB[32]        | -      | 72.5  | 78.2   |
| CSReID         | 73.6   | 76.7  | 84.1   |

DukeMTMC-reID

| model          | mAP(%) | rank1 | rank10 |
|----------------|--------|-------|--------|
| Resnet[4]      | 35.1   | 65.7  | 75.1   |
| SOMAne[23]     | 37.8   | 67.9  | 70.8   |
| SVDNet[24]     | 46.8   | 68.7  | 72.5   |
| Triplet Loss[25]| 48.1   | 69.8  | 71.4   |
| Cam-GAN[26]    | 49.7   | 70.5  | 72.8   |
| MultiRegion[27]| 50.4   | 70.3  | 71.0   |
| PAR[28]        | 53.7   | 70.2  | 73.6   |
| MultiLoss[29]  | 58.9   | 68.4  | 73.1   |
| MultiScale[30] | 58.0   | 69.3  | 71.3   |
| HA-CNN[31]     | 58.3   | 71.4  | 73.4   |
| PCB[32]        | 61.5   | 74.4  | 78.6   |
| CSReID         | 67.6   | 79.9  | 82.7   |

From Table 1, We can come to the conclusion that our network outperform the current state-of-the-art. All model was training on the additional data set and test on CUHK03, Market-1501 and DukeMTMC. All the results have not employ re-ranking [20]. Therefore, We can confirm that the generalization ability and robustness of the model can be effectively enhanced through the combination of human semantic segmentation and local feature representation.

5. Conclusion
In this paper, we propose two improvement points for two challenge points. First, whether the ReID train on one dataset can achieve the same effect in another scenario dataset without extra training. Second, whether has a better method to extra local visual features. Through this paper, we showed that Local Regions Representation surpass extract representations from multiple image patches. We have also demonstrated that by taking advantage of our proposed human semantics and Local Regions Representation, the CSReID framework performance excellent when test on different datasets.
REFERENCES

[1] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, pages 59–64. ACM, 2011. 5

[2] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017. 3

[3] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), volume 3, pages 1–7. Citeseer, 2007. 5

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 2, 3, 4, 6

[5] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In Proc. Scandinavian Conference on Image Analysis (SCIA), 2011. 5

[6] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems, pages 2017–2025, 2015. 2, 3

[7] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision, volume 5, page 6, 2014. 5

[8] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 384–393,2017. 2, 8

[9] W. Li and X. Wang. Locally aligned feature transforms across views. In CVPR, 2013. 5

[10] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In ACCV, 2012. 5

[11] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 152–159, 2014. 1, 2, 3, 4, 5, 6, 7

[12] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multicamera tracking. In European Conference on Computer Vision Workshop on Benchmarking Multi-Target Tracking,2016. 4

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252,2015. 8

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016. 2, 3, 4, 6

[15] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Endto-end deep learning for person search. arXiv preprint arXiv:1604.01850, 2016. 5

[16] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.3

[17] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1077–1085, 2017. 2, 3, 8

[18] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, pages 1116–1124, 2015. 1, 2, 4, 5, 6
[19] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, 2017.

[20] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. arXiv preprint arXiv:1701.08398, 2017.

[21] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 1435–1443. IEEE, 2017.

[22] [F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian. Part-based deep hashing for large-scale person re-identification. IEEE Transactions on Image Processing, 2017.

[23] Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking beyond appearances: Synthetic training data for deep cnns in re-identification. arXiv preprint arXiv:1701.03153 (2017)

[24] Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: ICCV(2017)

[25] Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv: 1703.07737 (2017)

[26] Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. arXiv preprint arXiv:1711.10295 (2017)

[27] Ustinova, E., Ganin, Y., Lempitsky, V.: Multiregion bilinear convolutional neural networks for person re-identification. arXiv preprint arXiv:1512.05300 (2015)

[28] Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)

[29] Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multiloss classification. In: IJCAI (2017)

[30] Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: In International Conference on Computer Vision, Workshop on Cross-Domain Human Identification (CHI) (2017)

[31] Li, W., Zhu, X., Gong, S.: Harmonious attention network for person reidentification. arXiv preprint arXiv:1802.08122 (2018)

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling. In ECCV. In press.