Segment attention-guided part-aligned network for person re-identification

Wen Wang,1,2 Yongwen Liu,1,2,3 and Gaoyun An1,2,3
1Institute of Information Science, Beijing Jiaotong University, Beijing, China
2Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
3Email: gyan@bjtu.edu.cn

Part misalignment of the human body caused by complex variations in viewpoint and pose poses a fundamental challenge to person re-identification. This letter examines Res2Net as the backbone network to extract multi-scale appearance features. At the same time, it uses the human parsing model to extract part features, which can be used as an attention stream to guide part features re-calibration from the spatial dimension. Additionally, in order to ensure the diversity of features, SAG-PAN effectively integrates the global appearance features of person image with part fine-grained features. The experimental results on the Market-1501, DukeMTMC-reID and CUHK03 datasets show that the proposed SAG-PAN achieved superior performance against the existing state-of-the-art methods.

Introduction: Person re-identification is a cross-camera retrieval task. The emergence of this task is due to the increasing public safety requirements and large-scale camera networks in public areas. However, because of the misalignment of the human body parts caused by cross-view, multipose, illumination, and occlusion, this research is still challenging.

Most methods for person re-identification assume that all features of image’s different parts are equally important, the entire image sharing one filter bank. But in reality, contributions of different parts to person re-identification are different. Recent researches mainly focus on using attention mechanism to relocate features. Li et al. [2] proposed harmonious attention network (HAN) to combine soft pixel attentive and hard zone attention through harmonious attention module. Guo et al. [3] proposed P2-Net. The kernel module of P2-Net is dual-part-aligned block (DPB), which uses human analytic model to generate body part covering mould. Those methods only focus on local features but ignore the global features. This letter, in order to solve this problem, introduces our segment attention-guide part-aligned network (SAG-PAN) where the human body analytical model is introduced as attention to extract local features while preserving the global features.

Segment attention-guided part-aligned network: Our proposed network, SAG-PAN, enables network selectively extract more accurate depth features from spatial dimension by human part parsing model which introduces part attention. The specific structure is shown in Figure 1, including global branch and human part branch. Human part branch extracts part feature and then converts it into a human body part mask to capture the local feature. Global branch is used to extract global feature. During training, local feature and global feature are trained independently. Since global features and local features are complementary to a certain extent. The method proposed in this article effectively merges the global features and local features. During testing, local feature and global feature are spliced and used as the final feature representation.

Human part branch: Human part branch is a dual-stream structure network. Res2Net-50 is used as the backbone network of the appearance feature extractor. The human part parsing model extracts the part features. Then, it converts the part features into a part mask and weights the appearance features to obtain local features.

Human parsing model adopts the parsing branch of augmented context embedding with edge perceiving (A-CE2P) [4] to capture high-level semantic perception information. The specific structure of human parsing model includes backbone network and context coding module. The backbone network consists of three convolutional layers, a pooling layer and four groups of ResNet-101 sub-modules. The context coding module using global context information to identify fine-grained category information to extract high-level features of different parts. Human parsing model models seven parts of the human image: background, head, torso, upper arm, lower arm, upper leg, and lower leg, of which the kernel module is context coding module.

The binary mask \( f_{bp} \in R^{H \times W} \) can be obtained by calculating following equation:

\[
f_{bp} = \text{Sigmod} \left( \sum_{i=1}^{C_d} p_i \right), \quad f_{bp} \in R^{H \times W}
\]

where \( f_{bp} \) represents the feature map of \( f_{bp} \)’s rth channel, \( f_{bp} \in R^{H \times W \times C_d} \) represents appearance feature and \( f_{bp} \in R^{H \times W \times C_p} \) represents part feature, \( H \) and \( W \) represent the height and width of feature map respectively, \( C_d \) and \( C_p \) represent the number of exterior feature channels and the number of part feature channels respectively.

Element multiplication is used to integration appearance feature and the binary mask of part feature. Then, global mean pooling layer is used to obtain feature \( f' \):

\[
f' = G4P(f \times \text{broadcasting}(f_{bp}))
\]

broadcasting makes \( f_{bp} \) get the same number of channels as \( f \), through self-replication. Binary mask completes appearance feature relocation in the spatial dimension. Human part feature map reflects feature values of human body’s various parts. The normalized part feature value can be considered as attention to guide appearance feature extractor to learn part feature efficiently and weaken the influence of background information.

Lastly, BN layer is used for part feature \( f_{local} \) to separate features:

\[
f_{local} = BN(f)
\]

During training, \( f \) and \( f_{local} \) are used to calculate triplet loss and identification loss respectively. During testing, \( f_{global} \) is used as the description of sample’s feature and cosine distance is used to measure the similarity between samples.

Our loss function is the combination of the identification loss, centre loss and triplet loss objectives, which can be expressed in the following equation as,

\[
L = L_{ID} + L_{triplet} + \lambda L_{centre} + L_{part} + L_{triplet}
\]

where \( \lambda \) represents the trade-off factor of the global branch loss function. \( L_{ID} \) is a N-class cross entropy loss,

\[
L_{ID} = -\frac{1}{|B|} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{ij} \log(p_{ij})
\]

where \( |B| \) represents the data size, \( N \) represents the number of classes, \( p_{ij} \) represents the probability that the sample belongs to the ith class.
Experimental results: In order to evaluate the effectiveness of the proposed SAG-P AN, we conducted extensive experiments on three datasets: Market-1501, DukeMTMC-reID, and CUHK03.

Experimental setup: The inputs to SAG-P AN are normalized RGB images with the spatial size 256 × 128. The data enhancement methods used in the training phase include random erasure and random horizontal flip. The parameters of the global branch backbone network in SAG-P AN and the appearance feature extractor backbone network in the part branch are initialized with the model parameters of Res2Net-50 trained on the ImageNet dataset. The pedestrian analysis model of SAG-P AN is initialized with the model parameters of A-CE2P trained on the Pascal Person Part dataset. The parameters of the BN layer in SAG-P AN are initialized using the same initialization method in [1] initialization method. The initialization parameters of the fully connected layer follow a normal distribution with a mean value of 0 and a standard deviation of 0.001. The ADAM algorithm is used to optimize the model. The initial learning rate is 0.00035, and the weight decay is 0.0005. When the epoch is 80 and 160, the learning rate becomes 1/10 of the previous stage. The batch size is 64. The margin of the triplet loss function is set to 0.3. The learning rate of the centre loss is initialized to 0.5, and the loss weight is 0.001. The ADAM algorithm is used to optimize the model. The initial learning rate is 0.00035, and the weight decay is 0.0005. When the epoch is 80 and 160, the learning rate becomes 1/10 of the previous stage. The batch size is 64. The margin of the triplet loss function is set to 0.3. The learning rate of the centre loss is initialized to 0.5, and the loss weight is 0.0005, which is optimized by using the SGD algorithm.

Conclusion: This paper proposes a SAG-P AN which is composed of global branch and human part branch. Global branch captures the global features of person image. Human part branch uses human parsing model to capture high-level semantic information of predefined parts to extract local features, and effectively combine global features with local features. The experimental results on the three benchmark datasets Market-1501, DukeMTMC-reID and CUHK03 show that the proposed SAG-P AN model achieves the state-of-the-art results.

Acknowledgements: This work was supported partly by the National Natural Science Foundation of China (62072028 and 61772067). This work was supported partly by the National Natural Science Foundation of China (62072028 and 61772067).

References
1. He, K. et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: IEEE International Conference on Computer Vision (ICCV 2015), Santiago (2015)
2. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City (2018)
3. Guo, J. et al.: Beyond human parts: Dual-part-aligned representations for person re-identification. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul (2019)
4. Li, P. et al.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence 1–1 (2019)
5. Zhou, K. et al.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul (2019)
6. Zheng, Z. et al.: Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach (2019)
7 Fang, P. et al.: Bilinear attention networks for person retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul (2019)
8 Chen, T. et al.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul (2019)
9 Zhang, Z. et al.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach (2019)
10 Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul (2019)
11 Zhao, S. et al.: Do not disturb me: Person re-identification under the interference of other pedestrians. In: European Conference on Computer Vision (ECCV), Glasgow (2020)
12 Zhou, J., et al.: Cross-correlated attention networks for person re-identification. *Image Vision Comput.* **100**, 103931 (2020)