Pose-aware Person Re-Identification with Spatial-temporal Attention

Qi Zhu
Beijing Institute of Technology
qzhu.working@gmail.com

Abstract. The Person re-ID task objective is to search for a specified target pedestrian image in a large gallery. With the development of deep learning technology, the accuracy of personnel re-identification has been greatly improved. However, previous methods lacked sufficient attention to local features when focusing on global features, causing background and occlusion objects to affect the foreground. The framework we proposed specifically designed this problem to optimize local features through human pose key-points and self-attention mechanisms. And the effectiveness of the method is demonstrated in multiple datasets.

Index Terms—Re-Identification, Pose, Attention, Spatial-temporal

1. Introduction
Person re-identification has been viewed as one of the critical problems of the general recognition task. It aims to identify a person of interest from a large scale gallery image database which captured from distinctively cameras in large time. By given a probe image of the person, Person re-identification algorithm will estimate the visual similarities between the probe image and gallery image database. Then the gallery images can be ranked descended according to the similarities as re-identification results. Person re-identification is a very challenging task. Between cameras, the illumination conditions, background clutters, occlusion, visible human body parts, and perceived posture of the person could be completely different. Even within same camera, as the person moves and covered by items, the features can vary dramatically. All of the factors generate a huge intra-class variation and become the limitation of traditional re-identification approaches.

Recently, following the development of deep learning [1, 2, 3] and better classification models [3, 4], person re-identification obtain more robust end-to-end representative features through deep learning networks. Compared with shallow learning, these deep features effectively improve the performance of person re-identification in the commonly used datasets [5, 6, 7].

Therefore, the deep feature cares more about global information. It may ignore some of the local features which include semantic information. [8] designs a simple splitting method to acquire different local features to aid the global features. [9] proposed a new approach that processes the image as different slices, then re-arrange all these pieces as new images. In the latest researches, some methods use pose information to guide the splitting of persons in the images and obtain better results.

In this paper, we present a new human re-identification framework that includes a new branch to explore feature information within the image. In this new branch, we extract image partitions with semantic information by extracting key-points of the human pose.
The spatial-temporal information extracted according to the key-point information can adequately capture the local features and supplement the difficult-to-identify parts in the global feature analysis. The main contributions of this work can be summarized as the following three-folds:

- We propose a novel coarse-to-fine framework to learn precise pedestrian features in person re-identification task. It mainly depends on pose estimation and self-attention mechanism.
- The proposed spatial-temporal framework obtained accurate aligned parts information, and consider the different human parts based on self-attention network without extra training.
- Our approach achieves the state-of-the-art performance on various large-scale person Re-ID dataset such as Market-1501 [5], and demonstrates its effectiveness and generalization ability.

2. Related Work

2.1. Recurrent Neural Network

Recurrent Neural Network (RNN) is the deep neural network that considers temporal information as the input. It needs to memorize past information and captures the contextual dependency of the sequential inputs. Long Short Term Memory (LSTM) [1] networks are proposed to apply in different tasks. In [10], Pyramidal LSTM was proposed to solve the person re-identification.

2.2. Pose Detection

Pose estimation is a traditional research [11] in computer vision, and nowadays lots of deep learning methods [12] are proposed to solve this question. Recently, [13, 14] proposed some methods to combine the multi-scale features and learning mechanisms.

2.3. Self-Attention

The self-attention mechanism is soft attention [15] that uses attention scores to measure all salient features. The main calculation of the self-attention network is matrix multiplication without iteration, meaning that it can be easily accelerated. Recently, more and more tasks have successfully applied the self-attention mechanism and achieved excellent performance.

2.4. Person Re-identification

In recent years, many state-of-the-art Re-ID results are relying on features embedding deep learning. Due to its superior performance. Deep learning based methods have been demonstrating the re-id community in the past several years. In the two early Works [8, 16], two images are feed into the siamese model at the same time to determine if there is the same person appeared in both of the images. After these works, approaches were improved in different Ways by [17, 18, 19]. After that, deep learning methods become a popular option in re-ID.

3. The Proposed Re-ID Method

We introduce a new deep framework for the person re-identification task. As shown in Fig. 1, our framework has global and local branches. And the local branch contains three main parts such as pose estimation, representation learning and self-attention mechanism. We will give more details of each part in the following subsections.

3.1. Framework

We can get the overall structure information of the framework for our proposed deep Re-ID from Fig 1. For one training sample, the input images are sent into two branches.

The global branch is intended to extract features with global image information. In this branch, we feed the images into Resnet-50 structure and obtain global features $F_{global}$ from the 512-dimensional fully connected layer results at the end of this branch.

In the local branch, the image is transmitted into a pre-trained pose estimation network to extract key-points of the human body. We use the obtained human body key-points to locate and separate the
original image to get the sub-image we need. Then we use Residual blocks from Resnet to extract the sub-feature group. After that, with the sub-feature group as input, we can get self-correlation features from the self-attention network. Finally, through the LSTM structure, we will take the spatial-temporal self-attention features of the sub-images.

After getting all the results of the two branches, we concatenate the global features, sub-features, and self-attention features as the final feature results.

**Figure 1.** Overview of our framework. We take the query image as input. Our algorithm consists of two branches, a global branch and a sub-feature branch, resulting in global features, sub-features, and self-attention features.

### 3.2. Pose estimation

In the pose estimation task, given a RGB image \( I \) containing a human, we aim to estimate a set of joint coordinates \( Y \in \mathbb{R}^{2D} \). These joints \( J \) composed as the human skeleton \( Y_{2D} \in \mathbb{R}^{J \times 2} \), where the \( J \) is the number of human joint. The two coordinates are pixel coordinates in \( I \).

In our approach, this paper adopts an off-the-shelf model named Stacked Hourglass Networks [14]. This method produces 16 key-points for each Re-ID image prediction. These points divide the image into three parts, head-shoulder, upper body and lower body. According to the selected key-points, we can roughly divide the original image into three parts, in which redundant key-points ensure that we don't lose the required components while dealing with human images in different positions.

Suppose the image \( I_{H \times W} \) with size \( H \times W \), for all the anchor points in each part, \( P_{1:N} = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \). We crop the regions with Eq. i.e.

\[
R^{S, u, d} = \left[ y_c - \frac{w}{2}, y_c + \frac{w}{2} \right] \\
y_c = \frac{y_{\min} + y_{\max}}{2} \\
w = \max(\max(y_1, y_2, \ldots, y_N) - \min(y_1, y_2, \ldots, y_N), \epsilon)
\]

The \( \epsilon \) is minimum dicing height to prevent inaccurate key-points locating. On the general 128 \( \times \) 64 pixel size dataset, \( \epsilon \) takes 32 as the minimum value.

After obtaining the sub-image from input image, we unify the image size to 64 \( \times \) 64 pixels, so that the convolutional neural network group can easily extract the sub-features. Since the texture information of the image needs to be obtained here, we adopt the design of the stacked residual blocks in Resnet. Since the texture features at different locations have similarities, the network parameters between different stacked residual blocks can be shared. Reducing the quantity of parameters can help the network to converge stably. Here we obtain the local sub-feature group named \( F_{\text{sub}} \).

### 3.3. Self-attention Mechanism
Attention mechanism is always used to weight the feature maps, it highlights the key-points of human and alleviates the weights of backgrounds. In self-attention module, we introduce the dual self-attention network structure improved on [20]. Here the sub-features are fed into two similar modules. In the spatial self-attention module, the feature $F_{sub}$ is reduced by a convolutional layer to obtain the feature $F_{sub}'$ of the $H \times W \times C$ dimension. $F_{sub}'$ is reshaped and multiplied with a matrix of dimension $H \times W$ to get the spatial domain attention weight, and broadcast back to the input size. Then $F_{sub}'$ is multiplied with the weight to obtain the attention feature. Structure of the channel self-attention Module is similar to the spatial one’s, only weight matrix is changed to $C \times C$.

3.4. Spatial-Temporal Features

Based on the pose detection method, we obtain different parts of the human in the images. The global representation features of the image focus more on global information such as the appearance, color and pose. In such cases, only some parts are different such as the head, upper body or legs and shoes. For solving these kinds of situations, it has been proposed many methods based on part-alignment representation structure. The part-based local representation of human is an effective complementary to the global information. Then, through integrating the global and LSTM features, we finally get a spatial-temporal discriminative features for our person re-identification.

In our Re-ID framework, we choose bidirectional LSTM as the concrete implementation. The Bidirectional LSTM network we designed include three layers. The first and the third layer network are forward, and the second is reverse. Bidirectional LSTM inputs 512-dimensional sub-features group generated by the self-attention model, and outputs 512 dimension features $F_{att}'$ which contains spatial-temporal information from self-attention and bidirectional LSTM. After that, $F_{att}'$ concatenate with $F_{global}$ and $F_{sub}$ as the final feature $F_{final}$ of the input image.

4. Experimental Results

To evaluate the effectiveness of our approach on person re-identification performance, we conduct several groups of experiments on CUHK03-NP, Market-1501, and DukeMTMC-reID

4.1. Baseline

When choosing the baseline, we compared the identification model structure with the verification model structure. Overall, the two can achieve similar performance. The verification method incorporates hard-mining acceleration training, but even this identification model structure still has an advantage in convergence speed and total training time. This may be related to the former using all the label information in training, while the latter is only related to the weak label information. [8].

So in this paper, we use the identification model structure as the form of the baseline. The baseline model uses the Residual-50 structure. After the original fully connected layer, another fully connected layer of the category of the training set is connected for the softmax loss calculation. In the baseline training process, we started training from the ImageNet pre-trained network with the initial learning rate set to 0.1 for a total of 60 epochs.

4.2. Implementation Details

In our experiments, the image data uniformly reshape to 256 * 128 pixels in preprocessing for ease of use. And we simply use flipping and random-cutting to augment data.

For the global branch, we use Resnet-50 network as the backbone pre-trained on ImageNet, replace the max pooling layer with average pooling layer, and training with the classification loss. In the local branch, we predict 14 key-points for each pedestrian image. The sub-images are extracted by the shared stacked residual block to obtain 128-dimensional $F_{sub}$. Next, the $F_{sub}$ send to the dual self-attention network to extract the attention information and output 512-dimensional $F_{sub}'$. The features then flow into the 3-layer Bidirectional LSTM network to obtain 512-dimensional $F_{att}'$.

All experiments performed on Pytorch. In training, we set batch size to 32, max epoch 60, pre-training part learning rate 0.01, and the remaining part learning rate 0.1. Training is optimized by
SGD optimizer, which weight decay set to $5 \times 10^{-4}$ and momentum set to 0.9. All training and verification perform on one GTX 1080Ti graphics card.

4.3. Evaluation with Baseline
We validated our algorithm performance on the Market-1501, DukeMTMC-reID, and CUHK03-NP datasets with default training-testing separation scheme using indicators Rank-1 and MAP. In each training and testing experiment, we adopt the default hyper parameters without specific adjustments. Results is shown in Table 1.

|                | Market-1501 |            | CUHK03-NP(manual labeled) |            |
|----------------|-------------|------------|--------------------------|------------|
| Method         | Rank 1      | MAP        | Rank 1                   | MAP        |
| baseline       | 0.879       | 0.700      | baseline                 | 0.477      |
| **Ours**       | **0.909**   | **0.750**  | **Ours**                 | **0.489**  |

|                | CUHK03-NP(auto labeled) |            | DukeMTMC-reID | Rank 1 | MAP |
|----------------|-------------------------|------------|--------------|--------|-----|
| baseline       | 0.419                   | 0.388      | baseline     | 0.761  | 0.582|
| **Ours**       | **0.440**               | **0.404**  | **Ours**     | **0.773** | **0.604** |

4.4. Comparison with Related Works
We compare our method with several state-of-the-art methods on Market-1501 dataset. As shown in Table 2., our method achieves competitive results with the state of the art.

| Method         | Rank 1 | MAP | Method         | Rank 1 | MAP |
|----------------|--------|-----|----------------|--------|-----|
| BoW+kissme     | 0.444  | 0.208| PAN            | 0.827  | 0.634|
| MR CNN         | 0.456  | 0.266| Transfer       | 0.837  | 0.655|
| Gated SCNN     | 0.659  | 0.396| DML            | 0.877  | 0.688|
| SOMAnet        | 0.739  | 0.479| SVDNet+REDA    | 0.871  | 0.711|
| ReRank         | 0.771  | 0.636| baseline       | 0.879  | 0.700|
| Verif-Identif  | 0.795  | 0.599| **Ours**       | **0.909** | **0.750** |

5. Conclusions
In this paper, we propose a new network framework for the pedestrian re-identification task. Expand from the global branch, we introduce a new branch based on the self-correlation mechanism and LSTM network structure to extract local features. This network can be implemented with limit hardware resources and trained stably. For verification, we performed experiments on Market-1501, CUHK03-NP and DukeMTMC-reID, and achieved visible performance improvements in pedestrian re-identification tasks.

In the future, we will focus on the video field and improve the performance and speed by improving the structure.

References
[1] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[2] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[3] Iandola F, Moskewicz M, Karayev S, et al. Densenet: Implementing efficient convnet descriptor pyramids[J]. arXiv preprint arXiv:1404.1869, 2014.
[4] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.
[5] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.

[6] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 152-159.

[7] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision. Springer, Cham, 2016: 17-35.

[8] Zheng L, Huang Y, Lu H, et al. Pose invariant embedding for deep person re-identification[J]. arXiv preprint arXiv:1701.07732, 2017.

[9] Su C, Li J, Zhang S, et al. Pose-driven deep convolutional model for person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3960-3969.

[10] Zhang X, Luo H, Fan X, et al. Alignedreid: Surpassing human-level performance in person re-identification[J]. arXiv preprint arXiv:1711.08184, 2017.

[11] Cheng D S, Cristani M. Person re-identification by articulated appearance matching[M]//Person Re-Identification. Springer, London, 2014: 139-160.

[12] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1653-1660.

[13] Tompson J J, Jain A, LeCun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]//Advances in neural information processing systems. 2014: 1799-1807.

[14] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision. Springer, Cham, 2016: 483-499.

[15] Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding[J]. arXiv preprint arXiv:1703.03130, 2017.

[16] Byeon W, Breuel T M, Raue F, et al. Scene labeling with lstm recurrent neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3547-3555.

[17] Ustinova E, Ganin Y, Lempitsky V. Multi-region bilinear convolutional neural networks for person re-identification[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017: 1-6.

[18] Varior R R, Shuai B, Lu J, et al. A siamese long short-term memory architecture for human re-identification[C]//European Conference on Computer Vision. Springer, Cham, 2016: 135-153.

[19] Zheng L, Yang Y, Hauptmann A G. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.

[20] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[J]. arXiv preprint arXiv:1809.02983, 2018.