Unsupervised Spatial-Temporal Model Based on Region Alignment for Person Re-identification

Wei Li\textsuperscript{1,a,*}, Meibin Qi\textsuperscript{1,b}, Ning Yang\textsuperscript{1,c}, Guowu Zhou\textsuperscript{2,d} and Yubing Yang\textsuperscript{2,e}

\textsuperscript{1}DSP Laboratory, School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China
\textsuperscript{2}ChiZhou Public Security Bureau, ChiZhou, 247100 China

\textsuperscript{a}lw2017110926@163.com; \textsuperscript{b}qimeibin@163.com; \textsuperscript{c}hfutxcdq123@126.com; \textsuperscript{d}119816599@qq.com; \textsuperscript{e}15105662588@qq.com

Abstract. Person re-identification(re-ID) refers to find a specific pedestrian across disjoint camera views. Recently, person re-ID rely on supervised learning to train network by labeled information. Resulting poor generalization in real-world environment because of the lack of pedestrian labels. At the same time, person images are easily affected by background, illumination and pose variations. And these factors make it difficult to extract discriminative features to distinguish different pedestrians. In order to resolve this research problem, we proposing an unsupervised learning alignment method called Region Alignment of Spatial-Temporal Fusion(RASTF) which joints global features with local aligned features to get more discriminative features. Local features are aligned by calculating the shortest distances between regions. Our proposed framework integrates a novel region alignment method in unsupervised network and the experiment results indicate that can outperform the state-of-the-art unsupervised methods.

1. Introduction

Person re-ID is a very challenging task in recent year. When the video surveillance is applied to the real scenarios, accurately identifying a specific pedestrian is an important cornerstone to solve the re-identification problem. The recognition results mainly affected by pedestrian appearance, and other pedestrians with similar clothing. Therefore, there are still many problems to be solved in person re-ID. To address these problems, the current person re-ID algorithms are divided into traditional methods \cite{1,2,3} and deep learning methods\cite{4,5}. The traditional person re-ID method mainly focuses on learning discriminative descriptors to represent different people’s appearance and learning to enlarge the distances between images of the different persons and to reduce the distances between images of the same persons. However, it could be hard to extract more discriminative information with traditional feature methods and traditional distance measurement algorithms have complex computational processes, which failed to achieve the need of research. In contrast, deep learning can extract deeper semantic features through training, which makes deep learning methods be a mainstream.

Recently, many deep learning studies have focused on semi-supervised learning and unsupervised transfer learning, such as SPGAN\cite{5}, and TAUDL\cite{6}. However, unsupervised person re-ID also face pedestrian misalignment which is a very challenging problem. As a person appearance in pose variations and detection algorithm results are imperfect, the body parts of each person image are
usually not aligned. In order to solve these problems, The algorithms[7,8] use the body joint point
detection method to align the pedestrians. Although it is effective, but requires additional supervision
information to train and the result of the body joint point must be accurate. Considering the above
person re-ID insufficient, an effective part-aligning method is obliged to be proposed.

To solve above problems, our contributes are as follows:
1. This paper provide a new unsupervised network called RASTF and propose a local alignment
method, which solves the misalignment problem of discuss above.
2. This paper adds a local alignment branch to the network in the training stage. The misalignment
problem in unsupervised learning can be solved in this way. Local feature learning aligns local
features by calculating the shortest distance between them, which greatly improves the efficiency of
global feature learning.
3. A large number of experiments are carried out on two public datasets. the experiment results
show the RASTF framework presents competitive recognition rate compared to the most advanced
methods. Therefore, our contribution is valuable and we conclude this algorithm has potential to be
more robust and effective in practical applications.

![Figure 1. A Framework of RASTF Network](image)

2. The Proposed Method
This paper proposes the RASTF network and its framework is shown in Figure 1.

2.1. Framework
To extract the person feature, we choose ResNet-50[9], the network weights has been trained on
ImageNet[10], and then use the Siamese network[11] to fine-tune ResNet-50 on the source dataset by
measuring the matching probability of two input images. Given a probe image $i_S$, the fusion model
selects one image as a positive sample $s^+_{i_S}$ from the first n images of the ranking result, and selects the
other image as a negative sample $s^-_{i_S}$ from the range of the (n, 2n) of the ranking result. Then the
global branch network extracts the global image features, and the local branch network extracts the
local image features, in which the alignment method is applied to solve the misalignment problem.
Finally, triplet loss is used to optimize the network weight and the global branch learns the ability to
extract accurate local feature. In this work, the weights of the ResNet-50 are shared with the Siamese
network, which not only can save training time and resources and prevent over-fitting from being
trained, but also can make the network training better. At the same time, the updated weights trained
by Siamese network are shared with the fusion model.
In the global branch, the image feature maps are extracted by ResNet-50. Then the feature maps are converted into single column vectors as global features by using global average pooling and the euclidean distance of positive and negative sample pairs are calculated respectively, the loss denoted as Loss_\text{f}. In the local branch, feature maps with size of $7 \times 7 \times 2048$ are converted into $7 \times 1 \times 512$ by convolution layers and horizontal pooling layers. It means that segmenting the image to 7 parts (T=7). The Euclidean distances of local regions of positive and negative pairs are calculated respectively. We use the Algorithm 1 (Section 2.3) to get a shortest distance. In our paper, the local and the global loss are merged by add and the network training is performed by the triplet loss.

### 2.2. Structure of Baseline in RASTF

As shown in Figure 1, the main framework of the RASTF network in this paper is based on the TFusion [12]. The framework makes use of the information of the target datasets and merges the spatio-temporal information of unlabeled target datasets with a classifier $\lambda$, which directly transfers the source datasets model to the target datasets model. Fusion model is a spatio-temporal fusion model of Bayesian probability inference, based on the fusion of visual and spatio-temporal features of the target dataset.

### 2.3. Region Alignment Detail

As we all know, the misalignment of the images is common in datasets and these misalignment can affect the recognition rates. We proposed a novel alignment algorithm base on the idea of literature[13] to address this problem. The algorithm mainly makes use of the relationship between horizontal picture strips. Firstly, need to determine the correspondence between the local regions of the two pictures by the shortest path method. Specifically, the idea of dynamic programming is determine the alignment relationship between all sub-regions of two images from top to bottom, which is traversing all the values of the distance matrix formed by the two images. And the shortest distance $\Theta_{T-1, T-1}$ between the two images can be obtained. Secondly, need to judge whether the path contains irrelevant areas based the shortest distance $\Theta_{T-1, T-1}$. Finding the starting position of the head of the target pedestrian at first. If the head regions are both located in the first area of the two images, it indicates that there is no invalid correspondence in the path, and the final shortest distance between the two pictures is $\Theta_{T-1, T-1}$, which is also the local loss Loss_\text{p}. On the contrary, it means that the shortest distance contains the invalid distance value. In order to remove these invalid distance values, the last region that contains the pedestrian’s head need to be found and then the invalid items in the path are determined. The resulting distance value can be obtained by removing the values of the invalid items, denoted as Loss_\text{p} between two images. As shown as in Figure 2, the first region of image A corresponds to the third region of image B, and the distance between the first region of A and the first region of B and the distance between the first region of image A and the second region of B are removed. The distances between irrelevant areas represented by variable $\gamma$ in the Alg.1. If the first region of the picture B is the head of the pedestrian, the conditional judgment is not established and $\gamma$ is set to 0, which indicates that there is no corresponding unrelated area on the pedestrian picture. Finally, we define Loss_\text{p} as shortest distance in distance matrix $\Psi$. The relationship of image alignment is shown in Figure 2.

---

**Algorithm 1: Region Alignment Processing**

Input: Image A = \{ $\partial_1$, $\partial_2$, $\partial_3$ ... $\partial_T$ \} and Image B = \{ $\beta_1$, $\beta_2$, $\beta_3$ ... $\beta_T$ \}, $\partial$, $\beta$ represents local features of image A and B respectively.

Output: Shortest distance Loss_\text{p}.

1. $i = 0$, $j = 0$, $k = 0$, $\gamma = 0$;
2. Initialize matrix $\Theta$ to zero, size is $T \times T$;
3. Compute the distance matrix $\Psi$ of each part in image A and B and normalize the value to [0,1];
4. for i = 0 → T-1 do
5.     for j = 0 → T-1 do
6.         if i == 0 and j == 0 then
7.             \( \Theta_{i,j} \leftarrow \Theta_{i,j} + \Psi_{i,j} \)
8.         else if i == 0 and j > 0 then
9.             \( \Theta_{i,j} \leftarrow \Theta_{i-1,j} + \Psi_{i,j} \)
10.        else if i > 0 and j == 0 then
11.            \( \Theta_{i,j} \leftarrow \Theta_{i,j-1} + \Psi_{i,j} \)
12.        else if i > 0 and j > 0 then
13.            \( \Theta_{i,j} \leftarrow \min (\Theta_{i-1,j}, \Theta_{i,j-1}) + \Psi_{i,j} \)
14.            if \( \min (\Theta_{i,j}, \Theta_{i-1,j}) = \Theta_{0,j} \) then
15.                Save \( \Theta_{0,j} \) to \( \gamma \)
16.            \( \gamma \leftarrow \gamma + \Theta_{0,j} \)
17.            else if \( \min (\Theta_{r,j}, \Theta_{r,j-1}, \Theta_{r,j-1}) = \Theta_{r-1,j} \) then
18.                Save until to \( \Theta_{r-1,j-1} \)
19.            for l = j → T-1 do
20.                \( \gamma \leftarrow \gamma + \Theta_{r,l} \)
21.        Assign the last element of matrix \( \Theta \) to \( \text{Loss}_p \)
22.        Loss p \( \leftarrow \) Loss p - \( \gamma \)

3. Experiment Result

3.1. Datasets
In this work, the datasets are Market-1501[14], GRID[15], VIPeR[16], DukeMTMC-reID[17]. Market-1501 is composed of 32668 images of 1501 identities and each identity is captured by one camera with low resolution and five cameras with high resolutions. DukeMTMC-reID contains 1404 pedestrians, which are collected by 8 cameras with 16522 images to train, 2228 query images and 17661 gallery images. GRID contains 1275 images, which comes from two cameras with non-overlapping views, including 250 pedestrians. VIPeR captures images in different illumination with two different angle cameras, including 632 pedestrians.

![Figure 2](image)

**Figure 2.** Example of the proposed region aligned method. Each endpoint represents an distance value.

3.2. Performance Comparison
We compare the RASTF Network with other algorithms on Market-1501. Through table 1, the Rank-1 of our algorithm is increased by 6.55% and the mAP is increased by 1.2% compared with SPGAN. The final results are much better than unsupervised algorithms SPGAN which is dataset background style transfer unsupervised learning methods.
Table 1. Comparison with the unsupervised methods in Market-1501 dataset.

| Method     | Rank-1 | mAP  |
|------------|--------|------|
| CAME[18]   | 54.5   | 26.3 |
| SPGAN+LMP[5] | 57.7   | 26.7 |
| TJ-AIDL[3] | 58.2   | 26.5 |
| HHL[19]    | 62.2   | 31.4 |
| TAUDL[6]   | 63.7   | 41.2 |
| Ours       | 64.3   | 25.7 |

We compare the RASTF Network proposed in this paper with other algorithms on DukeMTMC-reID. From table 2, Rank-1 result of this paper is 52.43% and mAP of this paper is 24.11%, which exceeds other unsupervised methods. Rank-1 accuracy increases by 6.03% compared to SPGAN, the result of RASTF also has a great improvement.

Table 2. Comparison with the unsupervised methods in DukeMTMC-reID dataset.

| Method     | Rank-1 | mAP  |
|------------|--------|------|
| TJ-AIDL[3] | 44.3   | 23   |
| PUL[20]    | 44.7   | 20.1 |
| SPGAN[5]   | 46.4   | 26.2 |
| HHL[19]    | 46.9   | 27.2 |
| Ours       | 52.43  | 24.11|

3.3. Comparison of Alignment effectiveness

In the experiment, we use our alignment algorithm to have a comparison with the alignment method of [13]. As shown in table 3. The Rank-1 result of our method is 60.38% and 60.48% when source datasets are VIPeR, GRID and target dataset is Market-1501, increased by 0.84% and 1.17%, respectively. This shows the effectiveness of our alignment algorithm. Therefore, the improvement measures to literature [12] is effective for promoting re-ID recognition rates.

Table 3. Comparison of alignment effectiveness on Market-1501 dataset.

| Method     | Source | Rank-1 | Rank-5 |
|------------|--------|--------|--------|
| TFusion[12]+[13] | GRID | 59.31  | 72.36  |
|             | VIPER | 59.54  | 73.31  |
| Ours       | GRID  | 60.48  | 72.93  |
|            | VIPER | 60.38  | 73.78  |

4. Conclusion

In this paper, our work mainly focuses on unsupervised local alignment methods. The RASTF network is better solve the challenges of occlusion and inaccurate detection of borders. The experimental results show our algorithm significantly improves the performance and has the possibility of being applied to actual scenes. However, our method of obtaining local regions of images is relatively simple. Future work can accurately locate the various regions of the pedestrian in the picture. At the same time, the more effective unsupervised person re-ID benchmark model need to further research, making our network more simple and efficient for larger dataset tasks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China Grant (61771180, 61876056), Anhui Province Key Research and Development Projects (1704d0802183).
References

[1] Liao, S.; Hu, Y.; Zhu, X. & Li, S. Z. (2015). Person re-identification by Local Maximal Occurrence representation and metric learning., in 'CVPR', IEEE Computer Societ.

[2] Chen D, Yuan Z, Chen B, et al. Similarity learning with spatial constraints for person re-identification[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 0.

[3] Gou M, Xiong F, Camps O, et al. Person Re-identification Using Kernel-Based Metric Learning Methods[C]// Computer Vision–ECCV 2014. Springer, Cham, 2014.

[4] Wang J, Zhu X, Gong S, et al. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification[J]. 2018.

[5] Deng W, Zheng L, Ye Q, et al. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification[J]. 2017.

[6] Li, Minxian & Zhu, Xiatian & Gong, Shaogang. (2018). Unsupervised Person Re-identification by Deep Learning Tracklet Association.

[7] Zhao, Haiyu & Tian, Maoqing & Sun, Shuyang & Shao, Jing & Yan, Junjie & Yi, Shuai & Wang, Xiaogang & Tang, Xiaou. (2017). Spindle Net: Person Re-id with Human Body Region Guided Feature Decomposition and Fusion.

[8] L. Wei, S. Zhang, H. Yao, W. Gao and Q. Tian, "GLAD: Global–Local-Alignment Descriptor for Scalable Person Re-Identification," in IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 986-999, April 2019.doi: 10.1109/TMM.2018.2870522

[9] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016.

[10] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems.

[11] Zheng Z, Zheng L, Yang Y. A Discriminatively Learned CNN Embedding for Person Re-identification[J]. Acm Transactions on Multimedia Computing Communications & Applications, 2016, 14(1).

[12] Lv J, Chen W, Li Q, et al. Unsupervised Cross-dataset Person Re-identification by Transfer Learning of Spatial-Temporal Patterns[J]. 2018.

[13] Zhang X, Luo H, Fan X, et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification[J]. 2017.

[14] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1116-1124.doi: 10.1109/ICCV.2015.133

[15] Loy C C, Xiang T, Gong S. Multi-camera activity correlation analysis[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.

[16] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking.

[17] Ristani E, Solera F, Zou R S, et al. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking[J]. 2016.

[18] Yu H X, Wu A, Zheng W S. Cross-view Asymmetric Metric Learning for Unsupervised Person Re-identification[J]. 2017.

[19] Zhong Z., Zheng L., Li S., Yang Y. (2018) Generalizing a Person Retrieval Model Hetero- and Homogeneously. ECCV 2018. Lecture Notes in Computer Science.

[20] H. Fan, L. Zheng, Y. Yang, Unsupervised person re-identification: Clustering and fine-tuning, CoRR abs/1705.10444.