Learning Instance-level Spatial-Temporal Patterns for Person Re-identification

Min Ren¹ ², Lingxiao He³, Xingyu Liao³, Wu Liu³, Yunlong Wang², Tieniu Tan²
¹University of Chinese Academy of Sciences
²CRIPAC NLPR, Institute of Automation Chinese Academy of Sciences
³JD AI Research
{min.ren, yunlong.wang} @cripac.ia.ac.cn, {helingxiao3, liaoxingyu5, liuwu1} @jd.com, tnt@nlpr.ia.ac.cn

Abstract

Person re-identification (Re-ID) aims to match pedestrians under dis-joint cameras. Most Re-ID methods formulate it as visual representation learning and image search, and its accuracy is consequently affected greatly by the search space. Spatial-temporal information has been proven to be efficient to filter irrelevant negative samples and significantly improve Re-ID accuracy. However, existing spatial-temporal person Re-ID methods are still rough and do not exploit spatial-temporal information sufficiently. In this paper, we propose a novel Instance-level and Spatial-Temporal Disentangled Re-ID method (InSTD), to improve Re-ID accuracy. In our proposed framework, personalized information such as moving direction is explicitly considered to further narrow down the search space. Besides, the spatial-temporal transferring probability is disentangled from joint distribution to marginal distribution, so that outliers can also be well modeled. Abundant experimental analyses are presented, which demonstrates the superiority and provides more insights into our method. The proposed method achieves mAP of 90.8% on Market-1501 and 89.1% on DukeMTMC-reID, improving from the baseline 82.2% and 72.7%, respectively. Besides, in order to provide a better benchmark for person re-identification, we release a cleaned data list of DukeMTMC-reID with this paper: https://github.com/RenMin1991/cleaned-DukeMTMC-reID/

1. Introduction

Person re-identification aims to retrieve pedestrians across non-overlapping camera views. Most existing person re-identification methods focus on the visual feature representations of pedestrian images [6, 11, 14, 27, 32, 35, 38, 36, 42], such as appearance, clothes, and textures. The auxiliary information of person images is also adopted recently, such as parsing information [5, 13, 15, 22, 25], pose of the pedestrians [3, 19, 20], or human body key points [33]. However, the performances of these methods are still far from the requirements of real-world situations. Because it is hard for visual representations to discriminate pedestrian with similar appearance and clothes.

Recent methods model spatial-temporal patterns [8, 18, 21, 34] to filter out the irrelevant candidates and narrow down the search space. Specifically, these methods mainly formulate spatial-temporal pattern as a joint distribution $P(S_{c_i,c_j}, T)$, where $S_{c_i,c_j}$ means moving from camera $i$ to camera $j$, $T$ means time interval. It has been proven to be efficient to significantly improve re-identification accuracy. However, there are two problems of the existing methods. Firstly, the existing spatial-temporal methods only consider camera-level but neglect instance-level information. The state information of each pedestrian is neglected while it is essential for spatial-temporal patterns of the person. Secondly, existing methods formulate spatial-temporal patterns as a joint distribution, meaning that only those candidates

---

*This work is done when Min Ren is an intern at JD AI Research.
matching both spatial and temporal priors can be matched. They are not robust to the outliers.

To solve these problems, we propose a novel method named Instance-level and Spatial-Temporal Disentangled Re-ID (InSTD) to model the instance-level and spatial-temporal disentangled patterns. Firstly, the traditional spatial-temporal pattern is updated to be conditional on instance-level state information. Its formulation looks like $p(S_{c_i,c_j}, T | P)$, where $P$ is instance-level pedestrian information. The walking direction of the pedestrian, which is the key instance-level state information, is taken into consideration in this paper. The walking direction of a pedestrian is complimentary information of pedestrian detection and tracking. It is useful because it is highly correlated with spatial-temporal patterns. For example, a pedestrian, who is walking towards the west in the view of a camera, is more probable to appear in the view of the western cameras later, rather than the eastern cameras. Meanwhile, it is economical because pedestrian detection and tracking are necessary steps before person re-identification in practice.

Secondly, we disentangle the spatial-temporal pattern by constructing their marginal distribution, i.e., transmission probability $P(S_{c_i,c_j}, P)$ and time interval distribution $P(T | P)$. They are modeled separately and adaptively combined to handle outliers. If the temporal (spatial) pattern of a pedestrian is unusual, the person may be normal in the term of spatial (temporal) patterns. The similarity metric should focus on the spatial (temporal) pattern. For example, a runner, who is moving faster than most pedestrians, is an outlier from the view of temporal pattern. But the runner can be quite normal in terms of spatial transmission perspective. It is harmful to model this runner by joint distribution of spatial and temporal patterns. To this end, we propose a novel fusion approach to adaptively combine the spatial and temporal patterns. The spatial patterns and temporal patterns are complementary, rather than in conflict as existing methods, so that outliers can also be well modeled.

The contributions of this paper can be summarized as follows:

- We present a novel instance-level method to model spatial-temporal patterns for person re-identification. The proposed method provides personalized predictions by leveraging the instance-level state information of each pedestrian.

- The instance-level spatial-temporal patterns are decoupled into transmission probabilities and time interval distributions between cameras in the proposed method. The spatial and temporal patterns become complementary rather than in conflict as existing methods.

- Without bells and whistles, the proposed method surpasses the baseline model based on visual features by 16.9% on DukeMTMC-reID and 8.6% on Market-1501 in the term of mAP, and outperforms the state-of-the-art method based on spatial-temporal patterns by 4.8% on DukeMTMC-reID and 2.2% on Market-1501.

2. Related Work

2.1. Visual Features based Re-ID

Person re-identification addresses the problem of matching pedestrian images across non-overlapping camera views [29, 40]. Many studies exploit discriminative visual features [2, 26, 37].

Deep learning algorithms foster significant improvements in the field of person re-identification. Some researchers attempt to explore effective convolutional neural networks [1, 4, 6, 9, 16, 17, 10, 30, 31, 35, 36, 41, 44]. Some studies explore training strategies and loss functions for person re-identification [1, 9, 10, 31, 43]. Recently, some studies leverage the structure information of person images, such as parsing information [5, 15, 22, 25], pose of the pedestrians [3, 19, 20, 24], or human body key points [33].

However, appearance-based methods are still far from practical applications. They are not discriminative enough in complex scenarios where pedestrians may exhibit similar appearance and clothes. It is hard to further improve the performance using only appearance-based features.

2.2. Spatial-temporal Person Re-ID

There are some researchers who have paid attention to the topology of cameras since the spatial-temporal patterns implied in the topology are essential for cross-camera retrieval. The spatial-temporal constraints are utilized to filter out the irrelevant gallery images [8, 18, 21, 34]. Huang et al. [34] propose a method to take both visual feature representation and spatial-temporal constraints into consideration for person re-identification. However, this method makes a strong assumption that the time intervals between cameras follow Weibull distribution. This assumption is invalid for complex scenarios. Cho et al. [21] propose a framework to integrate camera network topology into person re-identification. However, the temporal constraints are simply realized by time thresholds, which cannot handle massive gallery images and complex cases in practice. Lv et al. [18] propose a method that leverage the spatial-temporal constraints for cross-dataset person re-identification. The spatial-temporal constraints improve the performance on the target dataset by enhancing the pseudo label during training. It is not proper to be directly applied to general person re-identification tasks. Wang et al. [8] propose a two-stream architecture to apply spatial-temporal constraints to person re-identification. However, the spatial-temporal constraints are coupled together in this method.
Figure 2. Spatial-temporal patterns are implied in the topology of cameras. The spatial-temporal pattern between two cameras of a pedestrian are highly correlated with his/her moving direction. Pedestrians in the view of camera 2 may appear in camera 1, camera 3, or camera 5 after a certain time lapse. However, pedestrians with different states will appear in different cameras at different times. For example, the pedestrian in the red bounding box is much more likely to appear in Camera 3 than Camera 1, because he is moving towards the field of view of Camera 3. In the proposed method, the instance-level state information is adopted rather than modeling the spatial-temporal patterns on camera-level as the existing methods.

which is harmful to recalling positive samples. All these spatial-temporal person Re-ID methods establish the patterns based on the camera-level information, which means they can not provide fine-grained constraints for each instance.

Different from existing spatial-temporal methods, our method models spatial-temporal patterns at the instance-level to filter out more irrelevant gallery images and provide personalized predictions. And the spatial-temporal patterns are decoupled into transmission probabilities and time interval distributions to make them mutually beneficial rather than in conflict.

3. Method

The instance-level spatial constraint i.e. transmission probabilities, and temporal constraint i.e. time interval distributions are detailed separately in this section. Then the adaptive combined metric is presented.

3.1. Instance-level Spatial Constraint

The spatial constraint between two cameras is described by the transmission probability of the cameras, which means how tightly the two cameras correlate. Formally, we model the transmission probability by a conditional probability:

\[ p_{i,j} = \Pr(C_l = j | C_e = i) \]  

3.2. Time Interval Distribution

The instance-level temporal constraints are modeled by the time interval distribution of appearances in multiple cameras. Specifically, the conditional probability is calculated as:

\[ p_{i,j} = \frac{\Pr(C_l = j, C_e = i)}{\Pr(C_e = i)} \]  

The higher the conditional probability means the person in the view of camera \( i \) is more likely to appear in the view of camera \( j \) after a certain time interval.
of camera \( j \) later. The time interval between camera \( i \) and \( j \) is not involved here. Note that \( p_{i,j} \neq p_{j,i} \) in most cases.

However, the spatial patterns of persons appear in the same camera can be different, as shown in Fig. 2. To address this problem, we introduce the instance-level state information of a person into the conditional probability:

\[
p_{i,j}^s = \text{Pr}(C_l = j|C_e = i, s_e = s)
\]

where \( s_e \) is the state of a person in the view of \( C_e \), \( s \in S_i \), \( S_i \) is the set of states:

\[
S_i = \{ s_1, s_2, \ldots, s_{n_i} \}
\]

where \( n_i \) is the number of states of camera \( i \).

The instance-level states are represented by walking directions of pedestrians. For example, the view of the first camera of DukeMTMC-reID [7] is shown in Fig. 3. The state set of this camera contains two states: walking towards the red zone and walking towards the blue zone. The state sets of the rest cameras are defined similarly, and the illustrations of other cameras can be found in the supplementary material.

Hence, the instance-level transmission probability can be calculated:

\[
p_{i,j}^s = \frac{\text{Pr}(C_l = j, C_e = i, s_e = s)}{\text{Pr}(C_e = i, s_e = s)}
\]

### 3.2. Instance-level Temporal Constraint

The temporal constraint is described by the time interval distribution, which represents the time lapse for a pedestrian to transfer between two cameras. Formally, we model the time interval distribution by a conditional probability density function:

\[
f_{i,j}(\delta) = \frac{dF_{i,j}(\delta)}{d\delta}
\]

where \( \delta \) is the transfer time, \( F_{i,j}(\delta) \) is the cumulative distribution function, which is a conditional probability:

\[
F_{i,j}(\delta) = \text{Pr}(\Delta \leq \delta|C_e = i, C_l = j)
\]

It can be harmful to recalling positive samples that fitting \( f_{i,j}(\delta) \) or \( F_{i,j}(\delta) \) into a closed-form probability distribution. Hence, a non-parameter estimation method is adopted in our method. Specially, we use Parzen window with Gaussian kernel to estimate \( f_{i,j}(\delta) \):

\[
f_{i,j}(\delta) = \frac{1}{Z_{i,j}} \sum_n H_{i,j}(\delta)K(n - \delta)
\]

\[
H_{i,j}(\delta) = \begin{cases} 1 & \delta \in D_{i,j} \\ 0 & \text{otherwise} \end{cases}
\]

where \( K(\cdot) \) is the kernel function, \( Z_{i,j} = \sum_{\delta} H_{i,j}(\delta) \) is a normalized factor, \( D_{i,j} \) is the time interval set between camera \( i \) and camera \( j \) of training samples.

However, the time interval distribution of persons appear in the same camera can be different, as we have mentioned before. Similar to the instance-level transmission probabilities, the instance-level state information of a person is introduced into the conditional probability to address this problem:

\[
f_{i,j}^s(\delta) = \frac{dF_{i,j}^s(\delta)}{d\delta}
\]

\[
F_{i,j}^s(\delta) = \text{Pr}(\Delta \leq \delta|C_e = i, C_l = j, s_e = s)
\]

where \( s_e \) is the instance-level state of a person in the view of \( C_e \). The moving direction is also considered as the key state information. Similar with Eq. 8, instance-level time interval distribution \( f_{i,j}^s(\delta) \) can be estimated:

\[
f_{i,j}^s(\delta) = \frac{1}{Z_{i,j}} \sum_n H_{i,j}^s(\delta)K(n - \delta)
\]

\[
H_{i,j}^s(\delta) = \begin{cases} 1 & \delta \in D_{i,j}^s \\ 0 & \text{otherwise} \end{cases}
\]

where the normalized factor \( Z_{i,j}^s = \sum_{\delta} H_{i,j}^s(\delta) \), \( D_{i,j}^s \) is a subset of \( D_{i,j} \), it contains the samples subject to \( s_e = s \).

### 3.3. Joint Metric

Given the transmission probability and time interval distribution affiliated to instance-level state information, the spatial-temporal probability is the fusion of them:

\[
P = F(p_{spa}, p_{tem})
\]

where \( p_{spa} = p_{i,j}^s \) is the instance-level transmission probability of two images in Eq. 5, and \( p_{tem} = f_{i,j}^s(\delta) \) is the instance-level time interval probability in Eq. 10. And the final joint metric of two images is:

\[
S \cdot P = S \cdot F(p_{spa}, p_{tem})
\]

where \( S \) is the visual feature similarity.

A straightforward way to fuse both components is multiplying \( p_{spa} \) and \( p_{tem} \) together. However, the constraint realized by directly multiplying is too strict for person re-identification. If a spatial/temporal pattern of a pedestrian is unusual, the person may be normal in terms of temporal/spatial patterns. This kind of samples should not be removed recklessly. Hence, the spatial-temporal probability \( F(p_{spa}, p_{tem}) \) should be fairly high when only one of them is high. Fusion by multiplying directly is not proper here obviously.

The spatial-temporal probability in our method is defined as:

\[
P = \frac{1}{1 + e^{-(\alpha p_{spa} + \beta p_{tem})}}
\]

where \( \alpha \) and \( \beta \) are scaling parameters of similarity fusion. The spatial and temporal constraints are adjusted by \( \alpha \) and \( \beta \) separately.
The spatial-temporal factor is scaled into $[0.5, 1]$. The constraint is relaxed properly when the spatial-temporal probability is low. And the value of $P$ stays stable when $p_{spa}$ or $p_{tem}$ is low.

### 3.4. Implementation Details

More details are presented in this subsection. The moving direction of a pedestrian is complimentary information of pedestrian detection and tracking, which is a necessary step before person re-identification in practice. There is no need to predict the moving direction by an extra model or manual annotations. In our experiments, the moving directions of samples in the field of one camera are confirmed by tracking them in the original video of this camera. Actually, the moving direction of a pedestrian can be confirmed within five consecutive frames in most cases, which can be easily derived from existing tracking methods.

A pretrained ResNet-50 is adopted as baseline for feature extraction. We set the standard deviation of the Gaussian kernel for distribution estimation to 100. As for the scaling parameters, $\alpha$, $\beta$ in Eq. 16 are set to 0.15 and 1 respectively.

### 4. Experiments

In this section, we evaluate our method on two large scale person re-identification benchmark datasets, i.e., Market-1501 [23] and DukeMTMC-reID [7]. Then, more experimental analysis is presented.

#### 4.1. Datasets and Evaluation Protocol

Market-1501 dataset [23] is collected at a university campus. A total of six cameras are used, including 5 high-resolution cameras, and one low-resolution camera. It contains 32,968 annotated bounding boxes of 1,501 identities, plus a distractor set of over 500K images. The pedestrians are detected by Deformable Part Model (DPM). Among them, the training set consists of 12,936 images from 751 identities, the gallery set contains 19,732 images from other 750 identities and all the distractors. 3,368 hand-drawn bounding boxes from 750 identities are used as the query images. In this dataset, each image contains its camera index and time stamp.

DukeMTMC-reID [7] is a subset of DukeMTMC dataset for image-based person re-identification. There are eight cameras in total. 1,404 identities appear in more than one camera and 408 identities (distractor) appear in only one camera. 702 identities are used for training, and the other 702 identities plus distractors are used for testing. One image for each identity in each camera is picked as a query, and the other images are put into the gallery. Each image contains its camera index and time stamp.

We use two performance indexes as in most person re-identification literature. The first is mean average precision (mAP). The average precision (AP) of a query is the area under the Precision-Recall curve, which means both precision and recall rate is taken into consideration. Hence, the mean average precision among all query images is a comprehensive performance index for person re-identification. The second is the cumulative matching characteristic (CMC) i.e. the top-k accuracy. Hence, the cumulative matching characteristic emphasizes precision rather than recall rate.

#### 4.2. Comparisons to the State-of-the-Art

The proposed method is compared with fourteen existing state-of-the-art methods, which can be categorized into four groups. The first group of methods extract visual features directly from the person images, including PCB [38], VPM [39], and BOT [12]. These methods explore various aspects of visual feature extraction, including the structure of convolutional neural networks, training strategy, data augmentation, and loss function. The second group of methods adopt human parsing information for person re-identification, including SPRiD [25], MGCAM [5], MaskReID [22] and FPR [15]. The third group of methods leverage the pose or key points of person images, including PDC [3], Pose-transfer [20], PSE [28], PGFA [19] and HOREID [33]. The fourth group of methods utilize the spatial-temporal information to enhance the person re-identification, including TFusion-sup [18] and st-ReID [8]. These methods use hard or soft constraints to narrow the number of gallery images.

<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCB [38]</td>
<td>77.4%</td>
<td>92.3%</td>
<td>97.2%</td>
<td>98.2%</td>
</tr>
<tr>
<td>VPM [39]</td>
<td>80.8%</td>
<td>93.0%</td>
<td>97.8%</td>
<td>98.8%</td>
</tr>
<tr>
<td>BOT [12]</td>
<td>85.9%</td>
<td>94.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SPRiD [25]</td>
<td>81.3%</td>
<td>92.5%</td>
<td>97.15%</td>
<td>98.1%</td>
</tr>
<tr>
<td>MGCAM [5]</td>
<td>74.3%</td>
<td>83.8%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MaskReID [22]</td>
<td>75.4%</td>
<td>90.4%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FPR [15]</td>
<td>86.6%</td>
<td>95.4%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PDC [3]</td>
<td>63.4%</td>
<td>84.1%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pose-transfer [20]</td>
<td>68.9%</td>
<td>87.7%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PSE [28]</td>
<td>69.0%</td>
<td>87.7%</td>
<td>94.5%</td>
<td>96.8%</td>
</tr>
<tr>
<td>PGFA [19]</td>
<td>76.8%</td>
<td>91.2%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HOREID [33]</td>
<td>84.9%</td>
<td>94.2%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Baseline</td>
<td>82.2%</td>
<td>93.6%</td>
<td>98.4%</td>
<td>99.0%</td>
</tr>
<tr>
<td>Baseline+st-ReID [8]</td>
<td>88.6%</td>
<td>96.9%</td>
<td>99.2%</td>
<td>99.5%</td>
</tr>
<tr>
<td>Baseline+InSTD</td>
<td><strong>90.8%</strong></td>
<td><strong>97.6%</strong></td>
<td><strong>99.5%</strong></td>
<td><strong>99.7%</strong></td>
</tr>
</tbody>
</table>


which is a ResNet-50, our method improves the mAP by 8.6% on Market-1501 and 16.5% on DukeMTMC-reID, improve the Rank-1 accuracy by 4% on Market-1501 and 10% on DukeMTMC-reID. Our method achieves significant improvements especially in terms of mAP.

Comparing to the methods in the first three groups, the advantage of our method is obvious. Besides, the compared methods in the second and third groups need expensive annotations, such as key points, pixel-wise parsing maps, and masks, to match the query and gallery images. Our method adopts economical information i.e. camera ID, timestamp, and state information.

The disadvantages of the methods in the fourth group have been interpreted in Sec. 2. And the interpretations have been demonstrated by the results of experiments in this subsection. Given the same baseline model, our method outperforms the st-ReID by a remarkable margin, especially in terms of mAP (2.2% on Market-1501 and 4.8% on DukeMTMC-reID). These results indicate that our method has evident advantages over existing spatial-temporal methods.

To show the effect of spatial-temporal constraint, an example from DukeMTMC-reID is presented in Fig. 4. The appearance of the pedestrian in the red bounding box, who is mistakenly ranked first, is similar to the query image. The visual representation cannot distinguish it from the correct identites as shown in Fig. 4 (a). The incorrect pedestrian, which is difficult to discriminate for the visual representation, is filtered out by the spatial-temporal constraint as shown in Fig. 4 (b).

The effect of instance-level information are shown in Fig. 5. The spatial-temporal constrains may be misguided in complex scenarios, as shown in Fig. 5 (a). The instance-level information can make the spatial-temporal constrains more reliable for person re-identification. The incorrect pedestrians are filtered out by the instance-level state as shown in Fig. 5 (b).

4.3. Experiments on Different Feature Extractors

The proposed method can be applied to different feature extractors. To verify its effectiveness, we evaluate the proposed method based on other two feature extractors: PCB [38], and VPM [39].

The results are shown in Tab. 3. Our method consistently improves the performance of all feature extractors. Our method gains significant 20%/11% improvement in mAP/rank-1 accuracy for PCB [38], and 16%+/10%+ improvement for the other two feature extractors. Comparing to st-ReID [8], which is also based on spatial-temporal constraints, our method achieves consistent improvements too. Our method outperforms st-ReID [8] by 4%+/0.6%+ improvement in mAP/rank-1 accuracy for all of the feature extractors.

The results show that our method can be generalized to different feature extractors. Moreover, the results demonstrate the advantages of our method comparing to the existing spatial-temporal based method.

4.4. Analysis of Scaling Parameters

To investigate the impact of two scaling parameters, $\alpha$ and $\beta$ in Eq. 16, we conduct two sensitivity analysis experiments on $\alpha$ and $\beta$. The results are shown in Fig. 6. When
<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCB [38]</td>
<td>66.1%</td>
<td>81.7%</td>
<td>89.7%</td>
<td>91.9%</td>
</tr>
<tr>
<td>PCB [38]+st-ReID [8]</td>
<td>80.9%</td>
<td>92.1%</td>
<td>95.4%</td>
<td>96.6%</td>
</tr>
<tr>
<td>PCB [38]+InSTD</td>
<td>86.1%</td>
<td>92.7%</td>
<td>96.5%</td>
<td>97.6%</td>
</tr>
<tr>
<td>VPM [39]</td>
<td>72.6%</td>
<td>83.6%</td>
<td>91.7%</td>
<td>94.2%</td>
</tr>
<tr>
<td>VPM [39]+st-ReID [8]</td>
<td>84.9%</td>
<td>94.2%</td>
<td>96.1%</td>
<td>96.9%</td>
</tr>
<tr>
<td>VPM [39]+InSTD</td>
<td>89.3%</td>
<td>95.1%</td>
<td>97.0%</td>
<td>97.9%</td>
</tr>
</tbody>
</table>

Table 3. Effects on different feature extractors. The experiments are conducted on DukeMTMC-reID [7].

In the fourth protocol, the instance-level state information is excluded based on the third protocol:

\[
p_{st} = ̂f_{i,j}(δ)
\]  

The results of these four protocols are shown in Tab. 4. The results show that the instance-level state information of pedestrians and the decoupling of spatial and temporal are both useful to improve the performance. Besides, the instance-level state information of pedestrians is more helpful to improve precision by narrow the number of gallery images. The combination of these two strategies achieves the best performance.

To demonstrate the effect of introducing instance-level state information, the time interval distributions between camera 1 and camera 2 of DukeMTMC-reID are shown in Fig. 7. The instance-level states are not taken into consideration in Fig. 7 (a). On the other hand, distributions of two states are shown separately in Fig. 7 (b). The distribution in Fig. 7 (a) is split into two distributions, which means more irrelevant gallery images can be filtered out according to the instance-level state information.

In the coupled case, the spatial pattern is conveyed by the areas under distribution curves as shown in Fig. 8 (a). The areas under distribution curves are the same as shown in Fig. 8 (c), and the spatial pattern is decoupled from the time interval distributions as transmission probabilities.

### 4.5. Ablation Study

The ablation study on the instance-level state information of pedestrians and the decoupling of spatial-temporal patterns is presented in this part.

Four protocols are taken into consideration. The first one is the proposed method itself. In the second protocol, \( p_{spa} \) and \( p_{tem} \) in Eq. 14 are replaced by \( p_{i,j} \) (Eq. 2) and \( f_{i,j}(δ) \) (Eq. 8). The instance-level state information is excluded in this protocol.

In the third protocol, the spatial pattern and temporal pattern are coupled together. The normalized factor in Eq. 12 is replaced by \( ̂Z \):

\[
̂f_{i,j}(δ) = \frac{1}{Z^s} \sum_n H_{i,j}(δ)K(n − δ)
\]

\[
̂Z^s = \max_{i,j} Z^s_{i,j}
\]

which means all time interval distributions share an identical denominator. The numerical relations of the area under curves indicate the transmission probabilities between cameras. And \( p_{spa} \) and \( p_{tem} \) are replaced by \( p_{st} \) Eq. 16:

\[
P = \frac{1}{1 + e^{βp_{st}}},
\]

The results show that the instance-level state information of pedestrians is more helpful to Rank-1 accuracy, and the decoupling of spatial and temporal patterns is more contributive to mAP. These results indicate that the decoupling of spatial and temporal patterns is more helpful to improve mAP by recalling more hard positive samples. The instance-level state information of pedestrians is more helpful to improve precision by narrow the number of gallery images. The combination of these two strategies achieves the best performance.

4.6. Failure Analysis

In this part, we analyze the failure cases of the proposed method on DukeMTMC-reID. We find that the failures can be categorized as four cases:
Figure 7. The time interval distributions between camera 1 and camera 2 of DukeMTMC-reID. (a): Time interval distribution without state information. (b): Time interval distributions with instance-level state information. The distribution in (a) is split into two distributions in (b), which means more irrelevant gallery images can be filtered out according to the instance-level state information.

Figure 8. The time interval distributions of camera 1 to camera 2 and camera 1 to camera 5 of DukeMTMC-reID. (a): Time interval frequencies. (b): Time interval distributions without spatial-temporal decouple. The spatial pattern is conveyed by the areas under distribution curves. (c): Time interval distributions with spatial-temporal decouple. The areas under distribution curves are the same. The spatial pattern is decoupled from the time interval distributions as transmission probabilities. (Instance-level state information is not shown for simplicity.)

Firstly, there are incorrect labels in DukeMTMC-reID. The proportion of failure cases caused by incorrect labels is 16.2%. It is harmful to keep the incorrect labels in the database. Hence, we release a cleaned data list of DukeMTMC-reID with this paper: https://github.com/RenMin1991/cleaned-DukeMTMC-reID/

Secondly, the feature extractor is fooled because of serious occlusions. For example, the upper part of two individuals is quite similar while the lower part is occluded. The proportion of this case in all failures is 56.9%.

In the third case, the visual feature is not discriminative enough to distinguish the hard negative samples. The proportion of this case in all failures is 23.5%.

In the last case, the proposed method outputs high probabilities due to spatial-temporal patterns. However, it improperly pushes up the final joint metric. The proportion of this case in all failures is 3.4%.

The failure analysis shows that serious occlusion is the main cause of mismatching (56.9%). The second important reason is that the feature extracted by the recognition model is not discriminative enough (23.5%). Incorrect labels also degrade the performance (16.2%). The proportion of failure samples caused by the improper spatial-temporal probability in all failures is quite small (3.4%).

5. Conclusion

In this paper, we propose a method to exploit spatial-temporal patterns for person re-identification. Different from the existing spatial-temporal person re-identification methods, the proposed method adopts the walking direction of each pedestrian, as key instance-level state information, to provide personalized predictions. In addition, the spatial-temporal patterns are decoupled into transmission probabilities and time interval distributions between cameras. The spatial-temporal patterns become mutually beneficial rather than in conflict with each other as current methods. A novel joint metric is proposed to fuse the instance-level spatial constraint, temporal constraint, and visual feature similarity. The superiority of our method is demonstrated by extensive contrast experiments. And adequate experimental analyses provide more insights into our method.

Acknowledgments

The authors would like to thank Guanan Wang for his valuable help. This work is funded by National Natural Science Foundation of China (Grant No.62006225, 61806197, 62071468, U1836217) and supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27040700), Tianjin Science and Technology Plan Project (17ZXRGX00110).

References


in Computer Vision & Pattern Recognition, 42(7):301–313, 2014. 2


[34] Huang Wenxin, Hu Ruimin, Liang Chao, Yu Yi, and Zhang Chunjie. Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations. In International Conference on Multimedia Modeling, 2016. 1, 2


[38] Sun Yifan, Zheng Liang, Yang Yi, Tian Qi, and Wang Shengjin. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). European Conference on Computer Vision (ECCV), 2018. 1, 5, 6, 7


14939