A spatial and temporal features mixture model with body parts for video-based person re-identification

Jie Liu · Cheng Sun · Xiang Xu · Baomin Xu · Shuangyuan Yu

Published online: 12 April 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

The goal of video-based person re-identification is to recognize a person at different camera settings. Most previous methods use features from the full body to represent a person. In this paper, we propose a novel Spatial and Temporal Features Mixture Model (STFMM). Unlike previous approaches, our model first horizontally splits human body into N parts, which include the information of head, waist, legs and so on. The feature of each part is then integrated in order to achieve more expressive representation for each person. Experiments conducted on the iLIDS-VID and PRID-2011 datasets demonstrate that our approach outperforms the existing video-based person re-identification methods and significantly improves stability. Our model achieves a rank-1 CMC accuracy of 73.6% on the iLIDS-VID dataset and a rank-1 CMC accuracy of 47.8% for the cross-data testing.

Keywords Video-based person re-identification · Siamese network · Temporal series feature · Partial features

1 Introduction

The person re-identification task is to continuously track a person that appears in non-overlapping cameras at distinct times [4]. Recently, it has drawn a lot of attention due to its applications in security and surveillance systems. Despite many successes, this is still a challenging problem because of the different occlusions, backgrounds, illuminations and view points that appear across multiple frames.

The study of person re-identification in still images is abundant and has advanced a lot. Usually, there are two steps in the processing of still images: feature learning [3, 9, 12, 14, 17, 18, 32] and metric learning [2, 8, 11, 13, 21, 25, 27, 31, 34]. Feature learning focuses on extracting discriminative features and building an invariant representation for each person, such as pose and clothing color. Metric learning, on the other hand, aims to minimize the variance of the same person and maximize that of different ones. Compared to single image, video sequence is much closer to the real-world scenario where data are captured by a surveillance camera. Certainly, video sequence carries more information. For example, a person’s gait can be extracted to help with the accuracy of person re-identification. However, a lot of redundant and distracting information is also contained in the video, such as changing backgrounds and occlusions.

Recently, a growing number of researchers are interested in video-based person re-identification [10, 15, 19, 24, 28, 29, 35]. As deep neural network moves ahead in generic object recognition, it has also been successfully applied in this task. Given a video sequence, convolutional neural network works as a feature extractor for each frame, while recurrent neural network exploits the temporal information from the video sequences. Siamese network [5] is a popular deep neural network architecture in person recognition tasks, with which a lot of methods based on CNN-RNN-scheme have achieved reasonable successes [19, 28, 30]. Since the mode of CNN-RNN-scheme based on Siamese is so popular, we adopt it for video-based person re-identification.

Figure 1 shows the overall pipeline for our proposed STFMM model. The model actively divides the frame into several parts where each part serves as a unique feature for person identification. Compared with previous methods, our approach effectively guides the model to focus on
Aspatial and temporal features mixture model with body parts for video-based person re-identification

Fig. 1 Each sequence is split into \( N \) continuous part sequences, including head part (green box), shoulder part (brown box), ..., and shank part (red box).

different regions that are crucial for identifying a person. This idea is more simple than using an attention mechanism [1, 20] where an extra attention map has to be learned while serving the same purpose so that the model can learn to use different regions from a human for identification. Given a pair of video sequences, we first split each frame into \( N \) parts, and the adjacent parts are overlapped with several pixels. Next we utilize our STFMM method to process part sequences concurrently and calculate the similarity of two corresponding features of each part. Finally, we propose an algorithm to mix the \( N \) part features to generate the discriminative representation of each person. The extensive experiments are carried out on two datasets, iLIDS-VID and PRID-2011. The results demonstrate that our approach outperforms the existing methods for video-based person re-identification.

We summarize the main contributions of this paper as follows:

(i) We propose a novel Spatial and Temporal Features Mixture Model (STFMM) that makes full use of the information from individual human body parts and combine them into a discriminative representation for each person.

(ii) To satisfy the input of our network, we develop the Siamese network architecture in which the number of input is adjusted from 2 to \( 2^N \).

(iii) For the video-based person re-identification of complicated scenes, our proposed model can effectively improves the accuracy.

(iv) By combining the STFMM and developed Siamese network, our method achieves better results and has improved stability to a large degree than state-of-the-art methods.

2 Related work

Traditional methods for person re-identification mainly consist of two aspects: feature learning and metric learning. Many feature learning methods have been proposed for person re-identification in still images or video sequences. Matsukawa et al. [18] presented a descriptor on a hierarchical distribution of pixel features and used Gaussian distribution to describe a local image region. Liao et al. [12] utilized the horizontal occurrence of local features and maximized the occurrence for stable representation, called Local Maximal Occurrence (LOMO). Wang et al. [24] proposed a Discriminative Video Ranking model (DVR) to select the most discriminative video fragments, from which more reliable space-time features can be extracted. The method of Bag-of-Words (BoW) [33] aimed to learn a mapping function that converted frame-wise features to a global vector. The metric learning methods have also been widely invested and made some positive achievements, such as Relaxed Pairwise Learning (RPL) [8], Large Margin Nearest-Neighbour (LMNN) [25], Relevance Component Analysis (RCA) [2], Locally Adaptive Decision Function (LADF) [11], and RankSVM [32].

In recent years, deep neural network (DNN) has achieved significant successes in computer vision, and the DNN-based methods have been studied and applied in person re-identification task [19, 22–24, 26, 28, 29, 33, 35]. The DNN-based models are trained with a pair of inputs for learning a direct mapping from image or video sequence to feature space. McLaughlin et al. [19] combined CNN, RNN and Siamese network together, which is the first time to apply DNN to the video-based re-identification. Attention mechanism [1, 20] has attained huge achievement in deep learning. Xu et al. [28] proposed a joint Spatial and Temporal Attention Pooling Network (ASTPN) to extract sequence-level features by selecting informative frames and notable regions of each frame. Zhou et al. [35] used Temporal Attention Model (TAM) to measure the importance of each frame in video sequence and applied Spatial Recurrent Model (SRM) to explore contextual information.

However, most of the CNN-RNN-based methods mainly lay emphasis on extracting the feature of full body for creating sequence-level representation. Motivated by the traditional methods of GOG [18] and LOMO [12] which make full use of local region information from an image,
we propose a novel Spatial and Temporal Features Mixture Model (STFMM) to learn the significant representation consisting of the features of part sequences for video-based person re-identification.

3 Method

The architecture we proposed is shown in Fig. 2, and the details of each layer will be introduced in the following subsections. For our model, each video sequence is first split into \( N \) part sequences as the input of STFMM. Convolutional neural network (CNN) works as feature extractor, then the CNN features from each frame are fed into recurrent neural network (RNN), which is a powerful model to deal with the temporal sequences. After RNN, we adopt the method of Temporal Pooling (TP) \([19]\) to average the spatial features over time-steps. Finally, the part-level features are combined by an algorithm for mixing features and the sequence-level representation is formed.

In order to satisfy the input of our model STFMM, we develop the Siamese network architecture \([5]\). Traditional Siamese network consists of two identical neural networks which have the same parameters and weights. And the last layers of the two networks are fed into a contrastive loss function to calculate the similarity between the two inputs. In our model, we first develop the Siamese network architecture in which the number of identical neural networks is adjusted from 2 to \( 2N \) in order to calculate \( 2N \) input sequences. Each identical neural network has the same parameters and weights, and consists of two parts: CNN and RNN, as shown in Fig. 2. Secondly, the outputs of the last layers of \( 2N \) networks join the two features respectively, which represent the final representations of two original video sequences. Finally, those two features are fed into the loss function to calculate the similarity between the two original video sequences. In the following subsections we will explain each component of STFMM in detail.

3.1 Input data processing

The input data of the network consists of optical flow and color information. Optical flow is composed of the horizontal and vertical channels, and it is the pattern of apparent motion of image objects between two consecutive frames. Color information has three color channels that encode the information of a person’s appearance. In our method, we split each frame of video sequence into \( N \) parts, where each part frame consists of five channels, two being for optical flow and three being for color information.

Given a video sequence \( S = \{s^1, s^2, \ldots, s^T\} \), where \( T \) is the sequence length and \( s^t \) represents the frame of time \( t \), \( 1 \leq t \leq T \). Assuming that the parts number is \( N \), and the number of overlapping pixels is \( p \), the height of each part can be calculated as

\[
h = \left\lfloor \frac{1}{N} (H + (N - 1)p) \right\rfloor,
\]

where \( H \) is the height of original frame. According to (1), we can obtain the input data \( V \) as follows,

\[
V = \{S_1, S_2, \ldots, S_N\},
\]

\[
S_n = \{s^1_n, s^2_n, \ldots, s^T_n\}, 1 \leq n \leq N,
\]

where \( s^i_n \) is the frame of \( n \)-th part sequence at time \( t \).

3.2 Network

In our architecture, CNN consists of three convolutional layers \( l_1, l_2 \) and \( l_3 \). We define \( C_i \) as the operation function of \( i \)-th convolutional layers for each part sequence, then

\[
C_{l_1}(s^i_n) = O_1 = \text{Maxpool}(\text{tanh}(\text{Conv}(s^i_n))),
\]

\[
C_{l_2}(O_1) = O_2 = \text{Maxpool}(\text{tanh}(\text{Conv}(O_1))),
\]

\[
C_{l_3}(O_2) = O_3 = \text{Spp}(\text{tanh}(\text{Conv}(O_2))),
\]

where \( O_i \) is the output of \( l_i \) layer. For the last layer \( l_3 \), we employ the spatial pyramid pooling (SPP) \([6]\) to replace the max-pooling. In our model, we utilize SPP to generate a fixed-length representation with multi-level spatial bins \( 8 \times 8, 4 \times 4, 2 \times 2 \) and \( 1 \times 1 \). We define \( \text{SPP}_n \) as the \( n \)-th part comprehensive image-level features extracted by the spatial pooling layer, then

\[
\text{SPP}_n = \{\text{sp}_{p^1_n}, \text{sp}_{p^2_n}, \ldots, \text{sp}_{p^n_n}\},
\]

where \( \text{sp}_{p^n_n} \) is the output of \( \text{Spp} \) for the \( n \)-th part sequence at time \( t \).

After the spatial pyramid pooling layer, the \( n \)-th part features \( \text{SPP}_n \) will be passed on to the recurrent network. Recurrent neural network (RNN) is able to extract feature of temporal sequences, which can help to produce an output based on both current input and previous information at each time-step. The recurrent layer can be written as follows:

\[
o^i_n = U \text{sp}_{p^1_n} + W r_{n-1}^t,
\]

\[
r_n^t = \text{tanh}(o^i_n),
\]

where \( U \) and \( W \) are the parameters of RNN unit, \( o^i_n \) represents a linear combination of the current input \( \text{sp}_{p^1_n} \) and \( r_{n-1}^t \). In order to capture the long-term information, we employ the temporal pooling (TP) \([19]\) to process the output of RNN. After the TP layer, we get the part-level features of one video \( T = \{tp_1, tp_2, \ldots, tp_N\} \), where \( tp_n \) is the \( n \)-th part features. In order to comprehensively extract the representation of video sequence from these part-level
features, we propose an algorithm to mix features in $\mathcal{T}$ as follows:

$$v_p = t p_1^p \oplus t p_2^p \oplus \cdots \oplus t p_N^p, \quad v_g = t p_1^g \oplus t p_2^g \oplus \cdots \oplus t p_N^g,$$

where the operator $\oplus$ joints the part-level features together, $t p_n^p$ and $t p_n^g$ represent the $n$-th part features of the probe and gallery video sequences respectively, $v_p$ and $v_g$ represent the final representation of the probe and gallery video sequences.

### 3.3 Loss function

We train STFMM with three loss functions, including two losses of person’s identification and Siamese loss. Given the probe and the gallery video sequences, we utilize STFMM to capture the discriminative representations $v_p$ and $v_g$. $I(v_p)$ and $I(v_g)$ are defined by the person’s identification losses, which employ softmax regression and the standard cross-entropy loss. The Siamese loss of two sequences $E(v_p, v_g)$ is defined as follows:

$$E(v_p, v_g) = \begin{cases} \sum_{n=1}^{N} \| t p_n^p - t p_n^g \|^2, & p = g, \\ \max(0, m - \sum_{n=1}^{N} \| t p_n^p - t p_n^g \|^2), & p \neq g, \end{cases}$$

where $\| \cdot \|^2$ is the Euclidean distance of two vectors, $m$ is a margin to separate features of different people, and its value will be discussed in Section 4.2. We define the overall training objective as $L(v_p, v_g)$, which combines the person’s identification losses and Siamese loss,

$$L(v_p, v_g) = I(v_p) + I(v_g) + E(v_p, v_g).$$

In our experiments, the detailed parameters of each layer of CNN are shown in Table 1.

### 4 Experiments

#### 4.1 Data preparation and experiment settings

The iLIDS-VID dataset [24] contains a total of 300 pairs of video sequences that are captured by two non-overlapping cameras.
cameras at an airport arrival hall under CCTV networks. Each person is represented by two video sequences with the lengths ranging from 23 to 192 frames. The PRID-2011 dataset [7] consists of 749 persons in which each person is captured by two non-overlapping cameras, and the lengths of sequences range from 5 to 675 frames. We only use the first 200 persons in PRID-2011 dataset who appear in both cameras. We notice that the video sequences in iLIDS-VID are more challenging than those in PRID-2011, for example, we select 4 video sequences randomly as shown in Fig. 3.

To train and test the proposed network, we evenly split each dataset into two subsets randomly, in which one is for training and the other is for testing. We repeat the experiments 10 times with different train/test splits, and calculate the average of the results to get stable results. Since the deep neural network requires a large amount of data during the process of training, we perform data augmentation by random mirroring and cropping on the part sequences to increase the diversities of data.

We randomly choose sub-sequence of $k = 16$ consecutive frames from probe and gallery datasets for the training at each epoch. A full epoch consist of all positive pairs and the same number of negative pairs. In each epoch, the positive pairs and negative pairs are used alternately. The positive pairs consist of the two sub-sequences of the same person A from camera 1 and camera 2, and the negative pairs consist of two sub-sequences of person A and person B captured from two cameras respectively. During the testing process, we treat the video sequences of camera 1 as the probe sets and camera 2 as the gallery sets, where the data augmentation operation is also applied to all image sequences.

Before being passed on to the network, video sequences are converted to YUV color space and each color channel is normalized to have zero mean and unit variance. We use the Lucas-Kanade method [16] to calculate the horizontal and vertical optical flow channels. The optical flow channels are normalized to the range from -1 to 1. Both optical flow and color information are used as input data, which consists of five channels in training and testing processes.

We train the network by using the stochastic gradient descent with batch size of one and learning rate of $1e^{-3}$. Our model was trained for 700 epochs on a Nvidia GTX-1080.

### 4.2 Margin, parts number and overlapping

In (11), there is a margin value $m$ which plays an important role in Siamese loss function. Based on a large number of experiments, our model with the value of $m = 2$ can have better performance. In Section 3.1, we split each frame of video sequence in a specific way, in which there are two main parameters: the parts number $N$ and the overlapping pixels $p$. In this section, we exploit the effect of the parameters $N$ and $p$.

We set the number of parts $N$ as one of $\{1, 2, 3, 4, 5, 6\}$ and split each frame horizontally. The overlapping $p$ of two adjacent parts is also an important parameter, and we choose it according to the percentages of the height of the original frame at five levels 5%, 10%, 15%, 20%, and 25%. In our experiments, the height of the original frame is 128, therefore the value of $p$ is set as one of $\{7, 13, 20, 26, 32\}$.

The experimental results for the rank-1 CMC re-identification accuracy with different values of $N$ and $p$ are shown in Fig. 4. We can see that models with the input from part sequences have better performance compared with models that use the full frames, and that a number of overlapping pixels can improve the accuracy for person re-identification. The model with $N = 3$ and $p = 13$ outperforms the one with $N = 1$ (full frame sequence) by 11.3% of rank-1 accuracy on the iLIDS-VID dataset. The model with $N = 2$ and $p = 20$ outperforms the one with $N = 1$ by 4.0% of rank-1 accuracy on the PRID-2011 dataset. The results demonstrate that STFMM can significantly improve the accuracy for person re-identification. Comparing the results on two
datasets in Fig. 4, we can find that our model with \( N = 3 \) has a good performance on the rank-1 CMC re-identification accuracy on the iLIDS-VID dataset. Meanwhile it works better on the PRID-2011 dataset with \( N = 2 \) and \( N = 4 \).

In order to further explore the effect of \( N \) and \( p \), we list the top one CMC curves with the best value of \( p \) under different \( N \in \{1, 2, 3, 4, 5, 6\} \) as shown in Fig. 5. We can see that STFMM with \( N = 3 \) and \( p = 13 \) works best on the rank-5 CMC accuracy with the accuracy of 90.2\% on iLIDS-VID, and 94.1\% on PRID-2011. Therefore, if we aim to provide several candidate results for person re-identification, the model with \( N = 3 \) and \( p = 13 \) is a good choice. The number of parameters and the training time of STFMM with different \( N \) and \( p \) are shown in Table 2. Since the iLIDS-VID dataset contains more pairs of video sequences than the PRID-2011 dataset, we spend more time training the model on it. With the increasing of \( N \) and the change of \( p \), the growth of training time is nonlinear, and the growth change slows down. When the value of \( N \) is relatively large, the training time changes little with the different values of \( p \).

As shown in Fig. 6, we evaluate the performance of our model with different sequence lengths in testing. From the
Table 2  The training time of STFMM with different \(N\) and \(p\)

| \(N\) | \(p\) | iLIDS-VID Average time(s/epoch) | PRID-2011 Average time(s/epoch) |
|-------|-------|---------------------------------|---------------------------------|
| 1     | N/A   | 10.53                           | 6.47                            |
| 2     | 7     | 23.15                           | 14.37                           |
| 2     | 13    | 24.62                           | 15.83                           |
| 2     | 20    | 30.04                           | 19.74                           |
| 2     | 26    | 35.67                           | 25.57                           |
| 2     | 32    | 43.04                           | 27.58                           |
| 3     | 7     | 52.89                           | 23.73                           |
| 3     | 13    | 52.68                           | 24.78                           |
| 3     | 20    | 53.08                           | 27.99                           |
| 3     | 26    | 53.84                           | 29.27                           |
| 3     | 32    | 64.80                           | 31.24                           |
| 4     | 7     | 69.18                           | 41.33                           |
| 4     | 13    | 71.19                           | 42.37                           |
| 4     | 20    | 72.42                           | 43.08                           |
| 4     | 26    | 72.73                           | 44.18                           |
| 4     | 32    | 73.21                           | 44.28                           |
| 5     | 7     | 76.25                           | 51.23                           |
| 5     | 13    | 77.13                           | 52.47                           |
| 5     | 20    | 77.62                           | 54.08                           |
| 5     | 26    | 78.33                           | 55.18                           |
| 5     | 32    | 78.71                           | 56.08                           |
| 6     | 7     | 80.18                           | 60.13                           |
| 6     | 13    | 80.69                           | 60.57                           |
| 6     | 20    | 80.97                           | 61.16                           |
| 6     | 26    | 81.53                           | 61.88                           |
| 6     | 32    | 81.91                           | 62.13                           |

CMC curves we can conclude that increasing the lengths of input sequence can improve the re-identification accuracy. However, we find that the accuracy of CMC curves with \(T = 64\) is similar to the one with \(T = 128\). Moreover, the length of the video sequence has a great impact on the results especially on the more challenging dataset, iLIDS-VID.

In Fig. 7, we display some images of persons which have been successfully identified and unsuccessfully identified, and those images are listed in order of CMC curves. We can see that our model can powerfully identify the persons even if these people wear similar clothes. However, when the color of a person’s clothes changes in the two cameras with the different illuminations, it is still a challenging problem to identify the exact person for our model.

4.3 Comparison with the state-of-the-art methods

Except for the methods DVR [24], CNN+RNN [19], ASTPN [28], and TAM+SRM [35] described in Section 2, there are several other methods for video-based person re-identification. AFDA [10] is an algorithm which can hierarchically cluster video sequences and utilize the representative frames to learn a feature subspace maximizing the Fisher criterion. STA [15] utilizes the spatio-temporal body-action model to exploit the periodicity exhibited by a walking person and build a spatio-temporal appearance representation for pedestrian re-identification. In RFA [29], Yan et al. proposed a
Fig. 7 The results of our model with person successfully identified (a) and unsuccessfully identified (b). The correct results are marked with green boxes.

network based on LSTM to aggregate the frame-wise representation of human and yielded a sequence level representation.

The comparison of our STFMM method and previous methods is presented in Table 3. We can see that our method outperforms them significantly with the rank-1, rank-5, and rank-10 CMC accuracy of 73.6%, 89.5% and 95.2% respectively on the dataset iLIDS-VID. In particular, it exceeds the most recently proposed ASTPN by 11.3% on the rank-1 CMC accuracy. On the PRID-2011 dataset,
our method improves the rank-1 CMC accuracy from 77.3% to 81.3% compared with ASTPN. And the standard deviations on the rank-1 CMC accuracy of our model are 0.0051 and 0.0048 on the dataset iLIDS-VID and PRID-2011 separately. The results demonstrate that our method performs better than the methods that only use the input of full body sequences. In addition, we see that the STFMM model performs better on iLIDS-VID dataset by enhancing 11.3% rank-1 CMC accuracy but only 4.0% on PRID-2011 dataset. Since iLIDS-VID is more complicated and contains variations of occlusions, complicated backgrounds and view points, we conclude that our STFMM method improves the accuracy more effectively for complicated video-based person re-identification due to its ability to extract detailed features.

4.4 Cross-dataset testing

Considering that one model may be over-fitting to a particular scenario, cross-dataset testing would be a better way to evaluate its stability, in which the model is trained on dataset A but tested on dataset B. We perform the cross-dataset testing by using 50% persons of iLIDS-VID for training and 50% persons of PRID-2011 for testing. Referring to the conclusion of Section 4.2, our model is trained with $N = 3$, $p = 13$ and $m = 2$. As presented in Table 4, STFMM outperforms the previous methods significantly with the rank-1, rank-5, rank-10 and rank-20 CMC accuracy of 47.8%, 76.5%, 90.2% and 94.2% respectively, which exceeds the most recent method ASTPN by 17.4%, 18.3%, 19.3% and 8.8% on the corresponding levels. We also perform the cross-dataset testing by using 50% persons of PRID-2011 for training and 50% persons of iLIDS-VID for testing. As presented in Table 5, STFMM outperforms the previous methods significantly with the rank-1, rank-5, rank-10 and rank-20 CMC accuracy of 24.3%, 41.5%, 79.2% and 92.3% respectively. The results show that our model is more stable for practical applications.

Table 3 Comparison of STFMM method with previous state-of-the-art methods on iLIDS-VID and PRID-2011 in terms of Rank CMC(%)

| Datasets | iLIDS-VID | PRID-2011 |
|----------|-----------|-----------|
| CMC Rank | R = 1 | R = 5 | R = 10 | R = 20 | R = 1 | R = 5 | R = 10 | R = 20 |
| AFDA [10] | 37.8 | 62.4 | 72.9 | 81.7 | 43.0 | 72.5 | 85.7 | 91.8 |
| DVR [24] | 34.1 | 56.2 | 67.6 | 77.5 | 41.5 | 64.8 | 77.6 | 88.7 |
| STA [15] | 44.2 | 71.6 | 83.6 | 91.8 | 64.2 | 87.3 | 89.8 | 91.9 |
| RFA [29] | 48.7 | 71.5 | 85.2 | 93.1 | 70.1 | 90.4 | 94.9 | 97.3 |
| RNN + CNN [19] | 58.4 | 83.7 | 91.2 | 96.4 | 79.6 | 94.8 | 98.5 | 99.6 |
| TAM + SRM [35] | 55.7 | 86.3 | 91.8 | 97.3 | 77.3 | 95.3 | 98.9 | 99.7 |
| ASPTN [28] | 62.3 | 85.9 | 93.8 | 97.6 | 81.3 | 94.6 | 99.2 | 99.7 |
| STFMM | 73.6 | 89.5 | 95.2 | 99.2 | 73.6 | 89.5 | 95.2 | 99.2 |

The bold entries indicates the better results under the corresponding standard of CMC

Table 4 Cross-dataset testing accuracy tested on PRID-2011 in terms of Rank CMC (%)

| Model | Trained on | R = 1 | R = 5 | R = 10 | R = 20 |
|-------|------------|-------|-------|--------|--------|
| AFDA [10] | iLIDS-VID | 6.2 | 13.2 | 34.3 | 41.2 |
| DVR [24] | iLIDS-VID | 6.7 | 21.2 | 44.2 | 49.2 |
| STA [15] | iLIDS-VID | 7.2 | 23.3 | 45.1 | 51.2 |
| RFA [29] | iLIDS-VID | 19.7 | 45.2 | 61.2 | 75.2 |
| RNN + CNN [19] | iLIDS-VID | 27.6 | 57.3 | 68.9 | 80.5 |
| TAM + SRM [35] | iLIDS-VID | 26.7 | 58.3 | 69.8 | 81.3 |
| ASPTN [28] | iLIDS-VID | 30.4 | 58.2 | 70.9 | 85.4 |
| STFMM | iLIDS-VID | 47.8 | 76.5 | 90.2 | 94.2 |

The bold entries indicates the better results under the corresponding standard of CMC.

Table 5 Cross-dataset testing accuracy tested on iLIDS-VID in terms of Rank CMC (%)

| Model | Trained on | R = 1 | R = 5 | R = 10 | R = 20 |
|-------|------------|-------|-------|--------|--------|
| AFDA [10] | PRID-2011 | 4.2 | 11.2 | 30.3 | 36.2 |
| DVR [24] | PRID-2011 | 5.7 | 16.2 | 31.2 | 38.2 |
| STA [15] | PRID-2011 | 6.1 | 17.1 | 34.1 | 40.2 |
| RFA [29] | PRID-2011 | 10.7 | 23.2 | 51.3 | 71.2 |
| RNN + CNN [19] | PRID-2011 | 13.7 | 29.9 | 59.6 | 73.5 |
| TAM + SRM [35] | PRID-2011 | 14.2 | 32.3 | 64.3 | 76.3 |
| ASPTN [28] | PRID-2011 | 18.4 | 36.4 | 66.9 | 79.4 |
| STFMM | PRID-2011 | 24.3 | 41.5 | 79.2 | 92.3 |

The bold entries indicates the better results under the corresponding standard of CMC.
5 Conclusion

In this paper, we propose a novel spatial and temporal features mixture model (STFMM) in which the detailed features of human body are extracted for creating sequence-level representation. Different from the traditional methods which mainly focus on the local region information of an image, the extracting method of the local features is simplified in our model in order to deal with the video sequences.

In our model, we first split the human body to \( N \) parts in horizontal direction in order to obtain more specific information, and the adjacent parts are overlapped with \( p \) pixels. In order to satisfy the input of our model, we develop the Siamese network architecture in which the number of input is adjusted from 2 to 2\( N \). After choosing the appropriate values of the parameters \( m, N \) and \( p \), we evaluate our model on the iLIDS-VID and PRID-2011 datasets. The experimental results demonstrate that our approach outperforms the existing methods for video-based person re-identification. Specifically, it achieves a rank-1 CMC accuracy of 73.6% on the iLIDS-VID dataset. In order to evaluate the stability of our model, we do the cross-data testing which is trained on the iLIDS-VID dataset and tested on the PRID-2011 dataset. The results show that our model achieves the rank-1, rank-5, and rank-10 CMC accuracies of 47.8%, 76.5%, 90.2% respectively. For future work, we will consider applying our method to more complicated target tracking or detection systems.

Acknowledgements This research was supported by the National Natural Science Foundation of China (NSFC 61572005, 61672086, 61702030, 61771058), and Key Projects of Science and Technology Research of Hebei Province Higher Education (ZD2017304).

References

1. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. International conference on learning representations (ICLR)
2. Bar-Hillel A, Hertz T, Shental N, Weinshall D (2005) Learning a mahalanobis metric from equivalence constraints. J Mach Learn Res (JMLR) 6(6):937–965
3. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2360–2367
4. Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE international workshop on performance evaluation for tracking and surveillance (PETS), pp 1–7
5. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1735–1742
6. He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European conference on computer vision (ECCV), pp 346–361
7. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on image analysis (SCIA), pp 91–102
8. Hirzer M, Roth PM, Stinger M, Bischof H (2012) Relaxed pairwise learned metric for person re-identification. In: European conference on computer vision (ECCV), pp 780–793
9. Kviatkovsky I, Adam A, Rivlin E (2013) Color invariants for person reidentification. IEEE Trans Pattern Anal Mach Intell (PAMI) 35(7):1622–1634
10. Li Y, Wu Z, Karanam S, Radke RJ (2015) Multi-shot human re-identification using adaptive fisher discriminant analysis. In: British machine vision conference (BMVC), pp 73.1–73.12
11. Li Z, Chang S, Liang F, Huang TS, Cao L, Smith JR (2013) Learning locally-adaptive decision functions for person verification. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 3610–3617
12. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2197–2206
13. Liao S, Li SZ (2015) Efficient psd constrained asymmetric metric learning for person re-identification. In: IEEE international conference on computer vision (ICCV), pp 3685–3693
14. Liu C, Gong S, Chen CL, Lin X (2012) Person re-identification: what features are important? In: European conference on computer vision (ECCV), pp 391–401
15. Liu K, Ma B, Zhang W, Huang R (2015) A spatio-temporal appearance representation for video-based pedestrian re-identification. In: IEEE international conference on computer vision (ICCV), pp 3810–3818
16. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: International joint conference on artificial intelligence (IJCAI), pp 674–679
17. Ma B, Su Y, Jurie F (2012) Local descriptors encoded by fisher vectors for person re-identification. In: European conference on computer vision (ECCV), pp 413–422
18. Matsukawa T, Okabe T, Suzuki E, Sato Y (2016) Hierarchical gaussian descriptor for person re-identification. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1363–1372
19. McLaughlin N, Rincon JMD, Miller P (2016) Recurrent convolutional network for video-based person re-identification. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1325–1334
20. Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. In: Advances in neural information processing systems (NIPS), pp 2204–2212
21. Paistikriangkrai S, Shen C, Hengel AVD (2015) Learning to rank in person re-identification with metric ensembles. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1846–1855
22. Subramaniam A, Chatterjee M, Mittal A (2016) Deep neural networks with inexact matching for person re-identification. In: Advances in neural information processing systems (NIPS), pp 2667–2675
23. Varior RR, Shuai B, Lu J, Xu D, Wang G (2016) A siamese long short-term memory architecture for human re-identification. In: European conference on computer vision (ECCV), pp 135–153
24. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: European conference on computer vision (ECCV), pp 688–703
25. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res (JMLR) 10(2):207–244
26. Wu L, Shen C, Hengel A (2016) Deep recurrent convolutional networks for video-based person re-identification: an end-to-end approach. IEEE conference on computer vision and pattern recognition (CVPR)
27. Xiong F, Gou M, Camps O, Sznajer M (2014) Person re-identification using kernel-based metric learning methods. In: European conference on computer vision (ECCV), pp 1–16
28. Xu S, Cheng Y, Gu K, Yang Y, Chang S, Zhou P (2017) Jointly attentive spatial-temporal pooling networks for video-based person re-identification. IEEE International Conference on Computer Vision (ICCV), pp 4743–4752
29. Yan Y, Ni B, Song Z, Ma C, Yan Y, Yang X (2016) Person re-identification via recurrent feature aggregation. In: European conference on computer vision (ECCV), pp 701–716
30. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: International conference on pattern recognition (ICPR), pp 34–39
31. Zhang Z, Chen Y, Saligrama V (2015) Group membership prediction. In: IEEE International conference on computer vision (ICCV), pp 3916–3924
32. Zhao R, Ouyang W, Wang X (2014) Learning mid-level filters for person re-identification. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 144–151
33. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: IEEE international conference on computer vision (ICCV), pp 1116–1124
34. Zheng WS, Gong S, Xiang T (2013) Reidentification by relative distance comparison. IEEE Trans Pattern Anal Mach Intell (PAMI) 35(3):653–668
35. Zhou Z, Huang Y, Wang W, Wang L, Tan T (2017) See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 6776–6785

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.