A Hybrid 2D and 3D Convolution Based Recurrent Network for Video-Based Person Re-identification

Li Cheng[1], Xiao-Yuan Jing[1,2(✉)], Xiaoke Zhu[1,3], Fumin Qi[1], Fei Ma[1], Xiaodong Jia[1], Liang Yang[1], and Chunhe Wang[1]

[1] School of Computer, Wuhan University, Wuhan, China
cjackl@126.com, jingxy_2000@126.com, qfm120@163.com, mafei0603@163.com, jxdshimon@gmail.com, 1256951152@qq.com, 1173538302@qq.com

[2] College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China

[3] School of Computer and Information Engineering, Henan University, Kaifeng, China
henuzxk@163.com

Abstract. Video-based person re-identification (re-id), which aims to match people through videos captured by non-overlapping camera views, has attracted lots of research interest recently. In this paper, we propose a novel hybrid 2D and 3D convolution based recurrent neural network for video-based person re-id task, which can simultaneously make use of the local short-term fast-varying motion information and the global long-term spatial and temporal information. Specifically, the 3D convolutional module is able to explore the local short-term fast-varying motion information, while the recurrent layer performed can learn global long-term spatial and temporal information. We evaluate the proposed hybrid neural network on the publicly available PRID 2011, iLIDS-VID and MARS multi-shot pedestrian re-identification datasets, and the experiment results demonstrate the effectiveness of our approach on the task of video-based person re-id.

Keywords: 3D convolution
Short-term fast-varying motion information
Spatial and temporal information

1 Introduction

Person re-identification (re-id) is an important task in automated video surveillance and forensics, which aims to recognize an individual in a large set of candidates captured by different non-overlapping cameras. Since there usually exist visual ambiguity and spatio-temporal uncertainty in person's appearances across different cameras (which is usually caused by some external factors e.g., changes in lightness, viewpoint and resolution), person re-id is a challenging task in practice [14,18,28].

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11301, pp. 439–451, 2018.
https://doi.org/10.1007/978-3-030-04167-0_40

Person re-id techniques can be categorized into two main groups, single-shot methods and multiple-shot methods [4]. The single-shot methods try to associate pairs of images, each containing one instance of an individual. Most of existing methods can be classified as the single-shot methods [2,13]. For example, a semi-coupled low-rank discriminant dictionary learning (SLD^{2}L) method is developed in [13] for image-based super-resolution person re-identification, which aims to transform the feature of LR image into discriminative HR feature. To match individual images of the same person captured by different non-overlapping camera views against significant and unknown cross-view feature distortion, the CRAFT framework [2] performs cross-view adaptation by automatically measuring camera correlation from cross-view visual data distribution.

The multiple-shot methods extract features from multiple images of the same person to achieve a robust representation of the person. A significant amount of works has gone into the problem of multiple-shot person re-id over the years [1,21]. In [1], a set of frames of an individual were condensed into a highly informative signature, called the Histogram Plus Epitome (HPE), which incorporates complementary global and local statistical descriptions of the human appearance. Visual-spatial saliency, which represents the visual and spatial relationship among small regions segmented from multiple pedestrian images, is incorporated in region-based matching to improve the performance of person re-id [21]. Video-based person re-id methods are some special multiple-shot methods which require the multiple images of the same person to be a period of continuous video frames or a video clip [26]. Given a video clip of a person captured by one camera (probe person), video-based person re-id tries to find the corresponding person among a video gallery of people captured by other cameras in the surveillance systems. In this paper, we focus on the video-based person re-id problem.

1.1 Motivation

In general, there are two kinds of spatial and temporal information contained in a video clip of one walking person: (1) Global long-term spatial and temporal information; (2) Local short-term fast-varying motion information. The global long-term spatial and temporal information refers to the global long-term motion mode (e.g., speed and gait analysis) which is more abstract than local short-term fast-varying information. While the local short-term fast-varying motion information refers to the quick movements which occur in the partial limbs in a short time (e.g., optical flow and micro gestures). These movements always exist in multiple adjacent frames and can be obtained from detailed (raw) frames [11]. In practice, each pedestrian usually has some unique local short-term fast-varying motions, and thus making full use of these motion information is helpful to improving the discriminability and robustness of the features extracted from pedestrian videos. However, most existing video-based person re-id methods mainly focus on capturing the long-term spatial and temporal information, and ignore the local short-term fast-varying motion information, which will limit the person re-id performance of these methods.

Motivated by the above analysis, we intend to design an approach, which can simultaneously use the local short-term fast-varying motion information and global long-term spatial and temporal information contained in the person videos such that the person re-id performance can be further improved.

1.2 Contribution

Overall, the contributions of this study are mainly in three aspects:

(1) We design a hybrid 2D and 3D convolution based recurrent network (HCRN) for the video-based person re-id task. Specifically, HCRN simultaneously makes use of the local short-term fast-varying motion information and the global long-term spatial and temporal information.
(2) We introduce 3D convolutional operation to capture the local short-term fast-varying motion information contained in multiple adjacent frames of the pedestrian videos. To the best of our knowledge, this is the first work introducing 3D convolutional operation for the video-based person re-id task.
(3) We evaluate the performance of our approach on the public iLIDS-VID, PRID 2011 and MARS pedestrian sequence datasets. Extensive experimental results demonstrate the effectiveness of the proposed approach.

The rest of this paper is organized as follows. The next section briefly reviews the most recent and related developments with this work. Details of the proposed hybrid 2D and 3D convolutional and recurrent network are described in Sect. 3. Experimental results are provided in Sect. 4 to show the accuracy and applicability of the proposed approach. Finally, some concluding remarks are given in Sect. 5.

2 Related Works

In this section, we briefly review two types of works that are related to our approach: (1) Recurrent neural networks, (2) 3D convolutional networks.

2.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a powerful family of feedforward neural networks that can model global long-term temporal dependencies contained in inputs which consist of sequences of points that are not independent. There have been a number of works attempt to learn global long-term temporal dependencies contained in the input sequence to address different problems. A hierarchical recurrent neural network is proposed in [24] to capture long-term temporal information for tackling the video captioning problem. Recently, some works [16,29] apply recurrent neural network to extract spatio-temporal features from pedestrian videos for person re-id task.

Although recurrent neural networks have been widely used in many computer vision tasks, they mainly focus on learning global long-term temporal dependencies and ignore the local short-term fast-varying motion features (information). We are the first one which simultaneously use local short-term fast-varying motion information and global long-term spatial and temporal information for the video-based person re-id task.

2.2 3D Convolutional Networks

Deep learning technique has been successfully applied in many areas of computer vision, such as object detection [6], terrain perception [25] and face recognition [17]. Specially, 3D convolutional operation can extract spatial and temporal information from sequence data (e.g. video data) which is very useful for sequence data based recognition targets. Several 3D CNN models are developed in [5,12] to capture the motion information encoded in multiple adjacent frames for the action recognition problem. Authors in [11] designed a bidirectional recurrent convolutional network based on 3D convolution to capture local short-term fast-varying motion information contained in local adjacent frames for the video super-resolution task.

The major differences between our approach and these methods are twofold: (1) These methods apply 3D convolution to address the action recognition and video super-resolution tasks, while our approach employs the 3D convolution to solve the video-based person re-identification task. (2) In these methods, researchers mainly focus on local short-term fast-varying motion information encoded in multiple adjacent frames. Different from these methods, our approach not only utilizes the local short-term fast-varying motion information, but also can make use of the global long-term information existed in the whole video clips.

3 The Proposed HCRN Network

A diagram of the proposed HCRN network is shown in Fig. 1. The HCRN network consists of a 3D convolutional module, a 2D ResBlock module and a recurrent layer. Specially, we first perform three 3D convolutional layers (3D convolutional module) on raw frames to capture local short-term fast-varying motion information encoded in multiple adjacent frames. Then feature maps produced by 3D convolutional module will be processed by 2D ResBlock module. The 2D ResBlock module consists of three 2D ResBlock block units, which is used to explore high-level feature vectors for each frame. To further explore the global long-term temporal information contained in pedestrian video, we apply a recurrent layer (RNN) to the feature vectors which are produced by the 2D ResBlock module. A temporal pooling layer is adopted at the end of the RNN layer, such that feature vectors for all time-steps are aggregated to give a single feature vector which represents the whole sequence. Finally, we use the 3D convolutional module, 2D ResBlock module, RNN layer and temporal pooling layer as a feature extractor and adopt two loss functions including hinge embedding loss and cross-entropy loss to train the feature extractor in the Siamese architecture. In the following section we will give the details of each component in the proposed hybrid 2D and 3D convolution based recurrent network.

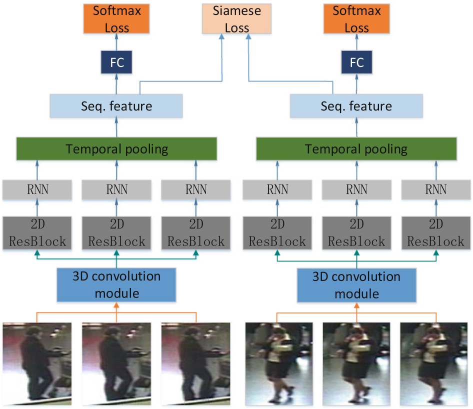


Fig. 1. Architecture of the proposed HCRN network

3.1 Input and Data Augmentation

Let $x^A_{(i)} = x_{(i,1)} \dots x_{(i,T)}$ be a video sequence, of length T, corresponding to the i^{th} person, where A means the full sequence. Similar to [16], we train the network using a Siamese architecture. For each epoch in the training phase, the input of the Siamese network is a pair of video sequences, $\left(x_{(i)}, x_{(j)}\right)$, where $x_{(i)}$ and $x_{(j)}$ are the randomly selected subset of 16 consecutive frames over the full sequence $x^A_{(i)}$ and $x^A_{(j)}$, respectively. Note that i and j may refer to the same or different person at each epoch. Specially, when $i = j$, the video sequence pair should be selected from the video clips that are captured from the same person by using two different cameras. When $i \neq j$, the video sequence pair can be selected from video clips captured by the same or different cameras of different persons.

To increase the diversity of the available datasets, we apply several data augmentation methods including randomly mirror all the frames contained in a video clip, and randomly change the brightness, contrast and saturation of each frame in the training phase. In the testing phase, we simply extract feature vector for each video clip (full sequence) from the raw video without any data augmentation.

3.2 3D Convolutional Module

3D convolutional operation has been demonstrated to be a powerful technique for capturing local short-term fast-varying motion information from video [12].

This motivate us to integrate three 3D convolutional layers at the head of the proposed HCRN network to capture local short-term fast-varying motion information (features) encoded in multiple adjacent frames. To perform a 3D convolution operation, we should first stack multiple contiguous frames together to form a cube, then a 3D kernel will be applied to convolve with the cube. In this way, the feature maps in the convolutional layer are connected to multiple contiguous frames in the previous layer, such that local short-term fast-varying motion information can be captured. Given a 3D convolutional operation, we can calculate the value of position (x, y, z) on the j^{th} feature map in the i^{th} layer as follows:

$$v_{ij}^{xyz} = b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)} \tag{1}$$

where R_i refers to the size of the 3D kernel along the temporal dimension, w_{ijm}^{pqr} denotes the $(p, q, r)^{th}$ value of the kernel connected to the mth feature map in the previous layer.

The reason why we only adopt three 3D convolutional layers at the head of our proposed network is that 3D convolution contains more parameters than 2D convolution which require large-scale dataset to train the network.

3.3 2D ResBlock Module

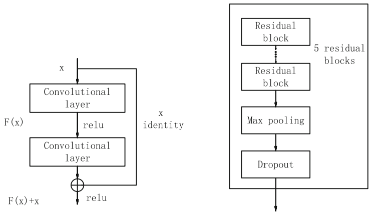
A typical residual block is showed in Fig. 2(a). The core idea of the residual block is the "shortcut connection" which can be formulated as $F(x)+x$. Several works [8,9,19] have demonstrated that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping, and can greatly improve the network's ability of feature extraction. Figure 2(b) is a diagram of the 2D ResBlock unit which we propose for the person re-id task. In each 2D ResBlock unit, we first stack five typical residual block as we have showed in Fig. 2(a). Then a max-pooling operation is applied to the feature maps which is produced by the residual blocks, such that the dimension of the feature maps can be reduced. Finally, we adopt a 1D dropout layer at the end of the 2D ResBlock unit to avoid over-fitting problem, which is the main difference between the common residual networks and the proposed 2D ResBlock unit. We stack three 2D ResBlock units in our 2D ResBlock module.

3.4 RNN

Let $c_{(i)} = c_{(i,1)} ... c_{(i,T)}$ be the output of the 2D ResBlock module corresponding to the input of $x_{(i)}$. The RNN [16] can learn the global long-term spatial and temporal information existed in $x_{(i)}$ on the following operations:

$$o_{(i,t)} = W_k c_{(i,t)} + W_l r_{(i,t-1)} \tag{2}$$

$$r_t = \tanh\left(o_{(i,t)}\right). \tag{3}$$



(a) (b)

Fig. 2. (a) A typical residual block, (b) Our 2D ResBlock unit

The output, $o_{(i,t)}$, at time step t is a linear combination of the vectors, $c_{(i,t)}$ denotes the output of 2D ResBlock module at time-step t, and, $r_{(i,t-1)}$ is used to remember the information on the RNN's state at the previous time-step.

Then a mean-pooling operation is performed on the $o_{(i)} = o_{(i,1)}...o_{(i,T)}$ over the temporal dimension, such that a single feature vector v_i representing the person's appearance averaged over the video clip can be produced. The mean-pooling operation can be formulated as follows:

$$v_i = \frac{1}{T}\sum_{t=1}^{T} o_{(i,t)}. \tag{4}$$

3.5 Joint Loss Function

The proposed HCRN network illustrated in Fig. 1 is a Siamese network architecture [7]. It consists of two feature extractors with identical weights which we showed above. Given a pair of video sequences $\left(x_{(i)}, x_{(j)}\right)$, we can get feature vectors $f_i = R\left(x_{(i)}\right)$ and $f_j = R\left(x_{(j)}\right)$, respectively, through the feature extractor. Then the Siamese network training objective function of the feature vectors (f_i, f_j) can be written as follows:

$$S\left(f_i, f_j\right) = \begin{cases} \frac{1}{2}\left\|f_i - f_j\right\|^2 & i = j \\ \frac{1}{2}\left[\max\left(m - \left\|f_i - f_j\right\|, 0\right)\right]^2 & i \neq j \end{cases}, \tag{5}$$

where d is the margin, which means that if a negative pair (f_i, x_j) is already separated by d, then there is no penalty for that pair and $S\left(f_i, f_j\right) = 0$. We set the margin d to 2 in our experiments.

Similar to the approach suggested in [16], we also apply the standard cross-entropy loss to optimize the feature extractor network. A cross-entropy loss can be formulated as follows:

$$I\left(fs_i\right) = \frac{exp\left(W_p fs_i\right)}{\sum_{q=1}^{Q} exp\left(W_q fs_i\right)} \tag{6}$$

$$fs_i = FC\left(f_i\right), \tag{7}$$

where Q is the number of identities contained in the training set, FC represents a fully connected layer which maps the output of the temporal pooling layer into the classification space,W_p and W_q refer to the p^{th} and q^{th} column of W, the softmax weight matrix, respectively. Finally, we can write the overall training objective G for the given pair of video sequences, $\left(x_{(i)}, x_{(j)}\right)$ as follows:

$$\begin{split} G = w_1 S\left(R\left(x_{(i)}\right), R\left(x_{(j)}\right)\right) + \\ w_2\left(I\left(FC\left(R\left(x_{(i)}\right)\right)\right) + I\left(FC\left(R\left(x_{(i)}\right)\right)\right)\right), \end{split} \tag{8}$$

where w_1 and w_2 are the weight for the corresponding loss function, which we set both of them to 1 in our experiments.

4 Experimental Results

4.1 Datasets

The PRID 2011 person sequence dataset [10] was captured by two disjoint cameras (Cam-A and Cam-B) in an outdoor street scenario with clean background and rare occlusions. 385 and 749 person sequences were recorded in Cam-A and Cam-B, respectively. Among all persons, only 200 persons were captured in both Cam-A and Cam-B. In our experiments, only these 200 persons who appear in both cameras were considered. The iLIDS-VID dataset [20] consists of 600 video sequences for 300 randomly sampled people with one pair of sequences for each person, which is created based on two non-overlapping camera views at a crowed airport arrival hall under a CCTV network. The MARS dataset [27] is a large-scale video re-id dataset containing 1,261 identities in over 20,000 video sequences. This dataset was collected by six near-synchronized cameras placed in the campus of Tsinghua university, and each identity was captured by at least two cameras.

4.2 Experimental Settings

We follow the evaluation protocol in [16] for both iLIDS-VID and PRID 2011 datasets. In particular, we randomly split all sequence pairs into two sets of equal size, with one for training and the other for testing. Then we further select sequences from the first camera in the testing set to form the probe set, and those from the other camera are used as the gallery set. While for MARS dataset, we follow the evaluation protocol in [22]. We first randomly chose two camera viewpoints of the same person, then set one of them as gallery set and the other as probe set. We employ the standard cumulated matching characteristics (CMC) curve as our evaluation metric for all three datasets, and report the rank-k average matching rates of 10 trials with different train/test splits.

4.3 Compared Methods

To evaluate the proposed HCRN network, we compared it against eight video-based person re-id methods including **DVR** [20], **STFV3D** and its enhancement method **STFV3D+KISSME** [15], **TDL** [23], **SI^{2}DL** [30], **RCN** [16], **TSS** [3] and **ASTPN** [22]. Experiment details will be presented in the following sections.

4.4 Comparison with State-of-the-Art Methods

We compare the proposed HCRN network against these eight video-based person re-id methods mentioned above on iLIDS-VID, PRID 2011 and MARS datasets in Table 1. One can observe that HCRN network always outperforms all the compared video-based person re-id methods on the three datasets. For example, when compare to the second best ASTPN model, the rank-1 matching rates are improved by 10.8% ((68.7 − 62.0)/62.0), 2.6% ((80.0 − 78)/78) and 6.8% ((47 − 44)/44) on iLIDS-VID, PRID 2011 and MARS datasets, respectively. Note that all the RCN, TSS, ASTPN and the proposed HCRN network use Siamese architecture, while the proposed HCRN network is the only one which don't use optical flow features, but with similar performance. The possible reason is that the 3D convolutional module can explore the motion information contained in multiple adjacent frames which play the same role as optical flow features. Among the eight compared methods, the RCN and ASTPN are the most similar methods to the proposed HCRN. The major differences between HCRN and these two methods are two-fold: (1) We apply a 3D convolutional module at the head of the network to explore motion information which is contained in multiple adjacent frames. However, these methods haven't used 3D convolution technique; (2) We adopt a deep residual network (2D ResBlock module) instead of shallow network used in these methods. Several works [8,9,19] have demonstrated that the deep residual architecture is a powerful architecture for

Table 1. Top r ranked matching rates (%) on iLIDS-VID, PRID 2011 and MARS datasets

Method/Rank	iLIDS-VID				PRID 2011				MARS			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Ours	**68.7**	**90.0**	**96.3**	**98.7**	**80.0**	94.0	97.0	**99.0**	**47.0**	**72.0**	**77.0**	**85.0**
RCN	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0	40.0	64.0	70.0	77.0
TSS	60.0	86.0	93.0	97.0	78.0	94.0	97.0	99.0				
ASTPN	62.0	86.0	94.0	98.0	77.0	95.0	**99.0**	99.0	44.0	70.0	74.0	81.0
TDL	56.3	87.6	95.6	98.3	56.7	80.0	87.6	93.6	37.4	62.5	67.7	76.3
SI^{2}DL	48.7	81.1	89.2	97.3	76.7	**95.6**	96.7	98.9	31.1	61.4	68.5	74.5
STFV3D	37.0	64.3	77.0	86.9	42.1	71.9	84.4	91.6	25.5	49.8	59.3	69.9
STFV3D + KISSME	49.7	78.3	84.7	91.7	66.2	87.3	88.4	89.4	31.7	58.6	62.2	73.1
DVR	23.3	42.4	55.3	68.4	28.9	55.3	65.5	82.8	15.9	31.5	41.6	50.4

extracting discriminative features. We have experimentally verified the effects of 3D convolutional module and 2D ResBlock module. The experimental results show that each module has played the expected role. Due to limited space, the related experimental results are not reported in this paper. Overall, the CMC performance improvements on three datasets demonstrate that HCRN network can extract more robust and discriminative features (information) than all the other compared methods.

4.5 Cross Dataset Testing

The generalization capability of person re-id methods always can be estimated by cross dataset testing. Based on the three datasets, we conducted two sets of cross dataset testing experiments where these two large and diverse datasets including iLIDS-VID and MARS were used for training, and testing were performed on 50% of the PRID 2011 dataset. It is evident from Table 2 that the CMC scores of the proposed method (HCRN) always slightly exceeds that of all compared methods. For instance, when the HCRN is trained on MARS, the proposed method achieves approx. 15.4% ((30.0 – 26.0)/26.0) performance advantage, at rank-1, over RCN. When trained on iLIDS-VID, the proposed method achieves approx. 6.7% ((32.0 – 30.0)/30.0) performance advantage, at rank-1, over ASTPN.

Table 2. Cross-dataset testing accuracy in terms of top r ranked matching rates (%): trained on MARS and iLIDS-VID, then tested on PRID 2011

Trained on	Method/Rank	r=1	r=5	r=10	r=20
MARS	HCRN	30.0	62.0	70.0	79.0
	RCN	26.0	57.0	68.0	78.0
iLIDS-VID	HCRN	32.0	61.0	73.0	86.0
	RCN	28.0	57.0	69.0	81.0
	ASTPN	30.0	58.0	71.0	85.0

5 Conclusion

In this paper, we develop a new hybrid 2D and 3D convolution based recurrent network for video based person re-id task. The use of 3D convolution layer allows us to explore the local short-term fast-varying motion information contained in multiple adjacent frames, while three 2D ResBlock units of each followed by a dropout layer further extract high-level information from each frame. Finally, the global long-term spatial and temporal information contained in the whole videos are learned by an RNN layer. Experiment results on three public video-based person re-id datasets show that the proposed hybrid network surpass any other methods in the video-based person re-id literature.

Acknowledgments. The authors would like to thank the editors and anonymous reviewers for their constructive comments and suggestions. This work was supported by NSFC-Key Project of General Technology Fundamental Research United Fund No. U1736211, the National Key Research and Development Program of China under Grant No.2017YFB0202001, the National Nature Science Foundation of China under Grant Nos. 61672208, U1504611,41571417, the Natural Science Foundation Key Project for Innovation Group of Hubei Province under Grant No.2018CFA024, the Science and Technique Development Program of Henan under Grant Nos. 172102210186, 182102311066, the Medical Education Research Project of Henan No. Wjlx2016095.

References

1. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recogn. Lett. **29**(1), 898–903 (2008)
2. Chen, Y., Zhu, X., Zheng, W., Lai, J.: Person re-identification by camera correlation aware feature augmentation. IEEE Trans. Pattern Anal. Mach. Intell. **40**(2), 392–408 (2018)
3. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: International Conference on Computer Vision, ICCV, pp. 1992–2000. IEEE Computer Society (2017)
4. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Computer Vision and Pattern Recognition, CVPR, pp. 2360–2367. IEEE Computer Society (2010)
5. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Computer Vision and Pattern Recognition, CVPR, pp. 1933–1941. IEEE Computer Society (2016)
6. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition, CVPR, pp. 580–587. IEEE Computer Society (2014)
7. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Computer Vision and Pattern Recognition, CVPR, pp. 1735–1742 (2006)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition, CVPR, pp. 770–778. IEEE Computer Society (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
10. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21227-7_9
11. Huang, Y., Wang, W., Wang, L.: Video super-resolution via bidirectional recurrent convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 1015–1028 (2018)
12. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)

13. Jing, X.Y., et al.: Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In: Computer Vision and Pattern Recognition, CVPR, pp. 695–704. IEEE Computer Society (2015)
14. Li, S., Shao, M., Fu, Y.: Person re-identification by cross-view multi-level dictionary learning. IEEE Trans. Pattern Anal. Mach. Intell. (2017)
15. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: International Conference on Computer Vision, ICCV, pp. 3810–3818. IEEE Computer Society (2015)
16. McLaughlin, N., del Rincón, J.M., Miller, P.C.: Recurrent convolutional network for video-based person re-identification. In: Computer Vision and Pattern Recognition, CVPR, pp. 1325–1334. IEEE Computer Society (2016)
17. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Computer Vision and Pattern Recognition, CVPR, pp. 815–823. IEEE Computer Society (2015)
18. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: IEEE International Conference on Computer Vision, ICCV. pp. 3739–3747. IEEE Computer Society (2015)
19. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4278–4284. AAAI Press (2017)
20. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 688–703. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_45
21. Xie, Y., Yu, H., Gong, X., Dong, Z., Gao, Y.: Learning visual-spatial saliency for multiple-shot person re-identification. IEEE Sig. Process. Lett. **22**(11), 1854–1858 (2015)
22. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: International Conference on Computer Vision, ICCV, pp. 4743–4752. IEEE Computer Society (2017)
23. You, J., Wu, A., Li, X., Zheng, W.: Top-push video-based person re-identification. In: Computer Vision and Pattern Recognition, CVPR, pp. 1345–1353. IEEE Computer Society (2016)
24. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: Computer Vision and Pattern Recognition, CVPR, pp. 4584–4593. IEEE Computer Society (2016)
25. Zhang, W., Chen, Q., Zhang, W., He, X.: Video paragraph captioning using hierarchical recurrent neural networks. Neurocomputing **275**, 781–787 (2018)
26. Zhang, W., Yu, X., He, X.: Learning bidirectional temporal cues for video-based person re-identification. IEEE Trans. Circuits Syst. Video Technol. **28**(10), 2768–2776 (2018)
27. Zheng, L., et al.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52
28. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: Computer Vision and Pattern Recognition, CVPR, pp. 1741–1750 (2015)

29. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In: Computer Vision and Pattern Recognition, CVPR, pp. 6776–6785. IEEE Computer Society (2017)
30. Zhu, X., Jing, X., Wu, F., Feng, H.: Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI, pp. 3552–3559. IJCAI/AAAI Press (2016)