Two Stream Deep CNN-RNN Attentive Pooling Architecture for Video-Based Person Re-identification

W. Ansar[1(✉)], M. M. Fraz[1,2,3,5], M. Shahzad[1,5], I. Gohar[1], S. Javed[2], and S. K. Jung[4]

[1] School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan
wansar.mscsl6seecs@seecs.edu.pk
[2] Department of Computer Science, University of Warwick, Coventry, UK
[3] The Alan Turing Institute, British Library, London NW1 2DB, UK
[4] Kyungpook National University, Daegu, Republic of Korea
skjung@knu.ac.kr
[5] Deep Learning Laboratory, National Center of Artificial Intelligence (NCAI), Islamabad, Pakistan

Abstract. Person re-identification (re-ID), is the task of associating the relationship among the images of a person captured from different cameras with non-overlapping field of view. Fundamental and yet an open issue in re-ID is extraction of powerful features in low resolution surveillance videos. In order to solve this, a novel Two Stream Convolutional Recurrent model with Attentive pooling mechanism is presented for person re-ID in videos. Each stream of the model is a Siamese network which is aimed at extracting and matching most differentiated feature maps. Attentive pooling is used to select most informative video frames. The output of two streams is fused to formulate one combined feature map, which helps to deal with major challenges of re-ID e.g. pose and illumination variation, clutter background and occlusion. The proposed technique is evaluated on three challenging datasets: MARS, PRID-2011 and iLIDS-VID. Experimental evaluation shows that the proposed technique performs better than existing state-of-the-art supervised video based person re-ID models. The implementation is available at https://github.com/re-identification/Person_RE-ID.git.

Keywords: Person re-identification · Spatial stream · Temporal stream

1 Introduction

Person re-ID is a task to recognize an individual in the images/videos captured from the cameras of disjoint non over-lapping field of view. It has its application areas in action recognition, people tracking, and surveillance videos at public places (like airport, museums, train stations, shopping mall, roads, universities etc.) [1]. An automated surveillance system is required, when images/video data in enormous amount is produced by the surveillance camera network which is difficult to be monitored manually. Person re-ID is a non-trivial task because of the challenges associated with it, which

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 654–661, 2019.
https://doi.org/10.1007/978-3-030-13469-3_76

includes change in body posture, variation in light, cluttered background, occlusion and low resolution images [2]. Person re-ID can be classified in two categories: image-based re-iD and video-based re-ID. Nowadays, multiple cameras are used for surveillance and produced massive amount of data. Therefore, image-based models cannot perform efficiently on large datasets [3, 4]. In contrast, video-based methods boost performance of person re-ID task due to multiples frames provide rich information, which involves temporal information associated to person motion. Still in video-based re-ID, some issues are not solved: Most of the methods [5] use a single stream convolutional neural network (CNN) to acquire temporal and spatial features through concatenation of RGB and optical flow, which limits the network capability to acquire the spatial and temporal data adequately. In this architecture, max-pooling is used for down sampling of features, which only focus on features with maximum value, so this model misses visual cues e.g. shoes and handbags. However, in order to deal with inter-class variations these small scale attributes are very beneficial. Thus single stream models seems not at optimal choice for features learning in low resolution videos. Secondly, different kind of cameras are used to capture videos sequences and produce low resolution video sequences. It's difficult to capture spatial features from low resolution images. So far, person re-ID methods are single stream CNN, which first extract spatial features and then fed these to recurrent unit to extract temporal information. These model cannot perform well in this situations and unable to learn gait feature independently e.g. arm and leg motion which remain same even in low resolution videos frames. Thirdly, consecutive video frames contain redundant information, thus it's so time taking and memory consuming process to see and remember each video frames.

To solve the issues mentioned above, we present novel two stream convolutional recurrent model with attentive pooling mechanism, using RGB frames and optical flow separately parallel as input to the network. The first stream named as spatial-net, which handle RGB video frames to capture useful spatial features representations like intensity gradient and color for each person in the video frames. The second stream named as temporal-net, handle the optical flows as input and learn gait features even in low resolution video frames. This solve second issue mentioned above. To solve third problem, the attentive pooling is used which selectively focus on discriminative video frames contain human being and ignore non informative frames. The fusion of spatial-net and temporal-net is done using a weighted objective function to give more weight to temporal features, which is based on both data and model fusion to utilize learned features maps. Finally, Siamese model is deployed on attention vectors, to judge the extent of matching.

Remaining paper is structured as follow: In Sect. 2 related work for video-based re-ID will be discussed. Our novel proposed architecture will be discussed in Sect. 3. The result of proposed model on PRID-2011 [6], MARS [7] and iLIDS-VID [8] will be discussed and compared with other models in Sect. 4. At last conclusion will be discussed in Sect. 5.

2 Related Work

In literature, researchers have explored various deep neural network (DNN) for video-based re-ID, to handle large data volume [4, 9–13]. The DNN models for re-ID can be further classified into two aspects: (i) the methods focused on spatial image-level representations e.g. person appearance such as color, style and shape of clothes [8], However, these techniques are generally build dense and complicated networks. For instance, Zheng et al. [7] attempts to train a classification network, in which a single feature vector represents each single image. A deeper and complex hierarchy is inevitable to establish in their models. (ii) The models which are paying more focus to temporal sequence-level representations e.g. person movement such as arm swinging and gait information [14]. Recently video-based re-ID models have explored the use of both temporal and spatial information [5, 9]. Yi et al. [9] presented the first Siamese-based CNN (S-CNN) architecture, which leveraged S-CNN model to the covered parts of the individual picture. To extract temporal information optical flows are computed, and then simply concatenated with person image as input of model to assemble the spatio-temporal features. McLaughlin et al. [5] proposed another model combination of CNN and recurrent neural network (RNN), which is also capable of learning features representation from multiple frames. In their model CNN extracts spatial information and fed features to RNN to effectively learn temporal information between different time frames. However, these methods used single stream to extract spatial and temporal information simultaneously, which limited capacity to fully obtained and combined temporal and spatial information. Chung et al. [15], used CNNs only to extract both spatial and short term temporal information, without taking into account the long term temporal information.

On the other hand, when we look at a picture, the one part of the picture is the focal point of our attention and perception. Even though the rest of the picture is still in eyes, we pay less attention to them. In context of computer vision, this mechanism is called attention model. A wide range of applications use attention models, such as action recognition task, image caption generation and questioning-answering models. To the best of our knowledge, just a few techniques [14, 16] utilized attention algorithms for person re-ID. However, these methods only use either spatial or temporal attention in single stream model. At present, there is no method for re-ID which utilized both temporal attention and spatial attention simultaneously to examine individual re-ID. Therefore, use of both spatial and temporal information are equally supportive to fully express a person in videos frames.

To address aforementioned issues, we proposed an end-to-end two stream CNN-RNN model with attentive pooling. In this model, both streams learn different feature maps, and finally merge output of two stream to obtain union of characteristics. This model can act as efficient features extractor for video-based re-ID, as well as it produces hidden unit representations for measuring similarity score for time series input.

3 The Proposed Model

This paper presents an end-to-end two stream convolutional-recurrent model with attentive pooling for video based person re-ID. To capture local effective contextual information spatial stream is used with RGB images as input and to obtain motion information temporal stream is used with optical flow as input, shown in Fig. 1. Each stream is Siamese based and output of CNN in each stream is images level representation, which then fed into recurrent neural network to extract temporal information of video sequence over long time. Following subsections will discuss two stream model and attentive pooling in more detail.

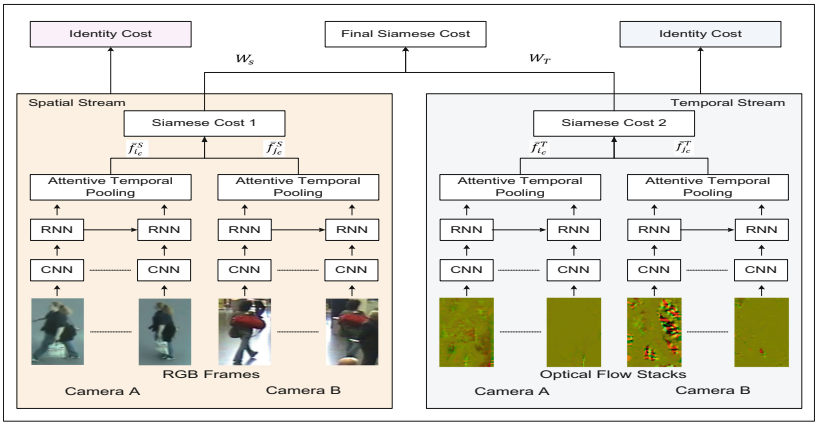

Fig. 1. The proposed two stream CNN-RNN architecture with attentive pooling

3.1 Two-Stream CNN-RNN

We express the input video sequence as V_s, s belongs to a, b used for camera A and camera B respectively. Input for first stream of our model are RGB frames: $V_s = \left(R^{(1)}...R^{(L)}\right)$ and for second stream optical flow images are used, $V_s = \left(T^{(1)} \cdots T^{(L)}\right)$. L is sequence length, since we have the gallery and probe sequences are of different size, so for training and testing we fixed the length of each person sequence to 16 and 128 respectively fairness of experimentations. Lucas-Kanade [17] technique is used to compute optical flows. Then by utilizing the convolutional network depict in Fig. 2; we obtain feature maps set $C_s = \left(C^{(1)}...C^{(L)}\right)$ and $C_T = \left(O^{(1)}...O^{(L)}\right)$ for RGB frames and Optical flow respectively. Unlike max pooling (only focus on maximum valued features) we used multiple convolutional layers in CNN for directly down sampling with stride of 2, it focuses on learned feature maps with fixed position and give us better performance. Output of CNN are then fed to RNN, recurrent layer is formulized by Eq. (1):

$$o^t = Ur^t + Ws^{t-1}, \quad s^t = tanh(o^t) \tag{1}$$

Where r^t is input of recurrent layer for time t, s^{t-1} is the hidden state which contain information for prior time step, and o^t is the output. Through matrix U recurrent layer implants high dimensional feature vector into low dimensional feature vector and $W \in R^{N \times N}$ which project s^{t-1} from R^1 to R^N. For first time step hidden state (s^0) is set to zero; tanh activation function is used to pass hidden state between different time steps. In the proposed model each stream helps other stream to learn multiple different aspect of feature maps, which is not possible in single stream models. Each batch of training images contains same number of positive and negative examples. We set the margin to 4 and learning rate to $2e^{-3}$ with stepwise decay for training the Siamese network. After training phase, we have stored the features vectors for all gallery sequences. During inference, the new sequence of probe persons is passed through the model to yield a feature vector. Afterwards, the probe feature vector is matched with the gallery features using a single matrix vector product.

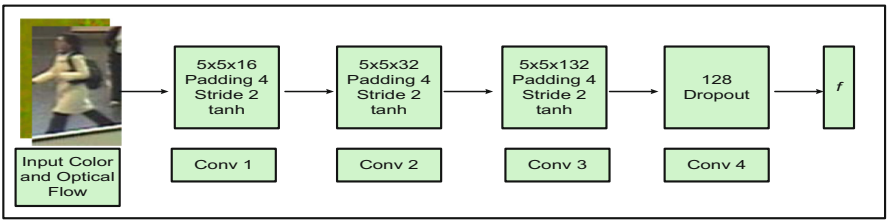

Fig. 2. The structure of CNN with hyper parameters.

3.2 Attentive Temporal Pooling Layer

There is many redundant information in consecutive video frames such as clothes and cluttered background. To tackle with this problem, we presented attentive temporal pooling for re-ID; which only focus on useful information. Attentive pooling layer is placed between RNN and distance measuring layer. By convolution and recurrent layer we obtained image and sequence level features, which are stored in P and G matrices for probe and gallery respectively; whose i^{th} row denotes output in the i^{th} time step of the recurrent. Then attention matrix A is computed by Eq. (2):

$$A = tanh(PUG^t) \tag{2}$$

In Eq. (2), both P *and* $Q \in R^{T \times N}$, $U \in R^{N \times T}$ and $A \in R^{T \times T}$. Matrix U is learned by network and intent for information sharing. Attention matrix calculates weight scores in temporal dimension and it is capable to have vision on both probe (t_p) and gallery (t_g) sequence features. Finally, in each stream, soft-max function is applied on temporal weight vectors. It transform the i^{th} weight $[t_p]_i$ and $[t_g]_i$ to the attention ratio $[a_p]_i$ and $[a_g]_i$ using the following Eq. (3):

$$[a_p]_i = \frac{e^{[t_p]_i}}{\sum_{j=1}^{T} e^{[t_p]_j}} \tag{3}$$

Where Eq. 4 applies to both two stream CNN-RNN similarly with changed sort of input. To acquire the sequence-level representation v_p and v_g, we applied dot product among the feature matrices P, G and attention vectors a_p as shown in Eq. (4):

$$v_p = P^T a_p, \quad v_g = G^T a_g \tag{4}$$

At last, we use a Siamese network for both two streams, which tries to decrease the distance between positive pairs and endeavor the separation between negative pairs during training, as shown in Fig. 1. From each stream we have two Siamese cost functions. Thusly, we characterize the joined cost function as Eq. (5):

$$\mathrm{D}\left(V_p, V_g\right) = \omega_s E\left(f_{i_{S_c}}^{-}, f_{j_{S_c}}^{-}\right) + \omega_T E\left(f_{i_{T_c}}^{-}, f_{j_{T_c}}^{-}\right) \tag{5}$$

ω_s, ω_T are the weights for Spatial-Net and Temporal-Net, E denote Euclidean distance and $f_{i_{S_c}}^{-}$ *and* $f_{j_{S_c}}^{-}$ are the attentive pooled feature vectors for person i and j separately. We utilized the unique weights for each stream to have the capacity to underline the spatial features when contrasted with the optical features for re-ID. Despite the fact that motions features include discriminative capacity to the accuracy even with low resolution image. Hence, we set the weights observationally with the condition $\omega_T \geq \omega_s$. Our ultimate training goal is grouping of Siamese and identity loss $\mathrm{L}\left(V_p, V_g\right) = \mathrm{D}\left(V_p, V_g\right) + \mathrm{I}\left(V_p\right) + \mathrm{I}\left(V_g\right)$.

4 Experimental Evaluation

The proposed model is evaluated on three publically available datasets: PRID-2011 [6], MARS [7] and iLIDS-VID [8].

4.1 Datasets

The PRID-2011 [6] dataset comprises of 749 persons, taken by two disjoint cameras. The dimension of frames differs from 5 to 675 for each person image sequence. It has simple backgrounds and less occlusions as compared with the iLIDS-VID dataset. We utilized only first 200 individuals captured by both cameras. MARS [7] is considered as the largest dataset for video-based person re-ID. It contains 1261 different persons, captured by 2 to 6 cameras and on average each person has 13 sequences. The iLIDS-VID [8] dataset consist of 300 persons, where every single individual is represented by two sequences taken at arrival hall of airport via two separate cameras. Size of sequences differs from 23 to 192 frames. It is very challenging dataset, because of dress resemblances for different persons, change in illumination, viewpoint deviations and arbitrary occlusions. For all three datasets, experiments were repeated 10 times with altered test and train splits to ensure constant results.

4.2 Quantitative Results and Comparison with Other Methods

In this section, we compared the results of our proposed architecture with state-of-the-art models for video-based re-ID. The proposed model achieves better performance than other methods on iLIDS-VID, PRID-2011 and MARS datasets in terms of Rank-1 (R1), Rank-5 (R5), Rank-10 (R10) and Rank-20 (R20) accuracies. The quantitative results are compared with other models in Table 1. The proposed model achieved matching rate of R1 = 72.6%, 84% and 56% on iLIDS-VID PRID-2011 and MARS dataset respectively, which is higher than other methods.

Table 1. Quantitative performance measures and comparison with other methods

Method	Dataset											
	iLIDS-VID				PRID-2011				MARS			
	R1	R5	R10	R20	R1	R5	R10	R20	R 1	R5	R 10	R20
Liu et al. [16]	44.3	71.7	83.7	91.7	64.1	87.3	89.9	92.0	–	–	–	–
Karanam et al. [18]	24.9	44.5	55.6	66.2	35.1	59.4	69.8	79.7	–	–	–	–
McLaughlin et al. [5]	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0	40.0	64.0	70.0	77.0
Xu et al. [14]	62.0	86.0	94.0	98.0	77.0	95.0	99.0	99.0	44.0	70.0	74.0	81.0
Yu et al. [4]	66.5	89.5	96.6	98.2	79.2	97.4	99.5	100	45.6	**72.4**	75.4	82.6
Boin et al. [3]	76.4	**95.3**	**98.0**	99.1	58.0	87.5	93.7	97.5	–	–	–	–
Ouyang et al. [19]	64.8	90.7	96.4	98.3	78.3	96.7	99.3	99.7	–	–	–	–
Proposed Model	**76.6**	90.8	96.6	**99.7**	**84.0**	**97.6**	**99.8**	**100**	**56**	67	**77**	**90**

5 Conclusion

In this paper, we have presented an end-to-end two streams CNN-RNN architecture with attentive pooling. The model uses two separate streams for the RGB frames and the optical flows to learn feature maps with different aspects. Only informative frames over full sequence are selected through attentive pooling followed by the concatenation of features. Quantitative results show that proposed model achieve greater performance to the existing state-of-the-art supervised re-ID models on PRID-2011, iLIDS-VID and MARS datasets. The proposed method performs better because the two streams emphasis on different feature maps and incorporate the temporal information for re-identification. The proposed architecture is simple to implement and generate best features even with low resolution input frames. In future we aim to extend the model using LSTM for solving the Open-world re-ID in real time.

Acknowledgements. This research was supported by development project of leading technology for future vehicle of the business of Daegu metropolitan city (No. 20180910). We are also thankful to NVIDIA Corporation for donating the TitanX GPU which is used in this research.

References

1. Karanam, S., et al.: A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets. IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 523–536 (2018)
2. Perwaiz, N., Fraz, M.M., Shahzad, M.: Person re-identification using hybrid representation reinforced by metric learning. IEEE Access **6**, 77334–77349 (2018)
3. Boin, J.-B., Araujo, A., Girod, B.: Recurrent neural networks for person re-identification revisited. arXiv preprint arXiv:1804.03281 (2018)
4. Yu, Z., et al.: Three-stream convolutional networks for video-based person re-identification. arXiv preprint arXiv:1712.01652 (2017)
5. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
6. 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
7. 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
8. 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
9. Yi, D., et al.: Deep metric learning for person re-identification. In: 22nd International Conference on Pattern Recognition (ICPR). IEEE (2014)
10. Mumtaz, S., et al.: Weighted hybrid features for person re-identification. In: 7th International Conference on Image Processing Theory Tools and Applications, Montreal (2017)
11. Mubariz, N., et al.: Optimization of person re-identification through visual descriptors. In: 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Funchal, Madeira, Portugal, pp. 348–355 (2018)
12. Khurram, I., Fraz, M.M., Shahzad, M.: Detailed sentence generation architecture for image semantics description. In: Bebis, G., et al. (eds.) ISVC 2018. LNCS, vol. 11241, pp. 423–432. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03801-4_37
13. Bashir, R.M.S., Shahzad, M., Fraz, M.M.: DUPL-VR: deep unsupervised progressive learning for vehicle re-identification. In: Bebis, G., et al. (eds.) ISVC 2018. LNCS, vol. 11241, pp. 286–295. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03801-4_26
14. Xu, S., et al.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. arXiv preprint arXiv:1708.02286 (2017)
15. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
16. Liu, K., et al.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
17. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision (1981)
18. Karanam, S., Li, Y., Radke, R.J.: Sparse re-id: block sparsity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2015)
19. Ouyang, D., Zhang, Y., Shao, J.: Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. Pattern Recogn. Lett. **117**, 153–160 (2018)