AN UNBIASED TEMPORAL REPRESENTATION FOR VIDEO-BASED PERSON RE-IDENTIFICATION

Xiu Zhang and Bir Bhanu

Center for Research in Intelligent Systems
University of California, Riverside, Riverside, CA 92521, USA

ABSTRACT

Person re-identification (re-id) aims to associate pedestrians across different camera views. As compared to the still image-based re-id, video-based re-id provides not only the spatial information but also the temporal dependency among frames. Most of the existing works apply the convolutional neural networks as a spatial feature extractor and then use backpropagation through time (BPTT) to train recurrent neural networks for temporal information. However, the long-term dependency is very hard to learn in RNNs via BPTT due to gradient vanishing or exploding. In the re-id task, the long-term dependency is quite common since the key information (identity of the pedestrian) exists most of the time along the given sequence. Thus, the importance of a frame should not be determined by its position in a sequence, which is usually biased in state-of-the-art models with RNNs. In this paper, we argue that long-term dependency can be very important and propose an unbiased siamese recurrent convolutional neural network architecture to model and associate pedestrians in a video. Experimental results on two public datasets demonstrate the effectiveness of the proposed method.

Index Terms— person re-identification (re-id), unbiased temporal representation, sparse attentive backtracking, recurrent neural networks (RNNs)

1. INTRODUCTION

Associating pedestrians across different camera views, known as person re-identification (re-id), has created significant interest in the image processing and computer vision communities. Re-id can be regarded as a promising and useful application to assist in many real-world scenarios such as human tracking, identifying individuals in crowded areas and criminal investigation [1, 2]. However, this task is still quite challenging due to variations in lighting conditions, human poses, occlusions and backgrounds.

In this work, we address the problem of video-based re-id. Unlike the previous image-based re-id settings, video-based re-id provides spatial appearance cues to create a more discriminative and robust feature representation. Besides, by using the sequences of the image frames, temporal information, such as gait, could also be utilized to distinguish people in complex situations.

Currently, most of the existing work solves video-based re-id with convolutional neural networks (CNNs) to extract the spatial features from each frame and recurrent neural networks (RNNs) to model the spatial-temporal correlation among frames [3, 4, 5]. They use the typical backpropagation through time (BPTT) to train RNNs. However, RNNs are well-known of having the vanishing and exploding gradient problems due to the exponential multiplication over time [6]. Although Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are proposed to alleviate this problem, it's still doubtful how much a fixed-length vector can memorize over a long sequence. All these limitations result in the difficulty in assigning enough or at least fair credit to the earlier timesteps in a long sequence, while looking at the entire sequence would be considerably better than relying on the last few frames only. This phenomenon is also common in re-id tasks. However, the importance of each image frame should not be biased by its position in the sequence. In this paper, we emphasize the importance of learning long-term dependencies in re-id, i.e., the global understanding from the entire sequence helps in associating the identity of a person. We propose a novel framework to effectively model the temporal correlations among frames by a sparse attentive backtracking mechanism [7, 8].

2. RELATED WORK

Re-id approaches mostly fall into two categories: feature representations [9, 10, 11] and metric learning [9, 12]. In recent years, deep learning methods have been successfully used in this area. Different CNNs [13, 14, 15] have been utilized for either robust features or learning a joint representation of features and similarity from the grouped image pairs or triplets.

For the video-based re-id, researchers put most of the efforts on using the temporal information such as gait [16, 17] and HOG3D descriptors [18]. Liu et al. [19] align video segments by using the gait information. More recently, McLaughlin et al. [3] incorporate CNNs and RNNs in a siamese architecture. The CNNs are used to capture the spatial representation and then the RNNs is applied to explore the temporal

information. Lastly, the temporal pooling layer is adopted to summarize the information. Varior et al. [4] and Zhang et al. [5] replace the regular RNNs with LSTMs and bi-directional RNNs, respectively. Xu et al. [20] use a similar architecture but add one spatial pooling layer to select regions from each frame, and add another attentive temporal pooling layer to select informative frames. All of these works first use BPTT to train RNNs and their variations, and then apply the pooling layer via either mean-pooling [3, 4, 5] or weighted pooling [20]. Thus, the bias is introduced along the time: even if the earlier frames convey better or more discriminative information than one in the latter frames, the model will still be likely memorize more about the latter frames. Inspired by Rosemary et al. [7], besides using BPTT, we apply sparse attentive backtracking mechanism to train RNNs to get the unbiased spatial-temporal representation for pedestrians.

3. APPROACH

In this section, we present our approach for video-based re-id, as shown in Fig. 1. It consists of two subsequent CNNs and RNNs in the siamese architecture. Given a video (a sequence of frames), CNNs are applied to extract the appearance representation of each frame independently. RNNs are further applied to learn the temporal dependency among the frames and generate a global description for the entire sequence. Instead of using the regular BPTT during the training of RNNs, the sparse attentive backtracking mechanism is leveraged to obtain better representation, which is unbiased in terms of the temporal information by capturing long-term dependency, for each input sequence.

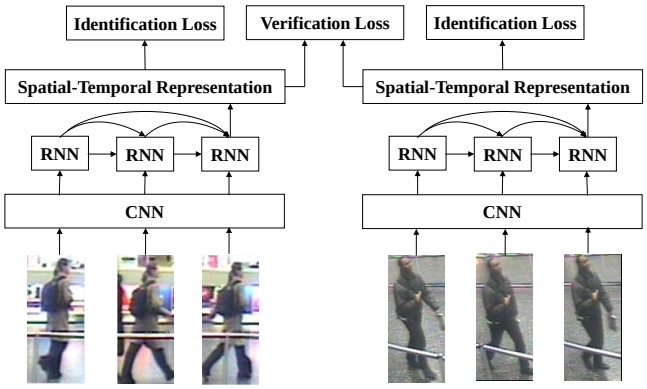

Fig. 1. The proposed architecture for video-based re-id

3.1. Spatial representation

As the resolution of given videos are relatively low (128×64) in most cases, we used the similar CNNs structure as in [3] to learn the spatial representation for each frame of a video. With higher resolution, we need to include more convolutional layers. The input consists of three color channels and two optical flow channels (horizontal and vertical optical flow) to encode appearance and motion, respectively. As shown in Fig. 2, the network consists of three convolutional + pooling layers and one fully connected layer. We use the hyperbolic-tangent ($tanh$) as the nonlinear activation function at each layer. To prevent over-fitting, we use the dropout for the CNNs and RNNs.

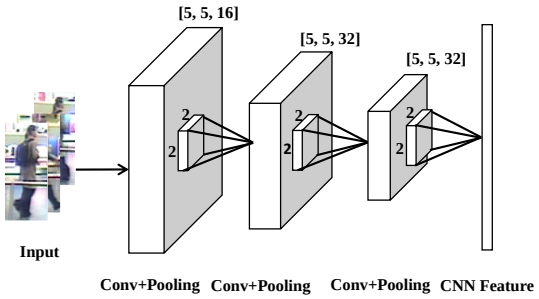

Fig. 2. CNN architecture for spatial representation. For each convolutional layer, the triplets shown in parenthesis represents filter size and the number of feature maps. The pooling window is 2×2.

3.2. Temporal Representation

RNNs are proposed to model temporal dynamics in sequences by mapping the input sequence to a fixed-length representation through hidden units. Given a video sequence $X = \{x_1, x_2, \cdots, x_T\}$, we use $v_i = \phi(x_i) \in \mathbb{R}^m$ to denote the description extracted by CNNs for the i-th frame, where q is the dimension of the output layer of CNNs. Then the recurrent connections by BPTT illustrated in Fig. 3(a) are defined as:

$$h_t = tanh(W_v v_t + W_h h_{t-1}) \quad (1)$$

where h_t is the hidden unit (representation) at time t. $W_v \in \mathbb{R}^{q \times n}$ and $W_h \in \mathbb{R}^{n \times n}$ are the parameters of RNNs, represented as the two fully connected layers, and n denotes the dimension of the hidden states in RNNs.

One can regard the hidden unit h_t in RNNs as a memory of the network which captures the important information about what happened across all the previous timesteps, i.e., $h_1, \cdots, h_{t-1}$, thus we can make the decision solely based on h_t. However, it is difficult for RNNs to learn long-term dependencies with BPTT in most of the real-world applications since multiplications of W_h over time give rise to the exponentially increasing or decreasing of related gradients, which is known as the exploding- or vanishing-gradient problem in RNNs. Many papers try to alleviate this by only back-propagate for few frames instead of the whole sequence, which

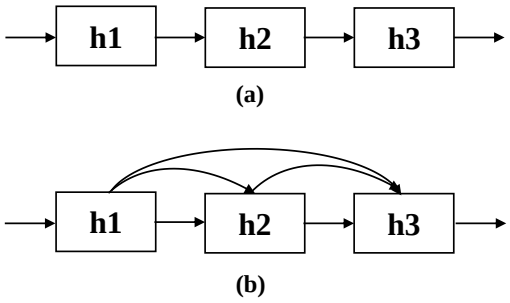

Fig. 3. Diagram illustrating the forward pass with BPTT and sparse attentive backtracking to compute the hidden unit h_3. In (a), the only way that h_3 gets the information from h_1 is through h_2 . In (b), h_3 could selectively choosing any previous hidden units (h_1 and h_2) for directly interaction.

further leads to the hidden representation from time T lean towards last few frames. To address the problem in RNNs mentioned above, sparse attentive backtracking mechanism is used in our approach to effectively exploit the information of the whole sequence and prevent biased representation.

We adapted the sparse attentive backtracking to train the RNNs [7]. To compute the hidden state at time t, we split the input into two sources: 1) the hidden unit from last timestep h_{t-1} as in standard RNNs and 2) all the hidden units prior to t as shown in Fig. 3(b). To deal with the variable number of input in 2), attention mechanism is applied here to generate a weighted combination of $\{h_1, \cdots, h_{t-1}\}$. A weight [21] is generated for each timestep prior to t as:

$$h_t = tanh(W_v v_t + W_h h_{t-1} + W_{\tilde{h}} \sum_{i=1}^{t-1} \alpha_i h_i)$$
$$\alpha_i = w_{\tilde{h}}[h_i; h_t] \tag{2}$$

where $W_{\tilde{h}} \in \mathbb{R}^{n \times n}$ and $w_{\tilde{h}} \in \mathbb{R}^{2n}$.

As described above, the sparse attentive backtracking mechanism explicitly model the correlation between the current hidden unit and all former hidden units to capture the long term dependency along the sequence. In video-based re-id task, the output of the last hidden unit is used as the final spatial-temporal representation of the given video.

3.3. Training Strategy

The objective of training is to minimize the loss of the joint identification and verification [3, 22]. Given a pair of videos of pedestrians i and j, we apply the siamese neural networks to get the spatial-temporal representations f_i for i and f_j for j, respectively. We take the Euclidean distance to measure the similarity between the video pairs. The closer the distance is, the more similar the videos are. Thus, the verification loss $V(f_i, f_j)$ is defined as:

$$V(f_i, f_j) = \begin{cases} ||f_i - f_j||^2 & \text{i = j} \\ max\{0, m - ||f_i - f_j||^2\} & \text{i} \neq \text{j} \end{cases} \tag{3}$$

where $||f_i - f_j||^2$ represents the Euclidean distance between two features, and m is the margin that separates the features of different pedestrians. When $i = j$, the two sequences indicate the same person and we push the two features to be close. On the contrary, When $i \neq j$, the video pairs belong to different persons and we pull the features away. Then, we use the cross-entropy loss to predict the identity of the person. The identification cost $I(f_i)$ is computed as follows:

$$I(f_i) = P(y = k|i) = \frac{exp(W_k f_i)}{\sum_{p=1}^{P} exp(W_p f_i)} \tag{4}$$

where y is the identity of the person, W is the softmax matrix with W_k and W_p referring to the k_{th} and p_{th} column of W respectively. P is the total number of the training identities. Finally, the overall training objective is to simultaneously optimize the joint identification and verification loss as:

$$L(f_i, f_j) = V(f_i, f_j) + I(f_i) + I(f_j) \tag{5}$$

Here we assign equal weights to the siamese loss and the identification loss. We train the whole architecture end to end with the sparse attentive backtracking method. For testing, we discard both siamese loss and identification loss functions, and use the network as a feature extractor and measure the similarity with Euclidean distance.

4. EXPERIMENTS

We evaluate the proposed approach on two of the most popular public datasets: iLIDS-VID [18] and PRID 2011 [23]. We compare our method with state-of-the-art methods especially those focusing on the spatial-temporal representations.

4.1. Datasets

The iLIDS-VID dataset consists of 300 persons. Each person is represented by two sequences taken by two non-overlapping cameras at an airport arrival hall. The length of these videos range from 23 to 192 with an average of 73. This dataset is very challenging due to the lighting and viewpoint variations, high clothing similarities among pedestrians, complex background and occlusions.

The PRID 2011 dataset contains 749 persons captured by two non-overlapping cameras, and only the first 200 persons are captured by both cameras. Their sequence lengths vary from 5 to 675, with an average of 100 frames. Compared with the iLIDS-VID dataset, this dataset has relatively simple background and rare occlusions, and thus becomes less challenging.

4.2. Experimental Setup

For each dataset, we randomly split it into training set and testing set with equal size. With different splits, the experiments are repeated 10 times. To evaluate the performance, for each test sequence, we compute its Euclidean distance against each video in the gallery and get the top n most-similar identities. The average of the Cumulative Matching Characteristics (CMC) plot is reported.

To train the Siamese network, we set the hinge margin of similarity to 2. In order to increase the diversity of the data, we randomly crop and flip each frame of videos in both the training and the testing settings to augment the data. For the fairness of comparison with [3, 5], we fix the length of the testing sequence to be 128. We set the initial learning rate to $1e-3$, momentum of 0.9, dropout rate of 0.6, and the total number of epochs to be 1000.

4.3. Experimental Results and Discussion

We compare the performance of our proposed models with the following state-of-the art methods: Recurrent Convolutional Neural Networks (RCNN) [3], Bidirectional Recurrent Convolutional Neural Networks (BRCNN) [5], Jointly Attentive Spatial-Temporal Pooling Network (ASPTN) [20], A two stream siamese CNN (TSC) [24], and Temporally Aligned Pooling Representation (TAPR) [25] . In Table 1 and Table 2, we show the recognition rates at rank 1, 5, 10 and 20 respectively for both datasets.

Methods	r=1	r=5	r=10	r=20
Ours (UTRCNN)	**62.7** $\pm$0.37	86	93.6	**98**
RCNN [3]	58	84	91	96
BRCNN [5]	55.3	85	91.7	95.1
ASPTN [20]	62	86	**94**	**98**
TSC [24]	60	86	93	97
TAPR [25]	55	**87.5**	93.8	97.2

Table 1. Comparison of the recognition rates at different ranks (%) on iLIDS-VID dataset. The first row shows performances for our Unbiased-Time RCNN (UTRCNN). The standard deviation value is along with the average rank one recognition rate. For each rank, the highest recognition rate is bold.

Methods	r=1	r=5	r=10	r=20
Ours (UTRCNN)	73 $\pm$0.41	92.7	95	98
RCNN [3]	70	90	95	97
BRCNN [5]	72.8	92	95.1	97.6
ASPTN [20]	77	**95**	**99**	**99**
TSC [24]	**78**	94	97	99
TAPR[25]	68.6	94.6	97.3	98.9

Table 2. Comparison of recognition rates at different ranks (%) on PRID 2011 dataset.

Frame 1 **Frame 5** **Frame 9** **Frame 12** **Frame 16**

Fig. 4. The training sequences of one successful instance in our UTRCNN model .

We achieve rank 1 recognition rates of 62.7% and 73.2%, with the improvement 8.1% and 4.2% compared to RCNN [3] with BPTT for iLIDS-VID dataset and PRID 2011 dataset, respectively. This means that using attentive backtracking mechanism to train RNN outperforms the standard BPTT in the video-based re-id task. We illustrate this by one example shown in Fig. 4. The RCNN fails because it is biased by the last few frames with BPTT. The red suitcase and the appearance of another man lead to an incorrect representation. On the other hand, our method explicitly use attentions to select the important frames to capture the better representation.

In particular, our rank 1 identification rate outperforms all the other compared methods in the iLIDS-VID dataset but get a fair performance on the PRID 2011 dataset. If we examine the two datasets, the iLIDS-VID dataset includes more occlusions and complex backgrounds. It is more suitable to learn the long-term dependency for a more complete and unbiased representation than to rely on the last few frames via BPTT in iLIDS-VID dataset. For PRID 2011 dataset, we still achieve improvement compared to RCNN and BRCNN methods which use the similar architecture.

5. CONCLUSION

In this paper, we proposed an unbiased siamese recurrent convolutional neural network architecture for video-based person re-identification task. Different from existing works, we apply sparse attentive backtracking mechanism, instead of typical BPTT, during the training of RNNs. This allows us to model long term dependencies and learn an unbiased temporal representation of any given video.

6. ACKNOWLEDGMENT

This work was supported in part by NSF grants 1552454 and 1330110. The contents of the information do not reflect the position or policy of US Government.

7. REFERENCES

[1] Shun Zhang, Jinjun Wang, Zelun Wang, Yihong Gong, and Yuehu Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognition*, vol. 48, no. 2, pp. 580–590, 2015.

[2] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 3, pp. 334–352, 2004.

[3] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016.

[4] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016.

[5] Wei Zhang, Xiaodong Yu, and Xuanyu He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[6] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[7] Nan Rosemary Ke, Anirudh Goyal, Olexa Bilaniuk, Jonathan Binas, Laurent Charlin, Chris Pal, and Yoshua Bengio, "Sparse attentive backtracking: Long-range credit assignment in recurrent networks," *arXiv preprint arXiv:1711.02326*, 2017.

[8] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[9] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[10] Shengcai Liao, Guoying Zhao, Vili Kellokumpu, Matti Pietikäinen, and Stan Z Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *CVPR*, 2010.

[11] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *ICCV*, 2013.

[12] Martin Hirzer, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[13] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[14] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[15] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

[16] Ju Man and Bir Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[17] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[18] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by video ranking," in *ECCV*, 2014.

[19] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *ICCV*, 2015.

[20] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," *arXiv preprint arXiv:1708.02286*, 2017.

[21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014.

[23] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*, 2011.

[24] Dahjung Chung, Khalid Tahboub, and Edward J Delp, "A two stream siamese convolutional neural network for person re-identification," in *CVPR*, 2017.

[25] Changxin Gao, Jin Wang, Leyuan Liu, Jin-Gang Yu, and Nong Sang, "Temporally aligned pooling representation for video-based person re-identification," in *ICIP*, 2016.