The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Multi-Rate Gated Recurrent Convolutional Networks for Video-Based Pedestrian Re-Identification

Zhihui Li,[1] **Lina Yao,**[2] **Feiping Nie,**[3*] **Dingwen Zhang,**[4] **Min Xu**[5]
[1]Beijing Etrol Technologies Co., Ltd.
[2]School of Computer Science and Engineering, University of New South Wales.
[3]Centre for OPTical Imagery Analysis and Learning, Northwestern Polytechnical University.
[4]School of Automation, Northwestern Polytechnical University.
[5]School of Electrical and Data Engineering, University of Technology Sydney.

Abstract

Matching pedestrians across multiple camera views has attracted lots of recent research attention due to its apparent importance in surveillance and security applications. While most existing works address this problem in a still-image setting, we consider the more informative and challenging video-based person re-identification problem, where a video of a pedestrian as seen in one camera needs to be matched to a gallery of videos captured by other non-overlapping cameras. We employ a convolutional network to extract the appearance and motion features from raw video sequences, and then feed them into a multi-rate recurrent network to exploit the temporal correlations, and more importantly, to take into account the fact that pedestrians, sometimes even the same pedestrian, move in different speeds across different camera views. The combined network is trained in an end-to-end fashion, and we further propose an initialization strategy via context reconstruction to largely improve the performance. We conduct extensive experiments on the iLIDS-VID and PRID-2011 datasets, and our experimental results confirm the effectiveness and the generalization ability of our model.

Introduction

Human re-identification (re-id), that is, matching pedestrians across multiple non-overlapping camera views (Farenzena et al. 2010), has attracted much research attention in the computer vision and machine learning communities due to its apparent critical role in surveillance and security applications such as people tracking and forensic search. Major challenges in person re-id include camera view changes, poor lighting conditions, and severe background clutter and occlusion; see some example illustrations in Figure 1. Various methods have been proposed in recent years to address this challenging problem, most of which concentrate on static-image-based person re-id (Jing et al. 2015; Liao et al. 2015) and can be divided into two groups: feature learning (Kviatkovsky, Adam, and Rivlin 2013; Yang et al. 2014a; 2014b; Zhang, Chen, and Saligrama 2014) and distance metric learning (Liao et al. 2015; Liao and Li 2015; Su et al. 2015; Varior et al. 2016b; Xiong et al. 2014).

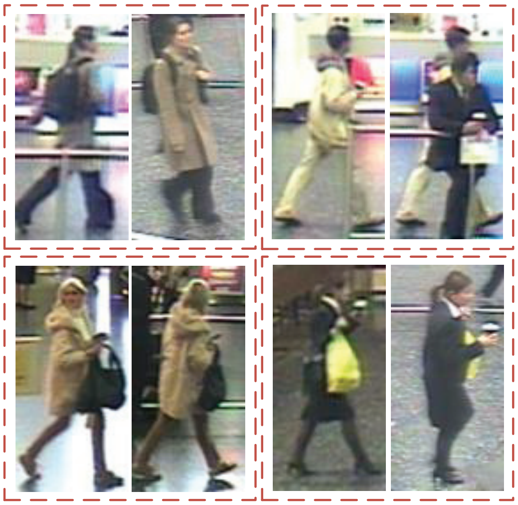


Figure 1: Person re-identification remains a challenging problem due to background clutter and occlusion, camera-view changes and lighting conditions.

Despite of the significant progress on still-image-based person re-identification, such existing methods still fall short of fully meeting the requirements of real-world applications, due to the following reasons: First, most naturally a pedestrian is captured in a video rather than in a single still image but still-image-based person re-id methods cannot exploit the rich temporal information related to a pedestrian's motion, such as his gait and perhaps even the way his clothing moves. Such information, if properly leveraged, can help disambiguate difficult matchings (McLaughlin, del Rincón, and Miller 2016). Second, there are significantly more and diverse appearance cues in a video sequence than in a static image, enabling us to extract more robust and discriminative appearance features. Third, video-based person re-id is a continuous process hence can largely reduce the negative effect due to occlusion and background clutter.

As a result, video-based person re-id has gained growing attention, and promising results on recent benchmark datasets (Hirzer et al. 2011; Wang et al. 2014) have been achieved. Existing methods in this setting can be divided into the following categories: (1) Key shot/fragment representation (Wang et al. 2014), which automatically selects the most discriminative fragments from the flow energy profile.

[*]Corresponding author. Email: feipingnie@gmail.com

However, in the matching process, since only one fragment is selected to represent the video sequence, temporal information is largely discarded. (2) Feature fusion/encoding, which encode frame-level feature vectors into a single vector via bag-of-words, but fail to consider the spatial-temporal information in video sequences. (3) Spatial-temporal appearance model (Liu et al. 2015), whose computation is unfortunately too expensive to be applicable in real-world applications. (4) Recurrent neural network based methods (Haque, Alahi, and Fei-Fei 2016; McLaughlin, del Rincón, and Miller 2016), which embed the inherent temporal hierarchy in the form of short, middle and long-term memory. Our work will largely follow the last category by combining a convolution network with a multi-rate recurrent unit.

Thus, the fundamental challenge in video-based pedestrian re-id is on how to *effectively* encode hence exploit the spatial-temporal information contained in a video sequence (Chang et al. 2017). Our work is based on a crucial novel observation: Different pedestrians, or sometimes even the same pedestrian, move in various speeds across non-overlapping camera views. This fact, to the best of our knowledge, has not been explicitly taken into account in existing works but intuitively can be very helpful for person re-id. To this end, we propose a novel siamese multi-rate gated recurrent network for video-based pedestrian re-id, which enables information sharing between different encoding rates and which collaboratively learns a multi-resolution representation that is robust to the motion rates of pedestrians. We train the entire network in an end-to-end fashion, and we initialize the network via context reconstruction—a strategy we found to work very well empirically.

Contributions. We summarize our contributions to video-based pedestrian re-identification as follows:

- Methodologically, we propose to leverage on a multi-rate recurrent network to extract a multi-resolution feature representation that is robust to the motion rates of pedestrians. Combined with a convolutional network, our model can very effectively exploit the spatial-temporal information in the video inputs.
- Procedurally, we propose to initialize our network via context reconstruction, in order to facilitate training and avoid poor local minima. Our empirical results confirm the effectiveness of this initialization strategy.
- Experimentally, we evaluate the performance of the proposed model on the iLIDS-VID and PRID 2011 datasets. Our results compare favorably against existing state-of-the-art alternatives, sometimes with a large margin.

Related Work

In this section, we briefly review two branches of works that are related to ours: (1) person re-identification, (2) recurrent neural networks.

Peson Re-Identification

Existing works on person re-id focus on discriminative feature learning (Kviatkovsky, Adam, and Rivlin 2013; Yang et al. 2014a; 2014b; Zhang, Chen, and Saligrama 2014) and distance metric learning (Liao et al. 2015; Liao and Li 2015; Su et al. 2015; Varior et al. 2016b; Xiong et al. 2014). Discriminative features that are invariant to camera view changes, lighting conditions, background clutter and occlusion play a vital role in boosting the performance of person re-id. However, traditional features alone are usually not sufficient to distinguish a person from similar ones, and one possibility is to combine a few features together to generate a more informative representation (Ma, Su, and Jurie 2012) or to select the most informative feature to represent a pedestrian (Farenzena et al. 2010).

When video sequences are used for person re-id, several new challenges arise. For instance, video sequences may have different length and/or frame-rates, and training an accurate appearance model with unknown partial or full occlusions within the video sequences is extremely difficult (Liu et al. 2017; You et al. 2016; Zhu et al. 2016). To address such issues, pioneering works explore space-time information to build spatial-temporal representations, resulting in more expressive features such as 3D HOG (Kläser, Marszalek, and Schmid 2008) and 3D SIFT. Other works have tried to build more discriminative representations. For example, Wang *et al.*(Wang et al. 2014) present a novel approach to automatically select the most discriminative video fragments from noisy image sequences of pedestrians, whie simultaneously learn a video ranking model for pedestrian re-id.

When highly discriminative features are available, numerous metric learning and ranking algorithms have been proposed to address the pedestrian re-id problem. For example, Zheng *et al.*(Zheng, Gong, and Xiang 2013) formulated person re-id as a relative distance comparison problem.

Recurrent Neural Networks

Recurrent Neural Network (RNN) is capable of capturing the context information in sequence date by maintaining some internal states. RNNs, particularly Long Short-Term Memory (LSTM) (Ng et al. 2015), have achieved remarkable success in natural language processing, machine translation (Karpathy and Li 2015; Sutskever, Vinyals, and Le 2014), and computer vision (McLaughlin, del Rincón, and Miller 2016; Varior, Haloi, and Wang 2016; Varior et al. 2016a; Yan et al. 2016). The fundamental idea behind RNN/LSTM is that through connections with previous states the network is able to “memorize” information from past inputs and thereby capture the contextual dependency in sequence data. Yan *et al.*(Yan et al. 2016) propose a novel recurrent feature aggregation framework for person re-id, which can learn discriminative sequence-level representation from simple frame-wise features. McLaughlin *et al.*(McLaughlin, del Rincón, and Miller 2016) introduce a novel temporal deep neural network architecture, and use optical flow, recurrent layers, and mean-pooling to embed the inherent temporal hierarchy in the form of short, middle and long-term temporal information, respectively. Varior *et al.*(Varior et al. 2016a) present a novel siamese LSTM architecture, which can selectively propagate relevant contextual information and thus enhance the discriminative capacity of the local features.

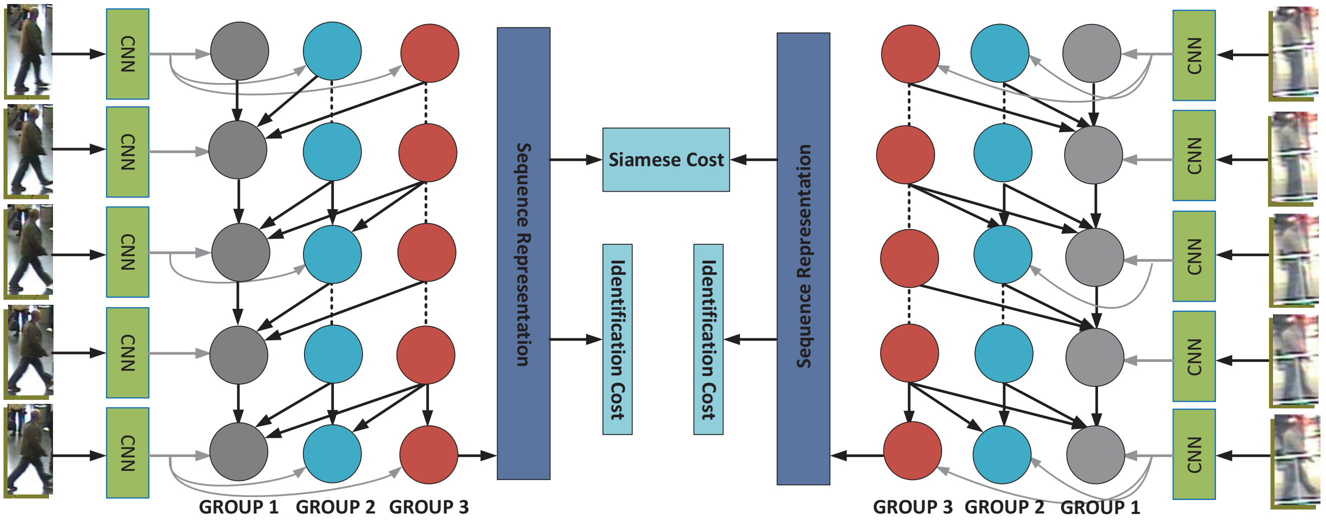


Figure 2: An overview of the proposed video-based re-id framework. We first process each sequence using a convolutional neural network and generate a feature vector to represent a pedestrian at a single time step. Then, we use a multi-rate recurrent neural network to encode sequences of representations in different resolutions. Lastly, we train the entire network end-to-end based on both the Siamese objective and the prediction cost of each pedestrian's identity.

However, a major limitation of existing works is that the input frames are encoded with a fixed sampling rate when training the RNNs. In other words, they fail to consider the fact that the motion speeds of different pedestrians can vary a lot. Sometimes even the same pedestrian may move in different speeds across different camera views. Our main goal in this work is to fill this gap.

The Proposed Approach

In this section we give a detailed description of the proposed model.

Overview

The goal of our model is to match sequences of same pedestrians obtained from different cameras. The proposed Siamese architecture consists of two copies of multi-rate recurrent networks that share the same weights. The fundamental idea behind a two-branch Siamese network is that it takes a pair of pedestrian images/sequences as input and aims to learn deep identity-discriminative representations so that images/sequences of the same pedestrian can be correctly matched whilst different pedestrians can be distinguished. Similar to (McLaughlin, del Rincón, and Miller 2016), we train the entire network end-to-end based on both the Siamese objective and the prediction cost of each pedestrian's identity.

An overview of our proposed architecture is shown in Figure 2. In our framework, each sequence is first processed by a convolutional neural network to generate a feature vector that represents a pedestrian at a single time step. Then, we use a multi-rate recurrent neural network to encode sequences of pedestrian representations in different resolutions. Since initialization is key to train a deep network and to avoid poor local minima, we further propose an effective initialization strategy via context reconstruction.

Input

To capture appearance and motion information for video-based re-id, our proposed model makes use of both color and optical flow information, which, as shown in (McLaughlin, del Rincón, and Miller 2016), will allow the network to better exploit short-term temporal information.

Before passing the images to the convolution network, we first convert them to the YUV color space and normalize each color channel to have zero mean and unit variance. The Lucas-Kanade algorithm has been widely used for optical flow computations, and we use it to compute the horizontal and vertical optical flow channels between each pair of frames. Thus, the first layer of the convolutional neural network uses five input channels: 3 for colors and 2 for optical flows.

Convolutional Network

As shown in Figure 2, we process each frame by a convolutional neural network (CNN) at each time step. The CNN takes an image I as an input, passes through multiple layers, and generates a vector x as an output. We refer this procedure as a function $\mathbf{x} = \texttt{CNN}(I)$. A traditional CNN includes several layers, such as convolution, pooling, and non-linear activation steps. Our model can take advantage of any CNN architecture. In this work we use ResNet-50 (He et al. 2016), which is different from most existing deep Re-ID networks (McLaughlin, del Rincón, and Miller 2016). Our motivation here is to choose an existing network that is competitive in the ImageNet classification benchmark and that has been widely used in many other vision problems. Among the recently proposed networks that achieved good classification performance on ImageNet, ResNet-50 achieved state-of-the-art performance. In the experiments, we conduct extensive experiments to evaluate the influence of CNN architectures on the final performance of our model.

The parameters of all CNNs are shared across all time-steps. Then, the output is fed into the multi-rate recurrent network, to encode the sequence of representations in different resolutions. To avoid overfitting, we also add a dropout layer between the CNN and the multi-rate recurrent layer.

Multirate Gated Recurrent Unit

We first revisit the basic Gated Recurrent Unit (GRU), which is a particular type of RNN and was proposed to allow each recurrent unit to adaptively capture dependencies of different time scales (Cho et al. 2014). It does not have any mechanism to control the degree to which its state is exposed, but rather expose the whole state each time.

More formally, at each time step t, given a frame representation $\mathbf{x}_t$ and previous state $\mathbf{h}_{t-1}$, the GRU cell generates a hidden state $\mathbf{h}_t$ and an output $\mathbf{o}_t$ iteratively as follows:

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}), \tag{1}$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}), \tag{2}$$

$$\bar{\mathbf{h}}_t = \tanh(\mathbf{W}_{\bar{h}} \mathbf{x}_t + \mathbf{U}_{\bar{h}}(\mathbf{r}_t \odot \mathbf{h}_{t-1})), \tag{3}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \bar{\mathbf{h}}_t \tag{4}$$

$$\mathbf{o}_t = \mathbf{W}_o \mathbf{h}_t, \tag{5}$$

where σ is the sigmoid activation function, $\mathbf{r_t}$ is the reset gate, $\mathbf{z}_t$ is the update gate, $\bar{\mathbf{h}}_t$ is the internal state, $\mathbf{W}_*$ and $\mathbf{U}_*$ are weight matrices and $\odot$ is the element-wise multiplication. When the reset gate is close to 0, it effectively forces the unit to act as if it is reading the first symbol of an input sequence, hence allows it to forget the previously computed state (Chung et al. 2014). The output $\mathbf{o}_t$ is calculated by a linear transformation from the state $\mathbf{h}_t$. For simplicity, neuron biases are omitted in the equations. We can write the entire iteration compactly as:

$$\mathbf{h}_t = \mathbf{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}),\ \mathbf{o}_t = \mathbf{W}_o \mathbf{h}_t. \tag{6}$$

After a maximum of S iterations, we get the final state $\mathbf{h}_S$ of the last step.

Multirate Gated Recurrent Unit. Next, we discuss the multirate extension of GRU as in (Koutník et al. 2014; Zhu, Xu, and Yang 2016). The clockwork RNN (Koutník et al. 2014) has delayed connections and units operating at different time-scales. The novelty of clockwork RNN is that its states and weights are divided into a few groups to capture temporal information at different rates. Following (Zhu, Xu, and Yang 2016), we divide state $\mathbf{h}_t$ into k groups, and each group g_i has a clock period T_i, where $i \in \{1, \ldots, k\}$. Empirically, we set $k = 3$ and $T_1, T_2, T_3 = 1, 3, 6$. Formally, at each step t, weight matrices of the group i with $(t \mod T_i) = 0$ are activated and are used to calculate the next state as follows:

$$\mathbf{r}_t^i = \sigma(\mathbf{W}_r \mathbf{x}_t + \sum_{j=b}^{e} \mathbf{U}_r^{i,j} \mathbf{h}_{t-1}^j), \tag{7}$$

$$\mathbf{z}_t^i = \sigma(\mathbf{W}_z^i \mathbf{x}_t + \sum_{j=b}^{e} \mathbf{U}_z^{i,j} \mathbf{h}_{t-1}^j), \tag{8}$$

$$\bar{\mathbf{h}}_t^i = \tanh(\mathbf{W}_{\bar{h}}^i \mathbf{x}_t + \sum_{j=b}^{e} \mathbf{U}_{\bar{h}}^{i,j} (\mathbf{r}_t \odot \mathbf{h}_{t-1}^j)), \tag{9}$$

$$\mathbf{h}_t^i = (1 - \mathbf{z}_t^i) \odot \mathbf{h}_{t-1}^i + \mathbf{z}_t^i \odot \bar{\mathbf{h}}_t^i, \tag{10}$$

where the state weight matrices $\mathbf{U}_*$ are divided into k row-blocks and each row-block is partitioned into k column-blocks. The input weight matrices $\mathbf{W}_*$ are divided into k row-blocks and $\mathbf{W}_*^i$ denotes the weights in row-block i. There are two modes for state transition, and depending on which mode we operate, we have

$$\begin{cases} b = 1, e = i, & \text{Fast} \rightarrow \text{slow mode} \\ b = i, e = k, & \text{Slow} \rightarrow \text{fast mode} \end{cases}. \tag{11}$$

In the fast to slow mode, states of faster groups (*i.e.*larger T_i) includes previous slower states (*i.e.*smaller T_i). Thus, the faster states incorporate information not only at the current rate but also information that is slower and more refined. The intuition for the fast to slow mode is that when it is activated, we can take advantage of the information already encoded in the slower states. Empirically, in this paper, we use the fast to slow mode for its better performance.

When $(t \mod T_i \neq 0)$, the previous state is directly passed over to the next state, *i.e.*,

$$\mathbf{h}_t^i = \mathbf{h}_{t-1}^i. \tag{12}$$

We illustrate the state transition process in Figure 2. We note that training is much faster than traditional GRU with the same number of hidden nodes since not all previous modules are evaluated at every time step.

Network Initialization

Context reconstruction has been demonstrated to play a vital role in different language modeling applications (Kiros et al. 2015), which inspires us to adapt it for initializing our network in video sequence modeling. We use two decoders to predict the context sequences of the inputs, *i.e.*, reconstructing the frame-level representations of the previous sequence and the next sequence.

We denote $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ as the previous sequence of the current input sequence $\mathbf{X}$, and $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ as the next sequence. The decoder is a GRU conditioned on the encoder outputs $\mathbf{o}_1, \ldots, \mathbf{o}_S$ and the final state $\mathbf{h}_S$ of the last step of the encoder. Since soft attention mechanism has been shown to be quite effective in several sequence modeling tasks, we utilize the attention mechanism at each step to help the decoder decide which frames in the input sequence might be related to the next frame reconstruction. The core of the soft attention mechanism is that instead of just inputting the original sequence $\mathbf{y}$ to the GRU

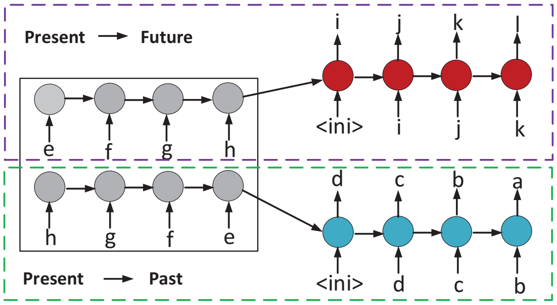


Figure 3: The architecture of network initialization. For initialization, two decoders are used to predict surrounding contexts by reconstructing previous frames and next frame sequences. The "<ini>" is the initial input at step 0. In this paper, we use a zero vector. Each decoder is chosen with a probability of 0.5 for reconstruction.

layer, dynamic weights are used to generate a new sequence. At step t,

$$\mathbf{y}_t^{\text{attn}} = W_a \mathbf{y}_t + W_b \mathbf{a}_{t-1}, \tag{13}$$

where the attention weights W_a and W_b measure the relevance between the i-th element $\mathbf{y}_i$ of the input sequence and the history information recorded by the GRU $\mathbf{h}_{t-1}$.

$$\mathbf{h}_t^{\text{dec}} = \text{GRU}(\mathbf{y}_t^{\text{attn}}, \mathbf{h}_{t-1}^{\text{dec}}), \;\; \mathbf{o}_t^{\text{attn}} = W_o^{\text{dec}} \mathbf{h}_t^{\text{dec}} \tag{14}$$

In details, the relevance score is calculated as follows:

$$g_t^i = \mathbf{v}^\top \tanh(\mathbf{W}_{hg}\mathbf{h}_t^{\text{dec}} + \mathbf{W}_{og}\mathbf{o}_i), \tag{15}$$

where $\mathbf{v}$, $\mathbf{W}_{hg}$, $\mathbf{W}_{og}$ are all weight parameters. Then, the attention vector $\mathbf{a}_t$ can be obtained by:

$$\mathbf{a}_t = \sum_{i=1}^{S} \frac{\exp(g_t^i)}{\sum_{j=1}^{S}\exp(g_t^j)} \mathbf{o}_i, \tag{16}$$

i.e., the weighted average of the encoder outputs $\mathbf{o}_i$, with weights proportional to the relevance score g_t^i.

Finally, the decoder ϕ generates the prediciton $\mathbf{o}_t^{\text{dec}}$ by calculating

$$\mathbf{o}_t^{\text{dec}} = \text{Linear}(\mathbf{o}_t^{\text{att}}, \mathbf{a}_t), \tag{17}$$

where $\text{Linear}(\mathbf{m}, \mathbf{n}) = \mathbf{W}_m\mathbf{m} + \mathbf{W}_n\mathbf{n}$. Different from the classic seq2seq model, we use two decoders here: one for the past sequence reconstruction and the other for the future sequence reconstruction. These two decoders do not share weights. The decoders are trained to minimize the reconstruction loss of two sequences, which is defined as follows:

$$\sum_t \ell(\phi(\mathbf{y}_{<t}, \mathbf{o}_1, \ldots, \mathbf{o}_S, \mathbf{h}_S), \mathbf{y}_t) \;\; + \tag{18}$$
$$\sum_{t'} \ell(\phi(\mathbf{z}_{<t'}, \mathbf{o}_1, \ldots, \mathbf{o}_S, \mathbf{h}_S), \mathbf{z}_{t'}),$$

where we use the Huber loss that has been widely demonstrated to be effective:

$$\ell(y, \bar{y}) = \begin{cases} \frac{1}{2}(y-\bar{y})^2, & \text{if } |y-\bar{y}| \le \delta \\ \delta|y-\bar{y}| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases}. \tag{19}$$

To minimize the information lag, we reverse the input order as well as the target order for the past reconstruction. We train the two decoders with the encoder via backpropagation and regularize the network by randomly dropping one decoder for each batch (as shown in Figure 3).

Final Representation Learning

After training, we discard the Siamese and identification cost functions and retrain the CNN and multi-rate recurrent neural networks for feature extraction. The multi-rate recurrent neural networks generate multirate states at each step. There are many ways to pool the states to obtain a global sequence representation. Xu *et al.*(Xu, Yang, and Hauptmann 2015) demonstrated that Vector of Locally Aggregated Descriptors (VLAD) encoding outperforms the other alternatives by a large margin. Hence, we also apply VLAD to encode the RNN representations.

VLAD encoding can be regarded as a simplified version of Fisher vector encoding. With inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and K coarse centers $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ generated by K-means, we can obtain the difference vector regarding center c_k by:

$$\mathbf{u}_k = \sum_{i:\text{NN}(\mathbf{x}_i)=\mathbf{c}_k} (\mathbf{x}_i - \mathbf{c}_k), \tag{20}$$

where $\text{NN}(\mathbf{x}_i)$ indicates $\mathbf{x}_i$'s nearest neighbors among K coarse centers. By concatenating $\mathbf{u}_k$ over all K centers, we obtain the feature vector of size DK where D is the dimension of $\mathbf{x}_i$. Different normalization methods have been used to improve performance. Signed square rooting (SSR) is usually used to convert each element $\mathbf{x}_i$ into $\text{sign}(\mathbf{x}_i)\sqrt{|\mathbf{x}_i|}$. The intra-normalization method normalizes representations for each center, followed by the ℓ_2 normalization for the whole feature vector. The final normalized feature representations are used for pedestrian re-identification.

Experiments

In this section, we conduct extensive experiments to evaluate the proposed approach in terms of both effectiveness and generalization ability. We also conduct experiments to study the affects of different components in the initialization procedure.

Datasets and Experimental Setup

We carry out our experimental comparisons on the following two real-world datasets:

iLIDS-VID dataset (Wang et al. 2014). The iLIDS-VID dataset consists of 600 image sequences of 300 distinct individuals. It is created based on two disjoint camera views in public open space. Each image sequence has variable length, ranging from 23 frames to 192 frames, with an average length of 73. This dataset is challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background, and random occlusions.

PRID 2011 dataset (Hirzer et al. 2011). The PRID 2011 dataset contains 400 image sequences of 200 randomly sampled people from two cameras. Each image sequence has

Table 1: Performance comparison against state-of-the-art alternatives on ILIDS-VID and PRID-2011 datasets. Cumulative Matching Characteristics (CMC) curve is used as an evaluation metric. Larger value indicates better performance.

	iLIDS-VID				**PRID-2011**			
	Rank1	**Rank5**	**Rank10**	**Rank20**	**Rank1**	**Rank5**	**Rank10**	**Rank20**
PaMM (Cho and Yoon 2016)	30.3	56.3	70.3	82.7	45.0	72.0	85.0	92.5
SI^{2}DL (Zhu et al. 2016)	48.7	81.1	89.2	97.3	76.7	**95.6**	96.7	98.9
CNN+XQDA (Zheng et al. 2016)	53.0	81.4	–	95.1	77.3	93.5	–	99.3
STFV3D+KISSME (Liu et al. 2015)	44.3	71.7	83.7	91.7	64.1	87.3	89.9	92.0
DVR (Wang et al. 2016)	39.5	61.1	71.7	81.0	40.0	71.7	84.5	92.2
RCN (McLaughlin, del Rincón, and Miller 2016)	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
TDL (You et al. 2016)	56.3	87.6	95.6	98.3	56.7	80.0	87.6	93.6
Ours (CaffeNet)	58.8	86.9	95.5	98.2	77.2	93.8	96.8	98.3
Ours (VGG16)	58.9	87.8	95.8	98.7	77.8	94.2	97.2	98.8
Ours (ResNet-50)	**60.8**	**89.2**	**97.2**	**99.5**	**78.4**	94.8	**97.9**	**99.4**

variable length consisting of 5 to 675 image frames, with a variable number of 100. The dataset was captured in uncrowded outdoor scenes with rare occlusions and simple background. However, these two camera views have significant viewpoint, illuminations, and color inconsistency.

Following (Wang et al. 2014; McLaughlin, del Rincón, and Miller 2016), we randomly split each dataset into 50% of persons for training and 50% of persons for testing for all experiments. During testing, we use the first camera as the probe set and the second camera as the gallery set. For all the datasets, the performance is evaluated by the average Cumulative Matching Characteristics (CMC) curves after 10 random training-test splits. The CMC curve, at rank score k, gives the percentage of the test queries whose target is within the top k closest match.

Experimental Setup. We implement the proposed model using the framework released by (McLaughlin, del Rincón, and Miller 2016) based on Torch. We will release our code and trained models upon acceptance. We train the network using an Nvidia TitanX Pascal with 12GB memory. The hyperparameters of the convolutional network were pre-trained on the ImageNet dataset. The network was trained using stochastic gradient decent with a learning rate of 1e-3, and a batch size of 1, and the input to the Siamese network is alternated between positive and negative sequence pairs, as in (McLaughlin, del Rincón, and Miller 2016). We train the network for 500 epochs. When we get the vector representation for each pedestrian, we evaluate the proposed model using the same metric as in (McLaughlin, del Rincón, and Miller 2016).

Test Initialization Strategy

First, we verify the advantage of our initialization strategy in § by comparing it against an encoder using random initialization. To ensure a fair comparison, both methods are trained and tested using the same train/test splits. The results on iLIDS-VID and PRID-2011 datasets are reported in Figure 5 and confirm that our initialization strategy can significantly improve the re-id performance.

Study of Different Combinations of the Proposed Model

We compare several variants in the initialization step and investigate the roles of different components. Our method is compared with a model without attention, a model without context, and a model without multirate. The experimental results on iLIDS-VID and PRID-2011 datasets are shown in Figure 4.

To begin with, we evaluate the importance of context reconstruction. In a model without context reconstruction, *i.e.*, only one decoder is used, neither past nor future context information is considered, *i.e.*, "Ours w/o context" in Figure 4. The results show that with context prediction, the encoder is able to take into account temporal information around neighboring sequences, hence can model the temporal structures in a better way.

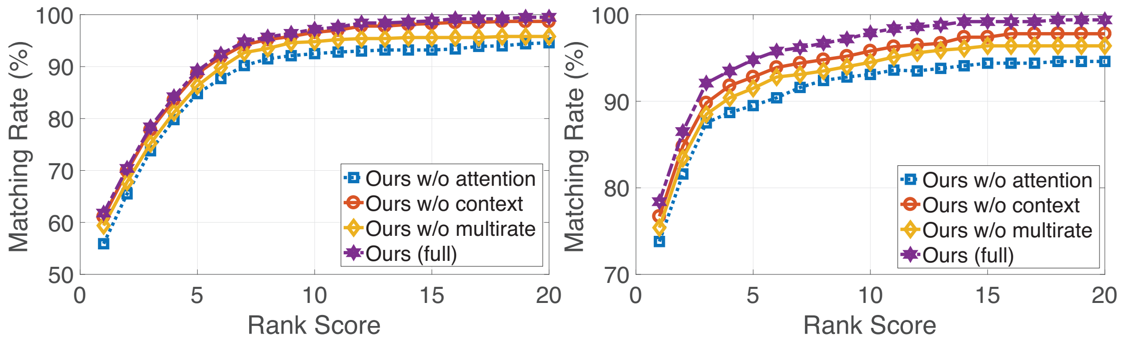
Then, we investigate the effect of the attention mechanism. We compare the proposed model with the same model without the attention mechanism, where temporal attention is removed and the decoder is forced to perform reconstruction based on only the last encoder state, *i.e.*, "Ours w/o attention" in Figure 4. The results confirm that the attention mechanism is important for learning good representations and also helps the learning process of the encoder.

Lastly, we demonstrate the necessity of the multi-rate GRU. We compare with the standard GRU, *i.e.*, "Ours w/o multirate" in Figure 4. The results shows that the proposed model encodes multirate sequence information, and is capable of learning more robust and discriminative representations from the pedestrian sequences.

Comparison with the State-of-the-art

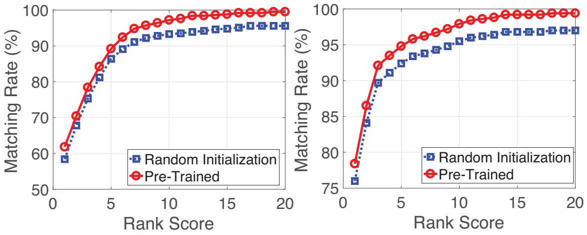
We now compare the re-id performance of the proposed approach against state-of-the-art methods in the literature. As described in the proposed model, numerous CNN architectures can be used in our model, so we also include results of the proposed model with different CNN architectures (CaffeNet, VGG16 and ResNet-50).

In Table 1, we report the CMC results of all the compared algorithms on the iLIDS-VID and PRID-2011 datasets.



(a) iLIDS-VID (b) PRID-2011

Figure 4: CMC curves for iLIDS-VID and PRID-2011 datasets, for different variants in the initialization step. Best viewed in color.



(a) iLIDS-VID (b) PRID-2011

Figure 5: CMC curves for iLIDS-VID and PRID-2011 datasets, comparing the same structure but with different initialization methods. The experimental results demonstrate that our initialization strategy largely boosts the performance of person re-id.

When we compare the results of our model with different CNN architectures (CaffeNet, VGG19 and ResNet-50), we find that the architecture that achieved better performance on ImageNet also performs better on our person re-id datasets, demonstrating the generalization ability of these networks. Comparing all the experimental results in Table 1, we can see that the proposed method outperforms other alternatives by a large margin for both datasets: For instance, on iLIDS-VID, our model achieved rank-1 score 60.8 while the second best achieved 58 and the last only achieved 30.3, whereas on PRID-2011 our model achieved rank-1 score 78.4 while the second best achieved 77.3 and the last only achieved 43. This observation is also consistent among the entire ranking profile.

Generalization Evaluation

Finally, we evaluate the generalization of our model by cross-dataset testing, which may also serve as a good way to avoid over-fitting. Following (McLaughlin, del Rincón, and Miller 2016), we use the large and diverse iLIDS-VID dataset for training and 50% of the PRID2011 dataset for testing. The Recurrent Convolutional Network (RCN) (McLaughlin, del Rincón, and Miller 2016) is used as a baseline. The experimental results are reported in Table 2, and should be compared with Table 1. From the experiments in Table 2, we can observe that the proposed model outperforms RCN with Rank1 CMC of 32.8 *vs*28.4, confirming the superiority of the proposed model in terms of generalization. Comparing the results of the proposed model with the results reported in Table 1, we observe again that the proposed model significantly outperforms the recent method PaMM (Cho and Yoon 2016).

Table 2: Performance comparison of generalization evaluation. Cumulative Matching Characteristics (CMC) curve is used as an evaluation metric. Larger value indicates better performance.

	Train on iLIDS-VID, Test on PRID-2011			
	Rank1	**Rank5**	**Rank10**	**Rank20**
RCN	28.4	57.6	69.2	81.5
Ours	32.8	61.4	72.6	84.3

Conclusion and Future Works

In this paper, we have proposed a novel siamese gated recurrent convolutional network for the video-based pedestrian re-identification problem. To explicitly embed short term and medium term temporal information into the network structure, we propose to use both optical flow and color features together with a recurrent layer. The use of multi-rate gated recurrent unit allows the system to be able to accommodate pedestrians with different motion speeds. We have also introduced an effective initialization strategy for our network via context reconstruction. Experimental results are reported on two standard datasets, and confirm the superiority of the proposed model. In the future, we plan to apply the current framework to other related challenging tasks, *e.g.*, memory question answering.

References

Chang, X.; Yu, Y.; Yang, Y.; and Xing, E. P. 2017. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(8):1617–1632.

Cho, Y., and Yoon, K. 2016. Improving person re-identification via pose-aware multi-shot matching. In *CVPR*.

Cho, K.; van Merrienboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR* abs/1409.1259.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

Haque, A.; Alahi, A.; and Fei-Fei, L. 2016. Recurrent attention models for depth-based person identification. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *Image Analysis*.

Jing, X.; Zhu, X.; Wu, F.; You, X.; Liu, Q.; Yue, D.; Hu, R.; and Xu, B. 2015. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *NIPS*.

Kläser, A.; Marszalek, M.; and Schmid, C. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*.

Koutník, J.; Greff, K.; Gomez, F. J.; and Schmidhuber, J. 2014. A clockwork RNN. In *ICML*.

Kviatkovsky, I.; Adam, A.; and Rivlin, E. 2013. Color invariants for person reidentification. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(7):1622–1634.

Liao, S., and Li, S. Z. 2015. Efficient PSD constrained asymmetric metric learning for person re-identification. In *ICCV*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Liu, K.; Ma, B.; Zhang, W.; and Huang, R. 2015. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*.

Liu, H.; Jie, Z.; Jayashree, K.; Qi, M.; Jiang, J.; Yan, S.; and Feng, J. 2017. Video-based person re-identification with accumulative motion context. *CoRR* abs/1701.00193.

Ma, B.; Su, Y.; and Jurie, F. 2012. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*.

McLaughlin, N.; del Rincón, J. M.; and Miller, P. C. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*.

Ng, J. Y.; Hausknecht, M. J.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *CVPR*.

Su, C.; Yang, F.; Zhang, S.; Tian, Q.; Davis, L. S.; and Gao, W. 2015. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Varior, R. R.; Shuai, B.; Lu, J.; Xu, D.; and Wang, G. 2016a. A siamese long short-term memory architecture for human re-identification. In *ECCV*.

Varior, R. R.; Wang, G.; Lu, J.; and Liu, T. 2016b. Learning invariant color features for person reidentification. *IEEE Trans. Image Processing* 25(7):3395–3410.

Varior, R. R.; Haloi, M.; and Wang, G. 2016. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*, 688–703.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2016. Person re-identification by discriminative selection in video ranking. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(12):2501–2514.

Xiong, F.; Gou, M.; Camps, O. I.; and Sznaier, M. 2014. Person re-identification using kernel-based metric learning methods. In *ECCV*.

Xu, Z.; Yang, Y.; and Hauptmann, A. G. 2015. A discriminative CNN video representation for event detection. In *CVPR*.

Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; and Yang, X. 2016. Person re-identification via recurrent feature aggregation. In *ECCV*.

Yang, Y.; Liao, S.; Lei, Z.; Yi, D.; and Li, S. Z. 2014a. Color models and weighted covariance estimation for person re-identification. In *ICPR*.

Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; and Li, S. Z. 2014b. Salient color names for person re-identification. In *ECCV*.

You, J.; Wu, A.; Li, X.; and Zheng, W. 2016. Top-push video-based person re-identification. In *CVPR*.

Zhang, Z.; Chen, Y.; and Saligrama, V. 2014. A novel visual word co-occurrence model for person re-identification. In *ECCV 2014 Workshops*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. MARS: A video benchmark for large-scale person re-identification. In *ECCV*.

Zheng, W.; Gong, S.; and Xiang, T. 2013. Reidentification by relative distance comparison. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(3):653–668.

Zhu, X.; Jing, X.; Wu, F.; and Feng, H. 2016. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*.

Zhu, L.; Xu, Z.; and Yang, Y. 2016. Bidirectional multi-rate reconstruction for temporal modeling in videos. *CoRR* abs/1611.09053.