Two-Stream Multi-Rate Recurrent Neural Network for Video-Based Pedestrian Re-Identification

Zhiqiang Zeng, Zhihui Li*, De Cheng, Huaxiang Zhang, Kun Zhan and Yi Yang

***Abstract*—Video-based pedestrian re-identification is an emerging task in video surveillance and is closely related to several real-world applications. Its goal is to match pedestrians across multiple non-overlapping network cameras. Despite the recent effort, the performance of pedestrian re-identification needs further improvement. Hence, we propose a novel two-stream multi-rate recurrent neural network for video-based pedestrian re-identification with two inherent advantages: (1) capturing the static spatial and temporal information; (2) dealing with motion speed variance. Given video sequences of pedestrians, we start with extracting spatial and motion features using two different deep neural networks. Then we explore the feature correlation which results in a regularized fusion network integrating the two aforementioned networks. Considering that pedestrians, sometimes even the same pedestrian, move in different speeds across different camera views, we extend our approach by feeding the two networks into a multi-rate recurrent network to exploit the temporal correlations. Extensive experiments have been conducted on two real-world video-based pedestrian re-identification benchmarks: iLIDS-VID and PRID 2011 datasets. The experimental results confirm the efficacy of the proposed method. Our code will be released upon acceptance.**

***Index Terms*—Video Surveillance, Person Re-Identification, Recurrent Neural Networks**

I. Introduction

THE extensive deployment of close-circuit television cameras (CCTV) has made surveillance video acquisition convenient for the general public [1], [2], [3], [4], [5], [6]. Consequently, the number of surveillance videos has grown at an unprecedented rate. How to effectively interpret these data has become an important challenge for multimedia and computer vision communities. One of the most challenging tasks is to re-associate a specific pedestrian across non-overlapping network cameras, which is known as pedestrian re-identification (re-id) [7], [8], [9]. It has received a lot of research attention since it can be applied to real-world video surveillance [10], [11], [12], [13], but remains a challenging problem and needs more effort for performance improvement.

Most existing work of pedestrian re-identification focuses on image based person re-id problem [14], [15], which generally falls into two categories. The first group aims to extract discriminative and informative features that are invariant to viewpoint and background modification [16], [17], [18], [19]. For example, some researchers propose to project the original features into a new space with higher discriminative ability [20]. Given extracted features, the second group employs metric learning methods that emphasize inter-pedestrian distance and de-emphasize intra-pedestrian distance [21]. The final decision is made based on the learnt metric. Various methods have been proposed in this direction, *e.g.*, Relevance Component Analysis (RCA) [22], Large Margin Nearest-Neighbour (LMNN) [23], Relaxed Pairwise Learning (PRLM) [21], *etc*. Although researchers have achieved promising results, pedestrian re-identification remains a challenging problem. On one hand, the performance is yet to be robust because there are usually significant changes of a pedestrian's appearance across different camera views due to the changes in body pose, view angle and illumination. On the other hand, in a real-world scenario pedestrians always appear in a video sequence instead of a still image. These traditional image based re-id algorithms fail to consider the temporal information in the video sequences.

Regarding video sequences of pedestrians, researchers have proposed several algorithms to consider the rich temporal information contained in them and re-associate pedestrians in the sequence level [24], [25], [26], [27], [28]. For example, [29], [30], [31] first extract spatial-temporal features to represent each pedestrian sequence and then conduct pedestrian re-identification with these features. Specifically, they break down each video to

Corresponding author: Zhihui Li.

Zhiqiang Zeng was in part supported by the National Natural Science Foundation of Fujian Province, China (Grant Nos. 2016 J01324, 2017 J01511), Scientific Research Fund of Fujian Provincial Education Department (Grant No. JA15385). Huaxiang Zhang was in part supported by National Natural Science Foundation of China under grant number 61772322.

Zhiqiang Zeng is with the College of Computer and Information Engineering, Xiamen University of Technology. Email: lbxzzq@163.com.

Zhihui Li is with Beijing Etrol Technologies Co., Ltd. Email: zhihuilics@gmail.com.

De Cheng is with the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University. Email: chengde19881214@stu.xjtu.edu.cn.

Huaxiang Zhang is with the School of Information Science and Engineering, Shandong Normal University. Email: huaxzhang@hotmail.com

Kun Zhan is with the School of Information Science and Engineering, Lanzhou University. Email: ice.echo@gmail.com

Yi Yang is with the Centre for Artificial Intelligence, University of Technology Sydney and the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China. Email: yi.yang@uts.edu.au

Manuscript received July 25 2017, revised August 29 2017, revised September 30 2017, accepted October 5 2017.

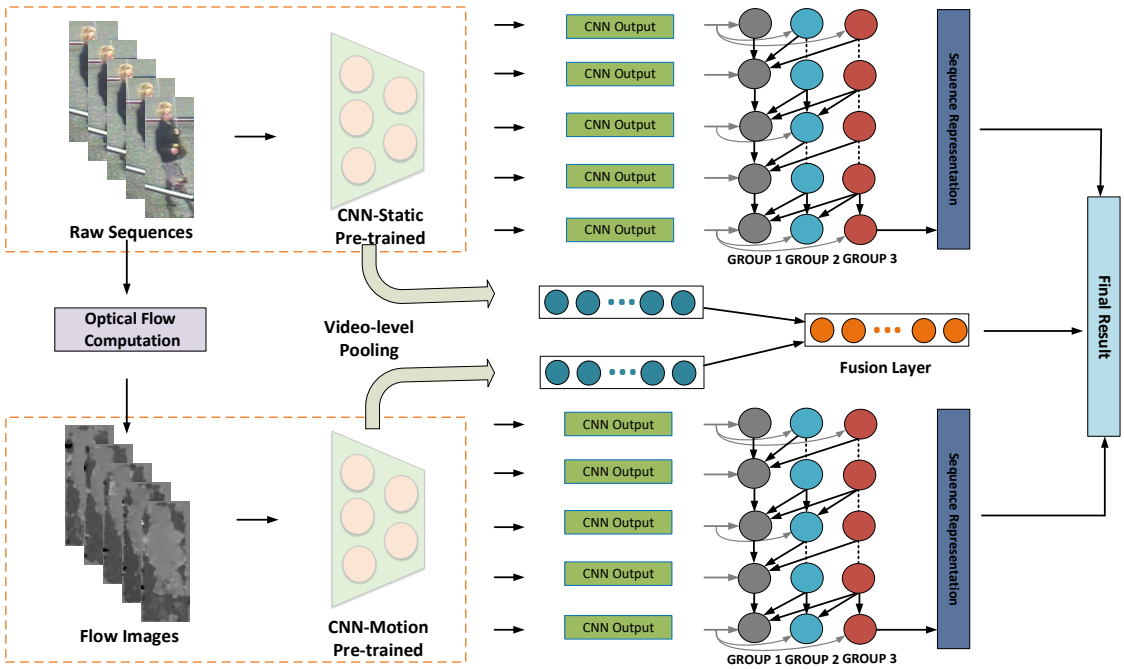


Fig. 1: An overview of the proposed two-stream multi-rate recurrent neural network for video-based pedestrian re-identification.

generate multiple fragments (walking cycles), and extract spatial-temporal feature from each fragment, and then represent each pedestrian with a set of the extracted spatial-temporal features. Hence, they cast the vide-based pedestrian re-id as a set-to-set matching problem. Researchers have also attempted to utilize deep learning to simultaneously conduct feature representation learning and metric learning [32], [33], [28]. McLaughlin *et al*. employ a Siamese network for discriminative feature learning and a RNN to exploit the interaction between different video frames [32]. Yan *et al*. adopt the Long-Short Term Memory (LSTM) to merge the frame-level features for video-based re-id [33]. Although these algorithms have acheived promising results on standard benchmark datasets, they still have the following drawbacks. First, most of these algorithms use a single convolutional neural network (CNN) for both spatial and temporal feature extraction. Nonetheless, the learning ability of spatial and temporal information is limited with a single CNN. Second, they fail to consider that pedestrians, sometimes even the same pedestrian, move in different speed across different camera views.

To address the above limitations, we propose a two-stream multi-rate recurrent neural network for video-based pedestrian re-id, which can not only capture both spatial and temporal information sufficiently, but also enable information sharing between different encoding rates, resulting in a multi-resolution representation that is robust to the motion rate of pedestrians. Figure 1 gives an overview of the proposed two-steam multi-rate recurrent neural network for video-based pedestrian re-id. We first extract spatial and motion features using two types of CNN, which are trained from static frames and stacked motion optical flows respectively. Then we feed these features into two sets of multi-rate recurrent neural networks (a GRU to be specific), which can encode sequences of a pedestrian with different intervals. This learning process enables the system to be more capable of dealing with motion speed variance. To step further, we adopt a regularized fusion layer to combine these two features. Finally, we combine the output of the fusion layer and the multi-rate GRU, attaining in the final results.

Contributions. The contributions of this paper can be summarized as follows. (1) We propose a novel two-stream multi-rate recurrent neural network for video-based pedestrian re-id problem. It can not only model spatial and temporal information, but also deal with motion speed variance. (2) To explore feature correlation, we employ a regularized fusion network to merge the spatial and motion features for pedestrian re-id. (3) To validate the effectiveness of the proposed algorithm, we conduct extensive experiments on two benchmark datasets: iLIDS-VID [29] and PRID-2011 [20]. Compared to the state-of-the-art alternatives, the proposed algorithm consistently achieves the best performance.

II. Related Work

A. Image-Based Pedestrian Re-ID

Previous work on image-based pedestrian re-id can be grouped into two categories: invariant feature representation learning and distance metric learning. The first category aims to learn discriminate features that are invariant to view-point and environmental changes.

It plays a vital role in pedestrian re-id problem. [34] proposes to simultaneously learn an ensemble of informative local features and classifiers to combine spatial and color information. It also shows how to use the AdaBoost algorithm to learn both the object class specific representation and the discriminative recognition model. [17] adopts pictorial structures to localize the parts, extracts and matches their descriptors. The algorithm learns the appearance of an individual and improves the localization of its parts, thus obtaining more reliable visual features for pedestrian re-id. [35] employs Fisher Vectors (FV) to encode the local descriptors, resulting in a global representation of an image. Based on the logchromaticity (log) color space, [16] proposes a new illumination-invariant feature representation and indicates that using color as a single cue shows promising performance for pedestrian re-id under greatly varying imaging conditions.

Given extracted discriminative feature representations, researchers use distance metric learning to make the distance between the same pedestrian close while keeping different pedestrians separated. [23] propose a large margin nearest neighbor metric (LMNN) for the traditional k-NN classification. Their framework mades no parametric assumptions about the structure or distribution of the data and scales naturally to problems with large number of classes. [36] reformulates the pedestrian re-id problem as a ranking problem and learns a subspace where the potential true match is given highest ranking rather than any direct distance measure. [37] formulates pedestrian re-id as a relative distance comparison learning problem to learn the optimal similarity metric between a pair of pedestrian images. [38] analyzes the horizontal occurrence of local features and maximizes the occurrence to make a stable representation against viewpoint changes. It learns a discriminant low dimensional subspace by cross-view quadratic discriminant analysis, and simultaneously learns the distance metric based on the derived subspace.

B. Video-Based Pedestrian Re-ID

In many realistic scenarios, pedestrians always appear in a video rather than in an image. The existing image-based methods fail to make full use of the temporal sequence information in surveillance videos. Several algorithms have been proposed to solve the multi-shot pedestrian re-id problem, *i.e.*, to match pedestrians at the video level. [39] proposes to solve the video-based pedestrian re-id problem by employing Dynamic Time Warping (DTW). [40] adopts conditional random field (CRF) to ensure the final labeling gives similar labels to detections that are similar in feature space. [29] derives a multi-fragment based space-time feature representation of image sequence of pedestrians, based on which a discriminative video ranking model is developed for cross-view re-identification by simultaneously selecting and matching more reliable space-time features from video fragments. [41] hypothesizes that the feature vector of a probe image approximately lies in the linear span of the corresponding gallery feature vectors in a learned embedding space, and formulates the re-id problem as a block sparse recovery problem. [31] proposes a spatio-temporal appearance representation method together with the extraction of feature vectors that encode the spatially and temporally aligned appearance of the pedestrian in a walking cycle. [42] proposes a top-push distance learning (TDL) model to address the video-based pedestrian re-id problem and introduces a top-push constraint to quantify ambiguous video representation.

Deep learning based methods have also been employed for pedestrian re-id problem by simultaneously learning feature representation and distance metric learning. They are trained based on pairs [43] or triplets [44] of input images. A deep network, *i.e*. Siamese network [45], is employed for feature mapping from raw images to a feature space where images from the same pedestrians are close while images from different pedestrians are well separated. [32] introduces a new temporal deep neural network architecture for video-based re-id problem. It utilizes optical flow, recurrent layers and mean-pooling to embed the temporal hierarchy inherent to the problem in the form of short, middle and long term temporal information respectively. Different from other multi-shot pedestrian re-id methods that use complex feature descriptors or design complex matching metrics, [33] aims to learn discriminate sequence level representation from simple frame-wise features using the Long-Short Term Memory (LSTM) network to aggregate the features in a recurrent manner. [46] proposes an end-to-end Accumulative Motion Convext Network (AMOC) based method addressing video-based pedestrian re-id problem through joint spatial appearance learning and motion context accumulating from raw video frames. However, these approaches have the following limitations. First, none of these approaches consider using optical flow ConvNet to capture temporal motion features. Second, all of them fail to consider the fact that pedestrians, sometimes even the same pedestrian, may move in different speed across different camera views. To overcome these limitations, in this paper we propose a novel two-stream multi-rate recurrent neural network for vide-based pedestrian re-identification. It can not only model spatial and temporal information, but also deal with motion speed variance.

III. The Proposed Approach

In this section, we first extract the spatial and motion features using two types of CNN. Then we feed these features into two sets of Gated Recurrent Unit (GRU)-based temporal modeling. GRU has been firstly used for video analysis in [47]. We further adopt a regularized fusion layer to combine these two features. Finally, we combine the output of the fusion layer and the multi-rate GRU.

A. Overview

Figure 1 shows an overview of the proposed two-stream multi-rate recurrent neural network for video-based pedestrian re-identification. We first use the two-stream CNN approach [48] to extract spatial and motion features, and then feed them into a multi-rate GRU for temporal modeling, which is capable of dealing with speed variance. This model is built upon the observation that pedestrians, sometimes even the same pedestrian, move in different speed across different camera views. We also employ an average pooling approach to combine the spatial and motion features, and fuse them using a regularized fusion layer. Finally, we combine the output of the fusion layer and the multi-rate GRU, attaining the final results.

B. Spatial and Motion CNN Features

In cognitive science, researchers claim that human visual system processes what we see through the ventral pathway and the dorsal pathway. The ventral pathway focuses on the spatial information, such as shape and color, while the dorsal pathway focuses on the motion information. Similarly, the video sequence of a pedestrian can be decomposed into the spatial and temporal components. Specifically, the pedestrian in the sequence frame belongs to the spatial component. The complementary temporal component contains motion information across sequences. We adopt the recent two-stream neural network approach [48], [49], [50] to extract the spatial and motion features. Different from existing work on video-based pedestrian re-id [32] which uses stacked frames in short time windows, we decouple the sequences into spatial and motion streams modeled by two CNNs individually. We build the spatial stream on the sequence frames following the standard CNN-based image classification pipeline that is capable of exploring the static appearance information contained in the video frames. We build the motion stream on the stacked optical flows, which are computed between each pair of adjacent video sequence frames.

C. Temporal Modeling with Multi-Rate GRU

We first revisit the basic GRU, which is a particular type of RNN and was proposed to allow each recurrent unit to adaptively capture dependencies of different time scales [51]. It does not have any mechanism to control the degree to which its state is exposed, but rather exposes 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 allowing it to forget the previously computed state [52]. 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.

a) Multi-rate Gated Recurrent Unit.: Next, we discuss the multi-rate extension of GRU as in [53], [13]. The clockwork RNN [53] 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 [13], 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 $(t \bmod T_i) = 0$ 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) include previous slower states (*i.e.* smaller T_i). Thus, the faster states incorporate not only information 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 1. 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.

D. Fusion Strategies

Fusing multiple features is a standard technique in video analysis, which can often lead to better performance. This is because correlations may exist between the spatial and the temporal features of the same pedestrian sequence. Feature fusion method should be capable of exploring the correlations while maintaining the unique characteristics to generate a better fused representation. Instead of using a naive late fusion method, we utilize a regularized fusion layer to combine the results. This is shown in Figure 1. To begin with, we employ average pooling to fuse the frame-level features and achieve the video-level representations. We non-linearly map the input features to a specific layer and then fuse these features using a regularized fusion layer.

Let N be the total number of training sequences with both the spatial and the motion representations. For the n-th sequence sample, it can be represented by $(\mathbf{x}_n^s, \mathbf{x}_n^m, \mathbf{y}_n)$, where $\mathbf{x}_n^s$ and $\mathbf{x}_n^m$ represent the averaged spatial and motion feature respectively. $\mathbf{y}_n$ represents the label of the n-th sequence. We denote $f(\mathbf{x})$ as the mapping function of the neural network from the input $\mathbf{x}$ to the output.

Following [49], we propose to preserve the unique discriminative information so that the complementary information can be explored to improve re-id performance. Hence, we add another regularizer to the objective function, arriving at:

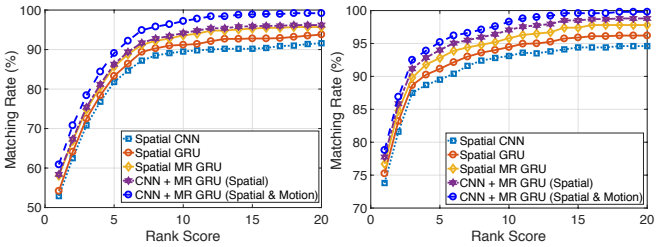
$$\min_{\mathbf{W}} \mathcal{L} + \lambda_1 \Phi(\mathbf{W}) + \frac{\lambda_2}{2} \|\mathbf{W}^E\|_{2,1} + \lambda_3 \|\mathbf{W}^E\|_{1,1}. \tag{13}$$

We use the last term as a complement of the $\ell_{2,1}$-norm. It provides the robustness of the $\ell_{2,1}$-norm by sharing incorrect features among different feature representations. In this way, we can allow different representations to emphasize different hidden neurons. We optimize the objective function using the gradient descent method. We will introduce the parameter tuning later.

IV. Experiments

A. Dataset Description

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



(a) iLIDS-VID (b) PRID 2011

Fig. 2: Performance of GRU, MR GRU and the CNN models trained with the spatial and short-term motion features on ILIDS-VID and PRID-2011 datasets.

a) iLIDS-VID dataset [29].: The iLIDS-VID dataset consists of 600 video sequences of 300 distinct individuals. Each video 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.

b) PRID 2011 dataset [20].: The PRID 2011 dataset contains 400 video sequences of 200 randomly sampled people from two cameras. Each video sequence has variable length ranging from 5 to 675 frames, with an average length of 100. Following [33], [29], we use the sequence pairs with more than 21 frames in all the experiments.

B. Experimental Setup

To extract features, we adopt the recently proposed VGG [54] and CNN_M [48] networks. These two networks have achieved promising results on the ImageNet validation set. We first train the spatial CNN on ImageNet and then fine-tune it on the re-id datasets. The resulted spatial CNN is better than training it from scratch on the re-id datasets. The motion CNN, by contrast, is trained from scratch. Following [48], we also use simple data augmentation methods like cropping and flipping to train motion CNN.

When training the neural networks, we employ mini-batch stochastic gradient descent. The momentum is fixed at 0.9. The spatial CNN is fine-tuned, with its learning rate starting from 10^{-3}, decreasing to 10^{-4} after 14K iterations, and decreasing to 10^{-5} after 20K iterations. When training the temporal network, we start the learning rate from 10^{-2} and decrease it to 10^{-4} after 200K iterations. The experiments are implemented using Torch. We run the experiments on NVIDIA Titan X Pascal. The network weights of multi-rate GRU are trained with ADAM [55]. We fix the learning rate at 10^{-4} and clip the global gradients at norm 10. The cell size is set to 1,024.

Four layers of neurons are used for the regularized fusion network. As shown in the middle of Figure 1, we

use one layer with 200 neurons for spatial and temporal features respectively. After that, we use one fusion layer with the regularization norms. When training the fusion network, we start the learning rate from 0.7. To avoid over-fitting problem, we set the regularization parameter λ_1 to 3×10^{-5}. λ_2 and λ_3 are tuned using cross-validation.

Following [29], [32], we randomly split each dataset into 50% of pedestrians for training and 50% of pedestrians 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 used datasets, we evaluate the performance by the average CMC (Cumulative Matching Characteristics).

C. Effect of Temporal Modeling

First, we conduct experiments to evaluate the multi-rate GRU (MR GRU) in terms of modeling the long-term temporal features for pedestrian re-id. We report the experimental results in Figure 2. Spatial CNN is used as a baseline, which does not consider the temporal order information of the sequence frames. It is observed that Spatial GRU performs much better than Spatial CNN on both datasets, which confirms the benefit of modeling temporal information for pedestrian re-id. To demonstrate the superiority of MR GRU over the classic GRU, we further compare their performance on both datasets. We observe that MR GRU performs much better than GRU using CMC as an evaluation metric. We attribute this improvement to the fact that MR GRU is capable of considering motion variance.

Comparing to other baseline methods, we observe that the proposed approach improves the re-id performance on both datasets. On both datasets, it is noticed that the best performance is obtained when the multi-rate GRU is utilized with both spatial and motion features used. The performance gain of using both spatial and motion features over that of using spatial feature only demonstrates that motion feature is able to provide complementary information for pedestrian re-id problem.

D. Effect of Feature Fusion

In this section, we compare the regularized fusion network with the alternative fusion methods in terms of pedestrian re-id problem. We use the spatial CNN and motion CNN features. We present the experimental results in Table I. We first report the performance of individual features extracted from the fc6 layer of the CNN models, based on which SVM is applied. We use them as baseline methods to show the benefit of SVM based fusion methods. We also compare with other neural network based fusion methods.

From the experimental results reported in Table I, we can see that the re-id performance is significantly improved using the SVM based fusion methods, which demonstrates that motion feature can provide complementary information for re-id. We also notice that the performance of motion features is worse than that of spatial features. This indicates that appearance feature is more important than motion feature for pedestrian re-id problem.

The regularized fusion network achieves consistently better results than the other alternative neural network based fusion methods. The performance gain of the regularized fusion network validates its advantage for pedestrian re-id. To evaluate the benefit of the regularizer, we also compare with a baseline method without the regularizer. We can see that the regularizer helps improve the re-id performance by 1.3% on iLIDS-VID in terms of Rank-20 accuracy as shown in Table I.

E. Comparison with State of the Art

We now compare the performance of the proposed video-based re-id framework to the state-of-the-art approaches, including DVDL [15], AFDA [56], PaMM [57], SI^{2}DL [58], CNN+XQDA [59], STFV3D+KISSME [31], DVR [30], RCN [32], andTDL [42]. To make a fair comparison, we test the approaches using the same re-id datasets and the same training/test splits.

In Table II, we report experimental results on the iLIDS-VID and PRID-2011 datasets in terms of CMC, comparing with other state-of-the-art video-based pedestrian re-id systems. From the experimental results we can see that the proposed approach is more competitive than the other state-of-the-art related methods. For example, our method outperforms RCN [32] for 1.4% and 8.7% in terms of rank-1 matching rate on iLIDS-VID and PRID-2011 datasets respectively. Note that RCN uses optical flow and recurrent layers to embed the temporal hierarchy. Hence, we attribute our improvement to the fact that we utilize a temporal convnet to model temporal features and the multi-rate GRU is capable of dealing with motion variances.

V. Conclusions

We have proposed a novel two-stream multi-rate recurrent neural network for video-based pedestrian re-identification. This network can not only capture static visual and dynamic motion information, but also deal with speed variance. In the framework, we first extract the spatial and motion features using two types of deep neural networks. We feed these two features to separate multi-rate GRU network to capture temporal information. We employ a fusion layer to fuse the two features at sequence level to further boost the re-id performance. To validate the re-id performance of the proposed framework, we conduct extensive experiments on two benchmark datasets. The experimental results demonstrate that the proposed framework outperforms the other alternatives remarkably.

References

[1] T. Räty, "Survey on contemporary remote surveillance systems for public safety," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 40, no. 5, pp. 493–515, 2010.

TABLE I: Performance comparison on ILIDS-VID and PRID-2011 datasets using different fusion approaches to combine the spatial and short-term motion features.

	iLIDS-VID				PRID-2011			
	Rank1	**Rank5**	**Rank10**	**Rank20**	**Rank1**	**Rank5**	**Rank10**	**Rank20**
Spatial SVM	47.8	76.4	84.2	84.7	64.2	80.9	83.5	85.9
Motion SVM	28.5	54.3	62.7	63.4	43.8	59.4	62.8	64.1
SVM-EF	49.2	78.1	85.7	86.2	65.9	82.3	85.1	87.3
SVM-LF	50.4	79.4	86.8	87.4	67.2	84.1	85.9	88.5
SVM-MKL	50.8	80.8	87.4	87.9	68.1	84.8	86.3	88.8
NN-EF	49.8	78.5	86.4	87.8	67.8	84.7	86.3	88.6
NN-LF	49.5	78.1	85.9	87.2	67.1	84.3	85.8	87.9
Two-Stream CNN	52.4	82.8	88.2	90.2	70.4	88.4	90.4	91.8
Fusion Network w/o Regularizer	56.4	85.8	93.5	94.8	74.5	90.3	93.7	96.2
Fusion Network w/ Regularizer	**57.8**	**86.6**	**94.2**	**96.1**	**75.9**	**91.7**	**94.6**	**97.3**

TABLE II: Performance comparison with state-of-the-art alternatives on ILIDS-VID and PRID-2011 datasets.

	iLIDS-VID				PRID-2011			
	Rank1	**Rank5**	**Rank10**	**Rank20**	**Rank1**	**Rank5**	**Rank10**	**Rank20**
DVDL [15]	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6
AFDA [56]	37.5	62.7	73.0	81.8	43.0	72.7	84.6	91.9
PaMM [57]	30.3	56.3	70.3	82.7	45.0	72.0	85.0	92.5
SI^2DL [58]	48.7	81.1	89.2	97.3	76.7	**95.6**	96.7	98.9
CNN+XQDA [59]	53.0	81.4	–	95.1	77.3	93.5	–	99.3
STFV3D+KISSME [31]	44.3	71.7	83.7	91.7	64.1	87.3	89.9	92.0
DVR [30]	39.5	61.1	71.7	81.0	40.0	71.7	84.5	92.2
RCN [32]	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
TDL [42]	56.3	87.6	95.6	98.3	56.7	80.0	87.6	93.6
Ours	**59.4**	**89.8**	**97.3**	**99.1**	**78.7**	95.2	**97.6**	**99.2**

[2] Y. Jararweh, I. Obaidat, and B. B. Gupta, "Automated wireless video surveillance: an evaluation framework," *Journal of Real-Time Image Processing*, pp. 1–20, 2016.

[3] I. Ababneh, S. Bani-Mohammad, and M. A. Smadi, "Corner-boundary processor allocation for 3d mesh-connected multicomputers," *IJCAC*, vol. 5, no. 1, pp. 1–13, 2015.

[4] S. M. Hossain, G. Muhammad, W. Abdul, B. Song, and B. Gupta, "Cloud-assisted secure video transmission and sharing framework for smart cities," *Future Generation Computer Systems*, 2017.

[5] C. Deng, R. Ji, D. Tao, X. Gao, and X. Li, "Weakly supervised multi-graph learning for robust image reranking," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 785–795, 2014.

[6] Z. Zha, M. Wang, Y. Zheng, Y. Yang, R. Hong, and T. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, 2012.

[7] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting listwise similarities," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 4741–4755, 2015.

[8] X. Xu, W. Li, and D. Xu, "Distance metric learning using privileged information for face verification and person re-identification," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 12, pp. 3150–3162, 2015.

[9] P. Ghosh, S. Shakti, and S. Phadikar, "A cloud intrusion detection system using novel PRFCM clustering and KNN based dempster-shafer rule," *IJCAC*, vol. 6, no. 4, pp. 18–35, 2016.

[10] Y. Chen, W. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *IJCAI*, 2015.

[11] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[12] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[13] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," *CoRR*, vol. abs/1611.09053, 2016.

[14] S. Chen, C. Guo, and J. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.

[15] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015.

[16] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, 2013.

[17] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011.

[18] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.

[19] B. B. Gupta, R. C. Joshi, and M. Misra, "ANN based scheme to predict number of zombies in a ddos attack," *I. J. Network Security*, vol. 14, no. 2, pp. 61–70, 2012.

[20] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011.

[21] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012.

[22] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, pp. 937–965, 2005.

[23] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[24] X. Chang, Y. Yu, Y. Yang, and E. P. Xing, "Semantic Pooling for Complex Event Analysis in Untrimmed Videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1617–1632, 2017.

[25] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE Trans. Cybernetics*, vol. 47, no. 5, pp. 1180–1197, 2017.

[26] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *IEEE Trans. Image Processing*, vol. 26, no. 4, pp. 1746–1758, 2017.

[27] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe, "Weakly supervised photo cropping," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 94–107, 2014.

[28] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.

[29] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *ECCV*, 2014.
[30] T. Wang, S. Gong, and X. Zhu, "Person re-identification by discriminative selection in video ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, 2016.
[31] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *ICCV*, 2015.
[32] N. McLaughlin, J. M. del Rincón, and P. C. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016.
[33] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *ECCV*, 2016.
[34] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.
[35] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV*, 2012.
[36] B. J. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *BMVC*, 2010.
[37] W. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, 2013.
[38] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.
[39] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in *ECCV*, 2012.
[40] S. Karaman and A. D. Bagdanov, "Identity inference: Generalizing person re-identification scenarios," in *ECCV*, 2012.
[41] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *CVPR Workshops*, 2015.
[42] J. You, A. Wu, X. Li, and W. Zheng, "Top-push video-based person re-identification," in *CVPR*, 2016.
[43] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.
[44] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.
[45] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.
[46] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, "Video-based person re-identification with accumulative motion context," *CoRR*, vol. abs/1701.00193, 2017.
[47] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 409–421, 2017.
[48] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.
[49] Z. Wu, X. Wang, Y. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *ACM MM*, 2015.
[50] H. Ye, Z. Wu, R. Zhao, X. Wang, Y. Jiang, and X. Xue, "Evaluating two-stream CNN for video classification," in *ICMR*.
[51] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *CoRR*, vol. abs/1409.1259, 2014.
[52] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.
[53] J. Koutník, K. Greff, F. J. Gomez, and J. Schmidhuber, "A clockwork RNN," in *ICML*, 2014.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
[56] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," in *BMVC*, 2015.
[57] Y. Cho and K. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *CVPR*, 2016.
[58] X. Zhu, X. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *IJCAI*, 2016.
[59] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

Zhiqiang Zeng received M.Sc. and Ph.D. degrees in Computer Science from the Xi'an Jiaotong University and Zhejiang University, China, in 2004 and 2007, respectively. He holds a Bachelor degree in Automation from Sichuan University, China. In 2008, he joined the Computer Science Department of the Xiamen University of Technology as a Research Associate. His interests include pattern recognition and machine learning, in particular, support vector machines and general kernel methods.

Zhihui Li received the B.S. degree from Beijing University of Posts and Telecommunications in 2008. After that, she has been working as a Data Analyst in Beijing Etrol Technologies Co., Ltd. Her research interests include artificial intelligence, machine learning, and computer vision.

De Cheng received the B.S. degree in automation control from Xi'an Jiaotong University, Xi'an, China, in 2011, where he is currently working toward the Ph.D. degree with the Institute of Artificial Intelligence and Robotics. His research interests include pattern recognition and machine learning, specifically in the areas of object detection and image classification.

Huaxiang Zhang is currently a professor with the Schoolof Information Science and Engineering, Shandong Normal University, Shandong China. He received his Ph.D. from Shanghai Jiaotong University in 2004, and worked as an associated professor with the Department of Computer Science, Shandong Normal University from 2004 to 2005. He has authored over 170 journal and conference papers and has been granted 10 invention patents. His current research interests include machine learning, pattern recognition, evolutionary computation, cross-media retrieval, etc.

Kun Zhan received the B.S. and Ph.D. degrees from the School of Information Science and Engineering, Lanzhou University, Lanzhou, China, in 2005 and 2010, respectively. He was a visiting student with the Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS, Canada, from 2009 to 2010. He is currently with Lanzhou University. His current research interests include machine learning and neural networks.

Yi Yang received the PhD degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. He is also holding a honorary appointment with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China. He was a post-doctoral research in the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video semantics understanding.