

A Two Stream Siamese Convolutional Neural Network For Person Re-Identification

Dahjung Chung Khalid Tahboub Edward J. Delp
Video and Image Processing Laboratory (VIPER)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA

chung123@purdue.edu

ktahboub@purdue.edu

ace@ecn.purdue.edu

Abstract

Person re-identification is an important task in video surveillance systems. It can be formally defined as establishing the correspondence between images of a person taken from different cameras at different times. In this paper, we present a two stream convolutional neural network where each stream is a Siamese network. This architecture can learn spatial and temporal information separately. We also propose a weighted two stream training objective function which combines the Siamese cost of the spatial and temporal streams with the objective of predicting a person's identity. We evaluate our proposed method on the publicly available PRID2011 and iLIDS-VID datasets and demonstrate the efficacy of our proposed method. On average, the top rank matching accuracy is 4% higher than the accuracy achieved by the cross-view quadratic discriminant analysis used in combination with the hierarchical Gaussian descriptor (GOG+XQDA), and 5% higher than the recurrent neural network method.

1. Introduction

In recent years, the number of video surveillance systems has increased dramatically. According to a study by Cisco, Internet video surveillance traffic is projected to increase tenfold between 2015 and 2020 [1]. The continuous monitoring of surveillance data is practically impossible, making the automatic analysis of surveillance video the only plausible solution. Many video analytic methods have been proposed for person detection and tracking, action recognition, crowd analysis and anomaly detection.

One of the fundamental tasks associated with video surveillance systems is person re-identification (ReID). Person re-identification refers to tracking a person across a network of non-overlapping cameras [2, 3]. Given sin-

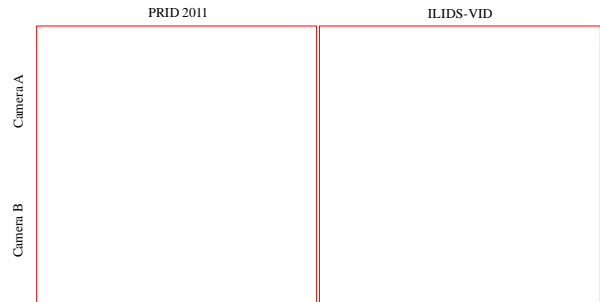


Figure 1: Sample images of two subjects captured from two different cameras in the PRID2011 [4] and the ILIDS-VID [5] datasets

gle/multiple images or a video sequence outlining a person's appearance in the field of view of a camera, person re-identification is the task of recognizing the same person across a network of cameras with non-overlapping fields of view. It can be formally defined as establishing the correspondence between images of a person taken from different cameras at different times [2].

The ReID task is a challenging problem. It remains an active research area due to inter/intra illumination changes, pose variations, occlusions, cluttered background, various scales and viewpoints [3]. Figure 1 shows sample images depicting the differences in camera viewpoints and illumination conditions.

Over the past years, the performance of ReID methods have improved by adopting new features, using metric learning techniques, and the use of semantic attributes and appearance models [6–13]. Most of the traditional approaches proposed for ReID uses low-level features in the form of color and texture histograms and exploit metric learning to find a distance function in which distances between images from the same class are minimized and

distances between different classes are maximized [9, 10, 14–16]. In addition, multiple datasets have been made available for testing, these include: Viewpoint invariant pedestrian recognition dataset (VIPeR) [6], person re-identification (PRID2011) dataset [4] and the iLIDS video re-identification (iLIDS-VID) dataset [5]. VIPeR contains a single image per appearance (single shot), whereas recent datasets, such as PRID2011 and iLIDS-VID, contain multiple images per appearance (multi-shot). ReID in multi-shot scenarios is also referred to as video-based ReID.

[17] demonstrates that multi-shot scenario is superior to single-shot scenario using empirical evidences. Their experiments show that multi-shot approaches are more favorable since both probe and gallery contain much richer visual information as compared to single image. In addition, combining spatial features using multi-shot helps address the challenges associated with viewpoint and pose invariance [5, 18]. Also, in real-life surveillance systems, human detection and tracking methods generate multiple images for each person appearance. Therefore, ReID in multi-shot scenario is more suitable for practical applications.

[19] shows that temporal features (e.g. gait pattern) can offer discriminative features for person identification even using low resolution video sequences. In ReID multi-shot scenarios, these temporal features can be used in combination with spatial features to create better feature representation. Temporal features can improve the accuracy of ReID methods in particular when the majority of clothing worn by subjects tends to be non-discriminative [5, 20].

In [20], a deep learning video-based ReID method using a recurrent convolutional neural network architecture is proposed to exploit both spatial and temporal features. A single network is used to learn a representation for both feature types. This poses a limitation which constrains the amount of information that the network can learn. To address this limitation, we propose the use of a two stream convolutional neural network (CNN) [21] with weighted objective function where each stream has Siamese structure [22].

The main contributions of this paper are:

- We propose a two stream CNN architecture where each stream is a Siamese network. This architecture can learn spatial and temporal information separately. By having two separate networks, each network can learn its own best feature representation.
- We propose a weighted two stream training objective function which combines the Siamese cost of the spatial and temporal streams with the objective to predict a person’s identity. For the ReID task, spatial features are more discriminative than temporal features [8]. The weighted cost function controls the individual contribution of the two streams accordingly. To our best knowledge, this is the first time a weighted

two stream cost function is proposed for ReID.

We evaluate our proposed method on two publicly available datasets. Our proposed method outperforms or shows comparable results to the existing best perform methods on both datasets.

2. Related Work

Recent ReID methods have focused on appearance modeling and metric learning to establish correspondences between people images. The input is assumed to be bounding boxes outlining persons appearances in two different cameras. Each appearance is represented by a single or multiple bounding boxes. We will also assume this in this paper. A common approach is to divide a bounding box into a number of horizontal strips and to extract low-level features from each strip. We will describe some of the features that have been proposed for use in ReID systems. In [6], an ensemble of local features (ELF) is constructed by using the eight color channels corresponding to the three separate channels of the RGB, YCbCr and HSV color spaces with the exception of the value (V) channel. Thirteen Schmid filters and six Gabor filters are also used to model texture. Sixteen bin histograms are constructed for each of the 19 filter responses and for the eight color channels. The histograms are concatenated to form a high dimensional feature vector for each image. Other approach is the use of the local maximal occurrence feature (LOMO) based on multi-scale Retinex to estimate HSV color histograms used for color features [8]. The scale invariant local ternary pattern (SILTP) descriptor is used to model illumination invariant texture [23]. In [7], a hierarchical Gaussian descriptor (GOG) is proposed and is based on the mean and covariance information of pixel features within patches and region hierarchies. Color and texture features are usually concatenated to form a high dimensional feature vector which is used as an input for learning methods.

We now describe some of the classification/learning methods that have been used in ReID. Metric or distance learning is used to find a distance function in which distances between images from the same class are minimized and in which distances between different classes are maximized [8–10, 14–16]. The keep it “simple and straightforward metric learning” method (KISSME) [9] and cross-view quadratic discriminant analysis (XQDA) [8] are widely used metric learning techniques for ReID. Both approaches belong to the class of Mahalanobis distance functions. KISSME, based on a likelihood ratio test, casts the problem in the space of pairwise differences and assumes a Gaussian structure of the difference space [9]. XQDA extends Bayesian faces and the KISSME approach by learning a subspace reduction matrix and a cross-view metric jointly. A closed-form solution is computed by formatting the prob-

lem as a generalized Rayleigh quotient and using eigenvalue decomposition [8].

In [24], a multi shot approach is based on the combination of random projections for dimensionality reduction and random forests for classification. A relative distance comparison model which maximizes the likelihood that a pair of correct match has a smaller distance than that of a wrong match pair along with an ensemble strategy is introduced in [25]. In [26], person re-id is formulated as a block sparse recovery problem and in [27] is formulated as a graph matching problem. In [28], images for a person trajectory are clustered hierarchically to mitigate the problems faced by Fisher Discriminate Analysis (FDA). A viewpoint-invariant descriptor along with sub-image rectification and poses estimation is proposed in [29].

When the majority of clothing worn tends to be non-discriminative, ReID becomes very challenging. Attributes-based re-identification methods try to solve this problem by incorporating semantic attributes. ‘Jacket’, ‘female’ and ‘carried object’ are all examples of semantic attributes. Semantic attributes are mid-level features learned from a larger dataset a priori [30]. In [31], semantic attributes are combined with the low level features and is shown to improve the performance of ReID.

Until recently, CNN architectures [32] have not been used for the ReID due to the small size of public datasets. With the release of larger datasets, recent methods have demonstrated the feasibility of the use of CNNs for ReID [33, 34]. A filter pairing neural network (FPNN) is proposed as a unified solution to extract features and learn photometric and geometric transforms in [33]. In [34], feature extraction layers are followed by a cross-input neighborhood difference layer to compute the differences in feature values across the camera views.

Very recent deep learning ReID methods extended [33, 34] and incorporate metric learning and part-based learning. In [35], a cosine layer connects two sub-networks and jointly learn color, texture and a similarity metric. In [36], multi channels part-based CNN is proposed to jointly learn both global and local body features of the person. The network is trained using triplet images and a triplet loss function is used to learn the network model. In [37], single image and cross-image representations are combined in a single network. A deep learning network for learning features from multiple domains is proposed in [38]. A domain guided dropout (DGD) method is shown to improve feature learning.

Most of the existing deep learning methods are based on a simple architecture and ignore the temporal information in a multi-shot scenario. To exploit the temporal information for ReID, the use of recurrent neural network (RNN) with the Siamese structure is proposed in [20]. An optical flow image is concatenated to the YUV image and comprises the

input to the deep learning network. For the remainder of this paper, we will refer to the ReID technique proposed in [20] as the RNN-ReID technique. Instead of using a single network to learn both spatial and temporal features, we propose the use of a two stream CNN architecture where each stream is a separate Siamese network.

3. Proposed Method

The overall architecture of our proposed method is shown in Figure 2. The method is motivated by the fact that both spatial and temporal features possess discriminative information useful for the ReID task. However, the best feature representation does not need to be the same for both types of features. Therefore, we propose a two stream Siamese CNN which processes spatial and temporal information separately. Siamese CNNs contain two identical sub-networks with shared weights and are suitable for tasks which involve finding the similarity between two comparable inputs [22]. CNNs typically process an image or multiple images and classify them into a single class, whereas Siamese CNNs process two images or two sequences of images and compute the similarity between them. In our proposed ReID system, the input to first stream are two sequences of RGB frames where each sequence is captured from a different camera. The second stream processes the optical flow information from both cameras as shown in Figure 2. The input is described in more details in Section 3.1. Each stream is based on the same network architecture. Throughout this paper, we will refer to the network associated with spatial content as SpatialNet and the network associated with temporal content as TemporalNet.

Both networks are composed of multiple CNNs with Siamese architecture, and all the CNNs within the same stream share the same parameters. We refer to this CNN as the “base CNN” and describe its structure in Section 3.2. The outputs of the base CNNs which processes images from the same camera view are combined using temporal pooling. The temporal pooling is described in Section 3.3. The outputs of the temporal pooling from both cameras are combined using the Siamese cost as described in Section 3.4. Finally, the two networks associated with both streams are fused together using a weighted cost function as described in Section 3.5.

3.1. The Inputs

We define the generic input sequence as I_c , where c is a, b for camera A and B, respectively. For the SpatialNet, the input sequence are RGB frames:

$$I_c = (S^{(1)}, \dots, S^{(t)}, \dots, S^{(L)}) \quad (1)$$

where L is the sequence length and $S^{(t)}$ is the RGB frame at time t .

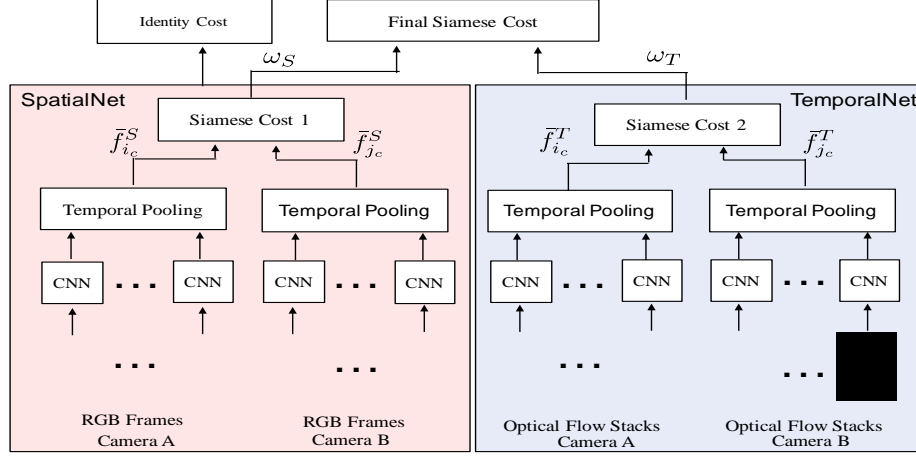


Figure 2: Overall Architecture of the proposed two stream ReID system

For TemporalNet, optical flow images are used as input:

$$\mathbf{I}_c = (\mathbf{T}^{(1)}, \dots, \mathbf{T}^{(t)}, \dots, \mathbf{T}^{(L)}) \quad (2)$$

where L is the sequence length and $\mathbf{T}^{(t)}$ is the input optical flow image at time t . To obtain $\mathbf{T}^{(t)}$, the displacement vectors in the horizontal and vertical directions between a pair of consecutive frames are computed using the Lucas-Kanade optical flow technique [39]. The effectiveness of using optical flow to learn temporal features are demonstrated in [20, 21].

3.2. The Base CNN Architecture

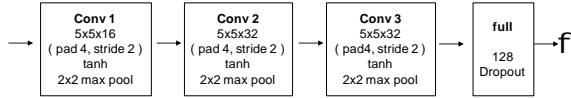


Figure 3: The structure of the base CNN and hyper-parameters

As shown in Figure 2, the input sequence \mathbf{I}_c is processed using the base CNN. Figure 3 shows the base CNN structure and the hyper-parameters associated with it. The CNN takes one input sample ($\mathbf{S}^{(t)}$ or $\mathbf{T}^{(t)}$) and produces the output feature vector \mathbf{f}^S or \mathbf{f}^T for SpatialNet and TemporalNet, respectively. Our base CNN is composed of three convolution layers where each layer has convolution, non-linear activation and max-pooling steps. We use hyperbolic-tangent (tanh) as non-linear activation function. At the end of the three convolution layers, a fully connected layer is placed to have mapping to all the activations from the last convolution layer. Dropout [40] is also used to reduce the model over-fitting.

3.3. Temporal Pooling

For the SpatialNet, the base CNN processes a single RGB frame out of the sequence of frames in the multi-shot scenario, or optical flow content in the case of the TemporalNet. Combining the spatial or temporal features using multiple frames helps address the challenges associated with various viewpoints and poses. To process the input sequence, each sub-network of the Siamese network in each stream utilizes L base CNNs and produces L feature vectors. The feature vectors produced by the L CNNs in each sub-network are combined into a single feature vector using temporal pooling. Max pooling, sum pooling and mean pooling are the most common techniques used to achieve this. In [20], the RNN-ReID method has shown that the mean pooling method is the most suitable temporal pooling technique for the ReID task. We adopt the same approach in our proposed method. If we denote the base CNN by the function $C()$, then the temporally pooled feature vector, $\bar{\mathbf{f}}_{i_c}$, is computed as follows:

$$\bar{\mathbf{f}}_{i_c} = \frac{1}{L} \sum_{t=1}^L C(\mathbf{I}_{i_c}^{(t)}) \quad (3)$$

where i is the person ID, $c \in \{a, b\}$ is the camera view and $\mathbf{I}_{i_c}^{(t)}$, $t = 1, \dots, L$, is one element (RGB image or optical flow vectors) of the input multi-shot sequence. The sequence of images are processed and temporally pooled to obtain the feature vector $\bar{\mathbf{f}}_{i_c}^S$ or $\bar{\mathbf{f}}_{i_c}^T$ for the SpatialNet and TemporalNet, respectively.

3.4. Siamese Cost

Siamese networks are composed of two sub-networks with shared weights [22]. While learning the features from each sub-network, Siamese networks compare the features

from the pair using Euclidean distance. Thus, in training process, the network tries to minimize the distance between feature pairs when they are from the same class and maximize the distance between feature pairs when they are from different classes. Due to this property, Siamese networks have been widely used for the ReID task since the goal is to find the similarity between a pair of sequences. As mentioned before, we use a Siamese network for both streams: SpatialNet and TemporalNet as shown in Figure 2. Furthermore, the generic Siamese cost of our proposed method can be defined as follows:

$$D(\bar{f}_{i_c}, \bar{f}_{j_c}) = \begin{cases} \frac{1}{2} \|\bar{f}_{i_c} - \bar{f}_{j_c}\|^2, & \text{if } i = j \\ \frac{1}{2} \{\max(m - \|\bar{f}_{i_c} - \bar{f}_{j_c}\|, 0)\}^2, & \text{if } i \neq j \end{cases} \quad (4)$$

where m is the Siamese margin and $\bar{f}_{i_c}, \bar{f}_{j_c}$ are the temporally pooled feature vectors for person i and j , respectively. Equation 4 applies to both SpatialNet and TemporalNet in the same way with different type of inputs.

3.5. Weighted Two Stream Joint Identification and Verification

During the training process, we build on the joint identification and verification approach from [41] to define our training objective. We use the softmax loss function to compute the identification cost as in [20]. Then, this cost is integrated into our final training objective function as explained later. The identification cost is defined as:

$$V(x) = P(q = c|x) = \frac{\exp(W_c x)}{\sum_k \exp(W_k x)} \quad (5)$$

where x is the feature vector and q is the person's identity. W_c and W_k indicate the c th and the k th column of the softmax matrix W , respectively. Note that the softmax matrix W is the matrix representation of the fully connected layer in the base CNN architecture.

From the RNN-ReID method, it was already observed that joining the identification with the Siamese cost is crucial to improve the ReID accuracy. We have two Siamese cost functions from each stream, whereas RNN-ReID has only one Siamese cost. Therefore, we define the combined cost function J_f as follows:

$$J_f = \alpha S D(\bar{f}_{i_c}^S, \bar{f}_{j_c}^S) + \beta T D(\bar{f}_{i_c}^T, \bar{f}_{j_c}^T) + V(\bar{f}_{i_c}^S) + V(\bar{f}_{j_c}^S) \quad (6)$$

where V is the standard softmax loss defined in Equation 5. α, β are the weights for SpatialNet and TemporalNet, respectively. Note that we only use the identification cost V which is computed using the spatial features since they contain more information regarding to the person label than the temporal features. We propose using different weights for each stream to be able to emphasize the

spatial features as compared to the temporal features. For ReID Task, even though walking motion adds discriminative power to the ReID solution, spatial features such as appearance, color or texture are relatively more important in terms of re-identifying people. Thus, we set the weights empirically with the condition $\alpha > \beta$.

3.6. Similarity Metric for Testing

The weighted two stream joint identification and verification objective function, which is used for training, incorporates the ability to predict a person's identity. However, during the evaluation, the goal is to find the similarity score (metric) between two sequences of images and to rank the gallery accordingly. Therefore, we modify Equation 6 to disregard the contribution of the standard softmax loss V and replace the Siamese cost D with the Euclidean distance. The Euclidean distances are computed using the temporally pooled feature vectors ($\bar{f}_{i_c}^S, \bar{f}_{i_c}^T, \bar{f}_{j_c}^S$ and $\bar{f}_{j_c}^T$) as follows:

$$d_S = \|\bar{f}_{i_a}^S - \bar{f}_{j_b}^S\| \quad (7)$$

$$d_T = \|\bar{f}_{i_a}^T - \bar{f}_{j_b}^T\| \quad (8)$$

Finally, d_S and d_T are combined using a weighted average to compute the final similarity metric d_F :

$$d_F = \frac{\alpha d_S + \beta d_T}{\alpha + \beta} \quad (9)$$

4. Experiments

In this section, we evaluate our proposed method using the publicly available datasets: Person re-identification (PRID2011) dataset [4] and the iLIDS video re-identification (iLIDS-VID) dataset [5]. We investigate our proposed method with different hyperparameter settings and evaluate the performance against the state-of-the-art ReID methods.

4.1. Datasets

Both datasets feature a multi-shot scenario in which a person trajectory is represented by a sequence of images. The PRID2011 dataset contains images from two non-overlapping static surveillance cameras. The sequence presents the significant differences in viewpoint, illumination and camera characteristics. It is composed of 385 person trajectories from one view and 749 from the other one, with 200 persons appearing in both views. Each image sequence has a variable length ranging from 5 to 675 image frames, with an average number of 100 images. We only consider the 200 persons appearing in both views as suggested in [5].

The iLIDS-VID dataset was created by observing pedestrians in two camera views. The outputs of two non-overlapping cameras were captured at a crowded airport ar-

rival hall. It consists of 600 image sequences of 300 individuals with one pair of sequences from two camera views for each person. Each image sequence has a variable length ranging from 23 to 192 image frames, with an average number of 73 images. It is one of the most challenging datasets due to the cluttered background and random occlusions.

4.2. Experiment Setup

Input images are pre-processed before being fed into the two stream Siamese CNN. Each color channel of the RGB image is normalized to introduce invariance to illumination changes. This is simply done by subtracting the mean and dividing by the standard deviation. Each horizontal and vertical optical flow channel is also normalized to the range of $[-1, 1]$.

The same data augmentation technique in [20] is used to add more variety to the data. Random mirroring and cropping are used for data augmentation. Note that a consistent data augmentation technique is applied to the images from the same sequence.

As suggested in [20], positive and negative pairs are alternatively fed into the network. Sequence pairs are randomly sampled from the all training identities. All training sequence lengths are set to 16 and the test sequence lengths are varied to investigate the significance of the sequence length as described in Section 4.4.1. Note that this sequence length can be arbitrary due to the network architecture.

The proposed network is trained for 1000 epochs using the stochastic gradient descent method. The batch size is set to 1, the learning rate to $1e^{-3}$ and the momentum to 0.9. The Siamese cost function margin is set to $m = 2$. The base CNN feature dimension is 128 with the dropout rate set to 0.5.

4.3. Evaluation Protocol

We follow the evaluation protocol described in [5]. The dataset is randomly split into two subsets with the same size. One is used for training and one for testing. For the testing, the sequences from the first camera are used as the probes while the sequences from the second camera are used as the gallery.

We validate the performance of our proposed method and compare the performance against other methods using the Cumulative Matching Characteristic (CMC) curve which indicates the probability of finding the correct match in the top K matches within the ranked gallery. The experiment is repeated five times by randomly splitting the dataset into training and testing and the average result is reported.

In our proposed method, we have two extra hyper-parameters (α_s, α_t). To see the effectiveness of proposed method, we perform experiments with various hyper-parameters settings. We perform experiments with $\alpha_s = 1$ when α_t is set to 0 or 1 in order to verify the individual

contribution of TemporalNet. We also perform experiments with $\alpha_s = 2, 3$ when $\alpha_t = 1$ to see the relative contribution of the spatial features as compared to the temporal features.

4.4. Results and Discussion

4.4.1 Probe and Gallery Sequence Length

Length	Rank	1	5	10	20
16		41	70	81	92
32		50	79	88	95
64		56	82	91	97
128		58	85	93	97

Table 1: Matching accuracies with various probe/gallery sequence lengths in iLIDS-VID

In this section, we investigate the significance of the sequence length during testing. An experiment is conducted to evaluate the ReID matching accuracy using various sequence lengths. Our proposed network shown in Figure 2 is trained with the sequence length set to 16 using the iLIDS-VID dataset. During evaluation, the matching accuracy is calculated using $\{16, 32, 64, 128\}$ as lengths for the probe and gallery sequences. In the case when the probe or gallery sequence is shorter than the test length, we use the entire sequence.

The matching accuracies for different sequence lengths are summarized in Table 1. The results clearly indicate that the matching accuracies are improved as the sequence length is increased. For instance, when we increase the sequence length from 16 to 128, the top rank matching accuracy is improved by 17%. This is an intuitive result since combining the spatial and temporal features using multiple images helps address the challenges associated with various viewpoints and poses.

4.4.2 Verification on Two Stream

To verify the usefulness of temporal information in ReID task, we perform the experiments with the different settings of the hyper-parameters (α_s, α_t). This also can verify the improvement gained by the use of a two stream CNN architecture. Note that α_s and α_t control the individual contributions of the SpatialNet and the TemporalNet, respectively. When $\alpha_t = 0$, the contribution of TemporalNet becomes totally 0 in training phase. This also applies to the test phase in the same way based on the Equation 9.

We then compare ReID matching accuracies for different hyper-parameter settings such as spatial only case ($\alpha_t = 0, \alpha_s = 1$) and Both Stream cases when α_t is fixed to 1 while α_s is varying from 1 – 3. As shown in Table 2, using both stream cases have 3-4% accuracy improvement

Streams	Rank	1	5	10	20
$s = 1, \tau = 0$		75	93	97	98
$s = 1, \tau = 1$		78	94	94	99
$s = 2, \tau = 1$		78	94	97	99
$s = 3, \tau = 1$		79	93	97	98

(a) PRID2011

Streams	Rank	1	5	10	20
$s = 1, \tau = 0$		57	60	91	95
$s = 1, \tau = 1$		58	86	93	97
$s = 2, \tau = 1$		60	86	93	97
$s = 3, \tau = 1$		56	86	92	96

(b) iLIDS-VID

Table 2: Matching accuracies with different stream settings

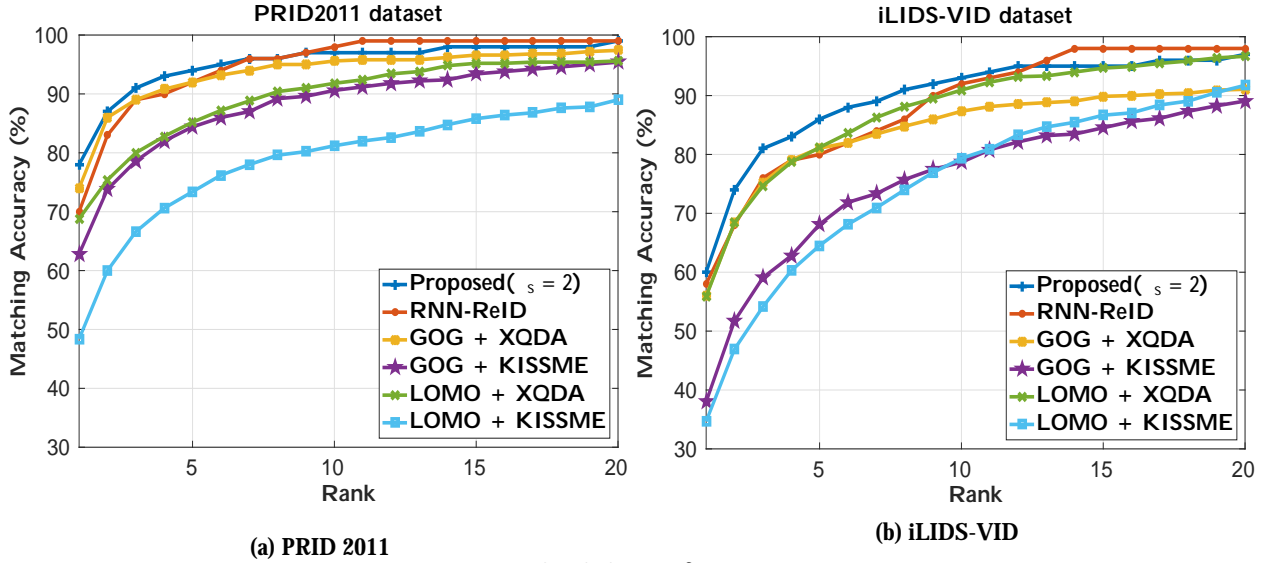
Methods	Rank	1	5	10	20
Proposed ($s = 2$)		78	94	97	99
RNN-ReID		70	90	95	97
GOG + XQDA		74	91	94	96
GOG + KISSME		57	80	89	94
LOMO + XQDA		67	86	92	94
LOMO + KISSME		48	72	82	91
ELF + XQDA		22	43	54	64
ELF + KISSME		15	32	42	56

(a) PRID2011

Methods	Rank	1	5	10	20
Proposed ($s = 2$)		60	86	93	97
RNN-ReID		58	84	91	96
GOG + XQDA		55	79	86	90
GOG + KISSME		38	67	79	89
LOMO + XQDA		53	79	88	95
LOMO + KISSME		35	65	79	90
ELF + XQDA		23	49	60	74
ELF + KISSME		15	40	55	70

(b) iLIDS-VID

Table 3: Matching accuracies comparison with previous methods



(a) PRID 2011

(b) iLIDS-VID

Figure 4: CMC Curves for comparison

in PRID2011 and 1-3% accuracy improvement in iLIDS-VID. This result demonstrates that by having two separate networks to represent the spatial and the temporal content, each network is able to learn the best feature representation and improves the ReID performance. In addition, based on the results for $s = 2$ cases, ReID performance improved in PRID2011 whereas it did not improve in iLIDS-VID for $s = 3$ case. We thus conclude that the optimal relative contribution of the spatial and temporal features is data de-

pendent.

4.4.3 Comparisons

We compare the performance of our proposed method against several of the best performing methods in a multi-shot ReID setting. We evaluate state-of-the-art metric learning methods (XQDA [8] and KISSME [9]) using state-of-the-art feature extraction methods: LOMO [8], GOG [7] and ELF [6]. Since we are evaluating multi-shot ReID

methods, we extract the features for each image in the sequence and compute the average which is used by the metric learning methods. To our best knowledge, the combination of GOG and XQDA achieves state-of-the-art performance and the RNN-ReID method is the best performing deep learning method [20].

The CMC curves are plotted in Figure 4a and 4b and the matching accuracies are summarized in Table 3a and 3b for the PRID2011 and the iLIDS-VID datasets, respectively. For the PRID2011 dataset, our proposed method outperforms all the other methods. The top rank matching accuracy is 4% higher than the accuracy achieved by the GOG+XQDA method and 8% higher than RNN-ReID.

For the iLIDS-VID dataset, the results show that our approach has comparable accuracy to the RNN-ReID method and is 5% higher than the accuracy achieved by the GOG+XQDA method as can be seen in Table 3b and Figure 4b. The top rank matching accuracy for the iLIDS-VID dataset is 18% lower than the case for the PRID2011 dataset. We believe this mainly due to the cluttered background and occlusions associated with the iLIDS-VID dataset. To make our method more robust to these challenging conditions, we plan to incorporate semantic attributes.

5. Conclusion

In this paper, we proposed a person re-identification method based on a two stream convolutional neural network where each stream is a Siamese network. This architecture can learn spatial and temporal information separately in a re-identification setting. Our proposed method is evaluated on the publicly available PRID2011 and iLIDS-VID datasets. We demonstrate that combining the spatial and temporal features using multiple images helps address the challenges associated with viewpoint and pose invariants. Our experimental results also demonstrate that by having two separate networks to represent the spatial and the temporal content, each network is able to learn the best feature representation and improves the ReID performance. In the future, we want to incorporate semantic attributes using a multi-stream approach to address the challenges associated with occlusions and cluttered background.

References

- [1] "Cisco visual networking index: Forecast and methodology, 2015/2020," Cisco Systems Inc., April 2016.
- [2] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, April 2014.
- [3] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. London: Springer, 2014.
- [4] M. Hirzer, C. Beleznaï, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," *Proceedings of the Scandinavian Conference on Image Analysis*, pp. 91–102, May 2011, Ystad, Sweden.
- [5] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," *Proceedings of the European Conference on Computer Vision*, pp. 688–703, September 2014, Zurich, Switzerland.
- [6] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Proceedings of the 10th European Conference on Computer Vision*, pp. 262–275, October 2008, Marseille, France.
- [7] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1372, June 2016, Las Vegas, NV.
- [8] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206, June 2015, Boston, MA.
- [9] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295, June 2012, Providence, RI.
- [10] F. Xiong, M. Gou, O. Camps, and M. Sznai, "Person re-identification using kernel-based metric learning methods," *Proceedings of the 13th European Conference on Computer Vision*, pp. 1–16, October 2014, Zurich, Switzerland.
- [11] R. Layne, T. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," *Proceedings of the British Machine Vision Conference*, vol. 2, no. 3, p. 8, September 2012, Guildford, United Kingdom.
- [12] R. Layne, T. M. T. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," *Proceedings of the European Conference on Computer Vision*, pp. 402–412, October 2012, Berlin, Heidelberg.
- [13] S. Khamis, C. Kuo, V. Singh, V. Shet, and L. Davis, "Joint learning for attribute-consistent person re-identification," *Proceedings of the European Conference on Computer Vision Workshops*, pp. 134–146, October 2014, Zurich, Switzerland.
- [14] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2666–2672, June 2012, Providence, RI.
- [15] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," *Proceedings of the 10th European Conference on Computer Vision*, pp. 780–793, October 2012, Florence, Italy.
- [16] J. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," *Proceedings of the 24th International Conference on Machine Learning*, pp. 209–216, June 2007, Corvallis, OR.

- [17] 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," *Proceedings of the European Conference on Computer Vision*, October 2016, Amsterdam, Netherlands.
- [18] J. You, A. Wu, X. Li, and W. Zheng, "Top-push video-based person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1345–1353, June 2016, Las Vegas, NV.
- [19] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 162–177, 2005.
- [20] N. McLaughlin, J. Martinez, and P. Miller, "Recurrent convolutional network for video-based person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1325–1334, June 2016, Las Vegas, NV.
- [21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 568–576, December 2014, Montreal, Canada.
- [22] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 25–44, January 1994.
- [23] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and Z. Stan, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1301–1306, June 2010, San Francisco, CA.
- [24] Y. Li, Z. Wu, and R. J. Radke, "Multi-Shot Re-Identification with Random-Projection-Based Random Forests," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 373–380, January 2015, Waikoloa, HI.
- [25] W. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, March 2013.
- [26] S. Karanam, Y. Li, and R. Radke, "Sparse re-id: Block sparsity for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 33–40, June 2015, Boston, MA.
- [27] K. Tahboub, B. Delgado, and E. J. Delp, "Person re-identification using a patch-based appearance model," *Proceedings of the IEEE Conference on Image Processing*, pp. 764–768, September 2016, Phoenix, AZ.
- [28] Y. Li, Z. Wu, S. Karanam, and R. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," *Proceedings of the British Machine Vision Conference*, September 2015, Swansea, United Kingdom.
- [29] Z. Wu, Y. Li, and R. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1095–1108, September 2014.
- [30] B. Delgado, K. Tahboub, and E. J. Delp, "Superpixels shape analysis for carried object detection," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pp. 1–6, March 2016, Lake Placid, NY.
- [31] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869–878, May 2015.
- [32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [33] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, June 2014, Columbus, OH.
- [34] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3908–3916, June 2015, Boston, MA.
- [35] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," *Proceedings of the International Conference on Pattern Recognition*, pp. 34–39, August 2014, Stockholm, Sweden.
- [36] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, June 2016, Las Vegas, NV.
- [37] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1288–1296, June 2016, Las Vegas, NV.
- [38] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1258, June 2016, Las Vegas, NV.
- [39] B. D. Lucas, T. Kanade et al., "An iterative image registration technique with an application to stereo vision," *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 81, pp. 674–679, 1981, Vancouver, Canada.
- [40] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, June 2014.
- [41] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1988–1996, December 2014, Montreal, Canada.