Video-based Person Re-identification Using Refined Attention Networks

Tanzila Rahman *
University of British Columbia
trahman8@cs.ubc.ca

Mrigank Rochan and Yang Wang
University of Manitoba
{mrochan, ywang}@cs.umanitoba.ca

Abstract

We consider the problem of video-based person re-identification. The goal is to identify a person from videos captured under different cameras. In this paper, we propose an efficient attention based model for person re-identifying from videos. Our method generates an attention score for each frame based on frame-level features. The attention scores of all frames in a video are used to produce a weighted feature vector for the input video. This video-level feature vector is refined iteratively for re-identifying persons from videos. Unlike most existing deep learning methods that use global or spatial representation, our approach focuses on attention scores. Extensive experiments on three benchmark datasets demonstrate that our method achieves the state-of-the-art performance.

1. Introduction

In this paper, our goal is to solve the problem of video-based person re-identification. Given a video containing a person, the goal is to identify the same person from other videos possibly captured under different cameras. Person re-identification is useful in a wide range of applications, e.g video surveillance, police investigation, etc. A common strategy for person re-identification is to formulate it as a metric learning problem. Given the query video and a candidate video, the goal is to develop algorithms to compute the distance between these two videos. If the distance is small, it means the two videos likely contain the same person. See Figure 1 for an illustration.

Previous work in person re-identification falls into two broad categories: image-based re-identification and video-based re-identification. Earlier work (e.g. [12, 18, 20, 21, 23, 27, 29, 33]) in this area focuses on the former, where the inputs to these systems are pairs of images and the goal is to identify whether they are images of the same person. Recently, video-based person re-identification is receiving increasing attention (e.g. [10, 13, 17, 22, 25, 26, 34, 35]). Compared with static images, video-based person re-identification is a more natural setting for practical applications such as video surveillance.

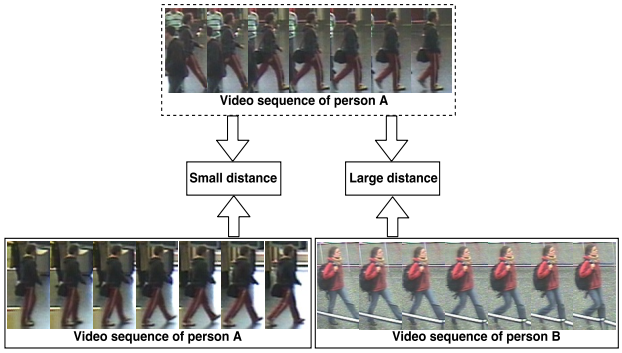

Figure 1. Illustration of the video-based person re-identification problem. In this case, our goal is to identify person A from two video sequences in the second row. If two videos contain the same person, we would like the distance between them to be small. Otherwise, we would like the distance to be large. Some frames in a video sequence may be affected by occlusions and are not informative about the person's identity. In this paper, we use an attention model to focus on informative frames for re-identification.

Person re-identification (either image or video based) is a challenging problem since the images/videos are often captured under different camera views. This can cause large variations in illumination, body pose, viewpoint, etc. Compared with static images, the temporal information in videos can potentially provide additional information that can help disambiguate the identity of a person. Previous work in this area has explored ways of exploiting this temporal information. A common strategy (e.g. [17, 25, 34]) is to use temporal pooling to combine frame-level features to represent the entire video sequence. Then this video-level feature vector can be used for re-identification.

Previous work (e.g. [25, 34]) has made the observation that not all frames in a video are informative. For example, if the person is occluded in a frame, ideally we would like the feature representation of the video to ignore this frame and focus on other "useful" frames. A natural way of solving this problem is to use the attention models [1, 19, 24]

*Work done while at the University of Manitoba.

978-1-5386-9294-3/18/$31.00

that have been popular in visual recognition recently. In [25, 34], RNN is used to model the temporal information of the frames and generate the attention score for each frame for person re-identification.

In this paper, we propose a new attention model for video-based re-identification. Compared with previous works [25, 34], our model has several novelties. First, instead of using RNN, we directly produce the attention score of each frame based on the image feature of this frame. Our experimental results show that this simpler method outperforms RNN-based attention method. Since the attention score of each frame is calculated based on the frame, the computation of attention scores over all frames can be easily made parallel and take full advantage of the GPU hardware. Second, the work in [25, 34] only calculates the attention scores once. In this paper, we introduce a new method to refine the attention scores based on the whole video features. We show that this attention refinement can improve the performance of our model.

Our contributions include:

1. A new attention mechanism for video-based person re-identification. Unlike previous work (e.g. [34]) that uses RNN to generate the attentions, our model directly generates attentions based on frame-based features. As a consequence, the computation of the attentions is much simpler and can be easily parallelized. In contrast, RNN has to process frames in a sequential order, so the computation cannot be made parallel. Despite of its simplicity, our model outperforms the more sophisticated RNN-based attention mechanism in [34].

2. We introduce an iterative refinement process to further improve the attentions. This allows the model to refine the attention scores over time. We show that this attention refinement improves the performance of the final model. In addition, we also study the effect of iterative refinement on the performance.

2. Related Work

There has been extensive work on person re-identification from static images. Early work in this area uses hand-crafted feature representations [3, 11, 15, 16, 31]. Most of these methods involve extracting feature representations that are invariant to viewpoint changes, then learning a distance metric to measure the similarity of two images.

Deep learning approaches, in particularly deep convolutional neural networks (CNNs), have achieved tremendous successes in various visual recognition tasks [8]. In many areas of computer vision, CNNs have replaced hand-engineering feature representations with features learned end-to-end from data. Recently, CNNs have been used for image-based person re-identification [12, 18, 20, 21, 23, 27, 29, 33]. These methods use deep network architecture such as Siamese network [4] to map images to feature vectors. These feature vectors can then be used for re-identification. Although the performance of image-based person re-identification has increased significantly, this is not a very realistic setting for practical applications.

To address the limitation of image-based re-identification, a lot of recent work has began to explore video-based re-identification [10, 13, 17, 22, 25, 26, 34, 35] since it is closer to real-world application settings. Compared with static images, videos contain temporal information that is potentially distinctive for differentiating a person's identity. Some prior work has explored ways of incorporating temporal information in deep convolutional neural network for re-identification. For example, McLaughlin *et al*. [17] use CNN on each frame in a video and incorporate a recurrent layer on the CNN features. Temporal pooling is then used to combine frame-level features into a single video-level feature vector for re-identification.

Our work is also related to a line of research on incorporating attention mechanism in deep neural networks. The attention mechanism allows the neural networks to focus on part of the input and ignore the irrelevant information. It has been successfully used in many applications, including machine translation [1], image captioning [24], visual question answering [19], etc. In video-based re-identification, the attention mechanism has also been explored [25, 34]. The intuition is that only a small portion of the video contains informative information for re-identification. So the attention mechanism can be used to help the model focus on the informative part of the video.

The work in [34] is the closest to ours. It uses an RNN to generate temporal attentions over frames, so that the model can focus on the most discriminative frames in a video for re-identification. In this paper, we uses temporal attentions over frames as well. But instead of using RNN-based models to generate attentions [34], we directly calculate the attention scores based on frame-based features. This makes the model much simpler and the computation of attention scores can be easily parallelized over frames. We also propose an attention refinement mechanism to iteratively refine the attention scores. We demonstrate that this attention refinement improves the performance of the final model.

3. Our Approach

Figure 2 shows the overall architecture of our proposed approach based on the Siamese network [4]. The input to the Siamese network is a pair of video sequences corresponding to the query video and the candidate video to be compared. The output of the Siamese network is a scalar value indicating how likely these two videos contain the

same person. Each video goes into one of the two branches of the Siamese network. Each branch of the Siamese network is a Convolutional neural network used to extract the features of the input video. The parameters of two branches of the Siamese network are shared. Finally, the features from the two input videos are compared to produce the final output.

When a video goes through one of the two branches of the Siamese network, we first extract per-frame features on each frame of the input video. Then we compute an attention score on each frame indicating how important this frame is for the re-identification task. The intuition is that not all frames in a video are informative. The attention scores enable our model to ignore certain frames and only pay attention to informative frames in the video. The attention scores are then used to aggregate per-frame visual features weighted by the corresponding attention score to form a feature vector for the entire video sequence. We also propose an iterative refinement mechanism that uses the feature vector of the video to further refine the attention scores. Here the intuition is that the initial attention score of a frame is computed purely based on the frame. It does not take into account of other frames in the video. Since the feature vector of the entire video encodes contextual information of the whole video sequence, we can use this feature vector to further refine the attention scores. We can repeat this process for several iterations (see Sec. 4.4), where each iteration produces attention scores that focus more on the informative frames. Finally, the features of two input videos are compared to produce the output.

3.1. Frame-Level Features

Similar to [17], we extract frame-level features using both RGB color and optical flow channels. The colors contain information about the appearance of a person, while the optical flows contain information about the movement of the person. Intuitively, both of them are useful to differentiate the identity of the person. As a preprocessing step, we convert all the input images (i.e. video frames) from RGB to YUV color space. We normalize each color channel to have a zero mean and unit variance. The Lucas-Kanade algorithm [14] is used to calculate both vertical and horizontal optical flow channels on each frame. We resize each frame to have a spatial dimension of 56×40. The optical flow field F of the frame is split into two scalar fields F_x and F_y corresponding to the x and y components of the optical flow. In the end, each frame is represented as a $56 \times 40 \times 5$ input, where the 5 channels correspond to 3 color channels (RGB) and 2 optical flow channels (x and y).

We fine-tuned CNN architecture of [17] to extract frame-level features for an input video. Note that we replace the fully connected in the end by two new fully connected layers that produce 1024 and 128 dimensional feature vectors respectively. Given an input video with T frames, we apply the CNN model on each frame of the input video. In the end, each frame $\mathbf{x}_i$ ($i = 1, 2, ..., T$) is represented as a 128 dimensional feature vector, i.e. $\mathbf{x}_i \in \mathbb{R}^{128}$.

3.2. Temporal Attention Network

Motivated by the recent success of attention based models [1, 2, 24, 28], we propose an attention based approach for re-identifying person from videos. The intuition behind the attention based approach is inspired by the human visual processing [25]. Human brains often pay attention to different regions of different sequences when trying to re-identify persons from videos. Based on this intuition, we propose a deep Siamese architecture where each branch generates attention scores of different frames based on the frame-level CNN features. The attention score of a frame indicates the importance of this frame for the re-identification task.

As shown in Figure 2, each input video sequence (sequence of frames with optical flow) is passed to the CNN to extract frame-level feature maps. Using fully connected layers, CNN generates feature vector for each video frame. The sequence of feature vectors are passed to the attention network to generate attention scores. More specifically, for each feature vector $\mathbf{x}_i$ corresponding to the i-th frame, we compute an attention score α_i indicating the importance of this frame. The attention score is obtained by applying a linear mapping followed by a sigmoid function. Here,we are use the same parameters for the linear mapping on all frames. Let θ be the vector of parameters for the linear mapping. Now the attention score α_i is calculated using the following equations:

$$z_i = \theta^T \mathbf{x}_i \tag{1a}$$

$$\alpha_i = \frac{1}{1 + \exp(-z_i)}, \quad \text{where } i = 1, 2, ..., T \tag{1b}$$

We have also tried using softmax instead of sigmoid function in Eq. 1 and found that it does not perform as good as the sigmoid function. Previous work [30] has made similar observations. Once we have obtained an attention score α_i for each frame in the video, we can then combine the attention scores α_i ($i = 1, 2, ..., T$) with frame-level feature vectors to create a weighted feature vector $\mathbf{f}$ as follows:

$$\mathbf{f} = \sum_{i=1}^{T} \alpha_i \mathbf{x}_i, \quad \text{where } i = 1, 2, ..., T \tag{2}$$

where $\mathbf{f}$ can be seen as a feature vector for the entire video which takes into account the importance of each frame in the video.

3.3. Attention Refinement

In principle, we can directly use the video-level feature vector in Eq. 2 for person re-identification, e.g. by com-

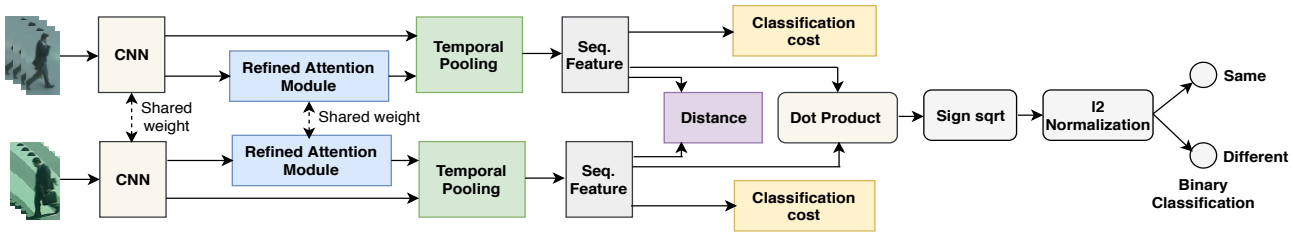

Figure 2. Overall architecture of our proposed Siamese network. It takes two input video sequences and pass to the Convolutional Neural Network (CNN) to extract features on each frame. The output from the CNN is fed to the attention module and generate an attention score for each frame. These attention scores combined with frame-level feature vectors to form a feature vector (i.e. temporal pooling) for the whole video. The video-level feature vectors are compared to decide whether the videos contain the same person.

paring the feature vectors of two videos. But one possible limitation is that the attention score in Eq. 1 is calculated on each frame in the video separately. In other words, the attention scores for frames in a video are independent of each other. This is not very intuitive – the attention score of a frame should depend on the visual information of the video, which in turn depends on all frames in the video. In this section, we introduce a strategy to refine the attention scores so that they are all coupled together in the end. In the experiment section, we will show that this attention refinement improves the performance of our model.

The basic idea of the attention refinement is to use the video-level feature vector $\mathbf{f}$ (Eq. 2) as one of the input to re-compute the attention score on each frame in the video. Since the video-level feature vector $\mathbf{f}$ depends on all frames in the video, the new attention score on a frame will implicitly depend on all frames in the video as well. The new attention scores can then be used to update the video-level feature vector. This process can be repeated for multiple iterations. Let us define α'_i as to be the new attention score. In this work, we simply concatenate $\mathbf{f}$ to each frame-level feature $\mathbf{x}_i$, then apply a linear mapping as follows:

$$z'_i = \theta'^T \text{concat}(\mathbf{x}_i, \mathbf{f}) \tag{3a}$$

$$\alpha'_i = \frac{1}{1+\exp(-z'_i)}, \quad \text{where } i = 1, 2, ..., T \tag{3b}$$

where concat$(\cdot)$ means the concatenation of two vectors. Then the new video-level feature vector $\mathbf{f}'$ can be computed as:

$$\mathbf{f}' = \sum_{i=1}^{T} \alpha'_i \mathbf{x}_i, \quad \text{where } i = 1, 2, ..., T \tag{4}$$

We alternate between updating attention scores (Eq. 3) and updating video-level feature vector (Eq. 4) for several iterations. Empirically, we have found 3 iterations give the best performance (see Sec. 4.4). Figure 3 shows the architecture of this attention refinement.

3.4. Model Learning

Our model is a form of the Siamese network (Figure 2). It has two identical branches with shared parameters. The detail architecture of each branch is shown in Figure 3. Each branch takes a video as its input and produces a feature vector of the video according to Eq. 4. Let $\mathbf{f}'_1$ and $\mathbf{f}'_2$ be the feature vectors of the two input videos to the Siamese network. We use Y_1 and Y_2 to denote the identity of the person in these two videos. Similar to [17, 25], we calculate Euclidean distance between these two feature vectors and use the following squared hinge loss(H_{loss}) as the loss function to train our network:

$$\mathcal{L}_{hinge} = \begin{cases} \frac{1}{2}\|\mathbf{f}'_1 - \mathbf{f}'_2\|^2, & Y_1 = Y_2 \\ \frac{1}{2}[\max(0, m - \|\mathbf{f}'_1 - \mathbf{f}'_2\|)]^2, & Y_1 \neq Y_2 \end{cases} \tag{5}$$

where m is a hyper-parameter that represents the margin of separating the two classes in $\mathcal{L}_{hinge}$. By minimizing this squared hinge loss, the distance between feature vectors will be small if the two videos contain the same person (i.e. $Y_1 = Y_2$). The distance will be large if the two videos contain two different persons (i.e. $Y_1 \neq Y_2$).

We also use a standard binary cross-entropy ($\mathcal{L}_{sim}$) that classifies the input videos to be same or different. For this, we firstly compute the inner product I of the video features and then perform a signed square-root step (i.e. $s \leftarrow sign(I)\sqrt{|I|}$). The resulting output is followed by a l_2 normalization ($N \leftarrow \frac{s}{\|s\|_2}$) and a softmax operation.

Following [17], we add an additional loss in each of the two branches of the Siamese network to predict the person's identity. Each branch uses the feature vector for the input video extracted from the network and applies a linear classifier to predict one of the K identities of the person. We use the softmax loss for the person identification classification. Let $\mathcal{L}_{id1}$ and $\mathcal{L}_{id2}$ be the loss functions of the two branches. The final loss function is the combination of the two identify classification losses, similarity loss and the squared hinge loss.

$$\mathcal{L}_{final} = \mathcal{L}_{id1} + \mathcal{L}_{hinge} + \mathcal{L}_{sim} + \mathcal{L}_{id2} \tag{6}$$

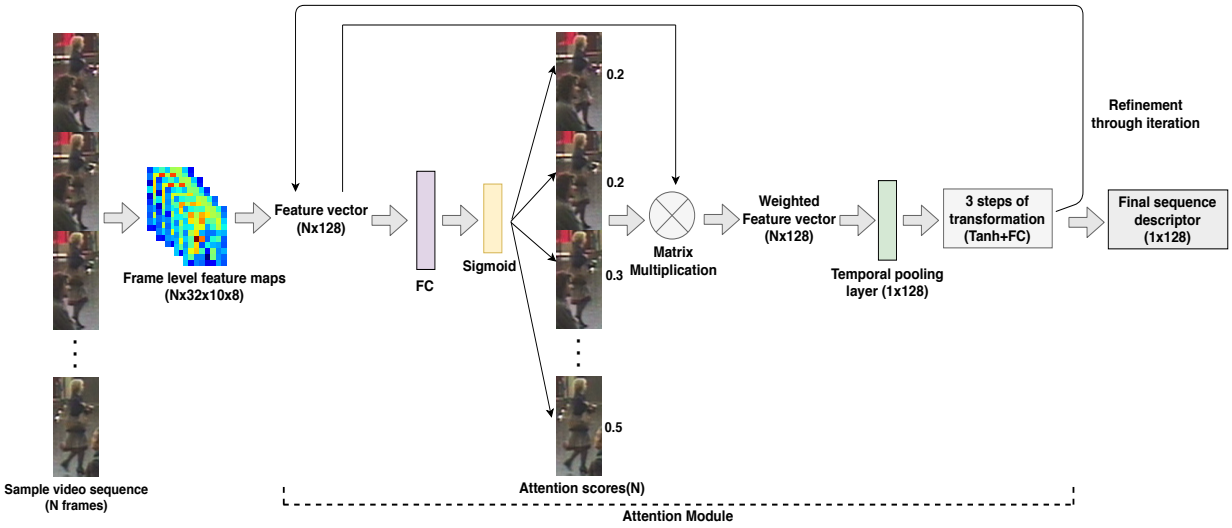

Figure 3. Illustration of our proposed refined attention network architecture. The input is a feature matrix of dimensions $N \times d$ where N is the number of frames in the sequence and d is the dimension of frame-level features. We generate N attention scores by applying linear mapping on the feature vectors followed by a sigmoid function. These attention scores are combined with frame-level features via temporal pooling to form a feature vector for the entire video. We use the video-level feature vector as one of the inputs to further refine the attention score on each frame. We then compute a new video-level feature vector using the new attention scores.

The network is trained end-to-end by optimizing the loss function in Eq. 6 using stochastic gradient descent. Following [17], we remove both classification losses, the squared hinge loss and similarity loss from the network after training is done. During testing, we only use the feature vectors generated by the two branches of the Siamese network and directly compare their distance for re-identification.

4. Experiments

In this section, we firstly introduce the datasets used in our experiments (Sec. 4.1). We then describe the experimental setup and some implementation details (Sec. 4.2). We present the results of experiment in Sec 4.3 and Sec 4.4.

4.1. Datasets

We conduct experiments on three benchmark datasets: iLIDS-VID [22], PRID-2011 [6] and MARS [32].

iLIDS-VID Dataset: This dataset consists of video sequences of 300 persons where each person is captured by a pair of non-overlapping cameras. The length of each video sequence varies from 23 to 192 frames with an average of 73 frames. The dataset is quite challenging due to lot of occlusions, illumination changes, background clutters and so on.

PRID-2011 Dataset: This dataset contains video sequences of 749 persons. For the first 200 persons (or identities), there are two video sequences captured by two different cameras. The remaining persons appear in only one camera. Each sequence contains between 5 to 675 frames, with an average of 100 frames. In terms of complexity this dataset is relatively simple than iLIDS-VID.

MARS Dataset: The Motion Analysis and Re-identification Set (MARS) is the largest video-based person re-identification dataset that contains 1,261 different pedestrians. Each pedestrian is captured by at least two cameras. DPM detector and GMMCP tracker are used to generate the tracklets. There are, on average, 13.2 tracklets for each pedestrian.

4.2. Setup and Implementation Details

We follow the experiment protocol of McLaughlin *et al.* [17]. On each of the two datasets (iLIDS-VID and PRID-2011), we randomly split the dataset into two equal subsets where one subset is used for training and remaining one for testing. For evaluating our proposed method, we use the Cumulative Matching Characteristics (CMC) curve which is a ranking based evaluation metric. In the ideal case, the ground-truth video sequence should have the highest rank. For each dataset, we repeat the experiment 10 times and report the average result over these 10 runs. In each run, we randomly split the dataset into training/test sets. Standard data augmentation techniques, such as cropping and mirroring, are applied to increase the amount of training data. We initialize the weights in the network using the initialization technique in [5]. For training our network, we consider equal numbers of positive and negative samples. We set the margin in the hinge loss (Eq. 5) as $m = 2$. The network

is trained for 1000 epochs with a batch size of one. The learning rate in the stochastic gradient descent is initially set to be $1e^{-3}$. We decrease the learning rate by a factor of 10 after 300 and 600 on the PRID-2011 dataset. Due to the variable-length of video sequences in both datasets, we use sub-sequences of 16 consecutive frames ($T = 16$) during training. Sometimes, this length is greater than the real sequence length. In that case, we consider the whole set of images (frames) as the sub-sequence. A full epoch consists of a pair of positive and negative sample. During testing, we consider a video sequence captured by the first camera as the probe sequence and a video sequence captured by the second camera as a gallery sequence. We use at most 128 frames in a testing video sequence. Again, if the length is greater than the real sequence, we consider the whole set of images as the video sequence. Similar strategies have been used in previous work [17]. For the MARS dataset, we follow the experimental protocol of state-of-the-art method by Xu *et al*. [25] which is different from [34].

4.3. Results

We present the results on the three benchmark datasets and compare with other state-of-the-art methods in Table 1, Table 2 and Table 3. From the CMC rank, we see that our method with attention refinement outperforms all other state-of-the-art methods by nearly 2% and 3% in terms of rank-1 accuracy on the iLIDS-VID and PRID-2011 dataset, respectively. On the MARS dataset, we outperform the state of the art by a big margin of 17% on rank-1 accuracy. Figure 5 shows some qualitative retrieval results after applying our proposed method on the challenging iLIDS-VID dataset. We also show some failure cases in Figure 4.

Dataset	iLIDS-VID			
Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**64**	**88**	**96**	**98**
Xu et al. [25]	62	86	94	98
Zhou et al. [34]	55.2	86.5	-	97.0
McLaughlin et al. [17]	58	84	91	96
Yan et al. [26]	49.3	76.8	85.3	90.1
STA [13]	44.3	71.7	83.7	91.7
VR [22]	35	57	68	78
SRID [7]	25	45	56	66
AFDA [9]	38	63	73	82

Table 1. Comparison of our proposed approach with other state-of-the-art methods on the iLIDS-VID dataset in terms of CMC(%) at different ranks.

4.4. Effect of Iterative Refinement

We conduct empirical study on the training set of the iLIDS-VID and MARS dataset to analyze the effect of the attention refinement (i.e. number of iterations) on the overall performance of the proposed network. We randomly divide the training dataset of iLIDS-VID into two parts: one for learning the model parameters and the other one for validation. We select 110 persons for training the model and the remaining 40 persons for validation. For MARS dataset, we select 400 identities for training and the remaining 225 identities for validation purpose. We train the model on the training videos and report the performance (CMC(%)) on the validation set for different number of iterations in Table 4 and Table 5 respectively. We observe that the performance gradually improves until iteration 3. After that, the performance starts to drop. Based on this empirical result, we choose 3 iterations in our experiments.

Dataset	PRID-2011			
Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**82**	**97**	**99**	99
Xu et al. [25]	77	95	99	99
Zhou et al. [34]	79.4	94.4	-	**99.3**
McLaughlin et al.[17]	70	90	95	97
Yan et al.[26]	58.2	85.8	93.7	98.4
STA [13]	64.1	87.3	89.9	92
VR[22]	42	65	78	89
SRID[7]	35	59	70	80
AFDA[9]	43	73	85	92

Table 2. Comparison of our proposed approach with other state-of-the-art methods on PRID-2011 dataset in terms of CMC(%) at different ranks.

Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**62**	**85**	**93**	**95**
[25]	44	70	74	81
[17] (obtained from [25])	40	64	70	77

Table 3. Comparison (CMC(%)) of our proposed approach with previous methods on the MARS dataset.

Figure 4. Examples of some failure case of our proposed method. The first row indicates the probe sequence where single image in second row represents retrieve gallery sequence of corresponding person.

5. Conclusion

In this paper, we have proposed an attention-based deep architecture for video-based re-identification. The attention module calculates frame-level attention scores, where the attention score indicates the importance of a particular

Figure 5. Qualitative retrieval results of our proposed method on the challenging iLIDS-VID dataset. The first column represents the probe video sequence. The remaining columns correspond to retrieved video sequences sorted by their distances to the probe video sequence. Here, we use a single image to represent each retrieved video sequence. The green boxes indicate the ground-truth matches. We can see that the ground-truth matches are ranked very high in the list.

iLIDS-VID				
# iterations	Rank-1	Rank-5	Rank-10	Rank-20
0 (No iteration)	60	92	97	100
1 (1 iteration)	70	95	97	100
2 (2 iterations)	62	95	97	100
3 (3 iterations)	**77**	**97**	**97**	**100**
4 (4 iterations)	70	97	97	100
5 (5 iterations)	65	97	97	100

Table 4. Validation performance for different number of iterations on the iLIDS-VID dataset. Again, we report the performance in terms of CMC (%).

MARS				
# iterations	Rank-1	Rank-5	Rank-10	Rank-20
0 (No iteration)	56	80	87	89
1 (1 iteration)	57	80	87	89
2 (2 iterations)	56	80	86	90
3 (3 iterations)	56	**80**	**88**	**90**
4 (4 iterations)	**58**	78	87	90
5 (5 iterations)	58	79	87	89

Table 5. Validation performance for different number of iterations on the MARS dataset.

frame. The output of the attention module can be used to produce a video-level feature vector which can be refined iteratively to generate rich feature information. We perform experiments on three benchmark datasets and compare with other state-of-the-art approaches. We demonstrate that our proposed method outperforms to other state-of-the-art approaches.

Acknowledgments: The authors acknowledge financial support from NSERC, MGS and UMGF funding. We also thank NVIDIA for donating some of the GPUs used in this work.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015. 1, 2, 3

[2] C. N. dos Santos, M. Tan, B. Xiang, and B. Zhou. Attentive pooling networks. *Computing Research Repository*, 2016. 3

[3] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008. 2

[4] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into recitifers: Surpassing human-level performance on imagenet classification. In *IEEE International Confernece on Computer Vision*, 2015. 5

[6] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011. 5

[7] S. Karanam, Y. Li, and R. J. Radke. Sparse re-id: Block sparsity for person re-identification. In *IEEE Conference on Computer Vision and Patten Recognition Workshop*, 2015. 6

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 2

[9] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *British Machine Vision Conference*, 2015. 6

[10] Y. Li, L. Zhuo, J. Li, J. Zhang, X. Liang, and Q. Tian. Video-based person re-identification by deep feature guided pooling. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 1, 2

[11] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[12] S. Liao and S. Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *IEEE International Conference on Computer Vision*, 2015. 1, 2

[13] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *IEEE International Conference on Computer Vision*, 2015. 1, 2, 6

[14] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *International joint conference on Artificial intelligence*, 1981. 3

[15] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision Workshop*, 2012. 2

[16] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[17] N. McLaughlin, J. M. del Rincon, and P. Miller. Recurrent convolutional neural network for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 3, 4, 5, 6

[18] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *IEEE International Conference on Computer Vision*, 2017. 1, 2

[19] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[20] E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, 2017. 1, 2

[21] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, 2016. 1, 2

[22] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, 2014. 1, 2, 5, 6

[23] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015. 1, 2, 3

[25] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 3, 4, 6

[26] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, 2016. 1, 2, 6

[27] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for person re-identification. In *IAPR International Conference on Pattern Recognition*, 2014. 1, 2

[28] W. Yin, H. Schutze, B. Xiang, and B. Zhou. ABCNN: Attention-based convolutional neural networks for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 2016. 3

[29] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[30] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[31] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[32] 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 *European Conference on Computer Vision*. Springer, 2016. 5

[33] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Salable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015. 1, 2

[34] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 6

[35] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *International Joint Conference on Artificial Intelligence*, 2016. 1, 2