Journal of Electronic Imaging

JElectronicImaging.org

CARF-Net: CNN attention and RNN fusion network for video-based person reidentification

Kajal Kansal
Subramanyam Venkata
Dilip K. Prasad
Mohan Kankanhalli

Kajal Kansal, Subramanyam Venkata, Dilip K. Prasad, Mohan Kankanhalli, "CARF-Net: CNN attention and RNN fusion network for video-based person reidentification," *J. Electron. Imaging* **28**(2), 023036 (2019), doi: 10.1117/1.JEI.28.2.023036.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

CARF-Net: CNN attention and RNN fusion network for video-based person reidentification

Kajal Kansal,[a,*] Subramanyam Venkata,[a] Dilip K. Prasad,[b] and Mohan Kankanhalli[c]
[a]Indraprastha Institute of Information Technology, Department of Computer Science and Engineering, Delhi, India
[b]Nanyang Technological University, Department of Computer Science and Engineering, Singapore, Singapore
[c]National University of Singapore, Department of School of Computing, Singapore

Abstract. Video-based person reidentification is a challenging and important task in surveillance-based applications. Toward this, several shallow and deep networks have been proposed. However, the performance of existing shallow networks does not generalize well on large datasets. To improve the generalization ability, we propose a shallow end-to-end network which incorporates two stream convolutional neural networks, discriminative visual attention and recurrent neural network with triplet and softmax loss to learn the spatiotemporal fusion features. To effectively use both spatial and temporal information, we apply spatial, temporal, and spatiotemporal pooling. In addition, we contribute a large dataset of airborne videos for person reidentification, named DJI01. It includes various challenging conditions, such as occlusion, illuminationchanges, people with similar clothes, and the same people on different days. We perform elaborate qualitative and quantitative analyses to demonstrate the robust performance of the proposed model. © *2019 SPIE and IS&T* [DOI: 10.1117/1.JEI.28.2.023036]

Keywords: attentions; convolutional neural network–recurrent neural network; person reidentification; surveillance.

Paper 181121 received Dec. 27, 2018; accepted for publication Apr. 2, 2019; published online Apr. 25, 2019.

1 Introduction

Person reidentification is a critical task in visual surveillance. For tasks such as threat detection, long-term tracking, and activity analysis, it is desirable to determine if an individual has already been observed in other cameras in the network. In particular, video-based person reidentification has attracted more attention recently.[1–4] However, to match a person across different views, there are several challenges, including large appearance changes caused by different lighting, low resolution, pose variation, occlusion, and background clutter.

Person reidentification has been investigated using both images and videos. Most of the early works are based on images. Some of these techniques focus on extracting features that are both discriminative and invariant to various challenges.[5–12] However, it is challenging to design features that are discriminative enough to distinguish people reliably and at the same time invariant to factors such as motion blur, view angle, pose change, and other factors. To this end, the use of video is a desirable approach to improve the performance of reidentification.[1] Video can naturally be attributed to spatial and temporal cues. The spatial part carries information about scenes and appearances of the person, such as color of the clothes, person's height and shape, whereas the temporal part encodes the walk pattern of the person, which is complementary to the spatial part. Thus, extracting these features from the videos can lead to better results.

Several approaches exploit spatiotemporal information for video-based person reidentification.[13–16] In Ref. 13, the authors use RGB frames and optical flow between consecutive frames to construct the spatiotemporal features. However, its single stream convolutional neural network (CNN) takes partial advantage of rich temporal information. In other tasks such as action recognition, two-stream CNN architecture has been shown to give better performance.[17] The video is split into two individual streams, i.e., sequence of still images and optical flow vectors, to separately learn better representative features. The information from the two streams is then fused at certain intermediate layers. Shallow models such as recurrent convolutional network (RCN)[13] and attentive spatial-temporal pooling networks (ASTPN)[18] perform well on small datasets, such as PRID-2011[19] and iLIDS-VID.[20] However, the performance does not generalize to large datasets such as MARS.[1] Dictionary learning techniques, such as those shown in Refs. 11 and 12, have also been exploited, which can be extended to video-based reidentification.

Person reidentification has also been investigated for both static and moving camera scenarios. A good survey of static camera-based person reidentification works can be found in Refs. 2, 3, and 21. In this paper, we focus on both static and moving camera-based person reidentification. Moving camera-based reidentification works can be found in Refs. 22–25. Schumann and Schuchert[22] proposed to use color and texture features to recognize individual persons in aerial video data and features are weighted based on their correlation to operator feedback in order to find possible matches to a query person track. However, the dataset captured with a moving platform requires view-specific discriminative training for obtaining good performance. This is because of the view variation in the dataset, where view variation is defined as the continuously varying view angle of the camera compared to the traditional static cameras.[25] The difficulty in collecting large view-specific data limits the learning ability of a model. In Ref. 25, the authors introduce a mobile platform database and test with support vector machine and metric learning techniques. However, it is captured at a very low altitude and does not have large variance per identity. In Refs. 23 and 24, authors exploit person reidentification in aerial images.

*Address all correspondence to Kajal Kansal, E-mail: kajalk@iiitd.ac.in

1017-9909/2019/$25.00 © 2019 SPIE and IS&T

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

In this paper, we propose a network for video-based person reidentification, which can account for learning representations under various challenges. The major contributions are as follows:

- We demonstrate that a shallow network can be trained in a manner such that it can attain accuracy comparable to deep networks. To this end, we learn rich representations from multiscale visually attentive regions, we use CNN and spatial pyramid pooling (SPP)[26] network. As discriminative frames may appear anywhere in the sequence, we exploit the spatiotemporal fusion of features by adding spatially pooled features of CNN with temporally pooled features of recurrent neural network (RNN). To train the network, we use a multiloss function. Our model requires only ∼15% of the parameters compared to the state-of-the-art deep models, such as LuNet.[27]
- Our paper contributes a large dataset of airborne videos for mobile platform video-based person reidentification. Our dataset is named DJI01. It is taken under various challenging conditions, such as view variation, occlusion, illumination changes, people with similar clothes, same people at different days, scale variation, and pose variation. To our knowledge, DJI01 is the largest airborne person reidentification video dataset to date.
- We conduct extensive experiments on DJI01, and publicly available datasets, such as MARS,[1] PRID-2011,[19] and iLIDS-VID[20] to demonstrate the state-of-the-art performance achieved by our method for video-based person reidentification. We also perform rigorous ablation study and qualitative analysis to show the effect of various network components. In addition, we also study the performance against noisy examples.

The rest of the paper is structured as follows. In Sec. 2, we briefly review related works in video-based person reidentification. In Sec. 3, we present our proposed approach, and analyze each component of the network. Section 4 gives information of the new dataset introduced in this work. Section 5 presents an extensive comparison with the state-of-the-art algorithms. Section 6 concludes the paper and discusses the future work.

2 Related Work

Here we discuss the state-of-the-art algorithms on video-based person re-identification. Over the past decade, deep learning methods[28–35] have shown a significant improvement over handcrafted features. They encode reliable features and corresponding similarity value for a pair or triplet of images or videos. Recently, deep learning architectures based on RNN[13] and long short-term memory (LSTM)[14–16] models have been explored. In Ref. 13, McLaughlin et al. focus on color appearance and optical flow, where the network jointly learns feature representation and similarity metric. RNN efficiently encodes the temporal information, but there is a limitation with the learning of long duration sequences of the inputs.[13] Graves[14] used an LSTM model for learning long duration dependencies through the use of memory cell units. Varior et al.[15] proposed a Siamese LSTM architecture that can process image regions sequentially and enhance the discriminative capability of local feature representation by leveraging contextual information. Feedback connections and internal gating mechanism of the LSTM cells enable the model to memorize the spatial dependencies and selectively propagate relevant contextual information through the network. Wu et al.[16] proposed a framework which combines time series modeling and metric learning to jointly learn relevant features and similarity measures between time sequences of person. However, LSTM requires a large training dataset due to the large number of parameters, for achieving good generalization. Yan et al.[36] investigated a recurrent feature aggregation (RFA) network with LSTM Layer. This network uses a set of handcrafted frame-level features extracted from each frame in the input sequence and produces sequence-level features. However, these features may exhibit limited discriminative nature. Dai et al.[37] proposed a temporal residual learning module which is equipped with two bidirectional LSTMs to simultaneously learn the generic and specific features of a video sequence. ASTPN[18] takes the advantage of the attention mechanism to extract features from informative frames. Zheng et al.[1] used Alexnet to extract features and metric learning to compute the similarity. Zhou et al.[38] proposed the temporal attention model (TAM) to focus on the discriminative frames. The TAM is jointly learned with the spatial recurrent model to integrate the surrounding information at different spatial locations for better similarity evaluation. Zhong et al.[39] proposed a reranking method in which given an input, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors into a single vector, which is used for reranking under the Jaccard distance. In Ref. 40, the authors use a two-stream Siamese network to learn spatiotemporal features separately for person reidentification. Yu et al.[41] explored different streams to learn different aspects of feature maps for attentive spatiotemporal fusion of video, and then merge them together to study some union features. Li et al.[42] proposed a deep context-aware feature (DCF) model, in which a multiscale context-aware network (MSCAN) is designed to learn the powerful features over full body and body parts, which can well capture the local context knowledge by stacking multiscale convolutions in each layer.[43] Li et. al.[43] propose spatial aligned temporal pyramid pooling (SATPP) model to leverage the rich visual-temporal cues for feature learning.

In addition to exploration of different features and architectures, various loss functions have also been investigated. Identification loss and verification loss are two types of losses used for training various reidentification models. Identification loss[1,44] limits the reidentification performance due to the avoidance of intraclass variance. Here, verification loss such as Siamese,[13] triplet loss,[27,45] and quadraplet[46] are found to be more stable for reidentification tasks. Other loss functions such as support neighbor loss[47] and center loss[48] have also shown effective performance. In Ref. 27, the authors highlight the fact that triplet loss-based training for pretrained models and for models trained from scratch can achieve state-of-the-art performance. The two models, TriNet and LuNet, are derived from pretrained ResNet-50 and ResNet-v2 architecture, respectively. Authors highlight that, due to the efficacy of triplet loss, even the LuNet model which is trained from scratch achieves a comparable performance to TriNet model. However, we show that a robust

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

performance comparable to the state of the art can be achieved even with a shallow network. Unlike in Ref. 27, we use a very shallow network of three layers followed by an attentive network and a RNN layer. In addition, we observe that the multiloss function can be utilized to enhance the reidentification performance.

We observe that the shallow networks work well in case of smaller datasets, such as PRID-2011 and iLIDS-VID. However, in the case of larger dataset, such as MARS, there is a large drop in accuracy. To deal with larger datasets, one can use deep pretrained models. Algorithms in Refs. 1, 42, and 27 use AlexNet, GoogleNet, and ResNet and achieve the state-of-the-art performance on MARS dataset. In addition, Hermans et al.[27] used LuNet model, which is trained from scratch, and show comparable accuracy with the use of triplet loss. However, LuNet uses 5M number of parameters, whereas our shallow network uses only 0.75M parameters with triplet loss and achieves comparable accuracy to that of LuNet.

Recently, attention models have shown good performance in various tasks. Haque et al.[49] exploited an attention model for the depth data to learn a specific local region. In this paper, we use a similar attention mechanism as in Ref. 26, where an attentive component to select discriminative regions is used from each frame. Our work is closest in spirit to ASTPN.[18] However, there are several differences between our model and ASTPN. First, we use two streams compared to a single stream, as shown in Ref. 18, which makes use of both RGB and optical flow data. As the two inputs are very different in nature, it is better to use respective streams and then fuse the features obtained from them. It has been shown that two streams give better performance compared to a single stream.[50,17] Second, we use a triplet network, whereas a Siamese network is used in Ref. 18, and our training objective is different as we incorporate triplet loss. Although there has been no concrete analysis between the performance of the Siamese loss or the performance of the triplet loss, in Ref. 27, it has been shown that a triplet loss performs better in the case of person reidentification even when the network is trained from scratch. Third, the spatial pooling and spatio-temporal fusion are introduced in the proposed network to effectively learn robust features. CNN features at lower layers have good localization details.[51] As the video frames always comprise persons in upright position, having spatial features would give the benefit as features would correspond to the respective spatial locations. In the case of feature vectors from deeper layers, the localization aspect is lost as it encodes only semantics. Thus, pooling of spatial and temporal features would result in a robust feature representation. In Sec. 5.3.7, we also show how each component qualitatively effects the performance of the network. We describe the proposed network in the following section.

3 Proposed Algorithm

In the proposed two-stream triplet network, the input consists of color channels and optical flow vectors, where optical flow is computed using the Lucas–Kanade algorithm.[52] Color channels encode details of a person's appearance and clothing, whereas optical flow encodes the temporal information. We first use a CNN module to extract the spatial information. CNN is followed by a spatial pyramid network which exploits visually attentive regions in the inputs. To encode the temporal information, we use RNNs. In addition, we perform the fusion of spatial features with recurrent layer features across time steps to form a discriminative video-level representation. We train the network using triplet softmax loss. This approach preserves the spatial information along with temporal information and also learns the most contributive regions toward reidentification. A diagram of our proposed architecture is shown in Fig. 1.

3.1 *Spatial Pyramid Pooling–Convolutional Neural Network*

The two streams in the architecture use spatial network (CNN) for learning feature representations from raw frames and optical flow, respectively. Let an input video sequence be denoted as s_t, $\forall\ t = 1,2,\ldots,T$, where T is the length of the sequence. Each layer of the convolutional network performs the following operation:

$$C_{l+1} = \tan h\{\text{maxpool}[\text{conv}(C_l)]\}, \tag{1}$$

where l denotes a layer and C_0 denotes the raw video frame or optical flow, and in deeper layers the input is the output feature map from the previous layer of the CNN. To avoid complexity and high computation, we use an efficient spatial network, as in Ref. 13. The spatial network (CNN) contains three convolutional layers and two max-pooling layers with a nonlinear tan h layer. Further, we use SPP layer,[26] which exploits the local spatial regions from each raw video frame or optical flow. This is necessary as only subregions of the entire frame contribute toward reidentification. The SPP generates multilevel spatial representations, which exploits the information from various scales. The layer comprises sizes 8×8, 4×4, 2×2, and 1×1. Both the streams follow the same architecture. Let the feature maps set be $F = \{F_1, F_2, \ldots, F_T\}$, obtained from the convolutional network. Each $F_i \in R^{c\times w\times h}$ is then fed into spatial pooling layer to get an image-level representation I_i, where c is the number of kernels and, w and h are the kernels' width and height, respectively. The SPP layer has spatial bins to generate multilevel spatial representations. The dimensions of spatial bins are $(B_w^j, B_h^j | j = 1,2,\ldots,n)$, the window size $\text{win}_j = (\lceil \frac{w}{B_w^j} \rceil, \lceil \frac{h}{B_h^j} \rceil)$ and pooling stride $\text{str}_j = (\lfloor \frac{w}{B_w^j} \rfloor, \lfloor \frac{w}{B_h^j} \rfloor)$ for j'th spatial bin. Finally, the resultant vector I_i is as follows:

$$v_{i,j} = f_{\text{reshape}}[f_{\text{pool}}(F_i; \text{win}_j, \text{str}_j)], \tag{2}$$

$$I_i = v_{i,1} \oplus v_{i,2} \oplus \ldots \oplus v_{i,n}, \tag{3}$$

where f_{pool} denotes the max pooling with window size win and stride str, f_{reshape} represents the reshape operation to reshape matrix into a vector, and $\oplus$ denotes the vector connection operation. Multilevel spatial representations generated from spatial bins are then combined into a fixed-length image-level representation. These representations involve the exact position of a person and multiscale spatial information. We use a four-level pyramid pooling for local spatial attention after the last convolution layer. The output of SPP layer is fed to the first fully connected (FC) layer. We use three FC layers. The output of the third FC layer is a 128-dimensional feature vector. Similarly, a feature vector is generated from the second stream. We fuse the two feature vectors from two streams to obtain a single representation.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

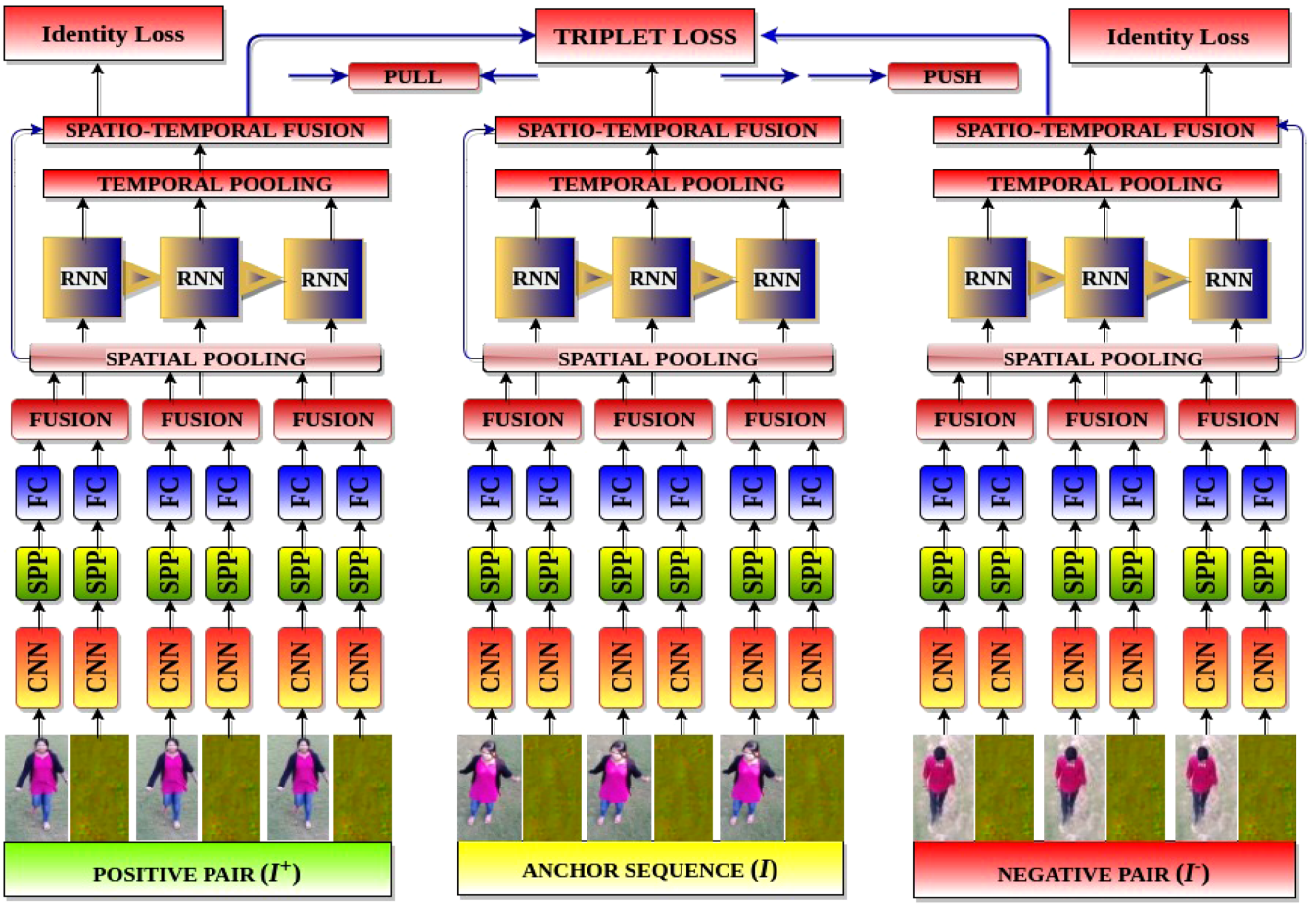

Fig. 1 Proposed architecture of CARF-Net. The input to the two-stream spatial network (CNN) comprises color image and optical flow. SPP refers to the spatial pyramid pooling. FC denotes three FC layers. Outputs of the last FC layer from the two streams are fused in the fusion layer and the resultant feature is given as an input to RNN. Simultaneously, the resultant features are also pooled together in the spatial pooling layer. The output of RNN layer is pooled in the temporal pooling layer. Further, outputs of the spatial pooling and the temporal pooling layer are fused in the spatiotemporal fusion layer.

We explain the fusion operation in Sec.3.3. The fusion is represented by the component "FUSION" in Fig. 1. Further, let $r = r_t \in R^{128\times1} | t = 1,2,\ldots,T$ be a sequential representation of the fused features from the two streams. To encode temporal information, we then pass r to the RNNs, which we explain in the next subsection.

3.2 *Recurrent Neural Network*

Given an input sequence r_t, $\forall\ t = 1,2,\ldots,T$, the output of the RNN is o_t which is the final feature vector obtained at time t. The RNN equation is given as

$$o_t = W_r r_t + W_h h_{t-1}, \tag{4}$$

$$h_t = \tan\ h(o_t), \tag{5}$$

where $o_t \in R^{q\times1}$ is the q-dimensional output of RNN at timestep t; $h_{t-1} \in R^{q\times1}$ contains the information on the RNN's state at the previous timestep, W_r and W_h are the respective weights for r_t and h_{t-1}. We add a temporal pooling layer after RNN to capture long-term information present in the whole sequence. Although RNNs are able to encode the temporal information, they are biased to current information.[13] However, discriminative frames may appear anywhere in the sequence. To preserve the information from discriminative frames, we apply the temporal pooling. We use an average-pooling over the temporal dimension to produce a single feature vector, which we explain in the next subsection.

3.3 *Fusion*

Here, we discuss fusion, spatial pooling, temporal pooling, and spatiotemporal fusion layers present in the architecture. Let f_{FCS1_t} and f_{FCS2_t} be the respective outputs of the FC layers from streams 1 and 2 at a given time instant t, respectively. Then the fusion of these features is given as

$$f_{\mathrm{FUS}_t} = \frac{1}{2}(f_{\mathrm{FCS1}_t} + f_{\mathrm{FCS2}_t}), \tag{6}$$

where f_{FUS_t} is the same as r_t.

Spatial pooling: The inputs to spatial pooling layer are f_{FUS_t}. The pooling equation is given as

$$f_{\mathrm{SP}} = \frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{FUS}_t}, \tag{7}$$

where T is the length of the sequence or timesteps.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Temporal pooling: Let f_{Temp} denote the motion information averaged over the whole input sequence, then

$$f_{\text{Temp}} = \frac{1}{T}\sum_{t=1}^{T} o_t. \tag{8}$$

Spatiotemporal fusion: We observe that the spatial information along with the temporal information can be a better representative for reidentification. Toward this, we pool the spatial and temporal features and this operation is given as[53]

$$F_{\text{SPT}} = f_{\text{SP}} + f_{\text{Temp}}. \tag{9}$$

In our experiments we find that there is 2% to 4% increase in the performance due to spatiotemporal fusion. In addition, we also tried max pooling and concatenation; however, we did not find any significant improvement compared to average pooling.

3.4 *Training Objective*

The triplet network consists of three subnetworks with identical weights. To train the triplet network to perform reidentification, our objective is to show similar and dissimilar input pairs, and learn to map those inputs to a feature space where similar inputs are close and dissimilar inputs are separated by a margin. We use a multiloss function, including triplet loss[54] and identity loss. Such combination of loss functions has been shown to perform well in verification tasks.[13,18,53] Let the positive samples be I_n^+ and the anchor sequence be I_n. Here, I_n can be represented by a feature; in our case, we use F_{SPT}. Let α be the margin between the positive and negative pairs to enhance the discriminative ability of learned features. Thus, we have

$$\|I_n - I_n^+\|_2^2 + \alpha < \|I_n - I_n^-\|_2^2. \tag{10}$$

Inspired from Ref. 55, we take all positive pairs and randomly sample negative examples. It is observed in Ref. 55 that using all positive pairs makes the model more stable and converge faster than selective sampling in a mini-batch. The proposed network can be trained end to end by using backpropagation, and the loss function for N triplets is given as

$$L_{\text{triplet}} = \frac{1}{N}\sum_{n=1}^{N}[\|I_n - I_n^+\|_2^2 - \|I_n - I_n^-\|_2^2 + \alpha]_+, \tag{11}$$

where $[.]_+$ is the hinge function.

Given the sequence feature vector I, we can determine the identity of the person in the sequence using the standard softmax function. Let Γ be the total number of identities, z is the predicted identity for the input person, and $S \in R^{M\times\Gamma}$ is the weight matrix used in the softmax function. Here, $S_c \in R^M$ and $S_\Gamma \in R^M$ denote the c'th and Γ'th column of the softmax weight matrix S, respectively. The softmax function is as follows:

$$L_{\text{softmax}} = \beta(I) = P(z = c|I) = \frac{\exp(S_c^T I)}{\sum_\Gamma \exp(S_\Gamma^T I)}. \tag{12}$$

To jointly train the network with both triplet loss and identification loss, the overall multiloss training function is as follows:

$$L_{\text{multi}} = L_{\text{triplet}} + L_{\text{softmax}}. \tag{13}$$

4 DJI01 Dataset

We introduce the DJI01 dataset for video-based person reidentification. (We plan to release the dataset upon acceptance of the paper.) It includes aerial videos captured through two drones in various challenging situations. It contains 200 identities, with video sequence length varying from 32 to 3000 frames. We show a comparison of this dataset with the existing datasets in Table 1.

4.1 *Data Acquisition and Processing*

Data collection is done using DJI Phantom 3 and DJI Phantom 4 Quadcopters. The dataset is captured in an outdoor environment with different backgrounds, altitudes ranging from 3 to 35 m, frame rate of 24 or 60 fps, and resolution of 1280×720. GMMCP tracker[56] is used for segmenting the person of interest from the video.

4.2 *Challenges*

In our new dataset, we collect rich information with large appearance variation for every single person. The camera's motion and orientation lead to a change of viewpoint, which adds challenges during reidentification. Further, the camera motion can also cause blurring effect. Thus, the performance of the algorithm may degrade in videos captured using drones. We show different examples of DJI01 in Fig. 2. It shows challenges of scale variation, pose variation, view variation, different altitude, occlusion, camera motion, and illumination variation. In addition, there are the same 10 people captured on different days with all the above-mentioned challenges. Moreover, the 10 subjects illustrate the challenge of people with very similar clothes.

4.3 *Evaluation Protocol*

Several approaches have been used for evaluating reidentification performance. We use the cumulative match characteristic (CMC) Top 1-5-10-20 accuracy to evaluate the performance of person reidentification. All CMC accuracies are given in percentage. The evaluation methodology used here is as follows. The set is split evenly into a training

Table 1 Comparison of DJI01 parameters with existing datasets (IDs denote the number of individual identities).

Datasets	IDs	Type	Cameras
PRID-2011[19]	200	Static	2
iLIDS-VID[20]	300	Static	2
MARS[1]	1261	Static	6
MRP[25]	84	Drone	1
DJI01	200	Drone	2

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Fig. 2 Various subjects showing different challenges in DJI01. First row: First subject refers to the illumination variation, low resolution, and clutters; second subject represents the scale variations and view variability; third subject shows the occlusion and illumination changes; and the fourth one shows the illumination changes and camera motion. Second row: First subject shows the scale variation, pose variation, and blurring due to camera motion; second subject shows the occlusion, altitude, scale variation, and pose variation; third subject shows the illumination changes, and the fourth one shows the pose and view variations.

and a test set. The CMC Top for the test set is calculated by selecting a probe video and matching it with a gallery video. This provides ranking for every video in the gallery with respect to the probe. This procedure is repeated for every probe video. The CMC Top is then the expectation of finding the correct match in the top n matches.

5 Experiments

We present a comprehensive evaluation of our approach by comparing it with state-of-the-art methods.[1,13,18,30,36,40,57–60] In Ref. 40, the authors use two-stream CNN, where each stream is a Siamese network, to learn the spatiotemporal features separately. ASTPN[18] takes the advantage of attention mechanism to extract the features from informative frames. In Ref. 1, the authors use Alexnet to extract the features and metric learning to compute the similarity[30] learns an intravideo and intervideo distance metric from the training videos. In Ref. 13, a RCN is used with temporal pooling. Spatiotemporal appearance (STA)[57] builds a spatiotemporal appearance representation for person reidentification. The RFA network[36] is based on LSTM.[58] use three-dimensional histogram of oriented gradients, color, and local binary pattern features, and learn a distance metric for matching. Adaptive fisher discriminant analysis[59] uses the representative data samples to learn a feature subspace maximizing the Fisher criterion.[60] learn a single dictionary to represent both gallery and probe images in the training phase. Three-stream convolution network SATPP (TSCN)[41] uses different streams to learn the different aspects of feature maps for attentive spatiotemporal fusion of video and then merges them together to study some union features. In DCF,[42] a MSCAN is designed to learn the powerful features over full body and body parts, which can well capture the local context knowledge by stacking multiscale convolutions in each layer[43] propose SATPP model to leverage the rich visual-temporal cues for feature learning.

5.1 *Datasets*

A brief description about PRID-2011,[19] iLIDS-VID,[20] and MARS[1] datasets is given in Table 1. PRID-2011 is captured in an uncrowded outdoor environment with stark difference in illumination, background clutter, and less occlusions. It contains 200 identities, with video sequence length varying from 5 to 675 frames. The iLIDS-VID dataset is more challenging due to occlusions, illumination changes, and viewpoint variations. The video sequence length varies from 23 to 192 frames. MARS is a much larger dataset when compared to others and sequence length varies from 32 to 20,000 frames.

Table 2 Comparison of our proposed approach (CARF-Net) with the state-of-the-art on PRID-2011.[19]

CMC Top@T	T-1	T-5	T-10	T-20
TSS-CNN[40]	78	94	97	99
ASTPN[18]	77	95	99	99
IDE[1]	77.3	93.5	—	99.3
SI^{2}DL[30]	76.7	95.6	96.7	98.9
RCN[13]	70	90	95	97
STA[57]	64	87	90	92
RFA[36]	58.2	85.8	93.4	97.9
TDL[58]	56.74	80	87.64	93.54
DTDL[60]	40.6	69.7	77.8	85.6
LBTC[61]	72.80	92.00	95.10	97.60
TLST[44]	73.07	98.53	99.41	99.41
SCPDL[31]	74.50	92.10	94.30	96.60
DSAN[62]	77.00	96.40	99.20	99.40
UTRCNN[63]	73	92.70	95	98
SLD^{2}L[32]	22.60	46.60	57.40	70.70
PHDL[34]	41.92	67.25	85.47	92.44
MRG[35]	78.4	94.8	97.9	99.4
CARF-NET	**79**	96	98	99

Note: Bold face represents the best accuracy.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

5.2 *Experimental Setup*

Our architecture is implemented using Torch. All the experiments are performed on Nvidia GTX 1080 GPU. It takes ~20 h for training with 1000 epochs. The values of the hyper-parameters are empirically set. We use the following values: learning rate = 0.001, momentum = 0.9, dropout ratio = 0.6, and the feature embedding-space dimension is 128. The dropout is applied on the last two FC layers. The value of α is varied between 0.2 to 0.5 and the best results are reported. For these experiments, each dataset is randomly split into 50% for training and 50% for testing. All experiments are repeated 10 times with different train/test set to ensure stable results. For the MARS dataset, we use the provided fixed training and test sets, containing 631 and 630 identities, respectively.

5.3 *Performance Comparison*

5.3.1 *PRID, iLIDS, DJI01, and MARS results*

We report the PRID-2011 results in terms of CMC top-1-5-10-20 accuracies in Table 2. With PRID-2011, we achieve 79% top-1 accuracy, 96% top-5 accuracy, 98% top-10 accuracy, and 99% top-20 accuracy. We observe that the proposed algorithm gives better results when compared to several popular algorithms in most of the cases.

With iLIDS-VID, top-1 accuracy is 65%, top-5 accuracy is 87%, top-10 accuracy is 93%, and top-20 accuracy is 98%. Here again, we observe that the performance of the proposed algorithm is significantly better when compared to the other schemes. The CMC top-1–5–10–20 accuracies of iLIDS-VID are reported in Table 3.

The results on MARS dataset are reported in Table 4. We observe that the performance of the proposed algorithm significantly outperforms the other existing algorithms in both top-1 and top-10 accuracies.

The results of DJI01 are reported in Table 5. A top-1 accuracy of 64% is obtained. We also test with other algorithms.[1,13,18] We can conclude that our network outperforms other networks by a significant margin for both top-1 and top-5 accuracy.

The average Top-1 accuracy of our algorithm across all four datasets shows a significant improvement of 6.1%,

Table 3 Comparison of our proposed approach (CARF-Net) with the state-of-the-art on iLIDS-VID.[20]

CMC Top@T	T-1	T-5	T-10	T-20
ASTPN[18]	62	86	**94**	**98**
TSS-CNN[40]	60	86	93	97
SATPP[43]	56.67	78.67	90.00	96.67
IDE[1]	53.0	81.4	–	95.1
SI^{2}DL[30]	48.7	81.1	89.2	97.3
RCN[13]	58	84	91	96
STA[57]	44	72	84	92
RFA[36]	49.3	76.8	85.3	90.0
TDL[58]	56.33	**87.60**	91	96
DTDL[60]	25.9	48.2	57.3	68.9
LBTC[61]	55.30	85.00	91.70	95.10
TLST[44]	59.20	86.06	99.00	99.00
SCPDL[31]	56.80	86.30	94.20	96.60
DSAN[62]	61.20	80.70	90.30	97.30
UTRCNN[63]	62.70	86	93.60	98
VSDL[64]	59.40	89.10	96.20	98.60
PHDL[34]	28.15	50.37	65.88	80.35
TCN[33]	60.6	83.8	91.2	95.8
MRG[35]	60.80	89.20	97.20	99.50
CARF-Net	**65**	87	93	**98**

Table 4 Results of MARS[1] dataset. RK: rerank.

CMC Top@T	T-1	T-5	T-10	T-20
DCF[42]	71.77	86.57	—	93.08
IDE(R) + (RK)[39]	70.51	—	—	—
SFT[38]	70	**90**	—	97
SATPP[43]	69.69	84.65	89.34	92.77
IDE[1]	65.0	81.1	—	88.9
TSCN[41]	45.6	72.4	75.4	82.6
ASTPN[18]	44	70	74	81
RCN[13]	40	64	70	77
TLST[44]	61.66	82.63	88.33	88.42
DSAN[62]	69.70	83.40	88.30	96.60
RQEN[65]	73.74	84.90	—	91.62
PHDL[34]	35.72	51.49	60.88	67.28
CARF-Net	**74**	83	**92**	**99**

Table 5 Results of our drone (DJI01) dataset.

CMC Top@T	T-1	T-5	T-10	T-20
ASTPN[18]	63	70	**85**	**98**
IDE[1]	59	65	82	84
RCN[13]	55	62	83	92
CARF-Net	**64**	**74**	83	95

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

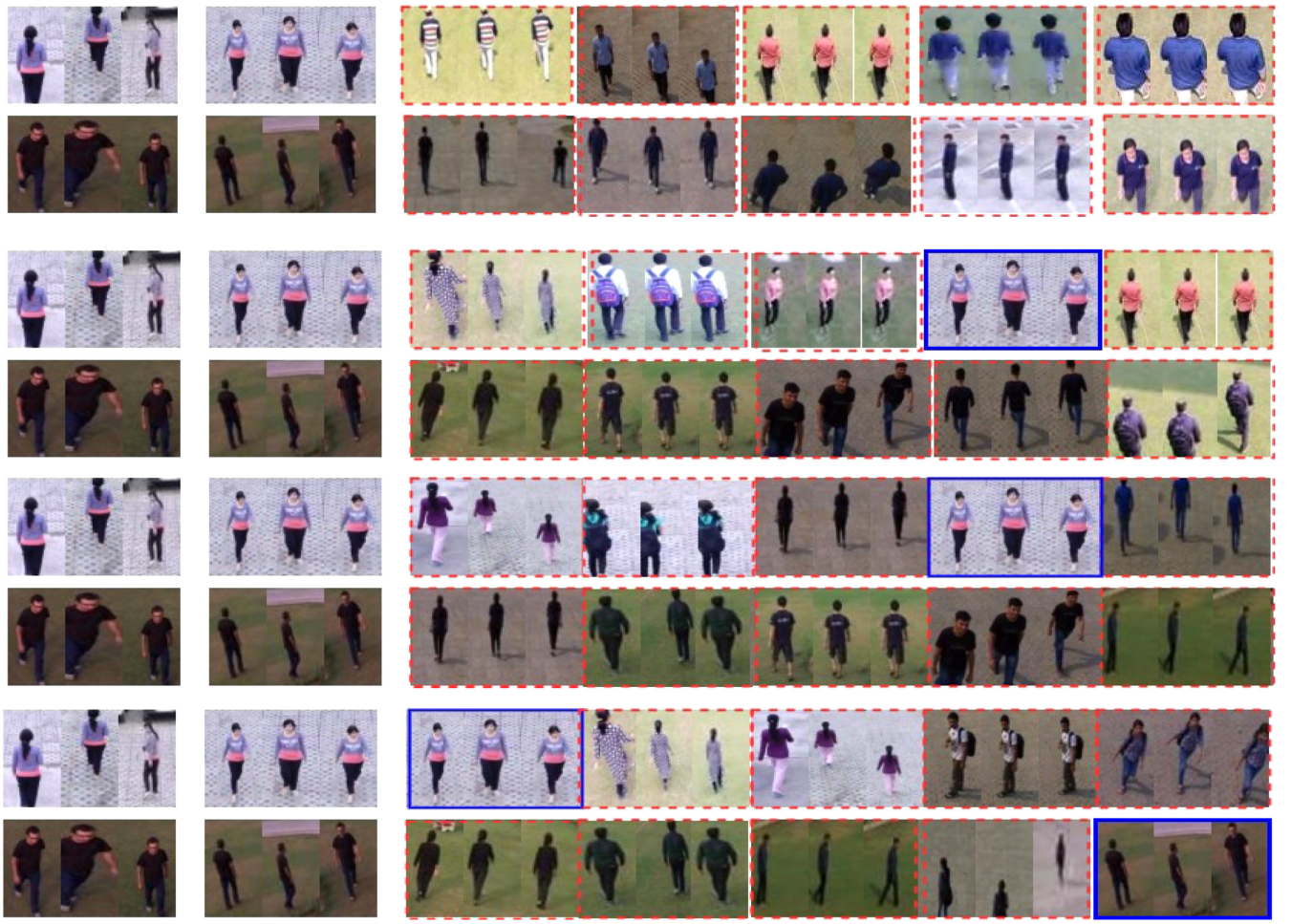

Fig. 3 Visualization of top-5 retrieval results on DJI01.

9%, and 14.25% over IDE,[1] ASPTN,[18] and RCN,[13] respectively. In addition, we use XQDA[66] metric learning method for similarity evaluation in feature vectors. The XQDA algorithm learns a discriminant subspace as well as a distance metric simultaneously, and it is able to perform dimension reduction and select the optimal dimensionality. With the incorporation of XQDA metric learning there is an improvement in the top-1 results: 1.20% in PRID-2011,[19] 1.86% in iLIDS-VID,[20] 0.90% in MARS,[1] and 2.40% in DJI01.

In Fig. 3, we visualize retrieval results of the DJI01 dataset using the existing algorithms[1,13,18] and compare it against the proposed CARF-Net algorithm. The videos in the first column are the query videos. The retrieved videos are sorted according to the similarity scores from left to right (from second column till last). Red dashed boundary indicates a negative match and blue shows a positive match. We randomly take probe videos as query and feed as an input to all the above algorithms and retrieve the corresponding gallery videos. In Fig. 3(a), we show the results of IDE.[1] We show the results of RCN[13] in Fig. 3(b). ASTPN[18] results are given in Fig. 3(c). We can see that RCN and ASTPN cover the correct results for one of the queries under top-5. In Fig. 3(d), we show the results of our proposed algorithm. We observe that in all the scenarios the algorithm covers the ground truth in top-5 retrieval results. In addition, the second row shows that the proposed algorithm is robust enough to discriminate people wearing similar clothes.

5.3.2 *Complexity*

We compare the number of parameters used in our network and in other algorithms in Table 6. CARF-NET uses only 0.75M parameters. On the other hand, methods such as IDE,[1] LuNet,[27] PBF,[67] and TLST[44] use deep networks such as CaffeNet, ResNetV2, ResNet-50, and 3D-VGGNet require extremely high number of parameters which increases the complexity of model. In addition, CARF-Net achieves all results with a feature size of 128, whereas others, such as IDE,[1] require 1024-dimensional features, and PBF[67]

Table 6 Comparison with deep networks in terms of number of parameters, feature size, and Top-1 accuracy of MARS dataset.

Method	Parameters	Feature size	Rank-1
IDE (CaffeNet)[1]	60M	1024	70.51
LuNet (ResNetV2)[27]	5M	128	75.56
PBF + ResNet-50[67]	25M	4096	72.64
PBF + VGG16[67]	134M	4096	67.27
TLST (3D-VGG)[44]	17M	4096	61.66
CARF-Net	0.75M	128	74.00

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

and TLST[44] use 4096-dimensional feature size. Thus, our model has twofold advantage of being lightweight while learning features with high discriminative ability in a much lesser dimension. We also observe that LuNet achieves an accuracy of 75.56% on MARS, whereas our network attains a comparable accuracy of 74.9% with metric learning. These observations are quite significant as our network is shallow and uses ∼15% of the parameters used by LuNet. Thus, based on the above experimental results, we can conclude that the attentive mechanism can efficiently utilize spatial and temporal human appearance to learn highly discriminative representations.

5.3.3 *Cross-dataset testing*

We discuss the comparison results in Table 7. Here, we train the system on iLIDS-VID and test on PRID-2011. It is evident that CARF-Net performs better except for being slightly inferior to TRL method at top-5. Thus, we can conclude that our model generalizes well to other datasets. We explore all the variants of cross dataset testing and report the results in Table 8.

Table 7 Comparison of cross dataset testing results, test data-PRID-2011, training data—iLIDS-VID.

CMC Top@T	T-1	T-5	T-20
ASTPN[18]	30	58	85
RCN[13]	28	57	81
TRL[37]	29.50	59.40	82.20
MRG[35]	32.80	**61.40**	72.60
CARF-NET	**36**	55	**87**

Table 8 Variants of cross dataset testing results, A-test data; B-training data.

A	B	Top-1	Top-5	Top-20
PRID-2011	iLIDS-VID	**36**	**55**	**87**
	MARS	25	35	51
	DJI01	22	27	44
iLIDS-VID	PRID-2011	**32**	**45**	**67**
	MARS	10	24	43
	DJI01	12	32	48
MARS	PRID-2011	**41**	**66**	**85**
	iLIDS-VID	15	28	43
	DJI01	9	21	38
DJI01	PRID-2011	22	35	59
	iLIDS-VID	**37**	**45**	**69**
	MARS	13	28	54

Table 9 Comparison on iLIDS-VID[20] (Top-1 accuracy) for sequence length of 1, 8,16, and 32 frames.

Seq. length	1/1	8/8	16/16	32/32
ASTPN[18]	16	35	**48**	**59**
RCN[13]	14	28	36	44
CARF-Net	**20**	**38**	46	55

5.3.4 *Fixed video sequence length*

We also analyze the performance of algorithm on iLIDS with fixed number of frame length sequence for both probe and gallery videos. We use a length of 1, 8, 16, and 32 frames. These results are shown in Table 9. Here, we observe a top-1 accuracy of 20% for 1/1 sequence length, 38% for 8/8 (probe/gallery length) sequence length, 46% for 16/16 sequence length, and 55% for 32/32 sequence length. It also indicates that as we decrease the probe and gallery sequence length, reidentification accuracy also reduces as smaller sequences will have less number of discriminative frames and insufficient temporal information.

5.3.5 *Ablation study*

We conduct an ablation study of some important factors of our method. We compare CMC rank results of all the variants of CARF-Net. The following are the different variants of our model. C-Net refers to the spatial pooling network, CA-Net refers to the attentive spatial pooling network, and CAR-Net refers to the attentive spatial pooling network with RNN. CARF-Net stands for the combination of CAR-Net and fusion of the spatial and temporal features. The results are shown in Fig. 4. The average top-1 accuracy of CA-Net over all four datasets shows a significant improvement of 9.25% over C-Net, CAR-Net by 14.20% over CA-Net, and CARF-Net by 4% over CAR-Net. Our experimental results show that the spatiotemporal fusion information obtained using attention mechanism is an important cue for person reidentification and is efficiently captured using the proposed framework (CARF-Net). The attention mechanism, multiple fusion operations, and the loss functions exploit the salient regions of the frames which leads to an increase in the accuracy.

5.3.6 *Robustness to Gaussian noise*

To test the robustness against noisy samples, we add Gaussian noise with zero mean and variance between 0.0001 to 0.003 (these examples are generated using imnoise command of MATLAB). These variance values serve the dual purpose of non-perceptually degrading the image as well as achieving substantial reduction in the model's performance. We show some clean and adversarial examples in Fig. 5. We test the noisy examples with the proposed algorithm and existing works.[1,13,18] The results are shown in Fig. 6 over all the four datasets. We can observe that for top-1 accuracy, CARF-Net performs better on PRID-2011 and DJI01, whereas ASTPN[18] performs better on iLIDS-VID, and IDE[1] performs better on MARS dataset. The average top-1 accuracy across all four datasets is 44.00% for

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

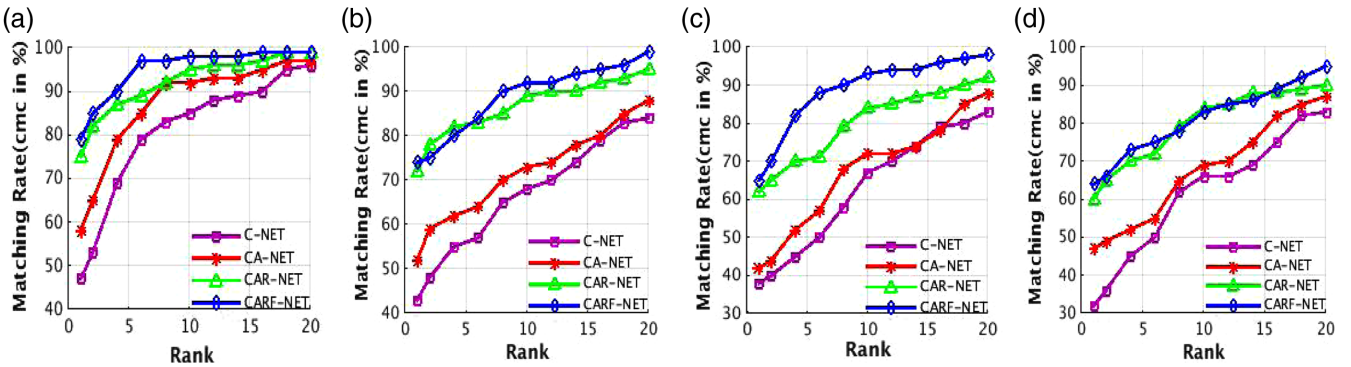

Fig. 4 Ablation study results of (a) PRID-2011,[19] (b) MARS[1], (c) iLIDS-VID,[20] and (d) DJI01.

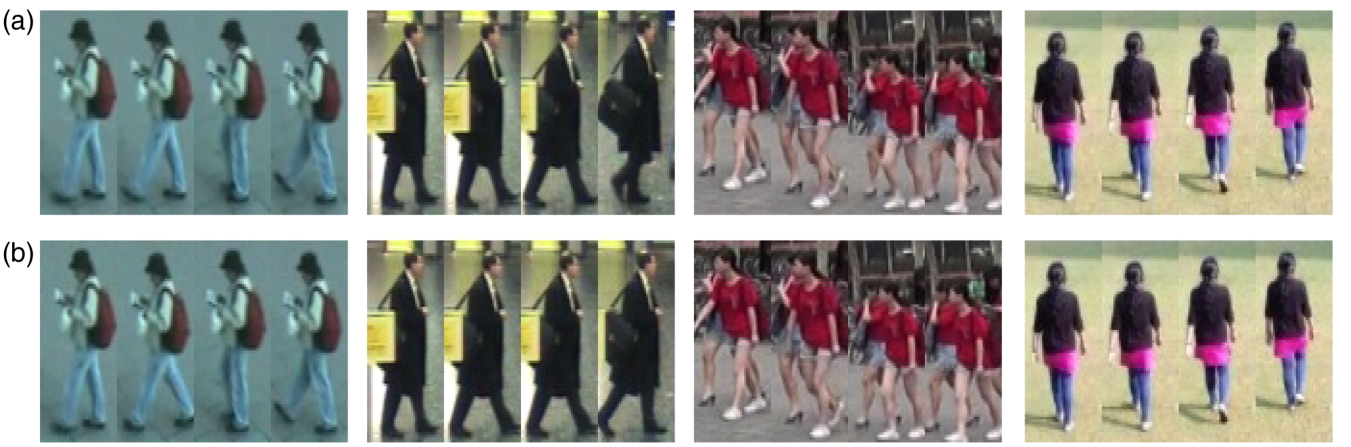

Fig. 5 (a) Shows clean examples and (b) shows examples perturbed with Gaussian noise.

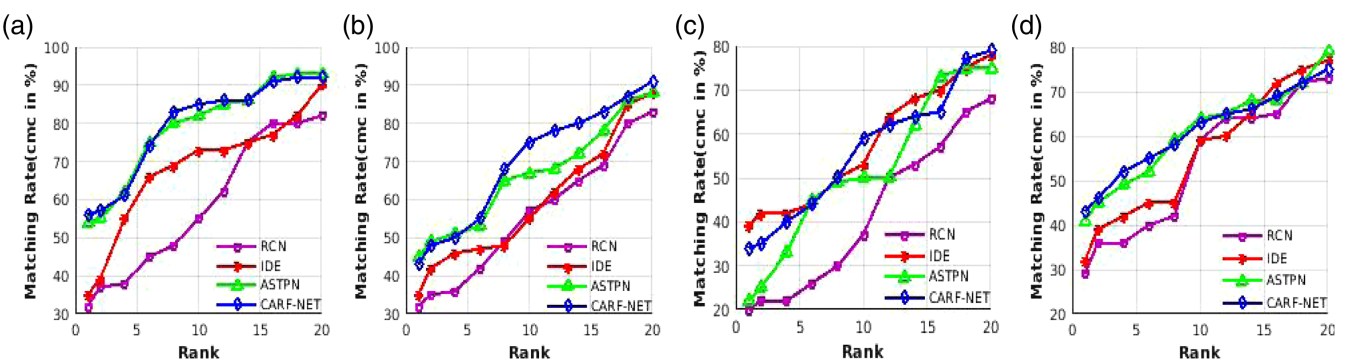

Fig. 6 Noise robustness results of (a) PRID-2011,[19] (b) iLIDS-VID,[20] (c) MARS,[1] and (d) DJI01.

CARF-Net, 40.50% for ASTPN,[18] 35.25% for IDE,[1] and 28.25% for RCN.[13]

As the vulnerability of RGB against noisy samples is well known, we investigate the performance degradation due to optical flow stream only. We perform experiments with noisy RGB and noisy optical flow, as well as with noisy RGB and clean optical flow. Here, noisy optical flow is the optical flow obtained from noisy RGB frames and clean optical flow is obtained from clean RGB frames. In the former case, average accuracy across all four datasets is 44.00% for CARF-Net and 40.50% for ASTPN.[18] In the latter case, average accuracy is 46.25% for CARF-Net and 41.25% for ASTPN.[18] We observe that the accuracy change between the two cases is low. This suggests that the optical flow inherently provides robustness when noise is added spatially. Optical flow gives temporal information and may not be as vulnerable as noisy RGB frames. Further, we also see that the increase in accuracy with clean optical flow is highest for the case of CARF-Net. This may also indicate the fact that two-stream networks provide an advantage over single-stream networks.

5.3.7 *t-SNE plots*

To qualitatively understand each component of CARF-Net, we analyze the learned embedding from C-Net, CA-Net, CAR-Net, and CARF-Net using t-SNE plots, a dimensionality reduction technique to visualize high-dimensional embeddings. We obtain the two resulting dimensions which

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

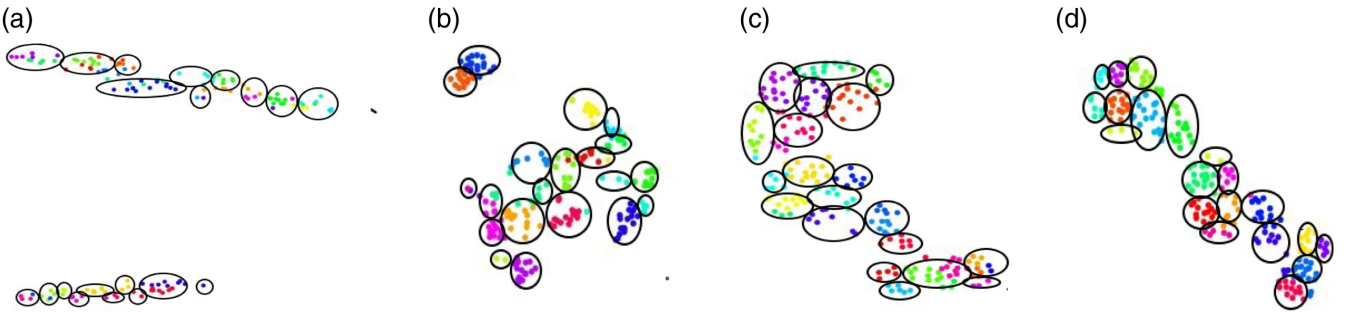

Fig. 7 Visualizations of the MARS[1] embeddings by t-SNE: embedding from (a) C-Net, (b) CA-Net, (c) CAR-Net, and (d) CARF-Net.

can be visualized by creating a scatter plot and by coloring each sample by its respective label. A total of 220 samples for 20 identities from MARS[1] dataset is used for the plot. We extract four types of embeddings from C-Net, CA-Net, CAR-Net, and CARF-Net, as shown in Fig. 7 (from left to right). **C-Net:** here, most of the dissimilar points cluster together incorrectly. **CA-Net:** we can see that similar points get more closer as compared to C-Net. It also shows the effect of SPP layer. **CAR-Net:** clusters can be easily differentiated as they are well aligned and well separated, though 40% of the clusters still contain samples of different identities. **CARF-Net:** we can see that the samples are very clearly clustered in their own compact group and >90% of similar points are correctly clustered.

6 Conclusion

In this paper, we propose an end-to-end joint learning network to address the problem of video-based person reidentification. Our proposed CARF-Net uses a two-stream triplet network which incorporates CNN, discriminative visual attention, RNN, and multiple pooling layers. To preserve information from discriminative frames and to learn the robust features, we apply multiple spatial, temporal, and spatiotemporal fusion operations. Further, we use a multiloss method as an objective function to train the network. We rigorously evaluate the performance of multiple variants of the proposed network and conclude that CARF-Net gives best results. This is because the network exploits most informative regions for person reidentification through visual attention on frames and optical flow vectors, whereas temporal and spatiotemporal pooling extracts most discriminative frames. The extensive comparison against several popular algorithms also shows the efficiency of the proposed network.

We also introduce an aerial video dataset with several challenges. In future, we plan to enrich the dataset and analyze reidentification based on a combination of an image and a video sequence.

References

1. L. Zheng et al., "Mars: a video benchmark for large-scale person re-identification," in *Eur. Conf. Comput. Vision*, pp. 868–884 (2016).
2. A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vision Comput.* **32**(4), 270–286 (2014).
3. R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: a survey," *ACM Comput. Surv.* **46**(2), 29 (2013).
4. F. Zhu et al., "A novel two-stream saliency image fusion CNN architecture for person re-identification," *Multimedia Syst.* **24**(5), 569–582 (2018).
5. S. B. P. Carr, "Deep spatial pyramid for person re-identification," in *Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6 (2017).
6. X. Yang et al., "Enhancing person re-identification in a self-trained subspace," *ACM Trans. Multimedia Comput. Commun. Appl.* **13**(3), 27 (2017).
7. Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput. Commun. Appl.* **14**(1), 13 (2017).
8. G. Lisanti, S. Karaman, and I. Masi, "Multichannel-kernel canonical correlation analysis for cross-view person reidentification," *ACM Trans. Multimedia Comput. Commun. Appl.* **13**(2), 13 (2017).
9. G. Watson and A. Bhalerao, "Person reidentification using deep foreground appearance modeling," *J. Electron. Imaging* **27**(5), 051215 (2018).
10. B. Yu and N. Xu, "Deep triplet-group network by exploiting symmetric and asymmetric information for person reidentification," *J. Electron. Imaging* **27**(3), 033033 (2018).
11. K. Li et al., "Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification," in *32nd AAAI Conf. Artif. Intell.* (2018).
12. S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(12), 2963–2977 (2018).
13. N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Comput. Vision and Pattern Recognit.*, pp. 1325–1334 (2016).
14. A. Graves, "Generating sequences with recurrent neural networks," arXiv:1308.0850 (2013).
15. R. R. Varior et al., "A Siamese long short-term memory architecture for human re-identification," in *Eur. Conf. Comput. Vision*, pp. 135–153 (2016).
16. L. Wu, C. Shen, and A. van den Hengel, "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach," arXiv:1606.01609 (2016).
17. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, pp. 568–576 (2014).
18. S. Xu et al., "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Int. Conf. Comput. Vision* (2017).
19. M. Hirzer et al., "Person re-identification by descriptive and discriminative classification," in *Scand. Conf. Image Anal.*, pp. 91–102 (2011).
20. T. Wang et al., "Person re-identification by video ranking," in *Eur. Conf. Comput. Vision*, pp. 688–703 (2014).
21. L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: past, present and future," arXiv:1610.02984 (2016).
22. A. Schumann and T. Schuchert, "Person re-identification in uav videos using relevance feedback," *Proc. SPIE* **9407**, 94070Z (2015).
23. A. Schumann and T. Schuchert, "Deep person re-identification in aerial images," *Proc. SPIE* **9995**, 99950M (2016).
24. A. Schumann and J. Metzler, "Adapted deep feature fusion for person re-identification in aerial images," *Proc. SPIE* **10643**, 106430L (2018).
25. R. Layne, T. M. Hospedales, and S. Gong, "Investigating open-world person re-identification using a drone," in *Eur. Conf. Comput. Vision*, pp. 225–240 (2014).
26. K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Eur. Conf. Comput. Vision*, pp. 346–361 (2014).
27. A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, (2017).
28. C. Su et al., "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognit.* **75**, 77–89 (2018).
29. L. Wei et al., "Glad: global-local-alignment descriptor for pedestrian retrieval," in *ACM Multimedia*, pp. 420–428 (2017).
30. X. Zhu et al., "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *Int. Joint Conf. Artif. Intell.*, pp. 3552–3559 (2016).

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

31. X. Zhu et al., "Semi-supervised crossview projection-based dictionary learning for video-based person re-identification," *IEEE Trans. Circuits Syst. for Video Technol.* **28**(10), 2599–2611 (2018).
32. X.-Y. Jing et al., "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 695–704 (2015).
33. Y. Wu et al., "Temporal-enhanced convolutional network for person re-identification," in *32nd AAAI Conf. Artif. Intell.* (2018).
34. X. Zhu et al., "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Trans. Inf. Forensics Secur.* **13**(3), 717–732 (2018).
35. Z. Li et al., "Multi-rate gated recurrent convolutional networks for video-based pedestrian re-identification," in *32nd AAAI Conf. Artif. Intell.* (2018).
36. Y. Yan et al., "Person re-identification via recurrent feature aggregation," in *Eur. Conf. Comput. Vision*, pp. 701–716 (2016).
37. J. Dai et al., "Video person re-identification by temporal residual learning," *IEEE Trans. Image Process.* **28**(3), 1366–1377 (2019).
38. Z. Zhou et al., "See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Comput. Vision and Pattern Recognit.*, pp. 6776–6785 (2017).
39. Z. Zhong et al., "Re-ranking person re-identification with k-reciprocal encoding," in *Comput. Vision and Pattern Recognit.*, pp. 3652–3661 (2017).
40. D. Chung, K. Tahboub, and E. J. Delp, "A two stream Siamese convolutional neural network for person re-identification," in *Int. Conf. Comput. Vision*, pp. 1983–1991 (2017).
41. Z. Yu et al., "Three-stream convolutional networks for video-based person re-identification," arXiv:1712.01652 (2017).
42. D. Li et al., "Learning deep context-aware features over body and latent parts for person re-identification," in *Comput. Vision and Pattern Recognit.*, pp. 384–393 (2017).
43. J. Li et al., "Lvreid: person re-identification with long sequence videos," arXiv:1712.07286 (2017).
44. K. Kansal and A. Subramanyam, "Transfer learning of spatio-temporal information using 3D-CNN for person re-identification," in *Proc. IEEE Conf. Syst. Man Cybern.* (2018).
45. D. Li, D. Zeng, and K. Zhao, "Dhnet: working double hard to learn a convolutional neural network-based local descriptor," *J. Electron. Imaging* **27**(4), 043008 (2018).
46. W. Chen et al., "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 403–412 (2017).
47. K. Li et al., "Support neighbor loss for person re-identification," in *ACM Multimedia Conf.*, ACM, pp. 1492–1500 (2018).
48. H. Jin et al., "Deep person re-identification with improved embedding and efficient training," in *IEEE Int. Joint Conf. Biometrics (IJCB)*, IEEE, pp. 261–267 (2017).
49. A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Comput. Vision and Pattern Recognit.*, pp. 1229–1238 (2016).
50. X. Wang et al., "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. Multimedia* **20**(3), 634–644 (2018).
51. C. Ma et al., "Hierarchical convolutional features for visual tracking," in *Int. Conf. Comput. Vision*, pp. 3074–3082 (2015).
52. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Int. Joint Conf. Artif. Intell.*, pp. 674–679 (1981).
53. L. Chen et al., "Deep spatial-temporal fusion network for video-based person re-identification," in *CVPRW*, pp. 63–70 (2017).
54. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *Comput. Vision and Pattern Recognit.*, pp. 815–823 (2015).
55. H. Liu et al., "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.* **26**(7), 3492–3506 (2017).
56. A. Dehghan, S. M. Assari, and M. Shah, "Gmmcp tracker: globally optimal generalized maximum multi clique problem for multiple object tracking," in *Comput. Vision and Pattern Recognit.*, pp. 4091–4099 (2015).
57. K. Liu et al., "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *Int. Conf. Comput. Vision*, pp. 3810–3818 (2015).
58. J. You et al., "Top-push video-based person re-identification," *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.* (2016).
59. Y. Li et al., "Multi-shot human re-identification using adaptive fisher discriminant analysis," in *Br. Mach. Vision Conf.*, pp. 1–12 (2015).
60. S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Int. Conf. Comput. Vision*, pp. 4516–4524 (2015).
61. W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Trans. Circuits Syst. Video Technol.* **28**(10), 2768–2776 (2018).
62. L. Wu et al., "Where-and-when to look: deep Siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia* (2018).
63. X. Zhang and B. Bhanu, "An unbiased temporal representation for video-based person reidentification," in *25th IEEE Int. Conf. Image Process. (ICIP)*, IEEE, pp. 838–842 (2018).
64. X. Zhu et al., "Simultaneous visual-appearance-level and spatial-temporal-level dictionary learning for video-based person re-identification," in *Neural Comput. Appl.*, pp. 1–13 (2018).
65. G. Song et al., "Region-based quality estimation network for large-scale person re-identification," in *32nd AAAI Conf. Artif. Intell.* (2018).
66. S. Liao et al., "Person re-identification by local maximal occurrence representation and metric learning," in *Comput. Vision and Pattern Recognit.*, pp. 2197–2206 (2015).
67. C. Liu, T. Bao, and M. Zhu, "Part-based feature extraction for person re-identification," in *Int. Conf. Mach. Learn. Comput.*, pp. 172–177 (2018).

Kajal Kansal received her BTech degree from PTU Giani Zail Singh Campus, Bathinda, Punjab, in 2014. Currently, she is pursuing her PhD with the Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Delhi, India. She works in the area of computer vision and deep learning. Her research interest includes person reidentification and robust feature extraction problem using learning-based techniques. She is a recipient of the Visvesvaraya Research Fellowship.

Subramanyam Venkata earned his PhD from Nanyang Technological University, Singapore, in 2013 and his BTech from Indian Institute of Technology (Indian School of Mines), Dhanbad, in 2007. Currently, he is an assistant professor at Indraprastha Institute of Information Technology, Delhi. He works in the area of multimedia and vision, information forensics and security, and machine learning. He regularly publishes as well as serves as a reviewer in top-ranked journals and conferences.

Dilip K. Prasad received his PhD and BTech degrees in computer science and engineering from Nanyang Technological University, Singapore and Indian Institute of Technology (Indian School of Mines), Dhanbad, India, in 2012 and 2003, respectively. Currently, he is a senior research fellow at Nanyang Technological University, Singapore. His current research interests include image processing, pattern recognition, and artificial intelligence. He was the founder of techaloo.com (successful exit and acquired by publicize.co) and gpl4you.com.

Mohan Kankanhalli obtained his BTech (electrical engineering) from the Indian Institute of Technology, Kharagpur, in 1986 and his MS and PhD (computer and systems engineering) from the Rensselaer Polytechnic Institute in 1988 and 1990, respectively. He subsequently joined the Institute of Systems Science in October 1990. During 1997–1998, he was a faculty member at the Department of Electrical Engineering, Indian Institute of Science, Bangalore. He has been with the NUS School of Computing since May 1998, where he is provost's chair professor of computer science. He is also the dean of NUS School of Computing. Before that, he was the vice provost (graduate education) for NUS during 2014–2016 and associate provost (graduate education) during 2011–2013. He is a fellow of IEEE.

Downloaded From: https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging on 12 Jun 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use