Transfer Learning of Spatio-Temporal Information using 3D-CNN for Person Re-identification

Kajal Kansal
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology
Delhi, India 110020
Email: kajalk@iiitd.ac.in.

A.V Subramanyam
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology
Delhi, India 110020
Email: subramanyam@iiitd.ac.in.

***Abstract*—Video based person re-identification has received significant attention in the recent past due to advancement in deep learning techniques. However, it is still a challenging problem because of inherent dynamic nature of videos. Additionally, lack of sufficient annotated dataset may lead to overfitting issues while training deep networks. In this paper, we propose a spatio-temporal transfer learning approach using 3D-CNN for video based person re-identification. To address the issue of insufficient labelled data and transfer the knowledge, we use a pre-trained 3D-CNN model of Sports-1M dataset and perform fine-tuning on multiple domain datasets such as PRID-2011, iLIDS-VID, MARS and an aerial video dataset simultaneously. Learning features from multiple domain data is of significant value because of large variation which otherwise is not possible to obtain from small individual datasets. In our experiments, we show that the fine-tuned transferred features encode robust representations and enhance the re-identification accuracy. Further, to boost the performance, we apply XQDA metric learning. Experiments conducted on all the four datasets show that the proposed framework outperforms the popular methods by an average improvement of 4% or more. In addition, we analyse the network's robustness against adversarial examples and show that the proposed 3D-CNN network has better resilience compared to 2D-CNN used in most of the existing algorithms.**

***Index Terms*—Air-borne Videos, 3D-CNN, Spatio-temporal transfer learning, Video-surveillance.**

I. Introduction

Person re-identification [2]–[5] has been well studied and widely used in computer vision applications such as long term tracking and video surveillance. Person Re-identification is to identify the same person across same or different cameras. It is a challenging task due to the large variations in pose, occlusion, background clutter, illumination change etc. Re-identification can be broadly classified into two main categories. The first one is feature extraction and learning discriminative models corresponding to that features. The learnt model is then used to re-identify a person. However, how to extract robust features is still an open problem today. On the other hand, deep learning based models have shown tremendous potential for learning feature embeddings that can lead to significant performance improvement.

Deep Learning based methods encode the reliable features and removes the need of extracting hand-crafted features. Deep Learning architectures based on Recurrent Neural Network (RCN) [6] and Long Short-Term Memory [7]–[10] models have been explored for person re-identification. In [6], 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 limitation with learning of the long duration sequences of the inputs and it can cause vanishing gradient problem [6]. Varior et al. [8] propose a siamese LSTM architecture that can process image regions sequentially and enhance the discriminative capability of local feature representation by leveraging contextual information. A drawback of the siamese model is that it does not utilize complete label information. The siamese model only considers if a pair of image or video is similar or not which is a weak label in re-identification. Another effective strategy is classification/identification mode. Xiao et al. [11] forms training set from multiple datasets and a softmax loss is employed in the classification network. However, this work is limited to images only. On larger datasets, the classification model achieves good performance. However, identification loss requires more training instances per ID for model convergence.

Training Convolutional Neural Networks (CNNs) with large number of parameters require sufficiently large amount of data. In cases, where the available data is small, it can cause overfitting issues. Addressing the problem of insufficient data to train a neural network, [12] introduced a large, automatically generated, database gathered from YouTube clips called Sports-1M. Due to availability of such dataset, one can utilize the pre-trained model on such large datasets and transfer the knowledge to required domain using Transfer Learning. This can be accomplished by replicating the learned weights and biases from one or more layers of a fully trained network to a different network. It can be used to overcome overfitting issues and to speed up the training process for a related task.

After feature extraction, various metric learning methods are used to find out the similarity between a pair of images or videos. Some of the prominent metric learning methods for re-identification are Cross-view Quadratic Discriminant Analysis (XQDA) [1], Relevance Component Analysis [13], Relaxed Pairwise Learning [14], Deep Metric Learning [15] and Large Margin Nearest Neighbour [16]. Liao et al. [1] propose a metric learning method called XQDA which learns a more discriminative distance metric and a low-dimensional subspace simultaneously. Hirzer et al. [14] learns a metric from pairs of samples from different cameras and once the metric has been

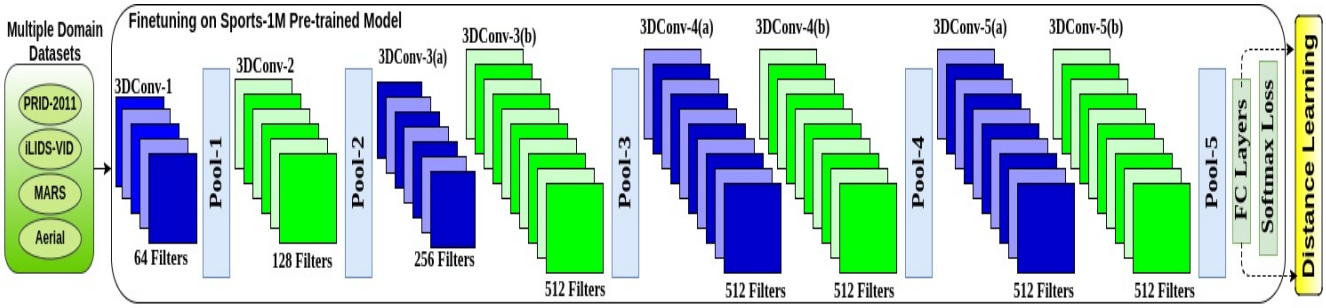

Fig. 1. Proposed Framework, 3DConv-1 refers to first convolutional layer, 3D-Conv2 refers to second Convolutional layer and so on. Pool-1 refers to first pooling layer, Pool-2 refers to second pooling layer and so on. Number of filters are shown corresponding to every layer, FC layers refers to three fully connected layers followed by Softmax Loss layer. We extract features from second fully conected layer, which is followed by Euclidean distance and XQDA [1] metric learning for computing similarity matrix.

learned, only linear projections are necessary at search time, where a simple nearest neighbor classification is performed to reidentify a person. Killian and Saul [16] introduce the metric which is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin.

In this paper, we propose an efficient deep 3D-CNN [17] based approach to extract optimal and robust spatio-temporal features for improving re-identification performance. We empirically show that these learned features can yield good performance in re-identification tasks. 3D-CNNs encapsulate information related to person appearance and motion in a video, making them useful for re-identification task. To overcome the need of large training data, we perform fine-tuning on multiple domain (different static-moving camera dataset at a time) data simultaneously. Our network outputs an overall feature representation by combining information from all input sequences, which is then followed by Euclidean distance and XQDA [1] metric learning for computing similarity matrix. We demonstrate the effectiveness of our approach through an extensive evaluation and experimental analysis.

II. Proposed Algorithm

In the proposed algorithm, we apply a 3D-CNN to encode spatio-temporal appearance of a person. We extract the features from second fully connected layer. Once the features are obtained, we use an Euclidean norm and XQDA metric learning method to compute similarity between a pair of features. In order to extract the robust features, we fine-tune the pretrained model of Sports-1M [12] using multiple domain dataset such as PRID-2011, iLIDS-VID, MARS and Aerial. In [12], an empirical evaluation of 3D-CNN has been performed on a large dataset with the goal of classification. Encouraged by the results, we use the pre-trained model to fine-tune our network.

A. Transfer Learning and Data Augmentation

In this section, we discuss the transfer learning and data augmentation part of our algorithm. To reducing the overfitting problem, we perform transfer learning of spatio-temporal information on multiple domains. The idea is to use pre-trained weights of Sports-1M [12] from an existing network, and then to tune them with multiple domain dataset. If we have more data, we can have more confidence that we would not overfit while fine-tuning through the full network. If the data is small, it is not a good idea to fine-tune the ConvNet due to overfitting concerns. We have different domain data as an input to get a fine-tuned model. Since we use multiple domain data, the features have good generic capability by which it can represent different types of videos well while being discriminative.

To reduce overfitting and data imbalance, we apply data-augmentation. It enlarges the dataset artificially. We augment the data by performing cropping of 112×112 on all the frames. We also perform resizing in width and height of frames to 128×171. As our dataset is not similar to the Sports-1M dataset, we finetune the network over all layers. In addition to multi-domain data and data-augmentation techniques, we use dropout with 0.5 probability, which is an efficient way to avoid overfitting issues.

B. Architecture of 3D-CNN

Our 3D-CNN architecture is inspired from 3D-VGGNets [17]. The architecture is shown in Fig.1. It has 8 convolutional layers, 5 max-pooling layers, and 3 fully connected layers, followed by softmax-loss layer. In [17], authors point out that a homogeneous architecture with 3x3x3 convolution kernels in all layers perform best for 3D-CNN. In our 3D-CNN, we also apply $3\times3\times3$ convolutional Kernels and $2\times2\times2$ pooling kernels. All 3D convolutional kernels are $3\times3\times3$ with stride 1 in both spatial and temporal dimensions. All pooling kernels are $2\times2\times2$, except for 1st pooling layer kernel which is $1\times2\times2$. We use Rectified linear unit (ReLU) as an activation function to accelerate the convergence of stochastic gradient descent. The convolutional layers are followed by three fully connected layers. First two fully connected layers have 4096 output units.

In [17], it is shown that after three convolutional layers, network starts loosing temporal information. In order to preserve temporal information, it is suggested to have a pair of sequential 3D-convolutional layers after two single 3D convolutional layers. This architecture results in preserving

temporal information better than a single convolution layer followed by a pooling layer.

C. Feature Extraction

We use the second fully connected layer for feature extraction. It gives feature vectors corresponding to the appearance and motion of a person in the video.

Let the feature vectors be $\overrightarrow{x_i}, \forall i \in 1, 2, 3...N$ corresponding to N persons. Then, we use Euclidean distance for computing similarity measures between feature vectors extracted from two different videos. Let D be this distance. Now, if i^{th} person is the query, then we can match $\overrightarrow{x_i}$ against $\overrightarrow{x_j}$, $\forall j \in 1, 2..N$ and $j \neq i$. Thus,

$$\mathrm{D}(\overrightarrow{x_i}, \overrightarrow{x_j}) = \|\overrightarrow{x_i} - \overrightarrow{x_j}\|_2^2, \quad \forall j \in 1, 2, 3...N \text{ and } j \neq i \qquad (1)$$

The person j whose distance is minimum to the person i is the closest match. This, gives us Top-1 accuracy. Similarly, we can compute Top-5, Top-10 and Top-20 accuracies.

In addition, we use XQDA [1] 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.

D. Training Objective

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. $S_c \in R^M$ and $S_\Gamma \in R^M$ denote the c^{th} and Γ^{th} column of the softmax weight matrix S, respectively. Hence, we use the following softmax function:

$$L_{softmax} = P(z = c|I) = \frac{exp(S_c I)}{\sum_\Gamma exp(S_\Gamma I)} \qquad (2)$$

III. Weight Visualization

A. Learned Weight Filters

Interacting with Deep Neural Networks (DNNs) can help in building our intuitions, which can help us in design better models. We show some samples of weight visualization in Fig 2 corresponding to first convolutional layer. Our visualizations suggest that many neurons represent abstract features such as edges, blobs, direction in change in intensity, and other discriminative information etc. For weight visualization, we use Direct encoding method [18].

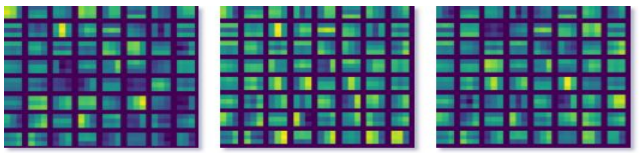

Fig. 2. Weight visualization results: 3DConv-1 Layer for all three channels

B. Visualizing Layers Response

In order to analyse what type of spatio-temporal patterns does the proposed model learn, we show visual results corresponding to the first layer and second layer in Fig. 3. We can see the activated neuron at the backend corresponding to the inputs. The first layer encodes direction and color. These directions and color filters then get combined into basic grid and spot textures. These textures gradually get combined into increasingly complex patterns. In Fig. 3, we can see that network is learning motion information at first layer.

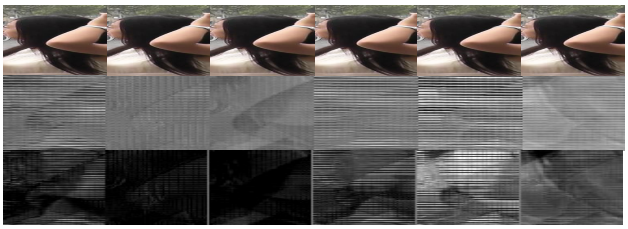

Fig. 3. Layer Responses at 3DConv-1 and 3DConv-2 Layer on samples MARS

IV. Experiments

We experimentally validate the model using PRID-2011, iLIDS-VID, MARS and Aerial dataset. Our experimental results show that spatio-temporal information of the video is an important cue for person re-identification and is efficiently captured using 3D-CNN.

A. Experimental set-up

Our architecture is implemented using Caffe [19]. We finetune two models. In first, we randomly select 50% of data from PRID-2011, iLIDS-VID and Aerial, and test it on same datasets as well as on MARS dataset to see cross dataset testing results. And then we again finetune the first model on 50% of MARS dataset and report the results for MARS dataset. First model takes approximately 20 hours on K40 Tesla GPU for fine-tuning with 10,000 iterations and another model took 5 days on K20 Tesla GPU with the same parameters. We repeat the experiments five times and report average accuracy. The values of the hyper-parameters are set based upon the experiments. We use the following values, batch size =50, learning rate=0.0001, momentum=0.9, and weight decay=0.0005. For network learning, we use stochastic gradient descent with minibatch.

B. Datasets

PRID-2011 consists of 200 identities captured by two cameras of sequence length 5 to 675. It is captured in uncrowded outdoor environment with stark difference in illumination, background clutter and less occlusions. iLIDS-VID dataset is more challenging due to occlusions, illumination changes and viewpoint variations. It has 300 identities captured from 2 cameras with sequence length 23 to 192 frames. MARS is much larger dataset with 1261 count captured from 6 cameras whose sequence length vary from 32 to 20,000. Aerial dataset

consists of 200 identities collected using two quadcopters in our campus whose sequence length vary from 32 to 3000 frames. There are significant challenges like occlusion, pose variations, illumination changes, scale variations, people with similar clothes and same people at different days. Few examples of Aerial dataset are shown in Fig. 4.

Fig. 4. Samples of Aerial Dataset

C. Performance Comparison

We compare the results of our algorithm with several popular techniques. We first compare it against already trained 3D-CNN [17] model on Sports-1M without fine-tuning. This model acts as a baseline for us. where as in Ours+Eucledian model, we have done finetuning over multiple domain dataset. We compare these finetuned results against baseline [17] to analyze the effect of transfer learning We report baseline [17] results using XQDA as it performs better than Euclidean. In [20], the authors use two stream CNN to learn spatio-temporal features separately. ASTPN [21] takes the advantage of attention mechanism to extract features from informative frames. In [22], the authors use Alexnet to extract features and metric learning to compute the similarity. [23] learns an intra-video and inter-video distance metric from the training videos. In [6], [24], [25], spatio-temporal features are used. [26] uses 3D HOG, color and LBP features, and learn a distance metric for matching.

1) PRID, iLIDS-VID, and Aerial Results: We report the results in terms of CMC Top 1-5-10-20 accuracies in Table I. We use first model to report CMC accuracies for PRID-2011, iLIDS-VID, Aerial and cross testing on MARS.

On PRID-2011, we achieve 73.09% Top-1 accuracy, 98.53% Top-5 accuracy, 99.41% Top-10 accuracy, and 99.41% Top-20 accuracy with Euclidean measure whereas with XQDA [1] metric learning method Top-1 results increase to 76.02%. We observe that the proposed algorithm gives better results when compared to several popular algorithms in most of the cases.

On iLIDS-VID, we again observe that the performance of the proposed algorithm is significantly better when compared with other schemes. In addition, with incorporation of XQDA metric learning there is an improvement of 14.42% in Top-1 result. Under Top-5, we see that TDL [26] gives better results. Interestingly, we observe that Euclidean measure gives better results for Top 10 and 20 accuracies.

TABLE I
COMPARISON OF PROPOSED APPROACH WITH THE STATE-OF-THE-ART ON PRID-2011 [27], ILIDS-VID [28] AND ON AERIAL DATASET IN TERMS OF CMC TOP 1-5-10-20 ACCURACY.

PRID-2011	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Ours+Euclidean	73.09	**98.53**	**99.41**	**99.41**
Baseline [17]	58.7	65.3	68.3	75.6
Chung et. al. [20]	**78**	94	97	99
ASTPN [21]	77	95	99	99
IDE [22]	77.3	93.5	–	99.3
SI^2DL [23]	76.7	95.6	96.7	98.9
RCN [6]	70	90	95	97
STA [24]	64	87	90	92
RFA [25]	58.2	85.8	93.4	97.9
TDL [26]	56.74	80	87.64	93.54
iLIDS-VID	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Ours+Euclidean	59.20	86.06	**99.00**	**99.00**
Baseline [17]	42.5	54.8	67.9	76.3
ASTPN [21]	**62**	86	94	98
Chung et. al. [20]	60	86	93	97
IDE [22]	53.0	81.4	–	95.1
SI^2DL [23]	48.7	81.1	89.2	97.3
RCN [6]	58	84	91	96
STA [24]	44	72	84	92
RFA [25]	49.3	76.8	85.3	90.0
TDL [26]	56.33	**87.60**	91	96
Aerial	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Ours+Euclidean	42.76	68.88	**86.66**	95.66
Baseline [17]	29.4	38.6	57.2	75.3
ASTPN [21]	**63**	**70**	85	**98**
IDE [22]	59	65	82	84
RCN [6]	55	62	83	92

TABLE II
CROSS TESTING AND COMPARISON RESULTS ON MARS [22] IN TERMS OF CMC TOP 1-5-10-20 ACCURACY.

Cross Testing	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Ours+Euclidean	27.9	32.8	35.11	42.56
Ours+XQDA [1]	**38.7**	42.6	48.9	**55.7**
Baseline [17]	35.3	**44.5**	**50.2**	54.9
Trained On MARS	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Ours + Euclidean	61.66	**82.63**	**88.33**	88.42
ASTPN [21]	44	70	74	81
IDE [22]	**68.3**	82.60	–	**89.4**
RCN [6]	40	64	70	77

Top-1 accuracy of 42.76% is obtained on Aerial dataset using Euclidean measure and an increase of 7.84% in Top-1 accuracy is observed using XQDA. Further, we also observe that the Baseline [17] performs poorly due to train/test domain mismatch.

2) Mars results: To evaluate the robustness of model, cross dataset testing experiment is designed where a model trained on dataset A, may not perform well on a statistically different dataset B which indicates over-fitting to the particular scenario. Therefore, we perform cross dataset testing on larger dataset i.e MARS to analyze the generalizability of model. We first report cross dataset testing results in Table 2 for MARS dataset using first model which is finetuned on PRID-2011, iLIDS-

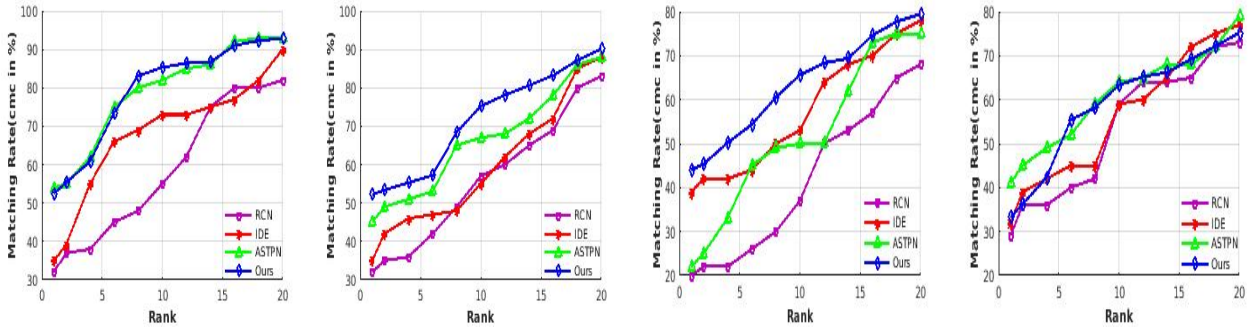

Fig. 5. Adversarial results on PRID-2011 [27], iLIDS-VID [28], MARS [22] and Aerial (from left to right)

VID and Aerial. The results in the cross dataset scenario are worse, as expected, probably due to dataset bias. Further, to improve the results on MARS dataset we perform training on 631 identities of this dataset. In [22], authors report best Top-1 accuracy of 68.3% with 2D-CNN Descriptor. Whereas we achieve an improved Top-1 accuracy of 75.17% by using 3D-CNN descriptor and XQDA. On the other hand, we observe Top-5 and Top-10 accuracies are 82.63% and 88.33% using Euclidean measure. However, [22] gives marginally better Top-20 results of 89.4% when compared to 88.42% obtained from our algorithm.

The average Top-1 accuracy of our algorithm across all four datasets shows a significant improvement of 4%, 7%, 13.25% and 27.8% over IDE [22], ASPTN [21], RCN [6] and baseline [17] respectively.

3) Exploiting Temporal Pattern: In order to better understand the temporal pattern encoding of the network, we analyze two specific cases in our Aerial dataset. First, where people wear similar clothes (a total of 5 people), and second where same people are captured on different days (a total of 8 people). In both the cases, we find that Top-10 results cover the ground truth for all 13 queries. This is still significant considering the variation in dataset.

In the first case where people wear similar clothes, then in the absence of strong cues such as visible face, we can infer that the temporal pattern is learned in a discriminative manner. In this scenario, temporal information is the most discriminative pattern available for people with similar clothes. This leads to the point that network is capable of encoding temporal cue which is predominantly walk pattern. Further, since the videos are captured from an altitude of 3m to 35m, it can be observed that other features might not be discriminative enough.

The second case strengthens our hypothesis that the network learns strong time series pattern. In this case, we match same people on different days. The two videos being matched have very different scales, non-discriminative spatial cues as face is not clearly visible, different angles, illumination variation and varying background along with added challenges due to camera motion. Thus, we can emphasize the fact that only the walk pattern is common on different days which is what the network encodes.

4) Fixed Video Sequence Length: We also analyse the performance of algorithm on iLIDS with fixed number of frame length sequence for both probe and gallery videos. We use 1, 8, 16 and 32 frames. These results are shown in Table III. Here, we observe that as we decrease probe and gallery sequence length re-identification accuracy also reduces.

TABLE III
COMPARISON ON ILIDS-VID FOR SEQUENCE LENGTH OF 1,8,16 AND 32 FRAMES FOR CMC TOP-1 ACCURACY

Sequence Length	**1/1**	**8/8**	**16/16**	**32/32**
Ours + Euclidean	**20**	33.86	**52.71**	56.80
ASTPN [21]	16	**35**	48	**59**
RCN [6]	14	28	36	44

V. ADVERSARIAL EXAMPLES

An adversarial example is a sample which is modified slightly with the aim of being misclassified by the classifier. These modifications can be subtle and may not be detected by a human eye.

We determine the performance of our model against adversarial examples generated by AWGN with zero mean and variance of 0.0001 to 0.003. The results are shown in Fig. 5 over all the four datsets. The average Top-1 accuracy across all four datasets is 45.20% for 3D-CNN, 40.50% for ASTPN [21], 35.25% for IDE [22] and 28.25% for RCN [6]. Since 3D-CNN exploits temporal information as well, it may be relatively less effected by the additive spatial noise, thereby leading to better accuracy. We also observe that our model works well in case of larger datasets iLIDS-VID and MARS. This can be explained by the fact that more data leads to better training.

VI. CONCLUSION

In this paper, we propose deep 3D-CNN based approach for learning robust spatio-temporal features for video based person re-identification. We use multiple domain data to overcome the need of large training data. Our experiments on PRID-2011, iLIDS-VID, MARS and an aerial dataset show a significant improvement of 4% or more on an average over other popular

algorithms. We also analyse the performance by incorporating a metric learning method, XQDA, and observe that the Top-1 accuracy shoots up by an average of 10% on all the four datasets. In addition, we empirically show that the proposed network is more robust to adversarial examples as compared to other algorithms.

In future, we plan to apply this model on more datasets. We also plan to explore various other distance metric learning methods and Siamese recurrent networks for computing similarity between feature vectors.

REFERENCES

[1] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in CVPR , 2015, pp. 2197-2206.

[2] L. Wu, Y. Wang, J. Gao, and X. Li, Deep adaptive feature embedding with local sample distributions for person re-identification,Pattern Recognition, vol. 73, pp. 275-288, 2018.

[3] Y.C. Chen, X. Zhu, W.S. Zheng, and J.-H. Lai, Person re-identification by camera correlation aware feature augmentation, IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 2, pp. 392-408, 2018.

[4] L. Zheng, Y. Yang, and A. G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv:1610.02984, 2016.

[5] A. Bedagkar Gala and S. K. Shah, A survey of approaches and trends in person re-identification, Image and Vision Computing, vol. 32, no. 4, pp. 270-286, 2014.

[6] N. McLaughlin, J. Martinez del Rincon, and P. Miller, Recurrent convolutional network for video-based person re-identification, in CVPR , 2016, pp. 1325-1334.

[7] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850, 2013.

[8] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, A siamese long short-term memory architecture for human re-identification, in ECCV, 2016, pp. 135-153.

[9] 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 preprint arXiv:1606.01595, 2016.

[10] J. Dai, P. Zhang, H. Lu, and H. Wang, Video person re-identification by temporal residual learning, arXiv preprint arXiv:1802.07918, 2018.

[11] T. Xiao, H. Li, W. Ouyang, and X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in CVPR, 2016, pp. 1249-1258.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, Large-scale video classification with convolutional neural networks, in CVPR, 2014, pp. 1725-1732.

[13] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, Learning a mahalanobis metric from equivalence constraints, Journal of Machine Learning Research, vol. 6, no. Jun, pp. 937-965, 2005.

[14] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof, Relaxed pairwise learned metric for person re-identification, in ECCV, 2012, pp. 780793.

[15] D. Yi, Z. Lei, and S. Z. Li, Deep metric learning for practical person re-identification, ICPR, 2014, pp.34-39.

[16] K. Q. Weinberger and L. K. Saul, Distance metric learning for large margin nearest neighbor classification,Journal of Machine Learning Research, vol. 10, no. Feb, pp. 207-244, 2009.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in ICCV, 2015, pp. 4489-4497.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, ICLR Workshop, 2014.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, Caffe: Convolutional architecture for fast feature embedding, ACM Multimedia, 2014, pp. 675-678.

[20] D. Chung, K. Tahboub, and E. J. Delp, A two stream siamese convolutional neural network for person re-identification, in ICCV , 2017, pp. 1983-1991.

[21] 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, ICCV, 2017.

[22] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, Mars: A video benchmark for large-scale person re-identification, in ECCV, 2016, pp. 868-884.

[23] X. Zhu, X.Y. Jing, F. Wu, and H. Feng, Video-based person re-identification by simultaneously learning intra-video and intervideo distance metrics. in IJCAI, 2016, pp. 3552-3559.

[24] K. Liu, B. Ma, W. Zhang, and R. Huang, A spatio-temporal appearance representation for viceo-based pedestrian re-identification, in ICCV, 2015, pp. 3810-3818.

[25] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, Person re-identification via recurrent feature aggregation, in ECCV, 2016, pp. 701-716.

[26] J. You, A. Wu, X. Li, and W.-S. Zheng, Top-push video-based person re-identification, CVPR, 2016, pp. 1345-1353.

[27] 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, pp. 91-102.

[28] T. Wang, S. Gong, X. Zhu, and S. Wang, Person re-identification by video ranking, in ECCV, 2014, pp. 688-703.