


Multimedia Tools and Applications (2020) 79:24537–24551
<https://doi.org/10.1007/s11042-020-09096-x>

Person Re-identification in Videos by Analyzing Spatio-temporal Tubes



Arif Ahmed Sekh¹  · Debi Prosad Dogra² · Heeseung Choi³ · Seungho Chae³ · Ig-Jae Kim³

Received: 22 October 2019 / Revised: 7 May 2020 / Accepted: 22 May 2020 /

Published online: 23 June 2020

© The Author(s) 2020

Abstract

Typical person re-identification frameworks search for k best matches in a gallery of images that are often collected in varying conditions. The gallery usually contains image sequences for video re-identification applications. However, such a process is time consuming as video re-identification involves carrying out the matching process multiple times. In this paper, we propose a new method that extracts spatio-temporal frame sequences or tubes of moving persons and performs the re-identification in quick time. Initially, we apply a binary classifier to remove noisy images from the input query tube. In the next step, we use a key-pose detection-based query minimization technique. Finally, a hierarchical re-identification framework is proposed and used to rank the output tubes. Experiments with publicly available video re-identification datasets reveal that our framework is better than existing methods. It ranks the tubes with an average increase in the CMC accuracy of 6–8% across multiple datasets. Also, our method significantly reduces the number of false positives. A new video re-identification dataset, named Tube-based Re-identification Video Dataset (TRiViD), has been prepared with an aim to help the re-identification research community.

Keywords Video-based Person Re-identification · Re-ranking · Person Re-identification

1 Introduction

Person re-identification (Re-Id) is useful in various intelligent video surveillance applications. The process can be considered as image retrieval problem, where a query image of a person (probe) is given and we search the person in a set of images extracted from different cameras (gallery). The task is difficult for various reasons. Firstly, face-based [24] and body movement-based identification [2] cannot be used due to the variations in CCTV camera positions. Secondly, complex nature of similarity measure and pose matching makes it

✉ Arif Ahmed Sekh
skarifahmed@gmail.com

harder. Recent advancement in object tracking [4] has opened up new possibilities. Video object trackers can be used to track people in real-time. These tracks containing humans can be passed to a ML framework to search for identification in other cameras. The query can be a single image [25] or multiple images [9]. Often multi-image query uses early fusion and generate an average query image [29]. The method thus consumes higher computational power as compared to single image-based methods. Video-based re-identification research is still evolving [6, 18]. Existing algorithms are sensitive to the query images or video segment. Choosing an improper image or video segment may lead to poor retrieval [25]. In this paper, we detect and track humans and construct spatio-temporal tubes that are used in the re-identification framework. We also propose a method for selecting an optimum set of key-pose frames. We apply a new learning framework to re-identify persons appearing in other cameras. To accomplish this, we have made the following contributions in this paper:

- We propose a learning-based method to select an optimum set of key-pose frames to reconstruct a query tube by minimizing its length.
- We propose a new hierarchical video Re-Id framework using detection, self-similarity matching, and temporal correlation that can be integrated with any image-based re-identification framework.
- We introduce a new video dataset, named Tube-based Re-identification Video Dataset (TRiViD) that has been prepared with an aim to help the re-identification research community.

Rest of the paper is organized as follows. In Section 2, we discuss the state-of-the-art of person re-identification research. Section 3 presents the proposed Re-Id framework with various components. Experiment results are presented in Section 4. Conclusion and future work are presented in Section 5.

2 Related work

Person re-identification applications are growing rapidly in numbers. However, the primary challenges are to handle large volume of data [33, 34], tracking in complex environment [21, 35], presence of group [7], occlusion [12], varying pose and style across different cameras [9, 17, 23, 36], etc. The process of Re-Id can be categorized as image-guided [1, 5, 7, 9] and video-guided [6, 8, 18, 28, 31]. The image-guided methods typically use deep neural networks for feature representation and re-identification, whereas the video-guided methods typically use recurrent convolutional networks (RNN) to embed the temporal information such as optical flow [8], sequence of pose, etc. Advancement of hardware and AI techniques, often re-identification tasks are solved using deep learning. In this area of research, Zhao et al. [32]. Liu et al. [15] have used saliency and attention-based learning to compute similarity among persons, Liu et al. [16] have proposed motion-based learning and Xu et al. [29] have proposed jointly learning of image and motion feature. It may be noted that the temporal information such as motion can be a good feature for re-identification. Chen et al. [6] and Zhang et al. [31] have used video sequence-based learning methods. Multiple information fusion-based methods have also been presented by various researchers. Chen et al. [7] have used fusion of local and global feature. Chung et al. [8] have proposed weighted fusion of spatio-temporal features. Considering the pose information, Zhong et al. [36] have used style transfer to learn similarity matching and Liu et al. [17] have augmented the pose to generate training data. Recently, late fusion of different scores [1, 20] has shown significant improvement over the final ranking. Our method is similar to a typical delayed

or late fusion guided method. We refine search results obtained using convolutional neural networks with the help of temporal correlation analysis.

3 Proposed approach

Our method can be regarded as tracking followed by re-identification. Moving persons are first tracked using Simple Online Deep Tracking (SODT) [3]. A tube is defined as the sequence of spatio-temporal frames. A gallery (G) contains a set of tubes $\{T_1, T_2, \dots, T_n\}$, where T_i represents the i^{th} tube. A tube (T) is created by arranging the image frames $\{I_0, I_1, \dots, I_k\}$ with respect to time.

The re-identification of a person in videos now can be defined by: “Given a query tube of a person, say T_q , we need to find out the tubes in other cameras that are likely to contain the queried person.”

First, the noisy frames are eliminated and the query tube is minimized. Next, the minimized query tube is passed through a 3-stage hierarchical re-ranking process to get the final ranking of the tubes in the gallery. The method is depicted in Fig. 1.

3.1 Simple Online Deep Tracking (SODT)

The tubes of the persons are extracted by tracking individuals in the video sequences. The tracking is a two-step process (i) detection and (ii) continuous track handling. The detection has been developed using YOLO [13] framework. The outputs of YOLO are the bounding boxes marking the persons in all frames. The model is pre-trained on ImageNet [22]. Next, the bounding boxes are linked as a sequence by preserving the identity (track number). Each track is represented by unique motion model. A liner constant motion model is used to estimate the inter-frame displacements of each object. The state of the target is modeled using using (1).

$$x = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T \tag{1}$$

where u and v are velocities at the center of the target’s bounding box along horizontal and vertical axes, s is the scale, and r represents aspect ratio of the bounding box. A detection from YOLO is associated with a target, and the initial bounding box is used to update the target’s state. The velocity components are solved using a Kalman filter framework [27].

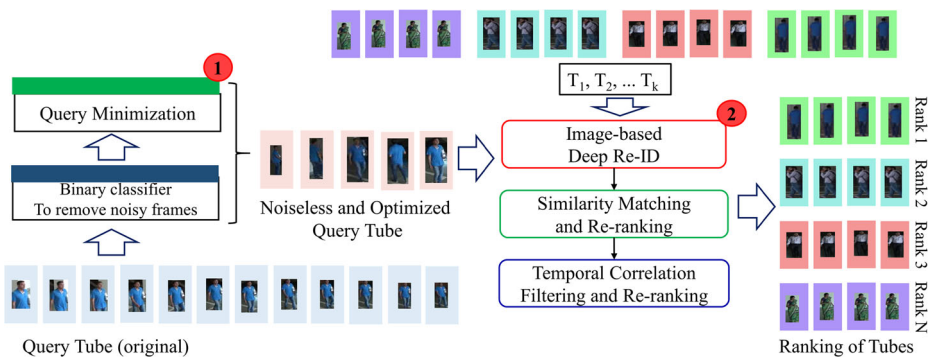


Fig. 1 The proposed method for Tube-to-tube Re-identification. The method takes a tube as query and ranks the tubes by best possible matching

The missing frames are updated by the prediction using the linear velocity model. The track assignment problem among the Kalman filter prediction and newly arrived detection are solved by Hungarian algorithm. We use the (squared) Mahalanobis distance to estimate the distance and optimum association of tracks. Finally, continuously detected bounding boxes preserving identity are included as the frames of a tube.

3.2 Query minimization

Selecting a set of frames that can uniquely represent a tube can be challenging. To address this, we have used a deep similarity matching architecture to select a set of representative frames based on pose dissimilarity. First, a query tube is passed through a binary classifier to remove noisy frames such as blurry, cropped, low-quality, etc. A ResNet [14] framework has been trained using a few query tubes containing similarly looking images. An energy function as discussed in (4) has been used to select a set of unique frames from the query tube (T_q). The two components of the energy function (ζ and γ) take into account the minimal impact of the closeness of each pair of frames and the maximal impact of the differences between each pair of frames in the query tube.

Overall closeness or similarity index of a frame, say i , is estimated as given in (2), where $\sigma(i, j)$ is a measure to quantify the similarity between two frames i and j in T_q .

$$\zeta_i = \min(\sigma(i, j)), \forall j \in T_q \quad (2)$$

Overall dissimilarity index of a frame, say i , is estimated as given in (3), where $\sigma(i, j)$ is a measure to quantify the similarity between two frames i and $j \in T_q$.

$$\gamma_i = \max(\sigma(i, j)), \forall j \in T_q \quad (3)$$

We now assume that the input tube contains k images and the output query tube contains l images such that $l \ll k$. Our objective is to choose query images from a tube such that most dissimilar images are taken and similar images are discarded. The optimal query energy (E) is defined in (4), where \hat{Q} is the set of images that are not included in the optimal query (Q) and ϕ is a weighting parameter. Increasing the weigh (ϕ) also increases number of images in the query.

$$E = \sum_{i \in Q} \phi \xi_i + \sum_{i \in Q, i \notin \hat{Q}} \gamma_i \quad (4)$$

3.3 Proposed Re-identification and ranking

At the beginning, SVDNet [25] has been used for re-identification at image-level. The network takes 128×64 images as input and produces a set of retrieved images. The outputs are then passed through late fusion layers for re-ranking the retrieved images. Figure 2 illustrates the process. Assume the set of retrieved images is denoted by T_{SVD} as given in (5). In the next stage, a network similar to ResNet50 has been trained to learn the self-similarity scores using the tubes of the query set.

$$T_{SVD} = \{I_1, I_2, \dots, I_p\} \quad (5)$$

During this process, similarity scores between every output image of SVD network up to rank p are calculated using a self similarity estimation layer. It learns to measure self-similarity using tracked tubes during training assuming single track contains images of the same person. We use ResNet50 [10] as the baseline. It takes a set of ranked images (SVDNet

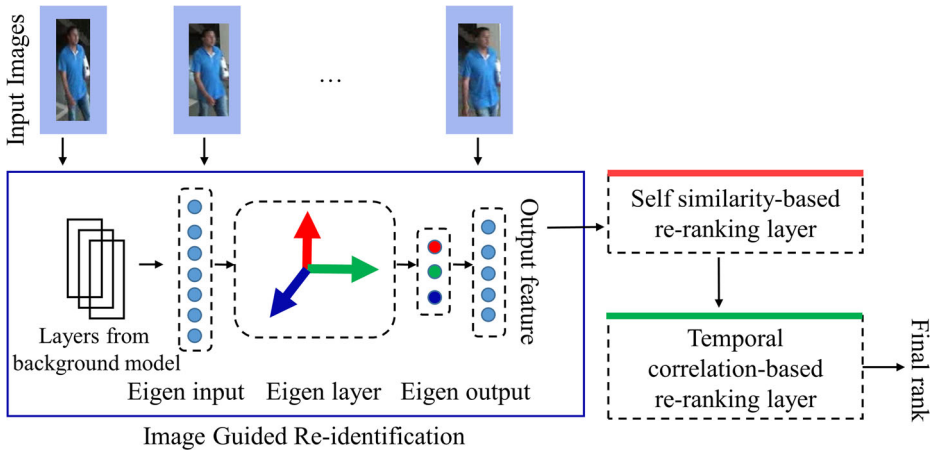


Fig. 2 The proposed re-ranking scheme adopted in our work

outputs) and produces a set of ranked images by introducing self-similarities between the retrieved images.

Finally, the scores are averaged and the images are re-ranked. This step ensures that the dissimilar images get pushed toward the end of the ranked sequence of the retrieved images.

3.4 Tube ranking by temporal correlation

Final step of the proposed method is to rank the tubes by temporal correlation among the retrieved images. Let the result matrix up to rank p ($p \leq k$) for the query tube T_q after the first two stages be given in (R) . Weight of an image (I_{jk}) in R is estimated using (7).

$$R = \begin{bmatrix} I_{11} & I_{12} & I_{13} & \dots & I_{1p} \\ I_{21} & I_{22} & I_{23} & \dots & I_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ I_{j1} & I_{j2} & I_{j3} & \dots & I_{jp} \end{bmatrix} \tag{6}$$

$$\alpha = \frac{1}{I_r}, \text{ where } I_r \text{ is the rank of the frame } I \tag{7}$$

Similarly, weight of a tube (T) is denoted by β and estimated using (8).

$$\beta = \frac{\# \text{ of images in } T \cap R}{\max(\# \text{ of images in } T \cap R), \forall T} \tag{8}$$

Finally, the temporal correlation cost (τ_I) of an image in R can be estimated as given in (9).

$$\tau_I = \alpha \times \beta, \text{ such that } I \in T \tag{9}$$

Based on the temporal correlation, the retrieved tubes are ranked, where higher rank tubes have higher weights. The final ranked images are extracted by taking the highest scoring images from the tubes. Figure 3 explains the whole process of tube ranking and selection

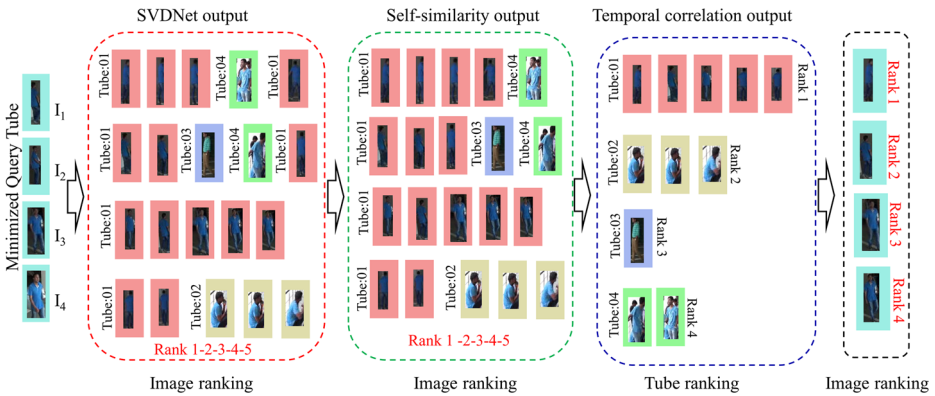


Fig. 3 Explanation of re-identification framework with the help of the proposed 3-stage framework depicted in Fig. 1

of final set of frames. The main motivation of the temporal correlation is to assign higher weights of the tubes containing a maximum number of retrieved images.

4 Experiments

We have evaluated our proposed approach on two public datasets, iLIDS-VID [26] and PRID-11 [11] that are often used for testing video-based re-identification frameworks. In addition to that, we have also prepared a new re-identification dataset. It has been recorded using 2 cameras in an indoor environment with human movements with moderately dense crowd (with more than 10 people appearing within 4-6 sq-mt), varying camera angles, and persons with similar clothing. Such situations have not been covered yet in existing re-identification video datasets. Details about these datasets are presented in Table 1. Several experiments have been carried out to validate our method. A through comparative analysis has also been performed.

Evaluation Metrics and Strategy: We have followed the well known experimental protocols for evaluating the method. For iLIDS-VID and TRiViD dataset videos, the tubes are randomly split into 50% for training and 50% for testing. For PRID-11, we have followed the experimental setup as proposed in [6, 19, 26, 29, 37]. Only first 200 persons

Table 1 Dataset used in our experiments. Only TRiViD dataset is tracked to extract tube

Dataset	Number of cameras	Persons re-appeared	Number of tubes in gallery	Challenges
PRID-11 [11]	2	245	475	Large volume
iLIDS-VID [26]	2	119	300	Clothing Similarity
TRiViD	2	47	342	Dense, Tracking, Similarity

In other dataset the given sequence of images are considered as tube

who appeared in both cameras of the PRID-11 dataset, have been used in our experiments. A 10-folds cross validation scheme has been adopted and the average results are reported. We have prepared Cumulative Matching Characteristics (CMC) and mean average precision (mAP) curves to evaluate and compare the performance.

4.1 Comparative analysis

As per the state-of-the-art, our work though unique in design has some similarities with video re-id methods proposed in [19, 30], multiple query-based method [25], and the re-ranking method [20]. The RCNN-based method [19] uses image level CNN and optical flow for learning and searching. Video-based feature learning method [30] uses sequence-based distance to compute similarity between query and gallery. SVDNet [25] is a typical image-based re-identification framework, it can be used a single image or multiple images as probe. Therefore, we have compared our approach with the above recently proposed methods. It has been observed that our method can achieve a gain up to 9.6% as compared to the state-of-the-art methods when top rank accuracy is estimated. Even if we compute the accuracy up to rank 20, our method performs better with a margin of 3%. This happens because our method tries to reduce the number of false positives which has not yet been addressed by the re-identification research community. Figures 4, 5 and 6 represent CMC curves and Table 2 summarizes the mAP up to rank 20 across the three datasets. Figure 7 shows a typical query and response applied on PRID-11 dataset.

4.2 Computational complexity analysis

Re-identification in real-time is a challenging task. All research work carried out so far presume the gallery as a pre-recorded set of images and they try to rank best 5, 10, 15, 20 images from the set. However, executing a single query takes considerable time when multiple images are involved in the query. We have carried out a comparative analysis on

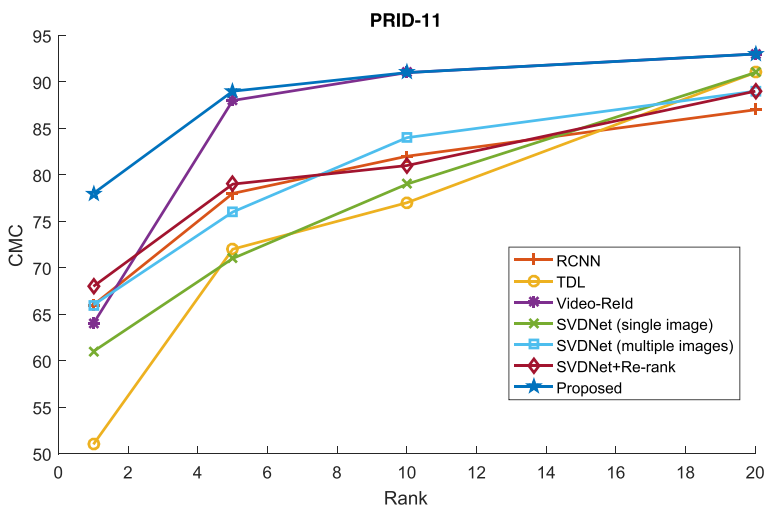


Fig. 4 The accuracy (CMC) in PRID-11 dataset using RCNN [19], TDL [30], Video re-id [19], SVDNet [25] (single image), SVDNet (multiple images), SVDNet+Re-rank [20]

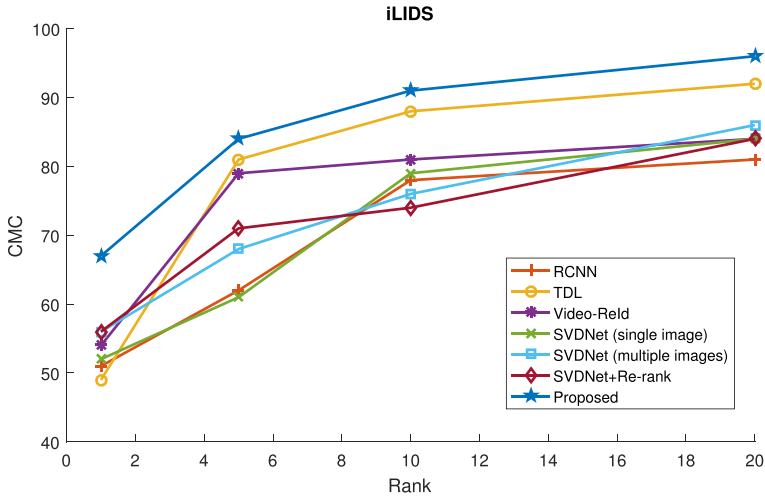


Fig. 5 The accuracy (CMC) in iLIDS dataset using RCNN [19], TDL [30], Video re-id [19], SVDNet [25] (single image), SVDNet (multiple images), SVDNet+Re-rank [20]

computation complexities across various re-identification frameworks including the proposed scheme. One Nvidia Quadro P5000 series GPU has been used to implement the frameworks. The results are reported in Fig. 8. We have observed that the proposed tube-based re-identification framework takes lesser time as compared to video re-id framework proposed in [19] and the multiple images-based re-id using SVDNet [25].

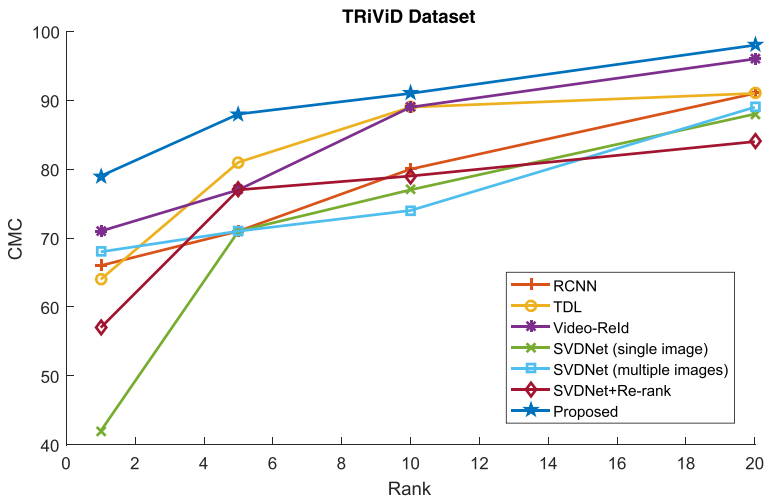


Fig. 6 The accuracy (CMC) using the TRiViD dataset with the help of RCNN [19], TDL [30], Video re-id [19], SVDNet [25] (single image), SVDNet (multiple images), SVDNet+Re-rank [20]

Table 2 mAP (%) up to rank 20 in across three video datasets

Method/Dataset	PRID	iLIDS	TRiViD
RCNN [19]	81.2	74.6	79.11
TDL [30]	78.2	74.1	80
Video-ReId [19]	73.31	64.29	83.22
SVD Net [25] (Single Image)	76.44	69	79.11
SVD Net [25] (Multiple Images)	79.21	66.71	82.66
SVD Net+Re Rank [20]	77.25	69.2	78.6
Proposed	86.17	79.22	91.66

Bold represent best performing results

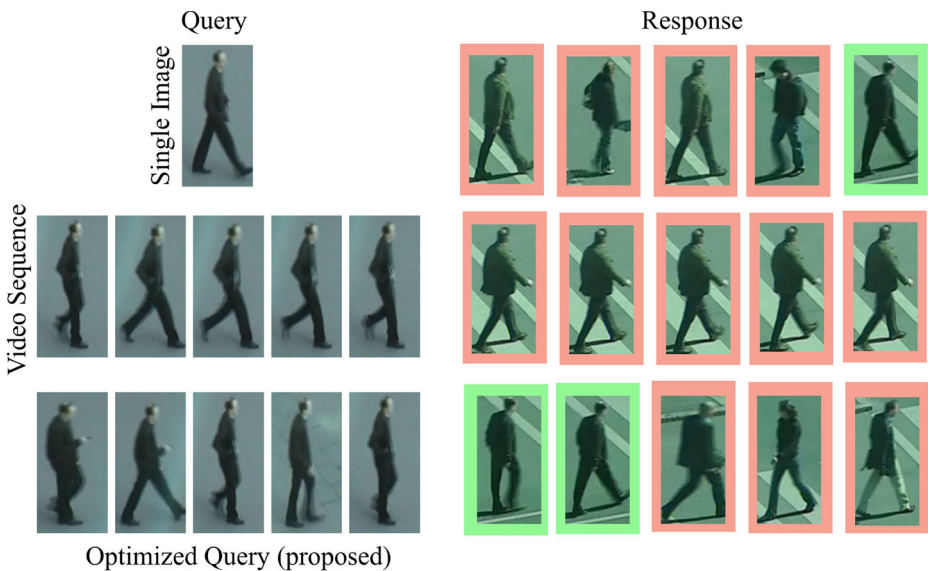


Fig. 7 Typical results obtained using PRID-11 dataset using single image query [25], video sequence [19], and using the proposed method. Green box indicates a correct retrieval

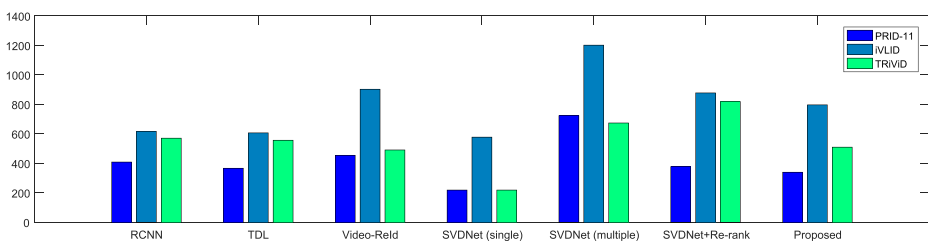


Fig. 8 Average response time (in seconds) for a given SVD query by varying the datasets. We have taken 100 query tubes in random and calculated the average response time using RCNN [19], TDL [30], Video re-id [19], SVDNet [25] (single image), SVDNet (multiple images), SVDNet+Re-rank [20]

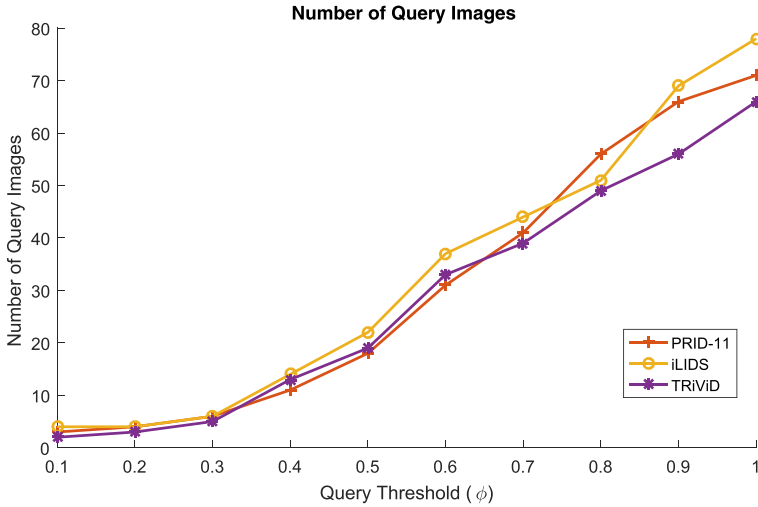


Fig. 9 Average number of query images by varying the query threshold (ϕ). We have taken 100 query sequences randomly and average number of optimized images, is reported. It may be observed that a higher ϕ produces more number of query images

4.3 Effect of ϕ

Our proposed method depends on the query threshold (ϕ). In this section, we present an analysis about the effect of ϕ on results. Figure 9 depicts the average number of query images generated from various query tubes. It may be observed that, higher ϕ produces more query images.

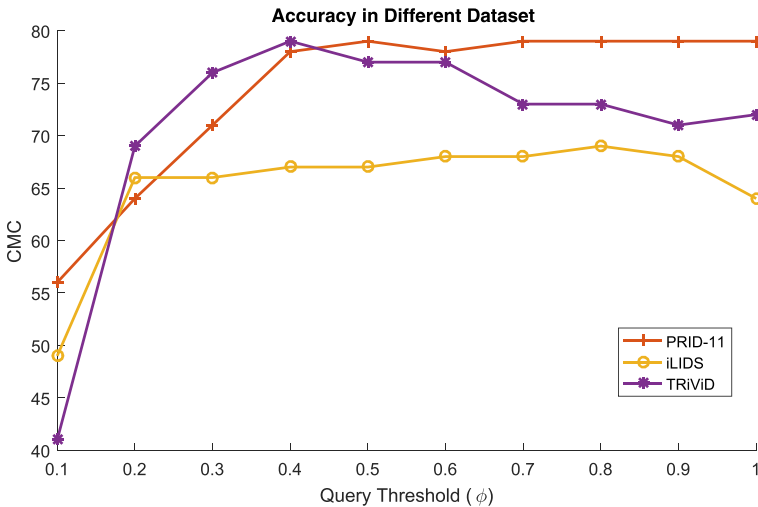


Fig. 10 Accuracy (CMC) by varying the query threshold (ϕ). We have taken 100 query sequences randomly and average is reported. It may be observed that a higher ϕ may not produce higher accuracy

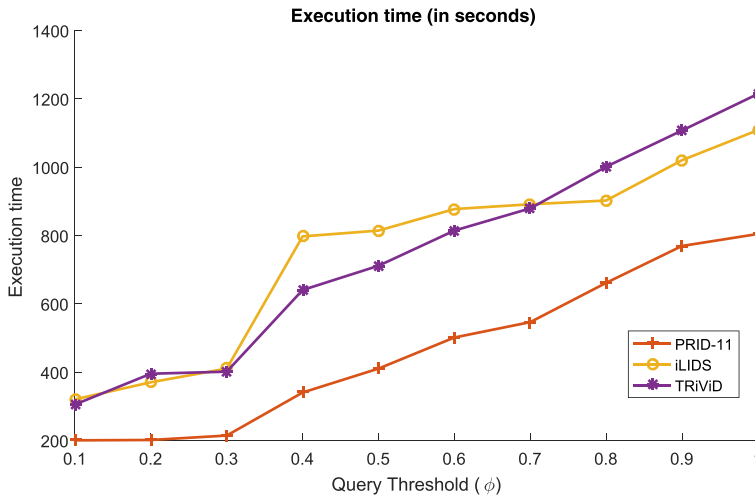


Fig. 11 Execution time by varying ϕ . It may be observed that a higher ϕ takes more time to execute as it produces more query images.

Figure 10 depicts average CMC by varying ϕ . It may be observed that the accuracy does not increase significantly when ϕ is increased above 0.4.

Figure 11 presents execution time (in seconds) by varying the query threshold. It can also be observed that an increase in ϕ leads to higher response time. Therefore, we have used $\phi = 0.4$ in our experiments.

4.4 Results after various stages

In this section, we present the effect of various stages of the overall framework on re-identification results. Table 3 shows the accuracy (CMC) in each step of the proposed method. It may be observed that the proposed method gains 11% rank-1 accuracy after the first stage and 7% rank-1 accuracy after the second step. The method gains 7% rank-20 accuracy in the first stage and 6% rank-20 accuracy after the second stage. Figure 12 shows an example of scores (true positives and false positives) during the self-similarity fusion. It may be observed that SVDNet output scores and similarity scores are high in case of true positives. Similarity scores are relatively low in case of false positives. More results can be found in the form of supplementary data.

Table 3 Accuracy (CMC) in each step of the proposed method

	PRID11 [11]				iLIDS [26]				TRiViD			
Method/Top Rank	1	5	10	20	1	5	10	20	1	5	10	20
SVD Net (Multi Image)	66	76	84	89	56	68	76	86	68	71	74	89
SVD Net+Self-similarity	69	77	84	89	61	71	79	86	71	77	76	91
SVD Net+Self-similarity+ Temporal Correlation (Proposed)	78	89	92	91	67	84	91	96	79	88	91	98

Bold represent best performing results



Fig. 12 Typical examples of failure cases using SVDNet + Self Similarity in TRiViD (first two rows) and PRID-11 [11] (last row)

5 Conclusion

In this paper, we propose a new person re-identification framework that is able to outperform existing re-identification schemes when applied on videos or sequence of frames. The method uses any CNN-based framework at the beginning (we have considered SVDNet). A self-similarity layer is used to refine the SVDNet scores. Finally, a temporal correlation layer is used to aggregate multiple query outputs and to match tubes. A query optimization has also been proposed to select an optimum set of images for a query tube. Our study reveals that the proposed method outperforms in several cases as compared to the state-of-the-art single image-based, multiple images-based, and video-based re-identification methods. The computational is also reasonably low. It can be noted that the method can rank the tubes with an average increase in the CMC accuracy of 6–8% across multiple datasets. Also, our method significantly reduces the number of false positives.

One straight extension of the present work is to fuse methods like camera pose-based [9], video-based [19], and description-based [5]. It may lead to higher accuracy in complex situations. Also, group re-identification can be tried with the similar concept of tube guided analysis.

Funding Information Open Access funding provided by UiT The Arctic University of Norway. The work has been funded under KIST Flagship Project (Project No.2E30270) executed at IIT Bhubaneswar under the Project Code: CP152. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P5000 GPU used for this research.

Compliance with Ethical Standards

Conflict of interests The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors. Informed consent: Informed consent was obtained from all individual participants included in the study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Barman A, Shah SK (2017) Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems. In: International conference on computer vision. IEEE, pp 1124–1133
2. Batchuluun G, Naqvi RA, Kim W, Park KR (2018) Body-movement-based human identification using convolutional neural network. *Expert Syst Appl* 101:56–77
3. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B (2016) Simple online and realtime tracking. In: 2016 IEEE International conference on image processing (ICIP). IEEE, pp 3464–3468
4. Cancela B, Ortega M, Fernández A, Penedo MG (2013) Hierarchical framework for robust and fast multiple-target tracking in surveillance scenarios. *Comput Vis Pattern Recogn* 40(4):1116–1131
5. Chang X, Hospedales TM, Xiang T (2018) Multi-level factorisation net for person re-identification. In: *Computer vision and pattern recognition*, vol 1, pp 2
6. Chen D, Li H, Xiao T, Yi S, Wang X (2018a) Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: *Computer vision and pattern recognition*, pp 1169–1178
7. Chen D, Xu D, Li H, Sebe N, Wang X (2018b) Group consistent similarity learning via deep crf for person re-identification. In: *Computer vision and pattern recognition*, pp 8649–8658
8. Chung D, Tahboub K, Delp EJ (2017) A two stream siamese convolutional neural network for person re-identification. In: *International conference on computer vision*
9. Deng W, Zheng L, Kang G, Yi Y, Ye Q, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In: *Computer vision and pattern recognition*, vol 1, pp 6
10. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Computer vision and pattern recognition*, pp 770–778
11. Hirzer M, Belezni C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: *Scandinavian conference on image analysis*. Springer, pp 91–102
12. Huang H, Li D, Zhang Z, Chen X, Huang K (2018) Adversarially occluded samples for person re-identification. In: *Computer vision and pattern recognition*, pp 5098–5107
13. Jensen MB, Nasrollahi K, Moeslund TB (2017) Evaluating state-of-the-art object detector on challenging traffic light data. In: *Computer vision and pattern recognition*. IEEE, pp 882–888
14. Lin H, Jegelka S (2018) Resnet with one-neuron hidden layers is a universal approximator. In: *Advances in neural information processing systems*, pp 6169–6178
15. Liu H, Feng J, Qi M, Jiang J, Yan S (2017a) End-to-end comparative attention networks for person re-identification. *IEEE Trans Image Process* 26(7):3492–3506
16. Liu Z, Wang D, Lu H (2017b) Stepwise metric promotion for unsupervised video person re-identification. In: *International Conference on Computer Vision*. IEEE, pp 2448–2457
17. Liu J, Ni B, Yan Y, Zhou P, Cheng S, Hu J (2018) Pose transferrable person re-identification. In: *Computer vision and pattern recognition*, pp 4099–4108
18. Lv J, Chen W, Li Q, Yang C (2018) Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: *Computer vision and pattern recognition*, pp 7948–7956
19. McLaughlin N, del Rincon JM, Miller P (2016) Recurrent convolutional network for video-based person re-identification. In: *Computer vision and pattern recognition*, pp 1325–1334

20. Paisitkriangkrai S, Shen C, Van Den Hengel A (2015) Learning to rank in person re-identification with metric ensembles. In: *Computer vision and pattern recognition*, pp 1846–1855
21. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: *European conference on computer vision*. Springer, pp 17–35
22. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252
23. Saquib Sarfraz M, Schumann A, Eberle A, Stiefelhagen R (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: *Computer vision and pattern recognition*
24. Soleymani R, Granger E, Fumera G (2018) Progressive boosting for class imbalance and its application to face re-identification. *Expert Syst Appl* 101:271–291
25. Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: *International conference on computer vision*. IEEE, pp 3820–3828
26. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: *European conference on computer vision*. Springer, pp 688–703
27. Weng S-K, Kuo C-M, Tu S-K (2006) Video object tracking using adaptive kalman filter. *J Vis Commun Image Represent* 17(6):1190–1208
28. Wu Y, Lin Y, Dong X, Yan Y, Ouyang W, Yang Y (2018) Exploit the unknown gradually One-shot video-based person re-identification by stepwise learning. In: *Computer vision and pattern recognition*, pp 5177–5186
29. Xu S, Yu C, Gu K, Yang Y, Chang S, Zhou P (2017) Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: *International conference on computer vision*. IEEE, pp 4743–4752
30. You J, Wu A, Li X, Zheng W-S (2016) Top-push video-based person re-identification. In: *Computer vision and pattern recognition*, pp 1345–1353
31. Zhang J, Wang N, Zhang L (2018) Multi-shot pedestrian re-identification via sequential decision making. In: *Computer vision and pattern recognition*
32. Zhao R, Oyang W, Wang X (2017) Person re-identification by saliency learning. *IEEE Trans Pattern Anal Mach Intell* 39(2):356–370
33. Zheng L, Bie Z, Sun Y, Wang J, Su C, Wang S, Qi T (2016) Mars a video benchmark for large-scale person re-identification. In: *European conference on computer vision*. Springer, pp 868–884
34. Zheng L, Shen L, Lu T, Wang S, Wang J, Qi T (2015) Scalable person re-identification a benchmark. In: *International conference on computer vision*, pp 1116–1124
35. Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q et al (2017) Person re-identification in the wild. In: *Computer vision and pattern recognition*, vol 1, pp 2
36. Zhong Z, Zheng L, Zheng Z, Li S, Yi Y (2018) Camera style adaptation for person re-identification. In: *Computer vision and pattern recognition*, pp 5157–5166
37. Zhou Z, Huang Y, Wang W, Wang L, Tan T (2017) See the forest for the trees Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: *Computer vision and pattern recognition*. IEEE, pp 6776–6785

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Arif Ahmed Sekh¹  · Debi Prosad Dogra² · Heeseung Choi³ · Seungho Chae³ · Ig-Jae Kim³

Debi Prosad Dogra
dpdogra@iitbbs.ac.in

Heeseung Choi
hschoi@kist.re.kr

Seungho Chae
seungho.chae@kist.re.kr

Ig-Jae Kim
drjay@kist.re.kr

¹ UiT The Arctic University of Norway, Tromsø, Norway

² Indian Institute of Technology Bhubaneswar, Bhubaneswar, India

³ Korea Institute of Science and Technology, Seoul, South Korea