Inteligencia Artificial 26(71), 25-33
doi: 10.4114/intartif.vol26iss71pp25-33

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Person Re-Identification by Siamese Network

Newlin Shebiah R*, Arivazhagan S, Amrith S G & Adarsh S

Centre for Image Processing and Pattern Recognition, Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India
*newlinshebiah@mepcoeng.ac.in

Abstract The re-identification of individuals aims to retrieve persons across multiple non-overlapping cameras. With the advancement of deep learning features and the increase in the number of surveillance videos, the computer vision community has experienced significant progress. However, person re-identification is still faced with various challenges such as low resolution images and pose variations. To overcome these challenges, state-of-the-art algorithms for person re-identification are supported by convolutional neural networks. This paper proposes the use of a Siamese network, which is a neural architecture that takes a pair of images or videos as input and predicts the similarity or dissimilarity of a person across two cameras. The output includes the prediction of similar and dissimilar persons along with their prediction scores. The proposed method was evaluated using iLIDS-VID and PRID 2011 datasets, and achieved recognition accuracy of 79.52% and 85.82%, respectively. These results demonstrate the effectiveness of the Siamese network for person re-identification tasks. Overall, this study contributes to the ongoing research on improving the accuracy of person re-identification across multiple cameras in surveillance videos.

Keywords: Siamese Network, Person re-identification, Surveillance videos, Deep learning

1 Introduction

In a video surveillance scenario, the task of identifying a person across multiple non-overlapping views captured through cameras deployed in surveillance spaces at the same or different points in time is termed "person re-identification." Human re-identification challenges include changes in scale, location, lighting, viewing angles, partial occlusions, and other variables that can produce considerable variations in a person's appearance across cameras. To overcome these challenges, it is desirable to analyze discriminative features and the most effective classifier ought to be used. With deep learning based models, discriminating feature vectors are learned effectively and with advanced distance or similarity metrics, the research on human re-identification has progressed successfully.

The majority of deep-learning methods used for human re-identification are based on Siamese architecture and are of the following two types: The first type takes image pairs as input and runs them through a series of convolutional layers with shared weights, and output similarity scores are used to determine whether the image pair depicts the same person or not. To reduce contrastive or triplet loss, the second method attempts to extract global characteristics from the top layer. These losses are to reduce positive pair distance and maximize negative pair distance. Siamese Network converts the input human image into a vector of its features. The vector should be comparable to that of the same human image and different from another. In short, the model learns to extract a person distinguishing traits. After obtaining the feature mapping, it can be matched to other person in a database.

ISSN: 1137-3601 (print), 1988-3064 (on-line)

The reason for using Siamese network is as follows: Siamese network is a one-shot classification model that can do prediction with just a single training example. Siamese network is more robust to class imbalance because it requires very little input. It is possible to apply it to a dataset in which just a small number of instances are available for some classes. The one-shot learning characteristic of the Siamese network does not rely on any domain-specific information but instead takes advantage of deep learning techniques. This research creates a Siamese model in which image pairs are analysed using the same CNN framework and weights. Each training step tells the model whether the input pair represents the same person. To summarize, our contributions are:

- Proposal of a Siamese network for human re-identification
- Network takes human images as input and returns their similarity measure
- Simultaneous learning of discriminative feature embedding and similarity metric using multiple CNN backbones
- Improved accuracy of human retrieval through the proposed approach

These contributions demonstrate the potential for the proposed Siamese network to advance the field of human re-identification. The results of this work have implications for the development of more accurate and efficient human tracking and identification systems.

2 Literature Survey

Zheng et al. [1] proposed a consistent Siamese attention network that localizes a person of interest while using a person identification label for supervision. It will contain attention constancy, allowing end-to-end learning of the same person's attentive visuals. CUHK03, Duke MTMC, and Market-1501 have identification accuracy of 71.5, 87.7, and 94.4 percent, respectively. Chung et al. [2] suggested a convolutional neural network with two Siamese networks. They proposed a two-stream training objective function that combines geographical and temporal costs to guess the same person. The author reports 98% and 96% accuracy for datasets PRID 2011 and iLIDS-VID. Wang et al. [3] presented bilinear channel fusion to reduce Resnet-20's bottleneck and fine-grained information to improve pedestrian feature robustness. Hard sample selection eliminates hard samples. Market-1501 and CUHK03 have 92.5 and 94.3 percent recognition accuracy, respectively.

Shen et al. [4] presented a Siamese model for applying distinct distance metrics to feature maps and a multi-level similarity perceptron method to increase performance. Multitask architecture was employed to optimise classification and similarity restrictions and the time-consuming process of matching input and extracting image features. CUHK03, CUHK01, and Market-1501 have recognition accuracy of 83.6%, 96.9%, and 81.9%, respectively. Khatun et al. [5] developed a four-stream Siamese deep convolutional neural network for human re-detection. It solves the constraints of standard triplet information by using four images, two of which are matched to raise inter-class variations and two mismatched to reduce intra-class variations. The author indicates that VIPeR, CUHK01, CUHK03, and PRID 2011 have recognition accuracy of 68.7%, 83.95 %, 85.5%, and 75%, respectively. Sun et al. [6] proposed a conditional transfer network (cTransNet) to transfer images to the perspective with the largest domain gap (GAN). Merging original image data with transferred image features and computing cosine distance to rank similarity creates hybrid person representation. The author's recognition accuracy for Market-1501, Duke MTMC, and MSMT17 is 91.1%, 85.1%, and 67.7%, respectively.

Pang et al. [7] presented a residual layer that learns high-level image attributes using "conv" and "identity" blocks. Global average pooling reduces model complexity and retrieval time in edge computing. The author shows that CUHK03 and Market-1501 have 84.6 and 80.5% recognition accuracy. Wu et al. [8] suggested a Siamese network architecture for video re-id that learns features and metrics simultaneously. The attention system selects the most important aspects and learns to focus on a particular region in the cross view, which helps re-identify the individual in stressful situations. iLIDS-VID, PRID 2011, and MARS had 61.2, 74.8, and 83.4 percent recognition accuracy, respectively. Song et al. [9] suggested a context interaction convolution neural network (CI-CNN) to address cross-scenario person re-identification by including a novel actor – critic agent in reinforcement learning embedded CNN that provides dynamic context and a telescoping perception field. By integrating a context policy network and a context critic network, a deep reinforcement-based learning multitask framework to learn context-agent was created. The author reports recognition accuracy of 87.49, 68.7, 94.2, 87.3, 87.6, and 71.23 for PRID 2011, PRW, MARS, Market-1501, DukeMTMC, and iLIDS-VID, respectively.

Zhang et al. [10] propose the HRNet-ReID feature extraction network, which combines the native HRNet-W32 backbone with a unique representation head to modify HRNet for individual person re-id. PS-HRNet blends

VDSR-CA and HRNet-ReID to explore feature space at a deeper level, eliminating the distribution gap between LR and HR image features and solving the cross resolution person ReID problem. The author reports recognition accuracy of 91.5%, 92.6%, 48.7%, MLR-DukeMTMC, and CAVIAR. Dual Attention Matching Network (DuATM) is an end-to-end trainable architecture that learns context-aware sequences and compares attention sequences. It uses intra-sequence and inter-sequence attention to improve and align features. The author achieves 91.42, 81.82, and 78.74% recognition accuracy with Market-1501, DukeMTMC, and MARS. Ke et al. [12] introduced an ID-adaption network to transform IDE characteristics to a common discriminative latent space. In this network, the representation of the training identities is enforced to adapt unobserved training identities. Take the probability distribution as a moment sequence and use core moments to calculate disagreement. Using Market-1501, CUHK03, and DukeMTMC, the author shows recognition accuracy of 81.59, 30, and 67.77%.

Mansouri et al. [13] used a Siamese convolutional neural network and a k-reciprocal nearest neighbour ranking method to compare two imagegraphs of the same individual. Combine S-CNN similarity with jaccard distance to revise the list. Market-1501 and DukeMTMC datasets had 82.42 and 74.28 percent recognition accuracy, respectively. Li et al. [14] used LSTM to learn the dependent relationship between pedestrian image block features and obtain feature representations via memory coding. Combining triplet loss and softmax recognition loss reduces intra-class and inter-class differences. The author shows 84.5, 89.3, and 95.2 percent recognition accuracy for PRID 2011, CUHK03, and Market-1501. Amouei et al. [15] developed a Siamese network to improve human re-identification. Siamese network uses its pre-trained model to extract features from image pairings. The author achieves 95.2% accuracy with CUHK01.For vehicle re-identification the features and methodology proposed by Anandhalli et al. [16-18] can be used. The paper [19] addresses the problem of potholes on Indian roads, which is a major cause of accidents and traffic congestion in the country. The proposed approach uses a CNN-based model, which is trained using a large dataset of images of Indian roads with and without potholes. The model uses an anchor-based approach to improve the accuracy of pothole detection. The anchor-based approach involves dividing the image into multiple regions and identifying the regions that contain potholes. The proposed approach achieves an accuracy of 95.6% in pothole detection, which is significantly higher than the accuracy achieved by existing methods. The paper also provides a detailed analysis of the performance of the proposed approach and compares it with existing methods. The proposed approach has the potential to improve road safety and reduce traffic congestion in India by enabling early detection and repair of potholes.

3 Proposed Work

Siamese networks are very helpful in situations in which there are a high number of classes but only a few observations of each. In situations like these, there is insufficient data to train a deep convolutional neural network to properly categorise images into the aforementioned categories. The Siamese network, on the other hand, is able to determine whether or not two images belong to the same class. For the purpose of training the network, the data must first be organized into pairs of images that are either identical or different from one another. In this context, image graphs that are similar have the same name since they depict distinct occurrences of the same character, whereas images that are dissimilar depict distinct individuals and have unique labels. For the purpose of the analysis, either a randomly selected batch of images or paired images is chosen, or the function then returns the label pairLabel. This label indicates whether the pair of images is similar to or distinct from one another. pairLabel is set to 1 for image pairs that are similar, whereas pairLabel is set to 0 for image pairs that are different.

When comparing two images, each image is run through one of two identical subnetworks that share their weights. This allows the images to be compared side-by-side. The M-by-N images are each converted to a feature vector of 4096 dimensions by the subnetworks. Images belonging to the same class have equivalent representations in all 4096 dimensions. After performing a subtraction in order to merge the output feature vectors from each subnetwork, the resulting vectors are then fed into a fully-connected operation that only produces a single output. This value is then converted using a sigmoid operation into a probability that falls between 0 and 1, which represents the network's prediction of whether or not the images are comparable to one another. During training, the neural network is kept up to date by using the binary cross-entropy loss that exists between the network prediction and the actual label. Table 1 has a detailed description of the architecture as well as the activations.

The modelLoss function requires the Siamese subnetwork net, the parameter structure for the fullyconnect operation, and a mini-batch of input data *X1* and *X2* with their labels pairLabels. It then performs the modelLoss calculation. This function provides the loss values as well as the loss gradients with regard to the learnable parameters of the network. The Siamese network's purpose is to differentiate between the two inputs known as X1 and X2 in order to achieve its goal. The output of the network is a probability that ranges from 0 to 1, with a value closer to 0 indicating a prediction that the images are different from one another and a value closer to 1 indicating that the images are comparable to one another. The loss is calculated using the binary cross-entropy between the predicted score and the actual value of the true label. The formula for the loss is as follows:

$$Loss=-t\,\log(y)-(1-t)\,\log(1-y) \tag{1}$$

where t is the actual label, which can either be 0 or 1, and y is the predicted label.

The Processing steps are as follows:

- *Load the datasets iLIDS-VID / PRID 2011*
- *Preprocess the datasets to extract features and remove noise*
- *Divide the datasets into training and testing sets*
- *Define a Siamese network architecture with two input branches, each processing an image or video from a different camera.*
- *Train the Siamese network using the training dataset with the following steps:*
  - *For each pair of images or videos, compute their feature embeddings using the Siamese network.*
  - *Compute the distance or similarity score between the embeddings.*
  - *Compute the loss using a contrastive loss function, which encourages the network to predict similar embeddings for images or videos of the same person and dissimilar embeddings for different persons.*
  - *Backpropagate the loss to update the network weights.*
- *Test the Siamese network using the testing dataset with the following steps:*
  - *For each pair of images or videos, compute their feature embeddings using the trained Siamese network.*
  - *Compute the distance or similarity score between the embeddings.*
  - *Compare the score with a threshold to predict whether the images or videos belong to the same person or different persons.*
- *Evaluate the performance of the Siamese network using metrics such as recognition accuracy, precision, and recall.*

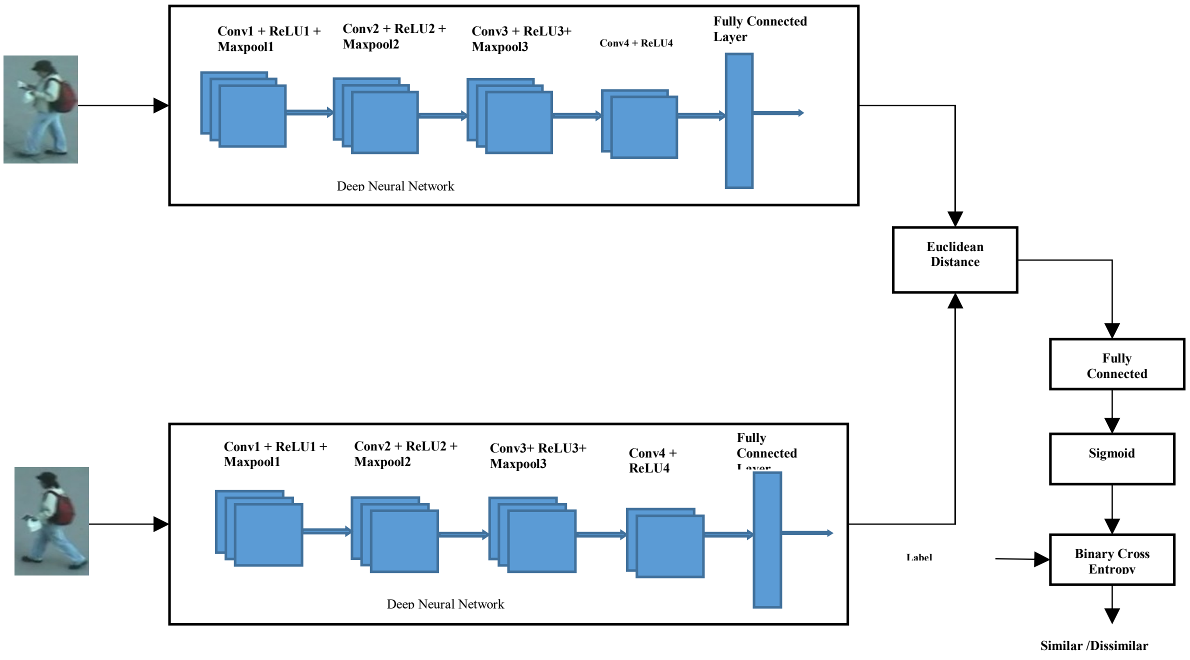

Figure 1. Architecture of Siamese Network for Human Re-identification

The Siamese network's main principle is to map the input images into a non-linear subspace in order to determine the Euclidean distance.

$$E(X1, X2) = ||X1 - X2||^2 \tag{2}$$

Table 3.1 Activation output sizes of layers in deep neural network

LAYERS	OUTPUT SIZE
Conv1	55 ×119 × 64
ReLU1	55 × 119 × 64
Maxpool1	27 × 59 × 64
Conv2	21 × 53 × 128
ReLU2	21 × 53 × 128
Maxpool2	10 × 26 × 128
Conv3	7 × 23 × 128
ReLU3	7 × 23 × 128
Maxpool3	3 × 11 × 128
Conv4	1 × 9 × 256
ReLU4	1 × 9 × 256
Fully Connected Layer	4096

4 Results and discussion

4.1 DATASETS

iLIDS-VID [20 -24] : In our project, we used the iLIDS-VID dataset. The iLIDS-VID dataset contains 600 image sequences for 300 persons who were randomly selected. Based on two non-overlapping camera views from the iLIDS multiple camera tracking scenario, this was constructed. The image sequences range in length from 23 to 192 images, with an average of 73 images. Due to the diversity in lighting and view point produced by cross-camera views, similar appearance among people, and clustering background, this dataset is extremely difficult to work with. In the iLIDS-VID dataset, there are two image sequence folders, cam 1 and cam 2. Cam1 serves as a training phase, while cam2 serves as a testing phase.

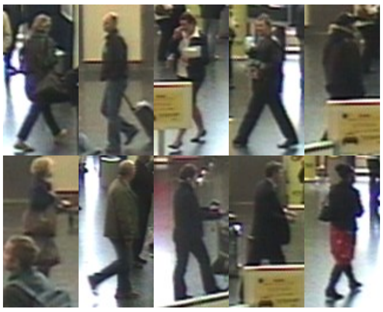

Figure 2. Sample images of iLIDS-VID dataset

PRID 2011 [25] : From two camera viewpoints on the opposite side, the PRID 2011 contains 400 image sequences for 200 people. The image sequences range from 5 to 675 frames, with an average of 100. This dataset was shot in natural settings with a clear background and few occlusions. There are noticeable colour shifts and shadows in this sample. Cam a and cam b are the names of the two cameras. Cam a performs the function of a training phase. Cam b serves as a testing ground.

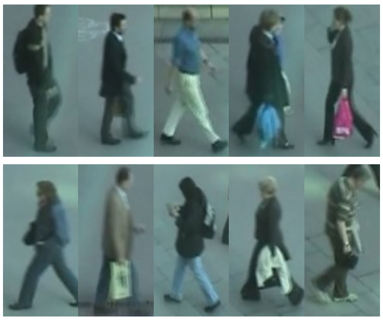

Figure 3. Sample images of PRID 2011 dataset

We aimed to establish a well-balanced training set that consists of both positive and negative labels for the proposed Siamese network. To achieve this, we generated pair data from all positive pairs, which were defined as a pair of images having the same identity. The suggested Siamese network was then trained with a total of one thousand epochs using the aforementioned balanced training set. To further increase the training set, false positive pairs were selected and added as supplement negative pairs by the original training set participants. Due to the complexity of the network and the large number of parameters that needed to be trained during the fine-tuning operation, additional examples were required for the starting dataset. To augment the dataset, we applied geometric transformations such as flipping, rotation, and centered and square cropping on the baggage image samples, which resulted in a significant increase in the total number of pairs. At the end of this procedure, the amount ratio of positive and negative pairs in the augmented training set should be approximately equal to each other. For Experimentation 1000 pairs of positive pairs and 1000 pairs of negative pairs are used for training. Testing is with another set of 1000 positive and 1000 negative pairs.

In the convolutional layers, all network weights were initially derived from a normal distribution with zero as the mean and a standard deviation of 10^{-1} as the parameter values. Similarly, biases were initialized from a normal distribution with a mean of 0.5 and a standard deviation of 10^{-2}. For the fully-connected layers, biases were initialized in the same manner as the convolutional layers' biases. However, the fully-connected layers' weights were selected from a much broader normal distribution with a zero-mean and a standard deviation of 2×10^{-1}.

The paper reports the results of experiments conducted on the iLIDS-VID dataset using different batch sizes while keeping the learning rate and iteration fixed at 1×10^{-4} and 1600, respectively. The evaluation shows that the recognition rate for the dataset is 75.66% when using a batch size of 20. However, increasing the batch size to 32, 64, 1024, and 2048 leads to higher recognition rates of 77.13%, 77.81%, 79.52%, and 79.45%, respectively. Notably, the cameras are also divided into training and testing phases, with Cam 1 being used for training and Cam 2 for testing. These results suggest that using a larger batch size can improve the recognition rate of the model for the iLIDS-VID dataset.

Table 2 Recognition rate by varying batch size in iLIDS-VID dataset

BATCH SIZE	LEARNING RATE	ITERATIONS	RECOGNITION RATE
20	1×10^{-4}	1600	75.66%
32	1×10^{-4}	1600	77.13%
64	1×10^{-4}	1600	77.81%
1024	$\mathbf{1\times10^{-4}}$	**1600**	**79.52%**
2048	1×10^{-4}	1600	79.45%

The paper presents an evaluation of the PRID 2011 dataset using different batch sizes and a fixed learning rate of 1×10^{-4} and iteration of 1000. The results show that a batch size of 64 yields the highest recognition rate of 85.82%, followed by a batch size of 1024 with a recognition rate of 85.23%. However, a batch size of 32 achieved a respectable recognition rate of 84.17%. On the other hand, a batch size of 128 resulted in a lower recognition rate of 83.22%. These findings suggest that the choice of batch size can have a significant impact on the performance of the model and should be carefully considered when developing and evaluating image recognition systems.

Table 3 Recognition rate by varying batch size in PRID 2011 dataset

Batch Size	Learning Rate	Iterations	Recognition Rate
32	1×10^{-4}	1000	84.17%
64	$\mathbf{1\times10^{-4}}$	**1000**	**85.82%**
128	1×10^{-4}	1000	83.22%
256	1×10^{-4}	1000	85.23%

The table shows the recognition rates achieved by different literature approaches on the iLIDS-VID dataset. The proposed method achieved the highest recognition rate of 79.52%, while other approaches achieved recognition rates ranging from 61.2% to 71.3%.

Table 4 Comparison with iLIDS-VID of other literatures

Authors	Recognition Rate (%)
Wu et al.[8]	61.2%
Song et al.[9]	71.3%
Zheng et al.[26]	62.4%
Liu et al.[28]	68.7%
Zhu et al.[29]	65.1%
Ours	**79.52%**

The table compares the recognition rates (%) of different literature on the PRID 2011 dataset. Khatun et al. achieved a recognition rate of 75%, Wu et al. achieved 74.8%, Li et al. achieved 84.5%, Zhu et al. achieved 80.3%, Ye et al. achieved 78.4% and our proposed method achieved the highest recognition rate of 85.82%. These results indicate that our Siamese network approach outperforms the other state-of-the-art algorithms in terms of recognition accuracy on the PRID 2011 dataset.

Table 5 Comparison with PRID 2011 of other literatures

Author	Recognition Rate (%)
Khatun et al.[5]	75%
Wu et al.[8]	74.8%
Li et al.[14]	84.5%
Zhu et al.[27]	80.3%
Ye et al.[28]	78.4%
Ours	85.82%

5 Conclusions

In this study, we have proposed the use of a Siamese network, a deep learning architecture for person re-identification. The network is able to predict the similarity or dissimilarity between two individuals in an image, and is supported by pedestrian properties, pose information, similarity metric enhancement, and group context to improve its accuracy. We evaluated our proposed method using the iLIDS-VID and PRID 2011 datasets, and achieved recognition rates of 79.52% and 85.82%, respectively. Our results demonstrate the effectiveness of the Siamese network for person re-identification tasks. Overall, this study contributes to the ongoing research on improving the accuracy of person re-identification across multiple cameras in surveillance videos.

References

[1] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-Identification with Consistent Attentive Siamese Networks," Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy NY, Siemens Corporate Technology, Princeton NJ.

[2] D. Chung, K. Tahboub, and E. J. Delp, "A Two Stream Siamese Convolutional Neural Network for Person Re-Identification," Video and Image Processing Laboratory (VIPER), School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA.

[3] G. Wang, S. Wang, W. Chi, S. Liu, and D. Fan, "A Person Reidentification Algorithm Based on Improved Siamese Network and Hard Sample," Mathematical Problems in Engineering, vol. 2020, article ID 3731848, 11 pages, Jun. 2020.

[4] C. Shen, Z. Jin, Y. Zhao, Z. Fu, R. Jiang, Y. Chen, and X.-S. Hua, "Deep Siamese Network with Multi-level Similarity Perception for Person Re-identification," in Proceedings of the ACM International Conference on Multimedia (MM'17), Oct. 23-27, 2017, Mountain View, CA, USA.

[5] A. Khatun, S. Denman, S. Sridharan, and C. Fookes, "A Deep Four-Stream Siamese Convolutional Neural Network with Joint Verification and Identification Loss for Person Re-detection," in 2018 IEEE Winter Conference on Applications of Computer Vision.

[6] R. Sun, W. Lu, Y. Zhao, J. Zhang, and C. Kai, "A Novel Method for Person Re-Identification: Conditional Translated Network Based on GANs," IEEE Transactions on Image processing, vol. 29, pp. 1966-1978, 2020.

[7] S. Pang, S. Iao, T. Song, J. Zhao, and P. Zheng, "An Improved Convolutional Network Architecture Based on Residual Modelling for Person Re-Identification in Edge Computing," IEEE Access, vol. 7, pp. 124938-124950, 2019.

[8] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-When to Look: Deep Siamese Attention Networks for Video-Based Person Re-Identification," IEEE Transactions on multimedia, vol. 21, pp. 1620-1633, 2019.

[9] W. Song, S. Li, T. Chang, A. Hao, Q. Zhao, and H. Qin, "Context-Interactive CNN for Person Re-Identification," IEEE Transactions on Image processing, vol. 29, pp. 5847-5859, 2020.

[10] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep High-Resolution Representation Learning for Cross-Resolution Person Re-Identification," IEEE Transactions on Image processing, vol. 30, pp. 2084-2096, 2021.

[11] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, and A. C. Kot, "Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[12] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Identity Adaptation for Person Re-Identification," IEEE Access, vol. 6, pp. 48147-48155, 2018.

[13] N. Mansouri, S. Ammari, and Y. Kessentini, "Improving Person Re-identification by Combining Siamese Convolutional Neural Network and Re-ranking Process," in International Conference on Machine Learning and Soft Computing.

[14] D. Li, R. Meng, and Y. Liu, "Person Re-identification Algorithm Based on Siamese LSTM," in 2021 3rd International Conference on Natural Language Processing (ICNLP).

[15] S. A. Sheshkal, K. Fouladi-Ghaleh, and H. Aghababa, "An Improved Person Re-identification Method by light-weight convolutional neural network," in 10th International Conference on Computer and Knowledge Engineering (ICCKE2020).

[16] M. Anandhalli, A. Tanuja, and V. P. Baligar, "Corner based statistical modelling in vehicle detection under various condition for traffic surveillance," Multimedia Tools and Applications, vol. 81, pp. 28849-28874, 2022.

[17] M. Anandhalli, A. Tanuja, and P. Baligar, "Geometric invariant features for the detection and analysis of vehicle," Multimedia Tools and Applications, vol. 81, no. 22, pp. 33549-33567, Feb. 2022. https://doi.org/10.1007/s11042-022-12919-8

[18] M. Anandhalli, P. Baligar, S. S. Saraf, et al., "Image projection method for vehicle speed estimation model in video system," Machine Vision and Applications, vol. 33, no. 7, pp. 1-10, Jan. 2022. https://doi.org/10.1007/s00138-021-01255-w

[19] M. Anandhalli, A. Tanuja, V. P. Baligar, et al., "Indian pothole detection based on CNN and anchor-based deep learning method," International Journal of Information Technology, vol. 14, no. 12, pp. 3343-3353, Mar. 2022. https://doi.org/10.1007/s41870-022-00881-5

[20] M. Li, X. Zhu, and S. Gong, "Unsupervised Tracklet Person Re-Identification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 6, pp. 1499-1513, 2019. DOI: 10.1109/TPAMI.2018.2847271.

[21] M. Li, X. Zhu, and S. Gong, "Unsupervised Person Re-Identification by Deep Learning Tracklet Association," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, September 2018, pp. 737-753. DOI: 10.1007/978-3-030-01270-0_45.

[22] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong, "Person Re-Identification by Unsupervised Video Matching," Pattern Recognition, vol. 65, pp. 197-210, May 2017. DOI: 10.1016/j.patcog.2016.11.006.

[23] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person Re-Identification by Discriminative Selection in Video Ranking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 12, pp. 2501-2514, December 2016. DOI: 10.1109/TPAMI.2016.2535198.

[24] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person Re-Identification by Video Ranking," in Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, September 2014, pp. 688-703. DOI: 10.1007/978-3-319-10578-9_45.

[25] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person Re-Identification by Descriptive and Discriminative Classification," in Proceedings of the Scandinavian Conference on Image Analysis (SCIA), 2011, pp. 23-28.

[26] A. Zheng, X. Zhang, B. Jiang, B. Luo, and C. Li, "A subspace learning approach to multishot person reidentification," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 50, no. 1, pp. 15-26, Jan. 2020.

[27] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," IEEE Transactions on Image Processing, vol. 27, no. 11, pp. 5543-5558, Nov. 2018.

[28] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2955-2967, June 2019.

[29] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, "Video-based person re-identification with accumulative motion context," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2732-2745, Oct. 2018.