Wireless Personal Communications
https://doi.org/10.1007/s11277-020-07948-1

Novel Similarity Metric Learning Using Deep Learning and Root SIFT for Person Re-identification

M. K. Vidhyalakshmi[1] · E. Poovammal[1] · Vidhyacharan Bhaskar[2] · J. Sathyanarayanan[3]

Accepted: 29 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
This paper deals with the person re-identification, intending to match the images of the person captured using disjoint cameras mounted in different locations. Such a task of matching the images remains a difficult issue as the appearance of the individual differs from the perspective of the various cameras. Inspired by the recent success of deep learning in the domain of person re-identification, a novel deep learning framework which combines deep features and Root Scale Invariant Features Transform (Root SIFT) features has been proposed. The conventional deep Convolutional Neural Network (CNN) can obtain significant features but does not take into account the spatial relationship between the features. Also, CNN requires an enormous number of instances to train the network. To address these issues, the proposed method combines Root SIFT features along with the CNN features. With the combination of Deep and Root SIFT features, the model can give improved performance over other CNN based models. Experiments were conducted on standard datasets CUHK 03 (labelled and detected), CUHK 01 and VIPeR and the matching rate is reported as 74.45% for CUHK 03 (labelled), 72.63% for CUHK 03 (detected), 76.12% for CUHK 01 and 48.45% for VIPeR dataset. The experiments demonstrate that the proposed algorithm has improved identification rate over the recent algorithms.

Keywords Person re-identification · CNN · Siamese network · Root SIFT · Bag of visual words

1 Introduction

In many places, camera monitoring systems have been installed for surveillance purpose. These monitoring systems have a set of cameras fixed at different locations. Such images of a person obtained from these cameras vary considerably. Because these images are taken by cameras under different lighting conditions, various pose changes, viewpoint variation,

✉ Vidhyacharan Bhaskar
vcharan@gmail.com

[1] Department of CSE, SRMIST, Kattangulathur, Chennai, India

[2] Department of Electrical and Computer Engineering, San Francisco State University, San Francisco, CA, USA

[3] R & D, Dev Technologies, Chennai, India

Published online: 28 November 2020

occlusions and low resolution; it is hard to associate the image of a person taken by one camera with the other. The problem has attracted a lot of researchers' attention to explore and find the best solution, this emerging problem is called person re-identification. Re-identification is a process of associating a person's images taken by different cameras positioned at different locations. Challenges to this issue include lighting adjustments, low-grade sensors, different camera settings and occlusions.

The established methods initially concentrated on extracting the hand-crafted invariant features from the images and use metric leaning techniques to learn the similarity among the images. With the success of deep learning networks, specifically, the Convolution Neural Networks (CNN) has substituted the hand-crafted feature representations with sophisticated features. CNN requires large training samples for accurate performance which literally requires thousands of images. Also, CNN pays attention to important features only and it does not take into account the spatial relationship between the features. The spatial relation between characteristics is important because the images of the same individual will differ according to their pose. On the other hand, Scale Invariant Feature Transform (SIFT) feature may not perform comparable to deep features but require only a few training samples. Also, the SIFT feature considers the spatial relation between the features. SIFT captures the local features present in the images. Root SIFT is the improved form of SIFT. To include the benefits of scale invariance, we have combined the Root SIFT feature with the deep feature to increase the performance of the model. The main contributions of this paper include:

- We propose a novel neural network model that utilizes both the local and global features of an image, generated using Root SIFT and deep CNN.
- This deep neural network effectively learns to preserve the relationships between the features thereby correctly identifying the person under various orientations and illumination conditions.
- Testing with the public datasets like CUHK03-NP, CUHK01 and VIPeR prove that the proposed model's performance is superior compared to recent methods.

2 Related Work

2.1 Methods Which Work with Features

Earlier works on person re-identification were focussing on developing the methods that extract features which are invariant to various changes that were incorporated due to cross camera views. Colour has a major role to identify a person. The research works [1–8] considered changes in colour due to lighting and colour invariants were built to handle different types of irregularity in the imaging environments. Colour features are represented as histograms in RGB, HSV or YUV spaces. Besides Colour features, texture features like Haar-like features [9] and Local binary Patterns [10] are also used in person re-identification. In [11, 12] salience is used as an important feature for better-discriminating capability. Farenzena et al. [13] suggested Symmetry Driven Accumulation of Local Features (SDALF) that represent parts of the person using a combination of texture and weighted colour histograms. These methods try to develop discriminative features which are used to separate the correct match and a wrong match of a person. But developing a discriminating feature is a difficult task and researchers started focussing on metric learning techniques.

2.2 Methods that Work with Metric Learning

The metric learning techniques are set of supervised learning techniques that learn a distance measure known as Mahalanobis distance which is a measure of similarity among the images. Metric learning tries to maximise the Mahalanobis distance between the images belonging to the different persons while minimising the distance between the image of the same persons. Kostinger et al. [14] introduced a non-iterative simple procedure Keep it Simple and Straight forward Metric Learning (KISSME) to match the sample pair using Gaussian distribution. Hizer et al. [15] utilized Large Margin Nearest-Neighbour (LMNN) classification problem that tries to bring together the instances of the person with the same label and forces the different labels away. The Relative Distance Comparison model is used for person re-identification in [16] to optimise the probability of a correct match which have smaller relative distance compared to an incorrect match. Leng [17] proposed a metric learning method which learns from the multiple views of images. By creating multiple views, training data is increased and solves the problem of non-availability of training data. Yang [18] used an improved logistic discriminant method which makes use of additional information obtained during training along with the original data. The additional data obtained during the training phase adds a piece of extra information to learn the distance metric.

2.3 Methods Work with Deep Learning

The influence of deep learning in several tasks of computer vision like image classification has contributed to its use in person re-identification. With the help of deep learning using CNN models, it is possible to extract sophisticated representations of features from images. Yi et al. [19] used Siamese network is trained to learn how to bring closer the images of a same person in the feature space and bring a considerable distance between the images of the different persons. Li et al. [20] used Siamese network and modelled person re-identification problem as a problem of verification to find out whether a given pair of images belong to the same class or not. Rama Varior et al. [21] proposed deep learning Convolution Neural Network architecture which is robust to illumination changes and camera views, learns the invariant colour features directly from the pixels of the image instead of handcrafted features like a histogram. Jiang et al. [22] proposed a deep learning framework which learns discriminatory features in multi-scale. Zheng et al. [23] proposed Consistent Attentive Siamese Network (CASN) which finds attention regions in the person image and provides a robust representation for the spatially localised regions. Ahmed et al. [24] proposed a deep Convolutional architecture which finds the cross-neighbourhood differences to find whether the given two images are the same or not. Da Xiang Li et al. [25] proposed a deep learning framework which uses global and regional features of the images of the person to improve the identification rate. The framework combines conventional CNN with the Siamese model to utilise the advantages of verification and identification data.

Recently hybrid deep learning models are used for person re-identification. The hybrid models combine both deep features and traditional features like LOMO, Histograms and so on. Many a time deep structure is not favourable for training as Convolution Neural Networks gives more attention to important features in the image but doesn't consider the spatial relevance of the feature. Yan et al. [26] combined the deep features with colour and LBP texture feature. The hy brid methods reduced the dependence of large training

datasets and gives good result even with a small training set. Haifeng Sang [27] proposed Multi-information flow convolutional neural network (MiF-CNN) which reuses the features by linking the input with the output of a convolutional layer to the successive layers. The method also combines the attribute information to increase the performance. Yu-Xin Yang et al. [28] used hybrid CNN by using with SIFT for facial expression recognition. Fangyi Liu [29] used SIFT features to locate the viewpoint invariant regions in the images of the person and used these features to regulate the CNN network. Inspired by these works we proposed a novel deep learning framework for person re-identification which uses Root SIFT features are combined with the CNN features to retain the spatial relevance of the features. Also, each of the two techniques utilized in our proposed model complements each other so the resultant output can be maximized.

3 Proposed Approach

3.1 Design

The Proposed work used the commonly utilized Siamese CNN model that takes two images as inputs and predicts if they belong to the same or different person, using the similarity measure between them. The framework of the proposed model is shown in Fig. 1. In the proposed model instead of giving two inputs, four inputs are given. Each arm of the Siamese network is fed with one of the image pair and its Histogram of Oriented Gradients (HOG)vector which is obtained using the Root SIFT descriptor. The detailed explanation of the process of extracting the HOG vector is given in Sect. 3.5.

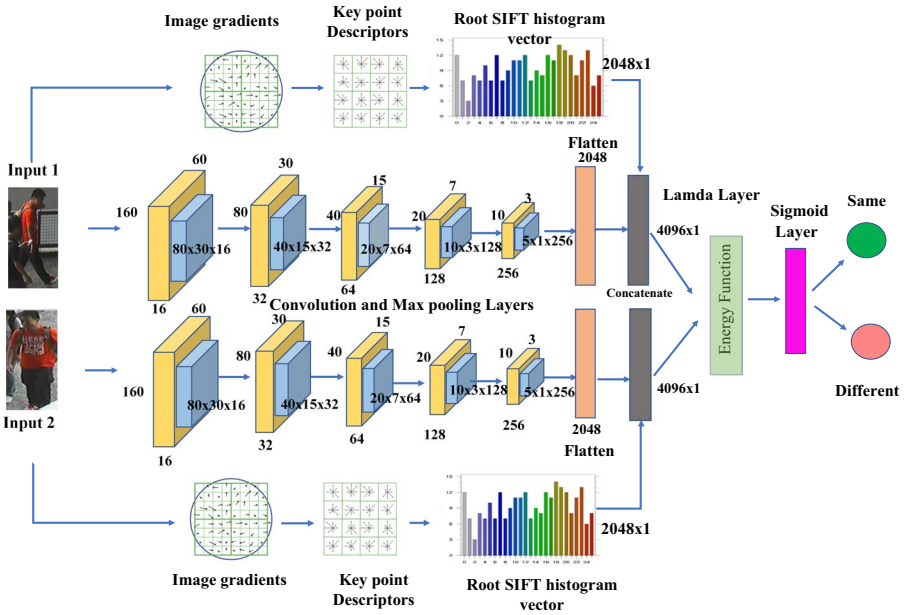

Fig. 1 Framework of the proposed deep learning model

The histogram vector and the flattened output of the deep CNN network are concatenated. The two emerging outputs are fed into an energy function. The similarity score is decided based on the Euclidean distance amongst these two inputs. As a final layer, the distance output is fed to the fully connected layer with activation function as Sigmoid. The resultant output predicts if the images belong to the same or different person.

3.2 Pre-processing

For the pre-processing step, the colour image is normalised by dividing the pixel value with 255 so that the range is between 0 and 1. The mean of the pixel values is subtracted from each pixel and divide by its standard deviation to standardise the data set. The augmentations of the images are implemented using the Keras deep learning neural network Library. The augmentation helps the deep learning model to train effectively with a wide range of inputs. The image augmentation has been carried out utilising the 'image data generator class' library to perform image transformations like randomly rotating, shifting, shearing, flipping and reordering the dimensions of the feature. Image sizes were retained to their initial value of 160×60 pixels.

3.3 Deep CNN Framework

In our customised deep learning network model, there are 5 convolution layers each followed by max-pooling, batch normalisation and drop out layers. In these convolution layers, the number of filters used is 16, 32, 64, 128 and 256. The size of the kernel is fixed as 3×3 for all the convolution layers. The padding has been set to same in order to maintain the spatial size of the image. Max pooling layers use nonlinear downsampling to reduce computation complexity. Max pooling layers are incorporated with a pool size of 2×2 and stride of 2. This layer provides translational invariance. A dropout layer of 0.5 helps in the regularisation and reduces the possibility of overfitting. To increase the network's ability to solve the complex problems some non-linear factors has to be introduced into the network. This is possible by introducing the activation layer in the network. Leaky Relu is chosen to be the activation function for all the convolution layers because of its superior capability as mentioned in [30].

Let X_A, X_B be the feature vectors of the given pair of training images A, B. $X_A = \{a_1 \text{ U } a_2\}$ where a_1 is deep feature vector and a_2 is the scale invariant feature vector of image (A) $X_B = \{b_1 \text{ U } b_2\}$ where b_1 is deep feature vector and b_2 is the scale invariant feature vector of image (B) Let Y denote binary label for this pair of images. The value of Y is 0 if X_A and X_B are similar pair, Y is 1 if X_A and X_B are dissimilar pair. Let G_w be the function defined by parameters W. To find the optimised value of parameter W, the objective function is defined as in the Eq. (1):

$$D_w(X_A, X_B) = \left\| G_w\ (X_A) - G_w\ (X_B) \right\|_2 \tag{1}$$

The contrastive loss function [31] can then be obtained as given in the Eq. (2):

$$L(W, Y, X_A, X_B) = (1 - Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}\{\max(0, m - D_w)\}^2 \tag{2}$$

The mode 'm' was chosen as 1 for our model and the similarity function is defined by contrastive loss function which is the Euclidean distance measure between the 2

concatenated feature vectors of the image pairs. The output layer was implemented with a fully connected layer with activation function as Sigmoid. Based on the L_2 norm distance between the 2 vectors a binary classification yields to predict whether the image pair is of the same person or a different person.

3.4 SIFT and Root SIFT Descriptors

Scale invariant feature transform (SIFT) was proposed by Lowe [32], to extract distinguishing invariable image features to achieve a good matching of images in different views. SIFT features are invariant to size and orientation and therefore resilient to change of perspective and other spatial distortions. The SIFT algorithm has two steps: Keypoint generation and Keypoint descriptors generation. For key point generation instead of using SIFT algorithm, Features from Accelerated Segment Test (FAST) algorithm proposed by Rosten and Drummond [33] has been used for speedy and efficient identification of the key points present in the image. After obtaining the key points, the descriptors for each key point is obtained using the SIFT algorithm.

With a key point as the centre, 16×16 window is placed over the image. The region under the 16×16 window is divided into 4×4 sub-regions. For all the pixels in the sub-regions, magnitude and direction are calculated. The orientation histogram is computed with 8 bins with each of the bin consisting of 45 degrees, for all the16 pixels in the sub-regions. For 4×4 window totally 16 histograms each with 8 bins are obtained. Thus, for a key point, the descriptor size is 128 elements. Root SIFT can improve the performance of the SIFT algorithm and are obtained by L_1 normalising the SIFT descriptors and taking the square root of each element [34].

3.5 Bag of Visual Words

Root SIFT descriptors obtained from the FAST key points of the image are used to create a pool of clusters using the K means algorithm. Thus, a bag of meaningful features is created as a reference model. With this reference model, the descriptors of each image are compared and appropriate values or frequencies are assigned. The more the number of features higher the magnitude of the histogram. Thus, a Histogram vector of size of 2048 is generated for every image and is used as a second input for the Deep CNN described earlier.

4 Implementation Details

All the experiments were carried out with Intel i7 9th Gen processor with NVIDIA GPU GeForce GTX 1660 Ti GPU processor. Implementation was carried out using Keras with Tensor flow 2.0 as the backend. Figure 2 describes the model of the proposed framework.

All convolution and FC layers use Leaky Relu as the activation function. Drop out layer with 50% is used after all Batch Normalization layers. Total parameters of the deep neural network are 44,973,420 with actual trainable parameters as 44,972,422 and non-trainable parameters as 998. The details of the layer parameters are listed in Table 1. Stochastic Gradient descent (SGD) has been chosen as the optimiser for the model. The learning rate is selected as 0.01 and momentum is selected as 0.9 as it yielded the best results.

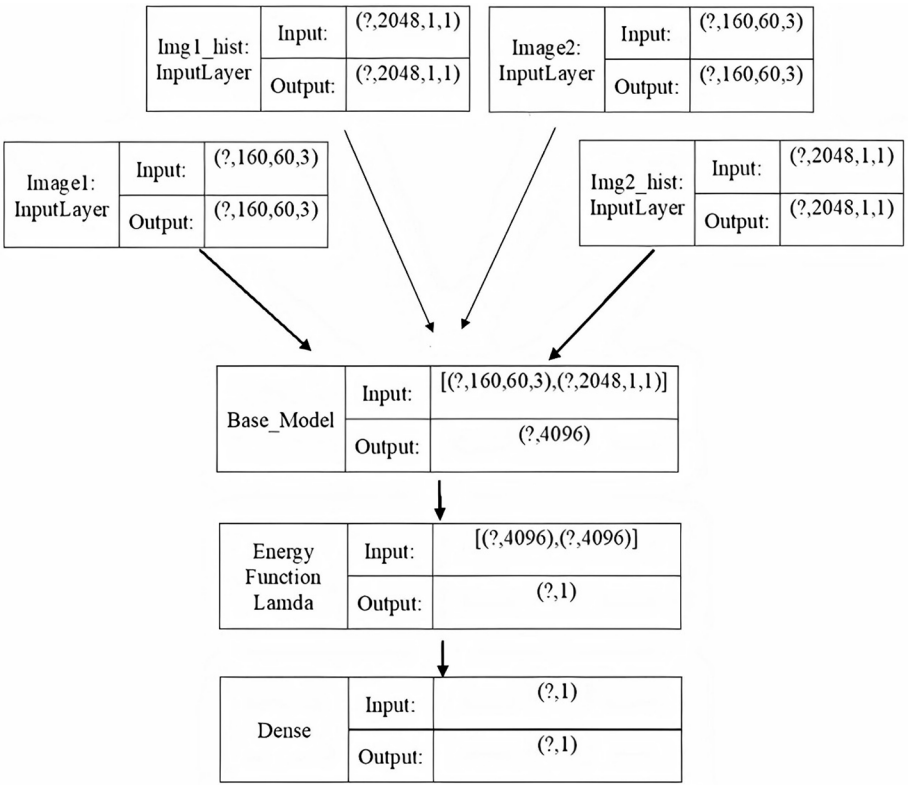

Fig. 2 Model of the proposed framework

4.1 The Protocol Used for Training

The given dataset is divided into training, validation and test sets according to the protocol defined by the respective datasets as detailed in the Sects. 5.1–5.3. The image size is chosen to be of standard size of 160×60 pixels. The training sets and validation sets are created with an equal number of positive and negative pairs for both the image and their Root SIFT histogram feature pairs. The number of samples in the training set play a crucial role in the performance of the model. The training set is fine-tuned with hyperparameters to give optimal results. The choice of Leaky Relu as the activation function is very much influential in improving the performance by way of faster learning. Augmentation was applied to the training and validation datasets to enhance the model's performance by way of different image transformations. The choice of batch size was also influential in increasing the model's accuracy.

4.2 The Protocol Used for Evaluation

The test set is divided into probe and gallery sets, with probe set having images obtained from one camera view and gallery set having images obtained from other camera views. An image

Table 1 Layer parameters of the deep learning network

Base model layer names	Output dimension (height,width, channel)	Filter size	Stride	Parameters
x_image_input (InputLayer)	(160, 60, 3)	–	–	–
Batch Normalization_0	(160, 60, 3)	–	–	12
convolution_1	(160, 60, 16)	16	1	448
Max Pooling_1	(80, 30, 16)	2	2	0
Batch Normalization_1	(80, 30, 16)	–	–	64
convolution _2	(80, 30, 32)	32	1	4640
Max Pooling_2	(40, 15, 32)	2	2	0
Batch Normalization_2	(40, 15, 32)	–	–	128
convolution _3	(40, 15, 64)	64	1	18,496
Max Pooling_3	(20, 7, 64)	2	2	0
Batch Normalization_3	(20, 7, 64)	–	–	256
convolution _4	(20, 7, 128)	128	1	73,856
Max Pooling_4	(10, 3, 128)	2	2	0
Batch Normalization_4	(10, 3, 128)	–	–	512
convolution _5	(10, 3, 256)	256	1	295,168
Max Pooling_5	(5, 1, 256)	2	2	0
Batch Normalization_5	(5, 1, 256)	–	–	1024
Fully Connected_1_image	–	2048	–	2,623,488
x_hist_input (InputLayer)	(2048, 1, 1)	–	–	–
Fully Connected _2_hist	–	2048	–	4,196,352
Concatenate_img&hist	4096	–	–	–
Fully Connected _3_image&hist	4096	4096	–	16,781,312
Fully Connected _4_image&hist	4096	4096	–	16,781,312

is randomly selected from the probe set and compared with the rest of the images in the gallery set. This procedure is carried out for 100 random samples of the probe. The identification rate at any rank k is the ratio of the number of correct probe images obtained at rank k to the total number of images in the gallery set. Cumulative identification rate is the sum of all identification rates obtained up to the level of kth rank. Cumulative Matching Characteristics (CMC) curves were used in the earlier person re-id algorithms for measuring their performance. As pointed out in [35], CMC is applicable only if the gallery contains a single ground truth match for any probe image. If the gallery contains more than one match for any given probe then it is inaccurate as it doesn't take into consideration the recall. The mean Average Precision (mAP) metric is considering to be the best measure for any object retrieval problem as it considers both the precision and recall statistics for any measure. Hence most of the recent person re-id models use mean Average Precision (mAP) as a measure for evaluating their performance. Following the trend, our model's performance is evaluated on all the datasets using the CMC and mAP measures.

Table 2 Comparison of identification rates and mAP for CUHK 03-NP dataset (Labelled)

Methods	Identification rate in %			mAP %
	Rank-1	Rank-5	Rank-10	
KISSME	14.17	37.47	52.20	–
FPNN	20.65	50.94	67.01	–
LOMO + XQDA	52.20	82.23	92.17	–
Ahmed	54.74	–	–	–
GST	71.20	–	–	68.60
LLMMGF	73.10	90.00	97.10	–
CASN	73.70	–	–	68.00
GreyReID	73.90	–	–	76.60
Deep + Root SIFT	74.45	91.43	98.52	77.39

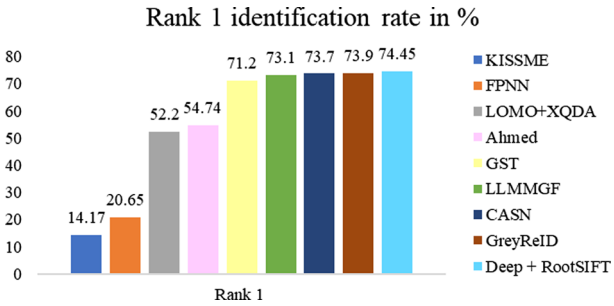

Fig. 3 Comparison of Rank 1 identification rates for CUHK03-NP (Labelled)

5 Experimental Results

5.1 Experimental Results for CUHK03-NP Dataset

We use the revised CUHK03-NP [20] dataset with the new protocol which contains 14,096 images of 1467 identities captured from 6 different cameras. Each person is captured by two cameras with disjoint views. The dataset has two types namely Labelled and Detected where Labelled is the one in which the images are manually labelled and Detected is the one in which the images are detected with bounding boxes using deformable part models (DPM). The images are not uniformly sized and had to be resized to 160×60 pixels before they are used for training and validation. As per the new protocol for CUHK03-NP dataset, the training set consists of 767 identities and test set consists of 700 identities. The proposed model is trained for 267,750 iterations using mini-batch size as 64.

For the dataset CUHK 03-NP (Labelled),the results of the proposed method are compared with the methods KISSME [14], FPNN [20], CASN [23], Ahmed [24], LLMMGF [25], LOMO + XQDA [36], GST [37] and GreyReID [38].The comparison results for identification rates and mAP for CUHK 03-NP (Labelled) are tabulated in Table 2. The diagrammatic comparison of identification rates at the level of Rank 1 is shown in Fig. 3 and the results of comparison of mAP for these methods are shown in Fig. 4. The identification rate at the level of Rank 1 and mAP is reported as 74.5% and 77.39% for CUHK03-NP dataset (Labelled).

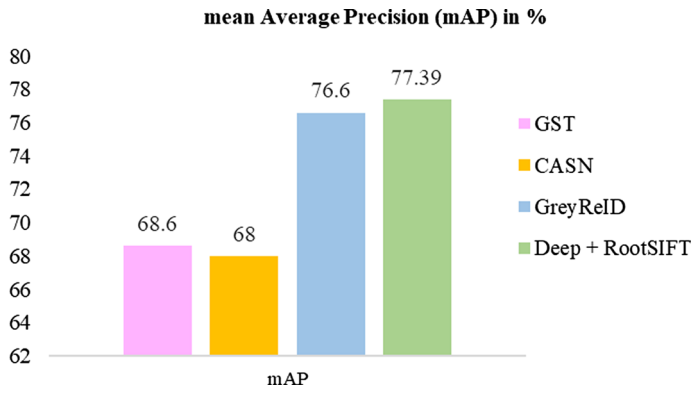

Fig. 4 Comparison of mAP CUHK03-NP (Labelled)

For the dataset CUHK 03-NP (Detected), the results of the proposed method are compared with the methods KISSME [14], FPNN [20], CASN [23], Ahmed [24], LLMMGF [25], VCFL + rerank [29], GST [37] and GreyReID [38]. The comparison results for identification rates and mAP for CUHK 03-NP (Detected) are tabulated in Table 3. The diagrammatic comparison of identification rates at the level of Rank 1 is shown in Fig. 5 and the comparison of the mAP for these methods are shown in Fig. 6. The identification rate at the level of Rank 1 and mAP are reported as 72.63% and 72.38% for CUHK03-NP dataset (Detected).

5.2 Experimental Results for CUHK01 Dataset

CUHK01 [39] dataset contains 971 identities each having 4 images from 2 different cameras. In total, there are 3884 images, all manually cropped to the size of 160 × 60 pixels. The proposed model is trained, validated and tested using two different scenarios. In test case-1 the dataset is divided into 771 training identities, 100 validation and remaining 100 as testing identities. With the mini-batch size of 64, the proposed model is trained for 43,350 iterations. In test case-2, the dataset is randomly split into 420 training identities,

Table 3 Comparison of identification rates and mAP for CUHK 03- NP dataset (Detected)

Methods	Identification rate in %			mAP %
	Rank-1	Rank-5	Rank-10	
KISSME	11.70	33.46	48.46	–
FPNN	19.90	50.00	64.00	–
Ahmed	44.96	–	–	–
GST	68.80	–	–	65.50
GreyReID	69.90	–	–	73.30
VCFL + rerank	70.36	–	–	70.44
LLMMGF	71.50	88.30	94.00	–
CASN	71.50	–	–	64.40
Deep + Root SIFT	72.63	89.62	95.46	72.38

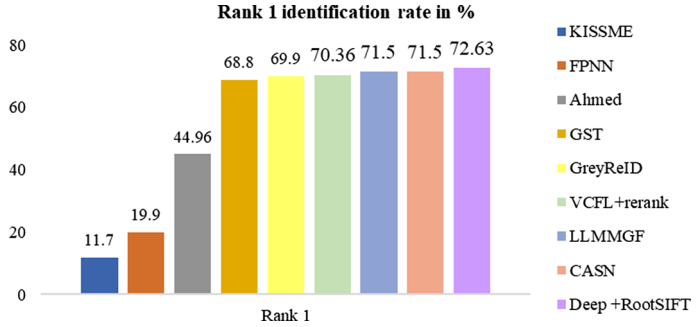

Fig. 5 Comparison of Rank 1 identification rate for CUHK03-NP (Detected)

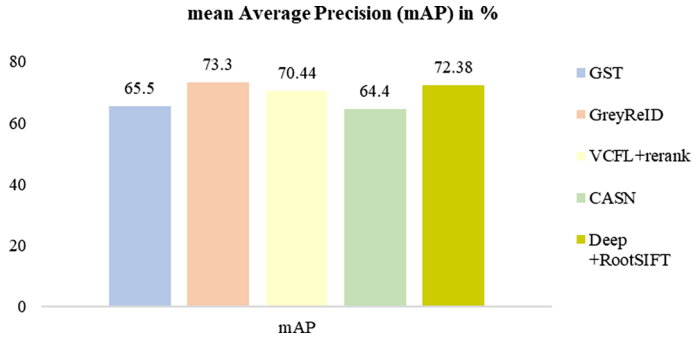

Fig. 6 Comparison of mAP for CUHK03-NP (Detected)

Table 4 Comparison of identification rates and mAP for CUHK01 test case-1

Methods	Identification rate in %			mAP %
	Rank-1	Rank-5	Rank-10	
SDALF	9.90	35.00	53.00	–
eSDC	22.83	42.00	57.00	–
KISSME	29.40	55.00	70.00	–
Ahmed	65.00	83.00	93.12	–
Deep + Root SIFT	81.35	87.32	94.04	75.96

65 validation and 486 testing identities. As the number of images available for training is very limited extended augmentation had to be performed to increase the number of training images. Also, for this dataset, the model was pre-trained with CUHK03-NP dataset.

For the test case-1, our proposed method is tested and its performance is compared with the methods eSDC [11] and SDALF [13]. KISSME [14] and Ahmed [24] The comparison results with other methods are shown in Table 4 and the Rank 1 identification rate is compared in Fig. 7. When compared with the other methods which use either CNN feature alone [24] or the other hand-crafted features [11, 13, 14], the performance of our model using both CNN and Root SIFT histogram features has shown a significant

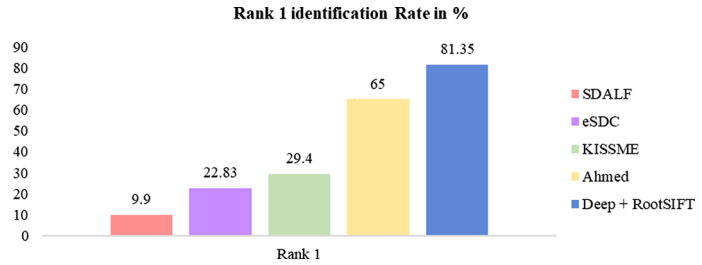

Fig. 7 Comparison of Rank 1 identification rates for CUHK01 test case-1

increase in the Rank1 identification rate. The Rank1 identification rate for test case-1 is 81.35% and the improvement over Ahmed [24] is by 16.35%. The mean Average Precision is 75.96%.

For the test case-2 our proposed method is compared with the methods SalMatch [12], SDALF [13], KISSME [14], Ahmed [24], LLMMGF [25], MiF + PARN [27] and LOMO + XQDA [36]. The comparison results with other methods are shown in Table 5 and the diagrammatic comparison of identification rate at the level of Rank 1 is shown in Fig. 8. The identification rate at Rank1 obtained for test case-2 is 76.12% and mAP is 73.89%.

Table 5 Comparison of identification rates and mAP for CUHK01 test case-2

Methods	Identification rate in %			mAP %
	Rank 1	Rank 5	Rank 10	
SDALF	9.90	22.00	30.20	–
KISSME	15.57	37.90	50.29	–
SalMatch	28.45	46.00	55.00	–
Ahmed	47.50	–	–	–
LOMO + XQDA	49.20	75.70	84.20	–
MiF + PARN	71.81	90.33	91.77	–
LLMMGF	72.00	88.10	91.20	–
Deep + Root SIFT	76.12	89.34	92.2	73.89

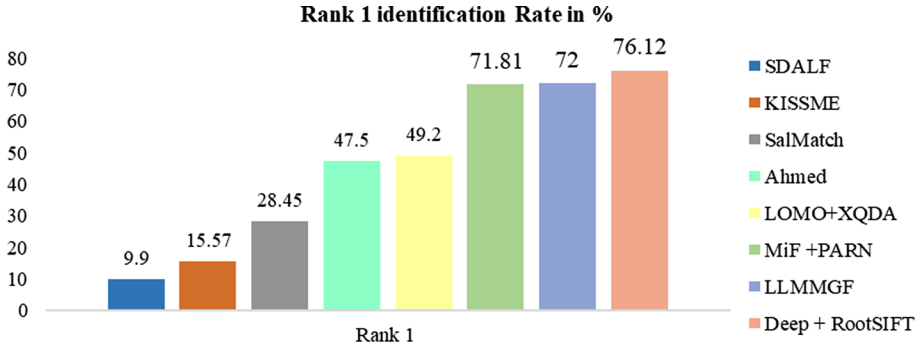

Fig. 8 Comparison of Rank 1 identification rates for CUHK01 test case-2

Table 6 Comparison of identification rates and mAP for VIPeR dataset

Methods	Identification rate in %			mAP %
	Rank 1	Rank 5	Rank 10	
SDALF	19.87	38.89	49.37	–
KISSME	22.63	50.13	63.73	–
SalMatch	30.16	52	65	–
MiF + PARN	46.84	74.68	83.23	–
Deep + Root SIFT	48.45	76.32	85.83	49.2

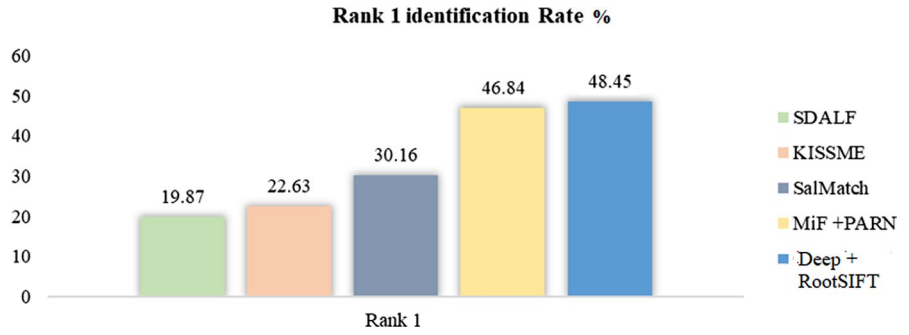

Fig. 9 Comparison of Rank 1 identification rates for VIPeR dataset

5.3 Experimental Results for VIPeR Dataset

VIPeR dataset contains images of 632 identities with 2 images available for each identity and thus totally 1264 images available in the dataset. The two images are captured using two different cameras. This dataset is quite challenging as there is only one image available per camera view. The dataset is divided into 316 training images and 316 testing images which are randomly selected. All the available images are equal in dimensions with 128×48 pixels. The results of the proposed method are compared with the methods Salmatch [12], SDALF [13], KISSME [14], and MiF + PARN [27]. The results are tabulated in Table 6. The first rank identification rate for VIPeR dataset is 48.45% which is 1.61% greater than MiF + PARN. The mean Average Precision is 49.2%. (Fig. 9)

6 Conclusion

We use a novel approach which combines the Root SIFT features along with CNN Deep features to enhance the performance of person re-identification model. By this method, we have proved how local features could complement the global features thereby enhancing the prediction metrics. This addresses the challenges arising due to viewpoint variation and also preserves the spatial relevance of deep features simplistically. Our method's performance surpasses the state of the art results in person reidentification problem with a minimal design. In future, additional features like colour ,texture can be included alon with

SIFT and deep features. To further enhance the model, transfer learning method can be done employed using the larger deep networks like ResNet ,VGGNet etc.These networks are large pre trained networks popularly used for image classification. By fine tuning these networks, the already learned weights can be used for the model's deep network. Though CNN includes translation variance it is resistant to rotational variance. To solve the pose variation problem, rotational invariance has to be incorporated. The use of Capsule networks can be explored to include rotational invariance.

References

1. Porikli, F. (2003). Inter-camera color calibration by correlation model function. In *Proceedings of the international conference on image processing. ICIP '03* (pp. II−133). Barcelona, Spain: IEEE.
2. Javed, O., Shafique, K., & Shah, M. (2005). Appearance modeling for tracking in multiple non-overlapping cameras. In *Proceedings of the computer society conference on computer vision and pattern recognition. CVPR '05* (pp. 26–33). San Diego, CA: IEEE.
3. Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Proceedings of the Scandinavian conference on image analysis* (pp. 91–102). Berlin, Heidelberg: Springer.
4. Gijsenij, A., Lu, R., & Gevers, T. (2012). Colour constancy for multiple light sources. *IEEE Transactions on Image Processing*, *21*(2), 697–707.
5. Kviatkovsky, Adam, A., & Rivlin, E. (2012). Colour invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1622–1634.
6. Kuo, C. H., Khamis, S., & Shet, V. (2013). Person re-identification using semantic colour names and rank boost. In *Proceedings of the workshop on applications of computer vision. WACV '13* (pp. 281–287). Clearwater Beach, FL: IEEE.
7. Li, P., Wu, H., Chen, Q., & Bi, C. (2015). Person re-identification using colour enhancing feature. In *Proceedings of the 3rd IAPR Asian conference on pattern recognition (ACPR)* (pp. 086–090). Kuala Lumpur, Malaysia: IEEE. https://doi.org/10.1109/ACPR.2015.7486471.
8. Varior, R. R., Wang, G., Lu, J., & Liu, T. (2016). Learning invariant colour features for person reidentification. *IEEE Transactions on Image Processing*, *25*(7), 3395–3410.
9. Bak, S., Corvee, E., Bremond, F., & Thonnat, M. (2010), Person reidentification using haar-based and DCD-based signature. In *Proceedings of the advanced video and signal based surveillance* (pp. 1–8). Boston, MA, USA.
10. Chahla, C., Snoussi, H., Abdallah, F., & Dornaika, F. (2017). Discriminant quaternion local binary pattern embedding for person re-identification through prototype formation and colour categorization. *Engineering Applications of Artificial Intelligence*, *58*, 27–33.
11. Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-Identification. In *Proceedings of the conference on computer vision and pattern recognition. CVPR '13* (pp. 3586–3593). Portland, Oregon: IEEE.
12. Zhao, R., Ouyang, W., & Wang, X. (2013). Person re-identification by salience matching. In *Proceedings of the international conference on computer vision. ICCV '13* (pp. 2528–2535). Sydney, Australia: IEEE.
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the computer society conference on computer vision and pattern recognition* (pp. 2360–2367). San Francisco, California: IEEE.
14. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Proceedings of the Computer Vision and Pattern Recognition. CVPR '12*. (pp. 2288–2295). IEEE: Providence, RI, USA.
15. Hirzer, M., Beleznai, C., Kostinger, M., Roth, P. M., & Bischof, H. (2012). Dense appearance modeling and efficient learning of camera transitions for person re-identification. In *Proceedings of the 19th international conference on image processing* (pp. 1617–1620). Los Vegas, Nevada: IEEE.
16. Zheng, W. S., Gong, S., & Xiang, T. (2012). Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(3), 653–668.
17. Leng, Q. (2018). Co-metric learning for person re-identification. *Advances in Multimedia*. https://doi.org/10.1155/2018/3586191.

18. Yang, X., Wang, M., & Tao, D. (2017). Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing, 27*(2), 791–805.
19. Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Deep metric Learning for Person Re-identification. In *Proceedings of the 22nd international conference on pattern recognition* (pp. 34–39). Stockholm, Sweden: IEEE.
20. Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: deep filter pairing neural network for person re-identification. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 152–159). IEEE: Massachusetts Ave, NW.
21. Chaudhary, D. D., & Jadhav, N. (2018). Learning invariant colour features for person reidentification. *International Journal of Engineering Technologies and Management Research*, *5*(5), 65–70.
22. Qian, X., Fu, Y., Jiang, Y. G., Xiang, T., & Xue, X. (2017). Multi-scale deep learning architectures for person re-identification. In *Proceedings of the international conference on computer vision* (pp. 5399–5408). Venice, Italy: IEEE.
23. Zheng, M., Karanam, S., Wu, Z., & Radke, R. J. (2019). Re-identification with consistent attentive siamese networks. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 5735–5744). Long Beach, CA: IEEE.
24. Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 3908–3916). Boston, MA: IEEE.
25. Li, D. X., Fei, G. Y., & Teng, S. W. (2020). Learning large margin multiple granularity features with an improved siamese network for person re-identification. *Symmetry, 12*(1), 92–99.
26. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., & Yang, X. (2016). Person re-identification via recurrent feature aggregation. In *European conference on computer vision* (pp. 701–716). Cham: Springer.
27. Sang, H., Wang, C., He, D., & Qing, L. (2019) Multi-information flow CNN and attribute-aided reranking for person reidentification. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2019/7028107.
28. Yang, Y. X., Wen, C., Xie, K., Wen, F. Q., Sheng, G. Q., & Tang, X. G. (2018). Face recognition using the SR-CNN model. *Sensors (Basel, Switzerland), 18*(12), 4237–4243.
29. Sang, H., Wang, C., He, D., & Liu, Q. (2019). View confusion feature learning for person re-identification. In *Proceedings of the international conference on computer vision* (pp. 6639–6648). Seoul, Korea: IEEE.
30. Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the international conference on machine learning* (Vol. 30(1), pp. 3–12).
31. Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proceedings of the computer society conference on computer vision and pattern recognition. CVPR '06*. 2 (pp. 1735–1742). New York, NY: IEEE.
32. Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 7th international conference on computer vision*. (Vol. 2, pp. 1150–1157). IEEE.
33. Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. In *European conference on computer vision* (pp. 430–443). Berlin, Heidelberg: Springer.
34. Arandjelović, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 2911–2918). Massachusetts Ave., NW: IEEE.
35. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the international conference on computer vision* (pp. 1116–1124). Santiago, Chile: IEEE.
36. Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition (pp. 2197–2206). IEEE: Boston, MA.
37. Zhang, J., Hu, X., Wang, M., Qiao, H., Li, X., & Sun, T. (2019). Person re-identification via group symmetry theory. *IEEE access : practical innovations, open solutions, 7*, 133686–133693.
38. Qi, L., Wang, L., Huo, J., Shi, Y., & Gao, Y. (2019). GreyReID: A Two-stream Deep Framework with RGB-grey Information for Person Re-identification. arXiv preprint arXiv:1908.05142http://arxiv.org/1908.05142.
39. Li, W., Zhao, R., & Wang, X. (2012). Human Reidentification With Transferred Metric Learning. In Asian Conference on Computer Vision. (pp. 31–44). Springer: Berlin, Heidelberg.

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

Mrs.M.K.Vidhyalakshmi has completed her B.E degree in Electronics and Communication Engineering from Bharathidasan University in the year 2002 and M.E degree in Communication Sytems from Trichy Anna University in the year 2011. She is working as an Assistant Professor in the Department of Electronics and Communication Engineering at Tagore Engineering College, Chennai. She has more than Eleven years of teaching experience in the field of Engineering. Currently, she is pursuing part-time research at SRM University, Kattangulathur. She has published more than six papers in international journals. Her area of interest is image processing and computer vision. She is a life member of professional bodies like ISTE, ISSE.

Dr. E. Poovammal is a Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Kattankulaur, Tamilnadu, India. She obtained her B.E. Degree in Electrical and Electronics Engineering from Madurai Kamaraj University in the year 1990, M.E degree in Computer Science and Engineering from Madras University in the year 2002, and a Ph.D. degree in Computer Science and Engineering from SRM University in 2011. Her research interests include Association Rule Mining, Web Mining, Text Mining, Image, and Video mining, Privacy and Security in the Cloud as well as Big Data Analytics. For the last 4 years, she is certified as Adjunct Faculty for the two courses, by the Institute of Software Research, Carnegie Mellon University, Pittsburgh, USA. She has published more than 30 referred journals and presented various international and national conferences. She is the recipient of the "Best Academic Dean award", by the Association of Scientists, Developers, and Faculties (ASDF), 2015 and Recipient of the "Women Engineer award", by IET-CLN, 2013. She is the Fellow IE(I) and a life member of ISTE and Indian Science Congress, IET, IEEE, ACM, and CSI.

Dr. Vidhyacharan Bhaskar received a B.Sc. degree in Mathematics from the University of Madras, Chennai, India in 1992, an M.E. degree in Electrical & Communication Engineering from the Indian Institute of Science, Bangalore in 1997, and M.S.E. and Ph.D. degrees in Electrical Engineering from the University of Alabama in Huntsville in 2001 and 2002, respectively. During 2002–2003, he was a Post-Doctoral fellow with the Communications research group at the University of Toronto, Canada, where he worked on the applications of space-time coding for wireless communication systems. He has served as an Associate Professor in the Department of Information Systems and Telecommunications at the University of Technology of Troyes, France, and Professor in the Department of Electronics and Communication Engineering at S.R.M. University, Kattankulathur, India. Since 2015, he is a Professor in the Department of Electrical and Computer Engineering at San Francisco State University, San Francisco, California, USA. His research interests include MIMO wireless communications, signal processing, error control coding, and queuing theory. He has published 133 Refereed Journal papers, presented around 72 Conference papers in various International Conferences over the past 20 years, published 3 books. He is an IEEE Senior member (SM-IEEE) and is a member of IET (M-IET, UK). He is a Fellow of the Institute of Electronics and Telecommunication Engineers (F-IETE), and a Fellow of the Institute of Engineers (F-IE), Kolkata, India. He is also a Life Member of the Indian Society of Technical Education (LM-ISTE) and a member of the Indian Science Congress (M-ISC).

Mr. J. Sathyanarayanan received his B.E degree in Electronics and Communication Engineering from Madras University and an M.Sc Degree in Physics and Computing with a specialization in Medical Imaging and Computer vision from the University of Manchester. He has more than two decades of experience in Biomedical Engineering and IT domains. He is credited with several professional certifications like CISSP, TOGAF 9.2, PMP, PMI-ACP, AWS Certified Solution Architect Associate, Oracle Certified Master in Java. His interests are in Deep learning and Computer vision. Currently, he is working as an R & D Consultant in Dev Technologies Pvt. Ltd Chennai.