Gated Siamese Convolutional Neural Network Architecture for Human Re-identification

Rahul Rama Varior, Mrinal Haloi, and Gang Wang(✉)

School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore, Singapore
{rahul004,mhaloi,wanggang}@ntu.edu.sg

Abstract. Matching pedestrians across multiple camera views, known as human re-identification, is a challenging research problem that has numerous applications in visual surveillance. With the resurgence of Convolutional Neural Networks (CNNs), several end-to-end deep Siamese CNN architectures have been proposed for human re-identification with the objective of projecting the images of similar pairs (i.e. same identity) to be closer to each other and those of dissimilar pairs to be distant from each other. However, current networks extract fixed representations for each image regardless of other images which are paired with it and the comparison with other images is done only at the final level. In this setting, the network is at risk of failing to extract finer local patterns that may be essential to distinguish positive pairs from hard negative pairs. In this paper, we propose a gating function to selectively emphasize such fine common local patterns by comparing the mid-level features across pairs of images. This produces flexible representations for the same image according to the images they are paired with. We conduct experiments on the CUHK03, Market-1501 and VIPeR datasets and demonstrate improved performance compared to a baseline Siamese CNN architecture.

Keywords: Human re-identification · Siamese Convolutional Neural Network · Gating function · Matching gate · Deep Convolutional Neural Networks

1 Introduction

Matching pedestrians across multiple camera views, also known as human re-identification, is a research problem that has numerous potential applications in visual surveillance. The goal of the human re-identification system is to retrieve a set of images captured by different cameras (gallery set) for a given query image (probe set) from a certain camera. Human re-identification is a very challenging task due to the variations in illumination, pose and visual appearance

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46484-8_48) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VIII, LNCS 9912, pp. 791–808, 2016.
DOI: 10.1007/978-3-319-46484-8_48

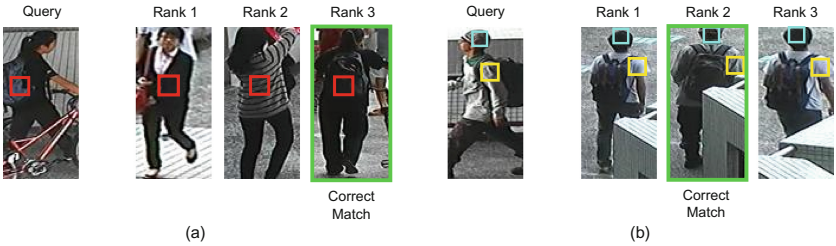

Fig. 1. Example case: Results obtained using a S-CNN. Red, Blue and Yellow boxes indicate some sample corresponding patches extracted from the images along the same horizontal row. See text for more details. **Best viewed in color** (Color figure online)

across different camera views. With the resurgence of Convolutional Neural Networks (CNNs), several deep learning methods [1,21,49] were proposed for human re-identification. Most of the frameworks are designed in a siamese fashion that integrates the tasks of feature extraction and metric learning into a single framework.

The central idea behind a Siamese Convolutional Neural Network (S-CNN) is to learn an embedding where similar pairs (i.e. images belonging to the same identity) are close to each other and dissimilar pairs (i.e. images belonging to different identities) are separated by a distance defined by a parameter called 'margin'. In this paper, we first propose a baseline S-CNN architecture that can outperform majority of the deep learning architectures as well as other handcrafted approaches for human re-identification on challenging human re-identification datasets, the CUHK03 [21], the Market-1501 [57] and the VIPeR [10] dataset.

The major drawback of the S-CNN architecture is that it extract fixed representations for each image without the knowledge of the paired image. This setting results in a risk of failing to capture and propagate the local patterns that are necessary to increase the confidence level (i.e., reducing the distances) in identifying the correct matches. Figure 1(a) and (b) shows two queries and the retrieved matches at the top 3 ranks using a S-CNN architecture. Even though there are obvious dissimilarities among the top 3 matches for a human observer in both the cases, the network fails to identify the correct match at Rank 1. For example, the patches corresponding to the 'bag' (indicated by red boxes) in Fig. 1(a) and the patches corresponding to the 'hat' (indicated by blue boxes) in Fig. 1(b) could be helpful to distinguish between the top retrieved match and the actual positive pairs. However, a network that fails to capture and propagate such finer details may not perform well in efficiently distinguishing positives from hard-negatives.

CNNs extract low-level features at the bottom layers and learn more abstract concepts such as the parts or more complicated texture patterns at the mid-level. Since the mid-level features are more informative compared to the higher-level features, the finer details that may be necessary to increase the similarity for

positive pairs can be more evident at the middle layers. Hence, we propose a gating function to compare the extracted local patterns for an image pair starting from the mid-level and promote (i.e. to amplify) the local similarities along the higher layers so that the network propagates more relevant features to the higher layers of the network. Additionally, during training phase, the mechanisms inside the gating function also boost the back propagated gradients corresponding to the amplified local similarities. This encourages the lower and middle layers to learn filters to extract more locally similar patterns that discriminate positive pairs from negative pairs. Hereafter, we refer to the proposed gating function as 'the Matching Gate' (MG).

The primary challenge in developing the matching gate is that it should be able to compare the local features across two views effectively and select the common patterns. Due to pose change across two views, features appearing at one location may not necessarily appear in the same location for its paired image. Since all the images are resized to a fixed scale, it is reasonable to assume a horizontal row-wise correspondence. Therefore, the matching gate first summarizes the features along each horizontal stripe for a pair of images and compares it by taking the Euclidean distance along each dimension of the obtained feature map. Once the distances between each individual dimensions are obtained, a Gaussian activation function is used to output a similarity score ranging from 0–1 where 0 indicates that the stripe features are dissimilar and 1 indicating that the stripe features are similar. These values are used to gate the stripe features and finally, the gated features are added to the input features to boost them thus giving more emphasis to the local similarities across view-points. Our approach does not require any part-level correspondence annotation between image pairs during the training phase as it directly compares the extracted mid-level features along corresponding horizontal stripes. Additionally, the proposed matching gate is formulated as a differentiable parametric function to facilitate the end-to-end learning strategy of typical deep learning architectures. To summarize, the major contributions of the proposed work are:

- We propose a baseline siamese convolutional neural network architecture that can outperform majority of the existing deep learning frameworks for human re-identification.
- To incorporate run time feature selection and boosting into the S-CNN architecture, we propose a novel matching gate that can boost the common local features across two views. This encourages the network to learn filters that can extract subtle patterns to discriminate hard-negatives from positive pairs. The proposed matching gate is differentiable to facilitate end-to-end training of the S-CNN architecture.
- We conduct experiments on the CUHK03 [21], Market-1501 [57] and the VIPeR [10] datasets for human re-identification and prove the effectiveness of our approach. The proposed framework also achieves promising results compared to the state-of-the-art algorithms.

2 Related Works

2.1 Human Re-identification

Existing research on human re-identification mainly concentrates on two aspects: (1) Developing a new feature representation [5,19,23,28,41,47,48,52] and (2) Learning a distance metric [20,22–24,31,32,38,46]. Novel feature representations were proposed [23,28,41] to address the challenges such as variations in illumination, pose and view-point. Scale Invariant Feature Transforms [27,53,54], Scale Invariant Local Ternary Patterns [23,25], Local Binary Patterns [30,46], Color Histograms [23,46,53,54] or Color Names [48,57] etc. are the basis of the majority of these feature representations developed for human re-identification. Several Metric Learning algorithms such as Locally adaptive Decision Functions (LADF) [22], Cross-view Quadratic Discriminant Analysis (XQDA) [23], Metric Learning with Accelerated Proximal Gradient (MLAPG) [24], Local Fisher Discriminant Analysis (LFDA) [31] and its kernel variant (k-LFDA) [46] were proposed for human re-identification achieving remarkable performance in several benchmark datasets. However, different from all the above works, our approach is modeled based on the Siamese Convolutional Neural Networks (S-CNN) [2,12] that can learn an embedding where similar instances are closer to each other and dissimilar images are distant from each other from raw pixel values.

Deep Learning for Human Re-identification: Convolutional Neural Networks have achieved phenomenal results on several computer vision tasks [13,35,36,39]. In the recent years, several CNN architectures [1,4,21,40,43,45,49] have been proposed for human re-identification. The first Siamese CNN (S-CNN) architecture for human re-identification was proposed in [49]. The system (DML) consists of a set of 3 S-CNNs for different regions of the image and the features are combined by using a cosine similarity as the connection function. Finally a binomial deviance is used as the cost function to optimize the network end-to-end. Local body-part based features and the global features were modeled using a Multi-Channel CNN framework in [4]. Deep Filter Pairing Neural Network (FPNN) was introduced in [21] to jointly handle misalignment, photometric and geometric transformations, occlusion and cluttered background. In [1], a cross-input neighborhood difference module was proposed to extract the cross-view relationships of the features and have achieved impressive results in several benchmark datasets. A recent work [43] also attempts to model the cross-view relationships by jointly learning subnetworks to extract the single image as well as the cross image representations. In [45], domain guided dropout was introduced for selecting the appropriate neuron for the images belonging to a given domain. A Long-Short Term Memory (LSTM) based architecture was proposed in [40] to model the contextual dependencies and selecting the relevant contexts to improve the discriminative capabilities of the local features. Different from all the above works, the proposed matching gate aims at comparing features at multiple levels (different layers) to boost the local similarities and enhance the discriminative capability of the propagated local features. The proposed gating

Table 1. Proposed Baseline Siamese Convolutional Neural Network architecture.

Input	Conv Block - P2	Max Pool	Conv Block - P1	Max Pool	Conv Block - P1	Max Pool	Conv Block	Conv Block	Conv Block	Conv Block
128 × 64	5 × 5 × 3 × 32	2 × 2	3 × 3 × 32 × 50	2 × 2	3 × 3 × 50 × 32	2 × 2	1 × 4 × 32 × 32	1 × 3 × 32 × 32	1 × 3 × 32 × 32	16 × 1 × 32 × 150

ConvBlock - Convolution -> Batch Normalization -> Parametric Rectified Linear Unit
P2 and P1 - zero padding the input with 2 pixels and 1 pixel on all sides respectively before convolution

function is flexible (in architecture) and differentiable to facilitate end-to-end learning strategy of deep neural networks.

2.2 Gating Functions

Gating functions have been proven to be an important component in deep neural networks [15,37]. Gating mechanisms such as the input gates and output gates were proposed in Long-Short Term Memory (LSTM) [15] cells for regulating the information flow through the network. Further, LSTM unit with forget gate [9] was proposed to reset the internal states based on the inputs. Inspired by the LSTM, Highway Networks [37] were proposed to train very deep neural networks by introducing gating functions into the CNN architecture. More recently, 'Trust Gates' were introduced in [26] to handle the noise and occlusion in 3D skeleton data for action recognition. However, the proposed matching gate is modeled entirely in a different context in terms of its architecture and purpose; i.e., the goal of the matching gate is to compare the local feature similarities of input pairs from the mid-level through the higher layers and weigh the common local patterns based on the similarity scores. This will enable the lower layers of the network to learn filters that can discriminate the local patterns of positive pairs from negative pairs. Additionally, to the best of our knowledge, the proposed work is the first of its nature to introduce differentiable gating functions in siamese architecture for human re-identification.

3 Proposed Model

In this section, we first describe our baseline S-CNN architecture and further introduce the Matching Gate to address the limitations of the baseline S-CNN architecture.

3.1 Model Architecture

Baseline Siamese CNN Architecture: The fundamental CNN architecture is modeled in a siamese fashion optimized by the contrastive loss function proposed in [12]. Table 1 summarizes the proposed Siamese CNN architecture. All the inputs are resized to a resolution of 128×64 and the mean image computed on the training set is subtracted from all the images. The description of the

proposed S-CNN layers is as follows. First, we limit the number of pooling layers to only 3 so that it results in less information loss as the features propagate through the network. Second, we also use asymmetric filtering in layers 4–6 to preserve the number of rows at the output of the third layer while reducing the number of 'columns' progressively to 1. This strategy is inspired by the technique introduced in [23] in which the features along a single row is pooled to make the final feature map to a shape (number of rows) $\times 1$. It also helps to reduce the number of parameters compared to symmetric filters. Further, this feature map is fed into a fully connected layer which is the last layer of our network. Finally, we also incorporate some of the established state-of-the-art techniques to the proposed S-CNN architecture. As suggested in VGG-Net [36], we use smaller convolutional filters to reduce the number of parameters to be learned while making the framework deeper. We also employ Batch Normalization [16] for standardizing the distribution of the inputs to each layer which helps in accelerating the training procedure. Parametric rectified linear unit (PReLU) [14] was used as the non-linear activation function as it has shown better convergence properties and performance gains with little risk of over-fitting. More results and analysis about the design choices are given in the supplementary material. The proposed S-CNN architecture outperforms majority of the existing approaches for human re-identification. However, as discussed in Sect. 1, the S-CNN model is not capable of adaptively emphasizing the local features that may be helpful to distinguish the correct matches from hard-negative pairs during run time. Therefore, we propose a matching gate to address this drawback. Below we give the details of the proposed module.

Matching Gate: The proposed matching gate (MG) receives input activations from the previous convolutional block, compares the local features along a horizontal stripe and outputs a gating mask indicating how much more emphasis should be paid to each of the local patterns. Figure 2 illustrates the proposed final architecture with the gating function. The various components of the proposed MG are given below.

1. **Feature summarization:** The feature summarization unit aggregates the local features along a horizontal stripe in an image. This is necessary due to the pose changes of the pedestrian images across different views. For instance, as shown in Fig. 1, the local features (indicated by red, blue and yellow boxes) appearing in one view may not be exactly at the same position in the other view, but it is very likely to be along the same horizontal region.

   Let $\mathbf{x_{r1}} \in \mathbb{R}^{1\times c\times h}$ be the input stripe features from the r^{th} row of a feature map at the input of the MG from one view point and $\mathbf{x_{r2}} \in \mathbb{R}^{1\times c\times h}$ be the corresponding input stripe features from the other view point. Here, c denotes the number of columns and h denotes the depth of the input feature map. Given $\mathbf{x_{r1}}$ and $\mathbf{x_{r2}}$, we propose to use a convolution strategy followed by the parametric rectified linear unit activation (PReLU) to summarize the features along the row resulting in feature vectors $\mathbf{y_{r1}}$ and $\mathbf{y_{r2}}$ respectively with dimensions $\mathbb{R}^{1\times 1\times h}$. The input features $\mathbf{x_{r1}}$ and $\mathbf{x_{r2}}$, are convolved with

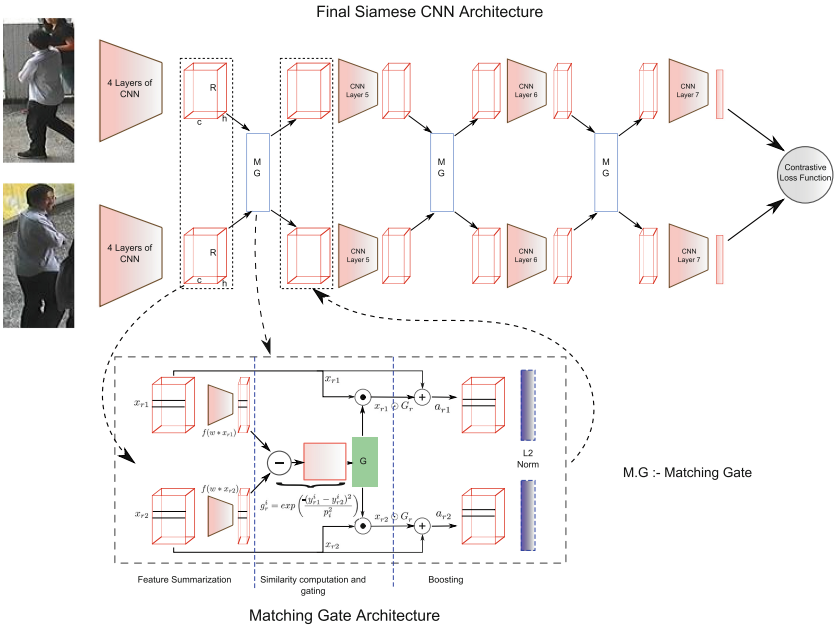

Fig. 2. Proposed architecture: The proposed architecture is a modified version of our baseline S-CNN proposed in Table 1. The matching gate is inserted between layers 4–5, 5–6 and 6–7. The detailed architecture of the gating function is also shown in the figure. See text for details. **Best viewed in color** (Color figure online)

filters $\mathbf{w} \in \mathbb{R}^{1\times c\times h\times h}$ without any padding. This will compute the combination of different extracted patterns along each of the feature maps of $\mathbf{x_{r1}}$ and $\mathbf{x_{r2}}$.

Mathematically, it can be expressed as

$$\mathbf{y_{r1}} = f(\mathbf{w} * \mathbf{x_{r1}}); \quad \mathbf{y_{r2}} = f(\mathbf{w} * \mathbf{x_{r2}}) \tag{1}$$

where '$*$' denotes the convolution operation and $f(.)$ denotes the PReLU activation function. The bias is omitted in Eq. (1) for brevity. The parameters $\mathbf{w}$ and bias of the summarization unit can be learned along with the other parameters of the matching gate through back-propagation.

2. **Feature Similarity computation:** Once the features along a horizontal stripe are summarized across the two views, the similarity between them is computed. The similarity is computed by calculating the Euclidean distance along each dimension 'h' of the summarized features. Computing the distance between each dimension is important as the gating function must have the flexibility to smoothly turn 'on' or turn 'off' each of the extracted patterns in the feature map. Once the distance is computed, a Gaussian activation function is used to obtain the gate values. The value of the Gaussian activation function varies from 0–1 and acts as a smooth switch for the input features.

It also helps the function to be differentiable which is essential for end-to-end training of the S-CNN framework. Mathematically the gating value for each of the dimensions along row 'r' can be obtained as given below;

$$\mathbf{g_r}^i = exp\left(\frac{-(\mathbf{y_{r1}}^i - \mathbf{y_{r2}}^i)^2}{\mathbf{p}_i^2}\right) \tag{2}$$

where $\mathbf{g_r}^i, \mathbf{y_{r1}}^i$ and $\mathbf{y_{r2}}^i$ denotes the i^{th} $(i = \{1, 2, \ldots, h\})$ dimension of the gate values $(\mathbf{g_r})$, $\mathbf{y_{r1}}$ and $\mathbf{y_{r2}}$ respectively for the r^{th} row. The parameter $\mathbf{p}_i$ decides the variance of the Gaussian function and the optimal value can be learned during the training phase. It is particularly important to set a higher initial value for $\mathbf{p}_i$ to ensure smooth flow of feature activations and gradients during forward and backward pass in the initial iterations of the training phase. Further, the network can decide the variance of the Gaussian function for each dimension by learning an optimal $\mathbf{p}_i$.

3. **Filtering and Boosting the features:** Once the gate values $(\mathbf{g_r})$ are computed, each dimension along a row of the input is gated with the corresponding dimension of $\mathbf{g_r}$. The computed gate values will be of dimensions $\mathbb{R}^{1\times1\times h}$ and is repeated c times horizontally to obtain $\mathbf{G_r} \in \mathbb{R}^{1\times c\times h}$ matrix and further an element wise product is computed with the input stripe features $\mathbf{x_{r1}}$ and $\mathbf{x_{r2}}$. This will 'select' the common patterns along a row from the images appearing in both views. To boost these selected common patterns, the input is again added to these gated values. Mathematically, each dimension of the boosted output can be written as

$$\mathbf{a_{r1}}^i = \mathbf{x_{r1}}^i + \mathbf{x_{r1}}^i \odot \mathbf{G_r}^i \tag{3}$$

$$\mathbf{a_{r2}}^i = \mathbf{x_{r2}}^i + \mathbf{x_{r2}}^i \odot \mathbf{G_r}^i \tag{4}$$

$$\mathbf{G_r}^i = [\mathbf{g_r}^i, \mathbf{g_r}^i, \ldots, \mathbf{g_r}^i]_{repeated\ c\ times} \tag{5}$$

where $\mathbf{a_{r1}}^i$, $\mathbf{a_{r2}}^i$, $\mathbf{x_{r1}}^i$, $\mathbf{x_{r2}}^i$, $\mathbf{G_r}^i \in \mathbb{R}^{1\times c\times 1}$. Once the boosted output $\mathbf{a_{r1}}$ and $\mathbf{a_{r2}}$ are obtained, we perform an $L2$ normalization across channels and the obtained features are propagated to the rest of the network. From Eqs. (3) and (4), we can understand that the gradients with respect to the 'selected' $\mathbf{x_{r1}}$ and $\mathbf{x_{r2}}$ will also be boosted during the backward pass. This will encourage the lower layers of the network to learn filters that can extract patterns that are more similar for positive pairs.

The key advantages of the proposed MG is that it is flexible in its architecture as well as differentiable. If the optimal variance factor $\mathbf{p}$ is learned to be high, it facilitates maximum information flow from the input to output and conversely if it is learned to be a low value, it allows only very similar patches to be boosted. The network learns to identify the optimal $\mathbf{p}$ for each dimension from the training data which results in a matching gate that is flexible in its functioning. Alongside learning an optimal $\mathbf{p}$, the network also learns the parameter $\mathbf{w}$ and the bias in Eq. (1) to summarize the features along a horizontal stripe. Additionally, the MG can be inserted in between any layers or multiple layers in the network as it is a differentiable function. This will also facilitate end-to-end learning strategy in deep networks.

Final Architecture: The final architecture of the proposed system is shown in Fig. 2. The baseline network is designed in such a way as to reduce the width of the feature map progressively without reducing the height from layers 4–6. This is essential to address the pose change of the human images across cameras while preserving the finer row-wise characteristics. As shown in the Fig. 2, we inserted the proposed MG between the last 4 layers once the number of rows of the propagated feature maps is fixed.

3.2 Training and Optimization

Input Preparation: Siamese networks take image pairs as inputs. Therefore, we first pair all the images in the training set with a label '1' indicating negative pairs and '0' indicating the positive pairs. For large datasets, the number of negative image pairs will be orders of magnitude higher than the number of positive pairs. To alleviate this bias in the training set, we sample approximately 5 times the number of positive image pairs, as negative image pairs, for each subject. The mean image computed from all the training images is subtracted from all the images and the input pairs are fed to the network.

Training: Both the baseline S-CNN model and the proposed architecture (Fig. 2) are trained from scratch in an end-to-end manner with a batch size of 100 pairs in an iteration. The weight parameters (i.e. filters) of the networks are initialized uniformly following [14]. The gradients with respect to the feature vectors at the last layer are computed from the contrastive loss function and back-propagated to the lower layers of the network. Once all the gradients are computed at all the layers, we use mini batch stochastic gradient descent (SGD) to update the parameters of the network. Specifically, we use the adaptive per-parameter update strategy called the RMSProp [6] to update the weights. The decay parameter for RMSProp is fixed to 0.95 following previous works [17] and the margin for the contrastive loss function is kept as 1. Training is done for 20 epochs with an early stopping strategy based on the saturation of the validation set performance. The initial learning rate is set to 0.002 and reduced by a factor of 0.9 after each epoch. The main hyper-parameter of the MG is the initial value of $\mathbf{p}$. We set this value to 4 initially and the network discovers the optimal value during learning. More details on parameter tuning and validation are given in the supplementary material.

Testing: During testing, each query image has to be paired with all the gallery images and passed to the network. The gating function can selectively boost the common patterns in each image pair. The Euclidean distance between the feature vectors obtained at the last layer is used to compare two input images. Once the distance between the query image and all the images in the gallery set are obtained, it is sorted in ascending order to find the top matches. The above procedure is done for all the query images and the final results are obtained. For an identity with multiple query images, the distances obtained for each query are rescaled in the range of 0–1 and then averaged.

4 Experiments

We provide a comprehensive evaluation of the proposed S-CNN architecture with the matching gate by comparing it against the baseline S-CNN architecture as well as other state-of-the-art algorithms for human re-identification. Majority of the human re-identification systems are evaluated based on the Cumulative Matching Characteristics by treating human re-identification as a ranking problem. However, in [57], human re-identification is treated as a retrieval problem and the mean average precision (mAP) is also reported along with the Rank - 1 accuracy (R1 Acc). For a fair comparison, we report both mAP as well as the performance at different ranks for CUHK03 dataset and mAP and R1 Acc for Market-1501 dataset. For VIPeR dataset, we report only the CMC as it is the relevant measure [57]. All the implementations are done in MATLAB-R2015b and we use the MatConvNet package [42] for implementing all the proposed frameworks. Experiments were run on NVIDIA-Tesla K40 GPU and it took approximately 40–50 minutes per epoch on the CUHK03 dataset.

4.1 Datasets and Settings

Experiments were conducted on challenging benchmark datasets for human re-identification, the Market-1501 [57] dataset, the CUHK03 [21] dataset and the VIPeR [10] dataset. Below, we give the details of the datasets.

Market-1501: The Market-1501 dataset contains 32668 annotated bounding boxes of 1501 subjects captured from 6 cameras and is currently the largest dataset for human re-identification. The bounding boxes for the pedestrian images are obtained by using deformable parts model detectors. Therefore, the bounding boxes are not as ideal as the ones generated by human annotators and there are also several mis-detections which make the dataset very challenging. Following the standard evaluation protocols in [57], the dataset is split into 751 identities for training and 750 identities for testing. We report the single-query (SQ) as well as the multi-query (MQ) evaluation results for this dataset. For multi-query evaluation, the matching scores obtained for each of the query images per identity are rescaled from 0–1 and averaged to obtain the final matching score.

CUHK03: CUHK03 dataset contains 13164 images of 1360 subjects collected on the CUHK campus. Authors of [21] provide two different settings for evaluating on this dataset, 'detected' with automatically generated bounding boxes and 'labeled' with human annotated bounding boxes. All the experiments presented in this paper follow the 'detected' setting as this is closer to the real-world scenario. Following the splitting settings provided in [21], evaluation is conducted 20 times with 100 test subjects and the average result obtained at different ranks is reported. We also use 100 identities from the training set for cross-validation leaving out 1160 identities for training the network.

Table 2. Performance Comparison of state-of-the-art algorithms for the Market-1501 dataset. Proposed baseline S-CNN architecture outperforms the previous works for Market-1501 dataset. The S-CNN architecture with the gating function advances the state-of-the-art results on the Market-1501 dataset.

Method	Rank 1	mAP
SDALF [8]	20.53	8.20
eSDC [54]	33.54	13.54
BoW [57] - (SQ)	34.40	14.09
DNS [50] - (SQ)	61.02	35.68
Ours - Baseline - S-CNN - (SQ)	**62.32**	**36.23**
Ours - With Matching Gate - (SQ)	**65.88**	**39.55**
BoW [57] - (MQ)	42.14	19.20
BoW + HS [57] - (MQ)	47.25	21.88
S-LSTM [40] - (MQ)	61.60	35.31
DNS [50] - (MQ)	71.56	46.03
Ours - Baseline - S-CNN - (MQ)	**72.92**	**45.39**
Ours - With Matching Gate - (MQ)	**76.04**	**48.45**

VIPeR: VIPeR dataset consists of 1264 images belonging to 632 subjects captured using 2 cameras. The dataset is relatively small and the number of distinct identities as well as positive pairs per identity for training are very less compared to the other datasets. Therefore, we conduct data augmentation as well as transfer learning from Market-1501 and CUHK03 datasets. For transfer learning, we remove the last fully connected layer in our baseline S-CNN architecture and then fine-tune the network using the VIPeR dataset. Removing the last fully connected layer was to avoid over-fitting by reducing the number of parameters. For the gated S-CNN framework, the MGs are inserted between layers 4–5 and 5–6. Other experimental settings are kept the same as in [1].

4.2 Results and Discussion

The results for the Market-1501, CUHK03 and VIPeR datasets are given in Tables 2, 3 and 4 respectively. The proposed baseline S-CNN architecture outperforms all the existing approaches for human re-identification for Market-1501 and CUHK03 datasets at Rank 1. We believe that the baseline S-CNN architecture sets a strong baseline for comparison of supervised techniques in future works for both datasets. However, for VIPeR dataset, even though our baseline S-CNN does not achieve the best results, it outperforms several other CNN based architectures [1,43,49]. Our final architecture with the MG improves over the baseline architecture by a margin of 4.2 % and 1.6 % at Rank 1 for CUHK03 and VIPeR datasets respectively. For Market-1501 dataset, our approach outperforms the baseline by a margin of 3.56 % at Rank 1 for single query (SQ) setting and 3.12 % at Rank 1 for multi query (MQ) setting.

Table 3. Performance Comparison of state-of-the-art algorithms for the CUHK03 dataset on the 'detected' setting. Proposed baseline S-CNN architecture outperforms all the previous state-of-the-art methods for CUHK03 dataset at Rank 1. The proposed variant of the S-CNN architecture with the gating function achieves the state-of-the-art results on CUHK03 benchmark dataset. In addition to the results at various ranks, we also provide the mean average precision to analyze the retrieval performance.

Method	Rank 1	Rank 5	Rank 10	mAP
SDALF [8]	4.9	21.0	31.7	_
ITML [7]	5.14	17.7	28.3	_
LMNN [44]	6.25	18.7	29.0	_
eSDC [54]	7.68	22.0	33.3	_
LDML [11]	10.9	32.3	46.7	_
KISSME [18]	11.7	33.3	48.0	_
FPNN [21]	19.9	49.3	64.7	_
BoW [57]	23.0	45.0	55.7	_
BoW + HS [57]	24.3	_	_	_
ConvNet [1]	45.0	75.3	55.0	_
LX [23]	46.3	78.9	88.6	_
MLAPG [24]	51.2	83.6	92.1	_
SS-SVM [51]	51.2	80.8	89.6	_
SI-CI [43]	52.2	84.3	92.3	_
DNS [50]	54.7	84.8	**94.8**	_
S-LSTM [40]	57.3	80.1	88.3	46.3
Ours - Baseline - S-CNN	**63.9**	**86.7**	92.6	**55.57**
Ours - With Matching Gate	**68.1**	**88.1**	94.6	**58.84**

For multi-camera networks, the mean average precision is a better measure for performance compared to the Rank - 1 accuracy [57] as it signifies how many of the correct matches are retrieved from various camera views. Therefore, compared to the improvement in Rank 1 accuracy, the mean average precision which indicates the retrieval accuracy may be more interesting for real-world applications with camera networks. Even though the mean average precision is not particularly important for CUHK03 dataset as it contains only two views, we report the mAP to compare the retrieval results of the proposed final architecture with the baseline S-CNN architecture. It can be seen that our final architecture with MG outperforms the mean average precision obtained by the baseline S-CNN by a margin of 3.32 %, 3.06 % and 3.27 % for Market-1501-Single Query, Market-1501-Multi Query and CUHK03 datasets respectively.

The visualization of the gating mechanism in the proposed matching gate is shown in Fig. 3. Figure 3(a) shows a query image and a hard negative image (example shown in Fig. 1(b)). The middle row shows the average feature

Table 4. Performance Comparison of state-of-the-art algorithms using an individual method for the VIPeR dataset. Proposed S-CNN framework outperforms several previous deep learning approaches for human re-identification [1,49]. Our S-CNN with MG achieves promising results compared to other approaches.

Method	Rank 1	Rank 5	Rank 10
LFDA [31]	24.1	51.2	67.1
eSDC [54]	26.9	47.5	62.3
Mid-level [55]	29.1	52.3	65.9
SVMML [22]	29.4	63.3	76.3
VWCM [52]	30.7	63.0	76.0
SalMatch [53]	30.2	52.3	65.5
QAF [56]	30.2	51.6	62.4
DML [49]	28.2	59.3	73.5
ConvNet [1]	34.8	63.7	75.8
CMWCE [47]	37.6	68.1	81.3
SCNCD [48]	37.8	68.5	81.2
LX [23]	40.0	68.1	80.5
PRCSL [33]	34.8	68.7	82.3
MLAPG [24]	40.7	69.9	82.3
MT-LORAE [38]	42.3	**72.2**	81.6
Semantic [34]	41.6	71.9	**86.2**
S-LSTM [40]	42.4	68.7	79.4
DGDropout [45]	38.6	_	_
SI-CI [43]	35.8	67.4	83.5
SS-SVM [51]	42.7	_	84.3
MCP-CNN [4]	47.8	74.7	84.8
HGD [29]	49.7	79.7	88.7
DNS [50]	51.7	82.1	90.5
SCSP [3]	**53.5**	**82.6**	**91.5**
Ours - Baseline - S-CNN	36.2	65.1	76.3
Ours - With Matching Gate	37.8	66.9	77.4

activations at the output of the 4^{th} convolutional block which is the input to the proposed gating function and the third row shows the obtained gate values using the proposed gating function. It can be seen that for the first few rows where the subject in the query is wearing a hat, the gate activations are low indicating lower similarity where as for a few middle rows, the gate activations are high indicating higher similarity. In Fig. 3(b), we show the image paired with its true positive, the layer 5 inputs and the gate values. It can be seen that for majority of the patches, the gate values are high indicating high similarity between the

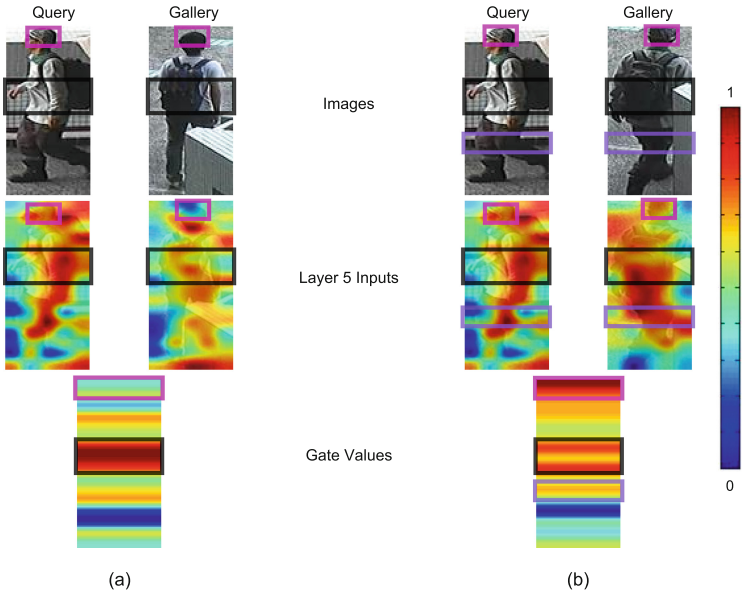

Fig. 3. Gate Visualization: (a) Query paired with its hard-negative (b) Query paired with its positive. Middle row shows the layer 5 input values of all the 4 images and last row shows the corresponding gate values obtained for both pairs. Boxes of same color indicates corresponding regions in the images. **Best viewed in color** (Color figure online)

image patches. This indicates that the gating function can efficiently extract relevant common information from the feature maps of both the images and boost them.

5 Conclusion and Future Works

We have proposed a baseline siamese CNN and a learnable Matching Gate function for siamese CNN that can vary the network behavior during training and testing for the task of human re-identification. The Matching Gate can compare the local features along a horizontal stripe for an input image pair during run-time and adaptively boost local features for enhancing the discriminative capability of the propagated features. The gating function is also designed to be a differentiable one with learnable parameters for adjusting the variance of the gate values as well as for summarizing the horizontal stripe features. This is essential for adjusting the amount of filtering at each stage of the network as well as to facilitate end-to-end learning of deep networks. We have conducted experiments on the Market-1501 dataset, the CUHK03 dataset and the VIPeR dataset to evaluate how run-time feature selection can enable the network to learn more discriminative features for extracting meaningful similarity information for an

input pair. The introduction of the gating function in between convolutional layers results in significant improvement of performance over the baseline S-CNN. Our S-CNN model with the matching gate achieves promising results compared to the state-of-the-art algorithms on the above datasets.

Acknowledgments. The research is supported by Singapore Ministry of Education (MOE) Tier 2 ARC28/14, and Singapore A*STAR Science and Engineering Research Council PSF1321202099.

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at Nanyang Technological University. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

We thank NVIDIA Corporation for their generous GPU donation to carry out this research.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems, vol. 6 (1994)
3. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
5. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference (BMVC) (2011)
6. Dauphin, Y.N., de Vries, H., Chung, J., Bengio, Y.: RMSProp and equilibrated adaptive learning rates for non-convex optimization. CoRR abs/1502.04390 (2015). http://arxiv.org/abs/1502.04390
7. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the International Conference on Machine Learning (ICML) (2007)
8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
9. Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. In: International Conference on Artificial Neural Networks, ICANN 1999 (1999)
10. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) (2007)
11. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: IEEE 12th International Conference on Computer Vision, ICCV 2009 (2009)

12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2006 (2006)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
14. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. CoRR abs/1502.01852 (2015). http://arxiv.org/abs/1502.01852
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
16. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167 (2015). http://arxiv.org/abs/1502.03167
17. Karpathy, A., Johnson, J., Li, F.: Visualizing and understanding recurrent networks. CoRR abs/1506.02078 (2015). http://arxiv.org/abs/1506.02078
18. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P., Bischof, H.: Large scale metric learning from equivalence constraints. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
19. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **35**, 1622–1634 (2013)
20. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37331-2_3
21. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 152–159, June 2014
22. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
23. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
24. Liao, S., Li, S.Z.: Efficient PSD constrained asymmetric metric learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3685–3693 (2015)
25. Liao, S., Zhao, G., Kellokumpu, V., Pietikainen, M., Li, S.: Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
26. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: European Conference on Computer Vision (ECCV) (2016)
27. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. (IJCV) **60**, 91–110 (2004)
28. Ma, B., Su, Y., Jurie, F.: BiCov: a novel image representation for person re-identification and face verification. In: Proceedings of the British Machive Vision Conference (BMVC) (2012)
29. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical Gaussian descriptor for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

30. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **24**, 971–987 (2002)
31. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
32. Rama Varior, R., Wang, G.: Hierarchical invariant feature learning with marginalization for person re-identification. ArXiv e-prints (2015)
33. Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., Wang, J.: Person re-identification with correspondence structure learning. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
34. Shi, Z., Hospedales, T.M., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
35. Shuai, B., Wang, G., Zuo, Z., Wang, B., Zhao, L.: Integrating parametric and non-parametric models for scene labeling. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4249–4258, June 2015
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556
37. Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 2377–2385. Curran Associates, Inc. (2015). http://papers.nips.cc/paper/5850-training-very-deep-networks.pdf
38. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), December 2015
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR abs/1512.00567 (2015). http://arxiv.org/abs/1512.00567
40. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: European Conference on Computer Vision (ECCV) (2016)
41. Varior, R.R., Wang, G., Lu, J., Liu, T.: Learning invariant color features for person re-identification. IEEE Trans. Image Process. **PP**(99), 1 (2016)
42. Vedaldi, A., Lenc, K.: Matconvnet – convolutional neural networks for matlab (2015)
43. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
44. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. (JMLR) **10**, 207–244 (2009)
45. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
46. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_1
47. Yang, Y., Liao, S., Lei, Z., Yi, D., Li, S.Z.: Color models and weighted covariance estimation for person re-identification. In: Proceedings of International Conference on Pattern Recognition (ICPR) (2014)

48. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part I. LNCS, vol. 8689, pp. 536–551. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_35
49. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: Proceedings of International Conference on Pattern Recognition (ICPR) (2014)
50. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
51. Zhang, Y., Li, B., Lu, H., Irie, A., Ruan, X.: Sample-specific SVM learning for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
52. Zhang, Z., Chen, Y., Saligrama, V.: A novel visual word co-occurrence model for person re-identification. In: European Conference on Computer Vision Workshop on Visual Surveillance and Re-identification (ECCV Workshop) (2014)
53. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: IEEE International Conference on Computer Vision (ICCV) (2013)
54. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
55. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identfiation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
56. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
57. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Bu, J., Tian, Q.: Scalable person re-identification: a benchmark. In: IEEE International Conference on Computer Vision (2015)