A Siamese Long Short-Term Memory Architecture for Human Re-Identification

Rahul Rama Varior[†], Bing Shuai[†], Jiwen Lu[§], Dong Xu[‡], and Gang Wang[†,⋆]

† School of Electrical and Electronic Engineering, Nanyang Technological University
§ Department of Automation, Tsinghua University
‡ School of Electrical and Information Engineering, University of Sydney
{rahul004,bshuai001,wanggang}@ntu.edu.sg lujiwen@tsinghua.edu.cn
dong.xu@sydney.edu.au

Abstract. Matching pedestrians across multiple camera views known as human re-identification (re-identification) is a challenging problem in visual surveillance. In the existing works concentrating on feature extraction, representations are formed locally and independent of other regions. We present a novel siamese Long Short-Term Memory (LSTM) architecture that can process image regions sequentially and enhance the discriminative capability of local feature representation by leveraging contextual information. The feedback connections and internal gating mechanism of the LSTM cells enable our model to memorize the spatial dependencies and selectively propagate relevant contextual information through the network. We demonstrate improved performance compared to the baseline algorithm with no LSTM units and promising results compared to state-of-the-art methods on Market-1501, CUHK03 and VIPeR datasets. Visualization of the internal mechanism of LSTM cells shows meaningful patterns can be learned by our method.

Keywords: Siamese Architecture, Long-Short Term Memory, Contextual Dependency, Human Re-Identification

1 Introduction

Matching pedestrians across multiple camera views which is known as human re-identification has gained increasing research interest in the computer vision community. This problem is particularly important due to its application in visual surveillance. Given a probe (query) image, the human re-identification system aims at identifying a set of matching images from a gallery set, which are mostly captured by a different camera. Instead of manually searching through a set of images from different cameras, automated human re-identification systems can save enormous amount of manual labor. However, human re-identification is a challenging task due to cluttered backgrounds, ambiguity in visual appearance, variations in illumination, pose and so on.

⋆ Corresponding author.

Fig. 1. An example of the human re-identification scenario. (a) Result obtained by our framework (b) Result obtained by the baseline algorithm. Correct result retrieved as the best match is shown in green box. It can be observed that, more visually similar images were retrieved by the proposed approach (together with the correct match). Images are taken from the VIPeR dataset [18]. **Best viewed in color.**

Most human re-identification works concentrate on developing a feature representation [33,58,59,63] or learning a distance metric [31,33,57]. With the recent advance of deep learning technologies for various computer vision applications, researchers also developed new deep learning architectures [1, 8, 29, 51, 52, 56, 60] based on Convolutional Neural Networks (CNNs) for the human re-identification task. Most of these handcrafted features as well as learned features have certain limitations. When computing histograms or performing convolution followed by max-pooling operation for example, the features are extracted locally and thus are independent of those features extracted from other regions [48].

In this paper, we explore whether the discriminative capability of local features can be enhanced by leveraging the contextual information, i.e. the features from other regions of the image. Recurrent Neural Network (RNN) architectures have been developed to successfully model such spatial correlations [47, 69] and adapt the local representations to extract more discriminative features. The self recurrent connections in RNNs enable them to learn representations based on inputs that it has previously 'seen'. Thus the features learned by RNNs at any point can encode the spatial dependencies and 'memorize' them. However, not all the spatial dependencies might be relevant for the image under consideration. Ideally, the network should be flexible to allow the propagation of certain contextual information that have discriminative capability and block the irrelevant ones. Thus the ambiguity of features can be reduced and more discriminative features can be learned. A variant of RNNs called Long Short-Term Memory (LSTM) [22] cells have been used to spot salient keywords in sentences [41] and speech inputs [13] to learn context (i.e., topic) relevant information. The advanced gating mechanisms inside the LSTM cell can regulate the information flowing into and out of the cell [19]. The extracted salient contextual information can further enhance the discriminative power of the learned local feature representations.

However, for human re-identification, in the embedded feature space, the feature vectors of similar pairs (i.e., from the same subject) must be 'close' to

each other while the feature vectors from dissimilar pairs should be distant to each other. To this end, we propose a siamese architecture based on LSTM cells. Siamese networks consist of two identical sub-networks joined at the output which are used for comparing two input fields [5]. For learning the network parameters, inputs are therefore given in the form of pairs. The network is optimized by a contrastive loss function [21]. The fundamental idea of the contrastive loss function is to 'attract' similar inputs towards each other and 'repel' dissimilar inputs. As a result, LSTM network can selectively propagate the contexts that can bring together the positive pairs and push apart the negative pairs.

The image is divided into several horizontal stripes and is represented as a sequence of image regions following [69] and starting from the first horizontal stripe, the LSTM cell progressively takes each of the horizontal stripes as inputs and decides whether to retain the information from the current input or discard it based on the information it captured from the current and previous inputs. Similarly, the LSTM cell can hide (or release) the contents of the memory from (or to) the other components of the network at each step. Detailed experimental evaluation of our proposed approach was conducted on three challenging publicly available datasets for human re-identification, Market-1501 [68], CUHK03 [30] and VIPeR [18]. Our approach outperforms a baseline without LSTM units and achieve promising results compared to the state-of-the-art algorithms on all these datasets. We also provide intuitive visualizations and explanations to demonstrate the internal interactions of LSTM cells and prove the effectiveness of our approach. We summarize the major contributions of this paper as follows.

- We adapt the LSTM to human re-identification that can leverage the contextual information to enhance the discriminative capability of the local features. This significantly differs from the traditional methods that perform feature extraction locally and independent of other regions.
- We propose a novel siamese LSTM architecture optimized by the contrastive loss function for learning an embedded feature space where similar pairs are closer to each other and dissimilar pairs are distant from each other.
- Our approach achieves better performance when compared to a baseline algorithm (without LSTM units) as well as promising results when compared to several state-of-the-art algorithms for human re-identification. We also evaluate the multiplicative internal interactions of the LSTM cells and provide intuitive visualizations to demonstrate the effectiveness of our approach.

To the best of our knowledge, this is the first siamese architecture with LSTM as its fundamental component for human re-identification task.

2 Related Works

2.1 Human Re-Identification

Most of the works on human re-identification concentrates on either developing a new feature representation [9, 27, 33, 36, 58, 59, 63] or learning a new distance

metric [28, 31, 33, 42, 57]. Color histograms [33, 57, 64, 65], Local Binary Patterns [39,57], Color Names [59,68], Scale Invariant Feature Transforms [35,64,65] etc are commonly used features for re-identification in order to address the changes in view-point, illumination and pose. In addition to color histograms, the work in [33] uses a Scale Invariant Local Ternary Pattern (SILTP) [34] features and computes the maximal local occurrence (LOMO) features along the same horizontal location to achieve view point invariance. Combined with the metric learning algorithm XQDA [33], LOMO features have demonstrated the state-of-the-art performance on both VIPeR [18] and CUHK03 [30] datasets. But, all the above features are extracted locally and without considering the spatial context which can enhance the discriminative capability of the local representation. Different from the above works, our proposed approach concentrates on improving the discriminative capability of local features by modeling the spatial correlation between different regions within the image. However, we use the state-of-the-art features (LOMO [33]) as the basic local features and further propose a new LSTM architecture to model the spatial dependency. Even though the proposed framework is optimized using the contrastive loss function [21], we would like to point out that any differentiable metric learning algorithms can be used to optimize the proposed siamese architecture.

Deep Learning for Human Re-Identification: Research in deep learning has achieved a remarkable progress in recent years and several deep learning architectures have been proposed for human re-identification [1,8,29,51,52,56,60]. The fundamental idea stems from Siamese CNN (SCNN) architectures [5]. The first SCNN architecture proposed for re-identification [60] consists of a set of 3 SCNNs for each part of the image. In [29], a convolutional layer with max-pooling is used to extract features followed by a patch matching layer which matches the filter responses across two views. A cross-input neighborhood difference module was introduced in [1] to learn the cross-view relationships of the features extracted by a 2-layer convolution and max-pooling layers. Cross-view relationships were also modeled in CNNs by incorporating matching gates [51] and cross-image representation subnetworks [52]. Domain guided dropout was introduced for neuron selection in [56]. Multi-Channel part based CNN was introduced in [8] to jointly learn both the global and local body-parts features. However, these architectures operate on convolved filter responses, which capture only a very small local context and is modeled completely independent of other regions. By using LSTM cells as the fundamental components in the proposed siamese architecture, we exploit the dependency between local regions for enhancing the discriminative capability of local features. Even though a recent work [38] uses RNN for human re-identification, they use it to learn the interaction between multiple frames in a video and not for learning the spatial relationships.

2.2 Recurrent Neural Networks

Recurrent Neural Network (RNN) is a type of deep neural network that has recurrent connections, which enables the network to capture the context in-

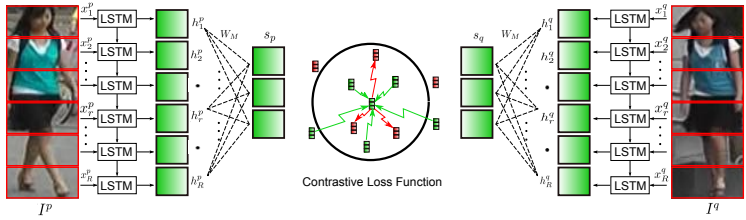

Fig. 2. A diagram showing the proposed siamese LSTM architecture. The LSTM network initially processes the image features ($\mathbf{x_r^p}$ and $\mathbf{x_r^q}$) sequentially to produce their hidden representations $\mathbf{h_r^p}$ and $\mathbf{h_r^q}$ respectively at each step. Once the hidden representations are obtained, a learned mapping ($\mathbf{W_M}$) combines the hidden representations $\mathbf{h_r^p}$ and $\mathbf{h_r^q}$ to obtain $\mathbf{s_p}$ and $\mathbf{s_q}$ respectively. A contrastive loss function is used to compute the loss. Note that dividing the image into 6 rows is merely for illustration and is not exactly the same in our experimental settings. **Best viewed in color.**

formation in the sequence and retain the internal states. RNNs have achieved remarkable success in several natural language processing [41,50], acoustic signal processing [4, 16], machine translation [24, 50] and computer vision [6, 44, 47, 69] tasks. The fundamental idea behind RNN is that the connections with previous states enables the network to 'memorize' information from past inputs and thereby capture the contextual dependency of the sequential data. Due to the difficulty in learning long range sequences (due to the vanishing gradient problem) [3], Long Short Term Memory (LSTM) [22] Networks were introduced and have been successfully applied to several tasks [6, 24, 41]. In addition to capturing the contextual dependency, LSTM can also selectively allow or block the information flow through the network by using its advanced multiplicative interactions in the cell. Several researchers have conducted an empirical evaluation of different RNN architectures and provided intuitive explanations for the internal interactions in these architectures. For more details, we refer the interested readers to [19, 23, 25, 41]. In [13, 41], it has been shown that LSTM cells can detect salient keywords relevant to a topic (context) from sentences or speech inputs. In [48], Pyramidal LSTM was proposed to segment brain images. However, the proposed work aims at building a siamese architecture with LSTM as its fundamental components for human re-identification. To the best of our knowledge, this is the first attempt to model LSTM cells in a siamese architecture for human re-identification.

3 Our Framework

Overview: The goal of our model is to match images of same pedestrians obtained from different surveillance cameras. The proposed siamese architecture consists of two copies of the Long-Short Term Memory network sharing the same set of parameters. The network is optimized based on the contrastive loss

function. Figure 2 illustrates the proposed siamese LSTM architecture. Below, we explain our motivation through the introduction of the Long-Short Term Memory networks. Our proposed siamese architecture is further explained in detail with the optimization methodologies.

3.1 Learning Contextual Dependency using LSTM

RNN architectures have been previously used in [6, 47, 69] to model the spatial dependency and extract more discriminative features for image classification and scene labeling. While encoding such spatial correlations, traditional RNNs do not have the flexibility to selectively choose relevant contexts. Our work is motivated by the previous works [13, 41] which have proven that the LSTM architectures can spot salient keywords from sentences and speech inputs. The internal gating mechanisms in the LSTM cells can regulate the propagation of certain relevant contexts, which enhance the discriminative capability of local features. With these key insights, we propose a siamese LSTM network with pairs of images as inputs for the human re-identification task. The network is modeled in such a way that it accepts the inputs (horizontal stripes) one-by-one and progressively capture and aggregate the relevant contextual information. The contrastive loss function is used to optimize the network parameters so that the learned discriminative features can successfully bring the positive pairs together and push apart the negative pairs. Below, we explain the dynamics of a single layer LSTM architecture without peephole connections.

Long Short-Term Memory Networks LSTM networks were introduced to address the vanishing gradient problem associated with RNNs and a more complete version was proposed in [14, 15, 17] with forget gate, peephole connection and full BPTT training. Mathematically, the update equations of LSTM cell at time t can be expressed as

$$\begin{pmatrix} \mathbf{i_t} \\ \mathbf{f_t} \\ \mathbf{o_t} \\ \mathbf{g_t} \end{pmatrix} = \begin{pmatrix} sigm \\ sigm \\ sigm \\ tanh \end{pmatrix} \mathbf{W_L} \begin{pmatrix} \mathbf{x_t} \\ \mathbf{h_{t-1}} \end{pmatrix} \tag{1}$$

$$\mathbf{c_t} = \mathbf{f_t} \odot \mathbf{c_{t-1}} + \mathbf{i_t} \odot \mathbf{g_t} \tag{2}$$

$$\mathbf{h_t} = \mathbf{o_t} \odot tanh(\mathbf{c_t}) \tag{3}$$

From the above equations, it can be seen that the hidden representation $\mathbf{h_t} \in \mathbb{R}^n$ obtained at each step, Eq. (3), is a function of the input at the current time step ($\mathbf{x_t} \in \mathbb{R}^d$) and the hidden state at the previous time step ($\mathbf{h_{t-1}} \in \mathbb{R}^n$). We use $\mathbf{h_t}$ at each step as the feature representation in our framework. The bias term is omitted in Eq. (1) for brevity. $\mathbf{W_L} \in \mathbb{R}^{4n \times (d+n)}$ denotes the LSTM weight matrix. Sigmoid ($sigm$) and hyperbolic tangent ($tanh$) are the non-linear activation functions which are applied element-wise. $\mathbf{c_t} \in \mathbb{R}^n$ denotes the memory state vector at time t. The vectors $\mathbf{i_t}, \mathbf{o_t}, \mathbf{f_t} \in \mathbb{R}^n$ are the *input*, the *output* and

the *forget* gates respectively at time t which modulates whether the memory cell is written to, reset or read from. Vector $\mathbf{g_t} \in \mathbb{R}^n$ at time t is added to the memory cell content after being gated by $\mathbf{i_t}$. Thus the hidden state vector $\mathbf{h_t}$ becomes a function of all the inputs $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_t}\}$ until time t. The gradients of RNN are computed by back-propagation through time (BPTT) [55].

Internal Mechanisms: The input gate can allow the input signal to *alter* the memory state or *block* it (Eq. (2)). The output gate can allow the memory contents to be *revealed* at the output or *prevent* its effect on other neurons (Eq. (3)). Finally, the forget gate can update the memory cell's state by *erasing* or *retaining* the memory cell's previous state (Eq. (2)). These powerful multiplicative interactions enable the LSTM network to capture richer contextual information as it goes along the sequence.

3.2 The Proposed Siamese LSTM Architecture

For human re-identification, the objective is to retrieve a set of matching gallery images for a given query image. Therefore, we develop a siamese network to take pairs of images as inputs and learn an embedding, where representations of similar image pairs are closer to each other while dissimilar image pairs are distant from each other. All the images in the dataset are paired based on the prior knowledge of the relationships between the images (i.e., similar or dissimilar pairs). Consider an image pair (I^p, I^q) as shown in Figure 2, corresponding to the i^{th} pair of images ($i = \{1, 2, .., N_{pairs}\}$). N_{pairs} indicates the total number of image pairs. Let $Y^i \in [0, 1]$ be the label of the i^{th} pair. $Y^i = 0$ indicates that the images are similar and $Y^i = 1$ indicates that they are dissimilar. Input features are first extracted from the images. Following previous works [33,57,68], the input image is divided into several horizontal stripes and thus, treated as a spatial sequence as opposed to a temporal sequence typically observed in acoustic or natural language inputs. Dividing the image into horizontal rows has the advantage of translational invariance across different view points. We use r as a suffix to denote the local features at a particular region (eg: $\mathbf{x_r}$; $r = \{1, 2, ..., R\}$). R indicates the total number of regions (rows).

Dividing the image into rows has the advantage of translational invariance, which is important for human re-identification and is a commonly adopted strategy to represent image features [33, 68].

Input Features: We extract the state-of-the-art features (LOMO) and Color Names [53] features from the images regions, corresponding to rows.

- Local Maximal Occurrence (LOMO): To extract the LOMO features, first, Color histogram feature and SILTP features [34] are extracted over 10×10 blocks with an overlap of 5 pixels. The feature representation of a row is obtained by maximizing the local occurence of each pattern at the same horizontal location. For an image with 128×64 pixels, this yields 24 rows. Following the same settings, the features are extracted at 3 different scales which resulted in 24, 11 and 5 rows per image.

– Color Names (CN) : Features are extracted for 4×4 blocks with a step-size of 4. Further a row-wise feature representation is obtained by combining the BoW features along the same horizontal location. We follow the same settings and codebook provided in [68] for feature extraction. The final feature representation yields 16 rows for each image with the size of 128×64.

Let the input features from the r^{th} region (row) from the image pairs (I^p, I^q) be $\mathbf{x_r^p}$ and $\mathbf{x_r^q}$ respectively. As shown in Figure 2, the input features $\mathbf{x_r^p}$ and $\mathbf{x_r^q}$ are fed into parallel LSTM networks. Each of these LSTM networks process the input sequentially and the hidden representations $\mathbf{h_r^p}$ and $\mathbf{h_r^q}$ at each step are obtained using Eq. (3). The hidden representation at a particular step is a function of the input at the current step and the hidden representation at the previous step. For example, $\mathbf{h_r^p}$ is a function of $\mathbf{x_r^p}$ and $\mathbf{h_{r-1}^p}$. In the proposed architecture, we use a single layer LSTM. Therefore, the hidden representations $\mathbf{h_r^p}$ and $\mathbf{h_r^q}$ obtained from the LSTM networks are used as the input representations for the rest of the network.

Once the hidden representations from all the regions are obtained, they are combined to obtain $\mathbf{s_p}$ and $\mathbf{s_q}$ as shown below.

$$\mathbf{s_p} = \mathbf{W_M^T}[(\mathbf{h_1^p})^T, ..., (\mathbf{h_r^p})^T, ..., (\mathbf{h_R^p})^T]^T;\ r = 1, 2, ..., R \tag{4}$$

$$\mathbf{s_q} = \mathbf{W_M^T}[(\mathbf{h_1^q})^T, ..., (\mathbf{h_r^q})^T, ..., (\mathbf{h_R^q})^T]^T;\ r = 1, 2, ..., R \tag{5}$$

where $\mathbf{W_M} \in \mathbb{R}^{(R*n)\times(R*n)}$ is the transformation matrix. $[.]^T$ indicates the transpose operator. The objective of the framework is that $\mathbf{s_p}$ and $\mathbf{s_q}$ should be closer to each other if they are similar and far from each other if they are dissimilar. The distance between the samples, $\mathbf{s_p}$ and $\mathbf{s_q}$ $(D_s(\mathbf{s_p}, \mathbf{s_q}))$ can be given as follows:

$$D_s(\mathbf{s_p}, \mathbf{s_q}) = ||\mathbf{s_p} - \mathbf{s_q}||_2 \tag{6}$$

Once the distance between the representations $\mathbf{s_p}$ and $\mathbf{s_q}$ is obtained, it is given as the inputs to the contrastive loss objective function. It can be formally written as:

$$L(\mathbf{s_p}, \mathbf{s_q}, Y^i) = (1 - Y^i)\frac{1}{2}(D_s)^2 + (Y^i)\frac{1}{2}\{max(0, m - D_s)\}^2 \tag{7}$$

where $m > 0$ denotes a margin which acts as a boundary (with the radius m). The intuition behind this loss function is that dissimilar pairs must be separated by a distance defined by m and similar pairs must be as close as possible (i.e., distance tends to 0). For more details regarding the loss function, we refer the interested readers to [21]. The total loss can be obtained by taking the sum of the losses for all pairs .

Network Training: The overall loss is minimized so that similar pairs are closer to each other and dissimilar pairs are separated by m. The system is trained by back-propagating the gradients of Eq. (7) with respect to $\mathbf{s_p}$ and $\mathbf{s_q}$ through the network. While generating the input image pairs, we do not consider all the negative images for a particular identity as it results in a biased

dataset. Following the previous works [1], the number of hard-negatives sampled is twice the number of positive pairs per image. To sample the hard-negatives, we consider the closest matching images in the input feature space (not in the raw image space). Even-though learning frameworks generalize better when trained with large datasets, we perform the training without any data augmentation or fine-tuning operations.

Optimization: We use the mini-batch stochastic gradient descent method with the batch size of 100 pairs. Weight parameters ($\mathbf{W_L}$ and $\mathbf{W_M}$) are initialized uniformly in the range of $[-a, a]$ where $a = \sqrt{1/(input\ size(d) + hidden\ size(n))}$ [2]. In traditional RNN/LSTM architectures, gradients are computed using the BPTT algorithm [55]. In the proposed siamese LSTM architecture, the gradients with respect to the feature vectors extracted from each pair of images are calculated and then back propagated for the respective branches independently using BPTT. As the parameters in each branch are shared, the gradients of the parameters are summed up and then the weights are updated. RMSProp [10] per parameter adaptive update strategy is used to update the weight parameters. Following the previous works [24, 25], we keep the decay parameter as 0.95 and clip the gradients element-wise at 5. These settings are fixed and found to be robust for all the datasets. The only tuned parameters were the hidden vector size (n), learning rate (lr) and margin (m) for the contrastive loss function (see Eq. (7)). Training is performed for a maximum of 20 epochs with an early stopping scheme if the cross-validation performance is found to be saturating. The optimal value for m is tuned by cross-validation and is fixed to 0.5 for all datasets. The optimal values for hidden size, learning rate as well as the learning rate decay coefficient (after every epoch) are dataset dependent.

Testing: During the testing process, the local features for all the query and gallery images are extracted and mapped using the proposed framework to obtain the corresponding representations $\mathbf{s_p}$ ($p = \{1, ..., N_{query}\}$) and $\mathbf{s_q}$ ($q = \{1, ..., N_{gallery}\}$), where N_{query} and $N_{gallery}$ denote the total number of images in the query set and the gallery set, respectively. The total number of query-gallery pairs will be $N_{query} \times N_{gallery}$. The final decision is made by comparing the Euclidean distance (i.e., matching scores) between all $\mathbf{s_p}$ and $\mathbf{s_q}$, $D_s(\mathbf{s_p}, \mathbf{s_q})$. When using multiple features, the matching scores obtained per query image with respect to all the gallery images for each feature are rescaled in the range of $0 - 1$ and then averaged. The final best match is the gallery image that has the least Euclidean distance based on the averaged scores.

4 Experiments

In this section, we present a comprehensive evaluation of the proposed algorithm by comparing it with a baseline algorithm as well as several state-of-the-art algorithms for human re-identification. In most existing human re-identification works, the Cumulative Matching Characteristics (CMC) results were reported. However, in [68], human re-identification is treated mainly as a retrieval problem, so the rank 1 accuracy (R1 Acc) and the mean average precision (mAP) are used

for performance evaluation. For a fair comparison with the baseline as well as the state-of-the-art algorithms, we report both CMC and mAP on all three datasets.

Baseline: To evaluate the performance contribution of the proposed LSTM based siamese network, we implement a baseline method without using LSTM, i.e., with a mapping $\mathbf{W}$ alone. Features from all rows were concatenated and given as input in contrast to concatenating the hidden features from LSTM. Formally, the equations for obtaining $\mathbf{s_p}$ and $\mathbf{s_q}$ using a single layer baseline can be given as follows:

$$\mathbf{s_p} = f(\mathbf{W^T}[(\mathbf{x_1^p})^T, ..., (\mathbf{x_r^p})^T, ..., (\mathbf{x_R^p})^T]^T) \tag{8}$$

$$\mathbf{s_q} = f(\mathbf{W^T}[(\mathbf{x_1^q})^T, ..., (\mathbf{x_r^q})^T, ..., (\mathbf{x_R^q})^T]^T) \tag{9}$$

where $f(.)$ is a non-linear activation function and $\mathbf{W}$ is the parameter matrix that is to be learned. The above system was optimized based on the same contrastive loss function in Eq. (7). We also report the results using a multi-layer baseline which can be obtained by extending the above framework to multiple layers.

4.1 Datasets and experimental settings

The experiments were conducted on 3 challenging human re-identification datasets, Market-1501 [68], CUHK03 [30] and VIPeR [18].

Market-1501: The Market-1501 dataset is currently the largest publicly available dataset for human re-identification with 32668 annotated bounding boxes of 1501 subjects. The dataset is split into 751 identities for training and 750 identities for testing as done in [68]. We provide the multi-query evaluation results for this dataset. For multi-query evaluation, the matching scores for each of the query images from one subject are averaged.

CUHK03: The CUHK03 dataset is a challenging dataset collected in the CUHK campus with 13164 images of 1360 identities from two camera views. Evaluation is usually conducted in two settings 'labelled' with human annotated bounding boxes and 'detected' with automatically generated bounding boxes. All the experiments presented in this paper use the 'detected' bounding boxes as this is closer to the real-world scenario. Twenty random splits are provided in [30] and the average results over all splits are reported. There are 100 identities for testing while the rest of the identities are used for training and validation. For multi-query evaluation, the matching scores from each of the query images belonging to the same identity are averaged.

VIPeR: The VIPeR dataset is another challenging dataset for human re-identification which consist of 632 identities captured from two cameras. For each individual, there is only one image per camera view. A stark change in illumination, pose and environment makes this dataset challenging for evaluating human re-identification algorithms. The dataset is split randomly into equal halves and cross camera search is performed to evaluate the algorithms.

Table 1 shows the performance comparison of the proposed algorithm with the baseline algorithm. It can be seen that the proposed LSTM architecture outperforms the single-layer and multi-layer baseline algorithms for all the datasets.

Table 1. The CMC and mAP on the Market-1501, CUHK03 and VIPeR datasets.

Dataset	Market 1501		CUHK03				VIPeR			
	Rank 1	mAP	Rank 1	Rank 5	Rank 10	mAP	Rank 1	Rank 5	Rank 10	mAP
Baseline (LOMO) - 1 Layer	46.9	21.3	49.1	76.0	85.3	40.1	35.8	62.3	75.0	42.8
Baseline (LOMO) - 2 Layer	47.8	23.9	50.4	77.6	85.9	40.9	36.3	63.6	75.3	42.9
Baseline (LOMO) - 3 Layer	48.4	24.8	51.1	78.3	86.1	41.8	34.8	63.3	75.1	41.5
LSTM (LOMO) - 1 Layer	**51.8**	**26.3**	**55.8**	**79.7**	**88.2**	**44.2**	**40.5**	**64.9**	**76.3**	**45.9**
Baseline (LOMO + CN) - 1 Layer	52.1	27.1	51.6	76.6	85.8	42.1	36.1	64.9	75.6	43.0
LSTM (LOMO + CN) - 1 Layer	**61.6**	**35.3**	**57.3**	**80.1**	**88.3**	**46.3**	**42.4**	**68.7**	**79.4**	**47.9**

Results indicate that feature selection based on the contextual dependency is effective for re-identification tasks. The Rank 1 performance on VIPeR dataset for the 3-layer baseline is lower compared to the 2 layer and 1 layer approach. We believe that this may be due to over-fitting as the dataset is smaller compared to the CUHK03 and Market-1501 datasets. Comparison with the state-of-the-art algorithms is shown in Table 2, 3 and 4. For a fair evaluation, we compare our results to only individual algorithms and **not** to ensemble methods [40,66]. Some qualitative results are shown in Figure 3. It can be seen that the proposed algorithm retrieves visually similar images thereby improving the re-identification rate and mean average precision.

4.2 Parameter tuning

All the parameters are tuned by conducting cross-validation on the training data. In the supplementary material, we have shown the cross validation results using the LOMO [33] features on the Market-1501 dataset. It was observed that when setting the LSTM hidden vector dimension larger than 25, there is no significant improvement in the validation performance. Therefore, we set the hidden dimensions for Market-1501 dataset as 25. Similarly, we observed that the optimal hidden dimensions for CUHK03 and VIPeR datasets were 50. We also observed that the validation performance drops beyond margin $m = 0.75$. For our experiments, we set $m = 0.5$ as there is a slight advantage in the validation performance. For tuning the learning rate, we conduct a log-uniform sampling in the range $[10^{-9},\ 10^{-1}]$. For more detailed information on hyper-parameter search methods, we refer the interested readers to [19].

5 Analysis

5.1 Internal Mechanisms of the LSTM cell

Figure 3 (a) shows two example queries from a testset of the VIPeR dataset and the input gate activations of the **query** image. The retrieved matches for the query image are shown in Figure 3 (b). From the 'response' of the gates to certain inputs, we would like to answer the question whether the LSTM can select and 'memorize' the relevant contextual information and discard the irrelevant ones. The optimal hidden dimensions (n) for the VIPeR dataset was found to be 50 for the LOMO features. Therefore, the gate activations are 50 dimensional vectors

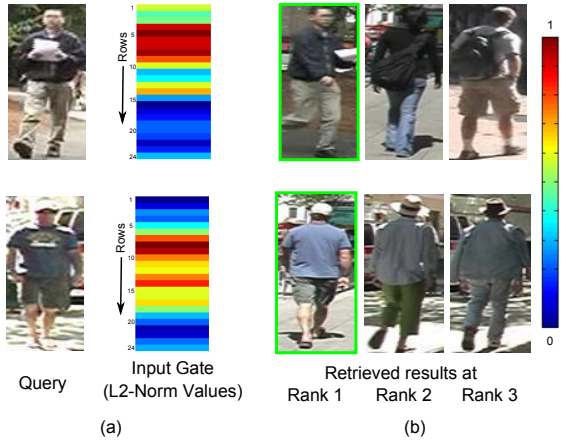

Fig. 3. Qualitative Results: To clearly distinguish between the different values, the gate activations are given as **heat maps.** (a) Two test queries and the $L2$ norm of the LSTM input gate activation values for the **query** image. (b) Retrieved results for the query images. Images shown in green box are the correct match. See text for more details. **Best viewed in color.**

whose values range from $0-1$. In Figure 3 (a), we show the L2 norm values of the gate activations at each step (24 steps in total corresponding to 24 local regions of the image). The L2 norm values are represented as a heat map where values close to 1 (right saturated - information is propagated) is represented by darker shades of red and values closer to 0 (left saturated - information is blocked) by deeper shades of blue. Kindly note that the $L2$ norm value is merely for better illustration and is not actually used in the system.

Input Gate: The input gate values evolve in such a way that the relevant contextual information is propagated and unimportant ones are attenuated. For example, in the Figure 3(a), for the first query image, we can see that the representations from top 3 rows, which mostly contains the background and head portion are attenuated with lower input gate activation values as the information is irrelevant to the context of the image (i.e., the visual appearance of the identity in this case). However, rows $4-9$ which mostly contains information from upper part of the body are selected and propagated. For more step-by-step explanation of the internal mechanisms, we refer the readers to the supplementary material.

5.2 State-of-the-art Comparison

Table 2, 3 and 4 shows the state-of-the-art comparisons on the Market-1501, CUHK03 and VIPeR datasets respectively. For Market-1501 dataset, a recent metric learning approach [61] outperforms ours. However, we believe that it can

Table 2. Performance Comparison of state-of-the-art algorithms for the Market-1501 dataset. Results for [12] and [65] are taken from [68].

Method	Rank 1	mAP
SDALF [12]	20.53	8.20
eSDC [65]	33.54	13.54
BoW + HS [68]	47.25	21.88
DNS [61]	**71.56**	**46.03**
Ours	61.60	35.31

Table 3. Performance Comparison of state-of-the-art algorithms for the CUHK03 dataset.

Method	Rank 1	Rank 5	Rank 10
SDALF [12]	4.9	21.0	31.7
ITML [11]	5.14	17.7	28.3
LMNN [54]	6.25	18.7	29.0
eSDC [65]	7.68	22.0	33.3
LDML [20]	10.9	32.3	46.7
KISSME [26]	11.7	33.3	48.0
FPNN [30]	19.9	49.3	64.7
BoW [68]	23.0	45.0	55.7
BoW + HS [68]	24.3	-	-
ConvNet [1]	45.0	75.3	55.0
LX [33]	46.3	78.9	88.6
MLAPG [32]	51.2	83.6	92.1
SS-SVM [62]	51.2	80.8	89.6
SI-CI [52]	52.2	84.3	92.3
DNS [61]	54.7	**84.8**	**94.8**
Ours	**57.3**	80.1	88.3

be complementary to our approach as the main contribution in this paper is on the feature learning aspect. For CUHK03 dataset, compared to other individual approaches, ours achieve the best results at Rank 1. For VIPeR dataset, several recent approaches [7, 8, 37, 61, 62] outperform our approach. We believe that the reason is lack of positive pairs per image (only 1) and also the lack of total number of distinct training identities compared to other larger datasets. However, to improve the performance on feature learning approaches such as ours, transfer learning from larger datasets or data-augmentation can be employed.

6 Conclusion and Future Works

We have introduced a novel siamese LSTM architecture for human re-identification. Our network can selectively propagate relevant contextual information and thus enhance the discriminative capacity of the local features. To achieve the aforementioned task, our approach exploits the powerful multiplicative interactions within the LSTM cells by learning the spatial dependency. By examining the activation statistics of the input, forget and output gating mechanisms in the LSTM cell, we show that the network can selectively allow and block the context propagation and enable the network to 'memorize' important information. Our

Table 4. Performance Comparison of state-of-the-art algorithms using an individual method for the VIPeR dataset.

Method	Rank 1	Rank 5	Rank 10
LFDA [43]	24.1	51.2	67.1
eSDC [65]	26.9	47.5	62.3
Mid-level [66]	29.1	52.3	65.9
SVMML [31]	29.4	63.3	76.3
VWCM [63]	30.7	63.0	76.0
SalMatch [64]	30.2	52.3	65.5
QAF [67]	30.2	51.6	62.4
SCNN [60]	28.2	59.3	73.5
ConvNet [1]	34.8	63.7	75.8
CMWCE [58]	37.6	68.1	81.3
SCNCD [59]	37.8	68.5	81.2
LX [33]	40.0	68.1	80.5
PRCSL [45]	34.8	68.7	82.3
MLAPG [32]	40.7	69.9	82.3
MT-LORAE [49]	42.3	72.2	81.6
Semantic Representation [46]	41.6	71.9	86.2
DGDropout [56]	38.6	-	-
SI-CI [52]	35.8	67.4	83.5
SS-SVM [62]	42.7	-	84.3
MCP-CNN [8]	47.8	74.7	84.8
HGD [37]	49.7	79.7	88.7
DNS [61]	51.7	82.1	90.5
SCSP [7]	**53.5**	**82.6**	**91.5**
Ours	42.4	68.7	79.4

approach is evaluated on several challenging real-world human re-identification datasets and it consistently outperforms the baseline and achieves promising results compared to the state-of-the-art.

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. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2012)

3. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks (1994)
4. Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In: International Conference on Machine Learning (ICML) (2012)
5. 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 6 (1994)
6. Byeon, W., Breuel, T.M., Raue, F., Liwicki, M.: Scene labeling with lstm recurrent neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
7. 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)
8. 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)
9. 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)
10. 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
11. 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)
12. 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)
13. Fernndez, S., Graves, A., Schmidhuber, J.: An application of recurrent neural networks to discriminative keyword spotting. In: Artificial Neural Networks ICANN 2007. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2007)
14. Gers, F., Schmidhuber, J.: Recurrent nets that time and count. In: Proceedings of the International Joint Conference on Neural Networks, (IJCNN) 2000 (2000)
15. Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with lstm. In: International Conference on Artificial Neural Networks, (ICANN) 1999 (1999)
16. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning (ICML) 2014) (2014)
17. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks (2005)
18. 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)
19. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: A search space odyssey. CoRR abs/1503.04069 (2015), http://arxiv.org/abs/1503.04069

20. 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)
21. 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)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)
23. Jzefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: Proceedings of the International conference on Machine learning, (ICML) (2015)
24. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
25. Karpathy, A., Johnson, J., Li, F.: Visualizing and understanding recurrent networks. CoRR abs/1506.02078 (2015), http://arxiv.org/abs/1506.02078
26. 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)
27. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2013)
28. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Proceedings of Asian Conference on Computer Vision (ACCV) (2012)
29. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
30. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 152–159 (June 2014)
31. 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)
32. 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 (ICCV) (2015)
33. 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)
34. 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)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision (IJCV) (2004)
36. 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)
37. 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)

38. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
39. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2002)
40. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
41. Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R.K.: Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval. CoRR abs/1502.06922 (2015), http://arxiv.org/abs/1502.06922
42. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
43. 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)
44. Pinheiro, P.H.O., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: Proceedings of the 31st International Conference on Machine Learning (ICML) (2014)
45. 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)
46. 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)
47. Shuai, B., Zuo, Z., Wang, G., Wang, B.: Dag-recurrent neural networks for scene labeling. CoRR abs/1509.00552 (2015), http://arxiv.org/abs/1509.00552
48. Stollenga, M., Byeon, W., Liwicki, M., Schmidhuber, J.: Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. CoRR abs/1506.07452 (2015), http://arxiv.org/abs/1506.07452
49. 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) (2015)
50. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. CoRR abs/1409.3215 (2014), http://arxiv.org/abs/1409.3215
51. 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)
52. 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)
53. van de Weijer, J., Schmid, C., Verbeek, J.: Learning color names from real-world images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2007)
54. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research (JMLR) (2009)
55. Werbos, P.: Backpropagation through time: what does it do and how to do it. In: Proceedings of IEEE (1990)

56. 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)
57. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: European Conference on Computer Vision (ECCV) (2014)
58. Yang, Y., Liao, S., Lei, Z., Yi, D., Li, S.Z.: Color models and weighted covariance estimation for person re-identification. Proceedings of International Conference on Pattern Recognition (ICPR) (2014)
59. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: European Conference on Computer Vision (ECCV) (2014)
60. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. Proceedings of International Conference on Pattern Recognition (ICPR) (2014)
61. 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)
62. 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)
63. 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)
64. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: IEEE International Conference on Computer Vision (ICCV) (2013)
65. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
66. 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)
67. 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)
68. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Bu, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Computer Vision, IEEE International Conference on (2015)
69. Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., Chen, Y.: Convolutional recurrent neural networks: Learning spatial dependencies for image representation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2015)