BI-DIRECTIONAL LONG SHORT-TERM MEMORY ARCHITECTURE FOR PERSON RE-IDENTIFICATION WITH MODIFIED TRIPLET EMBEDDING

Weilin Zhong, Huilin Xiong, Zhen Yang, Tao Zhang

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China
Institute for Sensing and Navigation, Shanghai Jiao Tong University, China

ABSTRACT

Matching a specific person across non-overlapping cameras, known as person re-identification, is an important yet challenging task owing to the intra-class variations of the images from the same person in pose, illumination, and occlusion. Most existing body-parts based deep methods simply concatenate the features or scores obtained from spatial parts and ignore the complex spatial correlation between them. In this paper, we present a bi-directional Long Short-Term Memory (Bi-LSTM) architecture that can process the spatial parts sequentially, and enable the messages of different parts to go through in a bi-directional manner. Therefore, the spatial and contextual visual information can be modeled efficiently by the bi-directional connections and the internal gating function in LSTM. Furthermore, we propose a modified triplet loss to learn more discriminative features to distinguish positive pairs from negative pairs. Experiments on CUHK01 and CUHK03 datasets are carried out to demonstrate the effectiveness of the proposed method.

Index Terms— bi-directional information flow, spatial correlation, Long-Short Term Memory, modified triplet

1. INTRODUCTION

Person re-identification aims to identify a specific person among a large number of images obtained across multiple non-overlapping cameras. Recent years, it has drawn increasing attention due to its important and broad applications in visual surveillance. However, person re-identification is also a challenging task because of the large variations of the images from the same person in pose, illumination and background occlusion.

Basically, person re-identification involves two aspects of computation, that is: i) extraction of discriminative features; ii) similarity metric learning. Hand crafted features [1-4] are designed to be robust to illumination change and variations of appearance caused by different camera views. Metric learning methods based on Mahalanobis distance [5-10] are shown to be effective in matching person images to separate the positive pairs from negative pairs.

(a) Examples in CUHK03 [12] (b) Bi-directional information flow

Fig.1. (a) Images in the same row come from the same person. Persons undergo large variations across non-overlapping camera, whereas share similar appearance in the same camera view. Thus triplets with different hard-level caused by camera view should be treated differently. (b) Spatial information can be either passed from top to bottom (Red arrow) or from bottom to top (Green arrow) to verify whether they are the same person.

With the recent advance of deep learning methods in various pattern recognition applications, researchers also develop new deep learning architectures [11-14] based on Convolutional Neural Networks (CNNs) to handle the person re-identification task, in which the feature representation and metric learning are usually jointly learned. However, most of the existing deep methods take the whole image as input [13, 15, 17] and focus only on the global information. As a consequence, the performances of such approaches may still suffer from such factors as illumination variance and occlusion. Inspired by the success of the spatial stripe representation in the hand-crafted features exaction [1, 4], several deep methods are proposed to concentrate on local regions or body-parts [11, 16]. However, simply concatenating features or scores obtained from body-parts and viewing different parts independently do not work well for person re-identification. Recently, Rahul Rama Varior [13] propose a siamese Long Short-Term Memory (S-LSTM) architecture, aiming to enhance the discriminative capability of feature representation such as LOMO feature [1].

Motivated by S-LSTM [13] and the latest deep methods in person re-id [11, 12, 17], we present a bi-directional Long

Short-Term Memory (Bi-LSTM) architecture in combination with modified triplet embedding to model the bi-directional spatial correlations and meanwhile, enhance the discriminative power of deep feature representation. Different from S-LSTM [13], we investigate whether the deep features of spatial parts can be enhanced by leveraging contextual information. Moreover, we modify the triplet loss according to the homogeneity of the image pairs, which means, if the image pairs comes from the same cameras, a more delicate loss function is employed.

2. METHODOLOGY

In this section, we describe the proposed architecture in detail. Our structure firstly splits the input image into four spatial parts and utilizes two convolutional layers to extract deep feature of each spatial part. The weights of the CNN network are shared for every part. Then a bi-directional Long Short-Term Memory (Bi-LSTM) architecture is proposed to bi-directionally pass the spatial information simultaneously, *i.e.*, from top to bottom and from bottom to top. The overall flowchart is illustrated in Fig. 2.

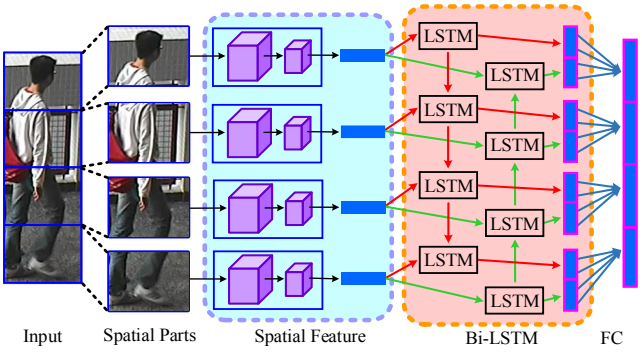


Fig. 2. Diagram of our proposed Bi-LSTM architecture. Here FC denotes fully connected layer.

2.1. Spatial Feature

Here we employ two convolutional layer to extract the deep feature of spatial parts as shown in Fig. 3. Different from the classic deep nets [18, 19], our network generates small feature map and directly flattens all the feature maps to obtain the final spatial feature.

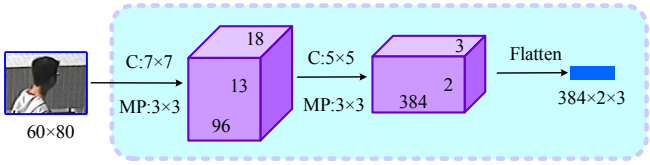


Fig. 3. CNN feature extraction for the spatial part. Here C means the convolution kernel size, and MP means max pooling size.

2.2. Bi-directional LSTM

Long Short-Term Memory Networks (LSTM) is successfully introduced to tackle the gradient vanishing problem in time sequence [20, 21]. Recently, it has been proved that the LSTM architectures can capture and aggregate the relevant contextual information of hand-crafted features [13, 22]. The gating component LSTM is shown in Fig. 4. Apart from considering the current spatial part feature, LSTM also takes previous hidden state as input and learn transitions from previous to current hidden state. Gating connection is incorporated to memorize useful information (memory state) and erase the outdated information (forget gate).

The update of hidden state with previous state can be expressed as

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} M \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \qquad (1)$$
$$h_t = o_t \odot \tanh(c_t)$$

where i_t, f_t, c_t, o_t, h_t are the input gate, forget gate, cell state, output gate and hidden state respectively. Besides, σ is sigmoid function and M is the affine transformation.

We argue that spatial information can be either passed from top to bottom (Red arrow in Fig. 1(b)) or from bottom to top (Green arrow in Fig. 1(b)). As we can see from Fig. 1(b), most of the spatial parts are useful to verify whether a pair of images come from the same person. However, some region information should be selectively passed to next region such as the fourth part occluded by the background in the right-most image in Fig. 1(b). Thus it is critical to introduce gating function in LSTM to learn more discriminative and robust features.

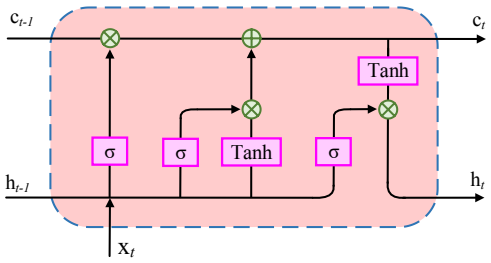


Fig. 4. Internal gating connections of LSTM.

After encoding the spatial correlations by Bi-LSTM, we propose to incorporate a fully connected layer in order to learn a fusing function to map bi-directional feature into fusing feature of each step. The whole feature encoding of Bi-LSTM can be expressed as

$$\begin{cases} h_t^f = \mathrm{LSTM}\left(x_{t-1}, h_{t-1}^f, M^f\right) \\ h_t^b = \mathrm{LSTM}\left(x_{t+1}, h_{t-1}^b, M^b\right) \\ h^t = W\left[cat\left(h_t^f, h_t^b\right)\right] + b \end{cases} \quad (2)$$

where h_t^f , h_t^b , h^t is the forward feature, backward feature, and fusing feature of *t-th* spatial part respectively, (W, b) is learnable parameters of the fusing function.

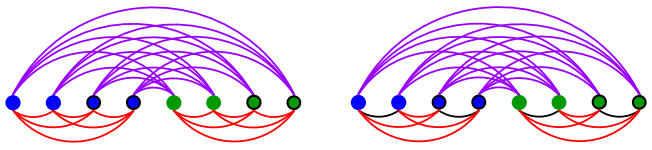
2.3. Modified Triplet

Similar to the works [11, 13, 23], we adopt the triplet loss which pull the distance of positive pairs while push the negative pairs to train our network in an end-to-end manner. Here, we denote anchor, positive and negative image in a triplet as $\left\{\left(x_a^{(i)}, x_p^{(i)}, x_n^{(i)}\right)\right\}$ respectively. The original triplet loss function requires that distance of positive pair be smaller than negative pair by a predefined margin and uses the following loss to penalize those do not satisfy the constraint:

$$L_{triplet} = \frac{1}{N}\sum_i^N \left[D_{ia,ip} - D_{ia,in} + m\right]_+ \quad (3)$$

where $D_{ia,ip} = \left\| f\left(x_a^{(i)}\right) - f\left(x_p^{(i)}\right)\right\|_2$, $D_{ia,in} = \left\| f\left(x_a^{(i)}\right) - f\left(x_n^{(i)}\right)\right\|_2$, $f(\cdot)$ is the concatenated fusing feature from the network, $[\cdot]_+$ operation means the hinge function $\max(0, \cdot)$, N stands for the number of all the triplets in a mini-batch, and m denotes the predefined margin.

In lifted structure [24], anchor and positive $\left(x_a^{(i)}, x_p^{(i)}\right)$ can come from the same camera view as in Fig. 5(a). However, for a specific person, most of the images looks similar in the same view and undergo large variations in cross camera view (Fig. 1(a)). It will convergence badly since many triplets in lifted structure easily satisfy the constraint in Eq. (3) (Such as the dark arcs in Fig. 5(b)). Thus we propose a modified triplet structure which takes into account different hard levels of positive pairs caused by the camera view. For those harder positives (Red arcs in Fig. 5(b)), we impose Eq. (3) on them.



(a) Lifted Triplet Structure (b) Modified Triplet Structure

Fig. 5. Illustration of the generation of triplet in a training mini-batch. Circles with same color denote the same person, and those with dark edges indicate that they come from another camera view. Red arcs and purple arcs represent positive and negative pairs respectively.

In [25], center loss is proposed to successfully assist softmax loss to obtain discriminative deep features. Here, we propose an adapted center loss to impose the following constraint on those easy positives (Dark arcs in Fig 5(b)):

$$L_{center} = \frac{1}{M}\sum_{(j,k)} D_{j,k} \quad (4)$$

where $D_{j,k} = \left\| f\left(x_j\right) - f\left(x_k\right)\right\|_2$, $\left(x_j, x_k\right)$ is the easy positive pair, *i.e.*, they come from the same camera, and M is the number of easy positive pairs.

Different from computing each class center within a training batch in center loss [25], our adapted center loss simply computes the distance of positive pairs to reduce the computation complexity while achieving the same effect of penalizing intra-class variations.

With a balance introduced between triplet loss and center loss, the total modified triplet loss can be written as follows:

$$\begin{aligned} L_{total} &= L_{triplet} + \lambda L_{center} \\ &= \frac{1}{N}\sum_i^N \left[D_{ia,ip} - D_{ia,in} + m\right]_+ + \lambda \frac{1}{M}\sum_{(j,k)} D_{j,k} \end{aligned} \quad (5)$$

where λ is the balance between triplet loss and center loss.

3. EXPERIMENT

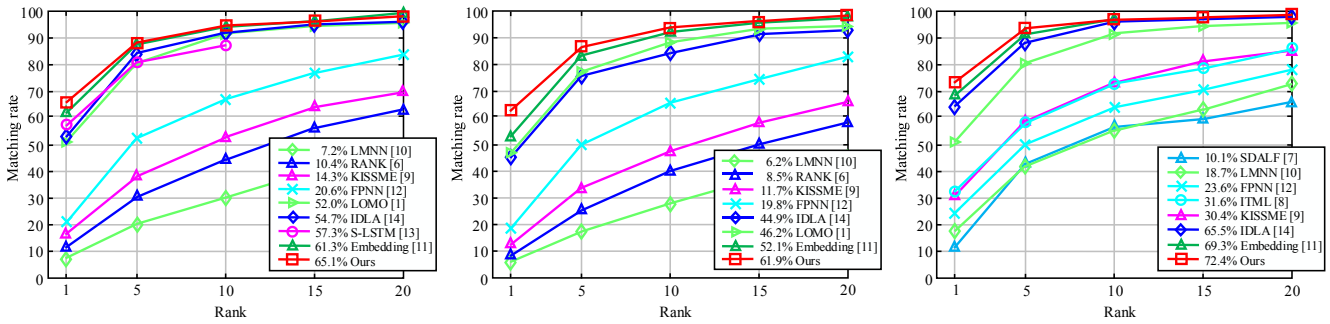
Our proposed method is implemented on Caffe framework [26] and we conduct experiments on two commonly used public datasets: CUHK03 [12] and CUHK01 [27]. We fine-tune the parameters of the first two convolutional layers from Alexnet. In triplet embedding, the L2-norm is incorporated in the last layer to make the network converge faster.

3.1. Experiment Settings and Parameter

Following the same evaluation metric in [11], we adopt single shot experiment setting, to randomly select one image from each view and report Cumulative Match Characteristic (CMC) curve on CUHK03 and CUHK01. CUHK03 contains 1360 persons and there are 13164 images at all. CUHK03 provides two versions, one has manually labeled images, and the other has detected images [12]. CUHK01 is a smaller dataset and contains 971 persons, each of which has 4 images [27]. All images are resized into 250×90. Also they are randomly cropped into 240×80 and vertically mirrored to perform data augmentation.

For fair comparison, 100 person identities are randomly selected and the rest are used for training. The experiment is repeated for ten random splits and the average matching rate is reported to avoid occasionality.

The parameters of our network is nearly the same as [11] and we intuitively set the pre-defined margin as 0.3 and balance as 0.7.



(a) CUHK03 labeled dataset (b) CUHK03 detected dataset (c) CUHK01 dataset

Fig. 6. CMC curves and performance comparison with state-of-the-art approaches on CUHK03 labeled, CUHK03 detected [12] and CUHK01 datasets [27] respectively. Our method gets the best performance.

3.2. Analysis of Proposed Method

To demonstrate the performance of Bi-LSTM and modified triplet embedding, we implement and train three extra models and the performances on CUHK03 labeled dataset are shown in Table 1.

Model	Rank 1	Rank 5	Rank 10	Rank 20
Spatial+Softmax baseline	48.5	76.6	89.4	92.6
Spatial+Bi-LSTM+Softmax	55.4	84.2	91.3	96.3
Spatial+Bi-LSTM+LT	63.8	88.5	94.7	98.2
Spatial+Bi-LSTM+MT	**65.1**	**89.7**	**95.2**	**99.4**

Table 1. Rank1, Rank5, Rank10 and Rank20 matching rate (%) of different models on CUHK03 labeled dataset. Here LT and MT is the abbreviation of lifted triplet and modified triplet respectively.

First, we remove Bi-LSTM and train spatial CNNs with softmax loss whose outputs correspond to the person identities (Spatial + Softmax) as a baseline model. The softmax baseline model achieves 48.5% Rank 1 matching rate. After embracing Bi-LSTM into the network (Spatial + Bi-LSTM + Softmax), the matching rate obtain a large improvement to 55.4%. This indicates that deep feature of spatial parts can be significantly enhanced if spatial correlation is well learned. Besides, performance gains vast increase to 63.8% when the network is trained with lifted triplet structure (Spatial + Bi-LSTM + LT). Rather than simply viewing person re-id as a person classification task in softmax loss, triplet embedding introduce a predefined margin and effectively learn more discriminative features to distinguish positive pairs from negative pairs especially in small datasets. Finally, our Bi-LSTM with modified triplet embedding (Spatial + Bi-LSTM + MT) obtains high performance and achieve 65.1% Rank 1 matching rate. This implies that triplets with different hard-level caused by camera view should be treated differently to better optimize the network.

3.3. Comparison with State-of-the-art Approaches

We compare our model with both traditional and deep state-of-the-art methods. Traditional methods include LOMO [1], RANK [6], SDALF [7], ITML [8], KISSME [9], LMNN [10]. Deep methods include Embedding [11], FPNN [12], S-LSTM [13], IDLA [14].

As shown in Fig. 6, our proposed Bi-LSTM architecture with modified triplet embedding obtains the best performance against state-of-the-art methods on two challenging datasets: CUHK03 (both labeled and detected version) and CUHK01. Especially in CUHK03 detected dataset, our method achieve nearly 10% Rank 1 improvement compared with the second best approach (Embedding [11]). As we know, CUHK03 detected dataset is collected by a pre-trained person detector DPM [28]. Thus persons across camera view share much larger variations than the humanly labeled version. Our modified triplet embedding structure carefully tackle this problem by introducing an adapted center loss. Note that in CUHK03 labeled dataset, our Bi-LSTM without modified triplet embedding (Spatial + Bi-LSTM + LT in Table 1) also achieve better performance than the second best approach.

4. CONCLUSIONS

In this paper, we present a bi-directional Long Short-Term Memory (Bi-LSTM) architecture in combination with modified triplet embedding for person re-identification. Extensive experiments demonstrate the effectiveness of the proposed two strategies, where Bi-LSTM enhances the discriminative capacity of the deep feature of spatial parts and modified triplet structure learns more robust features.

5. ACKNOWLEDGE

The work is partially supported by the National Science Foundation of China under Grant 61673274.

5. REFERENCES

[1] S. Liao, Y. Hu, X. Zhu, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197-2206.

[2] T. Matsukawa, T. Okabe, E. Suzuki, "Hierarchical gaussian descriptor for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1363-1372.

[3] Z. Zhang, Y. Chen, V. Saligrama, "A novel visual word co-occurrence model for person re-identification," in *European Conference on Computer Vision*, 2014, pp. 122-133.

[4] W. S. Zheng, S. Gong, T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 649-656.

[5] F. Xiong, M. Gou, O. Camps, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision*, 2014, pp. 1-16.

[6] B. McFee, G. R. Lanckriet, "Metric learning to rank," in *International Conference on Machine Learning*, 2010, pp. 775-782.

[7] M. Farenzena, L. Bazzani, A. Perina, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360-2367.

[8] J. Davis, B. Kulis, P. Jain, "Information-theoretic metric learning," in *International conference on Machine learning*, 2007, pp. 209-216.

[9] M. Koestinger, M. Hirzer, P. Wohlhart, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288-2295.

[10] M. Hirzer, P. M. Roth, H. Bischof. "Person re-identification by efficient impostor-based metric learning," in *International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 203-208.

[11] H. Shi, Y. Yang, X. Zhu, "Embedding deep metric for person re-identification: A study against large variations," in *European Conference on Computer Vision*, 2016, pp. 732-748.

[12] W. Li, R. Zhao, T. Xiao, "Deepreid: Deep filter pairing neural network for person re-identification" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152-159.

[13] R. Varior, B. Shuai, J. Lu, "A siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*, 2016, pp. 135-153.

[14] E. Ahmed, M. Jones, T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908-3916.

[15] S. Z. Chen, C. C. Guo, J. H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353-2367, 2016.

[16] J. Wang, Z. Wang, C. Gao, "DeepList: learning deep features with adaptive listwise constraint for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, issue 99, 2016.

[17] L. Lin, G. Wang, W. Zuo, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, issue 99, 2016.

[18] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097-1105.

[19] K. He, X. Zhang, S. Ren, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[20] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[21] F. A. Gers, J.Schmidhuber, "Recurrent nets that time and count," in *International Joint Conference on Neural Networks*, 2000, pp. 189-194.

[22] S. Fernández, A. Graves, J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *International Conference on Artificial Neural Networks*, 2007, pp. 220-229.

[23] F. Schroff, D. Kalenichenko, J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815-823.

[24] H. O. Song, Y. Xiang, S. Jegelka, "Deep metric learning via lifted structured feature embedding," *arXiv preprint arXiv: 1511.06452*, 2015.

[25] Y. Wen, K. Zhang, Z. Li, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016, pp. 499-515.

[26] Y. Jia, , E. Shelhamer, , J. Donahue, "Caffe: convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*, 2014, pp. 675-678.

[27] W. Li, R. Zhao, X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*, 2012, pp. 31-44.

[28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, "Object detection with discriminatively trained part-based models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645.