

PersonNet: Person Re-identification with Deep Convolutional Neural Networks

Lin Wu, Chunhua Shen, Anton van den Hengel

Abstract—In this paper, we propose a deep end-to-end neural network to simultaneously learn high-level features and a corresponding similarity metric for person re-identification. The network takes a pair of raw RGB images as input, and outputs a similarity value indicating whether the two input images depict the same person. A layer of computing neighborhood range differences across two input images is employed to capture local relationship between patches [1]. This operation is to seek a robust feature from input images. By increasing the depth to 10 weight layers and using very small (3×3) convolution filters, our architecture achieves a remarkable improvement on the prior-art configurations. Meanwhile, an adaptive Root-Mean-Square (RMSProp) gradient decent algorithm is integrated into our architecture, which is beneficial to deep nets. Our method consistently outperforms state-of-the-art on two large datasets (CUHK03 and Market-1501), and a medium-sized data set (CUHK01).

Index Terms—Person re-identification, Convolutional neural networks, Deep metric learning.

I. INTRODUCTION

THE task of person re-identification (re-id) is to match pedestrian images observed from multiple non-overlapping camera views with varied visual features [2]–[10]. Challenges are presented in the form of compounded variations in visual appearance across different camera views, human poses, illuminations, background clutter, occlusions, relatively low resolution and the different placement of the cameras. Some examples are shown in Fig. 1.

Person re-identification is essentially to measure the similarity for pairs of pedestrian images in a way that a pair is assigned a high similarity score in case of depicting the same identity and a low score if displaying different identities. This typically involves constructing a robust feature representation and an appropriate similarity measure in order to estimate accurate similarity scores. To this end, many methods focusing on feature representation and distance function learning are designed separately or jointly to deal with person re-id problem. Low-level features such as color and texture can be used for this purpose. Some studies have obtained more distinctive and reliable feature representations, including symmetry-driven accumulation [11], horizontal partition [12], [13], and saliency matching [7], [10]. However, it is still difficult to design a type of feature that is discriminative and invariant to severe changes in terms of misalignment across disjoint camera views. Another pipeline of person re-id system

is to learn a robust distance or similarity function to deal with complex matching problem. Many metric learning algorithms are proposed for this purpose [6], [8], [9], [13]–[17]. In practice, most of metric learning methods exhibit a two-stage processing which typically extract hand-crafted features and subsequently learn the metrics. Thus, these approaches often lead to sub-optimal solutions.

Convolutional Neural Networks (CNNs) have proven highly successful at image recognition problems and various surveillance applications including pedestrian detection [18], [19], and tracking [20]. However, little progress is witnessed in person re-id, except a few works in [1], [21], [22]. By applying CNNs, a joint feature representation and metric learning can be achieved. The FPN algorithm [21] makes the first attempt to introduce patch matching in CNNs, followed by an improved deep learning framework [1] where layers of cross-input neighborhood differences and patch summary are added. These two methods both evaluate the pair similarity early in the CNN stage, so that it could make use of spatial correspondence of feature maps. In fact, spatial misalignment is very notable in person re-id due to similar appearance or occlusion [23]. As a result, a more deep model is demanded to well address this challenge by faithfully capturing non-linear relationship between patches. We aim to improve the state-of-the-art architecture of [1] to achieve better accuracy. Specifically, we increase the depth of the network in [1] by adding more convolutional layers, which is feasible due to the use of very small (3×3) convolution filters in all layers.

By increasing the depth of AlexNet [24] using an architecture with very small convolution filters, VGG network [25] has shown a significant improvement over the prior-art configurations and generalise well to other databases. Encouraged by these positive results from VGG model, we deepen a state-of-the-art network [1] on person re-id task while achieving notable improvement. Our PersonNet consists of ten layers with weights and very small 3×3 receptive fields throughout the whole net. In training stage, we dynamically sample pairs of images in an online manner where the magnitudes of gradients can vary widely for different layers, especially in very deep nets. To this end, we introduce an adaptive root-mean-square (RMS) gradient decent algorithm, RMSProp [26], [27], which works by dividing the gradient by a running average of its recent magnitude. This adaptive gradient decent algorithm is more suitable to deep layers, and converges much faster than Stochastic Gradient Descent (SGD). We illustrate the architecture in Section III.

The main contributions of this paper are three-fold:

- We present a deep network of increasing depth using



Fig. 1: Typical samples of pedestrian images in person re-identification from CUHK03 data set [21]. Each column shows two images of the same individual observed by two different camera views.

an architecture with very small (3×3) convolution filters, which shows a notable improvement on person re-id by pushing the depth to 7-10 weight layers.

- We employ a different mini-batch gradient decent algorithm in back propagation, RMSProp, which adjusts the magnitudes of the gradients in each mini-batch sampled online. The integration of RMSProp makes our network reach convergence more quickly.
- Extensive experiments are conducted on benchmark datasets to validate the effectiveness of our architecture. We achieve *the best reported results* on three popular benchmark datasets.

II. RELATED WORK

Many recent studies on person re-identification attempt to generate robust feature representation which is discriminative and robust for describing a pedestrian's appearance under various changes and conditions [2], [4], [7], [11], [28]–[30]. Bazzani *et al.* [28] represent a person by a global mean color histogram and recurrent local patterns through epitomic analysis. Farenzena *et al.* [11] propose the symmetry-driven accumulation of local features (SDALF) which exploits both symmetry and asymmetry, and represents each part of a person by a weighted color histogram, maximally stable color regions and texture information. Gray and Tao [29] propose to use AdaBoost to select good features out of a set of color and texture features. Schwartz and Davis propose a discriminative appearance based model using partial least squares, in which multiple visual features: texture, gradient and color features are combined [31]. Recently, saliency information has been investigated for person re-id [7], [10], leading to a novel feature representation and improved discriminative power in person re-id. In [7], a method of (eSDC) is presented to learn salience for persons under deformation. Moreover, salience matching and patch matching can be integrated into a unified RankSVM framework (SalMatch [10]). They also propose mid-level filters (MidLevel) for person re-identification by exploring the partial area under the ROC curve (pAUC) score [4]. Lisanti *et al.* [32] leverage low-level feature descriptors to approximate the appearance variants in order to discriminate individuals by using sparse linear reconstruction model.

Metric learning approaches to person re-id is to essentially formalize the problem as a supervised metric/distance learning where a projection matrix is sought out so that the projected Mahalanobis-like distance is small when feature vectors represent the same person and large otherwise. Among many metric learning methods, Large Margin Nearest Neighbor Learning (LMNN) [33], Information Theoretic Metric Learning (ITML) [34], and Logistic Discriminant Metric Learning (LDM) [35] are three representative methods. By applying these metric learning methods into person re-id, many effective approaches are developed [6], [8], [9], [13]–[17], [36]. Mignon *et al.* [16] proposed Pairwise Pairwise Constrained Component Analysis (PCCA) to learn a projection into a low dimensional space in which the distance between pairs of samples respects the desired constraints, exhibiting good generalization properties in the presence of high dimensional data. Zheng *et al.* [13] presented a Relative Distance Comparison (RDC) to maximize the likelihood of a pair of true matches having a relatively smaller distance than that of a mismatched pair in a soft discriminant manner. Koestinger *et al.* propose the large-scale metric learning from equivalence constraint (KISSME) which considers a log likelihood ratio test of two Gaussian distributions [15]. Li *et al.* propose the learning of locally adaptive decision functions (LADF), which can be viewed as a joint model of distance metrics and locally adapted thresholding rules [9]. The Cross-view Quadratic Discriminant Analysis (XQDA) algorithm learns a discriminant subspace and a distance metric simultaneously, which is able to perform dimension reduction and select the optimal dimensionality. To make the metric learning more efficient, they further present a positive semidefinite constrained method to reduce the computation cost and get more robust learned metric [23]. In [12], an efficient feature representation called Local Maximal Occurrence is proposed, followed by a subspace and metric learning method. Last but not least, learning to rank can be employed in person re-id, and approaches include ensembled RankSVM [37], Metric Learning to Rank (MLR) [38] and its application to person re-id [36] and structured metric ensembles [39].

Three deep learning based person re-id algorithms have been proposed [1], [21], [22]. Yi *et al.* [22] utilized a Siamese CNN with a symmetry structure comprising two independent sub-nets, and then employed cosine distance as their metric. Li *et al.* [21] designed a different network, which begins with a single convolution layer with max pooling, followed by a patch-matching layer that multiplies convolutional feature responses from the two inputs at a variety of horizontal offsets. The most similar work to us is JointRe-id [1] where a layer of computing cross-input neighborhood difference features is introduced after two layers of convolution and max pooling.

Our architecture differs substantially from these previous networks. The network is very deep with very small (3×3) convolution filters. This has a significant improvement based on the prior-art configuration by pushing the depth to 7-10 weight layers. Moreover, an adaptive gradient decent algorithm, RMSProp is used in our network, which can be immune to initialisation and the instability of gradient in deep nets. Consequently, our network outperforms all previous approaches

TABLE I: Layer parameters of PersonNet. The output dimension is given by height \times width \times width. FS: filter size for convolutions. Layer types: C: convolution, MP: max-pooling, FC: fully-connected. All convolution and FC layers use hyperbolic tangent as activation function.

Name	Type	Output Dim	FS	Stride
Conv0	C	$157 \times 57 \times 32$	3×3	1
Pool0	MP	$79 \times 29 \times 32$	2×2	2
Conv1	C	$76 \times 26 \times 32$	3×3	1
Pool1	MP	$38 \times 13 \times 32$	2×2	2
Conv2	C	$35 \times 10 \times 32$	3×3	1
Conv3	C	$32 \times 7 \times 32$	3×3	1
Difference	-	$32 \times 7 \times 32$	3×3	3
Conv4	C	$32 \times 7 \times 32$	3×3	1
Pool4	MP	$15 \times 2 \times 32$	2×2	2
FC1	FC	-	4096	-
FC2	FC	-	4096	-
FC3	FC	-	512	-

on the largest Market-1501 data set [40], CUHK03 [21] and smaller CUHK01 datasets [41].

III. THE ARCHITECTURE

During training, the input to our PersonNet is a pair of fixed-size 160×60 RGB images. The pair of images is passed through a stack of tied convolutional layers, where we use filters with a very small receptive field: 3×3 . The convolution stride is fixed to 1 pixel. Spatial pooling is carried out by three max-pooling layers, which follow some of the convolution layers (not all the convolution layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2. After a stack of convolution layers, we have three fully-connected layers where the first two have 4096 dimension and the third is 512, which is put through softmax to determine the pair is same or different. The overall architecture of the proposed PersonNet is shown in Fig.2. We show details of these layers in Table I.

A. Convolution and max pooling

The first two layers are convolutional and max-pooling layers. Given two pedestrian images I and J observed by two different camera views with three color channels and sized 160×60 , the convolutional layer outputs local features extracted by filter paired. The filters (W, V) applied to two camera views are shared. Given the input I and J , consisting of C channels of height H and width W , if we use K filters and each filter is in size of $m \times m \times C$, the output consists of a set of C' channels of height H' and width W' . We define the filter functions as $f, h : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H' \times W' \times C'}$

$$\begin{aligned} f_{ij}^k &= \sigma((W_k * I)_{ij} + b_k^I) \\ h_{ij}^k &= \sigma((V_k * J)_{ij} + b_k^J), \end{aligned} \quad (1)$$

where $k \in \mathbb{R}^{K \times K \times C}$. Rather than using relatively large receptive fields in the first two convolutional layers (e.g., 5×5 in [1], [21]), we use small receptive fields with 3×3 throughout the whole net to convolute with the input at every pixel with stride of 1. Apparently, a stack of two 3×3 convolution layers (without spatial pooling between them) amounts to working as a receptive field of 5×5 . By doing this, more non-linear activation functions are embedded which can make the decision function more discriminative [25].

Activation function can increase the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolutional layer. Instead of using ReLU, $\sigma(x) = \max(0, x)$, as the activation function in deep network, we choose the nonlinear activation function $\sigma(x)$ to be hyperbolic tangent function, $\sigma(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, which can rescale the linear output in the range $[-1, 1]$. Scaling the activation function to be $\sigma(x) = \tanh(\frac{3x}{2})$ is able to ensure the training is spread uniformly over each layer particularly helpful in networks with very deep layering.

The max-pooling layer is used to reduce the dimensionality of the output and variance in deformable objects to ensure that the same result will be obtained even when image features undergo slight translations. The max-pooling operation is applied on every pixel around its neighborhood.

B. Modeling neighborhood patch matching

The patch matching layer is to compute the differences in filter responses of local patches across two views. Since we have $f_I, g_J \in \mathbb{R}^{32 \times 7}$, the difference around the neighborhood of each feature location yields to a set of feature maps $K_i \in \mathbb{R}^{32 \times 7 \times 3 \times 3}$ ($i = 1, \dots, 32$), where 3×3 is the window size of neighborhood around a feature value. In other words, K_i indicates a 32×7 grid of 3×3 blocks, $K_i(x, y) \in \mathbb{R}^{3 \times 3}$ where $1 \leq x \leq 7$ and $1 \leq y \leq 32$. Following [1], we have

$$K_i(x, y) = f_I(x, y)\mathbb{I}(3, 3) - \mathcal{N}[h_J(x, y)] \quad (2)$$

where $\mathbb{I} \in \mathbb{R}^{3 \times 3}$ is a 3×3 indicator matrix with all elements being 1s, and $\mathcal{N}[h_J(x, y)] \in \mathbb{R}^{3 \times 3}$ is a 3×3 neighborhood of h_J centered at (x, y) . Here we use a small neighborhood of size 3×3 to model the displacement of body parts caused by pose and viewpoint variations [21]. Our architecture can avoid symmetric operation on computing K_i because we use online sampling to generate pairs of images.

The visualization of feature responses at each layer of the network are shown in Fig.3. We can see that after Conv0, the features respond to bright regions of the images. After a few convolutions and max-pooling, higher responses are given to body as a whole. In this process, part-based CNNs may be beneficial to further improve the accuracy of recognition since human body parts can be very different across camera views and matching different parts are helpful in matching. Recall from III-B and [1], the neighborhood layer is to compute the difference of corresponding feature maps across two views in a small range. This can robustly match some patches that undergo variations in viewpoints, and poses. For a negative pair, a neighborhood difference layer can highlight some local patches that are visually different, as shown in Fig.3 (a). By contrast, for a positive pair, the difference map is expected to be close to zero and nonzero values should be small and uniformly distributed across the map, as shown in Fig.3 (b). The difference layer is followed by another patch summary layer that extracts these difference maps into a holistic representation of the differences in each 3×3 block. Then, we use another convolution layer with max pooling to learn spatial relationships across neighborhood differences. The network ends up with three fully connected layers with softmax output.

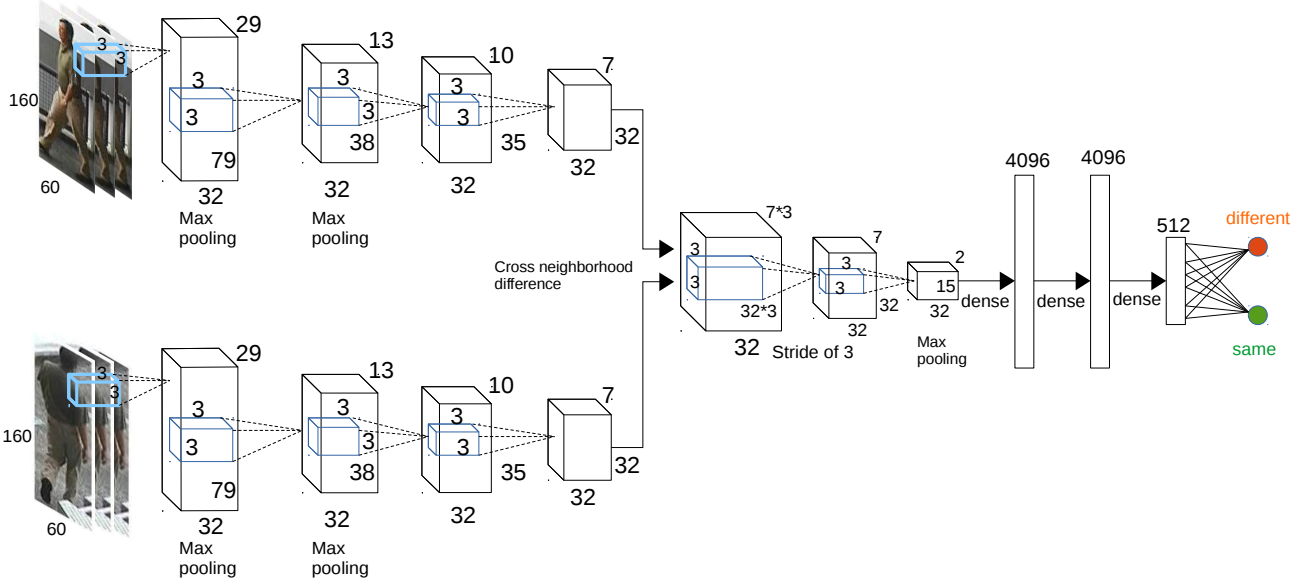


Fig. 2: The architecture of PersonNet. The network takes a pair of RGB images as input, which is put through a stack of convolution layers, matching layer, and higher layers computing relationships between them. The configurations of each layer are shown in Table I.

IV. TRAINING STRATEGIES

We use the hyperbolic tangent function as the non-linear activation function in our models. Training data are divided into mini-batches. Note that we employ RMSProp [26] instead of commonly used stochastic gradient decent (SGD) as a means of updating gradients for parameters.

A. RMSProp

RMSProp works by dividing the gradient by a running average of its recent magnitude. The main difference between SGD and RMSProp stems from the way of making use of gradient. The idea behind SGD is that when the learning rate is small, it averages the gradients over successive mini-batches. However, the magnitude of the gradient can be very different for different weights and can change during learning. Hence, it is impractical to choose a single learning rate. To work efficiently with mini-batches, RMSProp is developed to use the gradient but divide by a different number for each mini-batch and the number divided by is similar for adjacent mini-batches. Thus, RMSProp keeps the a moving average of the squared gradient for each weight:

$$\text{MeanSquare}(\mathbf{w}, t) = 0.9 * \text{MeanSquare}(\mathbf{w}, t - 1) + 0.1 * \left(\frac{\partial E}{\partial \mathbf{w}^{(t)}} \right)^2$$

$$\partial \mathbf{w}^{(t)} = \epsilon \frac{\partial E}{\partial \mathbf{w}^{(t)}} / \left(\text{MeanSquare}(\mathbf{w}, t)^{\frac{1}{2}} + \mu \right)$$

where \mathbf{w} is the weight parameter, t is the time step, ϵ is the learning rate, μ is a smoothing value for numerical convention, and E denotes the error surface. A recent study [26] has shown that dividing the gradient by $(\text{MeanSquare}(\mathbf{w}, t))^{\frac{1}{2}}$ makes the learning work much better. The introduction of RMSProp is beneficial to our architecture, which performs more robustly than SGD.

B. Data augmentation and data balancing

In the training set, the matched (positive) pairs are several orders fewer than non-matched (negative) pairs, which can lead to data imbalance and overfitting. To circumvent this issue, we augment data set by performing 2D translation on each pedestrian image. Specifically, following [1], [21] for an original image of size $H \times W$, five images of the same size are randomly sampled around the original image center, with translation drawn from a uniform distribution in the range $[-0.05H, 0.05H] \times [-0.05W, 0.05W]$. For CUHK01 dataset, we also horizontally reflect each image.

To achieve data balancing, we online sample the same number of negative and positive pairs instead of generating the proportion of negative pairs against positives in a fixed manner. For example, a common way as conducted in [1], [21] is to generate the same size of negatives and positives, then gradually increase the number of negative samples up to the ratio of 5:1. Such operation is unable to learn a robust and reliable network that tolerate the varied data distributions in each mini-batch. Nonetheless, our online sampling strategy can well address the aspect of data balance.

V. EXPERIMENTAL RESULTS

We implemented our network using Theano deep learning framework [42]. The training of the network converges in roughly 20-22 hours in NVIDIA GeForce GTX 980 GPU. The training is carried out by optimising the softmax objective using online sampling of each input pair of images with RMSProp gradient decent. The mini-batch size was set to 2. The training was regularised by L_2 penalty and dropout [24] regularisation for the first two fully-connected layers (dropout ratio set to 0.5) in order to alleviate over-fitting. The learning

rate was initially set to 0.05, and then decreased by a factor of 10 when the validation set accuracy stopped improving. In general, the learning was stopped within 100K iterations. We conjecture that in spite of the larger number of parameters and the greater depth of our nets compared with FPNN and JointRe-id, the nets required less iterations to converge due to (a) the implicit regularisation induced by greater depth and smaller convolution sizes; (b) adaptive gradient decent algorithm of RMSProp.

A. Experimental settings

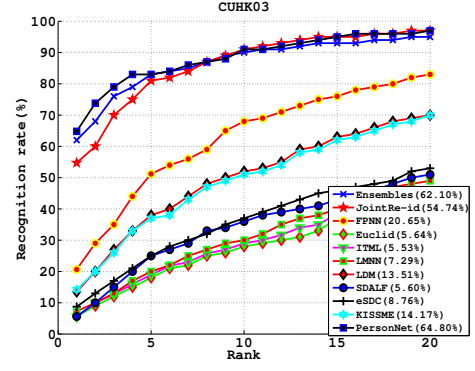
a) *Datasets.*: We perform experiments on three benchmarks: the CUHK03 dataset [21], the CUHK01 dataset [41] and the Market-1501 dataset [40].

b) *Evaluation protocol.*: We adopt the widely used single-shot modality in our experiment to allow extensive comparison. Each probe image is matched against the gallery set, and the rank of the true match is obtained. The rank- k recognition rate is the expectation of the matches at rank k , and the cumulative values of the recognition rate at all ranks are recorded as the one-trial Cumulative Matching Characteristic (CMC) results [39]. This evaluation is performed ten times, and the average CMC results are reported.

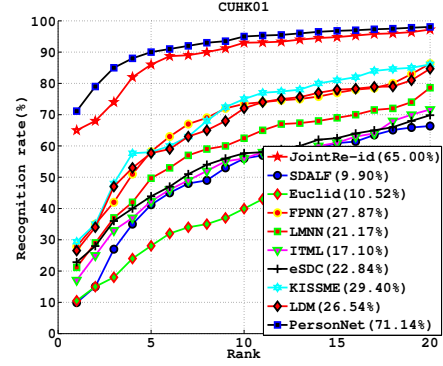
c) *Competitors.*: We compare our model with the following state-of-the-art approaches: SDALF [11], ELF [29], LMNN [17], ITML [34], LDM [35], eSDC [7], SalMatch [10], Generic Metric [41], Mid-Level Filter (MLF) [4], eBiCov [43], PCCA [16], LADF [9], kLFDA [6], rPCCA [6], RDC [13], RankSVM [44], Metric Ensembles (Ensembles) [39], KISSME [15], JointRe-id [1], FPNN [21].

B. Experiments on CUHK03 data set

The CUHK03 dataset includes 13,164 images of 1360 pedestrians. The whole dataset is captured with six surveil-

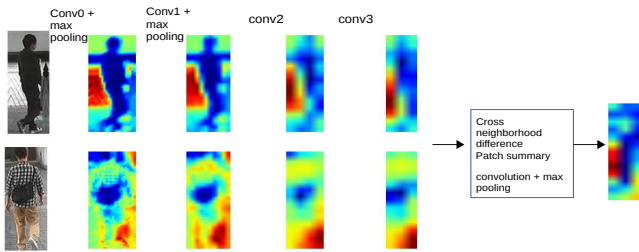


(a) CUHK03 data set

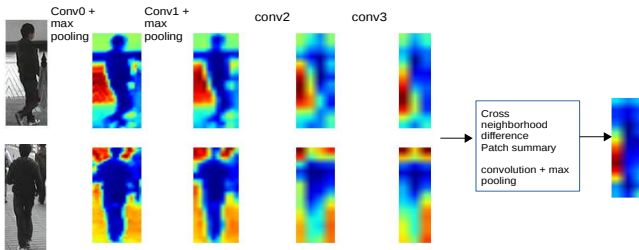


(b) CUHK01 data set

Fig. 4: Performance comparison with state-of-the-art approaches using CMC curves on CHK03 and CUHK01 datasets.



(a) A pair of negative images



(b) A pair of positive images

Fig. 3: Feature responses of each layer learned by our network. See Section III for details.

lance camera. Each identity is observed by two disjoint camera views, yielding an average 4.8 images in each view. This dataset provides both manually labeled pedestrian bounding

TABLE II: Rank-1, Rank-5, and Rank-10 recognition rate of various methods over CUHK03 dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
Ensembles [39]	62.10	87.81	92.30	97.20
JointRe-id [1]	54.74	86.42	91.50	97.31
FPNN [21]	20.65	51.32	68.74	83.06
Euclid	5.64	18.93	28.96	43.17
ITML [34]	5.53	18.89	29.96	44.20
LMNN [17]	7.29	21.00	32.06	48.94
LDM [35]	13.51	40.73	52.13	70.81
SDALF [11]	5.60	23.45	36.09	51.96
eSDC [7]	8.76	24.07	38.28	53.44
KISSME [15]	14.17	48.54	52.57	70.03
PersonNet	64.80	89.40	94.92	98.20

TABLE III: Rank-1, Rank-5, and Rank-10 recognition rate of various methods over CUHK01 dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
JointRe-id [1]	65.00	88.70	93.12	97.20
SDALF [11]	9.90	41.21	56.00	66.37
Euclid	10.52	28.07	39.94	55.07
FPNN [21]	27.87	58.20	73.46	86.31
LMNN [17]	21.17	49.67	62.47	78.62
ITML [34]	17.10	42.31	55.07	71.65
eSDC [7]	22.84	43.89	57.67	69.84
KISSME [15]	29.40	57.67	72.43%	86.07
LDM [35]	26.45	57.69	72.04%	84.69
PersonNet	71.14	90.07	95.00	98.06

boxes and bounding boxes automatically obtained by running a pedestrian detector [45]. In our experiment, we report results on labeled data set. As can be seen from Fig.4 (a) and Table II, our very deep PersonNet outperform the previous methods, which particularly improves from 62.1% (Ensemble [39]) to 64.8%.

C. Experiments on CUHK01 data set

The CUHK01 data set has 971 identities with 2 images per person in each view. We report results on the setting where 100 identities are used for testing, and the remaining 871 identities used for training, in accordance with FPNN [21]. Fig. 4 (b) and Table III compare the performance of our model with previous methods. Our approach is superior to the state of the art by a large margin with a rank-1 recognition rate of 71.14% against 65% by the next best method.

D. Experiments on Market-1501 data set

The Market-1501 dataset contains 32,643 fully annotated boxes of 1501 pedestrians, making it the largest person re-id dataset to date. Each identity is captured by at most six cameras and boxes of person are obtained by running a state-of-the-art detector, the Deformable Part Model (DPM) [46]. As conducted in, the dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively.

We compare our model with state-of-the-art methods in Table IV. The results are reported on single-shot and single-query. We can see that our deep network outperforms these methods notably on Market-1501 dataset.

E. Convergence study

In this section, we study the convergence speed of SGD and RMSProp and report empirical results in Fig. 5. It can

TABLE IV: Rank-1 and mAP of various methods over Market-1501 dataset.

Method	$r = 1$	mAP
SDALF [11]	20.53	8.20
eSDC [7]	33.54	13.54
Zheng <i>et al.</i> [40]	34.40	14.09
PersonNet	37.21	18.57

be seen that RMSProp is more stably and relatively faster to be converged than SGD. This is mainly because SGD as itself is solely depending on the given batch of instances of the present iteration. Therefore, it tends to have unstable update steps per iteration and convergence takes more time or even get stuck into local minima. By contrast, RMSProp keeps running average of its recent gradient magnitudes and divides the next gradient by this average so that loosely gradient values are normalized. Consequently, RMSProp works better on gradient updates in steps of different batches.

VI. CONCLUSION

In this work, we evaluated very deep convolutional networks (up to 10 weight layers) for person re-identification. It was demonstrated that the representation depth is beneficial to the recognition accuracy in person matching, and state-of-the-art performance on person re-id datasets including CUHK03, CUHK01, and Market-1501 datasets can be achieved using an effective matching based architecture [1], [21] with notably increased depth. Our experimental results justify the importance of depth in person identity matching.

REFERENCES

- [1] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015.
- [2] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2006.
- [3] P. M. Roth, M. Hirzer, M. Köstinger, C. Belezai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. Springer, 2014, pp. 247–267.
- [4] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.
- [5] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2011.
- [6] F. Xiong, M. Gou, O. Camps, and M. Sznajder, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comp. Vis.*, 2014.
- [7] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013.
- [8] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013.
- [9] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013.
- [10] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2013.
- [11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2010.

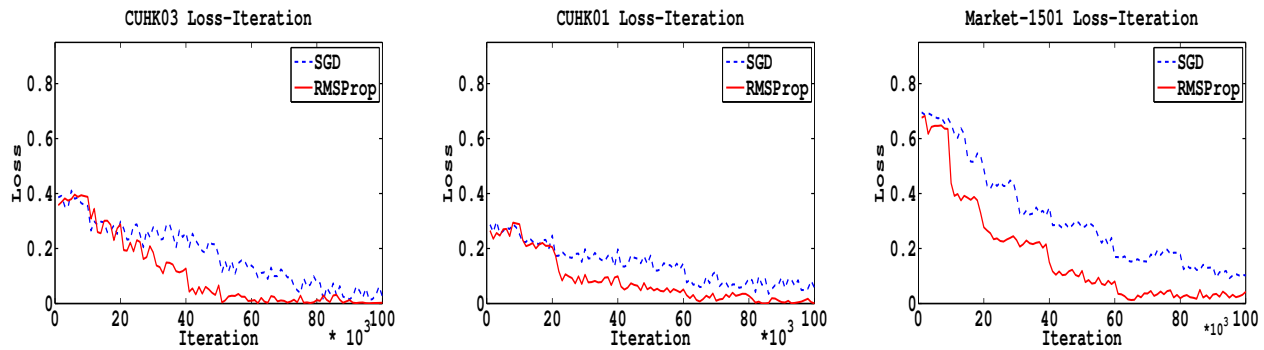


Fig. 5: Study on convergence speed of SGD and RMSProp.

- [12] S. L. Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015.
- [13] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 653-668, p. 3, 2013.
- [14] D. Kedem, S. Tyree, F. Sha, G. R. Lanckriet, and K. Q. Weinberger, "Non-linear metric learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012.
- [15] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012.
- [16] A. Mignon and F. Jurie, "Pcca: a new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 2666-2672.
- [17] M. Hirzer, P. Roth, and H. Bischof, "Person re-identification by efficient imposter-based metric learning," in *Proc. Int'l. Conf. on Adv. Vid. and Sig. Surveillance*, 2012.
- [18] X. Zheng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2013.
- [19] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2013.
- [20] N. W. D. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013.
- [21] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.
- [22] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. Int. Conf. Pattern Recogn.*, 2014.
- [23] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012.
- [25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
- [26] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," in *Technical report: Neural networks of machine learning*, 2012.
- [27] A. Graves, "Generating sequences with recurrent neural networks," in *arXiv:1308.0850*, 2014.
- [28] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Patt. Recogn.*, vol. 33, no. 7, pp. 898-903, 2012.
- [29] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comp. Vis.*, 2008.
- [30] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2007.
- [31] W. Schwartz and L. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. of SIBGRAPI*, 2009.
- [32] G. Lisanti, I. Masi, and A. D. B. Andrew D. Bagdanov, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, p. 1, 2015.
- [33] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006.
- [34] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Mach. Learn.*, 2007.
- [35] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2009.
- [36] Y. Wu, M. Mukunoki, T. Funatomi, M. Minoh, and S. Lao, "Optimizing mean reciprocal rank for person re-identification," in *Proc. Advanced Video and Signal-Based Surveillance*, 2011.
- [37] B. Prosser, W. S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. British Mach. Vis. Conf.*, 2010.
- [38] B. McFee and G. R. G. Lanckriet, "Metric learning to rank," in *Proc. Int. Conf. Mach. Learn.*, 2010.
- [39] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, Boston, USA, 2015.
- [40] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.
- [41] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comp. Vis.*, 2012.
- [42] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a cpu and gpu math expression compiler," in *Proc. Python for Scientific Computing Conference (SciPy)*, 2010.
- [43] B. Ma, Y. Su, and F. Jurie, "Bicov: A novel image representation for person re-identification and face verification," in *Proc. British Mach. Vis. Conf.*, 2012.
- [44] B. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. British Mach. Vis. Conf.*, 2010.
- [45] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627-1645, 2010.
- [46] B. Huang, J. Chen, Y. Wang, C. Liang, Z. Wang, and K. Sun, "Sparsity-based occlusion handling method for person re-identification," in *Multimedia Modeling*, 2015.