PERSON RE-IDENTIFICATION WITH DEEP DENSE FEATURE REPRESENTATION AND JOINT BAYESIAN

Shengke Wang, Lianghua Duan, Na Yang, Junyu Dong

Department of Computer Science and Technology, Ocean University of China, China

ABSTRACT

Person re-identification that aims at matching individuals across multiple camera views has become indispensable in intelligent video surveillance systems. It remains challenging due to the large variations of pose, illumination, occlusion and camera viewpoint. Feature representation and metric learning are the two fundamental components in person re-identification. In this paper, we present a Special Dense Convolutional Neural Network (SD-CNN) to extract the feature and apply Joint Bayesian to measure the similarity of pedestrian image pairs. The SD-CNN can preserve more horizontal information to against viewpoint changes, maximize the feature reuse and ensure feature distributing discriminative. Joint Bayesian models the extracted feature representation as the sum of inter- and intra-personal variations, and the joint probability of two images being a same person can be obtained through log-likelihood ratio. Experiments show that our approach significantly outperforms state-of-the-art methods on several benchmarks of person re-identification.

Index Terms— Person re-identification, Joint Bayesian, deep learning, Convolutional Neural Networks, verification

1. INTRODUCTION

Person re-identification focuses on the verification of two pedestrian images under surveillance cameras, and it still is a challenging problem in computer vision due to the complicated diversification in pose, occlusion, illumination, visual appearance and image resolution across different camera views.

The framework of existing methods usually consists of two critical components: extracting discriminative feature from pedestrian images and computing the similarity of image pairs by feature comparison. There are many works focusing on these two aspects. The traditional methods do a lot of on improving suitable hand-crafted features[1, 2, 3], or good metric for comparison [4, 5, 6]. The first aspect considers to find features that are robust to challenging factors, and the second aspect comes to the metric learning problem which generally minimizes the intra-personal distance while maximizing the inter-personal distance.

With the resurgence of Convolutional Neural Networks (CNNs), several deep learning methods were proposed for person re-id. In DeepReID [7], Li et al. first used CNN with two special layers to address the problems of viewpoint and pose variations. Shi et al. [8] used a Siamese architecture to deal with the variations of different images. Ahmed et al.[9] proposed an improved deep learning structure for person re-identification by using the Cross-Input Neighborhood Difference to learn the relationships between the two views. Yang et al. [10] improved a deep transfer metric learning (DTML) method to transfer cross-domain visual knowledge into target dataset, which increase the robust of result. Wu et al. [11] investigated the combination and complementary of a multi-color space hand-crafted features and deep feature with a feature fusion Network (FFN). Yi et al. [12] applied constrained deep metric learning for re-identification with a full connected layers replacing the process of metric learning, and got the good single-shot result on CUHK01 [13] dataset. Most of the frameworks are designed in a Siamese fashion that need amounts of procedure to train CNN and process data. Compared with the existing methods, our SD-CNN simplifies the process by using a single image as input and also improves the performance.

There is a consensus that the deeper the network, the better the results. However, with the depth of CNNs becoming more deeply, a new problem coming along: vanishing-gradient. The key to solve this problem is shortening the paths from early layers to later layers. Gao et al. [14] proposed a simple connectivity pattern named dense connectivity which can tackle this problem while ensuring maximum information flow between layers in the network. In this pattern, all layers are connected to each other directly, and each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers.

Joint Bayesian has been successfully used to model the joint probability of two faces being the same or different persons [15]. It models two faces jointly with an appropriate prior on the face representation and represent the feature of a face as the sum of inter- and intra-personal variations. Based on the leaned joint model, the distributions of faces can be obtained, and the verification can be efficiently performed by the derived closed-form expression of log likelihood ratio. The problem of person Re-ID is similar to face verification, and

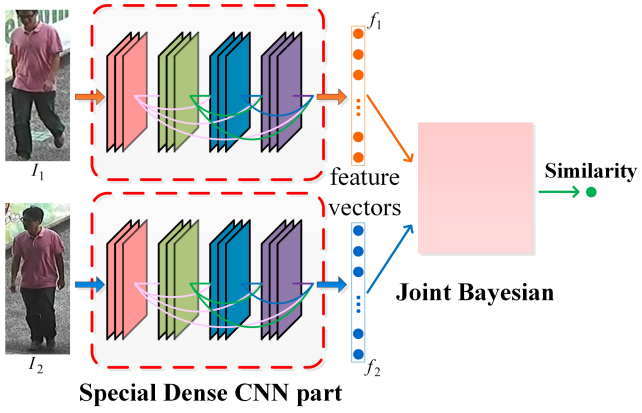


Fig. 1. The overview of our framework.

for the first time we apply Joint Bayesian in person Re-ID in our work.

In this paper, we propose a novel framework (See Fig. 1) to solve person re-identification problem that including a SD-CNN feature extractor and a Joint Bayesian model for distance metric. The major contributions of the proposed work are as follows: First, we propose a Special Dense Convolutional Neural Network (SD-CNN) architecture for feature extraction that can outperform majority of existing deep learning extractor for person re-identification. Second, basing on our efficient SD-CNN feature, we apply Joint Bayesian to person Re-ID problem for the first time and get a 2% performance improvement. Thirty, we conduct experiments on several datasets and evaluate our method by comparing it to the state-of-the-art approaches to prove the effectiveness.

2. METHOD

2.1. SD-CNN structure for feature extraction

There was no doubt that the core of superb method is learning the efficient deeply feature representation. Inspired by [14], we build a Special Dense Convolutional Neural Networks (SD-CNN) with several dense blocks to extract the pedestrian feature. An illustration of the SD-CNN structure is shown in Fig. 2 and the details are listed in Table 1. For convolution layers, Ci_j means the height and width of kernel is i and j, all the stride default to 1, each side of inputs is no-padded except for the C3_3 and C1_3 that padded one pixel to keep the feature-map size fixed respectively. m is the number of peoples in the dataset. Our SD-CNN has the following characteristics: First, due to the pose variations of one pedestrian across different views, the local features appearing in one view may not exactly at the same position in the other view while it is very similar along the same horizontal region. Therefore, we use asymmetric filtering in some critical convolution layers to preserve the horizontal features. Second, amount of 1×1 convolution are employed to reduce the dimension, interact and integrate the information across channels, and increase the nonlinear characteristics. Last but not least, softmax loss function can supervise feature distributing discriminative. For propagating this constraint better we abandon the activation function between the "Feature" layer and last full connected layer.

Table 1. Detail of SD-CNN structure. Note that each N corresponds the sequence BN-Scale-PReLU.

name	Output size	Type of layers
input	3 × 144× 56	data
Conv1	64× 71× 25	C3_7,stride:2
Pool1	64× 36× 13	MaxPooling
DB1	448× 36× 13	(N+C1_1+C3_3+Concat)*12
Trans1	448× 18× 7	N+C1_1*2+AvePooling
DB2	832× 18× 7	(N+C1_1+C3_3+Concat)*12
Trans2	832× 9× 4	N+C1_1+C1_3+AvePooling
DB3	1216× 9× 4	(N+C1_1+C3_3+Concat)*12
Trans3	1216× 5× 2	N+C1_1+C1_3+AvePooling
DB4	1600× 5× 2	(N+C1_1+C3_3+Concat)*12
Pool5	1600× 1× 1	N+Global ave pooling
Feature	64	FC+Dropout
Classify	m	FC+Softmax with loss

We extract a 64-dimensional feature vector at the "Feature" layer and the extraction process is denoted as $f = F(x, \theta_c)$, where $F(\bullet)$ is the propagating forward function, x is the input image of pedestrian, f is the extracted vector and θ_c denotes our SD-CNN parameters to be learned. We use an m-way softmax layer which outputs a probability distribution over m classes. The network is trained to minimize the cross-entropy loss:

$$L(f, y, \theta_{id}) = -\sum_{i=1}^{n} p_i \log q_i \tag{1}$$

where f is the feature vector, y is the target class, and θ_{id} denotes the softmax layer parameters. p_i is the true probability distribution that $p_i = 0$ for all i except $p_y = 1$ for the target class y , q_i is the predicted probability distribution. To classify all the classes simultaneously, the SD-CNNs must form discriminative identity-related features (including intra-class and inter-class information). The cost of Eq. 1 is minimized in the parameter space by using the stochastic gradient descent search, combined with error back-propagation.

2.2. Joint Bayesian for distance metric

We learned the Joint Bayesian model for distance metric based on the extracted SD-CNN feature. According to Joint Bayesian [15], the feature of a pedestrian image can be represented as the sum of inter- and intra-personal variations:

$$x = \mu + \varepsilon \tag{2}$$

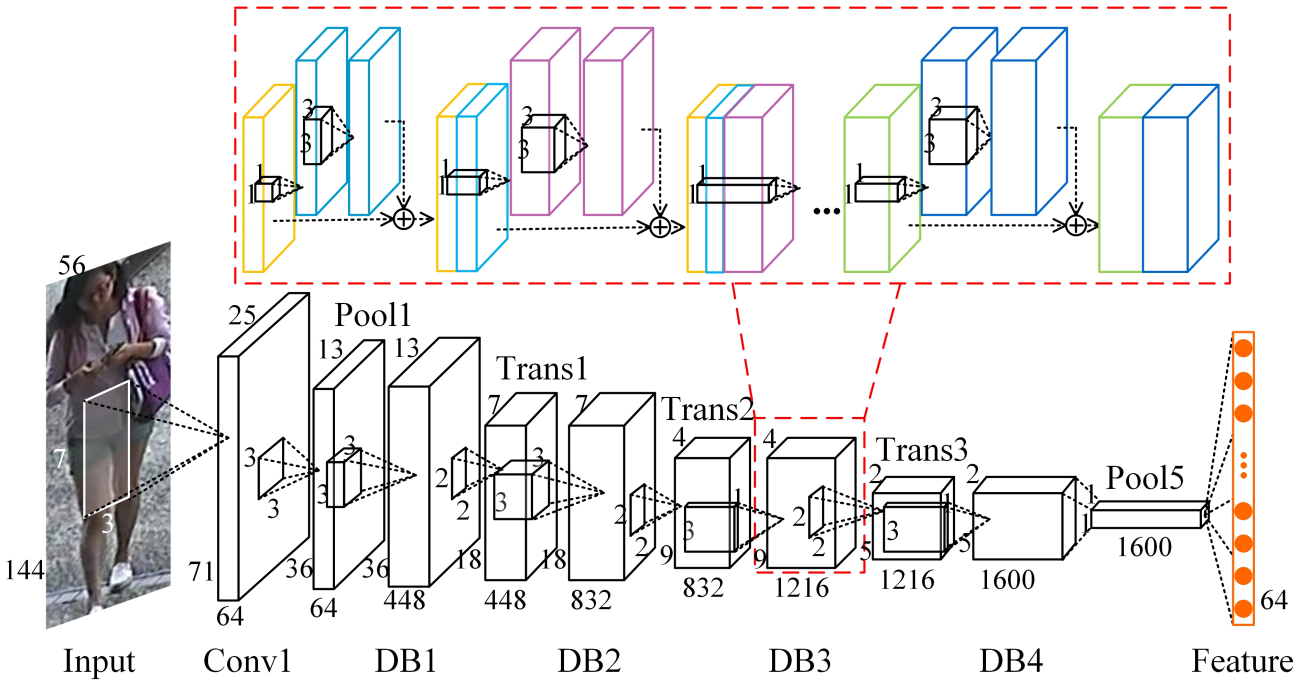


Fig. 2. The SD-CNN structure for feature extraction. Note that the red box is the example of dense block(DB).

where μ and ε follow two Gaussian distributions $N(0, S_\mu)$ and $N(0, S_\varepsilon)$ can be estimated from the training data.

Given two images, the features pair $\{x_1, x_2\}$ can be extracted by SD-CNN. Let H_I represents the intra-personal (same) hypothesis that two images belong to the same person, and H_E is the extra-personal (not same) hypothesis, then the person re-id problem amounts to classifying the difference $\Delta = x_1 - x_2$ as intra-personal variation or extra-personal variation. Based on the MAP(Maximum a Posterior) rule, the distance is made by testing a log-likelihood ratio :

$$d(x_1, x_2) = \log \frac{P(\Delta|H_I)}{P(\Delta|H_E)} = {x_1}^T A x_1 + {x_2}^T A x_2 - 2{x_1}^T G x_2 \tag{3}$$

where A and G can be estimated by the algorithm in Table 2.

3. EXPERIMENTS

We present a comprehensive evaluation of our framework by comparing it against the baseline SD-CNN feature with Euclidean distance as well as other state-of-art methods for person re-identification. All evaluations are based on the Cumulative Matching Characteristics (CMC).

3.1. Datasets

Experiments were conducted on the challenging benchmark datasets for person re-identification, CUHK03 [7], Market-1501 [16] and CUHK01 [13].

Table 2. The Joint Bayesian learning algorithm. Assume there are n identities and each identity has m_i images.

Input: Training data$\{x_i\}$, initialized parameters S_μ, S_ε by positive definite matrix,$t \leftarrow 0$

While not converge **do**

$t \leftarrow t + 1$.

$F = S_\varepsilon^{-1}, G = -(x_i S_\mu + S_\varepsilon)^{-1} S_\mu S_\varepsilon^{-1}$

$\mu_i = \sum_{j=1}^{m_i} S_u(F + m_i G)x_j, \varepsilon_{ij} = x_j + \sum_{j=1}^{m_i} S_\varepsilon G x_j$

Update the parameters S_μ by $S_\mu = \frac{1}{n}\sum_i \mu_i \mu_i^T$

Update the parameters S_ε by $S_\varepsilon = \frac{1}{n}\sum_i \sum_j \varepsilon_{ij}\varepsilon_{ij}^T$

end while

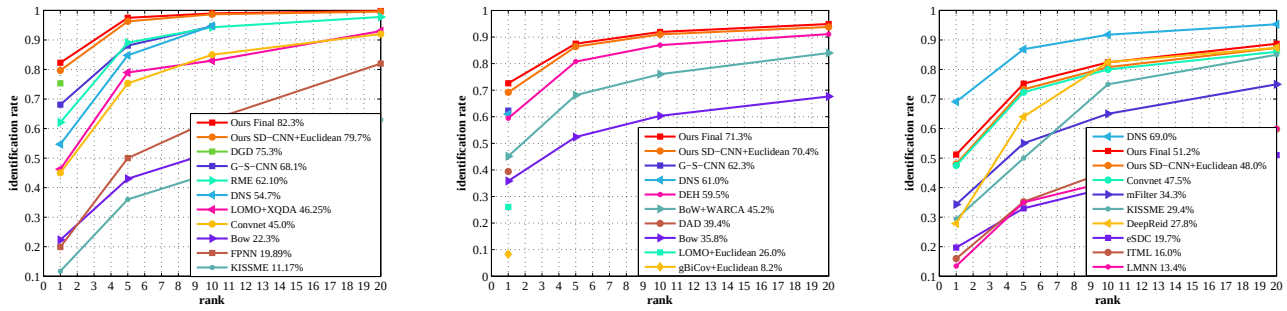
$F = S_\varepsilon^{-1}, G = -(2S_\mu + S_\varepsilon)^{-1} S_\mu S_\varepsilon^{-1}$

$A = (S_\mu + S_\varepsilon)^{-1} - (F + G)$

Output A, G

CUHK03: The CUHK03 is one of the largest published dataset captured by five different pairs of surveillance cameras. We use both manually cropped and automatically detected image, and it has more than 14,000 images of 1467 subjects, each identity has 10 images approximately. Following the splitting provided in [17], evaluation is conducted with 100 test subjects and 1367 identities for training the SD-CNN.

Market-1501: The Market-1501 dataset contains 32,668 annotated bounding boxes of 1,501 identities collected from six cameras. It employs Deformable Part Model (DPM) as



(a) CUHK03 (b) Market-1501 (c) CUHK01

Fig. 3. CMC curves of three datasets.

pedestrian detector that leaving several junk examples which make the dataset very challenging. Following the standard evaluation protocols in [16], the dataset is split into 751 identities for training and 750 identities for testing. We report the single-query (SQ) result for this dataset.

CUHK01: The CUHK01 dataset has 971 subjects, with 4 images per person under 2 camera views. According to the protocol in [17], the dataset is divided into 486 subject for training and 485 for testing.

3.2. Implement details

We implement our Special Dense Convolutional Neural Network (SD-CNN) architecture using the Caffe [18] deep learning framework and employ mini-batch stochastic gradient descent (SGD) for faster back propagation and smoother convergence [19]. In each iteration of training phase, the mini-batch is 50 images, learning rate=0.01 and decreased by every 20000 iteration. We train CUHK03 dataset firstly and then finetune it on other datasets.

3.3. Results

The CMC curves and the rank-1 identification rates for the CUHK03, Market-1501 and CUHK01 datasets are given in Fig. 3 respectively (some results of other methods is no reported). Compared with deep learning methods (DGD [17], G-S-CNN [20], Convernet [9]) and traditional methods (RME [21], DNS [22], LOMO+XQDA [23], Bow [16], FPNN [8], KISSME [24]), our SD-CNN features with simple Euclidean distance easily beats the other methods in CUHK03 dataset, and with the improvement of Joint Bayesian, our framework achieves the state-of-the-art result with rank-1 accuracy up to 82.3%. For Market-1501 dataset, our final framework gains 72.6% rank-1 accuracy, both Euclidean distance and Joint Bayesian with our SD-CNN feature are greatly better than the previous methods results (G-S-CNN [20], DNS [22], LOMO+Euclidean [16], DEH [25], BoW+WARCA [26], DAD [27], Bow [16], gBiCov+Euclidean [7]). However, for small dataset CUHK01, it would be insufficient to learn such a large capacity network from scratch, our framework with SD-CNN and Joint Bayesian fails to achieve the best result (compared with DNS [22]). But compared with others (Convnet [9], mFilter [1], KISSME [24], DeepReid [7], eSDC [28], ITML [29], LMNN [13]) our method get the best result.

Our baseline SD-CNN architecture with simple Euclidean metric outperforms all existing approach for person re-identification for CUHK03 and Market-1501 datasets at Rank 1. We also achieve a comparable result to the best even on small dataset (CUHK01). Our final framework with Joint Bayesian improve all the baseline about 2% at Rank 1, outperforms a large margin with the previous method and achieve the state-of-art result for CUHK03 and Market-1051dateset.

4. CONCLUSION

This paper proposes a novel framework for person Re-ID which consists of a convolutional neural network extractor named SD-CNN and a metric measure named Joint Bayesian. Experiments show that our approach significantly outperforms state-of-the-art methods on several benchmarks of person re-identification.

5. REFERENCES

[1] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[2] Ziming Zhang, Yuting Chen, and Venkatesh Saligrama, *A Novel Visual Word Co-occurrence Model for Person Re-identification*, Springer International Publishing, 2014.

[3] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li, *Salient Color Names for Person Re-identification*, 2014.

[4] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith, "Learning locally-adaptive decision functions for person verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[5] Martin K?stinger, Martin Hirzer, Paul Wohlhart, and Peter M. Roth, "Large scale metric learning from equivalence constraints," pp. 2288–2295, 2012.

[6] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision*, 2013, pp. 2528–2535.

[7] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[8] Hailin Shi, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Yang Yang, and Stan Z. Li, "Constrained deep metric learning for person re-identification," *Computer Science*, 2015.

[9] Ejaz Ahmed, Michael Jones, and Tim K. Marks, "An improved deep learning architecture for person re-identification," in *Computer Vision and Pattern Recognition*, 2015.

[10] Yang Yang, Shengcai Liao, Zhen Lei, and Stan Z. Li, "Large scale similarity learning using similar pairs for person verification," in *AAAI*, 2016.

[11] Shangxuan Wu, YingCong Chen, Xiang Li, AnCong Wu, JinJie You, and WeiShi Zheng, "An enhanced deep feature representation for person re-identification," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–8.

[12] Dong Yi, Zhen Lei, and Stan Z. Li, "Deep metric learning for practical person re-identification," *Computer Science*, pp. 34–39, 2014.

[13] Junlin Hu, Jiwen Lu, and Yap Peng Tan, "Deep transfer metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 325–333.

[14] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger, "Densely connected convolutional networks," 2016.

[15] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun, "Bayesian face revisited: a joint formulation," in *European Conference on Computer Vision*, 2012, pp. 566–579.

[16] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[17] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang, "Learning deep feature representations with domain guided dropout for person re-identification," 2016.

[18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *Eprint Arxiv*, pp. 675–678, 2014.

[19] Alexis Mignon and Frdric Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672.

[20] Rahul Rama Varior, Mrinal Haloi, and Gang Wang, "Gated siamese convolutional neural network architecture for human re-identification," 2016.

[21] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," *Computer Science*, pp. 1846–1855, 2015.

[22] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification," pp. 1239–1248, 2016.

[23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[24] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu, "Semi-supervised coupled dictionary learning for person re-identification," pp. 3550–3557, 2014.

[25] Evgeniya Ustinova and Victor Lempitsky, "Learning deep embeddings with histogram loss," 2016.

[26] Cijo Jose and Fran?ois Fleuret, "Scalable metric learning via weighted approximate rank component analysis," 2016.

[27] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian, "Deep attributes driven multi-camera person re-identification," 2016.

[28] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," vol. 9, no. 4, pp. 3586–3593, 2013.

[29] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon, "Information-theoretic metric learning," in *Machine Learning, Proceedings of the Twenty-Fourth International Conference*, 2007, pp. 209–216.