PERSON RE-IDENTIFICATION WITH COARSE-TO-FINE VISUAL ATTENTION

Zijie Zhuang[1], Haizhou Ai[1], Chong Shang[1], Lihu Xiao[2]

[1]Tsinghua National Lab for Info. Sci. & Tech., Depart. of Computer Sci. & Tech.,
Tsinghua University, Beijing, China. [2]Huawei Technologies, Beijing, China
zhuangzj15@mails.tsinghua.edu.cn, ahz@mail.tsinghua.edu.cn, shang-c13@mails.tsinghua.edu.cn
xiaolihu@huawei.com

ABSTRACT

The basic goal of person re-identification (re-id) tasks is to identify a person across disjoint camera views. It has been deeply explored in video surveillance but still remains a very challenging problem. In this paper, we introduce a novel model for re-id tasks with two components: an expressive feature fusion strategy that consists of high-level convolution features and the low-level optical information, and an improved recurrent attention model that performs a coarse-to-fine feature selection. To the best of our knowledge, experiments show that our model achieves the best performance on several benchmark datasets compared with all the other state-of-the-art approaches.

***Index Terms*—** person re-identification, coarse-to-fine feature selection, color histogram, attention model

1. INTRODUCTION

Identifying a person across disjoint camera views, also known as the person re-identification tasks (re-id) which has received increasing attention recently, is an essential component for video surveillance system. It is a very challenging task due to the visual ambiguity and spatiotemporal uncertainty in the appearance of a person across different views.

Deep neural network is a highly expressive model which can learn extremely complicated relationship between its input and output, which makes it applicable to handle complicated variations in re-id tasks. Although most of these complex relationships extracted from training data are well suited for re-id tasks, some of which could be regarded as the noise of sampling. These relationships only exist in the training set, negatively impacting the generalization ability of the model.As shown in Fig. 1, we list 2 groups of errors that a typical model [1] made. Images in the first column and the last column are from the same identity. Each image in the middle is a wrong prediction due to the distance between that image and the anchor image being less than the distance between the anchor and the positive image, although they can be easily distinguished by their appearances with human visual perception.

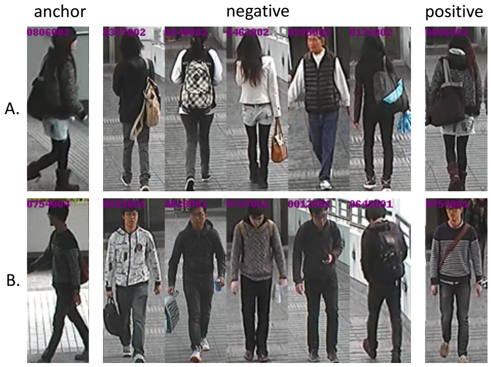

Fig. 1. Examples of some failure cases that a trained CNN model makes. Best viewed in color.

These failure cases motivate us to consider a discriminative feature fusion strategy and an improved recurrent attention model jointly for better performance. To extract better features, we propose a novel feature fusion strategy which combines high-level convolution feature representations and the low-level optical information. The basic idea of this fusion strategy is to preserve important intrinsic statistics of original images after intricate feature transformations.Furthermore, directly compressing entire images into static representations even could deteriorate the sampling noise problem. To allow salient features to come to the forefront dynamically, we improve the recurrent attention model (RAM) [1–3] by introducing a coarse-to-fine feature selection strategy to combine features from different layers. To train our model efficiently, we adopt the triplet loss and moderate data mining from the FaceNet [4] for superior generalization ability.

The main contributions of this paper include: 1) A novel feature fusion strategy which combines high-level features and the low-level optical information; 2) An improved recurrent attention model that generates robust and informative representations by coarse-to-fine feature selection; 3) A practical data mining strategy with triplet loss to streamline the training procedure.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

2. RELATED WORK

A typical person re-id system can be divided into two basic components: a discriminative feature to represent images and a robust distance measure to compare those features across images. There have been plenty of studies for obtaining efficient and discriminative features to handle large variations of the viewpoint, pose and illumination condition [5–9]. Many of them develop expressive features based on color information [5–7,9–14]. Meanwhile, a bunch of studies also focus on finding better metric learning and ranking algorithms, including Large Margin Nearest Neighbor (LMNN) [15, 16], Information Theoretic Metric Learning (ITML) [17], Logistic Discriminant Metric Learning (LDML) [18] and Mahalanobis metric learning(KISSME) [19]. Instead of hand-crafted feature representations and distance metric learning algorithms, the convolution neural network further boosts person re-id tasks [20,21] greatly by the end-to-end learning paradigm. Li et al. [20] use a set of filter pairs without parameter-sharing to generate more expressive features and a patch-matching layer to handle pose variations between two input images. Apart from single path convolutional neural network models, Liu et al. [1] adopt LSTM based recurrent attention models for re-id tasks and make it possible to seek discriminative parts of a given image.

3. OUR APPROACH

The overall network architecture is presented in Fig. 2. Given a person image, we extract both high-level coarse-to-fine convolution features from multiple layers and low-level color histogram features from raw images. These features are then subsequently passed through a recurrent attention model (RAM) [2]. We further develop this model to perform a coarse-to-fine feature selection, aiming to find the most discriminative information and eliminate noises.

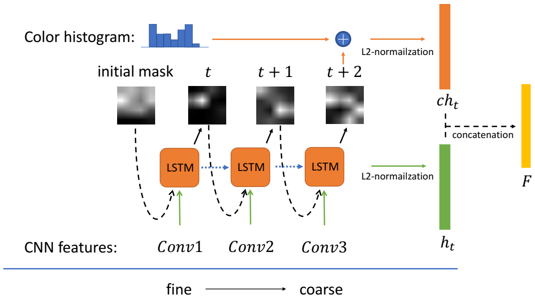

Fig. 2. The overall architecture of our model. Both CNN features from multiple layers and color histograms from raw images are extracted. Subsequently, these feature vectors are fed into our improved RAM to generate final representations.

3.1. Feature representation

3.1.1. Coarse-to-fine convolution feature descriptor

The main purpose of feature extraction is to find the most discriminative information in images. Since convolution neural networks encode each image into a more aggregated feature layer by layer, the detail information is lost during this process. To collect features from coarse to fine, we extract features from three different layers using truncated AlexNet [22]: *Max1*, *Max2*, *Max5*. The size of these three feature cubes are respectively $27 \times 27 \times 96$, $13 \times 13 \times 256$ and $6 \times 6 \times 256$. Then one extra procedure is executed to transform these three feature cubes to the same shape: a convolution layer with kernel 5×5 for *Max1* and another with kernel 2×2 for *Max2*. The output of these two layers and the $Max5$ layer are denoted as $Conv_1$, $Conv_2$ and $Conv_3$ respectively in the following sections.

3.1.2. Color histogram descriptor (HS)

We notice that without proper parameter initialization and carefully organizing training data, the CNN part learns from random fluctuations caused by sampling noise, leading to the overfitting problem. Fig. 1 shows that the CNN learns some specific transformations from the training set and makes wrong predictions during the testing phase. After thorough examinations on these errors, we discover a large difference between the appearances of these wrong predictions and corresponding anchor images. We argue that a descriptor for this difference is beneficial for boosting our model. Since the color histogram is robust to geometric changes and invariant to translation and rotation, we assert that the color histogram can be an important supplement to a pure CNN model.

For each person image, we use the same receptive filed as the CNN layers to generate color histograms corresponding to convolution features. The final color histogram descriptor HS are extracted from RGB color space and each channel of RGB color space is divided into 4 bins equally:

$$\mathbf{HS} = \begin{pmatrix} hs_{11} & \dots & hs_{16} \\ \vdots & \ddots & \vdots \\ hs_{61} & \dots & hs_{66} \end{pmatrix}, \tag{1}$$

where each $hs_{i,j}$ is a 64-dimensional vector and the size of HS is $6 \times 6 \times 64$.

3.2. Recurrent Attention Mechanism

After extracting features from raw images, we adopt the long short-term memory (LSTM) network as an attentional mechanism for generating final representations. In each time step, the LSTM can "look at" every location in the feature map by generating an attention mask. Parameters of LSTM cell are updated as follows:

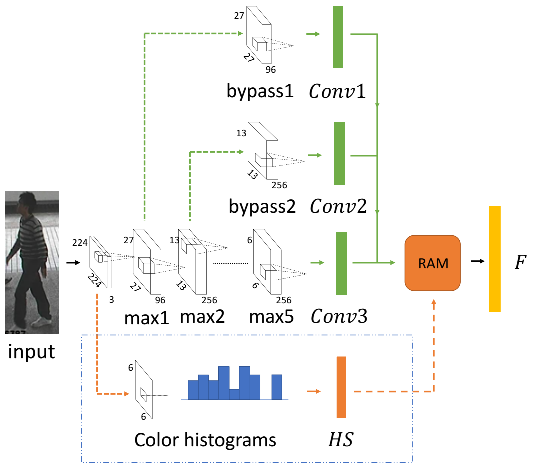

Fig. 3. Our improved recurrent attention model receives features from different layers and generates attention masks for each of them to perform a coarse-to-fine feature selection.

$$h_t, c_t = LSTM(x_t, h_{t-1}, c_{t-1}), \tag{2}$$

where h_t and c_t are the hidden state and memory in the time step t. And input features of LSTM in the time step t can be calculate as:

$$\begin{aligned} \mu_t &= softmax(W_t h_{t-1}), \\ x_t &= \mu_t Conv_t, \\ ch_t &= \mu_t HS_t, \end{aligned} \tag{3}$$

where W_i denotes the learned weight matrix for generating attention masks, μ_t is the "soft" attention mask and ch_t denotes the salient color histogram at the time t.

We implement CAN [1] as our baseline. In this experiment, we observe that when the same feature cube is fed into LSTM in every step, the LSTM always tends to "look at" the same region. Furthermore, each channel of input feature cubes indicates a unique kernel sliding all over the previous feature cubes.Applying the exact same mask to every channel may cause massive information loss.

These observations inspire us to further improve this attention mechanism by introducing a coarse-to-fine feature selection to combine salient information from different layers. This improved RAM architecture is shown in Fig. 3. Furthermore, to make full use of our attention masks, we split each feature cube of $Conv_1$, $Conv_2$, $Conv_3$ into n smaller one along the channel axis with the same size (where $n = 4$) and feed them into RAM sequentially.In this manner, we can extract salient information by using 4 different masks for each feature cube.

We initialize our improved RAM using the same method in [1] and extract the hidden state h_t and the salient color histogram ch_t at $t = 3, 6, 9, 12$. $l2$-normalization is applied to these two sequences separately before concatenating them together:

$$f = [h'_3, h'_6, h'_9, h'_{12}, \alpha ch'_3, \alpha ch'_6, \alpha ch'_9, \alpha ch'_{12}], \tag{4}$$

where h'_t and ch'_t denote the normalized hidden state and the normalized color histogram respectively. α is the coefficient to control the weight of salient color histograms.

3.3. Triplet Selection

After the feature extraction, we simply calculate the Euclidean distance between features to determine the similarity of them.

During the training phase, we follow the instruction introduced in [4]. For every identity I, we select several images and extract features from them: feature F^{anc} from the anchor image, F^{pos} from a positive image of the same identity and F^{neg} from a negative sample of a different identity.

The goal of the training procedure is to minimize the loss:

$$loss = \frac{1}{N}\sum_{i=1}^{N}[\|F_i^{anc} - F_i^{pos}\|_2^2 - \|F_i^{anc} - F_i^{neg}\|_2^2 + \beta]_+, \tag{5}$$

where the margin between the positive distance and the negative distance is denoted as β and $[\cdot]_+$ indicates that we truncate values at zero.

In practice, always selecting the most difficult samples can lead to bad local minima early on in the training phase. Shi et al. [23] propose a moderate positive mining method, which tries to preserve the intrinsic property of pedestrian data while minimizing the intra-class variance.We argue that for some relatively small datasets, this moderate positive mining method may not be able to work effectively since the positive samples are scarce. As a result, we use all anchor-positive pairs and propose a moderate negative mining method for generating negative samples from candidate images.

In each mini-batch with $size = 2 \times n$, we select negative samples which satisfy the following condition:

$$\|F_i^{anc} - F_i^{neg}\|_2^2 - \|F_i^{anc} - F_i^{pos}\|_2^2 > 0, \tag{6}$$

where i is the identifier of a certain pedestrian.

4. EXPERIMENTS

We implement our network using TensorFlow [24]. Our network architecture is evaluated on two challenging benchmark datasets: CUHK01 and CUHK03. Each experiment is repeated 10 times and results are evaluated in the form of Cumulated Matching Characteristic (CMC) curve.

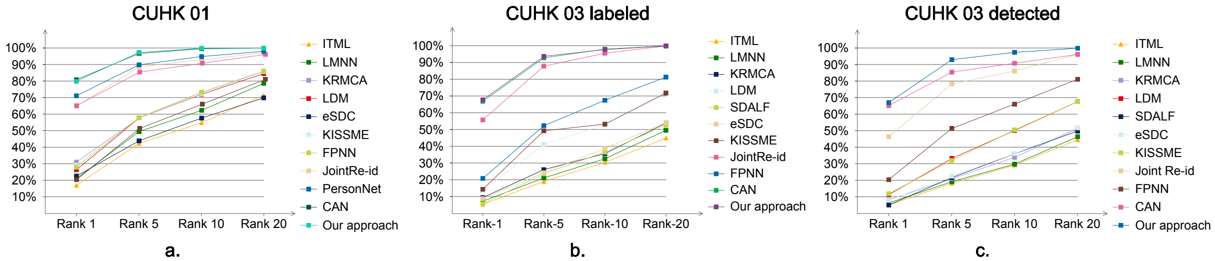

Fig. 4. We compare our model with other methods using CMC curves on CUHK01 and CUHK03 datasets. Our model outperforms previous state-of-the-art methods and achieves the 100% accuracy at Rank-10 on the CUHK01 dataset.

4.1. Experiments on CUHK01

The CUHK01 [25,26] dataset contains 971 identities with snapshots from two different cameras. We conduct experiments for 10 rounds. In each round, we randomly split the dataset into two subsets: one with 871 identities for training and another one with 100 identities for testing. We set the margin $\alpha = 0.3$ and $\beta = 0.1$.

We train our network using AdaDelta optimizer with base learning rate $\eta = 0.001$. The results are shown in Table 1 and Fig. 4. Our model outperforms all the other methods as far as we know and achieves the 100% accuracy at rank-10.

Table 1. Results on CUHK01

Model	Rank-1	Rank-5	Rank-10	Rank-20
ITML	17.10	42.31	55.07	71.65
LMNN	21.17	49.67	62.47	78.62
KRMCA	31.22	57.68	73.55	86.07
LDM	26.45	57.69	72.04	84.69
eSDC	22.84	43.89	57.67	69.84
KISSME	29.40	57.67	72.43	86.07
FPNN	27.87	58.20	73.46	86.31
JointRe-id	65.00	88.70	93.12	97.20
PersonNet	71.14	90.07	95.00	98.06
CAN	**81.04**	96.89	99.67	100.00
Our approach	80.13	**97.63**	**100.00**	**100.00**

4.2. Experiments on CUHK03

The CUHK03 [20] dataset is a relatively larger dataset for person re-id tasks. It contains 13164 images of 1360 identities. Each identity has at most ten images from two camera views. During experiments, the dataset is divided into a training set of 1260 identities and a testing set of 100 identities according to the test sequences provided by the dataset.

We set $\alpha = 1.0$ and $\beta = 0.1$ and start with $\eta = 0.01$ to make our model converge fast in the beginning. Then we change η to 0.001 until the end of training procedure.

Table 2 and Table 3 demonstrate that our model beats all the other methods as far as we know.

Table 2. Results on CUHK03 detected

Model	Rank-1	Rank-5	Rank-10	Rank-20
ITML	5.14	17.87	28.24	43.12
LMNN	6.25	18.68	29.07	45.03
KRMCA	8.14	20.31	32.96	49.96
LDM	10.92	32.25	48.78	65.63
SDALF	4.87	21.17	35.06	48.44
eSDC	7.68	21.86	34.96	50.03
KISSME	11.7	31.16	48.98	65.63
Joint Re-id	44.96	76.01	83.47	93.15
FPNN	19.89	50	64	78.5
CAN	63.05	82.94	88.17	93.29
Our approach	**65.07**	**90.18**	**94.28**	**96.83**

Table 3. Results on CUHK03 labeled

Model	Rank-1	Rank-5	Rank-10	Rank-20
ITML	5.53	18.89	29.96	44.20
LMNN	7.29	21.00	32.06	48.94
KRMCA	9.23	25.73	35.09	52.96
LDM	13.51	40.73	52.13	70.81
SDALF	5.60	23.45	36.09	51.96
eSDC	8.76	24.07	38.28	53.44
KISSME	14.17	48.54	52.57	70.53
JointRe-id	54.74	86.50	93.88	98.10
FPNN	20.65	51.50	66.50	80.00
CAN	65.65	91.28	**96.29**	98.17
Our approach	**66.58**	**92.30**	96.06	**98.24**

5. CONCLUSION

In this paper, we propose a novel model for person re-identification with three components: an efficient feature fusion strategy incorporating multiple CNN features and color histograms, an improved recurrent attention model to perform a coarse-to-fine feature selection and a practical triplet selection strategy. Experiments demonstrate that our model achieve state-of-the-art performance on several benchmarks.

6. ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China (Project Number 61521002).

7. REFERENCES

[1] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan, "End-to-end comparative attention networks for person re-identification," *arXiv preprint arXiv:1606.04404*, 2016.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, vol. 2, no. 3, pp. 5, 2015.

[3] Jason Kuen, Zhenhua Wang, and Gang Wang, "Recurrent attentional networks for saliency detection," *arXiv preprint arXiv:1604.03227*, 2016.

[4] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[5] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin, "Color invariants for person reidentification," *PAMI*, vol. 35, no. 7, pp. 1622–1634, 2013.

[6] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus, "Learning color names for real-world applications," *TIP*, vol. 18, no. 7, pp. 1512–1523, 2009.

[7] Raphael Felipe de Carvalho Prates and William Robson Schwartz, "Cbra: Color-based ranking aggregation for person re-identification," in *ICIP*, 2015.

[8] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.

[9] Mu Gao, Haizhou Ai, and Bo Bai, "A feature fusion strategy for person re-identification," in *ICIP*, 2016.

[10] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *ICCV*, 2013.

[11] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.

[12] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[13] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li, "Salient color names for person re-identification," in *ECCV*, 2014.

[14] Joost Van de Weijer and Cordelia Schmid, "Applying color names to image description," in *ICIP*, 2007.

[15] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, no. Feb, pp. 207–244, 2009.

[16] Kilian Q Weinberger and Lawrence K Saul, "Fast solvers and efficient implementations for distance metric learning," in *ICML*, 2008.

[17] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon, "Information-theoretic metric learning," in *ICML*, 2007.

[18] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009.

[19] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[21] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[23] Dong Yi, Zhen Lei, and Stan Z Li, "Deep metric learning for practical person re-identification," *arXiv preprint arXiv:1407.4979*, 2014.

[24] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[25] Wei Li, Rui Zhao, and Xiaogang Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.

[26] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.