DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification

Wei Li Rui Zhao Tong Xiao Xiaogang Wang*
The Chinese University of Hong Kong, Hong Kong
lwthucs@gmail.com, {rzhao, xiaotong, xgwang}@ee.cuhk.edu.hk

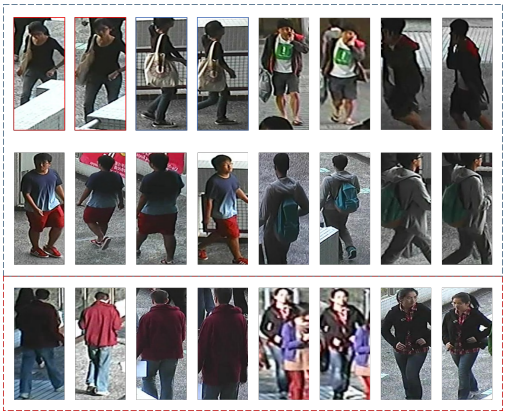
Abstract

Person re-identification is to match pedestrian images from disjoint camera views detected by pedestrian detectors. Challenges are presented in the form of complex variations of lightings, poses, viewpoints, blurring effects, image resolutions, camera settings, occlusions and background clutter across camera views. In addition, misalignment introduced by the pedestrian detector will affect most existing person re-identification methods that use manually cropped pedestrian images and assume perfect detection.

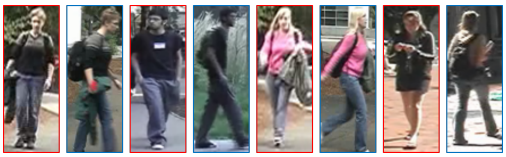
In this paper, we propose a novel filter pairing neural network (FPNN) to jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter. All the key components are jointly optimized to maximize the strength of each component when cooperating with others. In contrast to existing works that use handcrafted features, our method automatically learns features optimal for the re-identification task from data. The learned filter pairs encode photometric transforms. Its deep architecture makes it possible to model a mixture of complex photometric and geometric transforms. We build the largest benchmark re-id dataset with 13,164 images of 1,360 pedestrians. Unlike existing datasets, which only provide manually cropped pedestrian images, our dataset provides automatically detected bounding boxes for evaluation close to practical applications. Our neural network significantly outperforms state-of-the-art methods on this dataset.

1. Introduction

The purpose of person re-identification is to match pedestrians observed in non-overlapping camera views with visual features [13, 9, 35, 1, 7, 3, 51, 20, 19, 16, 29, 26, 24, 48, 2, 41]. It has important applications in video surveillance, such as cross-camera tracking [42], multi-camera event detection [27], and pedestrian retrieval [27]. This problem is extremely challenging because it is difficult to match the visual features of pedestrians captured in different camera views due to the large variations of lightings, poses, viewpoints, image resolutions, photometric settings of cameras, and cluttered backgrounds. Some examples are shown in Figure 1.



(a) Samples from our new dataset, CUHK03



(b) Samples from the VIPeR dataset [12]

Figure 1. Samples of pedestrian images observed in different camera views in person re-identification. The two adjacent images have the same identity.

The typical pipeline of a person re-identification system is shown in Figure 2. In practice, it should start with automatic pedestrian detection, which is an essential step for extracting pedestrians from long-hour recorded videos. Given a pedestrian detection bounding box, manually designed features are used to characterize the image region in all the existing works, although they may be suboptimal for the task of person re-identification. Image regions of the same person undergo photometric transforms due to the change of lighting conditions and camera settings. Their geometric transforms are caused by misalignment and the

*This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110, CUHK 417011, CUHK 429412)

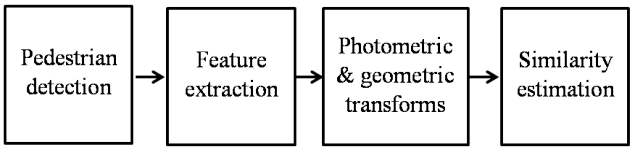


Figure 2. Pipeline of person re-identification.

change of viewpoints and poses. Such transforms could be normalized by learning mapping functions [33, 34] or similarity metrics [16, 51]. It is also supposed to be robust to occlusions and background clutter. All the existing works optimize each module in the pipeline either separately or sequentially. If useful information is lost in previous steps, it cannot be recovered later. Establishing automatic interaction among these components in the training process is crucial for the overall system performance.

The contribution of this paper is three-fold. Firstly, we propose a filter pairing neural network (FPNN) for person re-identification. This deep learning approach has several important strengths and novelties compared with existing works. (1) It jointly handles misalignment, photometric and geometric transforms, occlusions and background clutter under a unified deep neural network. During training, all the key components in Figure 2 are jointly optimized. Each component maximizes its strength when cooperating with others. (2) Instead of using handcrafted features, it automatically learns optimal features for the task of person re-identification from data, together with the learning of photometric and geometric transforms. Two paired filters are applied to different camera views for feature extraction. The filter pairs encode photometric transforms. (3) While existing works assume cross-view transforms to be unimodal, the deep architecture and its maxout grouping layer allow to model a mixture of complex transforms.

Secondly, we train the proposed neural network with carefully designed training strategies including dropout, data augmentation, data balancing, and bootstrapping. These strategies address the problems of misdetection of patch correspondence, overfitting, and extreme unbalance of positive and negative training samples in this task.

Thirdly, we re-examine the person re-identification problem and build a large scale dataset that can evaluate the effect introduced by automatic pedestrian detection. All the existing datasets [12, 37, 50, 27, 3, 16] are small in size, which makes it difficult for them to train a deep neural network. Our dataset has 13,164 images of 1,360 pedestrians; see a comparison in Table 1. Existing datasets only provide manually cropped pedestrian images and assume perfect detection in evaluation protocols. As shown in Figure 1, automatic detection in practice introduces large misalignment and may seriously affect the performance of existing methods. Our dataset provides both manually cropped images and automatically detected bounding boxes with a state-of-the-art detector [10] for comprehensive evaluation.

2. Related Work

A lot of studies aimed to improve individual components of the pipeline in Figure 2 [44]. The visual features used in the existing person re-identification systems are manually designed. Global features characterize the distributions of color and texture with the histograms of visual words [4, 43]. They have some invariance to misalignment, pose variation, and the change of viewpoints. However, their discriminative power is low because of losing spatial information. In order to increase the discriminative power, patch-based local features have been used [13, 9, 1, 25, 26, 48, 47, 49, 24]. When computing the similarity between two images, visual features of two corresponding patches are compared. The challenge is to match patches in two camera views when tackling the misalignment problem. Handcrafted features are difficult to achieve the balance between discriminative power and robustness. The optimal feature design depends on photometric and geometric transforms across camera views. For example, if the illumination variation is larger, the color space should be quantized at a coarser scale. It is hard to achieve such optimization if feature design is independent of other components in Figure 2. Although the features can be selected and weighted in later steps, the performance will decline if the feature pool is not optimally designed. *The right way is to automatically learn features from data together with other components. This is hard to achieve without deep learning.*

One could assume the photometric or geometric transform models and learn the model parameters from training samples [18, 33, 34]. For example, Prosser *et al.* [34] assumed the photometric transform to be bi-directional Cumulative Brightness Transfer Functions, which map color observed in one camera view to another. Porikli [33] learned the color distortion function between camera views with correlation matrix analysis. They assume transforms to be unimodal. *In our proposed filter pairing neural network, photometric transforms are learned with filter pairs and a maxout grouping layer. On the other hand, geometric transforms are learned with a patch matching layer, a convolutional-maxpooling layer and a fully connected layer. The proposed neural network can model a mixture of complex transforms.*

The effect of cross-camera transforms, occlusions and background clutter can be further depressed by learning a proper distance/similarity metric. Gray *et al.* [13] and Prosser *et al.* [35] use boosting and RankSVM, respectively, to select features and compute the distance between images. There are also many metric learning algorithms [51, 41, 20, 19, 16, 6, 14, 45, 30] designed for person re-identification. All the components in Figure 2 are optimized either separately or sequentially in the existing works.

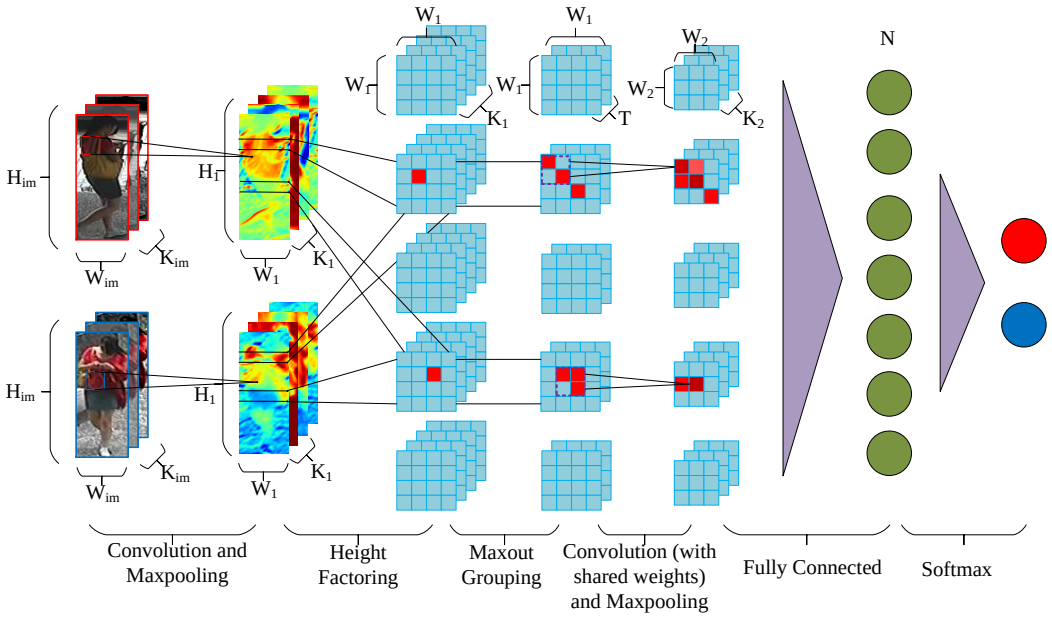


Figure 3. Filter pairing neural network.

Deep learning has achieved great success in solving many computer vision problems, including hand-written digit recognition [22], object recognition [17, 36], object detection [38, 31, 46, 28], image classification [23, 21], scene understanding [8], and face recognition [5, 39, 52, 40]. Although some deep learning works [32, 39] share the spirit of jointly optimizing components of vision systems, their problems, challenges, models and training strategies are completely different from ours. They did not design special layers to explicitly handle cross-view photometric and geometric transforms, misdetection of patch matching and background clutter. To our knowledge, this paper is the first work to use deep learning for person re-identification.

3. Model

The architecture of the proposed FPNN is shown in Figure 3. It is composed of six layers to handle misalignment, cross-view photometric and geometric transforms, occlusions and background clutter in person re-identification. The design of each layer is described below.

3.1. Feature extraction

The first layer is a ***convolutional and max-pooling layer***. It takes two pedestrian images $\mathbf{I}$ and $\mathbf{J}$ observed in different camera views as input. They have three color channels (RGB or LAB) and have the size of $H_{\text{im}} \times W_{\text{im}}$. *The photometric transforms are modeled with a convolutional layer that outputs local features extracted by filter pairs.* By convoluting a filter with the entire image, the responses at all the local patches are extracted as local features. The filters $(\mathbf{W}_k, \mathbf{V}_k)$ applied to different camera views are paired. If K_1 filter pairs are used and each filter is in size of $m_1 \times m_1 \times 3$, the output map for each image has K_1 channels and is in size of $H_0 \times W_0 \times K_1$, where $H_0 = H_{im} - m_1 + 1$ and $W_0 = W_{im} - m_1 + 1$. We define the filtering functions $f, g : \mathbb{R}^{H_{\text{im}} \times W_{\text{im}} \times 3} \rightarrow \mathbb{R}^{H_0 \times W_0 \times K_1}$

$$f_{ij}^k = \sigma((\mathbf{W}_k * \mathbf{I})_{ij} + b_k^I) \quad (1)$$

$$g_{ij}^k = \sigma((\mathbf{V}_k * \mathbf{J})_{ij} + b_k^J). \quad (2)$$

The convolution operation is denoted as $*$. A nonlinear activation function $\sigma(\cdot)$ is used to re-scale the linear output and chosen as $\sigma(x) = \max(x, 0)$. After filtering, each patch is represented by a K_1-channel feature vector. The activation function normalizes and balances different feature channels. *The parameters $\{(\mathbf{W}_k, \mathbf{V}_k, b_k^I, b_k^J)\}$ of the filter pairs are automatically learned from data. Two paired filters represent the same feature most discriminative for person re-identification. They are applied to different camera views and their difference reflects the photometric transforms.* The convolutional layer is followed by max-pooling, which makes the features robust to local misalignment. Each feature map is partitioned into $H_1 \times W_1$ subregions and the maximum response in each subregion is taken as the output. The output of the max-pooling layer is a $H_1 \times W_1 \times K_1$ feature map.

3.2. Patch matching

The second ***patch matching layer*** is to match the filter responses of local patches across views. Considering the geometric constraint, a pedestrian image is divided into M

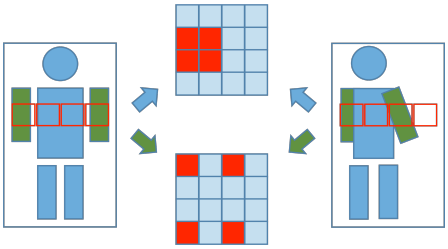


Figure 4. Illustration of patch matching in FPNN. One stripe generates two patch displacement matrices because there are two filter pairs. One detects blue color and the other detects green.

horizontal stripes(***height factoring*** in Figure 3), and each stripe has W_1 patches. Image patches are matched only within the same stripe. Since there are K_1 filter pairs representing different features, the outputs of the patch matching layer are $K_1 M$ $W_1 \times W_1$ patch displacement matrices. The output of the patch matching layer is

$$S^k_{(i,j)(i',j')} = f^k_{ij} g^k_{i'j'}, \tag{3}$$

These displacement matrices encode the spatial patterns of patch matching under the different features. An illustration is shown in Figure 4. If a matrix element $S^k_{(i,j)(i',j')}$ has a high value, it indicates that patches (i,j) and (i',j') both have high responses on a specific feature encoded by the filter pair $(\mathbf{W}_k, \mathbf{V}_k)$.

3.3. Modeling mixture of photometric transforms

Due to various intra- and inter-view variations, one visual feature (such as red clothes) may undergo multiple photometric transforms. In order to improve the robustness on patch matching, a ***maxout-grouping layer*** is added. The patch displacement matrices of K_1 feature channels are divided into T groups. Within each group, only the maximum activation is passed to the next layer. In this way, each feature is represented by multiple redundant channels. It allows to model a mixture of photometric transforms. During the training process, with the backpropagation algorithm, only the filter pair with the maximum response recieves the gradients and is updated. It drives filter pairs in the same group to compete for the gradients. Eventually, only one filter has large response to a training sample. Therefore, image patches have sparse responses with the learned filter pairs. It is well known that sparsity is a property to eliminate noise and redundancy. The output of the maxout grouping layer is TM $W_1 \times W_1$ displacement matrices. This is illustrated in Figure 5.

3.4. Modeling part displacement

Body parts can be viewed as adjacent patches. Another ***convolution and max-pooling layer*** is added on the top of patch displacement matrices to obtain the displacement matrices of body parts on a larger scale. It takes the MT $W_1 \times W_1$ patch displacement matrices as input and treat them as M $W_1 \times W_1$ images with T channels. Similar to the first convolutional layer, K_2 $m_2 \times m_2 \times T$ filters are applied to all the M images, and the output of this layer is M $W_2 \times W_2 \times K_2$ maps. The learned filters capture the local patterns of part displacements.

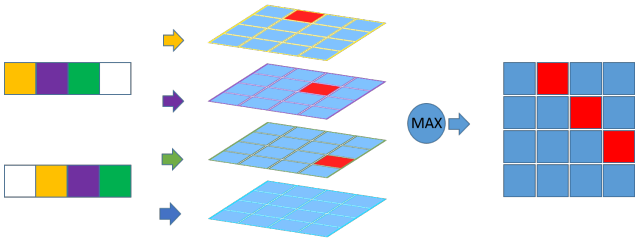


Figure 5. Maxout pooling. Left: Responses of patches to four filter pairs (indicated by the colors of yellow, purple, green and white) on two stripes. Middle: Four patch displacement matrices after passing the patch matching layer. Without maxout grouping, each matrix only has one patch with large response. Right: Group four channels together and take the maximum value to form a single channel output. A line structure is formed.

3.5. Modeling pose and viewpoint transforms

Pedestrians undergo various pose and viewpoint transforms. Such global geometric transforms can be viewed as different combinations of part displacement and their distributions are multi-modal. For example, two transforms can share the same displacement on upper bodies, but are different in the displacement of legs. Each output of a hidden node in the convolutional and maxpooling layer can be viewed as a possible part displacement detected with a particular visual feature. All of these hidden nodes form the input vector of the next ***fully connected layer***. In the next layer, each hidden node is a combination of all the possible part displacements and represents a global geometric transform. N hidden nodes are able to model a mixture of global geometric transforms.

3.6. Identity Recognition

The last ***softmax layer*** uses the softmax function to measure whether two input images belong to the same person or not given the global geometric transforms detected in the previous layer. Its output is a binary variable y defined as

$$p(y=i|\mathbf{a}_0, \mathbf{a}_1, b_0, b_1, \mathbf{x}) = \frac{e^{(\mathbf{a}_i \cdot \mathbf{x} + b_i)}}{\sum_i e^{(\mathbf{a}_i \cdot \mathbf{x} + b_i)}}. \tag{4}$$

Let $y = 1$ if two pedestrian images $(\mathbf{I}_n, \mathbf{J}_n)$ are matched, otherwise $y = 0$. $\mathbf{x}$ is the input from the previous layer. $\mathbf{a}_0$, $\mathbf{a}_1$, b_0 and b_1 are the combination weights and bias terms to be learned. Given the class labels of H training sample pairs, the negative log-likelihood is used as the cost for

training and could be written as

$$cost = -\sum_{n}^{H} y_n \log(p(y=1|\Phi,(\mathbf{I}_n,\mathbf{J}_n))) + (1-y_n)\log(1-p(y=1|\Phi,(\mathbf{I}_n,\mathbf{J}_n))). \quad (5)$$

It exerts large penalty for misclassified samples. For example, if $y_n = 0$ and $p(y = 1|\Phi,(\mathbf{I}_n,\mathbf{J}_n)) = 1$, $(1 - y_n)log(1-p(y=1|\Phi,(\mathbf{I}_n,\mathbf{J}_n))) \rightarrow -\infty$. Φ represents the set of parameters of the whole neural network to be learned.

4. Training Strategies

Our training algorithm adopts the mini-batch stochastic gradient descent proposed in [11]. The training data is divided into mini-batches. The training errors are calculated upon each mini-batch in the soft-max layer and get backpropogated to the lower layers. In addition, several carefully designed training strategies are proposed.

4.1. Dropout

In person re-identification, due to large cross-view variations, misalignment, pose variations, and occlusions, it is likely for some patches on the same person (but in different views) to be mismatched. To make the trained FPNN tolerable to misdetection of patch correspondences, the dropout strategy [15] is adopted. For each training sample as input at each training iteration, some outputs of the first convolutional layer (that is, extracted features with the filter pairs) are randomly selected and set as zeros. Gradients in backpropogation are calculated with those randomly muted filter responses to make the trained model more stable.

4.2. Data Augmentation

In the training set, the matched sample pairs (positive samples) are several orders fewer than non-matched pairs (negative samples). If they are directly used for training, the network tends to predict all the inputs as being non-matched. We augment data by simple translational transforms on each pedestrian image. For an original pedestrian image of size $H_{im} \times W_{im}$, five images of the same size are randomly sampled around the original image center and their translations are from a uniform distribution in the range of $[-0.05H_{im}, 0.05H_{im}] \times [-0.05W_{im}, 0.05W_{im}]$. The matched sample pairs are enlarged by a factor of 25.

4.3. Data balancing

Each mini-batch keeps all the positive training samples and randomly selects the same number of negative training samples at the very beginning of the training process. The network achieves a resonably good configuration after the initial training. As the training process goes along, we gradually increase the number of negative samples in each mini-batch up to the ratio of $5:1$.

4.4. Bootstrapping

After the network has been stabilized, we continue to select difficult negative samples, which are predicted as matched pairs with high probabilities by the current network, and combine them with all the positive samples to further train the network iteratively. Because of the large number of negative training samples, it is very time-consuming to re-predict all the negative samples with the current network after each epoch. We only re-predict hard samples selected in the previous epoch. Since these samples have been used to update the network, their predictions are expected to have larger changes than other samples after the update.

Each negative sample x is assigned with a score s_k after each epoch k. Samples with the smallest s_k are selected to re-train the network. At the beginning,

$$s_0 = 1 - p(x \text{ is a matched pair}|\Phi_0),$$

where Φ_0 is the configuration of the network. If x is selected as a hard sample for training in the previous epoch k, its score is updated as

$$s_k = \frac{1 - p(x \text{ is a matched pair}|\Phi_k) + s_{k-1}}{2},$$

where Φ_k is the configuration of the network trained after epoch k; otherwise, $s_k = \lambda s_{k-1}$. The diminishing parameter λ is set as 0.99. This increases the chance of those negative samples not being selected for a long time.

5. Dataset

All of the existing datasets are too small to train deep neural networks. We build a much larger dataset[1] which includes $13,164$ images of $1,360$ pedestrians. It is named CUHK03, since we already published two re-id datasets (CUHK01 [25] and CUHK02 [24]) in previous works. A comparison of the scales can be found in Table 1. The whole dataset is captured with six surveillance cameras. Each identity is observed by two disjoint camera views and has an average of 4.8 images in each view. Some examples are shown in Figure 1(a). Besides the scale, it has the following characteristics.

(1) Apart from manually cropped pedestrian images, we provide samples detected with a state-of-the-art pedestrian detector [10]. This is a more realistic setting and poses new problems rarely seen in existing datasets. From Figure 1(a), we can see that *misalignment, occlusions* and *body part missing* are quite common in this dataset. The inaccurate detection also makes the geometric transforms complex. We further provide the original image frames and researchers can try their own detectors on this dataset.

[1]The dataset is available at http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

Table 1. Compare the sizes of our dataset (CUHK03) and existing person re-identification datasets.

	CUHK03	VIPeR [12]	i-LIDS [50]	CAVIAR [3]	Re-ID 2011 [16]	GRID [27]	CUHK01 [25]	CUHK02 [24]
No. of images	13,164	1,264	476	1,220	2,450	500	1,942	7,264
No. of persons	1,360	632	119	72	245	250	971	1,816

(2) Some existing datasets assume a single pair of camera views and their cross-view transforms are relatively simple. In our dataset, samples collected from multiple pairs of camera views are all mixed and they form complex cross-view transforms. Moreover, our cameras monitor an open area where pedestrians walk in different directions, which leads to multiple view transforms even between the same pair of cameras.

(3) Images are obtained from a series of videos recorded over months. Illumination changes are caused by weather, sun directions, and shadow distributions even within a single camera view. Our cameras have different settings, which also leads to photometric transforms.

6. Experimental Results

Most of the evaluations are conducted on the new dataset, since existing datasets are too small to train the deep model. An additional evaluation is on the CUHK01 [25]. Our dataset is partitioned into training set (1160 persons), validation set (100 persons), and test set (100 persons). Each person has roughly 4.8 photos per view, which means there are almost $26,000$ positive training pairs before data augmentation. A mini-batch contains 512 images pairs. Thus it takes about 300 mini-batches to go through the training set. The validation set is used to design the network architecture (the parameters of which are shown in Table 2). The experiments are conducted with 20 random splits and all the Cumulative Matching Characteristic (CMC) curves are single-shot results.

Each image is preprocessed with histogram equalization and transformed to the LAB color space. It is normalized to the size of $(64 \times 32 \times 3)$, and subtracted with the mean of all the pixels in that location. Our algorithm is implemented with GTX670 GPU. The training process takes about five hours to converge.

We compare with three person re-identification methods (KISSME [20], eSDC [48], and SDALF [9]), four state-of-the-art metric learning methods (Information Theoretic Metric Learning (ITML)[6], Logistic Distance Metric Learning (LDM) [14], Largest Margin Nearest Neighbor (LMNN)[45], and Metric Learning to Rank (RANK)[30]), and directly using Euclidean distance to compare features. LMNN and ITML are widely used metric learning algorithms and have been used for person re-identification in [25]. RANK is optimized for ranking problems, while person re-identification is a ranking problem. LDM is specifically designed for face and person identification problems. When using metric learning methods and Euclidean distance, the handcrafted features of dense color histograms and dense SIFT uniformly sampled from patches are adopted. Through extensive experimental evaluation in [48], it has been shown that these local features are more effective on person re-identification than most other features and the implementation is publicly available.

Table 2. Settings of the filter pairing neural network.

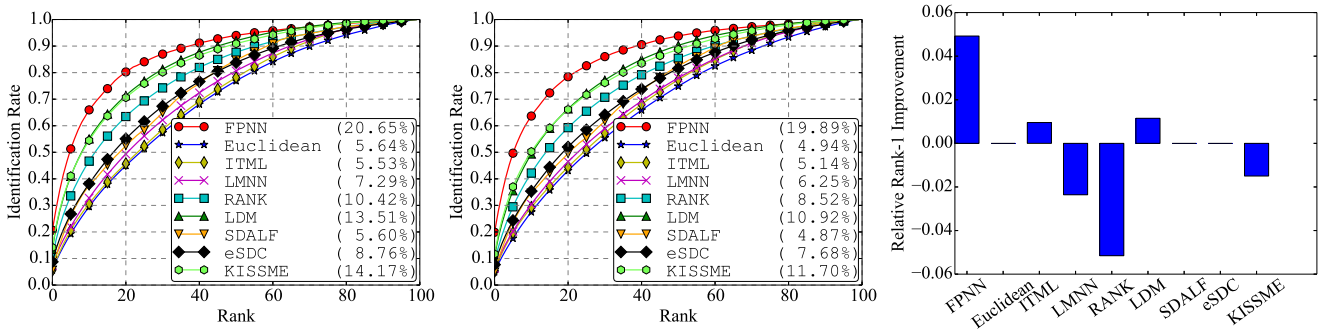
$H_{im} = 64$	$W_{im} = 32$	$K_1 = 64$	$m_1 = 5$
$H_0 = 60$	$W_0 = 28$	$H_1 = 20$	$W_1 = 9$
$M = 20$	$T = 16$	$m_2 = 3$	$W_2 = 3$
$K_2 = 16$	$N = 128$		

6.1. Experiments on our new dataset

On our CUHK03 dataset, we conduct comparisons using both manually labeled pedestrian bounding boxes and automatically detected bounding boxes. Figure 6(a) plots the CMC curves of using manually labeled bounding boxes. Our FPNN outperforms all the methods in comparison with large margins. The relative improvement on the Rank-1 identification rate is 46% compared with the best performing approach.

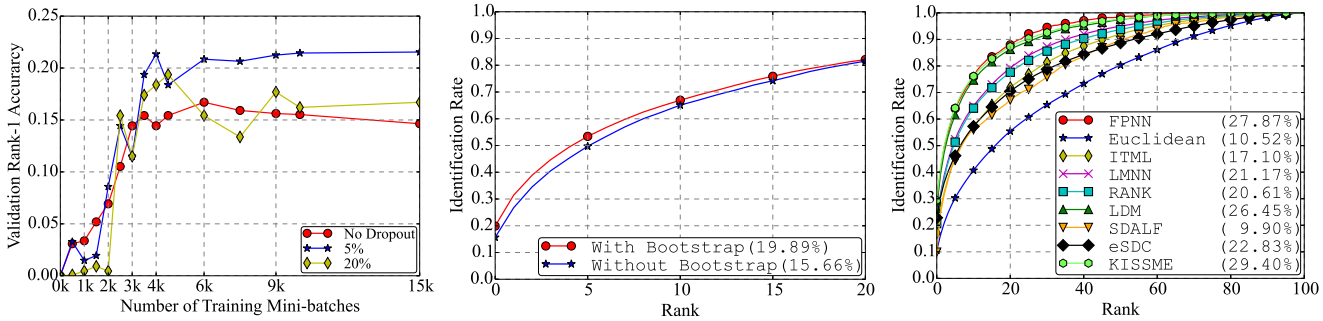
Figure 6(b) shows the results of using automatically detected bounding boxes, which cause misalignment. The performance of other methods drop significantly. For example, the Rank-1 identification rate of the best performing KISSME drops by 2.47%, while FPNN only drops by 0.76%. It shows that FPNN is more robust to misalignment.

In order to compare the learning capacity and generalization capability of different learning methods, we did another experiment by adding 933 images of 107 pedestrians to the training set, while keep the test set unchanged. Therefore, the training set has $1,267$ persons. These additional 933 images are captured from four camera views different from those in the test set. Adding training samples, which do not accurately match the photometric and geometric transforms in the test set, makes the learning more difficult. Figure 6(c) shows the changes of Rank-1 identification rates of different methods. It is observed that the performance of most of the methods drops, because their limited learning capacity cannot effectively handle a more complex training set and the mismatch between the training and test sets. On the contrary, the performance of our FPNN is improved because of its large learning capacity and also the fact that extra training samples improve the learned low-level features which can be shared by different camera settings.



(a) (b) (c)

Figure 6. Experimental results on our new dataset using manually labeled pedestrian bounding boxes (a) and automatically detected bounding boxes (b). Rank-1 identification rates are shown in parentheses. (c): After adding another 933 images of 107 persons to the training set, Rank-1 rate changes of different methods. The added images are collected from another four camera views different from those used in the test set. Automatically detected bounding boxes are used in (c).



(a) (b) (c)

Figure 7. (a): Rank-1 identification of FPNN on the validation set after different number of training mini-batches. (b): CMC curves of FPNN with and without bootstrap in training. Both (a) and (b) are evaluated on our new dataset. (c): CMC curves on the CUHK01 dataset.

6.2. Evaluation of training strategies

Experiments in Figure 7 (a) and (b) show the effectiveness of our dropout and bootstrapping training strategies. Figure 7(a) shows the Rank-1 identification rates after different numbers of training mini-batches on the validation set with dropout rates ranging from 0% to 20%. Without dropout, the identification rate decreases with more training mini-batches. It indicates that overfitting happens. With a 5% dropout rate, the identification rate is high and converges on the validation set. Dropout makes the trained FPNN tolerable to misdetection of patch correspondences and have good generalization power. If the dropout rate is high (e.g. 20%[2]), it cannot reach a good identification rate, even though the generalization power is good, because not enough features are passed to the next layer.

Figure 7(b) shows the CMC curves of FPNN with and without the bootstrapping strategy. Bootstrapping is effective in improving the Rank-1 identification rate from 15.66% to 19.89%. However, there is less difference on Rank-20. This may be attributed to the samples missed after Rank-20 are particularly difficult, while FPNN has given up fitting these extreme cases in order to be robust.

[2]In our case, 20% dropout in the first layer means on average roughly 36% of the patch matching layer outputs are set to zero due to Eqn 3.

6.3. Experiments on the CUHK01 dataset

We further evaluate FPNN on the CUHK01 dataset released in [25]. In this dataset, there are 971 persons and each person only has two images in either camera view. Again, 100 persons are chosen for test and the remaining 871 persons for training and validation. This dataset is challenging for our approach, since the small number of samples cannot train the deep model very well. There are only around 3,000 pairs of positive training samples on it (compared with 26,000 in our new dataset). Nevertheless, our FPNN outperforms most of the methods in comparison, except that its Rank-1 rate is slightly lower than KISSME. But its Rank-n ($n > 10$) rates are comparable to KISSME.

7. Conclusion

In this paper, we propose a new filter pairing neural network for person re-identification. This method jointly optimizes feature learning, photometric transforms, geometric transforms, misalignment, occlusions and classification

under a unified deep architecture. It learns filter pairs to encode photometric transforms. Its large learning capacity allows to model a mixture of complex photometric and geometric transforms. Some effective training strategies are adopted to train the network well. It outperforms state-of-the-art methods with large margins on a large scale benchmark dataset.

References

[1] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. Int'l Conf. Advanced Video and Signal Based Surveillance*, 2010. 1, 2

[2] S. G. C. Liu, C. C. Loy and G. Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013. 1

[3] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 1, 2, 6

[4] E. D. Cheng and M. Piccardi. Matching of objects moving across disjoint cameras. In *ICIP*, 2006. 2

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 3

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 2, 6

[7] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*. 2011. 1

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35:1915–1929, 2013. 3

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 2, 6

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *PAMI*, 2010. 2, 5

[11] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout networks. *CoRR*, 2013. 5

[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 1, 2, 6

[13] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 1, 2

[14] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 2, 6

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv e-prints*, 2012. 5

[16] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 1, 2, 6

[17] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *ICCV*, 2009. 3

[18] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *CVPR*, 2005. 2

[19] F. Jurie and A. Mignon. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 1, 2

[20] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 2, 6

[21] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[22] Y. LeCun and Y. Bengio. *The handbook of brain theory and neural networks*. MIT Press, 1998. 3

[23] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009. 3

[24] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 1, 2, 5, 6

[25] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 2, 5, 6, 7

[26] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *Proc. First Int'l Workshop on Re-Identification*, 2012. 1, 2

[27] C. C. Loy and T. Xiang. Multi-camera activity correlation analysis. In *CVPR*, 2009. 1, 2, 6

[28] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014. 3

[29] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012. 1

[30] B. Mcfee and G. Lanckriet. Metric learning to rank. In *ICML*, 2010. 2, 6

[31] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, 2014. 3

[32] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013. 3

[33] F. Porikli. Inter-camera color calibration by correlation model function. In *ICIP*, 2003. 2

[34] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer function. In *BMVC*, 2008. 2

[35] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 1, 2

[36] M. Ranzato, F.-J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, 2007. 3

[37] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least sqaures. In *SIBGRAPI*, 2009. 2

[38] P. Sermanet, K. Kavukcuoglu, and S. Chintala. Pedestrian detection with unsupervised and multi-stage feature learning. In *CVPR*, 2013. 3

[39] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 3

[40] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 3

[41] S. G. W. Zheng and T. Xiang. Re-identification by relative distance comparison. *PAMI*, 2013. 1, 2

[42] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34:3–19, 2013. 1

[43] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 2

[44] X. Wang and R. Zhao. *Person Re-Identification*, chapter Person Re-identification: System Design and Evaluation Overview, pages 351–370. Springer, 2014. 2

[45] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin distance metric learning for large margin. *JMLR*, 2009. 2, 6

[46] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, 2013. 3

[47] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 2

[48] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2, 6

[49] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2

[50] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 2, 6

[51] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 1, 2

[52] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity preserving face space. In *ICCV*, 2013. 3