Learning Mid-level Filters for Person Re-identification

Rui Zhao Wanli Ouyang Xiaogang Wang
Department of Electronic Engineering, the Chinese University of Hong Kong
{*rzhao, wlouyang, xgwang*}@ee.cuhk.edu.hk

Abstract

In this paper, we propose a novel approach of learning mid-level filters from automatically discovered patch clusters for person re-identification. It is well motivated by our study on what are good filters for person re-identification. Our mid-level filters are discriminatively learned for identifying specific visual patterns and distinguishing persons, and have good cross-view invariance. First, local patches are qualitatively measured and classified with their discriminative power. Discriminative and representative patches are collected for filter learning. Second, patch clusters with coherent appearance are obtained by pruning hierarchical clustering trees, and a simple but effective cross-view training strategy is proposed to learn filters that are view-invariant and discriminative. Third, filter responses are integrated with patch matching scores in RankSVM training. The effectiveness of our approach is validated on the VIPeR dataset and the CUHK01 dataset. The learned mid-level features are complementary to existing handcrafted low-level features, and improve the best Rank-1 matching rate on the VIPeR dataset by 14%.

1. Introduction

Person re-identification is to match pedestrian images observed from non-overlapping camera views based on appearance. It receives increasing attentions in video surveillance for its important applications in threat detection, human retrieval, and multi-camera tracking [23]. It saves a lot of human labor in exhaustively searching for a person of interest from large amounts of video sequences. Despite several years of research, person re-identification is still a very challenging task. A person observed in different camera views often undergoes significant variations in viewpoints, poses, and illumination. Background clutters and occlusions introduce additional difficulties. Moreover, since some persons share similar appearance, it is a big challenge to match a query pedestrian image with a large number of candidates from the gallery.

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).

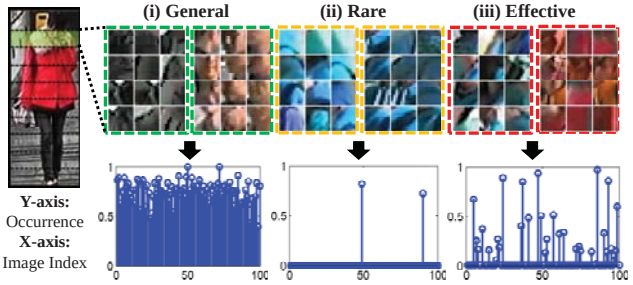


Figure 1. Three types of patches with different discriminative and generalization powers. Each dashed box indicates a patch cluster. To make filters region-specific, we cluster patches within the same horizontal stripe across different pedestrian images. See details in the text of Section 1. **Best viewed in color**.

Feature extraction is the most important component for the success of a person re-identification system. Different from existing approaches of using handcrafted features, which are not optimal for the task of person re-identification, we propose to learn mid-level *filters* from automatically discovered clusters of patches. A filter captures a visual pattern related to a particular body part. The whole work is motivated by our study on what are good filters for person re-identification and what are good patch clusters to train these filters, and how to quantify these observations for guiding the learning process.

(1) A good mid-level filter should reach the balance between discriminative power and generalization ability. As examples shown in Figure 1, we divide patches from training images into three categories. *General patches* - Patches in the green dashed boxes appear frequently in a large portion of pedestrian images. Filters learned from this type of patches are too general to discriminate pedestrian images. *Rare patches* - Patches in the yellow dashed boxes indicate patterns appearing in very few pedestrian images. Filters learned from this type of patches have very low generalization ability on new test images. Including the filters learned from these two types of patches in person re-identification increases computational cost. More importantly, they serve as noise channels and deteriorate the identification perfor-

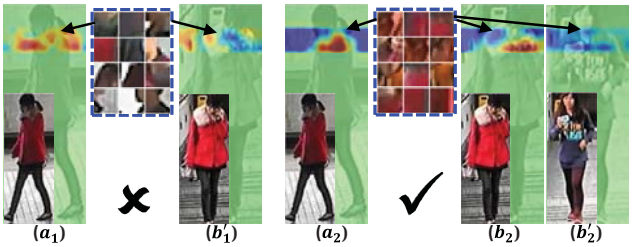


Figure 2. Filter in $(a_1)(b_1)$ is learned from a cluster with incoherent appearance and generates scattered responses in the two images. Filter in $(a_2)(b_2)$ is learned from a cluster with coherent appearance. It generates compact responses. It also has view-invariance. It matches (a_2) and (b_2) which are the same person in different views, while distinguishes (b_2) and (b'_2) which are different person in the same view. **Best viewed in color**.

mance. The last category is *Effective patches*. Patches in the red dashed boxes appear in an appropriate proportion of pedestrian images. Filters learned from them are representative in describing common properties of a group of persons, and effective in discriminating identities. This study will show how to quantify the discriminative and generalization powers and how to select *effective patches* as candidates for filter learning.

(2) A filter should be learned from a cluster of patches with coherent appearance. However, due to the imperfection of clustering algorithms, some patch clusters have mixed visual patterns as shown in Figure 2 $(a_1)(b_1)$. The filter learned from this patch cluster cannot accurately locate a specific visual pattern, and therefore generates scattered filter responses. In contrast, the filter learned from the patch cluster in Figure 2 $(a_2)(b_2)$ generates compact responses and can accurately locate the target pattern.

(3) Human is a well-structured object with body parts (e.g. head, and torso). We wish that the patches of a cluster come from the same body part, such that the learned filter can capture the visual pattern of a particular body part.

(4) The learned filters should be robust to cross-view variations caused by body articulation, viewpoint and lighting changes. The filter responses in Figure 2 $(a_2)(b_2)$ are view-invariant. This filter well matches (a_2) and (b_2) which are images of the same person in different views, while distinguishes (b_2) and (b'_2) which are different persons in the same view.

Based on above observations, we propose a new approach of learning mid-level filters for person re-identification. Its contributions are in the following aspects:

(i) Based on observation (1), partial Area Under Curve ($pAUC$) score is proposed to measure the discriminative power of local patches. Discriminative and representative patches are collected based on partial AUC quantization.

(ii) Based on observation (2), hierarchical clustering trees are built to exploit visual patterns from local patches. Through pruning the trees, we collect coherent cluster nodes as primitives for filter learning.

(iii) Based on observation (3), all the patch matching, patch clustering, and filter learning are done within the same horizontal stripe as a spatial constraint. Moreover, patches are clustered by including their locations as features, such that the clustered patches are spatially proximate and region-specific.

(iv) Based on observation (4), a simple but effective cross-view training strategy is proposed to learn SVM filters that are view-invariant and discriminative. Filter responses are sparsified to eliminate noise and redundancy.

(v) Finally, matching scores of filter responses are integrated with patch matching in RankSVM training.

The effectiveness of the learned mid-level filters is shown through experiments on the VIPeR [6] and CUHK01 [12] datasets. It achieves the state-of-the-art performance. Our learned mid-level filters well complement to existing handcrafted low-level features. By combining with LADF [14], it significantly enhances the state-of-the-art by 14% on Rank-1 matching rate on the VIPeR dataset.

2. Related Works

Existing person re-identification approaches work on three aspects: distance learning [29, 4, 12, 18, 30, 16, 14], feature design and selection [5, 3, 17, 15, 28, 7, 20, 25], and mid-level feature learning [11, 22, 10, 13]. A review can be found in [25].

In distance learning, distance metrics are discriminatively optimized for matching persons. Zheng *et al*. [30] introduced a Probabilistic Relative Distance Comparison (PRDC) model to maximize likelihood of true matches having a relatively smaller distance than that of a wrong match pair. Mignon and Jurie [18] proposed Pairwise Constrained Component Analysis (PCCA) to learn a projection from high-dimensional input space into a low-dimensional space where the distance between pairs of data points respects the desired constraints. It exhibits good generalization properties in presence of high-dimensional data. Liu *et al*. [16] presented a man-in-the-loop method to allow user quickly refine ranking performance, and achieved significant improvement over other metric learning methods. Li *et al*. [14] developed a Locally-Adaptive Decision Function (LADF) that jointly models a distance metric and a locally adaptive thresholding rule, and achieved good performance.

In feature design and selection, research works can be further divided into *unsupervised* [5, 3, 17, 15, 28, 25] and *supervised* [7, 20] approaches. (i) Unsupervised approaches: Farenzena *et al*. [5] proposed the Symmetry-Driven Accumulation of Local Features (SDALF) by exploiting the symmetry property in pedestrian images to handle view variation. Cheng *et al*. [3] utilized the Pictorial Structures to estimate human body configuration and also computed visual features based on different body parts to

cope with pose variations. Liu *et al*. [15] learned a bottom-up feature importance to adaptively weight features of different individuals rather than using global weights. (ii) Supervised approaches: Gray *et al*. [7] used boosting to select a subset of optimal features for matching pedestrian images. Prosser *et al*. [20] formulated person re-identification as a ranking problem, and learned global feature weights based on an ensemble of RankSVM.

Some research works on person re-identification have been done to learn reliable and effective mid-level features. Layne *et al*. [11] proposed to learn a selection and weighting of mid-level semantic attributes to describe people. Song *et al*. [22] used human attributes to prune a topic model and matched persons through Bayesian decision. However, learning human attributes require attribute labels for pedestrian images which cost human labor. It is much more costly than labeling matched pedestrian pairs, since each pedestrian image could have more than 50 attributes. Li *et al*. [13] proposed a deep learning framework to learn filter pairs, which encode photometric transforms across camera views for person re-identification. However, it requires larger scale training data. In this work, we automatically learn discriminative mid-level features without annotation of human attributes.

In a wider context, mid-level feature learning has been exploited in recent works on several vision topics. Singh *et al*. [21] and Jain *et al*. [8] learned mid-level features in scene classification and action recognition by patch clustering and measuring of the purity and discriminativeness with detection scores. Different from these works, we use hierarchical clustering and pruning to find coherent patch clusters, and jointly measures the representative and discriminative powers by proposed partial AUC quantization. Moreover, due to the nature of re-identification problem, our mid-level filter learning targets on cross-view invariance and considers constraints of body parts through patch matching. This is the first study of importance of mid-level filtering in person re-identification. Existing works on mid-level feature learning did not consider special challenges in person re-identification.

3. Patch Classification

Our mid-level filters are learned from local patches selected by their discriminative and representative powers at different locations. The discriminative power of a patch is quantified with its appearing frequency in pedestrian images. This is implemented by patch matching and computing **partial Area Under Curve** (**pAUC**) score. In this section, we introduce how to build dense correspondence of patches between images in different views, and how to perform partial AUC quantization based on matching results.

3.1. Dense Correspondence

Dense features. Local patches on a dense grid are extracted. The patch is of size 10×10 and the grid step is 5 pixels. 32-bin color histogram and 128-dimensional SIFT features in each of LAB channels are computed for each patch. To robustly capture the color information, color histograms are also computed on another two downsampled scales for each patch with downsampling factors 0.5 and 0.75. The color histograms and SIFT features are normalized with $L2$ norm, and are concatenated to form the final 672-dimensional dense local features.

Constrained Patch Matching. Dense local features for an image are denoted by $\mathbf{x}^{A,u} = \{x^{A,u}_{m,n}\}$, and $x^{A,u}_{m,n}$ represents the features of a local patch at the m-th row and n-th column in the u-th image from camera view A, where $m = 1, \ldots, M,\ n = 1, \ldots, N,\ u = 1, \ldots, U$. When patch $x^{A,u}_{m,n}$ searches for its corresponding patch in the v-th image from camera view B, *i.e.* $\mathbf{x}^{B,v} = \{x^{B,v}_{i,j}\}, v = 1, \ldots, V$, the constrained search set of $x^{A,u}_{m,n}$ in $\mathbf{x}^{B,v}$ is

$$\begin{aligned}\mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v}) = \{x^{B,v}_{i,j} \mid j = 1, \ldots, N, \\ i = \max(0, m-h), \ldots, \min(M, m+h)\},\end{aligned} \quad (1)$$

where h denotes the height of the search space. If all pedestrian images are well aligned and there is no vertical pose variation, h should be zero. However, misalignment, camera view change, and vertical articulation result in vertical movement of the human body in the image. Thus it is necessary to relax h to be greater than 0 for handling vertical movement. We choose $h = 2$ in our experiment setting.

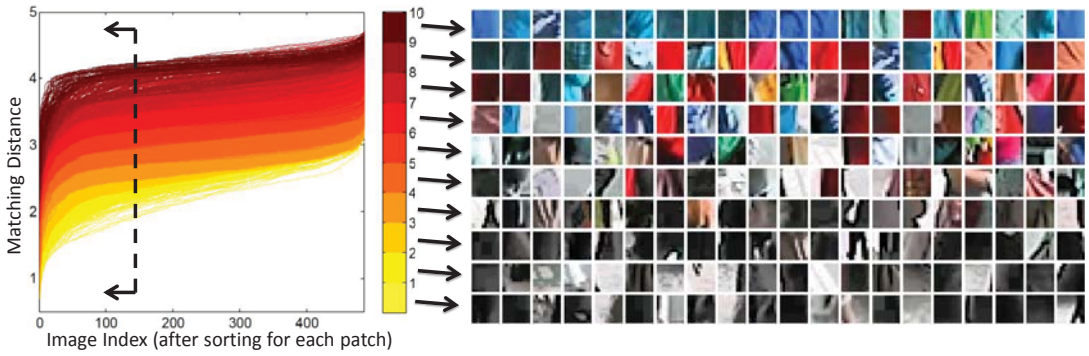
Patch matching is widely used, and many off-the-shelf fast algorithms [1] are available for faster speed. In this work, we simply do a nearest-neighbor search for patch $x^{A,u}_{m,n}$ in its search set $\mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v})$. For each patch $x^{A,u}_{m,n}$, a nearest neighbor (NN) is sought from its search set in every image from camera view B to build an NN set $X_{NN}(x^{A,u}_{m,n})$,

$$\begin{aligned}X_{NN}(x^{A,u}_{m,n}) = \{x | x = \underset{x^{B,v}_{i,j}}{\operatorname{argmin}}\ d(x^{A,u}_{m,n}, x^{B,v}_{i,j}), \\ x^{B,v}_{i,j} \in \mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v}), v = 1, \ldots, V\},\end{aligned} \quad (2)$$

where $\mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v})$ is the constrained search set defined in Eq.(1), and $d(x^{A,u}_{m,n}, x^{B,v}_{i,j})$ denotes the Euclidean distance between $x^{A,u}_{m,n}$ and $x^{B,v}_{i,j}$.

3.2. Partial AUC quantization

Partial AUC Score. To quantify the discriminative and generalization power of local patches in discriminating identities, we propose to compute $pAUC$ score based on the matching distances obtained in constrained patch matching. Because the matching distances between a patch $x^{A,u}_{m,n}$ and its closer neighbors in $X_{NN}(x^{A,u}_{m,n})$ are more meaningful to describe the ability of the patch in distinguishing similar



(a) (b)

Figure 3. (a): Each curve represents the sorted distances between a patch and patches in its nearest-neighbor set, and $pAUC$ score is computed by accumulating patch matching distances with the N_p closest nearest neighbors, as the black arrows indicated. Different color indicates different $pAUC$ level. (b): Example patches are randomly sampled from each $pAUC$ level for illustration, patches in low $pAUC$ levels are monochromatic and frequently seen, while those in high $pAUC$ levels are varicolored and less frequently appeared. Clearly the examples show the effectiveness of the $pAUC$ score in quantifying the discriminative power. **Best viewed in color**.

patches from other images, $pAUC$ score is defined as the cumulation of distances between the patch $x^{A,u}_{m,n}$ and its K nearest neighbors in $X_{NN}(x^{A,u}_{m,n})$,

$$s^{pAUC}(x^{A,u}_{m,n}) = \sum_{k=1}^{N_p} d_k(X_{NN}(x^{A,u}_{m,n})), \tag{3}$$

where d_k denotes the distance between patch $x^{A,u}_{m,n}$ and its k-th nearest neighbor in $X_{NN}(x^{A,u}_{m,n})$, and N_p is the number of nearest neighbors included in computing $pAUC$ score. We set $N_p = 0.3V$ in our experiments. Small s^{pAUC} implies that the patch $x^{A,u}_{m,n}$ has lots of similar patches from camera view B, and it is too general to describe a specific group of persons. Large s^{pAUC} implies that the patch $x^{A,u}_{m,n}$ is dissimilar with most of the patches from another view, and it can only describe few persons that have similar appearance. Median s^{pAUC} indicates that the patch $x^{A,u}_{m,n}$ is similar with a portion of patches from another view, and it has the ability to describe the common properties of a group of persons.

Quantization. To consider the patches in different body parts separately, we firstly divide local patches into N_Y horizontal stripes as follows,

$$S^A_y = \{x^{A,u}_{m,n} \mid m = (y-1) \times \left\lfloor \frac{M}{N_Y} \right\rfloor + 1, \ldots, y \times \left\lfloor \frac{M}{N_Y} \right\rfloor, \\ n = 1, \ldots, N,\ u = 1, \ldots, U.\},\ y = 1, \ldots, N_Y, \tag{4}$$

where S^A_y is the patch set in the y-th stripe, and N_Y is the number of stripes. Then, we uniformly quantize the patches within a stripe into N_L $pAUC$ levels according to their $pAUC$ scores as follows,

$$S^A_{y,l} = \Big\{x \mid x \in S^A_y, \\ s^{pAUC}(x) \in [s^A_{y,min} + (l-1)\frac{\Delta s^A_y}{N_L},\ s^A_{y,min} + l\frac{\Delta s^A_y}{N_L}]\Big\}, \tag{5}$$

where $s^A_{y,min} = \min\{s^{pAUC}(x) \mid x \in S^A_y\}$, $s^A_{y,max} = \max\{s^{pAUC}(x) \mid x \in S^A_y\}$, $\Delta s^A_y = s^A_{y,max} - s^A_{y,min}$, and $l = 1, \ldots, N_L$. In our experiments, patches are quantized into $N_Y = 10$ stripes and each stripe are quantized into $N_L = 10$ $pAUC$ levels, as illustrated in Figure 3.

4. Learning Mid-level Filters

4.1. Hierarchical Patch Clustering

Although local patches are classified using partial AUC quantization, patches with different visual information are still mixed together. Therefore, clustering is performed to group patches into subsets with coherent visual appearance.

In our task, patch features usually have high dimensions, and the distributions of data are often in different densities, sizes, and form manifold structures. Therefore, graph degree linkage (GDL) algorithm [26] is employed for clustering patches since it can well handle these problems. An ideal filter should be learned from a cluster of patches with coherent appearance. However, due to the imperfection of the clustering algorithm and the difficulty of determining appropriate cluster granularities, some patch clusters have mixed visual patterns. Filters learned from these patch clusters cannot accurately locate specific visual patterns. We propose to build a hierarchical tree by clustering patches from coarse to fine granularities, and find coherent patch clusters through pruning the tree. Given a set of patches $S^A_{y,l}$ in Eq.(5), we build a hierarchical clustering tree with order O_t and maximal depth D_t, *i.e.*, each parent node in the tree has O_t children and there are D_t layers of nodes. We set $O_t = 4$ and $D_t = 10$ in experiments. The root node contains all the patches in set $S^A_{y,l}$, and other nodes in the tree are patch clusters from coarse to fine granularities. As shown in Figure 4 (dashed box), shallow nodes (in black color) are decomposed into deep nodes (in blue color) in

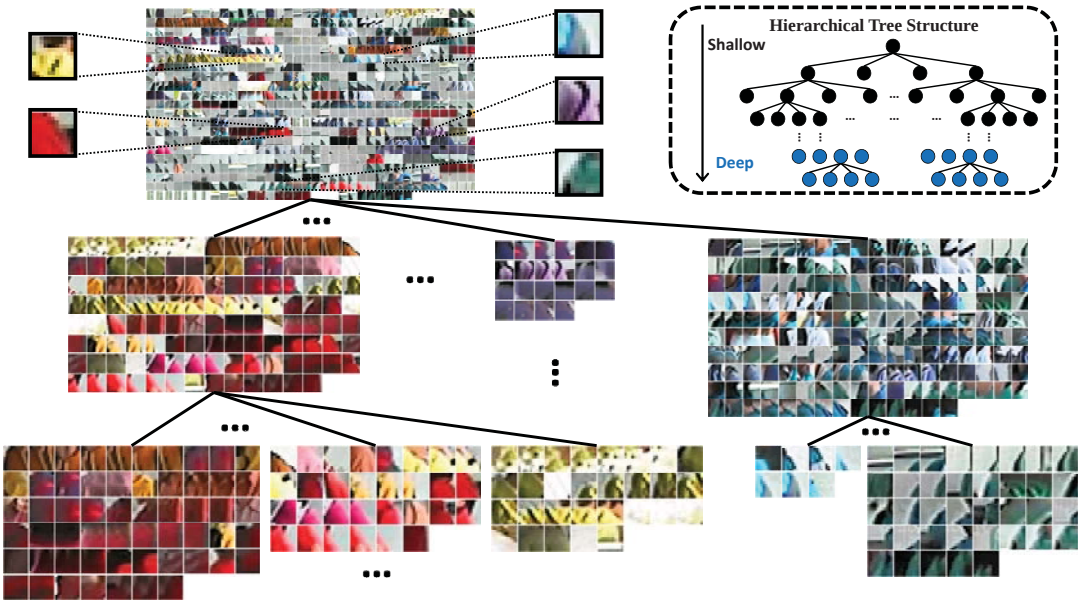


Figure 4. Illustration of hierarchical clustering tree structure, and examples of cluster nodes. As shown in the dashed box, patches in a parent node is divided into $O_t = 4$ children nodes, and shallow nodes (in black color) are decomposed into deep nodes (in blue color) in hierarchical clustering. The shallow nodes represent coarse clusters while the deep nodes denote finer clusters. Shallow nodes contain patches with different color and texture patterns while the patch patterns in the deep nodes are more coherent. **Best viewed in color**.

hierarchical clustering. The shallow nodes represent coarse clusters while the deep nodes denote finer clusters. To learn mid-level filters that can well describe a specific visual pattern and generate compact filter responses, we only retain the deep nodes in the hierarchical clustering tree, *i.e.* nodes are picked only when the number of patches in this node is below a threshold T_{max}. In addition, cluster nodes with very few patches are pruned with threshold T_{min}. Examples of patch clusters are shown in Figure 4. We set $T_{max} = 300$ and $T_{min} = 10$ in experiments. The set of retained cluster nodes are denoted by $\{Node_k\}_{k=1}^{N_{node}}$, where N_{node} is number of cluster nodes.

4.2. Learning Mid-level Filters

Given all the retained cluster nodes, we aim to learn filters that (1) are robust to appearance and lighting variations caused by viewpoint change, and (2) have the ability of distinguishing the most confusing images. We firstly perform an initial matching based on dense correspondence, and then learn mid-level filters in a supervised cross-view training strategy based on the initial matching.

Initial Matching. Since dense correspondence has been built for images from two camera views, images can be initially matched based on the patch matching scores,

$$s_0(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}) = \mathbf{w}_p^{\mathrm{T}} \mathbf{s}_p(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}), \tag{6}$$

$$\mathbf{w}_p^{\mathrm{T}} = [w_{p_1}, \ldots, w_{p_{MN}}],\ \mathbf{s}_p^{\mathrm{T}} = [s_{p_1}, \ldots, s_{p_{MN}}], \tag{7}$$

$$s_{p_i} = \exp\big(-d(x_{p_i}^{A,u}, x_{p_i'}^{B,v})/\sigma_p^2\big), \tag{8}$$

where p_i and p_i' are indices of a pair of matched patches in image $\mathbf{x}^{A,u}$ and $\mathbf{x}^{B,v}$ respectively, σ_p is a bandwidth parameter, and w_{p_i} is the weight for similarity score s_{p_i}. $\mathbf{w}_p$ will be automatically learned in an integrated matching model (see Section 4.3), and we initially set $w_{p_i} = \hat{s}^{pAUC}(x_{p_i}^{A,u})\hat{s}^{pAUC}(x_{p_i'}^{B,v})$, where the similarity scores are weighted by the consistence in normalized (unit-variance) $pAUC$ scores. The initial matching model will be used for cross-view training of mid-level filters.

Cross-view Training. The learned filters should be robust to cross-view variations caused by body articulation, viewpoint and lighting changes, and be discriminative to identify the same person from different persons. We learn a filter for each retained cluster node $Node_k$ in a simple but effective cross-view training scheme. All patches in $Node_k$ are put in a positive set X_k^+, and patches from the other cluster nodes in the same stripe are randomly sampled to form a negative set X_k^-. However, these are not enough to learn a robust and discriminative filter. To make sure that the learned filter robustly produce consistent responses in matched images from both views, for every patch $x_{p_i}^{A,u} \in Node_k$, its matched patch $x_{p_i'}^{B,u}$ in the matched image is added into an auxiliary positive patch set X_k^{aux+}, as shown by the red solid arrow in Figure 5. In this way, the learned filter can produce high filter responses in both views and be robust to cross-view variations. In another aspect, since initial matching model has some confusions in finding the true match from a portion of mismatched images, extra negative patches can be mined from most confusing mismatched images to avoid high filter responses on them,

i.e. we sample the matched patches of $x_{p_i}^{A,u}$ in mismatched images to build an auxiliary negative set X_k^{aux-} for learning filters, as shown by the blue dashed arrows in Figure 5. Since images in top ranks are more confusing, the sampling is based on a decreasing probability distribution, as shown in the bottom right of Figure 5.

After construction of positive and negative patch set, we simply train a linear SVM [2] filter $\{\mathbf{w}_k, b_k\}$ using the train data $\{X_k^+, X_k^-, X_k^{aux+}, X_k^{aux-}\}$ for each cluster node $Node_k$. Because the patches in cluster node $Node_k$ belong to the y_k-th stripe, the corresponding SVM filter is spatially constrained within the stripe. With a set of SVM filters $\{\mathbf{w}_k, b_k, y_k\}_{k=1}^{N_{node}}$, filter responses are computed for each image by max-pooling of detection scores within the y_k-th stripe. We denote the filtering responses for the u-th image in view A as $\mathbf{f}^{A,u} = \{f_k^{A,u}\}_{k=1}^{N_{node}}$.

Normalization and Sparsity. For each filter, its responses are firstly normalized with $L2$ norm along all images to ensure that it is equally active as other filters. Then, for each image, responses of all filters are normalized with $L2$ norm. We denote by $\hat{\mathbf{f}}^{A,u}$ the normalized filter responses of u-th image in view A. To suppress noise in the filter responses, sparsity is then enforced on $\hat{\mathbf{f}}^{A,u}$ by $\|\hat{\mathbf{f}}^{A,u}\|_0 \leq N_{sparse}$. As suggested by the evaluation result in 5.2, we set $N_{sparse} = 0.5N_{node}$ in experiments. The sparsified filter response is denoted by $\hat{\mathbf{f}}_*^{A,u}$. Similarly, filter responses $\hat{\mathbf{f}}_*^{B,v}$ in camera view B can also be obtained.

4.3. Integrated Matching Scores

We integrate filter responses $\hat{\mathbf{f}}_*^{A,u}$ and $\hat{\mathbf{f}}_*^{B,v}$ with the initial matching scores in Eq.(6) into a unified matching model,

$$s_{int}(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}) = \mathbf{w}^{\mathrm{T}} \mathbf{\Phi}(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}, \hat{\mathbf{f}}_*^{A,u}, \hat{\mathbf{f}}_*^{B,v}), \quad (9)$$

$$\mathbf{\Phi}(\mathbf{x}^{A,u}, \mathbf{x}^{B,v})^{\mathrm{T}} = [\mathbf{s}_p(\mathbf{x}^{A,u}, \mathbf{x}^{B,v})^{\mathrm{T}}, \mathbf{s}_f(\hat{\mathbf{f}}_*^{A,u}, \hat{\mathbf{f}}_*^{B,v})^{\mathrm{T}}], \quad (10)$$

$$\mathbf{s}_f(\hat{\mathbf{f}}_*^{A,u}, \hat{\mathbf{f}}_*^{B,v}) = [s_{f_1}, \ldots, s_{f_{N_{node}}}]^{\mathrm{T}}, \quad (11)$$

$$s_{f_k} = \exp\Big(-(\hat{f}_{k*}^{A,u} - \hat{f}_{k*}^{B,v})^2 / \sigma_f^2\Big), \quad (12)$$

where $\mathbf{s}_p(\mathbf{x}^{A,u}, \mathbf{x}^{B,v})$ is patch matching scores defined in Eq.(6), s_{f_k} is the matching score between the k-th filter responses $\hat{f}_{k*}^{A,u}$ and $\hat{f}_{k*}^{B,v}$, σ_f is a bandwidth parameter, and $\mathbf{w}$ is the unified weighting parameters which are learned by RankSVM training [27].

5. Experimental Results

5.1. Datasets and Evaluation Protocol

We evaluate our approach on two public datasets, *i.e.* the VIPeR dataset [6] and the CUHK01 dataset [12]. The VIPeR dataset is the mostly used person re-identification dataset for evaluation, and the CUHK01 dataset contains more images than VIPeR (3884 $vs.$ 1264 specifically). Both

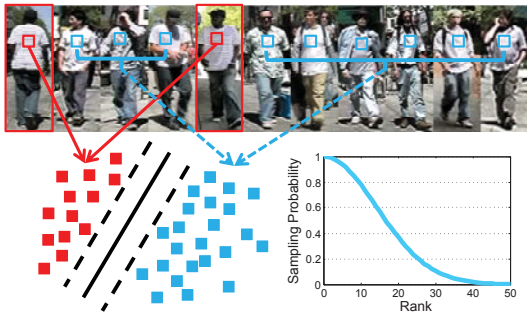


Figure 5. Scheme of learning view invariant and discriminative filters. Patches in red boxes are matched patches from images of the same person, while those in blue boxes are matched patches in most confusing images. Bottom right is the probability distribution for sampling auxiliary negative samples.

are very challenging datasets for person re-identification because they show significant variations in viewpoints, poses, and illuminations, and their images are of low resolutions, with occlusions and background clutters. All the quantitative results are reported in standard Cumulated Matching Characteristics (CMC) curves [24].

VIPeR Dataset[1] **[6]** was captured from two hand-carried cameras in outdoor academic environment. They were placed at many different locations forming different view pairs. It contains 632 pedestrian pairs, and each pair has two images of the same person observed from different camera views. Most of the image pairs show viewpoint change larger than 90 degrees. All images are normalized to 128×48 for evaluations.

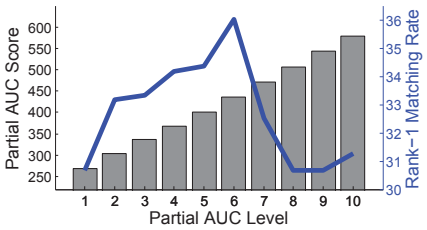
CUHK01 Dataset[2] **[12]** is also captured with two camera views in a campus environment, and it contains 971 persons, each of which has two images from two camera views. Camera A captures the frontal view or back view of pedestrians, while camera B captures the side view. All the images are normalized to 160×60 for evaluations.

5.2. Evaluations and Analysis

Evaluation of Partial AUC Quantization. We investigate the influence of partial AUC quantization on the rank-1 matching rate in re-identification. As shown by the results in Figure 6(a), median $pAUC$ levels have the highest Rank-1 performance because patches in these $pAUC$ levels are representative in describing common properties of a group of persons and effective in discriminating identities. Low $pAUC$ levels obtain lower performance than median levels since patches with low $pAUC$ score appear frequently in pedestrian images and are too general to discriminate identities. High $pAUC$ levels have the lowest performance because patches in these levels appear very few in pedestrian images and have low generalization power.

[1] http://vision.soe.ucsc.edu/projects

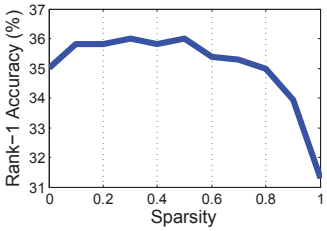
[2] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html



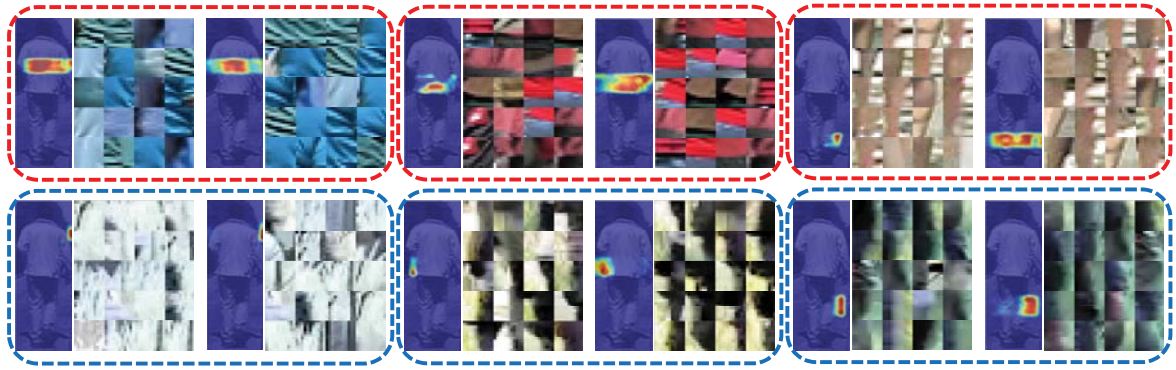
(a) Evaluation of Partial AUC quantization

CMC	*S1*	*S2*	*S3*	*S4*
Rank 1	34.09	34.71	35.12	**35.12**
Rank 5	53.64	54.05	53.64	**54.67**
Rank 10	63.49	62.90	63.93	**64.96**
Rank 20	75.86	75.45	74.42	**77.10**
Rank 30	81.01	80.60	81.21	**83.48**

(b) Evaluation of cross-view training

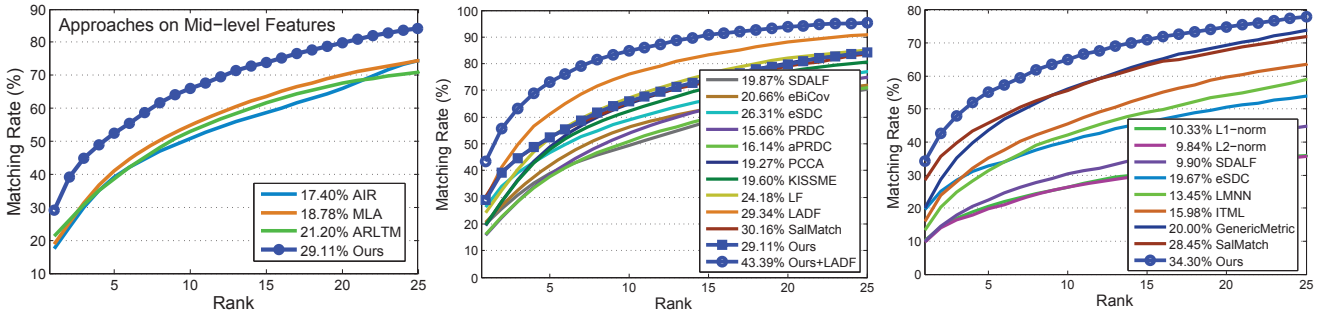


(c) Evaluation of sparsity



(d) Examples of positive training patches and high-score testing patches of high-weight and low-weight filters respectively

Figure 6. Evaluation and Analysis on the CUHK01 Dataset. (a): Average $pAUC$ score and rank-1 matching rate in each $pAUC$ level. Gray bars show the average $pAUC$ scores in 10 $pAUC$ levels, and the blue line indicates the rank-1 matching rates in (%). (b): Matching rate (%) in top ranks with different training strategies, *i.e.* *S1*:(X^+, X^-), *S2*:(X^+, X^-, X^{aux+}), *S3*:(X^+, X^-, X^{aux-}), and *S4*:$(X^+, X^-, X^{aux+}, X^{aux-})$. (c): Rank-1 Performance of enforcing different sparsity in filter response. Larger sparsity indicates there are more zero responses. (d): Example of learned filters. Each dashed box corresponds to a filter, and the contents in each dashed box are: average spatial distribution and examples of positive training patches, and average spatial distribution and examples of high-score testing patches. Examples in red dashes boxes (first row) have high weights while those in blue dashed boxes (second row) have low weights.



(a) VIPeR Dataset (b) VIPeR Dataset (c) CUHK01 Dataset

Figure 7. CMC on the VIPeR dataset and the CUHK01 Dataset. Rank-1 matching rate is marked before the name of each approach.

Evaluation of Cross-view Training. To validate the effectiveness of cross-view training in Section 4, we evaluate on the CUHK01 dataset with four controlled settings, *i.e.* *S1*:(X^+, X^-), *S2*:(X^+, X^-, X^{aux+}), *S3*:(X^+, X^-, X^{aux-}), and *S4*:$(X^+, X^-, X^{aux+}, X^{aux-})$, where X^+ (X^{aux+}) represents filter learning using (auxilary) positive samples, similarily for $X^-(X^{aux-})$. As shown in Figure 6(b), *S4* has better performance because it considers both view invariant property and ability in distinguishing confusing images. Thus, we adopt *S4* in training.

Evaluation of Sparse Filtering. We also evaluate the effectiveness of enforcing sparsity in filter response. As shown in Figure 6(c), rank-1 performance varies as the sparsity (percentage of zeros in filter responses) changes, and the performance is stable within $[0, 0.5]$.

Evaluation of Learned Filters. The learned weighting parameters $\mathbf{w}$ in Eq.(9) indicate the importance of response for each filter. As shown in Figure 6(d), each dashed box corresponds to a filter, and it contains examples of positive training patches with their average spatial distribution, and examples of high-score testing patches with their average spatial distribution. As seen from the spatial distributions, high-weight filters in red dashed boxes (first row) have discriminative visual patterns and mostly focus on human body part, while low-weight filters in blue dashed box (second row) either locate in background or have less meaningful

visual patterns.

5.3. Comparison with State-of-the-Arts

Evaluation Protocol. Our experiments on both datasets follow the evaluation protocol in [7], *i.e*. we randomly partition the dataset into two even parts, 50% for training and 50% for testing, without overlap on person identities. Images from camera A are used as probe and those from camera B as gallery. Each probe image is matched with every image in gallery, and the rank of correct match is obtained. Rank-k matching rate is the expectation of correct match at rank k, and the cumulated values of recognition rate at all ranks is recorded as one-trial CMC result. 10 trials of evaluation are conducted to obtain stable statistics, and the expectation is reported.

Result and Analysis. On the VIPeR dataset, we firstly compare with approaches on learning mid-level features, *i.e*. AIR [11], MLA[10], and ARLTM[22]. As shown in Figure 7(a), our approach significantly outperform all other methods in this category, which validates the effectiveness of our mid-level filters. We also compare our approach with benchmarking methods including SDALF [5], eBiCov [17], eSDC [28], PRDC [29], aPRDC [15], PCCA [18], KISSME [9], LF [19], SalMatch[27] and LADF[14]. As shown in Figure 7(b), our approach achieves rank-1 accuracy 29.11% and outperforms almost all the benchmarking methods. By combining with the best performing LADF under the same training / testing partitions, it significantly enhances the state-of-the-art by 14% on the rank-1 matching rate. On the CUHK01 dataset, our approach is compared with $L1$-norm distance, $L2$-norm distance, SDALF[5], eSDC[28], LMNN [12], ITML [12], GenericMetric [12], and SalMatch [27]. As the Figure 7(c) shows, our approach clearly outperform all previous methods on this dataset. One possible reason of the larger improvement compared with the results on the VIPeR dataset is that images in the CUHK01 dataset are of finer resolution, in which filters are better learned.

6. Conclusion

In this paper, we propose to learn mid-level filters for person re-identification. We explore different discriminative abilities of local patches by introducing $pAUC$ score. Discriminative and representative local patches are collected for learning filters. Coherent patch clusters are obtained by pruning hierarchical clustering trees, and a simple but effective cross-view training strategy is propose to learn filters that are view invariant and discriminative in distinguishing identities. Furthermore, matching scores of filter responses are integrated with patch matching in RankSVM training. Experimental results show the learned mid-level filters greatly improves the performance of person re-identification.

References

[1] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*, 2010. 3

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. In *ACM TIST*, 2011. 6

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

[4] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2011. 2

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2, 8

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

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 2, 3, 8

[8] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis. Representing videos using mid-level discriminative patches. In *CVPR*, 2013. 3

[9] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 8

[10] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *ECCV Workshops*, 2012. 2, 8

[11] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, 2012. 2, 3, 8

[12] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 2, 6, 8

[13] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2, 3

[14] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 2, 8

[15] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012. 2, 3, 8

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

[17] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012. 2, 8

[18] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2, 8

[19] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 8

[20] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2, 3

[21] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 3

[22] M. Song, C. Chen, J. Bu, X. Liu, Q. Zhao, and D. Tao. Attribute-restricted latent topic model for person re-identification. In *Pattern Recognition*, 2012. 2, 3, 8

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

[24] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 6

[25] X. Wang and R. Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*. Springer, 2014. 2

[26] W. Zhang, X. Wang, D. Zhao, and X. Tang. Graph degree linkage: Agglomerative clustering on a directed graph. In *ECCV*, 2012. 4

[27] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 6, 8

[28] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2, 8

[29] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 2, 8

[30] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. In *PAMI*, volume 35, pages 653–668, 2013. 2