Viewpoint Invariant Human Re-identification in Camera Networks Using Pose Priors and Subject-Discriminative Features

Ziyan Wu, *Student Member, IEEE,* Yang Li, *Student Member, IEEE,* and Richard J. Radke, *Senior Member, IEEE*

Abstract—Human re-identification across cameras with non-overlapping fields of view is one of the most important and difficult problems in video surveillance and analysis. However, current algorithms are likely to fail in real-world scenarios for several reasons. For example, surveillance cameras are typically mounted high above the ground plane, causing serious perspective changes. Also, most algorithms approach matching across images using the same descriptors, regardless of camera viewpoint or human pose. Here, we introduce a re-identification algorithm that addresses both problems. We build a model for human appearance as a function of pose, using training data gathered from a calibrated camera. We then apply this "pose prior" in online re-identification to make matching and identification more robust to viewpoint. We further integrate person-specific features learned over the course of tracking to improve the algorithm's performance. We evaluate the performance of the proposed algorithm and compare it to several state-of-the-art algorithms, demonstrating superior performance on standard benchmarking datasets as well as a challenging new airport surveillance scenario.

Index Terms—Human Re-Identification, Viewpoint Invariance, Camera Networks

1 Introduction

Recognizing the same human as he or she moves through a network of cameras with non-overlapping fields of view is an important and challenging problem in security and surveillance applications. This is often called the re-identification or "re-id" problem. For example, in an airport security surveillance system, once a target has been identified in one camera by a user or program, we want to learn the appearance of the target and recognize him/her when he/she is observed by the other cameras. We call this type of re-id problem "tag-and-track".

Unfortunately, current re-id algorithms are likely to fail in real-world tag-and-track scenarios for several reasons. The standard datasets used to evaluate re-id algorithms (see Figure 1) are all images taken from cameras whose optical axes have a small angle with (or are even parallel to) the ground plane, which is generally not the case in real-world surveillance applications. In the latter environments, the angle between the camera optical axis and the floor is usually large ($\sim 45^\circ$), causing serious perspective changes.

(a) (b)

Fig. 1. (a) Sample images from the VIPeR dataset [20]. (b) Sample images from the iLids dataset [38].

More importantly, most re-id algorithms approach matching across images using the same descriptors, regardless of camera viewpoint or human pose, which can induce serious error in the matching of target candidates.

In this paper, we propose a novel viewpoint-invariant approach to re-identify target humans in cameras that don't share overlapping fields of view. The approach is designed to be directly applicable to typical real-world surveillance camera networks. It improves the traditional person re-identification process with three contributions, as shown in Figure 2. First, we introduce a sub-image rectification method to cope with perspective distortion, which is common in surveillance cameras and may cause serious errors in matching. Second, we show that pairs of person descriptors from a traditional feature

• *The authors are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, 12180. E-mail: ziyan@alum.rpi.edu, liy21@rpi.edu, rjradke@ecse.rpi.edu*

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the U.S. Department of Homeland Security. Thanks to Edward Hertelendy and Michael Young for supplying the data in Section 6.4. Thanks to Fei Xiong for the PCCA implementation used in this paper.

extraction method vary significantly with viewpoint. Hence, we propose a viewpoint-invariant descriptor that takes into account the viewpoint of the human using what we call a pose prior learned from training data. Finally, complementing the traditional offline specification of target features, we show how person-specific discriminative features can be learned online for re-identification. The proposed algorithms can easily be introduced into current metric learning based re-id algorithms. We test our algorithms on both standard benchmarking datasets and a challenging new dataset acquired at a US airport, demonstrating that the proposed algorithm significantly improves the performance of current state-of-the-art metric learning based re-id algorithms.

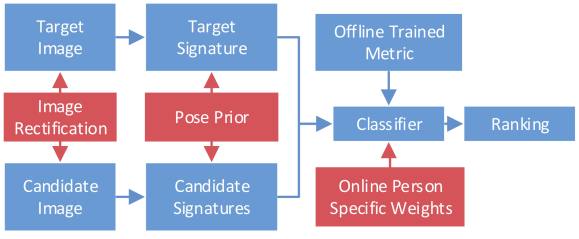


Fig. 2. Improvements to the re-id process with the contributions proposed in this paper. Blocks in blue are steps in a traditional re-id process while blocks in red are the new steps proposed in this paper.

This paper is organized as follows. Section 2 introduces related work in the field of human re-identification. Section 3 describes a sub-image rectification method used to eliminate the perspective distortion in images from typical surveillance cameras. Section 4 introduces the proposed pose prior and its application to feature extraction and offline classifier training. It also describes the online learning of subject-discriminative features and its integration with the offline classifier in descriptor matching. Section 5 describes how the proposed algorithms can be applied in combination with current metric learning methods in practical real-world scenario. Section 6 presents the experimental results on multiple standard datasets, as well as a new dataset collected from a surveillance camera network at an airport.

2 Related Work

Traditional biometric methods such as face [21], gait [46] or silhouette [44] recognition have been widely used in human identity recognition; however, they are difficult to apply to the re-id problem since most surveillance cameras' resolution is too poor to get a clear image of the target. This is quite different from the case of person identification with high-definition videos [16], [27], [42].

Instead, most recently proposed re-identification algorithms take a feature-based approach. These approaches mainly focus on feature representation and metric learning. There are several types of feature models used in re-id problems. Haar and DCD based features [3] are discriminative but are not robust to illumination and viewpoint changes. Color and texture histograms [20], [40], [50] are computationally efficient and considered to be more robust. Histogram Plus Epitome (HPE) [8], Quantized local feature histograms [43] and Mean Riemannian Covariance patches [5] take into account recurrent local patterns or accumulated local features to achieve more stable performance. However, they depend on multiple images. Exploiting symmetry and asymmetry features [17] can effectively discriminate candidates without an offline trained metric. However, the computational cost is high and is not suitable for real-time applications. Probabilistic color histograms [13] enhanced the robustness of color signatures and depend on the selection of the training set. Spatiotemporal appearance [19] part-based feature descriptors [4], [10], [11], [36] enable more invariant features. However, they require complex and precise pre-processing. More importantly, they are still not invariant to viewpoint changes. As we show below, always extracting feature descriptors in the same way when looking at people from different angles introduces errors that may significantly affect the descriptiveness of person signatures.

Many of these feature descriptors are high-dimensional and contain some unreliable features; hence metric model learning and feature selection are also critical modules for re-id. Many general metric machine learning approaches have been adopted in re-identification applications including Large Margin Nearest Neighbor (LMNN) [47], and Information Theoretic Metric Learning (ITML) [14]. Based on standard machine learning algorithms, much work has been done to improve the performance of these algorithms specifically for re-identification or recognition problems, such as the variants of LMNN [15], [23], Support Vector Ranking (RankSVM) [40], Logistic Discriminant Metric Learning (LDML) [21], Mahalanobis distance metric [24], local Fisher Discriminant Analysis [39], and boosting approaches [20], [22]. Also, some metric learning algorithms specifically targeting re-identification problem have been proposed, including Relative Distance Comparison (RDC) [49], dissimilarity profile analysis [30], and Pairwise Constrained Component Analysis (PCCA) [37]. These types of approaches can usually achieve better performance than traditional machine learning algorithms. There are also approaches to exploit discriminative or invariant features. Usually they are learned online, such as online feature selection [16], unsupervised salience learning [48], set-based methods [50] and covariance metric [2]. They can also be learned based on an offline learned dictionary, or using prototypes such as attribute-sensitive feature importance learning [31]. However, while these algorithms can to some ex-

tent extract discriminative and descriptive features, none of these metric learning algorithms are able to specifically extract viewpoint-invariant information from the feature descriptors. Among these metric learning algorithms, RDC [49] and RankSVM [40] are the most popular, and are extensively evaluated on most standard datasets. We selected these two methods to evaluate the performance improvement of our proposed algorithms, as well as the recent PCCA algorithm [37].

Based on the above feature model and metric learning approaches, additional techniques have also been introduced to improve the performance of re-id, including taking into account the spatial-temporal information of camera networks [19], [25], [29], [36], descriptive and discriminative classification [22], and panoramic appearance models [18].

The critical issue for our problem of interest is that these previous algorithms are not generally viewpoint invariant. Some re-id algorithms based on viewpoint invariant features [1], [27] such as SIFT [32] and SURF [7] may only work well with HD videos. Conte et al. [12] proposed the Multiview Appearance Model, which continuously updates an appearance model of the target for each quantized orientation. Signatures of candidates are only matched with the signatures of targets with the closest orientation. However, this algorithm is based on the assumption that many views of the target have been captured before starting re-id.

3 Sub-Image Rectification

A key issue for human re-identification in real-world camera networks is that in each viewpoint, the image of the target will be perspectively distorted, which may seriously affect algorithm performance. Hence, we first address the problem of eliminating perspective distortions in the region of the image containing the target, which we call rectification [28]. We assume the camera to be calibrated, which is straightforward in surveillance scenarios given the presence of parallel lines and tracked objects [33]. Our algorithm can also obtain necessary calibration information with rough hand-labeling or by observing target motion on the ground plane.

3.1 Rectification

We assume the XY plane of the world coordinate system coincides with the ground plane, so that for any 3D point on the floor, we have

$$\lambda \mathbf{z} = \mathbf{K}[\ \mathbf{r_1}\quad \mathbf{r_2}\quad \mathbf{t}\ ]\tilde{\mathbf{Z}} = \mathbf{H}\tilde{\mathbf{Z}}$$

in which $\mathbf{z}$ is the homogeneous coordinate of a point (u, v) on the image plane, $\tilde{\mathbf{Z}} = [\ x\quad y\quad 1\ ]^\top$, λ is a scalar, $\mathbf{K}$ is the camera intrinsic parameter matrix, $\mathbf{r_1}$ and $\mathbf{r_2}$ are the first and second columns of rotation matrix $\mathbf{R}$, $\mathbf{t}$ is the translation vector and $\mathbf{H}$ is the homography matrix between the image plane and the XY plane (i.e., the floor). Once a person is detected, the image position of his/her foot on the floor can be obtained as $\mathbf{z_f}$ roughly from the bounding box , and we can recover the 3D position where the person is standing on the floor as $\tilde{\mathbf{Z}}_\mathbf{f} \sim \mathbf{H}^{-1}\mathbf{z_f}$, in which $\tilde{\mathbf{Z}}_\mathbf{f} = [\ x_f\quad y_f\quad 1\ ]^\top$. With the image coordinates of the top of the person's head, $\mathbf{z_h}$, we can recover the 3D coordinates $\tilde{\mathbf{Z}}_\mathbf{h} = [\ x_f\quad y_f\quad h\ ]^\top$ of the top-of-head with

$$h\mathbf{m_3} = \mathbf{z_h} - [\ \mathbf{m_1}\quad \mathbf{m_2}\quad \mathbf{m_4}\ ]^\top \tilde{\mathbf{Z}}_\mathbf{h} \qquad (1)$$

in which $\mathbf{m_i}$ is the i^{th} column of $\mathbf{M} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$. From this information we can infer a 3D cylinder orthogonal to the ground plane with height h and radius w, which can be either a fixed value or estimated by the width of the tracking bounding box.

We now assume a person has been detected in an image within the polygon $\overline{\mathbf{n_a n_b n_c n_d}}$, with foot position $\mathbf{z_f}$ and head position $\mathbf{z_h}$. We want to obtain a new image corresponding to a head-on view of a corresponding rectangle in 3D given by $\overline{\mathbf{N_a N_b N_c N_d}}$ orthogonal to the ground plane.

Let $\mathbf{C}$ be the 3D location of the camera and $\mathbf{Z_c}$ be the projection of $\mathbf{C}$ on the XY plane. Now we want to get a rectified sub-image of the detected target. Let the projection of $\overline{\mathbf{C Z_f}}$ onto the ground plane be $\overline{\mathbf{Z_c Z_f}}$. The rectified sub-image $\overline{\mathbf{N_a N_b N_c N_d}}$ is parallel to the Z axis and orthogonal to $\overline{\mathbf{Z_c Z_f}}$. The desired 3D points in Figure 3 can then be determined as

$$\begin{aligned} N_a &= (x_f + w\cos\phi, y_f + w\sin\phi, 0) \\ N_b &= (x_f - w\cos\phi, y_f - w\sin\phi, 0) \\ N_c &= (x_f + w\cos\phi, y_f + w\sin\phi, h) \\ N_d &= (x_f - w\cos\phi, y_f - w\sin\phi, h) \end{aligned}$$

in which

$$\phi = \arctan\left(\frac{x_c - x_f}{y_c - y_f}\right).$$

The homography between the two planes defined by $\overline{\mathbf{n_a n_b n_c n_d}}$ and the projection of $\overline{\mathbf{N_a N_b N_c N_d}}$ on the image plane can then be computed and used to create a new rectified image in which the person appears to be vertical, as illustrated in Figure 3.

3.2 Viewpoint Estimation

As the person moves between two points on the floor, we obtain two 3D positions $\tilde{\mathbf{Z}}_{\mathbf{f_1}}$ and $\tilde{\mathbf{Z}}_{\mathbf{f_2}}$. The viewpoint angle of the person with respect to the camera can be estimated by:

$$\theta = \arccos\left(\frac{(\tilde{\mathbf{Z}}_{\mathbf{f_2}} - \tilde{\mathbf{Z}}_{\mathbf{f_1}})^\top(\mathbf{Z_c} - \tilde{\mathbf{Z}}_{\mathbf{f_1}})}{\|\tilde{\mathbf{Z}}_{\mathbf{f_2}} - \tilde{\mathbf{Z}}_{\mathbf{f_1}}\|\|\mathbf{Z_c} - \tilde{\mathbf{Z}}_{\mathbf{f_1}}\|}\right) \qquad (2)$$

This viewpoint angle is required for the pose prior approach in the next section.

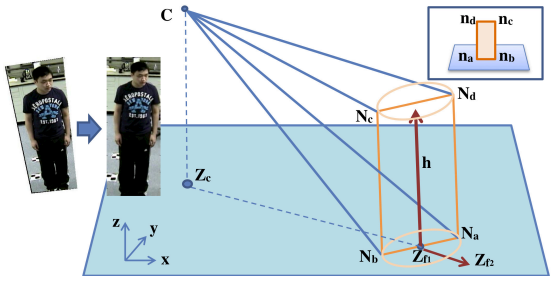


Fig. 3. Illustration of sub-image rectification.

Fig. 4. Example image pairs that may cause serious matching errors if the same descriptor is used without regard to pose. The yellow rectangles show the most difficult parts.

4 The Pose Prior

As in other re-identification methods, we divide the image of a human into horizontal strips or sub-regions. In each strip, histograms of color and texture information are extracted that form a feature vector (discussed in Section 4.1).

Let $\mathbf{X}_a$ and $\mathbf{X}_b$ be two descriptors extracted from the images of targets A and B, and $\mathbf{W}$ be a classifier trained to distinguish between A and B (discussed further in Section 4.3). A distance function between the two descriptors can thus be computed as:

$$f(\mathbf{X}_a, \mathbf{X}_b) = \mathbf{W}^\top |\mathbf{X}_a - \mathbf{X}_b|$$

While this approach is reasonable with images taken from the same point of view, when matching a pair of images of the same person from different viewpoints (e.g., $0°$ and $90°$), the descriptor distance is likely to be large. Several examples of this problem are shown in Figure 4.

Obviously, different viewpoints need to be considered differently, which motivates our proposed algorithm. We define a "pose prior", which is used to make the descriptor distance invariant to viewpoint changes, as represented by a new distance function:

$$f(\mathbf{I}_a, \mathbf{I}_b) = \mathbf{W}^\top |X(\mathbf{I}_a, \theta_a) - X(\mathbf{I}_b, \theta_b)| \qquad (3)$$

in which $\mathbf{I}_a$ and $\mathbf{I}_b$ are the images of targets A and B. θ_a and θ_b are the estimated viewpoint angles corresponding to targets A and B. $X(\mathbf{I}, \theta)$ is the converted descriptor of $\mathbf{I}$ with respect to the pose prior for angle θ. Instead of directly extracting descriptors for each strip of the target, $X(\mathbf{I}, \theta)$ weights the contribution at each pixel of a strip based on the estimated pose of the target, as described in detail in Section 4.2.

4.1 Features and Descriptors

Our approach to feature and descriptor extraction for a given image is similar to Gray and Tao [20] and uses color and texture histograms, which are widely used as features for human re-identification [40], [49], [50]. We first divide each image of a subject into six horizontal strips after rectifying the image as described in the previous section, so that the horizontal strips are parallel to the ground plane.

Since we want to re-identify human targets across different cameras, the color features used to represent the target should be invariant to lighting changes. Since RGB histograms are a poor choice [45], we use three histograms over the transformed RGB space defined as $(R', G', B') = \left(\frac{R-\mu_R}{\sigma_R}, \frac{G-\mu_G}{\sigma_G}, \frac{B-\mu_B}{\sigma_B}\right)$, where $\mu_{\{R,G,B\}}$ and $\sigma_{\{R,G,B\}}$ are the mean and standard deviation of each color channel respectively. This transformation is only applied to the person rectangles.

HSV histograms are also known to be descriptive for re-id and stable to lighting changes; thus we also include one histogram of the hue multiplied by the saturation at each pixel (since hue becomes unstable near the gray axis and its certainty is inversely proportional to saturation).

To represent texture, we compute the histograms of responses at each pixel of a strip to 13 Schmid filters and 8 Gabor filters [20]. In total 25 types of features (4 color and 21 texture) are computed at each strip pixel. We then compute the histogram over each feature in the strip using 16 bins, resulting in a 25·16 = 400-dimensional descriptor for each strip.

4.2 Learning the Pose Prior

We estimated the pose prior mappings required for (3) offline using laboratory training data (independent of the actual testing scenario). That is, for the i^{th} training subject, we acquired n_i images taken from different angles θ_j^i. These are used to estimate the relationship between the front view image of a subject and their appearance from angle $\theta \in [-180°, 180°]$. That is, we learn how to map the appearance of each strip of a target at angle θ to be directly comparable to the corresponding strip from the front view.

In our experiments, we trained the pose prior based on 20 subjects and 6–10 images of each subject (one of which is the front view at $\theta = 0$). All the images are normalized to 128×64 pixels. We estimate the angle of each image automatically with the method introduced in Section 3. Figure 5 illustrates example images used in learning the pose prior.

We first extract the descriptors for the six horizontal strips of the front view image, denoted $\mathbf{X}_F = \{\mathbf{x}_F^1, \ldots, \mathbf{x}_F^6\}$. Now we consider a sample image $\mathbf{I}_\theta$ at angle θ, and denote its strips $\{\mathbf{I}_\theta^1, \ldots, \mathbf{I}_\theta^6\}$. We let $\mathbf{I}_\theta^k(u)$ be a patch of pixels in strip k centered around column u (in our experiments we chose the patches to be 10 pixels wide). Note that the patches overlap,

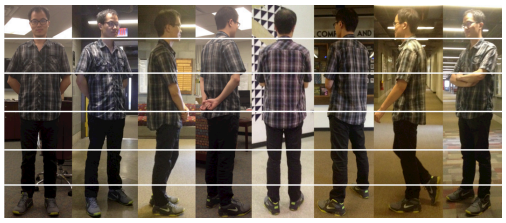


Fig. 5. Example images used in learning the pose prior.

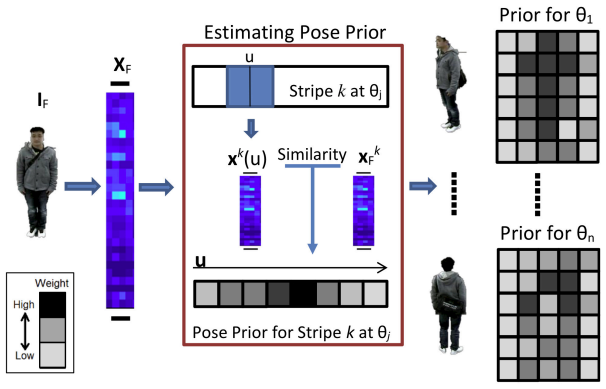


Fig. 6. Learning the pose prior.

and we only consider values of u where a full-sized patch exists (i.e., not along the left and right edges of the image). We denote $\mathbf{x}_\theta^k(u)$ as the descriptor computed within the patch centered around column u in strip k at pose angle θ. Please note that no background subtraction has been done in any step of the algorithm.

We now form a function of u that relates how similar the descriptors inside each patch are to the entire descriptor of the front view at this strip:

$$p_\theta^k(u) = \mathcal{S}(\mathbf{x}_F^k, \mathbf{x}_\theta^k(u)) \tag{4}$$

where $\mathcal{S}$ is a similarity measure. Any histogram similarity measure can be used. In this paper we used the covariance between the two descriptor vectors, normalized to the range [0,1]. The intuition is that we want to measure the similarity in shape of the two descriptors. If the pose angle θ is near 0 (that is, the image is close to a frontal view), then we expect $p_\theta^k(u)$ to be high for most values of u, while if θ is near 90° for example, we expect $p_\theta^k(u)$ to be high for u towards the right side of the image (i.e., corresponding to pixels on the front of the person) and low elsewhere. Figure 6 illustrates the idea.

Figure 7 illustrates that as we compute these functions of u for each of several angles $\{\theta_j^i, j = 1, \ldots, n_i\}$ for subject i, we can project them onto a circle to create the pose prior $\mathbf{P}_\theta$. That is, we wrap the collected similarity measures around a cylinder (the position and radius of which are computed from the apparent size of the person in the image, the fact that they are

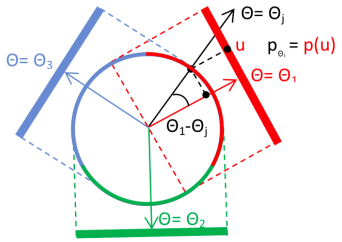


Fig. 7. Complete pose prior generated by projections of pose priors calculated from multiple viewpoints.

orthogonal to the ground plane, and the calibrated camera).

We obtain a different pose prior for each of the N training subjects; we merge them at each angle θ independently by taking the mode (i.e, the center value of the bin with highest frequency) of the distribution formed by the samples $\{\hat{P}_\theta^1, \ldots, \hat{P}_\theta^N\}$. When multiple bins have the same frequency, we simply average the modes if they are close to each other. In the rare case that the differences between modes are noticeable, which suggests that the robustness of the features to changes in viewpoint angle is low at this location, we choose the smallest one to be safest. The overall algorithm of learning the pose prior from training data is shown in Algorithm 1.

Algorithm 1: Learning the pose prior

Input: N: number of people used for training
$\{n_1, \ldots, n_N\}$: the number of images for each person
$\mathbf{I}_{F^i}$:Image of front view, $i = 1...N$
$\mathbf{I}_{\theta_j^i}$, $j = 1, \ldots, n_i$: Image taken from viewpoint θ_j for person i
Output: Pose prior $\mathbf{P}_\theta$
for $i = 1$ **to** N **do**
  Estimate the descriptors X_{F^i} for $\mathbf{I}_{F^i}$;
  for $j = 1$ **to** n_i **do**
    Calculate $p_{\theta_j^i}$ with (4);
  end
  for $\theta = -180$ **to** 180 **do**
    Find $\hat{P}_\theta^i$ as $p_{\theta_j^i}$ where $\theta_j^i = \arg\min |\theta - \theta_j^i|$;
  end
end
for $\theta = -180$ **to** 180 **do**
  Find $\mathbf{P}_\theta$ by taking the mode of the distribution formed by $\{\hat{P}_\theta^1, \ldots, \hat{P}_\theta^N\}$;
end

Figure 8 shows examples of the trained pose prior at 45, 90, 135 and 180 degrees.

4.3 Offline Training

After we learn the pose prior, we train a classifier offline to form a general understanding of what features are most discriminative for re-identification. As described above, when extracting the descriptor for a given image at an estimated angle, the pose prior $\mathbf{P}_\theta$ is used to weight the contributions of the columns of each strip when calculating the histograms for color and texture features.

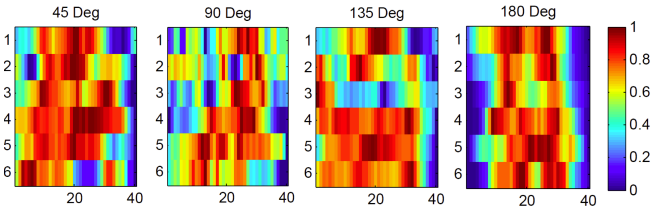


Fig. 8. Examples of the trained pose prior. It can be seen that the results match our assumptions. For example, at 90 degrees, the weights on the right side are much heavier than on the left side. At 135 and 180 degrees, the weights in the central area (i.e., the back of the person) are significantly lower than in the neighboring areas. We found this phenomenon to be generally caused by people carrying backpacks.

That is, the final descriptor for an image at estimated angle θ is $\mathbf{X}_\theta = \{\mathbf{x}^1, \ldots, \mathbf{x}^6\}$, where

$$\mathbf{x}^k(i) = \sum_{(u,v)\in \text{strip } k} \chi_{(u,v)}(i)\mathbf{P}^k_\theta(u) \tag{5}$$

where

$$\chi_{(u,v)}(i) = \begin{cases} 1 & \text{feature value at } (u,v) \text{ is in bin } i \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{P}^k_\theta(u)$ is the trained pose prior for stripe k at column u. The contribution of each feature is weighted by the horizontal position u along the strip as a function of the pose prior. The absence of a pose prior would correspond to letting $\mathbf{P}^k_\theta(u) = 1$ for all u.

The classifier training follows the standard approach [40]: that is, we want to maximize the norm of a weight vector $\mathbf{W}$ subject to the constraints that if $\mathbf{d}_{\text{same}} = |\mathbf{x}^{\mathbf{i}}_{\mathbf{j}} - \mathbf{x}^{\mathbf{i}}_{\mathbf{k}}|$ is the absolute difference of descriptors for two images of the same person i, and $\mathbf{d}_{\text{diff}} = |\mathbf{x}^{\mathbf{i}}_{\mathbf{j}} - \mathbf{x}^{\mathbf{l}}_{\mathbf{k}}|$ is the absolute difference of descriptors for images of two different people i and l, then

$$\mathbf{W}^\top \mathbf{d}_{\text{same}} < \mathbf{W}^\top \mathbf{d}_{\text{diff}}$$

for all possible pairs of same and different images. We use the absolute difference of descriptors, since it has been shown to yield more consistent distance comparison results [49]. Figure 9 shows an example comparison between descriptors of the same person from different viewpoints with and without the pose prior. We can see that the differences between some of the features have been reduced, making the descriptors more similar.

5 Online Learning and Matching

In this section, we introduce algorithms for online descriptor extraction and discriminative feature learning. The online algorithm leverages the pose prior $\mathbf{P}_\theta$ and classifier $\mathbf{W}$ learned offline using training data in the previous section.

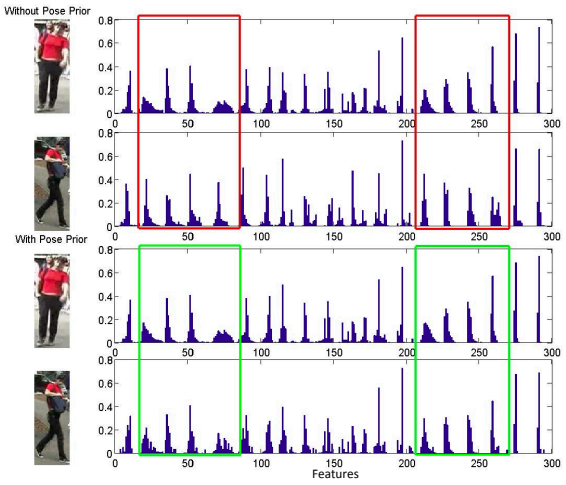


Fig. 9. The distance between descriptors of the same person from different viewpoints is reduced by applying the pose prior. Without the pose prior, the distance between the two descriptors is 1.4709, which has been reduced to 0.9738 after applying the pose prior. Without pose prior, the mean distance of all negative pairs containing this person is 5.2815, which increased to 5.6369 with pose prior.

5.1 Extracting Descriptors

When we identify (or "tag") a target for re-identification online, we must first extract its descriptors $\mathbf{X}_F$. However, at the time when the target is tagged, or even during the whole time the target was tracked in the current view, we may not be able to exactly observe the front (0°) view of the target. By using the pose prior, we can estimate these descriptors. With the camera calibrated to the floor plane, we can estimate the viewpoint angle θ. For a target image taken from viewpoint θ_j, we can estimate $\mathbf{X}_F$ by weighting the feature histograms by the pose prior as specified by (5).

5.2 Learning Subject-Discriminative Features

The classifier model we trained offline is a general model, which is universal for every human. However, learning discriminative features for the particular target being tracked may greatly boost the performance of the re-identification. That is, we update the classifier function to

$$f(\mathbf{d}) = \mathbf{W}^\top \mathbf{d} + \alpha \mathbf{s}^\top \mathbf{d} \tag{6}$$

where α is a weighting factor and $\mathbf{s}$ is a person-specific weight on the descriptor. We first model the distribution of each feature with a Gaussian, based on offline training data; denote the mean and variance of feature i in the descriptor $\mathbf{X}$ as $\hat{\mu}(i)$ and $\hat{\sigma}(i)$:

$$\hat{\mu}(i) = \frac{\sum_{j=1}^{N}(1+\cos\frac{\theta_j}{2})\mathbf{X}_j}{(\sum_{j=1}^{N}\cos\frac{\theta_j}{2})+N} \tag{7}$$

$$\hat{\sigma}^2(i) = \frac{\sum_{j=1}^{N}(1+\cos\frac{\theta_j}{2})(\mathbf{X}_j-\hat{\mu}(i))^2}{(\sum_{j=1}^{N}\cos\frac{\theta_j}{2})+N} \tag{8}$$

in which N is the total number of training samples for all subjects in the training set, and $\mathbf{X}_j$ and θ_j are the descriptor and viewpoint respectively of the j^{th} training sample. Here, $\cos\frac{\theta_j}{2}$ is a weighting term for images captured from different viewpoint angles θ_j, so that the front view $\theta_j = 0$ has unit weight and the back view $\theta_j = 180$ has zero weight. During online processing, once the target is tagged and descriptors $\mathbf{X}_F$ are extracted, we find the features i that have low likelihood with respect to the offline Gaussian distribution. That is, we find the person-specific features that we would not expect based on the training data; these are particularly discriminative for that person. We explicitly determine $\mathbf{s}$ by first computing:

$$\tilde{\mathbf{s}}(i) = \begin{cases} 1 - G(\mathbf{X}_F(i), \hat{\mu}(i), \hat{\sigma}(i)) & G(\cdot) < \tau \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

in which the Gaussian function $G(\cdot)$ is normalized to have height 1. For each dimension of a particular descriptor, this approach calculates its "distinctiveness" based on the Gaussian model learned from training data. The lower the likelihood based on the training model $G(\cdot)$, the more discriminative the feature is. In our experiments, we used $\tau = 0.1$. After processing all the features in $\mathbf{X}_F$, we compute

$$\mathbf{s} = \mathbf{W} \circ \frac{\tilde{\mathbf{s}}}{\|\tilde{\mathbf{s}}\|}$$

where $\circ$ is the element product. Finally, we learn α from the training data using

$$\alpha = \frac{1}{|\mathbb{O}_s|\ |\mathbb{O}_d|} \sum_{\substack{\mathbf{d}_{\text{same}} \in \mathbb{O}_s \\ \mathbf{d}_{\text{diff}} \in \mathbb{O}_d}} \delta\left(\mathbf{s}^\top \mathbf{d}_{\text{same}} < \mathbf{s}^\top \mathbf{d}_{\text{diff}}\right) \quad (10)$$

where $\mathbb{O}_s$ is the set of descriptor differences between all pairs of the same person from the training set. $\mathbb{O}_d$ is the set of descriptor differences between all pairs of different people, where one of them is the target being tracked. α can be viewed as the confidence in the trained discriminative coefficient vector $\mathbf{s}$. The better that $\mathbf{s}$ is able to distinguish image pairs from different people, the more confident we are, and the higher the weight α. Figure 10 illustrates the process of offline training and online learning of discriminative features. Note that the discriminative features can be learned with a single image. The overall process of learning the classifier for the target is shown in Figure 11.

5.3 Matching and Identification

After extracting the descriptors $\mathbf{X}_F^t$ and the discriminative vector $\mathbf{s}$ of the target, the target may leave the current view and enter another view. Now the task becomes matching candidates in the view with the target model.

First, by tracking the candidates, their viewpoints are estimated using the approach at the end of Section 3. Using the pose of each candidate, we can decide how the target model should be used to match the candidates' descriptors. For a candidate with viewpoint angle θ, we have $f(\mathbf{d}) = (\mathbf{W} + \alpha\mathbf{s})^\top \mathbf{d}$ with $\mathbf{d} = |\mathbf{X}_F^t - \mathbf{X}_F^c|$, where $\mathbf{X}_F^c$ is the descriptor of the candidate, normalized to the front view using the pose prior (5). The candidate with the highest matching result $f(\mathbf{d})$ is considered as the detection of the target.

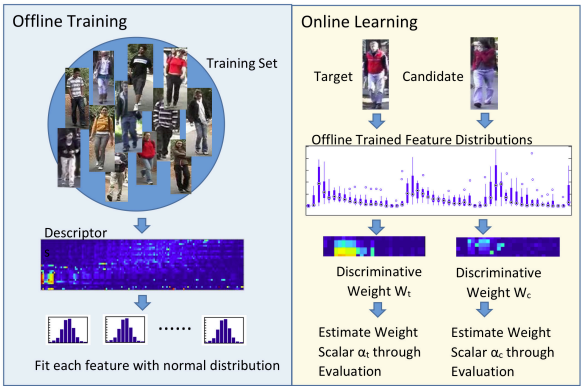


Fig. 10. Illustration of learning discriminative features.

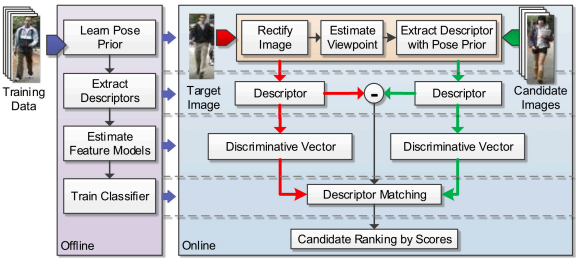


Fig. 11. Flowchart of learning and matching processes.

6 Experimental Results

In the following experiments, we applied the standard protocol for the re-id problem. That is, each dataset is randomly sampled so a certain number of person images are taken as the training set, and the remaining images form the testing set. The testing set is further divided into a probe set and a gallery set. The gallery set consists of one image from each person in the testing set, while all other images are in the probe set. The detailed training/testing number split will be specified for each dataset, following the standard splits from [49].

We use cumulative match characteristic (CMC) curves to report the experimental performance, specifically, the matching rate at rank n, where $n \in \mathbb{N}$. The rank n matching rate specifies the percentage of probe images that matched correctly with one of the top n images in the gallery set.

We use RDC [49], RankSVM [40] and PCCA [37] as our baseline comparison algorithms. These state-of-the-art algorithms have been shown to have high

performance, and they are metric learning techniques, which are suitable for demonstrating the improvement obtained through our algorithm. We applied the publicly available source code of Zheng et al. [49] for the RDC algorithm. We implemented the RankSVM algorithm using the library published by Joachims [26], which produces similar results as [40]. We also implemented the PCCA [37] algorithm with the χ^2 kernel, using which we obtained the best and most similar results compared with the original paper. Some numerical results differ slightly from those stated in the original papers.

We extensively tested our algorithms on the standard VIPeR [20], ETHZ [41], and i-LIDS MCTS [38] datasets. These three datasets were validated with similar configurations as in [49]. Note that since images from these datasets have only a small amount of perspective distortion, they were processed without rectification. We also use the recently proposed 3DPeS [6] and SAIVT-SoftBio [9] datasets to evaluate the performance gains for two key subcomponents of the algorithm, as well as a challenging new dataset collected at a US airport to evaluate our performance in a real-world scenario.

6.1 Evaluating Key Subcomponents

The 3DPeS [6] dataset is built from a multi-camera surveillance network with non-overlapping fields of view on a university campus. While this dataset is not yet widely used as a re-id benchmark, due to the accompanying side information it is well-suited to isolating the improvement brought by the rectification and pose prior steps of the proposed algorithm.

3DPeS contains two sets of images extracted from 8 different camera views. The first set contains 199 individuals with a total of 606 images and includes calibration for each camera. We refer to this set of images as 3DPeS-1, and use it to evaluate the image rectification method. The second set contains 193 individuals with a total of 1012 images and includes corresponding foreground masks for each subject. We refer to this set of images as 3DPeS-2, and use it to evaluate the pose prior method.

To evaluate the benefit of sub-image rectification, we randomly chose 99 people in the 3DPeS-1 dataset as the training set, and used the remaining 100 people as the testing set. We used RDC [49] as the baseline metric learning algorithm. Figure 12 illustrates the re-identification results with and without sub-image rectification. We observed a 2–5% increase in matching rate using the rectified images, indicating the value of the approach.

We next compared the pose prior approach to a simple method using the foreground mask of the subject when it is available. That is, the descriptors are only computed over the foreground pixels, not the entire rectangle. To evaluate the benefit of the pose prior, we

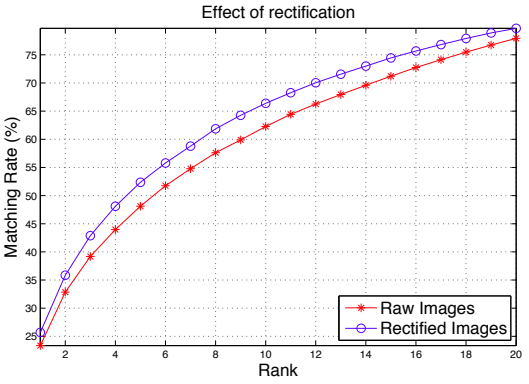


Fig. 12. The effect of sub-image rectification effect on 3DPeS-1, split 99/100.

randomly chose 95 people in the 3DPeS-2 dataset as the training set, and used the remaining 98 people as the testing set. We again used RDC [49] as the baseline metric learning algorithm. Figure 13 illustrates the re-identification results with and without the pose prior, and with and without the foreground mask. We can see that the pose prior visibly improves the matching rate when the foreground mask is not available. When the foreground mask is available, the matching rate is much better than using the pose prior alone, but can be further improved by applying the pose prior to the foreground pixels. This experiment suggests that foreground information should definitely be used when it is available; however, this information is not provided for the standard re-id benchmarking datasets discussed next.

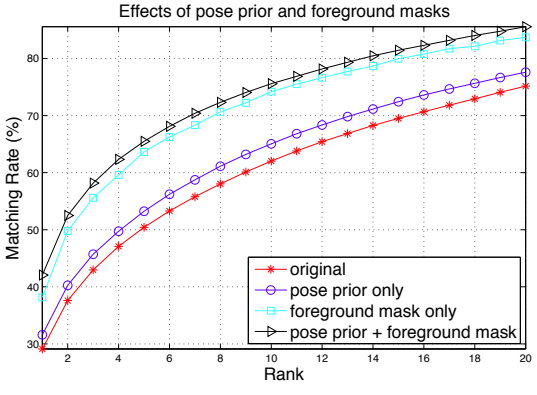


Fig. 13. The effects of the pose prior and the foreground mask on 3DPeS-2, split 95/98.

6.2 Performance on Benchmarking Datasets

We now demonstrate the matching rate improvement of the proposed method compared to the baseline algorithms, RDC [49], RankSVM [40] and PCCA [37], illustrated in Figure 14, Figure 15 and Figure 16 respectively. Each CMC chart compares our pose prior algorithm (baseline+PP), subject-discriminative

feature selection (baseline+DF) and their combination (baseline+PP+DF) with the original algorithm. Detailed numerical matching percentages are listed in Table 1. Performance comparisons with additional state-of-the-art algorithms are given in Figure 17 and Table 3. We repeated all the experiments 10 times to report averaged results. All the images are normalized to 128×48 pixels in the VIPeR dataset, and 128×64 pixels in the ETHZ and i-LIDS datasets. We discuss the results in more detail below.

6.2.1 VIPeR [20]

This dataset contains images of 632 people captured by 2 cameras from different viewpoints. The major challenges of this dataset are viewpoint change and illumination variation. The VIPeR images are labeled with viewpoint information that we used to train the pose prior. (In all other experiments, we use the pose prior trained with lab-based data as described in Section 4.2.)

In the experiment, we first randomly picked 316 people for the training set, and the other 316 people as the testing set. We also used the split of 100 people for training set, and 532 people for the testing set. The CMC curves of the three different baseline algorithms in the latter case are shown in Figure 14a, 15a, and 16a respectively. It can be seen that both the discriminative features and the pose prior improve the performance. The pose prior improves the matching rate by 3–5% in all cases, since in this dataset the image pair for each person contains significantly differing viewpoints. The VIPeR images are low resolution, so the discriminative features do not appear to be very distinctive for each person. However, for each split, the combined techniques improved the overall results. When the testing gallery size is 316, rank 1 of RDC shifts from 15.66% to 19.43%, rank 1 of RankSVM shifts from 15.78% to 21.35% and rank 1 of PCCA shifts from 16.07% to 21.31%. When the testing gallery size is 532, rank 1 of RDC shifts from 9.62% to 12.12%, rank 1 of RankSVM shifts from 9.06% to 12.91% and rank 1 of PCCA shifts from 8.28% to 11.25%.

6.2.2 ETHZ

This dataset consists of three video sequences captured by moving cameras in a street scene. Schwartz et al. [41] extracted an image set of pedestrians to perform appearance-based model learning, which essentially converts it into a re-identification dataset. The dataset includes 8555 images from 146 people. Since the extracted images are captured from moving cameras while people are walking, they contain illumination variation and serious occlusion. The cropped bounding box size also varies slowly as the person walks from one point to another. We roughly ground-truthed images into 5 viewpoint angles: 0, 45, 90, 135 and 180 degrees.

Instead of evaluating the three subsets of the ETHZ dataset individually, we combined them into one dataset as in [49], since ETHZ 2 and ETHZ 3 have only 20-30 people each. Adopting the configuration used in [49], we conducted the experiments with testing gallery sizes of 70 and 120. As in [49], we randomly picked 6 images from each person in the training set. The CMC curves of the three different baseline algorithms in the gallery-120 case are shown in Figure 14b, 15b, and 16b respectively. Again, the pose prior contributes the major improvement. Due to occlusion and the inconsistent size of the person in the image, the discriminative feature barely has any improvement compared with the baseline algorithm.

6.2.3 i-LIDS

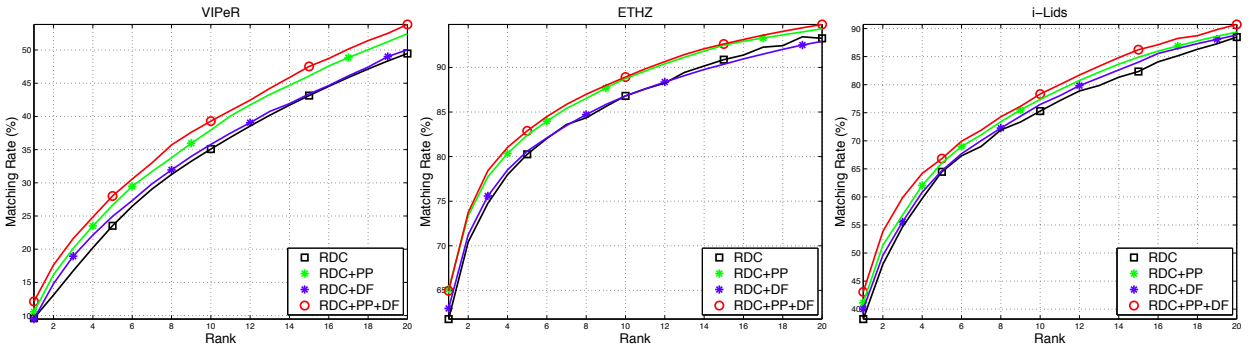
The i-LIDS MCTS [38] dataset is a real-world scenario captured in a busy airport arrival hall by multiple non-overlapping cameras. Zheng et al. [51] extracted a re-identification dataset containing 476 images of 119 people, with an average of around four images per person. The images undergo large viewpoint variation, severe occlusion and some illumination changes. We roughly ground-truthed the images into the same 5 viewpoint angles as above.

In our experiments, we took the gallery sizes of 50 and 80, as in [49]. The CMC curves of the three different baseline algorithms in the gallery-50 case are shown in Figure 14c, 15c, and 16c respectively. Again, the combined performance steadily improves the results.

6.2.4 SAIVT-SoftBio

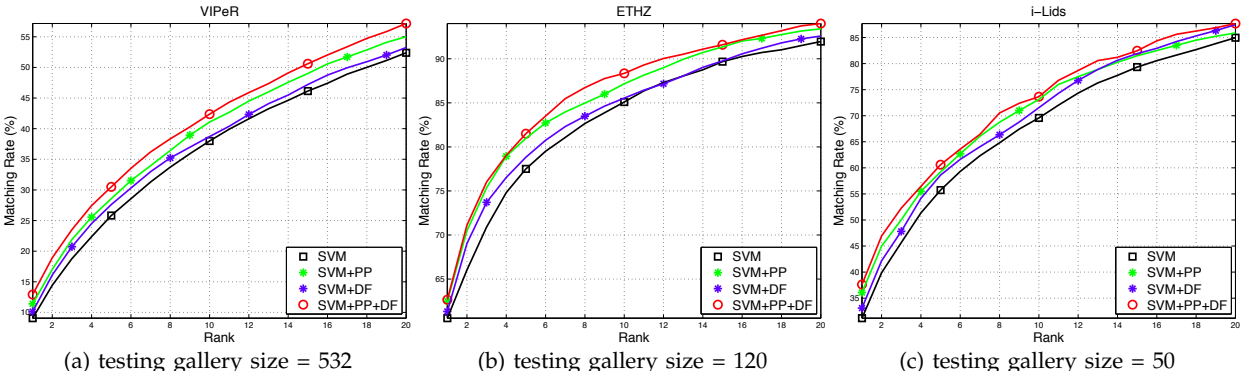
Bialkowski et al. [9] presented a new well-structured dataset designed for person re-identification. The images are extracted from multi-camera surveillance videos in a a real-world campus environment. This dataset easily enables performance evaluation as a function of different factors such as pose, viewpoint, and lighting condition. The dataset contains 152 individuals, each of whom is captured by up to eight cameras. Each individual is tracked from when they enter a building until they leave the view of the surveillance network. The dataset includes 64472 frames in total. The frames with occlusion are omitted.

For our experiment on this dataset, we chose a training set of 40 people and a testing gallery of 112 people. Since the dataset contains a massive number of images, for each person in the training set we randomly chose 6 images to perform metric learning. The gallery set consists of one image from each person in the test set; however, instead of using all the remaining images, 50 images were randomly selected from each person to form the probe set. To be consistent with the other experiments, we learned one metric from the entire training set. The pose of each sub-image is precisely estimated using the proposed algorithm in Section 3.2. The experimental results are



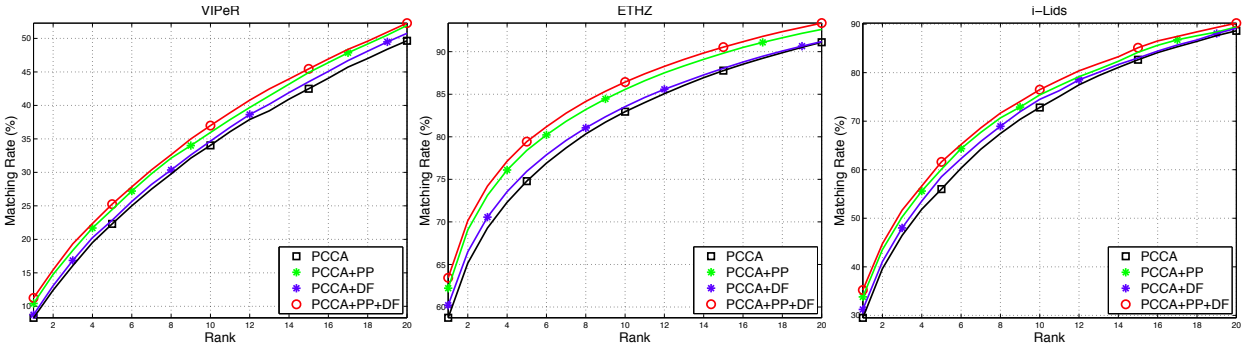
(a) testing gallery size = 532 (b) testing gallery size = 120 (c) testing gallery size = 50

Fig. 14. Ranking performance CMC curves on public datasets using RDC [49] based algorithms.



(a) testing gallery size = 532 (b) testing gallery size = 120 (c) testing gallery size = 50

Fig. 15. Ranking performance CMC curves on public datasets using RankSVM [40] based algorithms.



(a) testing gallery size = 532 (b) testing gallery size = 120 (c) testing gallery size = 50

Fig. 16. Ranking performance CMC curves on public datasets using PCCA [37] based algorithms.

shown in Table 2. Since camera calibration results are available, we roughly estimated the ground plane and applied rectification to all person rectangles. We also computed the performance of metric learning on the original images to evaluate the gains brought by rectification.

6.3 Discussion

In general, we can see that both the pose prior and the discriminative features improve the performance of re-identification. Combining the two methods, the proposed algorithm obtains the best overall results, with the pose prior having the main contribution. The discriminative features will be less stable when the training set is small.

To further compare the performance of our algorithm with state-of-the-art algorithms, numerical ranking results from a larger selection of algorithms are listed in Table 3. Please note that not all algorithms provide results for each dataset or each split. The most widely used benchmark in the community is VIPeR with the testing gallery size of 316, which can give a universal comparison with all algorithms. The CMC curves are displayed in Figure 17. Our algorithm significantly improves the baseline metric learning algorithms and gives results potentially bet-

TABLE 1
Result comparisons between the proposed algorithms and competitive algorithms. Each result shows the percentage of the time the correct match occurred in the top k results for a given classifier, where k is the rank.

	VIPeR								ETHZ								i-LIDS							
Train/Test	316/316				100/512				76/70				26/120				69/50				39/80			
Method	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
RDC Comb.	**19.4**	**44.1**	**58.5**	**74.2**	**12.1**	**28.0**	**39.3**	**53.9**	**73.6**	**88.9**	**93.9**	**97.8**	**64.9**	**82.9**	**88.9**	**94.8**	**43.1**	**66.8**	**78.3**	**90.7**	**35.6**	**57.3**	**69.6**	**82.5**
RDC+PP	18.5	41.8	56.6	73.2	10.6	26.6	37.9	52.4	72.1	87.7	93.4	97.7	65.1	82.4	88.7	94.3	41.1	66.0	77.4	89.4	34.7	56.8	67.2	81.7
RDC+DF	17.1	40.6	55.7	71.9	9.7	24.9	35.8	50.0	70.0	87.0	92.5	97.2	63.0	80.6	86.8	92.9	40.0	64.7	76.5	88.9	31.1	54.9	68.4	79.7
RDC [49]	15.7	37.5	53.5	69.9	9.6	23.5	35.1	49.5	69.0	85.8	92.2	96.9	61.6	79.7	86.7	93.3	37.8	63.7	75.1	88.4	32.6	54.6	65.9	78.3
SVM Comb.	**21.4**	**45.9**	**60.5**	**75.9**	**12.9**	**30.5**	**42.4**	**57.2**	**74.3**	**88.0**	**93.2**	**96.3**	**62.7**	**81.5**	**88.3**	**94.0**	**37.6**	**60.6**	**73.6**	**87.7**	**32.3**	**52.8**	**66.1**	**80.3**
SVM+PP	19.4	44.0	59.0	74.3	11.4	28.6	41.1	55.0	72.9	86.8	92.3	96.1	62.5	81.0	87.2	93.4	36.1	62.7	73.1	85.9	31.8	51.1	63.0	79.3
SVM+DF	18.7	42.2	56.5	73.4	10.1	30.3	38.7	53.2	71.2	85.9	91.4	95.4	61.3	78.8	85.5	92.6	33.1	58.6	71.5	87.5	29.5	52.0	65.6	78.7
SVM [40]	15.8	40.7	55.9	71.9	9.1	25.8	38.0	52.4	69.4	86.3	90.7	94.5	60.6	77.5	85.1	92.0	31.2	55.7	69.6	85.0	27.5	48.5	61.5	77.3
PCCA Comb.	**21.3**	**45.8**	**62.6**	**79.7**	**11.3**	**25.3**	**37.0**	**52.3**	**71.6**	**88.8**	**94.5**	**98.1**	**63.4**	**79.4**	**86.4**	**93.3**	**35.2**	**61.6**	**76.5**	**90.2**	**31.5**	**52.1**	**64.7**	**80.8**
PCCA+PP	19.8	45.4	61.7	79.5	10.4	24.4	36.0	51.9	70.3	87.9	93.9	97.8	62.2	78.4	85.7	92.6	33.8	60.1	75.4	89.4	29.8	51.0	64.6	80.5
PCCA+DF	17.3	43.6	60.6	78.3	8.73	22.8	34.7	50.8	67.7	86.8	93.3	97.6	60.2	75.9	83.5	91.2	31.2	58.5	74.5	89.1	28.9	50.7	64.1	79.8
PCCA [37]	16.1	41.8	59.8	77.3	8.28	22.3	34.0	49.6	65.9	85.6	92.6	97.8	58.7	74.8	83.0	91.1	29.5	56.0	72.8	88.6	26.7	48.6	62.4	79.2
ITM [14]	11.3	31.4	45.8	63.9	4.2	11.1	17.2	24.6	56.3	80.7	88.6	94.1	43.1	66.0	76.6	86.8	29.0	54.0	70.5	86.7	21.7	41.8	55.1	71.3
AdaBoost [20]	8.2	24.2	36.6	52.1	4.2	13.0	20.2	30.7	65.6	84.0	90.5	95.6	60.7	78.8	85.7	92.0	29.6	55.2	68.1	82.4	22.8	44.4	57.2	70.6
L^1-norm	5.8	14.3	21.3	32.7	5.1	11.1	15.6	24.1	56.4	77.1	85.1	92.1	51.2	71.2	79.1	86.7	18.8	41.7	57.9	75.9	14.8	33.4	46.0	62.7
L^2-norm	5.3	13.4	20.1	32.4	5.1	10.7	15.2	24.6	55.8	76.5	84.9	92.2	51.1	70.3	78.6	86.5	16.0	38.3	52.5	71.4	13.7	31.7	44.0	58.7

TABLE 2
Re-identification results on the SAIVT-SoftBio dataset.

Method	1	5	10	20
RDC+PP+DF	**22.3**	**33.6**	**40.2**	**54.9**
RDC+PP	20.7	31.5	38.9	52.2
RDC+DF	20.3	30.8	38.3	51.5
RDC	18.5	28.5	36.3	48.1
RDC (not rectified)	15.9	25.7	34.8	47.6
SVM+PP+DF	**23.9**	**34.3**	**40.1**	**52.8**
SVM+PP	22.1	32.5	39.1	50.4
SVM+DF	19.9	30.3	38.2	51.3
SVM	19.2	28.9	35.2	47.7
SVM (not rectified)	16.3	25.8	31.9	43.2
PCCA+PP+DF	**21.4**	38.2	**50.9**	**71.3**
PCCA+PP	20.3	**38.4**	47.7	67.2
PCCA+DF	18.9	35.2	46.9	64.5
PCCA	16.1	32.4	44.5	61.2
PCCA (not rectified)	14.3	29.8	42.1	59.9

TABLE 3
Rank matching rate comparison on VIPeR dataset, with training size = 316, testing size = 316.

Method	1	5	10	20
L^1-norm	5.8	14.3	21.3	32.7
ITM [14]	11.3	31.4	45.8	63.9
RDC [49]	15.7	38.4	53.9	69.9
SVM [40]	16.5	36.3	52.0	68.3
PCCA [37]	16.1	41.8	59.8	77.3
SDALF [17]	19.9	38.9	49.4	65.7
CPS [11]	21.8	44.0	57.2	71.0
eBiCov [34]	20.7	42.0	56.2	68.0
eLDFV [35]	**22.3**	**47.0**	60.0	71.0
L^1-norm+PP+DF	9.5	16.2	23.6	34.7
RDC+PP+DF	19.4	44.1	58.5	74.2
SVM+PP+DF	21.4	45.9	60.5	75.9
PCCA+PP+DF	21.3	45.8	**62.6**	**79.7**

ter than current state-of-the-art algorithms. Figure 17 also includes a comparison to using the raw L^1 norm between descriptors, with and without the pose prior and discriminative features, in the absence of any metric learning. We can see that metric learning is definitely required to achieve competitive re-id performance.

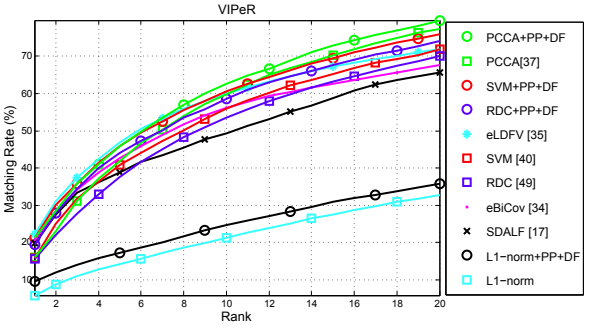


Fig. 17. Performance comparison on VIPeR dataset, with the split of 316/316.

To further illustrate the intuition behind our method, we describe an example showing the real improvement from the baseline algorithm. In Figure 18, the leftmost column is the probe image, and the rows illustrate the baseline algorithm, the result of applying discriminative features, applying the pose prior and applying both. Originally the correct person in the gallery is at rank 10; with discriminative features, it jumps to rank 4. If we look at the woman in the probe image, the discriminative features could be, for example, the clear separation between the torso and legs, the dark coat, and the pinkish trousers. By applying person-specific features, some people with simple-colored clothing, or a white pattern on their shirt are left behind. If we only add the pose prior, not only does the correct image shift to rank 2, but more side-view images also move to higher rank, because unrelated parts of the images play a less significant role in the feature vector. For example, the person at rank 4 in the first row, who got to rank 2 in the second row, doesn't even appear in the third row. The pose

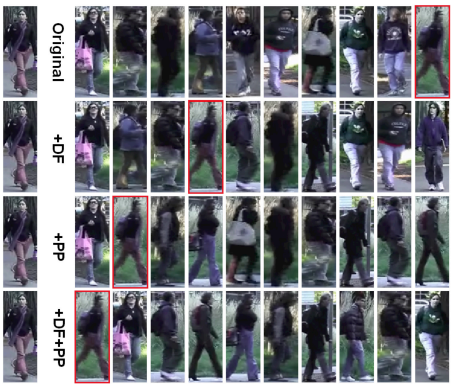


Fig. 18. Improvement brought by the proposed algorithms on one example from VIPeR.

prior can put less weighting on his/her backpack, which may be the main reason keeping him/her at top ranks in the first 2 rows.

6.4 Airport Surveillance Camera Experiment

We also tested our re-identification algorithm using video collected from a surveillance camera network at a medium-sized US airport. We analyzed three synchronized video streams from this camera network, with relative positions sketched in Figure 19. We select a target in one view, and then automatically extract descriptors of the target and detect its reappearance in the other cameras. We roughly calibrated all the cameras using the embedded function from a Bosch VIP1X encoder by manually labeling parallel lines on the floor. In this way we rectify the sub-images and obtain the pose angle of all the humans in the videos.

To demonstrate the performance of our algorithm, we collected cropped rectangles around people and built a new real-world airport surveillance scenario dataset. While the VIPeR dataset is designed to investigate viewpoint variation, it only has 2 images per person, and the background condition is also comparatively good. In our airport dataset, each person has multiple images with large viewpoint variation. Moreover, the surveillance cameras installed in the airport produce fairly low quality, low resolution video with serious illumination changes compared to other datasets. Some sample images are shown in Figure 19.

We extracted 113 people from the video recorded by the three cameras in Figure 19. The first 88 people have a total of 6 images per person across the 3 videos, while the other 25 people have between 1 and 15 images. In total, there are 625 images in this dataset. In this experiment, we applied the pose prior with two different sets of pose angle estimates to evaluate how the accuracy of the pose estimate affects re-id performance. In the first set (PP1), as we do for the standard datasets, the estimated pose of each image is categorized into one of five view angles: 0, 45, 90, 135 and 180 degrees. In the second set (PP2), unquantized, more accurate pose estimates are used. This dataset will be available for download from our website.

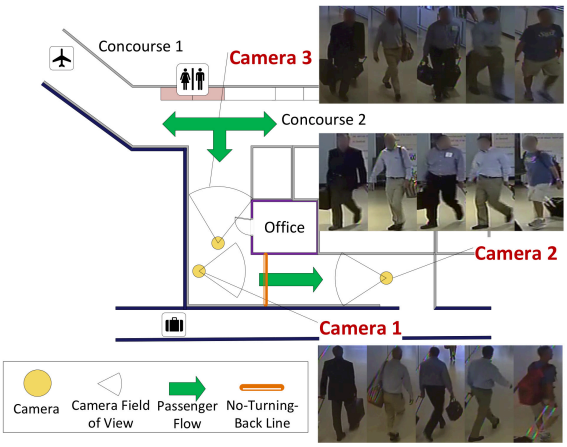


Fig. 19. Floor plan of the new airport camera network used for human re-identification, and sample images in three camera views, with large viewpoint and illumination variation.

We conducted experiments on this dataset following a similar protocol to the previous experiments. We chose 33 people out of the first 88 (as they appeared in all views) to form the training set. The testing gallery set consists of one image from each of the remaining people, which gives 80 images in total. The remaining images of the first 88 people are used as probe set. Again, we applied RDC, RankSVM and PCCA (both unrectified and rectified) as the baseline algorithms, and then applied the proposed improvements. The results are shown in Table 4 and CMC curves in Figure 20.

We can see that our algorithm boosts the baseline algorithms' performance significantly, especially at lower ranks. For example, rank 1 of RDC is improved from 11.16% to 21.85%, rank 1 of SVM is improved from 11.01% to 21.75% and rank 1 of PCCA is improved from 13.23% to 24.08%. The pose prior still makes the main contribution. The images with large viewpoint variation easily benefit from our algorithms. With the person-specific feature selection, the effects of illumination change and noise in the images are also suppressed. We can see that with more accurate pose estimates (PP2), slightly better re-id performance can be obtained. However, the algorithm does not highly depend on accuracy of the pose estimates.

7 Conclusion

We proposed a new pose prior technique that can effectively leverage the correlation between images from different viewpoints, significantly improving re-identification performance. Moreover, with the discriminative feature approach, the distinctiveness of a person can be better highlighted. Experimental results

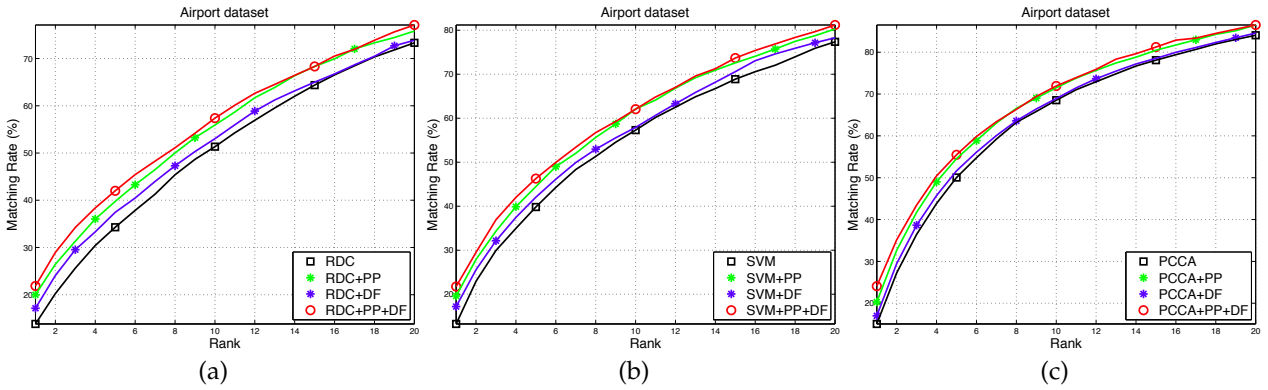


Fig. 20. Ranking performance CMC on airport dataset, with training size = 33, testing gallery size = 80. (a) Based on RDC (b) Based on RankSVM (c) Based on PCCA. The results shown here are based on raw pose estimates (corresponding to PP2 in Table 4).

TABLE 4
Re-identification results on the airport dataset.

Method	1	5	10	20
RDC+PP2+DF	**21.85**	**42.00**	**57.37**	**77.11**
RDC+PP1+DF	20.72	41.04	56.31	76.08
RDC+PP2	20.00	39.75	55.91	75.81
RDC+PP1	18.94	38.60	54.71	74.89
RDC+DF	17.14	37.44	53.04	73.86
RDC	13.82	34.28	51.34	73.32
RDC (not rectified)	11.16	32.08	48.92	70.17
SVM+PP2+DF	**21.75**	**46.28**	**62.05**	**81.19**
SVM+PP1+DF	20.55	45.07	61.10	80.08
SVM+PP2	19.60	44.47	62.11	80.27
SVM+PP1	18.38	43.37	60.87	79.21
SVM+DF	17.20	42.05	57.91	78.27
SVM	13.26	39.82	57.30	77.36
SVM (not rectified)	11.01	37.19	54.81	73.53
PCCA+PP2+DF	**24.08**	**55.48**	**71.93**	**86.51**
PCCA+PP1+DF	22.97	54.14	70.61	85.31
PCCA+PP2	20.29	54.57	71.51	86.51
PCCA+PP1	19.28	53.30	70.66	85.23
PCCA+DF	16.98	51.64	68.91	84.58
PCCA	15.05	50.05	68.52	84.05
PCCA (not rectified)	13.23	47.98	66.93	82.73

on challenging datasets suggest that the proposed algorithm can significantly improve performance and robustness in real-world re-identification problems with lighting and viewpoint changes. We note that some newly proposed metric learning algorithms [24], [37] claim higher performance than the ones we used for comparison. We plan to combine our proposed algorithm with such metric learning algorithms, with the hope that the re-id performance can be further elevated.

In a more general scenario, the proposed algorithms can be extended by using two pose priors corresponding to the front view and back view of a subject. For each person, both the front and back views can be estimated, along with measures of confidence in the two views based on the estimated pose. As we track the target and observe more images, the descriptors of the front and/or back views can be updated if the incoming image has a higher confidence than the current confidence index.

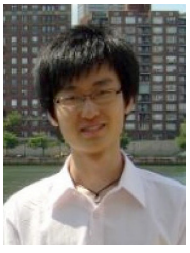
The performance of the proposed algorithms may be degraded when the appearances of all the candidates are similar, or if they move in unusual patterns. To address these issues, future work also includes adding dynamic characteristics to descriptors and continuously learning discriminative features. We also plan to investigate calibration-free pose/viewpoint estimation to make the algorithm more general.

REFERENCES

[1] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, June 2010.
[2] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. *ECCV*, 2012.
[3] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Person Re-identification Using Haar-based and DCD-based signature. *AVSS*, 2010.
[4] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. *AVSS*, 2010.
[5] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Boosted human re-identification using Riemannian manifolds. *Image and Vision Computing*, 30(6-7):443–452, Oct. 2011.
[6] D. Baltieri, R. Vezzani, and R. Cucchiara. 3DPeS: 3D people dataset for surveillance and forensics. *International ACM Workshop on Multimedia Access to 3D Human Objects*, 2011.
[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, Mar. 2008.
[8] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Lett.*, 33(7):898–903, May 2012.
[9] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes. A database for person re-identification in multi-camera surveillance networks. *Digital Image Computing: Techniques and Applications*,2012.
[10] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. *ICCV*, 2011.
[11] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. *BMVC*, 2011.
[12] D. Conte, P. Foggia, G. Percannella, and M. Vento. A multiview appearance model for people re-identification. *AVSS*, 2011.
[13] A. D'Angelo and J.-L. Dugelay. People re-identification in camera networks based on probabilistic color histograms. *SPIE Electronic Imaging*, 2011.
[14] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. *ICML*, 2007.
[15] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. *ACCV*, 2011.

[16] M. Eisenbach, A. Kolarow, K. Schenk, K. Debes, and H.-M. Gross. View invariant appearance-based person reidentification using fast online feature selection and score level fusion. *AVSS*, 2012.
[17] M. Farenzena, L. Bazzani, and A. Perina. Person re-identification by symmetry-driven accumulation of local features. *CVPR*, 2010.
[18] T. Gandhi and M. M. Trivedi. Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation. *Machine Vision and App.*, 18(3-4):207–220, Feb. 2007.
[19] N. Gheissari and T. Sebastian. Person reidentification using spatiotemporal appearance. *CVPR*, 2006.
[20] D. Gray. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *ECCV*, 2008.
[21] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. *ICCV*, 2009.
[22] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. *SCIA*, 2011.
[23] M. Hirzer, P. M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. *AVSS*, 2012.
[24] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. *ECCV*, 2012.
[25] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, Feb. 2008.
[26] T. Joachims. Training linear SVMs in linear time. *KDD*, 2006.
[27] K. Jungling and M. Arens. View-invariant person re-identification with an Implicit Shape Model. *AVSS*, 2011.
[28] Y. Li, B. Wu, and R. Nevatia. Human detection by searching in 3D space using camera and scene knowledge. *ICPR*, 2008.
[29] G. Lian, J. Lai, and W.-S. Zheng. Spatial-temporal consistent labeling of tracked pedestrians across non-overlapping camera views. *Pattern Recognition*, 44(5):1121–1136, May 2011.
[30] Z. Lin and L. S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. *Int. Symp. Visual Computing*, 2008.
[31] C. Liu, S. Gong, C. Loy, and X. Lin. Person re-identification: What features are important? *1st Int. Workshop Re-Identification*, 2012.
[32] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.
[33] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1513–8, Sept. 2006.
[34] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. *BMVC*, 2012.
[35] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by Fisher vectors for person re-identification. *1st Int. Workshop Re-Identification*, 2012.
[36] R. Mazzon, S. F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recognition Lett.*, 33(14):1828–1837, Oct. 2012.
[37] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. *CVPR*, 2012.
[38] UK Home Office. i-LIDS Multiple Camera Tracking Scenario Definition, 2008.
[39] S. Pedagadi, J. Orwell, and S. Velastin. Local Fisher Discriminant Analysis for pedestrian re-identification. *CVPR*, 2013.
[40] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-Identification by Support Vector Ranking. *BMVC*, 2010.
[41] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. *XXII Brazilian Symp. Computer Graphics and Image Processing*, 2009.
[42] M. Tapaswi. Knock! Knock! Who is it? Probabilistic person identification in TV-series. *CVPR*, 2012.
[43] L. F. Teixeira and L. Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Lett.*, 30(2):157–167, Jan. 2009.
[44] D.-N. Truong Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray. People re-identification by spectral classification of silhouettes. *Signal Process.*, 90(8):2362–2374, Aug. 2010.
[45] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–96, Sept. 2010.
[46] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1505–1518, Dec. 2003.
[47] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learning Research*, 10:207–244, Dec. 2009.
[48] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. *CVPR*, 2013.
[49] W. Zheng, S. Gong, and T. Xiang. Re-identification by Relative Distance Comparison. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):653–668, June 2012.
[50] W. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. *CVPR*, 2012.
[51] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. *BMVC*, 2009.

Ziyan Wu Ziyan Wu received a Ph.D. degree in Computer and Systems Engineering in the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute in 2014. He received a B.S. degree in Electrical Engineering and Automation and an M.S. degree in Measurement Technology and Instruments, both from Beihang University in Beijing, China in 2006 and 2009 respectively. As a graduate student, he was affiliated with the DHS Center of Excellence on Explosives Detection, Mitigation and Response (ALERT). His research interests include camera calibration, object tracking, anomaly detection and human re-identification.



Yang Li Yang Li is currently a Ph.D. student in the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute. He received a B.Eng. degree in Electrical Engineering from Hong Kong Polytechnic University in 2010. He received a HKSAR government scholarship in 2010. He is a graduate student affiliated with the DHS Center of Excellence on Explosives Detection, Mitigation and Response (ALERT). His research interests include pedestrian surveillance, multi-object tracking and human re-identification in non-overlapping camera networks.

Richard J. Radke Richard J. Radke joined the Electrical, Computer, and Systems Engineering department at Rensselaer Polytechnic Institute in 2001, where he is now a Full Professor. He has B.A. and M.A. degrees in computational and applied mathematics from Rice University, and M.A. and Ph.D. degrees in electrical engineering from Princeton University. His current research interests include computer vision problems related to modeling 3D environments with visual and range imagery, designing and analyzing large camera networks, and machine learning problems for radiotherapy applications. Dr. Radke is affiliated with the NSF Engineering Research Centers for Subsurface Sensing and Imaging Systems (CenSSIS) and Smart Lighting, the DHS Center of Excellence on Explosives Detection, Mitigation and Response (ALERT), and Rensselaer's Experimental Media and Performing Arts Center (EMPAC). He received an NSF CAREER award in March 2003 and was a member of the 2007 DARPA Computer Science Study Group. Dr. Radke is a Senior Member of the IEEE and an Associate Editor of *IEEE Transactions on Image Processing*. His textbook *Computer Vision for Visual Effects* was published by Cambridge University Press in 2012.