Person Re-identification: System Design and Evaluation Overview

Xiaogang Wang and Rui Zhao

Abstract Person re-identification has important applications in video surveillance. It is particularly challenging because observed pedestrians undergo significant variations across camera views, and there are a large number of pedestrians to be distinguished given small pedestrian images from surveillance videos. This chapter discusses different approaches of improving the key components of a person re-identification system, including feature design, feature learning and metric learning, as well as their strength and weakness. It provides an overview of various person re-identification systems and their evaluation on benchmark datasets. Mutliple benchmark datasets for person re-identification are summarized and discussed. The performance of some state-of-the-art person identification approaches on benchmark datasets is compared and analyzed. It also discusses a few future research directions on improving benchmark datasets, evaluation methodology and system desgin.

1 Introduction

Person re-identification is to match pedestrian images observed in different camera views with visual features. The task is to match one or one set of query images with images of a large number of candidate persons in the gallery in order to recognize the identity of the query image (set). It has important applications in video surveillance including pedestrian search, multi-camera tracking and behaviour analysis. Under the settings of multi-camera object tracking, matching of visual features can be integrated with spatial and temporal reasoning [29, 32, 8]. This chapter focuses

Xiaogang Wang
Department of Electronic Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong, e-mail: xgwang@ee.cuhk.edu.hk

Rui Zhao
Department of Electronic Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong, e-mail: rzhao@ee.cuhk.edu.hk

Fig. 1 The same 12 pedestrians captured in two different camera views. Examples are from the VIPeR dataset [21].

on visual feature matching. A detailed survey on spatial and temporal reasoning in object tracking can be found in [70]. People working on the problem of person re-identification usually assume that observations of pedestrians are captured in relatively short periods, such that clothes and body shapes do not change much and can be used as cues to recognize identity. In video surveillance, the captured pedestrians are often small in size, facial components are indistinguishable in images and face recognition techniques are not applicable. Therefore person re-identification techniques become important. However, its a very challenging task. Surveillance cameras may observe tens of thousands pedestrians in a public area in one day and many of them look similar in appearance. Another big challenge comes from large variations of lightings, poses, viewpoints, blurring effects, image resolutions, camera settings, and background across camera views. Some examples are shown in Figure 1. The appearance of same pedestrians observed in different camera views changes a lot.

This book chapter provides an overview of designing a person re-identification system, including feature design, feature learning and metric learning. The strength and weakness of different person re-identification algorithms are analyzed. It also reviews the performance of state-of-the-art algorithms on benchmark datasets. Some future research directions are discussed.

2 System Design

2.1 System Diagram

The diagram of a person re-identification system is shown in Figure 2. It starts with automatic pedestrian detection. Many existing works [47, 73] detect pedestri-

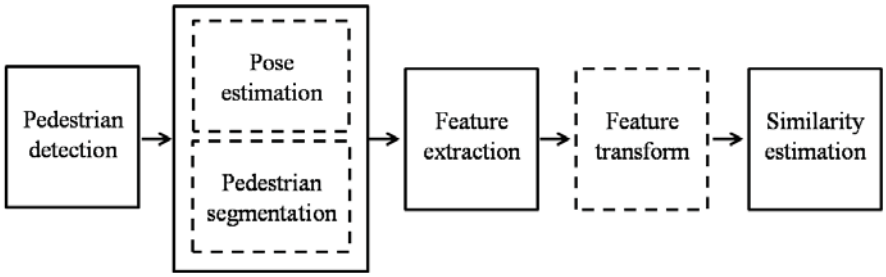


Fig. 2 Diagram of a person re-identification system. Dashed windows indicate steps which can be skipped in some person re-identification systems.

ans from videos captured by static cameras with background subtraction. However, background subtraction is sensitive to lighting variations and scene clutters. It is also hard to separate pedestrians appearing in groups. In recent years, appearance-based pedestrian detectors [11, 19, 63, 50] learned from training samples become popular. There is a huge literature on this topic. The details will be skipped in this chapter. All the existing person re-identification papers have ignored this step and assume perfect pedestrian detection by using manually cropped pedestrian images. However, perfect detection is impossible in real applications and misalignment can seriously reduce the person re-identification performance. Therefore this factor needs to be carefully studied in the future work.

The performance of person re-identification is largely affected by variations of poses and viewpoints, which can be normalized with pose estimation [69, 60]. The change of background also has negative effect on estimating the similarity of two pedestrians. Background can be removed through pedestrian segmentation [7, 55, 16]. Although significant research works have been done on these two topics, they are still not mature enough to work robustly in most surveillance scenes. The errors of pose estimation and segmentation may lead to re-identification failures. Some person re-identification systems skip the two steps and directly extract features from detection results.

Same pedestrians may undergo significant photometric and geometric transforms across camera views. Such transforms can be estimated through a learning process. However, it is also possible to overcome such transforms by learning proper similarity metrics without the feature transform step.

Person re-identification approaches generally fall into two categories: unsupervised [64, 17, 15, 9, 45, 26, 43, 44, 39, 72] and supervised [22, 54, 75, 37, 36]. Unsupervised methods mainly focus on feature design and feature extraction. Since they do not require manually labeling training samples, they can be well generalized to new camera views without additional human labeling efforts. Supervised methods generally have better performance with the assistance of manually labeled training samples. Most existing works [64, 22, 24, 17, 15, 54, 9, 25, 75, 37, 39, 35, 44, 43, 72] choose training and test samples from the same camera views and it is uncertain about their generalization capability on new camera settings. Only very recently,

people started to investigate the cases when training and test samples are from different camera views [36]. In surveillance applications, when the number of cameras is large, it is impractical to label training samples for every pair of camera views. Therefore, the generalization capability is important. The overview of designing each module of the person re-identification system is given below.

2.2 Low-level Features

Feature design is the foundation of the person re-identification system. Effective low-level features usually have good generalization capability to new camera views because their design does not rely on training. Most low-level features can be integrated with the learning approaches developed in the later steps. Good features are expected to discriminate a large number of pedestrians in the gallery and to be robust to various inter- and intra-camera view variations, such as background, poses, lighting, viewpoints and self-occlusions.

2.2.1 Color, shape and texture features

Like most object recognition tasks, the appearance of pedestrians from static images can be characterized from three aspects: color, shape and texture. Color histograms of the whole images are widely used to characterize color distributions [49, 51, 10]. In order to be robust to lighting variations and the changes of photometric settings of cameras, various color spaces have been studied when computing color histograms [64]. Some components in the color spaces sensitive to photometric transformations are removed or normalized. Instead of uniformly quantizing the color spaces, Mittal and Davis [46] softly assigned pixels to color modes with a Gaussian mixture model, and estimated the correspondences of color modes across camera views. Other color invariants [59, 10, 66] can also be used as features for person re-identification.

Color distributions alone are not enough to distinguish a large number of pedestrians since the clothes of some pedestrians could be similar. Therefore, it needs to be combined with shape and texture features. Shape context [4] is widely used to characterize both global and local shape structures. Its computation is based on edge or contour detection. Histogram of Oriented Gradients (HOG) has been widely used for object detection [11], and is also effective for person re-identification [64, 57]. It characterizes local shapes by computing the histograms of gradient orientations within cells over a dense grid. In order to be robust to lighting variations, local photometric normalization is applied to histograms. Shape features are subject to the variations of viewpoints and poses.

Many texture filters and descriptors have been proposed in object recognition literature, such as Gabor filer [12] and other linear filter banks [68, 62], SIFT [40], color SIFT [1], LBP [48], and region covariance [61]. Many of them can also be used in person re-identification [24]. A typical approach is to apply these filters

and descriptors to sparse interest points or on a dense grid, and then quantize their responses into visual words. The histograms of visual words can be used as features to characterize texture distributions. However, these features cannot encode spatial information. It is also possible to directly compare the responses on a fixed dense grid. But its is sensitive to misalignment, pose variation and viewpoint variation. Therefore, correlograms [28] and correlatons [56] are proposed to capture the co-occurrence of visual words over spatial kernels. They balance the two extreme cases.

2.2.2 Global, regional and patch-based features

Most of the visual features described above are global. They have some invariance to misalignment, pose variation, and the change of viewpoint. However, their discriminative power is not high because of losing spatial information. In order to increase the discriminative power, patch-based features are used [22, 17, 39, 37, 72, 36]. A pedestrian image is evenly divided into multiple local patches. Visual features are computed as each patch. When computing the similarity of two images, visual features of two corresponding patches are compared. The biggest challenge of patch-based methods is to find correspondences of patches when tackling the misalignment problem. Zhao *et al.* [72] divided an image into horizontal stripes and find the dense correspondence of patches along each stripe with some spatial constraints.

Some patches are more distinctive and reliable when matching two persons. Some examples are shown in Figure 3. In this dataset, it is easy for human eyes to match pedestrian pairs because they have distinct patches. Person (a) carries a backpack with tilted blue stripes. Person (b) holds a red folder. Person (c) has a red bottle in hand. These features can well separate one person from others and they can be reliable detected across camera views. If a body part is salient in one camera view, it should also be salient in another camera view. However, most existing approaches only consider clothes and trousers as the most important regions for person re-identification. Such distinct features may be considered as outliers to be removed, since some of them do not belong to body parts. Also, these features may only take small regions in the body parts, and have little effect on computing global features. Zhao *et al.* [72] estimated the salience of patches through unsupervised learning and incorperate it into person matching. A patch with higher salience value gains more weight in the matching.

Pedestrians have fixed structures. If different body constitutes can well detected with pose estimation and human parsing are available, region-based features can be developed and employed in person re-identification [20, 9]. Visual features are computed from each body part. Body alignment is naturally established. Cheng *et al.* [9] employed Custom Pictorial Structure to localize body parts, and matched their visual descriptors. Wang *et al.* [64] proposed shape and appearance context. The body parts are automatically obtained through clustering shape context. The shape and appearance context models the spatial distributions of appearance relative to body parts.

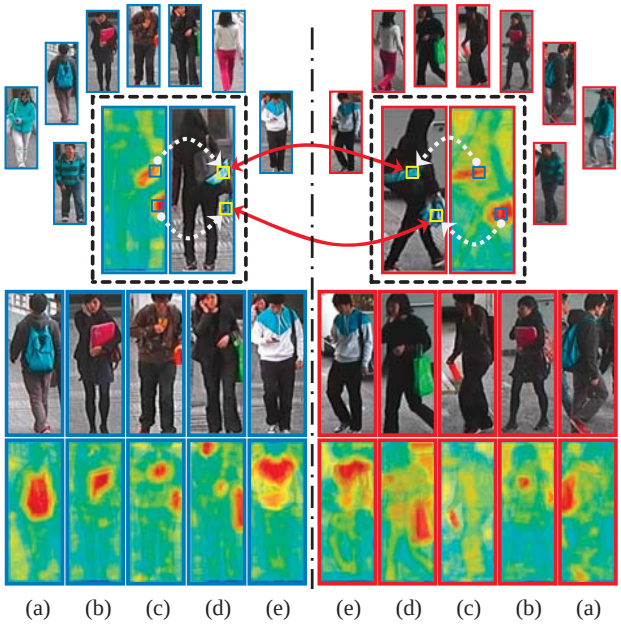


Fig. 3 Illustration of patch-based person re-identification with salience estimation. The dash line in the middle divides the images observed in two different camera views. The salience maps of exemplar images are also shown.

2.3 Semantic Features

In order to effectively reduce the cross-view variations, some high-level semantic features could be used for person re-identification besides the low-level visual features discussed above. The design of semantic features is inspired by the process of human beings recognizing person identities. For example, humans describe a person by saying "he or she looks similar to someone I know" or "he or she is tall and slim, has shot hair, wears a white shirt, and carries a baggage". Such high-level descriptions are independent of camera views and have good robustness. In the computer vision field, semantic features have also been widely used in face recognition [71], general object recognition [18], and image search [65].

Shan *et al.* [58, 23] proposed exemplar-based representations. An illustration is shown in Figure 4. The similarities of an image sample with selected representative persons in the training set are used as the semantic feature representation of the image. Suppose a and b are the two camera views to be matched. n representative pairs $\{(\mathbf{x}_1^a, \mathbf{x}_1^b), \ldots, (\mathbf{x}_n^a, \mathbf{x}_n^b)\}$ are selected as exemplars in the training set. $\mathbf{x}_i^a$ and $\mathbf{x}_i^b$ are the low-level feature vectors of the same person identity i, but are observed in different camera views a and b. If the low-level feature vector of a sample image $\mathbf{y}^a$ is observed in camera view a, it is compared against the n representative persons also

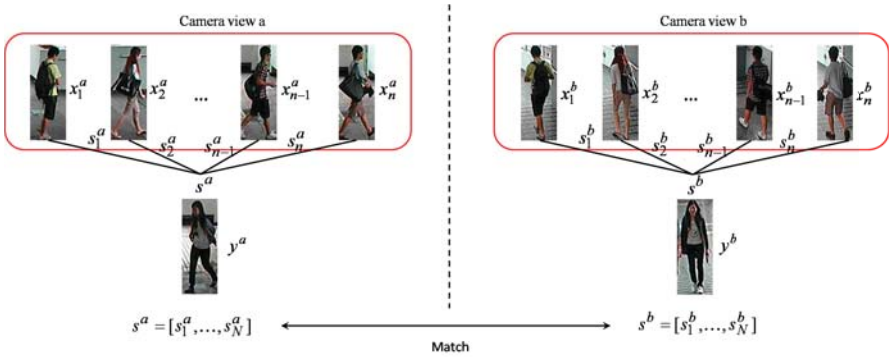


Fig. 4 Illustration of exemplar-based representation for person re-identification.

observed in a, and its semantic features are represented with a $n-$dimensional vector $\mathbf{s}^a = (s_1^a, \ldots, s_n^a)$, where s_i^a is the similarity between $\mathbf{y}^a$ and $\mathbf{x}_i^a$ by matching their low-level visual features. If a sample $\mathbf{y}^b$ is observed in camera view b, its semantic feature vector $\mathbf{s}^b$ can be computed in the same way. When computing $\mathbf{s}^a$ and $\mathbf{s}^b$, the low-level visual features are only compared under the same camera view, and therefore large cross-view variations are avoided. Eventually, the similarity between $\mathbf{y}^a$ and $\mathbf{y}^b$ are computed by comparing the semantic feature vectors $\mathbf{s}^a$ and $\mathbf{s}^b$. The underlying assumption is that if a person in test is similar to one of the representative persons i in the training set, its observations in camera views a and b should be similar to $\mathbf{x}_i^a$ and $\mathbf{x}_i^b$ respectively, and therefore both s_i^a and s_i^b are large no matter how different the two camera views are. Therefore, if $\mathbf{y}^a$ and $\mathbf{y}^b$ are the observations of the same object, $\mathbf{s}^a$ and $\mathbf{s}^b$ are similar.

Layne *et al.* [35] employed attribute features for person re-identification. They defined 15 binary attributes regarding to cloth-style, hair-style, carrying-object and gender. Attribute classifiers are based on low-level visual features. They are learned with SVM from a set of training samples whose attributes are manually labeled. The outputs of attribute classifiers are used as feature representation for person re-identification. They can also be combined with low-level visual features for matching. Since the training samples with the same attribute may come from different camera views, the learned attribute classifiers may have view invariance to some extent. Liu *et al.* [39] weighted attributes according to their importance in person re-identification. Attribute-based approaches require more labeling effort for training attribute classifiers. While in other approaches each training sample only needs one identity label, it requires all the M attributes to be labeled for a training sample.

2.4 Learning Feature Transforms across Camera Views

In order to learn the feature transforms across camera views, one could first assume the photometric or geometric transform models and then learn the model parameters from training samples [52, 30, 53]. For example, Prosser *et al.* [53] assumed the photometric transform to be bi-directional Cumulative Brightness Transfer Functions, which map color observed in one camera view to another. Porikli and Divakaran [52] learned the color distortion function between camera views with correlation matrix analysis. Geometric transforms can also be learned from the correspondences of interest points.

However, in many cases, the assumed transform functions cannot capture the complex cross-camera transforms which could be multi-model. Even if all the pedestrian images are captured by a fixed pair of camera views, their cross-view transforms may have different configurations because of many different possible combinations of poses, resolutions, lightings and background. Li and Wang [36] proposed a gating network to project visual features from different camera views into common feature spaces for matching without assuming any transform functions. As shown in Figure 5, it automatically partitions the image spaces of two camera views into subregions, corresponding to different transform configurations. Different feature transforms are learned for different configurations. A pair of images to be matched are softly assigned to one of the configurations and their visual features are projected on a common feature space. Each common feature space has a local expert learned for matching images. The features optimal for configuration estimation and identity matching are different and can be jointly learned. Experiments in [36] show that this approach not only can handle the multi-model problem but also have good generalization capability on new camera views. Given a large diversified training set, multiple cross-view transforms can be learned. The gating network can automatically choose a proper feature space to match test images from new camera views.

2.5 Metric Learning and Feature Selection

Given visual features, it is also important to learn a proper distance/similarity metric to further depress cross-view variations and well distinguish a large number of pedestrians. A set of reliable and discriminative features are to be selected through a learning process. Some approaches [38, 57] require that all the persons to be identified must have training samples. But this constraint largely limits their applications. In many scenarios, it is impossible to collect the training samples of pedestrians in test beforehand. Schwartz and Davis [57] learned discriminative features with Partial Least Square Reduction. The features are weighted according to the discriminative power based on one-against-all comparisons. Lin and Davis [38] learned the dissimilarity profiles under a pairwise scheme. More learned based approaches [22, 54, 75, 33, 31] were proposed to identify persons outside the training set. Zheng

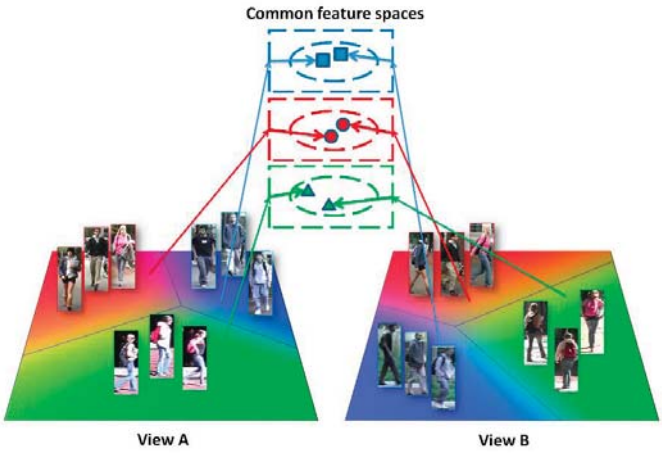


Fig. 5 Person re-identification in locally aligned feature transformations. The image spaces of two camera views are jointly partitioned based on the similarity of cross-view transforms. Sample pairs with similar transforms are projected to a common feature space for matching.

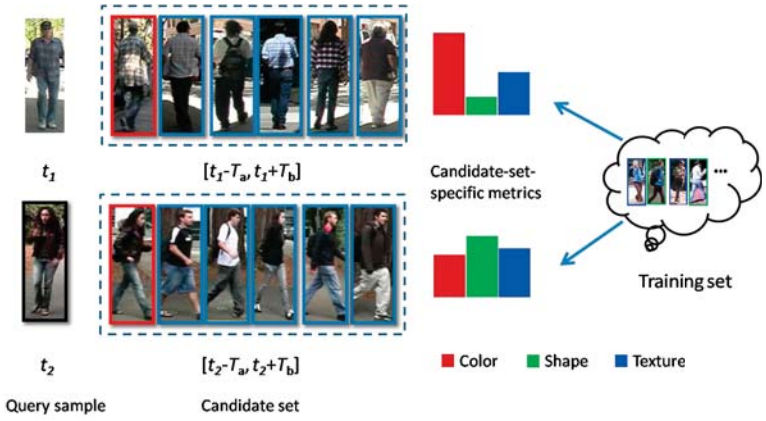


Fig. 6 Illustration of learning candidate-set-specific metric. A query sample i is observed at a camera view at time t_i. By reasoning the transition time, only the samples observed in another camera view during time window $[t_i - T_a, t_i + T_b]$ are considered at candidates. To distinguish persons in the first candidate set, color features are more effective. For the second candidate set, shape and texture could be more useful. Persons in the candidate sets do not have training samples. Candidate-set-specific metrics could be learned from a large training set through transfer learning.

et al. [75] learned a distance metric which maximizes the probability that a pair of true match has a smaller distance than a wrong match. Gray and Tao [22] employed boosting to select viewpoint invariant and discriminative features for person re-identification. Prosser *et al.* [54] formulated person re-identification as a ranking problem and used RankSVM to learn an optimal subspace.

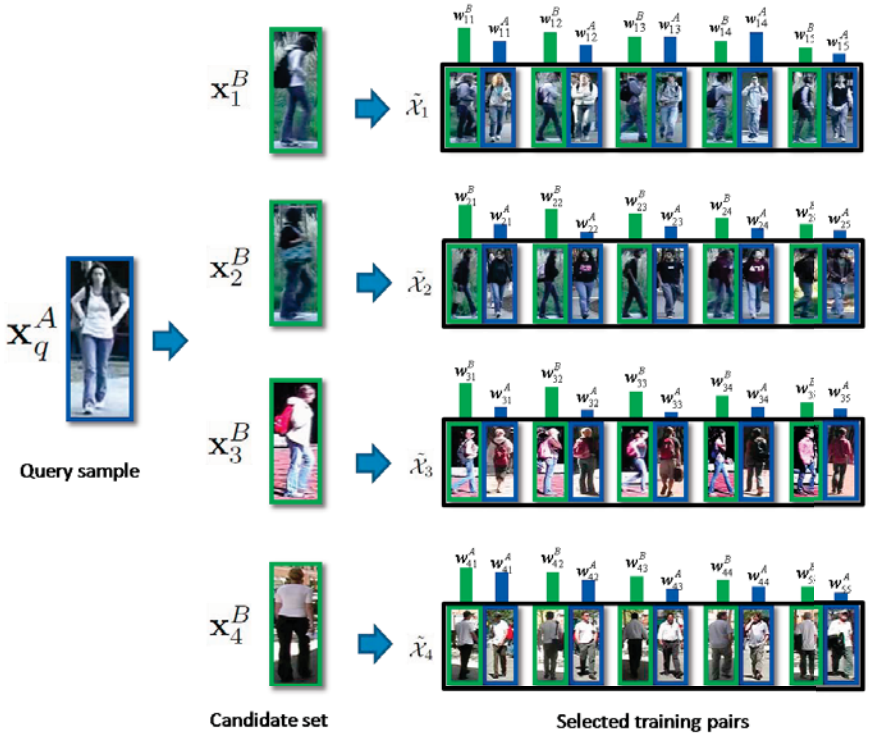


Fig. 7 Illustration of transfer learning for person re-identification proposed in [37]. Blue and green windows indicate samples observed in camera views A and B. $\mathbf{x}_q^A$ is a query sample observed in camera view A. $\mathbf{x}_1^B, \ldots, \mathbf{x}_4^B$ are samples of four candidate persons observed in camera view B. Each $\mathbf{x}_i^B$ finds five nearest neighbors in the same camera view B from the training set. Since the corresponds of training samples in camera views A and B are known, the paired samples of the nearest neighbors can be found to training candidate-set-specific metric. w_{ij}^A and w_{ij}^B are the weights assigned to each pair of training samples according to their visual similarities to the candidates and the query sample.

The difficulty of person re-identification increases with the number of candidates to be matched. In cross-camera tracking, given a query image observed in one camera view, the transition time across two camera views can be roughly estimated. This simple temporal reasoning can simplify the person re-identification problem by pruning the candidate set to be matched in another camera view. All the approaches discussed above adopt a fixed metric to match a query image with any candidate. However, if the goal is to distinguish a small number of subjects in a particular candidate set, candidate-set-specific distance metrics should be preferred. An illustration is shown in Figure 6. For example, the persons in one candidate set can be well distinguished with color features, while persons in another candidate set may be better distinguished with shape and texture. A better solution should tackle this problem through optimally learning different distance metrics for different can-

didate sets. Unfortunately, during online tracking, the correspondence of samples across camera views cannot be manually labeled for each person in the candidate set. Therefore, directly learning a candidate-set-specific metric is infeasible, since metric learning requires pairs of samples across camera views with correspondence information. Li *et al.* [37] tackled this problem by proposing a transfer learning approach. It assumes a large training set with pair training samples across camera views. This training set has no overlap with candidate sets on person identities. As shown in Figure 7, each sample in the candidate set finds its nearest neighbors in the training set based on visual similarities. Since the training set has ground truth labels, the corresponding training samples of the found nearest neighbors in another camera view are known. The selected training pairs are weighted according to their visual similarities to the samples in the candidate set and the query sample. Finally, the candidate-set-specific distance metric is learned from the selected and weighted training pairs.

3 Benchmark Datasets

Multiple benchmark datasets for person re-identification have been published in recent years. There are multiple factors to be considered when creating a benchmark dataset. (1) The number of pedestrians. As the number of pedestrians in the gallery increases, the person re-identification task becomes more challenging. On the other hand, when more pedestrians are included in the training set, the learned recognizer will be more robust at the test stage. (2) The number of images per person in one camera view. Multiple images per person can capture the variations poses and occlusions. If they are available in the gallery, person re-identification becomes easier. They also improve the training process. They are available in practical applications, if assuming pedestrians can be tracked in the same camera views. (3) Variations of resolutions, lightings, poses, occlusion and background in the same camera view and across camera views. (4) The number of camera views. As it increases, the complexity of possible transforms across camera views becomes more complicated.

VIPeR dataset [21] built by Gray *et al.* includes 632 pedestrians taken by two surveillance camera views. Each person only has one image per camera view. The two cameras were placed at many different locations and therefore the captured images cover a large range of viewpoints, poses, lighting and background variations, which makes image matching across camera views very challenging. Images were sampled from videos with compression artifacts. The standard protocol on this dataset is to randomly partition the 632 persons into two non-overlapping parts, 316 persons for training and the remaining ones for test. It is the most widely used benchmark dataset for person re-identification so far.

ETHZ dataset [57] includes $8,580$ images of 146 persons taken with moving cameras in a street scene. Images of a person are all taken with the same camera and undergo less viewpoint variation. However, some pedestrians are occluded due to

the crowdedness of the street scene. The number of images per person varies from 10 to 80.

i-LIDS MCTS dataset created by Zheng *et al.* [74] was collected from an airport arrival hall. It includes 476 images of 119 pedestrians. Most persons have four images captured by the same camera views or non-overlapping different camera views.

CAVIAR4REID created by Cheng *et al.* [9] collected 1,220 images of 72 pedestrians from a shopping center. 50 pedestrians were captured with two camera views and the remaining ones by one camera view. Compared with other datasets, its images have large variation on resolutions.

Person Re-ID 2011 Dataset created by Hirzer *et al.* [25] have 931 persons captured with two static surveillance cameras. 200 of them appear in both camera views. The remaining ones only appear in one of the camera views.

RGB-D Person Re-identification Dataset created by Barbosa *et al.* [3] has depth information of 79 pedestrians captured in an indoor environment. For each person, the synchronized RGB images, foreground masks, skeletons, 3D meshes and estimated floor are provided. The motivation is to evaluate the person re-identification performance for long-term video surveillance where the clothes can be changed.

QMUL underGround Re-IDentification (GRID) Dataset created by Loy *et al.* [41, 42] contains 250 pedestrian image pairs captured from a crowded underground train station. Each pair of images have the same identity and were captured by two non-overlapping camera views. All the images were captured by 8 camera views.

Besides the dataset discussed above, there are also some other datasets published recently such as the **CUHK Person Re-identification Dataset** [37, 36], the **3DPes Dataset** [2], the **Multi-Camera Surveillance Database** [6]. [2] and [6] also provided video sequences besides snapshots. The emergence of all these benchmark datasets clearly advanced the state-of-the-art on person re-identification. However, they also have several important drawbacks to be addressed in the future work.

First of all, the images in all the benchmark datasets are manually cropped. Most of the datasets even did not provide the original image frames. It means the assumption that images are perfectly aligned. Thereafter, all the developed algorithms and training process are based on this assumption. However, in practical surveillance applications, perfect alignment is impossible and pedestrian images need to be automatically cropped with pedestrian detectors [11, 19]. It is expected that the performance of existing person re-identification algorithms should drop significantly with the existence of misalignment. However, such effect has been ignored by almost all the existing publications. When building new benchmark datasets, automatically cropped image with state-of-the-art pedestrian detectors should be provided.

Secondly, the numbers of camera views in the existing datasets are small (the maximum number is 8). Moreover, in existing evaluation protocols, training and testing images are from the same camera views. The biggest challenge of person re-identification is to learn and depress cross-camera-view transforms. Given the fact that tens of thousands of surveillance cameras are available in large cities, in most surveillance applications, it is impossible to manually label training samples for every pair of camera views. Therefore, it is uncertain about the generalization

capability of existing algorithms given a pair new camera views in test without extra training samples from which.

Thirdly, the numbers of persons ($< 1,000$) and the numbers of images ($< 10,000$) in person re-identification datasets are still much smaller than the scales of existing benchmark datasets for other computer vision problems such as object detection, object recognition, and face recognition. For example, ImageNet [14] for object classification has more than 14 million images. LFW [27] for face recognition has more than $5,000$ people. Some powerful machine learning tools such as deep models [5] have shown superior performance on computer vision challenges [34] based on large scale training data. Therefore, it is desirable to build very large scale datasets covering a large set of diversified camera views for person re-identification, which could not only significantly boost the performance but also enhance to the generalization capability.

4 Evaluation

The accumulative matching characteristic (CMC) is the most widely used to evaluate the performance of person re-identification. It treats person re-identification as a ranking problem. Given one or one set of query images, the candidate images in the gallery are ranked according to their similarities to the query. CMC(k) measures the probability that the correct match has a rank equal or higher than k. As the gallery size increases, it becomes more difficult to find the correct match and CMC(k) becomes lower.

Single-shot person re-identification only analyzes a single image for each person assuming no tracking information is available. Therefore, the query or any person in the gallery only has one image. Multi-shot person re-identification assume multiple images are available for each person through tracking. Therefore, a query is a set of images and the images in the gallery are also grouped into sets according the identity information.

In Figure 8, we summarize the results of single-shot person re-identification on the VIPeR dataset. VIPeR has 632 persons. They are randomly partitioned, half of the persons for training and the remaining half for test. The existing approaches are divided into two groups (unsupervised and supervised methods) for comparison. Unsupervised methods include:

- symmetry-driven accumulation of local features (SDALF) [17];
- custom pictorial structures (CPS) [9];
- biologically inspired features and covariance descriptors (BiCov) [43];
- local descriptors encoded by Fisher vectors combined with other features (eLDFV) [44];
- and salient dense correspondence combined with other features (eSDC) [72].

Supervised methods are:

- ensemble of localized features learned with AdaBoost (ELF) [22];

- person re-identification by support vector ranking (PRSVM) [54];
- distance metric learning for large margin nearest neighbor classification (LMNN) [67];
- information theoretic metric learning (ITML) [13];
- probabilistic relative distance comparison (PRDC) [75];
- attribute sensitive feature importance combined with PRDC (ASFI+PRDC) [39];
- Pairwise Constrained Component Analysis (PCCA) [45];
- large margin nearest neighbor with rejection (LMNN-R) [15];
- relaxed pairwise learned metric (RPLM) [26];
- supervised local descriptors encoded by Fisher vectors (sLDFV) [44];
- and local aligned feature transforms (LAFT) [36].

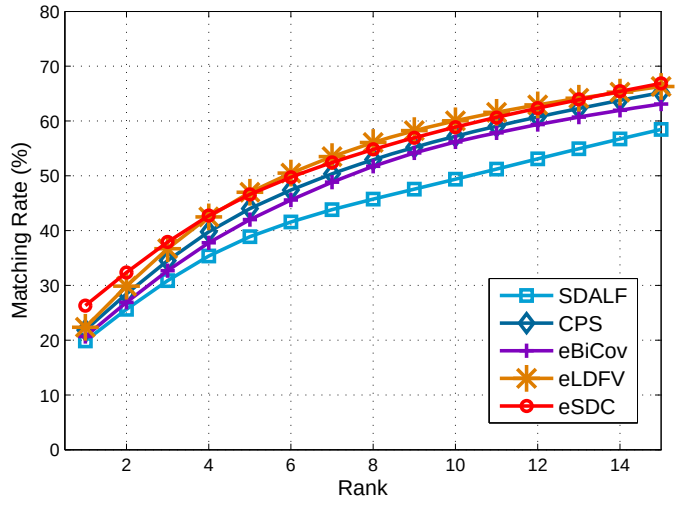
The unsupervised methods focus on feature design. All the top ranked state-of-the-art feature sets employ regional or patch-based features. It is also observed that the combination of different types of features can improve the performance. For example, both eLDFV and eSDC combine local and global features. The information of patch salience is useful in person re-identification.

Supervised methods focus on feature extraction, feature transform and metric learning. LMMNN and ITML are metric learning methods. When they are applied to person re-identification, the same visual features proposed in [75] are used. LAFT, RPLM and sLDFV perform the best on VIPeR. They all employ metric learning. LAFT locally aligns images observed in different camera views by projecting them to a common feature space. Different metrics are learned for different common feature spaces. sLFDV uses Fisher vectors for unsupervised feature learning and then employ PCCA [26] to learn the metric based on extracted features. RPLM employs a pairwise metric learning approach by relaxing hard constraints commonly used in other metric learning approaches.

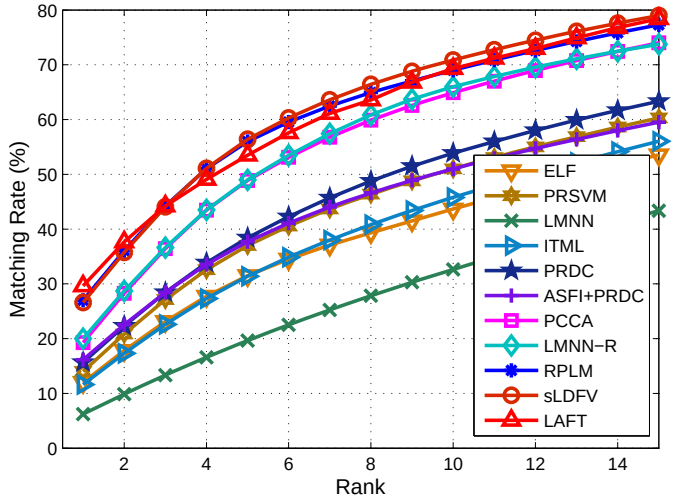
Figure 9 shows the results of multi-shot person re-identification on the ETHZ dataset. Not many papers have published their results on multi-shot person re-identification. We report the CMC curves of SDALF [17] and BiCov [43]. It is assumed that each each query person or each person in the gallery has M images. ETHZ has three video sequences and results are reported on each of them separately. Since both SDALF and BiCov are unsupervised methods, all the images in the dataset are used for test. It is observed that as M increases, the CMC curves get improved significantly.

5 Conclusions and Discussions

Person re-identification is an emerging research topic and significant research progress has been achieved in this field in the past five years. Multiple benchmark datasets and evaluation protocols have been published. This paper provides an overview of system design and evaluation. Most research works focus on feature design, feature learning, and metric learning. Different types of features characterize pedestrian images from multiple perspectives and all have their own strength



(a)



(b)

Fig. 8 CMC results of single-shot person re-identification on the VIPeR dataset. (a): unsupervised methods. (b): supervised methods.

and weakness. A lot of published results have shown that the integration of global, regional, and patch-based features, low-level visual features and high-level semantic features can improve the system performance. People start to pay attentions to high-level semantic features, importance of different attributes, and salience of local regions, not only because that they can improve the matching accuracy, but also

that they are interpretable by humans and can get human feedback involved in the recognition loop. The major challenge of person re-identification is the large variations across camera views. It is tackled by learning feature transforms and distance metrics. On a complex camera network, or even just between two camera views, the cross-view transforms are multi-modal, which cannot be handled with a single feature transform or a single distance metric. Mixture models are needed.

Person re-identification is still a very challenging problem and not well solved yet. On the VIPeR dataset, the rank-1 accuracy is still below 30%. There are multiple directions to be explored in the future. Existing works match manually cropped images. Automatic pedestrian detection should be included in the person re-identification pipeline, and the effect of misalignment caused by detection should be considered in the future research. It requires the development of new methodology as well as new benchmark datasets. When the camera network is large, it is impractical to manually label training samples for every pair of camera views. It is important to study the generalization capability of person re-identification algorithms to unseen camera views. Existing benchmark datasets are relatively small in the numbers of samples, pedestrians and camera views. The diversity of their scene coverage is also limited. In recent years, large-scale machine learning has achieved great success in many fields. It would be interesting to see its application to person re-identification. Besides online multi-camera tracking, person re-identification can be also applied to pedestrian retrieval over camera networks. Like general image retrieval systems, user feedback and linguistic descriptions can get involved in the search loop. This would be another interesting direction to be explored.

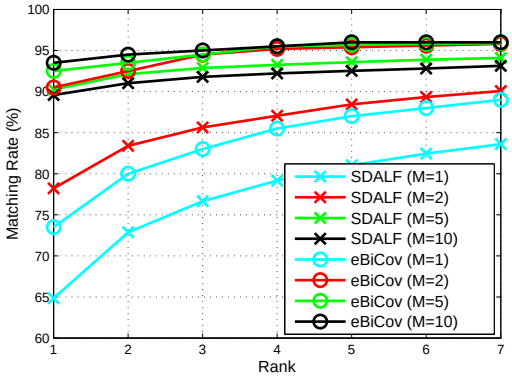
References

1. Abdel-Hakim, A.E., Farag, A.A.: Csift: A sift descriptor with color invariant characteristics. In: Proc. of European Conf. Computer Vision (2006)
2. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proc. of the 1st International ACM Workshop on Multimedia access to 3D Human Objects (2011)
3. Barbosa, B.I., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: First International Workshop on Re-Identification (2012)
4. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. on Pattern Analysis and Machine Intelligence **24**, 509–512 (2002)
5. Bengio, Y.: Learning Deep Architectures for AI. Now Publishers (2009)
6. Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P.: A database for person re-identification in multi-camera surveillance networks. In: Proc. Int'l Conf. Digital Image Computing Techniques and Applications (2012)
7. Bo, Y., Fowlkes, C.C.: Shape-based pedestrian parsing. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2011)
8. Cai, Y., Chen, W., Huang, K., Tan, T.: Continuously tracking objects across multiple widely separated cameras. In: Asian Conf. Computer Vision (2007)
9. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proc. of British Machine Vision Conf. (2011)
10. Cheng, E.D., Piccardi, M.: Matching of objects moving across disjoint cameras. In: Proc. of IEEE Int'l Conf. Image Processing (2006)

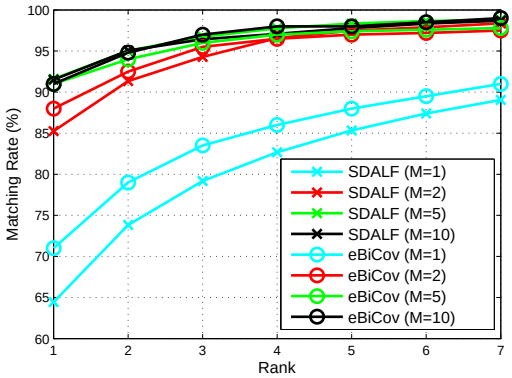
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2005)
12. Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America A **2**, 1160–1169 (1985)
13. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information theoretic metric learning. In: Proc. of Int'l Conf. Machine Learning (2007)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2009)
15. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Asian Conf. Computer Vision (2010)
16. Eslami, S.M.A., Williams, C.K.I.: A generative model for parts-based object segmentation. In: Proc. of NIPS (2012)
17. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2010)
18. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2009)
19. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2008)
20. Gheissari, N., Sebastian, T.B., Rittscher, J., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2006)
21. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance (2007)
22. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proc. of European Conf. Computer Vision (2008)
23. Guo, Y., Rao, C., Samarasekera, S., Kim, J., Kumar, R., Sawhney, H.: Matching vehicles under large pose transformations using approximate 3d models and piecewise mrf model. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2008)
24. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: Proc. of IEEE Conference on Distributed Smart Cameras (2008)
25. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: In Proc. Scandinavian Conf. on Image Analysis (2011)
26. Hirzer, M., M., R.P., Kostinger, M., Bischof: Relaxed pairwise learned metric for person re-identification. In: Proc. of European Conf. Computer Vision (2012)
27. Huang, G., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts, Amherst (2007)
28. Huang, J., Kumar, S.R., Mitra, M., Zhu, M., Zabih, R.: Image indexing using color correlograms. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (1997)
29. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2003)
30. Javed, O., Shafique, K., Shah, M.: Appearance modeling for tracking in multiple non-overlapping cameras. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2005)
31. Jurie, F., Mignon, A.: Pcca: A new approach for distance learning from sparse pairwise constraints. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2012)
32. Khan, S., Shah, M.: Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Trans. on Pattern Analysis and Machine Intelligence **25**, 1355–1360 (2003)

33. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P., Bischof, H.: Large scale metric learning from equivalence constraints. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2011)
34. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proc. of NIPS (2012)
35. Layne, R., Hospedales, T., Gong, S.: Person re-identification by attributes. In: Proc. of British Machine Vision Conf. (2012)
36. Li, W., Wang, X.: Locally aligned feature transforms across views. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2013)
37. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Asian Conf. Computer Vision (2012)
38. Lin, Z., Davis, L.: Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: Proc. Int'l Symposium on Advances in Visual Computing (2008)
39. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: What features are important? In: Proc. First International Workshop on Re-Identification (2012)
40. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
41. Loy, C., Xiang, T.: Multi-camera activity correlation analysis. Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2009)
42. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. International Journal on Computer Vision **90**, 106–129 (2010)
43. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: Proc. of British Machine Vision Conf. (2012)
44. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Proc. First International Workshop on Re-identification (2012)
45. Mignon, A., Jurie, F.: Pcca: A new approach for distance learning from sparse pairwise constraints. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2012)
46. Mittal, A., Davis, L.S.: M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. International Journal on Computer Vision **51**, 189–203 (2003)
47. Nakajima, C., Pontil, M., Heisele, B., Poggio, T.: Full-body recognition system. Pattern Recognition **36**, 1977–2006 (2003)
48. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence pp. 971–987 (2002)
49. Orwell, J., Remagnino, P., Jones, G.A.: Multiple camera color tracking. In: Poc. IEEE Workshop on Visual Surveillance (1999)
50. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2012)
51. Park, U., Jain, A., Kitahara, I., Kogure, K., Hagita, N.: Vise: Visual search engine using multiple networked cameras. In: Proc. of IEEE Int'l Conf. Pattern Recognition (2006)
52. Porikli, F.: Inter-camera color calibration by correlation model function. In: Proc. of IEEE Int'l Conf. Image Processing (2003)
53. Prosser, B., Gong, S., Xiang, T.: Multi-camera matching using bi-directional cumulative brightness transfer function. In: Proc. of British Machine Vision Conf. (2008)
54. Prosser, B., Zheng, W., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: Proc. of British Machine Vision Conf. (2010)
55. Rauschert, I., Collins, R.T.: A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In: Proc. of European Conf. Computer Vision (2012)
56. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2006)
57. Schwartz, W., Davis, L.: Learning discriminative appearance-based models using partial least sqaures. In: Proc. XXII SIBGRAPI (2009)
58. Shan, Y., Sawhney, H., Kumar, R.: Vehicle identification between non-overlapping cameras without direct feature matching. In: Proc. of IEEE Int'l Conf. Computer Vision (2005)

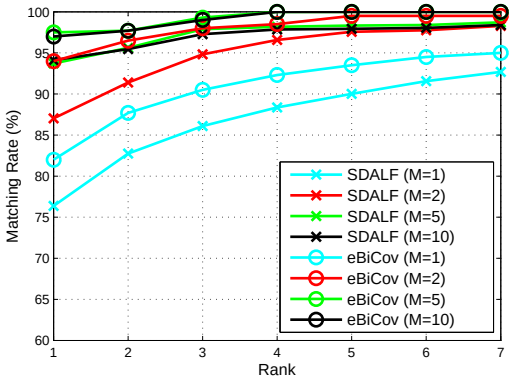
59. Slater, D., Healey, G.: The illumination-invariant recognition of 3d objects using local color invariants. IEEE Trans. on Pattern Analysis and Machine Intelligence **18**, 206–210 (1996)
60. Tian, Y., Zitnick, C.L., Narasimhan, S.G.: Exploring the spatial hierarchy of mixture models for human pose estimation. In: Proc. of European Conf. Computer Vision (2012)
61. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: Proc. of European Conf. Computer Vision (2006)
62. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal on Computer Vision **62**, 61–81 (2005)
63. Wang, M., Li, W., Wang, X.: Transferring a generic pedestrian detector towards specific scenes. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (wangLWcvpr12)
64. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: Proc. of IEEE Int'l Conf. Computer Vision (2007)
65. Wang, X., Qiu, S., Liu, K., Tang, X.: Web image re-ranking using query-specific semantic signatures. IEEE Trans. on Pattern Analysis and Machine Intelligence (2013)
66. Weijer, J., Schmid, C.: Coloring local feature extraction. In: Proc. of European Conf. Computer Vision (2006)
67. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: Proc. of NIPS (2006)
68. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Proc. of IEEE Int'l Conf. Computer Vision (2005)
69. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2011)
70. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys **38**, 1–45 (2006)
71. Yin, Q., Tang, X., Sun, J.: An associate-predict model for face recognition. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2011)
72. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2013)
73. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. IEEE Trans. on Pattern Analysis and Machine Intelligence **30**, 1198–1211 (2008)
74. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: Proc. of British Machine Vision Conf. (2009)
75. Zheng, W., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition (2011)



(a) Sequence 1



(b) Sequence 2



(b) Sequence 3

Fig. 9 CMC results of multi-shot person re-identification on the ETHZ dataset.