Person Re-identification using Deformable Patch Metric Learning

Sławomir Bąk Peter Carr
Disney Research
Pittsburgh, PA, USA, 15213
{slawomir.bak,peter.carr}@disneyresearch.com

Abstract

The methodology for finding the same individual in a network of cameras must deal with significant changes in appearance caused by variations in illumination, viewing angle and a person's pose. Re-identification requires solving two fundamental problems: (1) determining a distance measure between features extracted from different cameras that copes with illumination changes (metric learning); and (2) ensuring that matched features refer to the same body part (correspondence). Most metric learning approaches focus on finding a robust distance measure between bounding box images, neglecting the alignment aspects. In this paper, we propose to learn appearance measures for patches that are combined using a spring model for addressing the correspondence problem. We validated our approach on the VIPeR, i-LIDS and CUHK01 datasets achieving new state of the art performance.

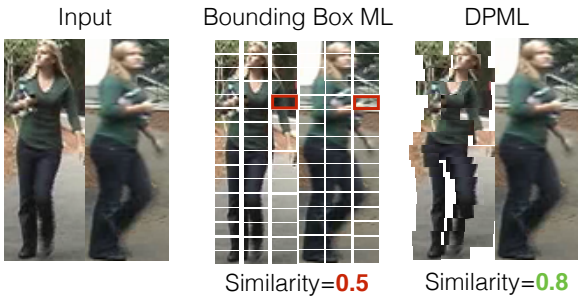


Figure 1: Full bounding box metric learning *vs*. deformable patch metric learning (DPML). The corresponding patches in the grid (highlighted in red) do not correspond to the same body part because of the pose change. Information from such misaligned features might be lost during the metric learning step. Instead, our DPML deforms to maximize similarity using metrics learned on a patch level.

1. Introduction

Person re-identification is the problem of recognizing the same individual across a network of cameras. In a real-world scenario, the transition time between cameras may significantly decrease the search space, but temporal information alone is not usually sufficient to solve the problem. As a result, visual appearance models have received a lot attention in computer vision research [3, 19, 20, 21, 22, 25, 27]. The underlying challenge for visual appearance is that the models must work under significant appearance changes caused by variations in illumination, viewing angle and a person's pose.

Metric learning approaches often achieve the best performance in re-identification. These methods learn a distance function between features from different cameras such that relevant dimensions are emphasized while irrelevant ones are discarded. Many metric learning approaches [6, 14, 18] divide a bounding box pedestrian image into a fixed grid of regions and extract descriptors which are then concatenated into a high-dimensional feature vector. Afterwards, dimensionality reduction is applied, and then metric learning is performed on the reduced subspace of differences between feature vectors. To avoid overfitting, the dimensionality must be significantly reduced. In practice, the subspace dimensionality is about three orders of magnitude smaller than the original. Such strong dimensionality reduction might result in the loss of discriminative information. Additionally, features extracted on a fixed grid (see Fig. 1), may not correspond even though it is the same person (*e.g*. due to a pose change). Metric learning is unable to recover this lost information.

In this paper, instead of learning a metric for concatenated features extracted from full bounding boxes from different cameras, we propose to learn metrics for 2D patches. Furthermore, we do not assume the patches must be located on a fixed grid. Our model allows patches to perturb their location when computing similarity between two images (see Fig. 1). This model is inspired from part-based object detection [8, 26], which decomposes the appearance model into local templates with geometric constraints (conceptualized as springs). Our contributions are:

- We propose to learn metrics locally, on feature vectors extracted from patches. These metrics can be combined into a unified distance measure.
- We introduce a deformable patch-based model for accommodating pose changes and occlusions. This model combines an appearance term with a deformation cost that controls relative placement of patches.

Our experiments illustrate the merits of patch-based comparison and achieve state of the art performance on multiple data sets.

2. Related work

Person re-identification approaches can be divided into two groups: *feature modeling* [2, 7] designs descriptors (usually handcrafted) which are robust to changes in imaging conditions, and *metric learning* [1, 6, 17, 14, 18, 29] searches for effective distance functions to compare features from different cameras. Robust features can be modeled by adopting perceptual principles of symmetry and asymmetry of the human body [7]. The correspondence problem can be approached by locating body parts [2, 4] and extracting local descriptors (color histograms [4], color invariants [15], covariances [2], CNN [21]). However, to find a proper descriptor, we need to look for a trade-off between its discriminative power and invariance between cameras. This task can be considered a *metric learning* problem that maximizes inter-class variation while minimizing intra-class variation.

Many different machine learning algorithms have been considered for learning a robust similarity function. Gray *et al*. employed Adaboost for feature selection and weighting [10], Prosser *et al*. defined the person re-identification as a ranking problem and used an ensemble of RankSVMs [23]. Recently features learned from deep convolution neural networks have been investigated [1, 17].

However, the most common choice for learning a metric remains the family of Mahalanobis distance functions. These include Large Margin Nearest Neighbor Learning (LMNN) [24], Information Theoretic Metric Learning (ITML) [5] and Logistic Discriminant Metric Learning (LDML) [11]. These methods usually aim at improving k-nn classification by iteratively adapting the metric. In contrast to these iterative methods, Köstinger [14] proposed the KISS metric which uses a statistical inference based on a likelihood-ratio test of two Gaussian distributions modeling positive and negative pairwise differences between features. Owing to its effectiveness and efficiency, the KISS metric is a popular baseline that has been extended to linear [19, 22] and non-linear [21, 25] subspace embeddings.

All of these approaches learn a Mahalanobis distance function for feature vectors extracted from bounding box images. Instead, we propose to learn dissimilarity functions for patches within bounding boxes, and then combine their scores into a robust distance measure. We show that our approach has clear advantages over existing algorithms.

3. Method

The Mahalanobis metric measures the squared distance between $\mathbf{x}_i$ and $\mathbf{x}_j$

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors extracted from bounding boxes taken from different cameras and $\mathbf{M}$ is a matrix encoding the basis for the comparison. $\mathbf{M}$ is usually learned in two stages: dimensionality reduction is first applied on $\mathbf{x}_i$ and $\mathbf{x}_j$ (*e.g*. principle component analysis - PCA), and then metric learning (*e.g*. KISS metric [14]) is performed on the reduced subspace. To avoid overfitting, the dimensionality must be significantly reduced to keep the number of free parameters low [12, 19]. In practice, $\mathbf{x}_i$ and $\mathbf{x}_j$ are high dimensional feature vectors and their reduced dimensionality is usually about three orders of magnitude smaller than the original [14, 19, 21]. Such strong dimensionality reduction might lose discriminative information, especially in case of misaligned features in $\mathbf{x}_i$ and $\mathbf{x}_j$ (*e.g*. highlighted patches in Fig. 1).

We propose to learn a metric for matching patches within the bounding box. We perform dimensionality reduction on features extracted from each patch. The reduced dimensionality is usually only one order of magnitude smaller than the original one, thus keeping more information (see Section 4).

3.1. Patch-based Metric Learning (PML)

We divide bounding box i into a dense grid with overlapping rectangular patches. From each patch location k, color and texture descriptors (*e.g*. color and gradient histograms) are extracted and concatenated into the patch feature vector $\mathbf{p}_i^k$. We represent bounding box image i as an ordered set of patch features $\mathcal{X}_i = \{\mathbf{p}_i^1, \mathbf{p}_i^2, \ldots, \mathbf{p}_i^K\}$, where K is the number of patches. Usually in standard metric learning approaches, these patch descriptors are concatenated into a single high dimensional feature vector [14, 20, 21, 22]. Instead, we learn a dissimilarity function for feature vectors extracted from patches. We define the dissimilarity measure as

$$\Phi(\mathbf{p}_i^k, \mathbf{p}_j^k) = (\mathbf{p}_i^k - \mathbf{p}_j^k)^T \mathbb{M}^{(k)} (\mathbf{p}_i^k - \mathbf{p}_j^k), \tag{2}$$

where $\mathbf{p}_i^k$ and $\mathbf{p}_j^k$ are the feature vectors extracted from patches at location k in bounding boxes i and j, from different cameras. Although, it is possible to learn one metric for each patch location k, this might be too many degrees of freedom. In practice, multiple patch locations might share a common metric, and in the extreme case a single $\mathbb{M}$ could be learned for all patch locations. We investigated re-identification performance with different numbers of patch

metrics $\mathbb{M}$ (see Section 4.2.2) and found that in some cases multiple $\mathbb{M}$'s might perform better than a single $\mathbb{M}$. Regions with statistically different amounts of background noise should have different metrics (*e.g.* patches close to the head contain more background noise than patches close to the torso). However, we also found that the recognition performance is a function of available training data (see Section 4.2.2), which limits the number of patch metrics that can be learned efficiently.

Learning $\mathbb{M}^{(k)}$: Given pairs of sample bounding boxes (i, j) we introduce the space of pairwise differences $\mathbf{p}_{ij}^k = \mathbf{p}_i^k - \mathbf{p}_j^k$ and partition the training data into $\mathbf{p}_{ij}^{k+}$ when i and j are bounding boxes containing the same person and $\mathbf{p}_{ij}^{k-}$ otherwise. Note that for learning we use differences on patches from the same location k.

To learn $\mathbb{M}^{(k)}$ we follow Köstinger [14] and assume a zero mean Gaussian structure on difference space and employ a log likelihood ratio test. This results in

$$\mathbb{M}^{(k)} = \Sigma_{k+}^{-1} - \Sigma_{k-}^{-1}, \tag{3}$$

where Σ_{k+} and Σ_{k-} are the covariance matrices of $\mathbf{p}_{ij}^{k+}$ and $\mathbf{p}_{ij}^{k-}$ respectively

$$\Sigma_{k+} = \sum(\mathbf{p}_{ij}^{k+})(\mathbf{p}_{ij}^{k+})^T, \tag{4}$$

$$\Sigma_{k-} = \sum(\mathbf{p}_{ij}^{k-})(\mathbf{p}_{ij}^{k-})^T. \tag{5}$$

Our dissimilarity score between patches is

$$\Phi(\mathbf{p}_i^k, \mathbf{p}_j^k) = (\mathbf{p}_{ij}^k)^T(\Sigma_{k+}^{-1} - \Sigma_{k-}^{-1})(\mathbf{p}_{ij}^k). \tag{6}$$

To compute the dissimilarity between two bounding boxes i and j, we combine patch dissimilarity scores by summing over all patches $\sum_{k=1}^{K} \Phi(\mathbf{p}_i^k, \mathbf{p}_j^k)$. This is equivalent to learning a block diagonal matrix

$$\left[\mathbf{p}_{ij}^1, \mathbf{p}_{ij}^2, \ldots, \mathbf{p}_{ij}^K\right] \begin{bmatrix} \mathbb{M}^1 & 0 & \ldots & 0 \\ 0 & \mathbb{M}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \mathbb{M}^K \end{bmatrix} \begin{bmatrix} \mathbf{p}_{ij}^1 \\ \mathbf{p}_{ij}^2 \\ \vdots \\ \mathbf{p}_{ij}^K \end{bmatrix} \tag{7}$$

where all $\mathbb{M}^{(k)}$ are learned independently. We refer to this formulation as **PML**.

A pair of bounding boxes corresponds to a single training example in the standard approach. Breaking a bounding box into a set of patches increases the amount of training data if a reduced number of $\mathbb{M}$ is learned (*e.g.* some locations k share the same metric). When a single $\mathbb{M}$ is learned, the amount of training data increases by combining $\mathbf{p}_{ij}^{k+}$ for all K locations into set $\mathbf{p}_{ij}^{+} = \sum_{k=1}^{K} |\mathbf{p}_{ij}^{k+}|$. In experiments we show that this can significantly boost performance when the training dataset is small (*e.g.* iLIDS dataset).

3.2. Deformable Model (DPML)

Pose changes and different camera viewpoints make re-identification more difficult. To overcome this issue we allow patches in one bounding box to perturb their locations (deform) when matching to another bounding box. We employ a model which approximates continuous non-affine warps by translating 2D templates [8, 26] (see Fig. 1). We use a spring model to limit the displacement of patches.

We define the deformable dissimilarity score for matching the patch at location k in bounding box i with bounding box j as

$$\psi(\mathbf{p}_i^k, j) = \min_l \left[\Phi(\mathbf{p}_i^k, \mathbf{p}_j^l) + \alpha_k \Delta(k, l)\right], \tag{8}$$

where patch feature $\mathbf{p}_j^l$ is extracted from bounding box j at location l.

Appearance term $\Phi(\mathbf{p}_i^k, \mathbf{p}_j^l)$ computes the feature dissimilarity between patches and is learned by employing our previously introduced **PML** (see Section 3.1).

Deformation cost $\alpha_k \Delta(k, l)$ refers to a spring model that controls the relative placement of patches k and l. $\Delta(k, l)$ is the squared distance between the patch locations. α_k encodes the rigidity of the spring: $\alpha_k = \infty$ corresponds to a rigid model, while $\alpha_k = 0$ allows a patch to change its location freely.

We combine the deformable dissimilarity scores $\psi(\mathbf{p}_i^k, j)$ into a unified dissimilarity measure

$$\begin{aligned} \Psi(i, j) &= \sum_{k=1}^{K} w_k \psi(\mathbf{p}_i^k, j) \\ &= \langle \mathbf{w}, \psi_{ij} \rangle, \end{aligned} \tag{9}$$

where $\mathbf{w}$ is a vector of weights and ψ_{ij} corresponds to a vector of patch dissimilarity scores.

Learning α_k and $\mathbf{w}$: Similarly to [21], we define the optimization problem as a relative distance comparison of triplets $\{i, j, z\}$ such that $\Psi(i, z) > \Psi(i, j)$ for all i, j, z; where i and j correspond to bounding boxes extracted from different cameras containing the same person, and i and z are bounding boxes from different cameras containing different people. Unfortunately, Eq. 8 is non-convex and we can not guarantee avoiding local minima. In practice, we use a limited number of unique spring constants α_k and apply two-step optimization. First, we optimize α_k with $\mathbf{w} = \mathbf{1}$, by performing exhaustive grid search (see Section 4.3) while maximizing Rank-1 recognition rate. Second, we fix α_k and determine the best $\mathbf{w}$ using structural SVMs [13]. This approach is referred to as **DPML**.

Figure 2: Sample images from **VIPeR** dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

Figure 3: Sample images from **i-LIDS** dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

4. Experiments

We carry out experiments on three challenging datasets: **VIPeR** [9], **i-LIDS** [28] and **CUHK01** [16]. The results are analyzed in terms of recognition rate, using the *cumulative matching characteristic* (CMC) [9] curve. The CMC curve represents the expectation of finding the correct match in the top r matches. The curve can be characterized by a scalar value computed by normalizing the area under the curve referred to as $nAUC$ value.

Section 4.1 describes the benchmark datasets used in the experiments. We explore our rigid patch metric model (PML) and its parameters in Section 4.2, then the deformable patch model (DPML) in Section 4.3. Finally, in Section 4.4, we compare our performance to other state of the art methods.

4.1. Datasets

VIPeR [9] is one of the most popular person re-identification datasets. It contains 632 image pairs of pedestrians captured by two outdoor cameras. VIPeR images contain large variations in lighting conditions, background, viewpoint, and image quality (see Fig. 2). Each bounding box is cropped and scaled to be 128×48 pixels. We follow the common evaluation protocol for this database: randomly dividing 632 image pairs into 316 image pairs for training and 316 image pairs for testing. We repeat this procedure 10 times and compute the average CMC curves for obtaining reliable statistics.

i-LIDS [28] consists of 119 individuals with 476 images. This dataset is very challenging since there are many occlusions. Often only the top part of the person is visible and usually there is a significant scale or viewpoint change as well (see Fig. 3). We follow the evaluation protocol of [21]: the dataset is randomly divided into 60 image pairs used for training and the remaining 59 image pairs are used for testing. This procedure is repeated 10 times for obtaining averaged CMC curves.

Figure 4: Sample images from **CUHK01** dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

CUHK01 [16] contains 971 persons captured with two cameras. For each person, 2 images for each camera are provided. The images in this dataset are better quality and higher resolution than in the two previous datasets. Each bounding box is scaled to be 160×60 pixels. The first camera captures the side view of pedestrians and the second camera captures the frontal view or the back view (see Fig. 4). We follow the single shot setting: randomly selecting 971 image pairs and randomly dividing it into 486 image pairs for training and 485 image pairs for testing. We repeat this procedure 10 times for computing averaged CMC curves.

4.2. Rigid Patch Metric Learning

In this section, we first compare our rigid patch model (PML) to the standard full bounding box approach (BBOX). BBOX is equivalent to the method presented in [14].

Each bounding box of size $w \times h$ is divided into a grid of $K = 60$ overlapping patches of size $\frac{w}{4} \times \frac{w}{2}$ and a 20×3 layout. Details of how a feature vector is extracted for each patch location are discussed in Section 4.2.1.

For the full bounding box case, we concatenate the extracted patch feature vectors into a high dimensional feature

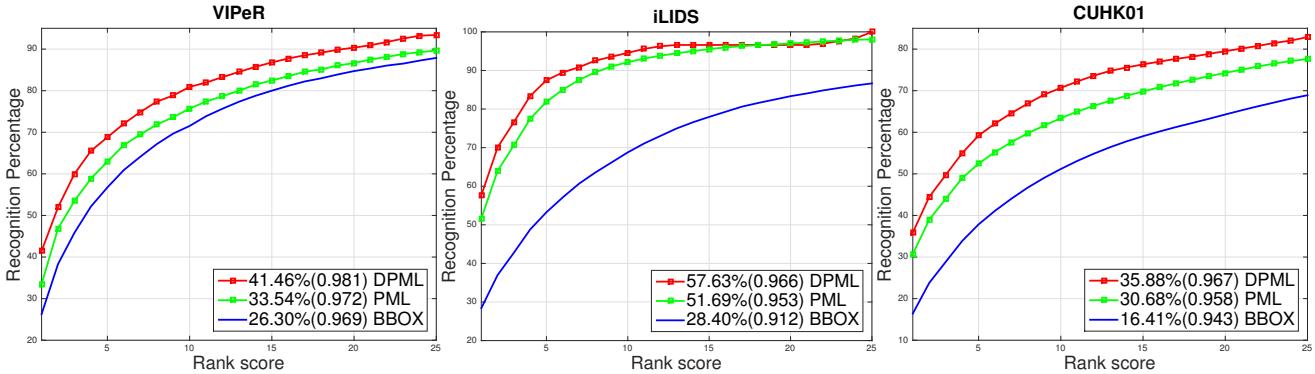


Figure 5: Performance comparison of Patch based Metric Learning (PML) and Deformable Patch based Metric Learning (DPML) *vs*. full bounding box metric learning (BBOX). Rank-1 identification rates as well as $nAUC$ values provided in brackets are shown in the legend next to the method name.

vector. PCA is applied to obtain a 62-dimensional feature space (where the optimal dimensionality is found by cross-validation). Then, the KISS metric [14] is learned in the 62-dimensional PCA subspace. For PML, instead of learning a metric for the concatenated feature vector, we learn metrics for patch features. In this way, we avoid undesirable compression. The dimensionality of the patch feature vector is reduced by PCA to 35 (also found by cross-validation) and metrics are learned independently for each patch location. Fig. 5 illustrates the comparison on three datasets. It is apparent that PML significantly improves the re-identification performance by keeping a higher number of degrees of freedom (35×60) when learning the dissimilarity function.

4.2.1 Patch Feature Vector

It is common practice in person re-identification to combine color and texture descriptors for describing an image. We evaluated the performance of different combinations of representations, including Lab, RGB and HSV histograms, each with 30 bins per channel. Texture information was captured by color SIFT, which is the SIFT descriptor extracted for each Lab channel and then concatenated. Fig. 6 illustrates the averaged CMC curves for VIPeR data set. The most informative color space is Lab, and the best performance is achieved by combining Lab, HSV and color SIFT. We use this representation in all experiments.

4.2.2 Patch Metrics

As mentioned earlier, our formulation allows one $\mathbb{M}$ to be learned per patch. In practice, there may be insufficient training data for this many degrees of freedom. We evaluate two extremes: learning 60 independent metrics (one per patch) and learning a single metric for all 60 patches

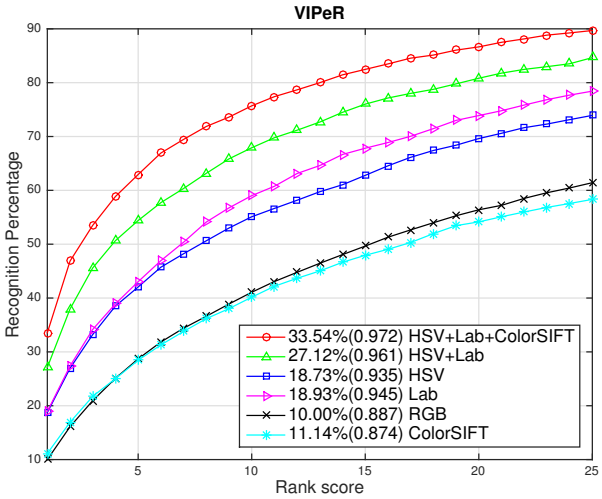


Figure 6: Performance comparison of different patch descriptors for VIPeR dataset.

(see Fig. 7). The results indicate that multiple metrics lead to significantly better recognition accuracy.

To understand the variability in the learned metrics, we setup the following experiment: learn a metric for a particular location k, and then apply this metric to compute dissimilarity scores for all other patch locations. We plot $nAUC$ values w.r.t. to the location of the learned metric in Fig. 8(a). It is apparent that metrics learned at different locations yield different performances. Surprisingly, higher performance is obtained by metrics learned on patches at lower locations within the bounding box (corresponding to leg regions). We believe that it is due to significant number of images in the VIPeR dataset having dark and cluttered backgrounds in the upper regions (see the last 3 top images in Fig. 2). Lower parts of the bounding boxes usually have

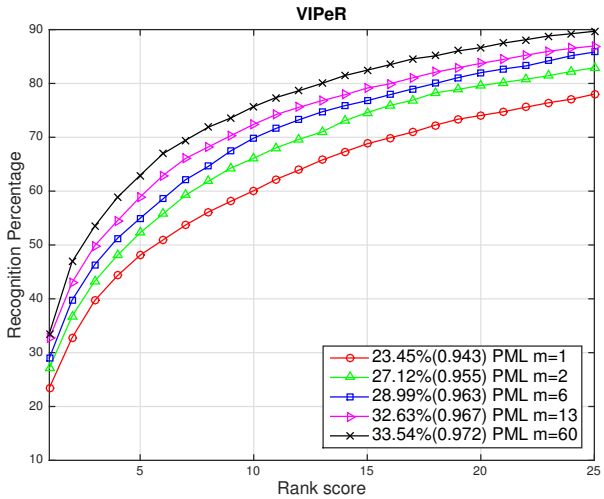
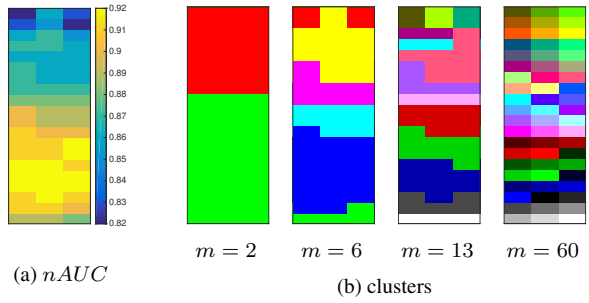


Figure 7: Performance comparison w.r.t. the number of $\mathbb{M}$. Using different metrics for different image regions yields better performance.



(a) $nAUC$ (b) clusters

Figure 8: Dividing image regions into several metrics $\mathbb{M}$. (a) $nAUC$ values w.r.t. a location of a learned metric; (b) clustering results for different number of clusters m.

more coherent background from sidewalks.

Additionally, we cluster patch locations spatially using hierarchical clustering (bottom-up), where similarity between regions is computed using $nAUC$ values. Fig. 8(b) illustrates clustering results w.r.t. to the number of clusters. Next, we learn metrics for each cluster of patch locations. These metrics are then used for computing patch similarity in corresponding image regions. Recall from Fig. 7 that the best performance was achieved with $m = 60$. In this circumstance, there appears to be sufficient data to train an independent metric for each patch location. We test this hypothesis by reducing the amount of training data and evaluating the optimal number of patch metrics when fewer training examples are available. Fig. 9 illustrates that the patch-based approach achieves high performance much faster than full bounding box metric learning. Interestingly, for a small number of positive pairs (less than 100), a reduced number of metrics gives better performance. When a common metric is learned for multiple patch locations, the amount of training data is effectively increased because features from multiple patches can be used as examples for learning the same metric (Section 3.1).

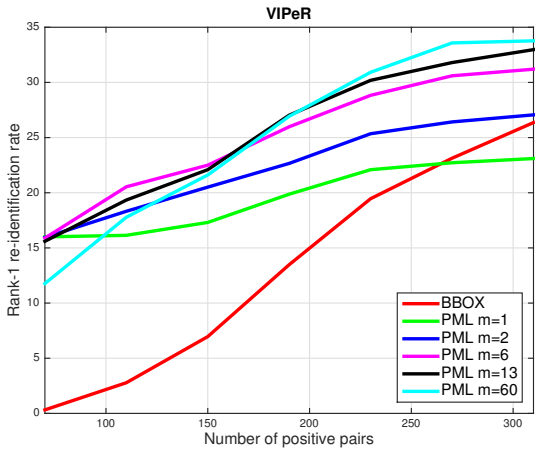


Figure 9: Rank-1 recognition rate with varying size of training dataset.

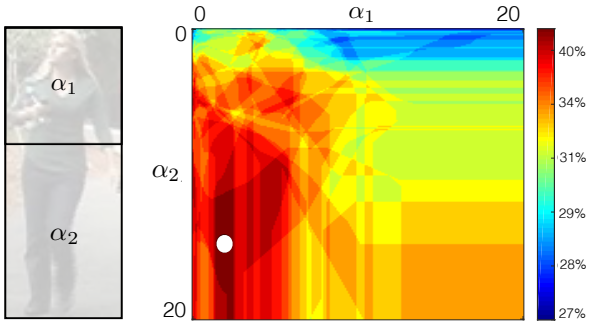


Figure 10: Exhaustive grid search over α_1 and α_2 coefficients for VIPeR. α_1 and α_2 correspond to patches locations w.r.t. to the left image. Grid search map illustrates Rank-1 recognition rate as a function of (α_1, α_2). The white dot highlights the optimal operating point.

4.3. Deformable Patch Metric Learning

We simplify Eq. 8 by restricting the number of unique spring constants. Two parameters α_1, α_2 are assigned to patch locations obtained by hierarchical clustering with the number of clusters $m = 2$ (see Fig. 10). α_k encodes the rigidity of the patches at particular locations. We perform an exhaustive grid search iterating through α_1 and α_2 while maximizing Rank-1 recognition rate. Fig 10 illustrates the recognition rate map as a function of both coefficients. Interestingly, rigidity (high spring constants) is useful for lower patches (the dark red region in the left-bottom corner of the map) but not so for patches in the upper lo-

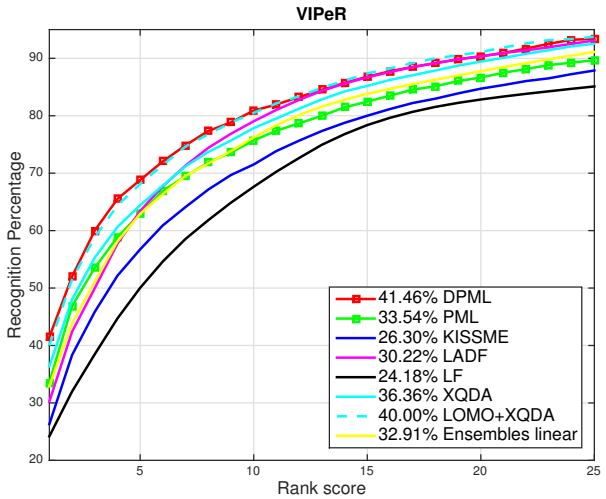


Figure 11: CMC curves and Rank-1 re-identification rates for VIPeR dataset. Comparison with state of the art approaches.

METHOD	$r = 1$	$r = 10$	$r = 20$
DPML	**41.46%**	**80.90%**	**90.46%**
PML	**33.54%**	**75.70%**	**86.65%**
KISSME[14]	19.60%	62.20%	84.72%
LF[22]	24.18%	67.12%	82.00%
LADF[18]	30.22%	78.92%	90.44%
Ensembles[21]	32.91%	76.65%	87.76%
XQDA[19]	36.36%	77.84%	89.43%
LOMO+XQDA[19]	40.00%	80.51%	**91.08%**
kLDFA[25]*	32.8%	79.10%	90.00%
MidLevel[27]*	29.11%	65.95%	79.87%
MidLevel+LADF[27]*	**43.39%**	**84.05%**	**92.37%**
Ensembles[21]*	**45.89%**	**88.90%**	**95.80%**
KernelMap[3]*	36.80%	83.70%	**91.70%**
DeepNN[1]*	34.81%	76.40%	-

Table 1: Comparison with state of the art on VIPeR dataset. * corresponds to non-linear models. Competitive results are highlighted in **bold**.

cations of the bounding box. This might be related to the fact that metrics learned on the lower locations have higher performance (compare with $nAUC$ values in Fig. 8). Fig. 5 clearly shows that introducing our deformable model improves the recognition accuracy in all datasets.

4.4. Comparison with Other Methods

VIPeR The performance of PML and DPML on VIPeR relative to other state of the art approaches is reported in Fig. 11 and Table 1. Fig. 11 illustrates that our DPML achieves the new state of the art among linear models. From

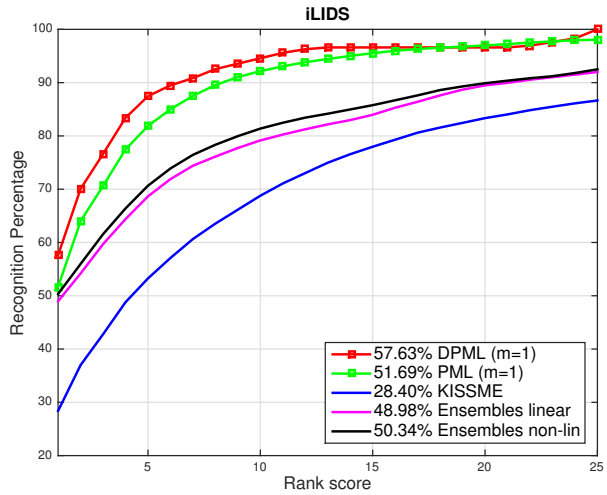


Figure 12: CMC curves for i-LIDS dataset. Our approaches outperform all existing person re-identification algorithms by the large margin.

METHOD	$r = 1$	$r = 10$	$r = 20$
DPML	**57.63%**	**95.61%**	**96.21%**
PML	**51.69%**	**92.49%**	**96.80%**
KISSME[14]	28.40%	68.90%	83.40%
PRDC[29]	37.83%	75.09%	88.35%
Ensembles[21]	48.98%	79.00%	89.00%
kLDFA[25]*	40.30%	78.10%	89.60%
Ensembles[21]*	50.34%	81.00%	90.00%

Table 2: Comparison with state of the art on i-LIDS dataset. * corresponds to non-linear models.

Table 1, we can observe that our approach outperforms all non-linear models except for non-linear ensembles [21] and a fusion of MidLevel filters [27] with LADF [18]. We are primarily concerned with the performance of linear models, as these are practical for deploying on large databases. Non-linear methods may be more accurate, but can be slower at making comparisons, which is a significant obstacle when deploying for camera networks.

i-LIDS This dataset contains a relatively small number of training samples (we use only 60 image pairs for training). Driven by our previous analysis (Section 4.2.2), we learn a single $\mathbb{M}$ for all patches, thus increasing the training set. As a result, PML and DPML significantly outperform full bounding box based metric learning methods (see Fig. 12). There are three aspects that make our approach more effective: (1) we are able to generate a significantly larger training set using $m = 1$, (2) occlusions in images pollute only a few patch scores in our similarity measure, while in

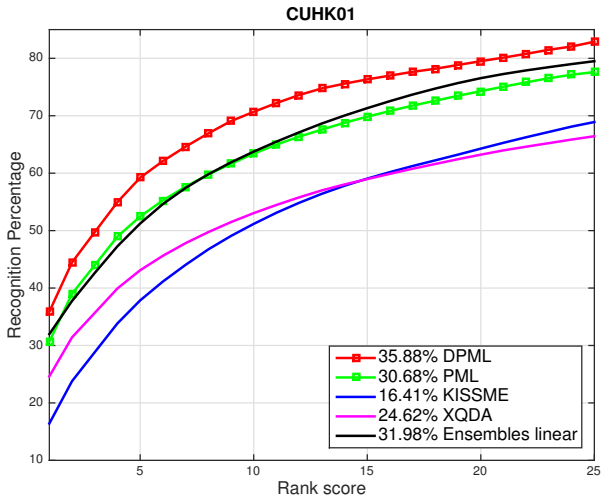


Figure 13: CMC curves for CUHK01 dataset. Comparison with state of the art approaches.

METHOD	$r = 1$	$r = 10$	$r = 20$
DPML	**35.88%**	**70.92%**	**79.51%**
PML	**30.68%**	**63.40%**	**74.28%**
KISSME[14]	16.41%	51.48%	64.29%
XQDA[19]	24.62%	53.07%	63.34%
Ensembles[21]	31.98%	63.77%	76.68%
MidLevel[27]*	34.30%	66.50%	76.00%
Ensembles[21]*	53.40%	84.40%	90.50%

Table 3: Comparison with state of the art on CUHK01 dataset. * corresponds to non-linear models.

case of full-image based metric learning they might have global impact on the final dissimilarity measure, (3) misaligned features can be corrected by our deformable model. DPML outperforms the second best one, Ensembles[21], by 7.29% in the first rank and by 14.61% in the tenth rank. Table 2 summarizes the comparison.

CUHK01 This dataset contains better quality and higher resolution images, thus it is not surprising that keeping a higher number of degrees of freedom improves the re-identification performance. Fig. 13 and Table 3 illustrate that PML and DPML achieve the new state of the art performance. PML and DPML outperform KISSME in the first rank by 14.27% and 19.47%, respectively. DPML achieves the best re-identification performance among all algorithms except non-linear ensembles [21]. In our experiments we followed a single shot setting and trained our models using only 486 image pairs. It is not clear whether this procedure is the same as the training method used in [21].

5. Summary

Re-identification must deal with appearance differences arising from changes in illumination, viewpoint and a person's pose. Traditional metric learning approaches do not address registration errors and instead only focus on feature vectors extracted from bounding boxes. In contrast, we propose a patch-based approach. Operating on patches has several advantages:

- Extracted feature vectors have lower dimensionality and do not have to be subject to the same levels of compression as feature vectors extracted for the entire bounding box.
- Multiple patch locations can share the same metric, which effectively increase the amount of training data.
- We allow patches to adjust their locations when comparing two bounding boxes. The idea is similar to part-based models used in object detection. As a result, we directly address registration errors while simultaneously evaluating appearance consistency.

Our experiments illustrate how these advantages lead to state of the art performance on well established, challenging re-identification datasets.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. 2015.

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010.

[3] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015.

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, pages 68.1–68.11, 2011.

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

[6] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, pages 501–512, 2010.

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

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010.

[9] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *PETS*, 2007.

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

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

[12] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, 2010.

[13] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009.

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

[15] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *TPAMI*, 2013.

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

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

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

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[20] N. Martinel, C. Micheloni, and G. Foresti. Saliency weighted features for person re-identification. In *ECCV Workshops*, 2014.

[21] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

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

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

[24] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[25] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[26] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *TPAMI*, 2013.

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

[28] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.

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