# Scalable Metric Learning via Weighted Approximate Rank Component Analysis

Cijo Jose François Fleuret

Idiap Research Institute
École Polytechnique Fédérale de Lausanne
{cijo.jose, francois.fleuret}@idiap.ch

**Abstract.** We are interested in the large-scale learning of Mahalanobis distances, with a particular focus on person re-identification. We propose a metric learning formulation called Weighted Approximate Rank Component Analysis (WARCA). WARCA optimizes the precision at top ranks by combining the WARP loss with a regularizer that favors orthonormal linear mappings and avoids rank-deficient embeddings. Using this new regularizer allows us to adapt the large-scale WSABIE procedure and to leverage the Adam stochastic optimization algorithm, which results in an algorithm that scales gracefully to very large data-sets. Also, we derive a kernelized version which allows to take advantage of state-of-the-art features for re-identification when data-set size permits kernel computation. Benchmarks on recent and standard re-identification data-sets show that our method beats existing state-of-the-art techniques both in terms of accuracy and speed. We also provide experimental analysis to shade lights on the properties of the regularizer we use, and how it improves performance.

**Keywords:** metric learning, orthonormal regularizer, person re-identification

## 1 Introduction

Metric learning methods aim at learning a parametrized distance function from a labeled set of samples, so that under the learned distance, samples with the same labels are nearby and samples with different labels are far apart [1]. Many fundamental questions in computer vision such as "How to compare two images? and for what information?" boil down to this problem. Among them, person re-identification is the problem of recognizing individuals at different physical locations and times, on images captured by different devices.

It is a challenging problem which recently received a lot of attention because of its importance in various application domains such as video surveillance, biometrics, and behavior analysis [2].

The performance of person re-identification systems relies mainly on the image feature representation and the distance measure used to compare them. Hence the research in the field has focused either on designing features [3, 4] or on learning a distance function from a labeled set of images [5–8, 4, 9].

It is difficult to analytically design features that are invariant to the various non-linear transformations that an image undergoes such as illumination, viewpoint, pose changes, and occlusion. Furthermore, even if such features were provided, the standard Euclidean metric would not be adequate as it does not take into account dependencies on the feature representation. This motivates the use of metric learning for person re-identification.

Re-identification models are commonly evaluated by the cumulative match characteristic (CMC) curve [6]. This measure indicates how the matching performance of the algorithm improves as the number of returned image increases. Given a matching algorithm and a labeled test set, each image is compared against all the others, and the position of the first correct match is recorded. The CMC curve indicates for each rank the fraction of test samples which had that rank or better. A perfect CMC curve would reach the value 1 for rank #1, that is the best match is always of the correct identity.

In this paper we are interested in learning a Mahalanobis distance by minimizing a weighted rank loss such that the precision at the top rank positions of the CMC curve is maximized. When learning the metric, we directly learn the low-rank projection matrix instead of the PSD matrix because of the computational efficiency and the scalability to high dimensional datasets (see § 3.1). But naively learning the low-rank projection matrix suffers from the problem of matrix rank degeneration and non-isolated minima [10]. We address this problem by using a simple regularizer which approximately enforces the orthonormality of the learned matrix very efficiently (see § 3.2). We extend the WARP loss [11, 12, 10] and combine it with our approximate orthonormal regularizer to derive a metric learning algorithm which approximately minimizes a weighted rank loss efficiently using stochastic gradient descent (see § 3.3).

We extend our model to kernel space to handle distance measures which are more natural for the features we are dealing with (see § 3.4). We also show that in kernel space SGD can be carried out more efficiently by using preconditioning [13, 5].

We validate our approach on nine person re-identification datasets: Market-1501 [14], CUHK03 [15], OpeReid [16], CUHK01 [17], VIPeR [18], CAVIAR [3], 3DPeS [19], iLIDS [20] and PRI450s [21], where we outperform other metric learning methods proposed in the literature, both in speed and accuracy.

## 2 Related Works

Metric learning is a well studied research problem [22]. Most of the existing approaches have been developed in the context of the Mahalanobis distance learning paradigm [23, 1, 24, 5, 6]. This consists in learning distances of the form:

$$\mathcal{D}^2_M(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j), \tag{1}$$

where $M$ is a positive semi-definite matrix. Based on the way the problem is formulated the algorithms for learning such distances involve either optimization

in the space of positive semi-definite (PSD) matrices, or learning the projection matrix $W$, in which case $M = W^T W$.

Large margin nearest neighbors [1] (LMNN) is a metric learning algorithm designed to maximize the performance of $k$-nearest neighbor classification in a large margin framework. Information theoretic metric learning [24] (ITML) exploits the relationship between the Mahalanobis distance and Gaussian distributions to learn the metric. Many researchers have applied LMNN and ITML to re-identification problem with varying degree of success [21].

Pairwise Constrained Component Analysis (PCCA) [5] is a metric learning method that learns the low rank projection matrix $W$ in the kernel space from sparse pairwise constraints. Xiong *et al.* [8] extended PCCA with a $L_2$ regularization term and showed that it further improves the performance.

Köstinger *et al.* [6] proposed the KISS ("Keep It Simple and Straight forward") metric learning abbreviated as KISSME. Their method enjoys very fast training and they show good empirical performance and scaling properties along the number samples. However this method suffers from of the Gaussian assumptions on the model.

Li *et al.* [7] consider learning a local thresholding rule for metric learning. This method is computationally expensive to train, even with as few as 100 dimensions.

The performance of many kernel-based metric learning methods for person re-identification was evaluated in [8]. In particular the authors evaluated PCCA [5], variants of kernel Fisher discriminant analysis (KFDA) and reported that the KFDA variants consistently out-perform all other methods. The KFDA variants they investigated were Local Fisher Discriminant Analysis (LFDA) and Marginal Fisher Discriminant Analysis (MFA).

Chen *et al.* [25] attempt to learn a metric in the polynomial feature map exploiting the relationship between Mahalanobis metric and the polynomial features. Ahmed *et al.* [26] propose a deep learning model which learns the features as well as the metric jointly. Liao *et al.* [4] propose XQDA exploiting the benefits of Fisher discriminant analysis and KISSME to learn a metric. However like FDA and KISSME, XQDA's modeling power is limited because of the Gaussian assumptions on the data. In another work Liao *et al.* [9] apply accelerated proximal gradient descent (APGD) to a Mahalanobis metric under a logistic loss similar to the loss of PCCA [5]. The application of APGD makes this model converge fast compared to existing batch metric learning algorithms but still it suffers from scalability issues because all the pairs are required to take one gradient step and the projection step on to the PSD cone is computationally expensive.

None of the above mentioned techniques explicitly models the objective that we are looking for in person re-identification, that is to optimize a weighted rank measure. We show that modeling this in the metric learning objective improves the performance. We address scalability through stochastic gradient descent (SGD) and our model naturally eliminates the need for asymmetric sample weighting as we use triplet based loss function.

There is an extensive body of work on optimizing ranking measures such as AUC, precision at $k$, $F_1$ score, etc. Most of this work focuses on learning a linear decision boundary in the original input space, or in the feature space for ranking a list of items based on the chosen performance measure. A well known such model is the structural SVM [27]. In contrast here we are interested in ranking pairs of items by learning a metric. A related work by McFee *et al.* [28] studies metric learning with different rank measures in the structural SVM framework. Wu *et al.* [29] used this framework to do person re-identification by optimizing the mean reciprocal rank criterion. Outside the direct scope of metric learning from a single feature representation, Paisitkriangkrai *et al.* [30] developed an ensemble algorithm to combine different base metrics in the structural SVM framework which leads to excellent performance for re-identification. Such an approach is complementary to ours, as combining heterogeneous feature representations requires a separate additional level of normalization or the combination with a voting scheme.

We use the WARP loss from WSABIE [12], proposed for large-scale image annotation problem, that is a multi-label classification problem. WSABIE learns a low dimensional joint embedding for both images and annotations by optimizing the WARP loss. This work reports excellent empirical results in terms of accuracy, computational efficiency, and memory footprint.

The work that is closely related to us is FRML [10] where they learn a Mahalanobis metric by optimizing the WARP loss function with SGD. However there are some key differences with our approach. FRML is a linear method using $L_2$ or LMNN regularizer, and relies on an expensive projection step in the SGD. Beside, this projection requires to keep a record of all the gradients in the mini-batch, which results in high memory footprint. The rationale for the projection step is to accelerate the SGD because directly optimizing low rank matrix may result in rank deficient matrix and thus result in non-isolated minimizers which might generalize poorly to unseen samples. We propose a computationally cheap solution to this problem by using a regularizer which approximately enforces the rank of the learned matrix efficiently.

**Table 1:** Notation

| | |
|---|---|
| $N$ | Number of training samples |
| $D$ | Dimension of training samples |
| $Q$ | Number of classes |
| $(x_i, y_i) \in \mathbb{R}^D \times \{1, \dots, Q\}$ | $i$-th training sample |
| $\mathbb{1}_{\text{condition}}$ | is equal to 1 if the condition is true, 0 otherwise |
| $\mathcal{S}$ | the pairs of indices of samples of same class |
| $\mathcal{T}_y$ | the indices of samples not of class $y$ |
| $\mathcal{F}_W$ | distance function under the linear map $W$ |
| $rank_{i,j}(\mathcal{F}_W)$ | for $i$ and $j$ of same label, no. of miss-labeled points closer to $i$ than $j$ is |
| $\mathcal{L}(W)$ | the loss we minimize |
| $L(r)$ | rank weighting function |

## 3 Weighted Approximate Rank Component Analysis

This section presents our metric learning algorithm, Weighted Approximate Rank Component Analysis (WARCA). Table 1 summarizes some important notations that we use in the paper.

Let us consider a training set of data point / label pairs:

$$(x_n, y_n) \in \mathbb{R}^D \times \{1, \dots, Q\},\ n = 1, \dots, N. \tag{2}$$

and let $\mathcal{S}$ be the set of pairs of indices of samples of same labels:

$$\mathcal{S} = \left\{(i,j) \in \{1,\dots,N\}^2,\ y_i = y_j\right\}. \tag{3}$$

For each label $y$ we define the set $\mathcal{T}_y$ of indices of samples of a class different from $y$:

$$\mathcal{T}_y = \{k \in \{1,\dots,N\},\ y_k \neq y\}. \tag{4}$$

In particular, to each $(i,j) \in \mathcal{S}$ corresponds a set $\mathcal{T}_{y_i}$.

Let $W$ be a linear transformation that maps the data points from $\mathbb{R}^D$ to $\mathbb{R}^{D'}$, with $D' \leq D$. For the ease of notation, we do not distinguish between matrices and their corresponding linear mappings. The distance function under the linear map $W$ is given by:

$$\mathcal{F}_W(x_i, x_j) = \|W(x_i - x_j)\|_2. \tag{5}$$

### 3.1 Problem Formulation

For a pair of points $(i,j)$ of same label $y_i = y_j$, we define a ranking error function:

$$\forall (i,j) \in \mathcal{S},\ err(\mathcal{F}_W, i, j) = L\left(rank_{i,j}\left(\mathcal{F}_W\right)\right) \tag{6}$$

where:

$$rank_{i,j}\left(\mathcal{F}_W\right) = \sum_{k \in \mathcal{T}_{y_i}} \mathbb{1}_{\mathcal{F}_W(x_i,x_k) \leq \mathcal{F}_W(x_i,x_j)}. \tag{7}$$

is the number of samples $x_k$ of different labels which are closer to $x_i$ than $x_j$ is.

Formulating our objective that way, following closely the formalism of [12], shows how training a multi-class predictor shares similarities with our metric-learning problem. The former aims at avoiding, for any given sample to have incorrect classes with responses higher than the correct one, while the latter aims at avoiding, for any pair of samples $(x_i, x_j)$ of the same label, to have samples $x_k$ of other classes in between them.

Minimizing directly the rank treats all the rank positions equally, and usually in many problems including person re-identification we are interested in maximizing the correct match within the top few rank positions. This can be achieved by a weighting function $L(\cdot)$ which penalizes more a drop in the rank at the top positions than at the bottom positions. In particular we use the rank weighting function proposed by Usunier *et al.* [11], of the form:

$$L(r) = \sum_{s=1}^{r} \alpha_s,\ \alpha_1 \geq \alpha_2 \geq \dots \geq 0. \tag{8}$$

For example, using $\alpha_1 = \alpha_2 = ... = \alpha_m$ will treat all rank positions equally, and using higher values of $\alpha$s in top few rank positions will weight top rank positions more. We use the harmonic weighting, which has such a profile and was also used in [12] as it yielded state-of-the-art results on their application.

Finally, we would like to solve the following optimization problem:

$$\underset{W}{\operatorname{argmin}} \ \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} L\left(rank_{i,j}\left(\mathcal{F}_W\right)\right). \tag{9}$$

### 3.2 Approximate OrthoNormal (AON) Regularizer

The optimization problem of Equation 9 may lead to severe over-fitting on small and medium scale datasets. Regularizing penalty terms are central in re-identification for that reason.

The standard way of regularizing a low-rank metric learning objective function is by using a $L_2$ penalty, such as the Frobenius norm [10]. However, such a regularizer tends to push toward rank-deficient linear mappings, which we observe in practice (see § 4.4, and in particular Figure 2a).

Lim *et al.* [10] in their FRML algorithm, addresses this problem by using a Riemannian manifold update step in their SGD algorithm, which is computationally expensive and induces a high memory footprint. We propose an alternative approach that maintains the rank of the matrix by pushing toward orthonormal matrices. This is achieved by using as a penalty term the $L_2$ divergence of $WW^T$ from the identity matrix $\mathbf{I}$ :

$$\|WW^T - \mathbf{I}\|^2. \tag{10}$$

This orthonormal regularizer can also be seen as a strategy to mimic the behavior of approaches such as PCA or FDA, which ensure that the learned linear transformation is orthonormal. For such methods, this property emerges from the strong Gaussian prior over the data, which is beneficial on small data-sets but degrades performance on large ones where it leads to under-fitting. Controlling the orthonormality of the learned mapping through a regularizer weighted by a meta-parameter $\lambda$ allows us to adapt it on each data-set individually through cross-validation.

Finally, with this regularizer the optimization problem of Equation 9 becomes:

$$\underset{W}{\operatorname{argmin}} \ \frac{\lambda}{2}\|WW^T - \mathbf{I}\|^2 + \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} L\left(rank_{i,j}\left(\mathcal{F}_W\right)\right). \tag{11}$$

### 3.3 Max-Margin Reformulation

The metric learning problem in Equation 11 aims at minimizing the 0-1 loss, which is a difficult optimization problem. Applying the reasoning behind the WARP loss to make it tractable, we upper-bound this loss with the hinge one with margin $\gamma$. This is equivalent to minimizing the following loss function:

$$\mathcal{L}(W) = \frac{\lambda}{2}\|WW^T - \mathbf{I}\|^2 + \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\sum_{k\in\mathcal{T}_{y_i}} L(rank^{\gamma}_{i,j}(\mathcal{F}_W))\frac{|\gamma + \xi_{ijk}|_+}{rank^{\gamma}_{i,j}(\mathcal{F}_W)}, \quad (12)$$

where:

$$\xi_{ijk} = \mathcal{F}_W(x_i, x_j) - \mathcal{F}_W(x_i, x_k) \quad (13)$$

and $rank^{\gamma}_{i,j}(\mathcal{F}_W)$ is the margin penalized rank:

$$rank^{\gamma}_{i,j}(\mathcal{F}_W) = \sum_{k\in\mathcal{T}_{y_i}} \mathbb{1}_{\gamma+\xi_{ijk}>0}. \quad (14)$$

The loss function in Equation 12 is the WARP loss [11, 12, 10]. It was shown by Weston *et al.* [12] that the WARP loss can be efficiently solved by using stochastic gradient descent and we follow the same approach:

1. Sample $(i, j)$ uniformly at random from $\mathcal{S}$.
2. For the selected $(i, j)$ uniformly sample $k$ in $\{k \in \mathcal{T}_{y_i} : \gamma + \xi_{ijk} > 0\}$, *i.e.* from the set of incorrect matches scored higher than the correct match $x_j$.

The sampled triplet $(i, j, k)$ has a contribution of $L(rank^{\gamma}_{i,j}(\mathcal{F}_W))|\gamma + \xi_{ijk}|_+$ because the probability of drawing a $k$ in step 2 from the violating set is $\frac{1}{rank^{\gamma}_{i,j}(\mathcal{F}_W)}$.

We use the above sampling procedure to solve WARCA efficiently using mini-batch stochastic gradient descent (SGD). We use Adam SGD algorithm [31], which is found to converge faster empirically compared to vanilla SGD.

### 3.4 Kernelization

Most commonly used features in person re-identification are histogram-based such as LBP, SIFT BOW, RGB histograms to name a few. The most natural distance measure for histogram-based features is the $\chi^2$ distance. Most of the standard metric learning methods work on the Euclidean distance with PCCA being a notable exception. To plug any arbitrary metric which is suitable for the features, such as $\chi^2$, one has to resort to explicit feature maps that approximate the $\chi^2$ metric. However, it blows up the dimension and the computational cost. Another way to deal with this problem is to do metric learning in the kernel space, which is the approach we follow.

Let $W$ be spanned by the samples:

$$W = AX^T = A\begin{pmatrix} x_1^T \\ \dots \\ x_N^T \end{pmatrix}. \quad (15)$$

which leads to:

$$\mathcal{F}_A(x_i, x_j) = \|AX^T(x_i - x_j)\|_2, \quad (16)$$
$$= \|A(\kappa_i - \kappa_j)\|_2. \quad (17)$$

Where $\kappa_i$ is the $i^{th}$ column of the kernel matrix $K = X^T X$. Then the loss function in Equation 12 becomes:

$$\mathcal{L}(A) = \frac{\lambda}{2}\|AKA^T - \mathbf{I}\|^2 + \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\sum_{k\in\mathcal{T}_{y_i}} L(rank^{\gamma}_{i,j}(\mathcal{F}_A))\frac{|\gamma + \xi_{ijk}|_+}{rank^{\gamma}_{i,j}(\mathcal{F}_A)}, \quad (18)$$

with:

$$\xi_{ijk} = \mathcal{F}_A(x_i, x_j) - \mathcal{F}_A(x_i, x_k). \quad (19)$$

Apart from being able to do non-linear metric learning, kernelized WARCA can be solved efficiently again by using stochastic sub-gradient descent. If we use the inverse of the kernel matrix as the pre-conditioner of the stochastic sub-gradient, the computation of the update equation, as well the parameter update, can be carried out efficiently. Mignon *et al.* [5] used the same technique to solve their PCCA, and showed that it converges faster than vanilla gradient descent. We use the same technique to derive an efficient update rule for our kernelized WARCA. A stochastic sub-gradient of Equation 18 with the sampling procedure described in the previous section is given as:

$$\nabla\mathcal{L}(A) = 2\lambda(AKA^T - \mathbf{I})AK + 2L(rank^{\gamma}_{i,j}(\mathcal{F}_A))A\mathbb{1}_{\gamma+\xi_{ijk}>0}\mathcal{G}_{ijk}, \quad (20)$$

where:

$$\mathcal{G}_{ijk} = \frac{(\kappa_i - \kappa_j)(\kappa_i - \kappa_j)^T}{d_{ij}} - \frac{(\kappa_i - \kappa_k)(\kappa_i - \kappa_k)^T}{d_{ik}}, \quad (21)$$

and:

$$d_{ij} = \mathcal{F}_A(x_i, x_j), \quad d_{ik} = \mathcal{F}_A(x_i, x_k). \quad (22)$$

Multiplying the right hand side of Equation 20 by $K^{-1}$:

$$\nabla\mathcal{L}(A)K^{-1} = 2\lambda(AKA^T - \mathbf{I})A + 2L(rank^{\gamma}_{i,j}(\mathcal{F}_A))AK\mathbb{1}_{\gamma+\xi_{ijk}>0}\mathcal{E}_{ijk}. \quad (23)$$

with:

$$\mathcal{E}_{ijk} = K^{-1}\mathcal{G}_{ijk}K^{-1} = \frac{(e_i - e_j)(e_i - e_j)^T}{d_{ij}} - \frac{(e_i - e_k)(e_i - e_k)^T}{d_{ik}}. \quad (24)$$

where $e_l$ is the $l^{th}$ column of the canonical basis that is the vector whose $l^{th}$ component is one and all others are zero. In the preconditioned stochastic sub-gradient descent we use the updates of the form:

$$A_{t+1} = (\mathbf{I} - 2\lambda\eta(A_t K A_t^T - \mathbf{I}))A_t - 2\eta L(rank^{\gamma}_{i,j}(\mathcal{F}_A))A_t K\mathbb{1}_{\gamma+\xi_{ijk}>0}\mathcal{E}_{ijk}. \quad (25)$$

Please note that $\mathcal{E}_{ijk}$ is a very sparse matrix with only nine non-zero entries. This makes the update extremely fast. Preconditioning also enjoys faster convergence rates since it exploits second order information through the preconditioning operator, here the inverse of the kernel matrix [13].

# 4 Experiments

We evaluate our proposed algorithm on nine standard person re-identification datasets. We first describe the datasets and baseline algorithms and then present our results. Our code will be made publicly available.

## 4.1 Datasets and Baselines

The largest dataset we experimented with is the **Market-1501** dataset [14] which is composed of 32,668 images of 1,501 persons captured from 6 different view points. It uses DPM [32] detected bounding boxes as annotations. **CUHK03** dataset [15] consists of 13,164 images of 1,360 persons and it has both DPM detected and manually annotated bounding boxes. We use the manually annotated bouding boxes here. **OpeReid** dataset [16] consists of 7,413 images of 200 persons. **CUHK01** dataset [17] is composed of 3,884 images of 971 persons, with two pairs of images per person, each pair taken from a different viewpoint.**VIPeR** [18] dataset has 1,264 images of 632 person, with 2 images per person. The **PRID450s** dataset [21] consists of 450 image pairs recorded from two different static surveillance cameras. The **CAVIAR** dataset [3] consists of 1,220 images of 72 individuals from 2 cameras in a shopping mall. The **3DPeS** dataset [19] has 1,011 images of 192 individuals, with 2 to 6 images per person. The dataset is captured from 8 outdoor cameras with horizontal but significantly different viewpoints. Finally the **iLIDS** dataset [20] contains 476 images and 119 persons, with 2 to 8 images per individual.

We compare our method against the current state-of-the-art baselines MLAPG, rPCCA, SVMML, FRML, LFDA and KISSME. A brief overview of these methods is given in section 2. rPCCA, MLAPG, SVMML, FRML are iterative methods whereas LFDA and KISSME are spectral methods on the second order statistics of the data. Since WARCA, rPCCA and LFDA are kernel methods we used both the $\chi^2$ kernel and the linear kernel with them to benchmark the performance. Marginal Fisher discriminant analysis (MFA) is proven to give similar result as that of LFDA so we do not use them as the baseline.

We did not compare against other ranking based metric learning methods such as LORETA [33], OASIS [34] and MLR [28] because all of them are linear methods. In fact we derived a kernelized OASIS but the results were not as good as ours or rPCCA. We also do not compare against LMNN and ITML because many researchers have evaluated them before [5–7] and found out that they do not perform as well as other methods considered here.

## 4.2 Technical Details

For the Market-1501 dataset we used the experimental protocol and features described in [14]. We used their baseline code and features. As Market-1501 is quite large for kernel methods we do not evaluate them. We also do not evaluate the linear methods such as Linear rPCCA and SVMML because their optimization algorithms were found to be very slow.

**Table 2:** Table showing the rank 1, rank 5 and AUC performance measure of our method WARCA against other state-of-the-art methods. Bold fields indicate best performing methods. The dashes indicate computation that could not be run in a realistic setting on Market-1501

**(a)** Rank 1 accuracy

| Dataset | WARCA-$\chi^2$ | WARCA-L | rPCCA-$\chi^2$ | rPCCA-L | MLAPG | FRML | SVMML | LFDA-$\chi^2$ | LFDA-L | KISSME |
|---|---|---|---|---|---|---|---|---|---|---|
| Market-1501 | – | **45.16±0.00** | – | – | – | – | – | – | 34.65±0.00 | 42.81±0.00 |
| CUHK03 | **78.38±2.44** | 62.12±2.07 | 76.74±2.06 | 59.22±2.65 | 44.90±1.57 | 53.87±2.31 | 47.89±2.59 | 69.94±2.21 | 46.02±1.55 | 47.88±1.80 |
| CUHK01 | **58.34±1.26** | 39.30±0.76 | 48.55±1.12 | 34.73±1.06 | 22.92±0.94 | 33.58±0.69 | 27.96±0.86 | 54.25±1.04 | 33.74±0.73 | 35.74±0.95 |
| OpeReid | **57.65±1.60** | 43.74±1.34 | 52.89±1.78 | 43.66±1.45 | 40.63±1.31 | 42.27±1.35 | 30.63±1.51 | 53.58±1.65 | 42.84±1.18 | 41.76±1.36 |
| VIPeR | **37.47±1.70** | 20.86±1.04 | 22.25±1.91 | 15.91±1.16 | 19.49±2.26 | 18.52±0.78 | 23.28±1.53 | 36.77±2.10 | 20.22±1.85 | 20.89±1.22 |
| PRID450s | **24.58±1.75** | 10.33±1.20 | 16.35±1.30 | 8.34±1.25 | 2.13±0.59 | 7.05±1.60 | 13.08±1.63 | 24.31±1.44 | 3.24±0.95 | 15.24±1.56 |
| CAVIAR | **43.44±1.82** | 39.35±1.98 | 37.56±2.17 | 27.26±2.15 | 36.74±1.96 | 35.40±2.67 | 26.82±1.64 | 41.29±2.25 | 37.72±2.08 | 31.99±2.17 |
| 3DPeS | **51.89±2.27** | 43.57±2.18 | 46.42±2.25 | 33.12±1.58 | 41.17±2.26 | 39.03±1.85 | 29.94±2.10 | 51.44±1.40 | 43.24±2.57 | 37.55±1.80 |
| iLIDS | **36.61±2.40** | 31.77±2.77 | 26.57±2.60 | 23.07±3.07 | 31.13±1.57 | 25.68±2.25 | 21.32±2.89 | 36.23±1.89 | 32.70±3.12 | 28.29±3.59 |

**(b)** Rank 5 accuracy

| Dataset | WARCA-$\chi^2$ | WARCA-L | rPCCA-$\chi^2$ | rPCCA-L | MLAPG | FRML | SVMML | LFDA-$\chi^2$ | LFDA-L | KISSME |
|---|---|---|---|---|---|---|---|---|---|---|
| Market-1501 | – | **68.23±0.00** | – | – | – | – | – | – | 52.76±0.00 | 62.74±0.00 |
| CUHK03 | **94.55±1.31** | 86.03±1.62 | 94.50±1.29 | 84.52±1.41 | 71.80±1.52 | 80.36±1.22 | 79.97±2.08 | 90.15±1.27 | 65.41±1.66 | 69.29±2.35 |
| CUHK01 | **79.76±0.69** | 61.84±0.98 | 73.29±1.32 | 56.67±1.20 | 48.48±1.49 | 55.27±0.83 | 53.11±0.78 | 74.60±1.00 | 49.73±0.91 | 53.34±0.69 |
| OpeReid | **80.43±1.71** | 67.39±1.02 | 77.95±1.82 | 67.68±1.25 | 61.45±1.61 | 66.08±1.30 | 60.32±1.31 | 75.34±1.76 | 59.70±1.37 | 61.74±1.55 |
| VIPeR | **70.78±2.43** | 50.29±1.61 | 53.82±2.32 | 42.71±2.02 | 46.49±2.23 | 46.15±1.62 | 55.28±1.99 | 69.30±2.23 | 45.25±1.90 | 47.73±2.28 |
| PRID450s | **55.52±2.23** | 31.73±3.08 | 43.82±2.18 | 26.89±2.21 | 11.29±1.66 | 24.16±3.04 | 38.38±1.77 | 54.58±2.06 | 12.55±1.41 | 37.22±1.81 |
| CAVIAR | **74.06±3.13** | 68.06±2.44 | 70.62±2.26 | 57.44±2.48 | 65.83±2.73 | 66.24±3.08 | 61.53±3.64 | 69.12±3.02 | 61.60±2.94 | 61.17±3.21 |
| 3DPeS | **75.64±2.80** | 68.26±1.91 | 73.54±2.26 | 58.34±2.31 | 65.06±1.89 | 65.20±2.15 | 59.52±2.62 | 75.36±1.91 | 65.64±1.91 | 60.22±2.05 |
| iLIDS | **66.09±2.31** | 59.27±3.12 | 57.07±2.93 | 51.55±3.59 | 57.31±3.12 | 53.42±2.17 | 51.45±4.30 | 65.20±2.68 | 59.66±2.51 | 54.08±3.63 |

**(c)** AUC score

| Dataset | WARCA-$\chi^2$ | WARCA-L | rPCCA-$\chi^2$ | rPCCA-L | MLAPG | FRML | SVMML | LFDA-$\chi^2$ | LFDA-L | KISSME |
|---|---|---|---|---|---|---|---|---|---|---|
| Market-1501 | – | **75.41±0.00** | – | – | – | – | – | – | 60.53±0.00 | 70.02±0.00 |
| CUHK03 | **93.94±0.76** | 89.67±0.80 | 93.92±0.81 | 89.17±0.69 | 82.30±1.01 | 86.64±0.65 | 86.64±1.07 | 91.66±0.68 | 74.23±1.51 | 77.68±1.83 |
| CUHK01 | **84.99±0.65** | 71.88±0.67 | 81.00±0.88 | 67.56±0.93 | 62.84±1.51 | 66.39±0.76 | 65.73±1.07 | 80.84±0.80 | 58.92±1.08 | 62.36±0.95 |
| OpeReid | **86.47±1.08** | 77.17±0.94 | 85.25±1.16 | 77.42±1.01 | 72.34±1.11 | 76.51±0.88 | 73.88±1.04 | 82.67±1.30 | 68.96±1.53 | 71.33±1.14 |
| VIPeR | **81.87±1.07** | 67.00±1.11 | 71.30±1.50 | 62.40±1.43 | 64.71±1.15 | 64.19±1.39 | 71.04±1.63 | 81.34±1.21 | 62.67±1.35 | 64.74±1.20 |
| PRID450s | **72.13±1.49** | 50.07±2.25 | 63.10±2.16 | 46.19±1.89 | 30.81±2.19 | 42.97±2.84 | 59.54±1.25 | 71.55±1.70 | 28.18±1.22 | 53.83±1.86 |
| CAVIAR | **85.76±1.48** | 83.01±1.44 | 84.41±1.28 | 76.57±1.29 | 81.58±1.50 | 81.88±1.85 | 79.38±2.19 | 81.94±2.32 | 76.76±1.69 | 78.85±1.54 |
| 3DPeS | **83.89±1.53** | 78.07±1.57 | 82.84±1.44 | 72.27±1.96 | 75.98±1.28 | 76.89±1.44 | 73.38±1.70 | 83.49±0.95 | 75.87±1.49 | 72.22±1.31 |
| iLIDS | **79.04±1.60** | 73.42±1.96 | 74.10±2.04 | 69.60±2.44 | 72.45±1.99 | 71.26±1.55 | 70.25±2.09 | 78.98±1.43 | 74.26±2.02 | 70.33±2.90 |

All other evaluations where carried out in the single-shot experiment setting [2] and our experimental settings are very similar to the one adopted by Xiong *et al.* [8]. Except for Market-1501, we randomly divided all the other datasets into two subsets such that there are $p$ individuals in the test set. We created 10 such random splits. In each partition one image of each person was randomly selected as a probe image, and the rest of the images were used as gallery images and this was repeated 10 times. The position of the correct match was processed to generate the CMC curve. We followed the standard train-validation-test splits for all the other datasets and $P$ was chosen to be 100, 119, 486, 316, 225, 36, 95 and 60 for CUHK03, OpeReid, CUHK01, VIPeR, PRID450s, CAVIAR, 3DPeS and iLIDS respectively.

We used the same set of features for all the datasets except for the Market-1501 and all the features are essentially histogram based. First all the datasets were re-scaled to 128×48 resolution and then 16 bin color histograms on RGB, YUV, and HSV channels, as well as texture histogram based on Local Binary Patterns (LBP) were extracted on 6 non-overlapping horizontal patches. All the

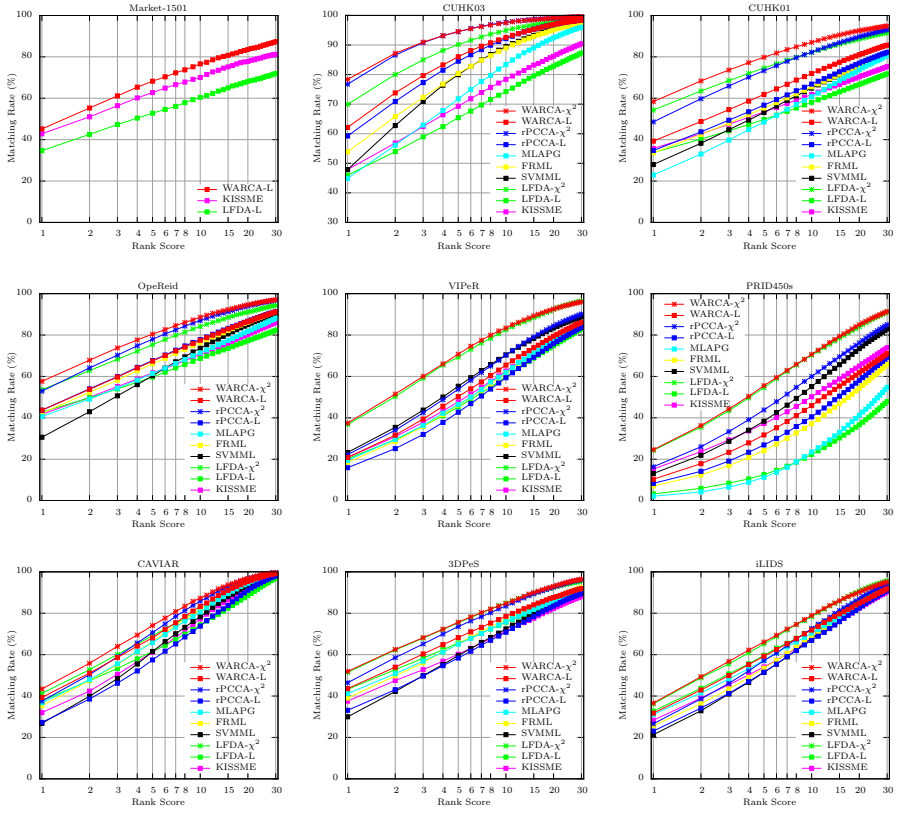


**Fig. 1:** CMC curves comparing WARCA against state-of-the-art methods on nine re-identification datasets

histograms are normalized per patch to have unit $L_1$ norm and concatenated into a single vector of dimension 2,580 [5, 8].

The source codes for LFDA, KISSME and SVMML are available from their respective authors website, and we used those to reproduce the baseline results [8]. The code for PCCA is not released publicly. A version from Xiong *et al.* [8] is available publicly but the memory footprint of that implementation is very high making it impossible to use with large datasets (e.g. it requires 17GB of RAM to run on the CAVIAR dataset). Therefore to reproduce the results in [8] we wrote our own implementation, which uses 30 times less memory and can scale to much larger datasets. We also ran sanity checks to make sure that it behaves the same as that of the baseline code. All the implementations were done in Matlab with mex functions for the acceleration of the critical components.

In order to fairly evaluate the algorithms, we set the dimensionality of the projected space to be same for WARCA, rPCCA and LFDA. For the Market-1501 dataset the dimensionality used is 200 and for VIPeR it is 100 and all

the other datasets it is 40. We choose the regularization parameter and the learning rate through cross-validation across the data splits using grid search in $(\lambda, \eta) \in \{10^{-8}, \ldots, 1\} \times \{10^{-3}, \ldots, 1\}$. Margin $\gamma$ is fixed to 1. Since the size of the parameter matrix scales in $O(D^2)$ for SVMML and KISSME we first reduced the dimension of the original features using PCA keeping 95% of the original variance and then applied these algorithms. In our tables and figures WARCA$-\chi^2$, WARCA-L, rPCCA$-\chi^2$, rPCCA-L, LFDA$-\chi^2$ and LFDA-L denote WARCA with $\chi^2$ kernel, WARCA with linear kernel, rPCCA with $\chi^2$ kernel, rPCCA with linear kernel, and LFDA with $\chi^2$ kernel, LFDA with linear kernel respectively.

For all experiments with WARCA we used harmonic weighting for the rank weighting function of Equation 8. We also tried uniform weighting which gave poor results compared to the harmonic weighting. For all the datasets we used a mini-batch size of 512 in the SGD algorithm and we ran the SGD for 2000 iterations (A parameter update using the mini-batch is considered as 1 iteration).

Tables 2a and 2b summarize respectively the rank-1 and rank-5 performance of all the methods, and Table 2c summarizes the Area Under the Curve (AUC) performance score. Figure 1 reports the CMC curves comparing WARCA against the baselines on all the nine datasets. The circle and the star markers denote linear and kernel methods respectively.

WARCA improves over all other methods on all the datasets. On VIPeR, 3DPeS, PRID450s and iLIDS datasets LFDA come very close to the performance of WARCA. The reason for this is that these datasets are too small and consequently simple methods such as LFDA which exploits strong prior assumptions on the data distribution work nearly as well as WARCA.

### 4.3 Comparison against State-of-the-art

We also compare against the state-of-the-art results reported using recent algorithms such as MLAPG on LOMO features [9], MLPOLY [25] and IDEEP [26] on VIPeR, CUHK01 and CUHK03 datasets. The reason for not including these comparisons in the main results is because apart from MLAPG the code for other methods is not available, or the features are different which makes a fair comparison difficult. Our goal is to evaluate experimentally that, given a set of features, which is the best off-the-shelf metric learning algorithm for re-identification.

In this set of experiments we used the state-of-the-art LOMO features [4] with WARCA for VIPeR and CUHK01 datasets. The results are summarized in the Table 3. We improve the rank1 performance by 21% on CUHK03 by 1.40% on CUHK01 dataset.

**Table 3:** Comparison of WARCA against state-of-the-art results for person re-identification

| Dataset | WARCA(Ours) | | | | MLAPG [9] | | | | MLPOLY [25] | | | | IDEEP [26] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | rank=1 | rank=5 | rank=10 | rank=20 | rank=1 | rank=5 | rank=10 | rank=20 | rank=1 | rank=5 | rank=10 | rank=20 | rank=1 | rank=5 | rank=10 | rank=20 |
| VIPeR | 40.22 | 68.16 | 80.70 | 91.14 | **40.73** | **69.94** | **82.34** | **92.37** | 36.80 | 70.40 | 83.70 | 91.70 | 34.81 | 63.61 | 75.63 | 84.49 |
| CUHK01 | **65.64** | 85.34 | 90.48 | **95.04** | 64.24 | **85.41** | **90.84** | 94.92 | - | - | - | - | 47.53 | 71.60 | 80.25 | 87.45 |
| CUHK03 | **78.38** | **94.5** | **97.52** | **99.11** | 57.96 | 87.09 | 94.74 | 98.00 | - | - | - | - | 54.74 | 86.50 | 94.02 | 97.02 |

### 4.4 Analysis of the AON regularizer

Here we present an empirical analysis of the AON regularizer against the standard Frobenius norm regularizer. We used the VIPeR dataset with LOMO features for the experiments shown in the first row of Figure 2. With very low regularization strength AON and Frobenius behave the same. As the regularization strength increases, Frobenius results in rank deficient mappings (Figure 2a), which is less discriminant and perform poorly on the test set (Figure 2b). The AON regularizer on the contrary pushes towards orthonormal mappings, and results in an embedding well conditioned, which generalizes well to the test set. It is also worth noting that training with the AON regularizer is robust over a wide range of the regularization parameter, which is not the case the Frobenius norm. Finally, the AON regularizer was found to be very robust to the choice of

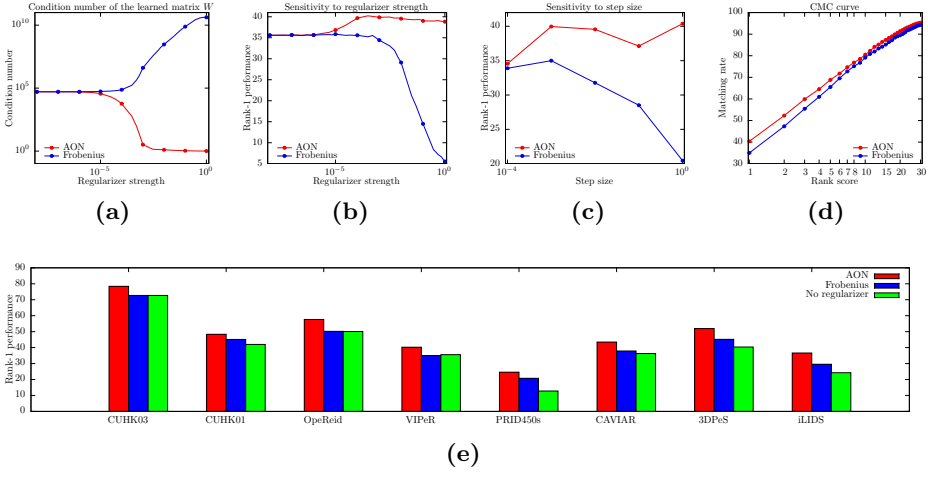


**Fig. 2:** Comparison of the Approximate OrthoNormal (AON) regularizer we use in our algorithm to the standard Frobenius norm ($L_2$) regularizer. Graph (a) shows the condition number (ratio between the two extreme eigenvalues of the learned mapping) vs. the weight $\lambda$ of the regularization term. As expected, the AON regularizer pushes this value to one, as it eventually forces the learning to chose an orthonormal transformation, while the Frobenius regularizer eventually kills the smallest eigenvalues to zero, making the ratio extremely large. Graph (b) shows the Rank-1 performance vs. the regularizer weight $\lambda$, graph (c) the Rank-1 performance vs. the SGD step size $\eta$, graph (d) CMC curve with the two regularizers and finally graph (e) shows the Rank-1 performance on different datasets

the SGD step size $\eta$ (Figure 2c) which is a crucial parameter in large-scale learning. A similar behavior was observed by Lim *et al.* [10] with their orthonormal

Riemannian gradient update step in the SGD but it is computationally expensive and not trivial to use with modern SGD algorithms such as Adam [31], and Nesterov's momentum [35].

### 4.5 Analysis of the Training Time

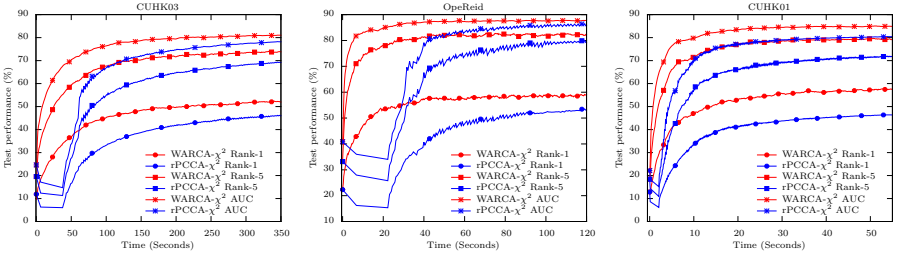


**Fig. 3:** WARCA performs significantly better than the state-of-the-art rPCCA on large datasets for a given training time budget

Figure 3 illustrates how the performance in test of WARCA and rPCCA increase as a function of training time on 3 datasets. We implemented both the algorithms entirely in C++ to have a fair comparison of running times. In this set of experiments we used 730 test identities for CUHK03 dataset to have a quick evaluation. Experiments with other datasets follow the same protocol described above. Please note that we do not include spectral methods in this plot because the solutions are found analytically. Linear spectral methods are very fast for low dimensional problems but the training time scales quadratically in the data dimension. In case of kernel spectral methods the training time scales quadratically in the number of data points. We also do not include iterative methods MLAPG and SVMML because they proved to be very slow and not giving good performance.

## 5 Conclusion

We have proposed a simple and scalable approach to metric learning that combines a new and simple regularizer to a proxy for a weighted sum of the precision at different ranks. The later can be used for any weighting of the precision-at-$k$ metrics. Experimental results show that it outperforms state-of-the-art methods on standard person re-identification datasets, and that contrary to most of the current state-of-the-art methods, it allows for large-scale learning.

**Acknowledgements** – This work was supported by the Swiss National Science Foundation under grant number, CRSII2-147693 WILDTRACK.

## References


1. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research **10** (2009) 207–244
2. Gong, S., Cristani, M., Yan, S., Loy, eds.: Person Re-Identification. Advances in Computer Vision and Pattern Recognition. Springer (2014)
3. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference (BMVC). (2011)
4. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2197–2206
5. Mignon, A., Jurie, F.: Pcca: A new approach for distance learning from sparse pairwise constraints. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2666–2672
6. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2288–2295
7. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3610–3617
8. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Computer Vision–ECCV 2014. Springer (2014) 1–16
9. Liao, S., Li, S.Z.: Efficient psd constrained asymmetric metric learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3685–3693
10. Lim, D., Lanckriet, G.: Efficient learning of mahalanobis metrics for ranking. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). (2014) 1980–1988
11. Usunier, N., Buffoni, D., Gallinari, P.: Ranking with ordered weighted pairwise classification. In: Proceedings of the 26th annual International Conference on Machine Learning (ICML-09), ACM (2009) 1057–1064
12. Weston, J., Bengio, S., Usunier, N.: WSABIE: Scaling up to large vocabulary image annotation. In: IJCAI. Volume 11. (2011) 2764–2770
13. Chapelle, O.: Training a support vector machine in the primal. Neural Computation **19**(5) (2007) 1155–1178
14. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Bu, J., Tian, Q.: Scalable person re-identification: A benchmark. Computer Vision, IEEE International Conference on (2015)
15. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 152–159
16. Liao, S., Mo, Z., Hu, Y., Li, S.Z.: Open-set person re-identification. arXiv preprint arXiv:1408.0872 (2014)
17. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: ACCV (1). (2012) 31–44
18. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Computer Vision–ECCV 2008. Springer (2008) 262–275

19. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, ACM (2011) 59–64
20. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference (BMVC). Volume 2. (2009) 6
21. Roth, P.M., Hirzer, M., Köstinger, M., Beleznai, C., Bischof, H.: Mahalanobis distance learning for person re-identification. In: Person Re-Identification. Springer (2014) 247–267
22. Yang, L., Jin, R.: Distance metric learning: A comprehensive survey. Michigan State Universiy **2** (2006)
23. Xing, E.P., Jordan, M.I., Russell, S., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: Advances in neural information processing systems. (2002) 505–512
24. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the 24th annual International Conference on Machine Learning (ICML-07), ACM (2007) 209–216
25. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1565–1573
26. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3908–3916
27. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: Proceedings of the 21st annual International Conference on Machine Learning (ICML-04), ACM (2004) 104
28. McFee, B., Lanckriet, G.R.: Metric learning to rank. In: Proceedings of the 27th annual International Conference on Machine Learning (ICML-10). (2010) 775–782
29. Wu, Y., Mukunoki, M., Funatomi, T., Minoh, M., Lao, S.: Optimizing mean reciprocal rank for person re-identification. In: Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on, IEEE (2011) 408–413
30. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
31. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
33. Shalit, U., Weinshall, D., Chechik, G.: Online learning in the embedded manifold of low-rank matrices. The Journal of Machine Learning Research **13**(1) (2012) 429–458
34. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. The Journal of Machine Learning Research **11** (2010) 1109–1135
35. Sutskever, I.: Training recurrent neural networks. PhD thesis, University of Toronto (2013)