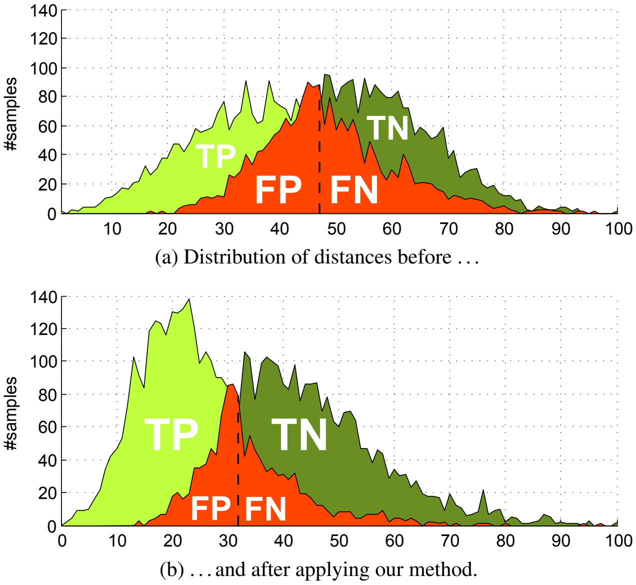
Large Scale Metric Learning from Equivalence Constraints *

Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology
{koestinger,hirzer,wohlhart,pmroth,bischof}@icg.tugraz.at

Abstract

In this paper, we raise important issues on scalability and the required degree of supervision of existing Mahalanobis metric learning methods. Often rather tedious optimization procedures are applied that become computationally intractable on a large scale. Further, if one considers the constantly growing amount of data it is often infeasible to specify fully supervised labels for all data points. Instead, it is easier to specify labels in form of equivalence constraints. We introduce a simple though effective strategy to learn a distance metric from equivalence constraints, based on a statistical inference perspective. In contrast to existing methods we do not rely on complex optimization problems requiring computationally expensive iterations. Hence, our method is orders of magnitudes faster than comparable methods. Results on a variety of challenging benchmarks with rather diverse nature demonstrate the power of our method. These include faces in unconstrained environments, matching before unseen object instances and person re-identification across spatially disjoint cameras. In the latter two benchmarks we clearly outperform the state-of-the-art.



(a) Distribution of distances before . . .

(b) . . . and after applying our method.

Figure 1: **Face verification on LFW [12]:** The challenging task shows the benefit of metric learning. Our method significantly increases the TPR at EER from 67.4% (a) to 80.5% (b). Training takes only **0.05** seconds and is thus orders of magnitudes faster than related methods.

1. Introduction

Learning distance or similarity metrics is an emerging field in machine learning, with various applications in computer vision. It can significantly improve results for tracking [22], image retrieval [11], face identification [9], clustering [21], or person re-identification [4]. The goal of metric learning algorithms is to take advantage of prior information in form of labels over simpler though more general similarity measures, illustrated in Figure 1.

A particular class of distance functions that exhibits good generalization performance for many machine learning problems is Mahalanobis metric learning. The goal is to find a global, linear transformation of the feature space such that relevant dimensions are emphasized while irrelevant ones are discarded. As there exists a bijection between the set of Mahalanobis metrics and the set of multivariate Gaussians one can think of it in terms of the corresponding covariance matrix. The metric adapts to the desired geometry by arbitrary linear rotations and scalings. After projection the plain Euclidean distance is measured.

Machine learning algorithms that learn a Mahalanobis metric have recently attracted a lot of interest in computer vision. These include Large Margin Nearest Neighbor Learning (LMNN) [19, 20], Information Theoretic Metric Learning (ITML) [3] and Logistic Discriminant Metric Learning (LDML) [8], which can be considered as state-of-the-art. LMNN [19, 20] aims at improving k-nn clas-

*The work was supported by the Austrian Science Foundation (FWF) project Advanced Learning for Tracking and Detection in Medical Workflow Analysis (I535-N23) and by the Austrian Research Promotion Agency (FFG) project SHARE in the IV2Splus program.

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

sification. It establishes for each instance a local perimeter. The perimeter surrounds the k-nns with similar label (target neighbors), plus a margin. To reduce the amount of instances with dissimilar label that invade the perimeter (impostors) the metric is iteratively adapted. This is done by strengthening the correlation to target neighbors while weakening it to impostors. Conceptually sound, LMNN is sometimes prone to over-fitting due to the lack of regularization. Davis *et al.* [3] avoid over-fitting by explicitly integrating a regularization step. Their formulation trades off between satisfying the given constraints on the distance function while minimizing the differential entropy to the initial prior distance metric distribution. Guillaumin *et al.* [8] introduce a probabilistic view on learning a Mahalanobis metric where the a posteriori class probabilities are treated as (dis)similarity measures. Thus, they propose to iteratively adapt the Mahalanobis metric to maximize the log-likelihood. The a posteriori probability is modeled by a sigmoid function that reflects the fact that instances share labels if their distance is below a certain threshold. In principle, all of these methods are able to generalize well to unseen data. They focus on robust loss functions and regularize solutions to avoid over-fitting.

Considering the ever growing amount of data, learning a Mahalanobis metric on a large scale dataset raises further issues on scalability and the required degree of supervision. Often it is infeasible to specify fully supervised labels for all data points. Instead, it is easier to specify labels in form of equivalence constraints. In some cases it is even possible to obtain this form of weak supervision automatically, *e.g.*, by tracking an object. Hence, to capitalize on large scale datasets one faces the additional challenges of scalability and the ability to deal with equivalence constraints.

To meet these requirements, we learn an effective metric just based on equivalence constraints. These are considered as natural inputs to distance metric learning algorithms as similarity functions basically establish a relation between pairs of points. Our method is motivated by a statistical inference perspective based on a likelihood-ratio test. We show that the resulting metric is not prone to over-fitting and very efficient to obtain. Compared to other approaches we do not rely on a tedious iterative optimization procedure. Therefore, our method is scalable to large datasets, as it just involves computation of two small sized covariance matrices. As analog to the KISS principle (*keep it simple and straightforward!*) we keep our method easy and efficient per design and will thus refer to it as *KISS metric*.

We demonstrate our method on various different benchmarks where we match or even outperform state-of-the-art metric learning approaches, while being orders of magnitudes faster in training. In particular, we provide results on two recent face recognition benchmarks (LFW [12], PubFig[13]). Due to the non-rigid nature and changes in pose, lighting and expression faces are a challenge for learning algorithms. Further, we study the task of person re-identification across spatially disjoint cameras (VIPeR [6]) and the comparison of before never seen object instances on ToyCars [15]. On VIPeR and the ToyCars dataset we improve even over the domain specific state-of-the-art. Further, for LFW we obtain the best reported results for standard SIFT features.

The rest of this paper is organized as follows. In Section 2 we discuss related metric learning approaches that motivate our approach. Succeeding, in Section 3 we introduce our KISS metric learning approach. Extensive experiments and evaluations on performance and scalability are conducted in Section 4. Finally, we conclude and summarize the paper in Section 5.

2. Learning a Mahalanobis Metric

Learning a distance or similarity metric based on the class of Mahalanobis distance functions has gained considerable interest in computer vision. In general, a Mahalanobis distance metric measures the squared distance between two data points $\mathbf{x}_i$ and $\mathbf{x}_j$:

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

where $\mathbf{M} \succeq 0$ is a positive semidefinite matrix and $\mathbf{x}_i$,$\mathbf{x}_j \in \mathbb{R}^d$ is a pair of samples (i, j). Further, for the following discussion we introduce a similarity label y_{ij}: $y_{ij} = 1$ for similar pairs, *i.e.*, if the samples share the same class label $(y_i = y_j)$ and $y_{ij} = 0$ otherwise. To motivate our approach, we give in the following an overview of the state-of-the-art in learning a Mahalanobis metric. In particular, we examine LMNN [19, 20], ITML [3] and LDML [8].

2.1. Large Margin Nearest Neighbor Metric

The approach of Weinberger *et al.* [19, 20] aims at improving k-nn classification by exploiting the local structure of the data. For each instance a local perimeter surrounding the k nearest neighbors sharing the same label (target neighbors) is established. Samples having a different label that invade this perimeter (impostors) are penalized. This is realized via the following objective function:

$$\epsilon(\mathbf{M}) = \sum_{j \rightsquigarrow i} \left[d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) + \mu \sum_l (1 - y_{il}) \xi_{ijl}(\mathbf{M}) \right]. \tag{2}$$

The first term minimizes the distance between target neighbors $\mathbf{x}_i$,$\mathbf{x}_j$, indicated by $j \rightsquigarrow i$. The second term denotes the amount by which impostors invade the perimeter of i and j. An impostor l is a differently labeled input $(y_{il} = 0)$ that has a positive slack variable $\xi_{ijl}(\mathbf{M}) \geq 0$:

$$\xi_{ijl}(\mathbf{M}) = 1 + d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) - d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l). \tag{3}$$

To estimate $\mathbf{M}$, gradient descent is performed along the gradient defined by the triplets (i, j, l) having positive slack:

$$\frac{\partial \epsilon(\mathbf{M}^t)}{\partial \mathbf{M}^t} = \sum_{j \rightsquigarrow i} \mathbf{C}_{ij} + \mu \sum_{(i,j,l)} (\mathbf{C}_{ij} - \mathbf{C}_{il}) \ , \tag{4}$$

where $\mathbf{C}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$ denotes the outer product of pairwise differences. Conceptually, for active triplets this formulation strengthens the correlation to target neighbors while weakening it to impostors.

2.2. Information Theoretic Metric Learning

Davis *et al.* [3] exploit the relationship between multivariate Gaussian distributions and the set of Mahalanobis distances. The idea is to search for a solution that trades off the satisfaction of constraints while being close to a distance metric prior $\mathbf{M}_0$, *e.g.*, the identity matrix for the Euclidean distance. The closeness of the solution to the prior is measured by the Kullback-Leibler divergence of the corresponding distributions. The prior can be considered as a regularization term to avoid over-fitting. The constraints enforce that similar pairs are below a certain distance $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq u$ while dissimilar pairs exceed a certain distance $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq l$. The optimization builds on Bregman projections [1], which project the current solution onto a single constraint via the update rule:

$$\mathbf{M}_{t+1} = \mathbf{M}_t + \beta \mathbf{M}_t \mathbf{C}_{ij} \mathbf{M}_t \ . \tag{5}$$

The parameter β involves the pair label and the step size. It is positive for similar pairs and negative for dissimilar pairs. Thus, for similar pairs the optimization is performed in direction of $\mathbf{C}_{ij}$ while for dissimilar pairs in the negative direction.

2.3. Linear Discriminant Metric Learning

Guillaumin *et al.* [8] offer a probabilistic view on learning a Mahalanobis distance metric. The a posteriori class probabilities are treated as (dis)similarity measures, whether a pair of images depicts the same object. For a given pair (i, j) the a posteriori probability is modeled as

$$p_{ij} = p(y_{ij} = 1 | \mathbf{x}_i, \mathbf{x}_j; \mathbf{M}, b) = \sigma(b - d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)) \ , \tag{6}$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is a sigmoid function and b is a bias term. Thus, to estimate $\mathbf{M}$ the Mahalanobis metric is iteratively adapted to maximize the log-likelihood:

$$L(\mathbf{M}) = \sum_{ij} y_{ij} \ln(p_{ij}) + (1 - y_{ij}) \ln(1 - p_{ij}). \tag{7}$$

The maximization by gradient ascent is obtained in direction of $\mathbf{C}_{ij}$ for similar pairs and in the negative direction for dissimilar pairs:

$$\frac{\partial L(\mathbf{M})}{\partial \mathbf{M}} = \sum_{ij} (y_{ij} - p_{ij}) \mathbf{C}_{ij} \ . \tag{8}$$

The influence of each pair on the gradient direction is controlled over the probability.

If we recapitulate the properties and characteristics of the described metric learning approaches we observe two commonalities. First, all methods rely on an iterative optimization scheme which can be computationally expensive for large scale datasets. Second, if we compare the update rules of the different methods, given in Eqs. (4), (5) and (8), we can see that the optimization is performed in direction of $\mathbf{C}_{ij}$ for similar pairs and in the negative direction of $\mathbf{C}_{ij}$ for dissimilar pairs. In the following, we introduce a non-iterative formulation, which builds on a statistical inference perspective of the space of pairwise differences. This allows us to face the additional challenges of scalability and the ability to learn from equivalence constraints. Our parameter-free approach is very efficient in training, enabling to exploit the constantly growing amount of data also for learning.

3. KISS Metric Learning

Our method considers two independent generation processes for observed commonalities of similar and dissimilar pairs. The dissimilarity is defined by the plausibility of belonging either to one or the other. From a statistical inference point of view the optimal statistical decision whether a pair (i, j) is dissimilar or not can be obtained by a likelihood ratio test. Thus, we test the hypothesis H_0 that a pair is dissimilar versus the alternative H_1:

$$\delta(\mathbf{x}_i, \mathbf{x}_j) = \log \left(\frac{p(\mathbf{x}_i, \mathbf{x}_j | H_0)}{p(\mathbf{x}_i, \mathbf{x}_j | H_1)} \right) \ . \tag{9}$$

A high value of $\delta(\mathbf{x}_i, \mathbf{x}_j)$ means that H_0 is validated. In contrast, a low value means that H_0 is rejected and the pair is considered as similar. To be independent of the actual locality in the feature space, we cast the problem in the space of pairwise differences ($\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$) with zero mean and can re-write Eq. (9) to

$$\delta(\mathbf{x}_{ij}) = \log \left(\frac{p(\mathbf{x}_{ij} | H_0)}{p(\mathbf{x}_{ij} | H_1)} \right) = \log \left(\frac{f(\mathbf{x}_{ij} | \theta_0)}{f(\mathbf{x}_{ij} | \theta_1)} \right) \ . \tag{10}$$

Whereby $f(\mathbf{x}_{ij} | \theta_1)$ is a pdf with parameters θ_1 for hypothesis H_1 that a pair (i, j) is similar ($y_{ij} = 1$) and vice-versa H_0 for a pair being dissimilar. Assuming a Gaussian structure of the difference space we can relax the problem and re-write Eq. (10) to

$$\delta(\mathbf{x}_{ij}) = \log \left(\frac{ \frac{1}{\sqrt{2\pi |\Sigma_{y_{ij}=0}|}} \exp(-1/2 \ \mathbf{x}_{ij}^T \ \Sigma_{y_{ij}=0}^{-1} \ \mathbf{x}_{ij}) }{ \frac{1}{\sqrt{2\pi |\Sigma_{y_{ij}=1}|}} \exp(-1/2 \ \mathbf{x}_{ij}^T \ \Sigma_{y_{ij}=1}^{-1} \ \mathbf{x}_{ij}) } \right) , \tag{11}$$

where

$$\Sigma_{y_{ij}=1} = \sum_{y_{ij}=1} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{12}$$

$$\Sigma_{y_{ij}=0} = \sum_{y_{ij}=0} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \,. \tag{13}$$

The pairwise differences $\mathbf{x}_{ij}$ are symmetric. Thus, we have zero mean and $\theta_1 = (\mathbf{0}, \Sigma_{y_{ij}=1})$ and $\theta_0 = (\mathbf{0}, \Sigma_{y_{ij}=0})$. The maximum likelihood estimate of the Gaussian is equivalent to minimizing the Mahalanobis distances from the mean in a least squares manner. This allows us to find respective relevant directions for the two independent sets. By taking the log, we can re-formulate the likelihood-test as

$$\begin{aligned}\delta(\mathbf{x}_{ij}) = {} & \mathbf{x}_{ij}^T\, \Sigma_{y_{ij}=1}^{-1}\, \mathbf{x}_{ij} + \log(|\Sigma_{y_{ij}=1}|) \\ & -\mathbf{x}_{ij}^T\, \Sigma_{y_{ij}=0}^{-1}\, \mathbf{x}_{ij} - \log(|\Sigma_{y_{ij}=0}|).\end{aligned} \tag{14}$$

Further, we strip constant terms as they just provide an offset and simplify to

$$\delta(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T (\Sigma_{y_{ij}=1}^{-1} - \Sigma_{y_{ij}=0}^{-1}) \mathbf{x}_{ij}. \tag{15}$$

Finally, we obtain our Mahalanobis distance metric that reflects the properties of the log-likelihood ratio test

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) \tag{16}$$

by re-projection of $\hat{\mathbf{M}} = \left(\Sigma_{y_{ij}=1}^{-1} - \Sigma_{y_{ij}=0}^{-1}\right)$ onto the cone of positive semidefinite matrices. Hence, to obtain $\mathbf{M}$ we clip the spectrum of $\hat{\mathbf{M}}$ by eigenanalysis.

4. Experiments

To show the broad applicability of our method we conduct experiments on various standard benchmarks with rather diverse characteristics. The goals of our experiments are twofold. First, we want to show that our method is able to generalize to unseen data as well or even better than state-of-the-art metric learning approaches. Second, we want to prove that we are orders of magnitudes faster. This is clearly beneficial for large scale or online applications. In Section 4.1 we first conduct experiments on faces in unconstrained environments. Succeeding, in Section 4.2 we study the problem of person re-identification across spatially disjoint cameras. Finally, in Section 4.3 we evaluate on the ToyCars dataset intended to compare before unseen object instances.

For the comparison to other metric learning approaches the numbers were generated with original code and same input data. The code was kindly provided by the respective authors. Further, we compare our method to related domain specific approaches. For all plots the numbers in parentheses denote the Equal Error Rate (EER) of the respective method.

4.1. Face Recognition

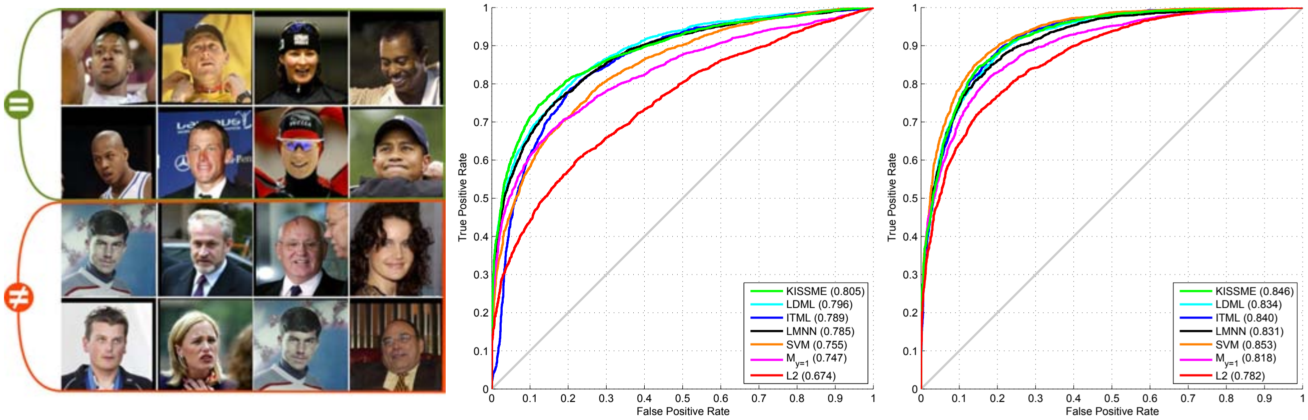
In the following, we demonstrate the performance of our method on two challenging face recognition datasets, namely on Labeled Faces in the Wild (LFW) [12] and Public Figures Face Database (PubFig) [13]. Hereby, the study of face recognition is divided into two different objectives: face identification (naming a face) and face verification (deciding if two face images are of the same individual). The nature of the face identification task requires a number of annotated faces per individual, with which these real-world databases not always comply with. In contrast, face verification needs less annotations and can be evaluated more seriously also on a large scale. Thus, compliant with previous work we focus on the face verification task.

4.1.1 Labeled Faces in the Wild

The Labeled Faces in the Wild dataset [12] contains 13,233 unconstrained face images of 5,749 individuals and can be considered as the current state-of-the-art face recognition benchmark. The database is considered as very challenging as it exhibits huge variations in pose, lighting, facial expression, age, gender, ethnicity and general imaging and environmental conditions. Some illustrative examples are given in Figure 2 (a). An important aspect of LFW is that per design the subjects are mutually exclusive in any split of the database. Thus, for the face verification task testing is done on individuals that have not been seen in training.

The data is organized in 10 folds that are used for cross-validation. Each fold consists of 300 similar and 300 dissimilar pairs. The result scores are averaged over the 10 folds. In the restricted protocol it is only allowed to consider the equivalence constraints given by the similar / dissimilar pairs. No inference on the identity of the subject, *e.g.*, to sample more training data, is allowed.

For our experiments we use the face representation proposed by Guillaumin *et al.* [8]. Basically, it extracts SIFT descriptors [14] at 9 automatically detected facial landmarks (corners of the mouth, eyes and nose), over three scales. The resulting descriptor is a 3,456 dimensional vector. To make it tractable for the distance metric learning algorithms we perform dimensionality reduction by PCA to a 100 dimensional subspace. To evaluate the different metric learning methods and enable a fair comparison we train the classifiers with exactly the same features and PCA dimensions. The influence of the PCA dimensionality is not too critical. Using different dimensionalities for all tested methods reveals that there is no significant change in performance. Except for the linear SVM we train directly on



(a) Example pairs.

(b) ROC/SIFT.

(c) ROC/Attributes.

Figure 2: **Face verification results on the LFW dataset**: (a) Examples of similar and dissimilar pairs. ROC curves for different feature types and learners: In (b) we report the performance for the SIFT features of [8] and in (c) for the "high-level" description of visual face traits [13]. For the SIFT features our method outperforms several metric learning approaches slightly. For the attributes it matches with the SVM based approach proposed by [13].

the face descriptors as this delivers better results.

In Figure 2 (b) we report a Receiver Operator Characteristic (ROC) curve for LDML [8], ITML [3], LMNN [19], SVM [2], our method (KISSME), the Mahalanobis distance of the similar pairs and the Euclidean distance as baseline. Please note that for LMNN we have to provide more supervision in form of the actual class labels (not just equivalence constraints) as it needs labeled triplets.

The Mahalanobis distance of the similar pairs ($\mathbf{M}_{y=1}$) performs already quite well in comparison to the Euclidean distance. It increases the performance by about 7%. Interestingly, LMNN is not really able to capitalize on the additional information over the other metrics. KISSME outperforms the other metrics slightly. It reaches with an Equal Error Rate of 80.5% the best reported results up to now for this kind of feature type. Of course recent state-of-the-art on LFW provides better results but also requires considerably more domain knowledge (*i.e.*, pose specific classifiers), as these approaches focus purely on faces.

When analyzing the training times given in Table 1 the main advantage of our method is obvious. In fact, compared to LMNN, ITML, and LDML our method is computationally much more efficient, however, still yielding competitive results.

Method	**LFW**	**PubFig**	**VIPeR**	**ToyCars**
KISSME	**0.05s**	**0.07s**	**0.01s**	**0.04s**
SVM	12.78s	0.84s	–	0.60s
ITML	24.81s	20.82s	8.60s	14.05s
LDML	307.23s	2868.91s	0.72s	1.21s
LMNN	1198.69s	783.66s	27.56s	0.79s

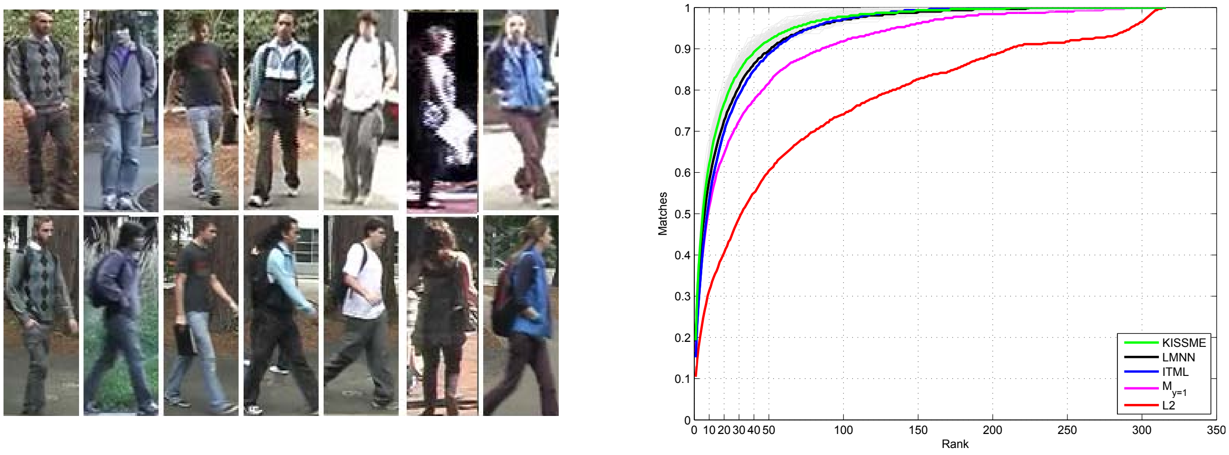
Table 1: **Average training times**. LDML, LMNN make use of multi-threading. Evaluated on a 3.06 GHz Xeon with 24 cores.

4.1.2 Public Figures Face Database

The PubFig dataset [13] has many commonalities with LFW. It is also an extremely challenging large-scale, real-world database, consisting of 58,797 images of 200 individuals. The images were gathered from Google images and FlickR. The face verification benchmark consists of 10 cross-validation folds with 1,000 intra and 1,000 extra-personal pairs each. Per fold the pairs are sampled of 14 individuals. Similar to the LFW benchmark individuals that appear in testing have not been seen before in training.

An interesting aspect of the database is that "high-level" features are provided that describe the presence or absence of visual face traits. The appearance is automatically encoded in either nameable attributes such as gender, race, age, hair *etc.* or "similes" that relate the similarity of face regions to specific reference people. This indirect description yields nice properties such as a certain robustness to image variations compared to low-level features. Further, it offers us a complementary feature type to evaluate the performance of the distance metric learning algorithms.

In Figure 4 we report ROC curves for LDML [8], ITML [3], LMNN [19], SVM [2], our method (KISSME) and two baselines. It can be seen that we outperform ITML, LMNN and match the state-of-the-art performance of the SVM based method proposed by Kumar *et al.* [13]. LDML delivers similar results to our algorithm while being



(a) Example Pairs

(b) CMC

Figure 3: **Person re-identification results on the VIPeR dataset**. Example image pairs are shown in (a). In (b) average Cumulative Matching Characteristic (CMC) curves over 100 runs are plotted. Our method (KISSME) slightly outperforms the other metrics. In light-gray all runs of our method are indicated.

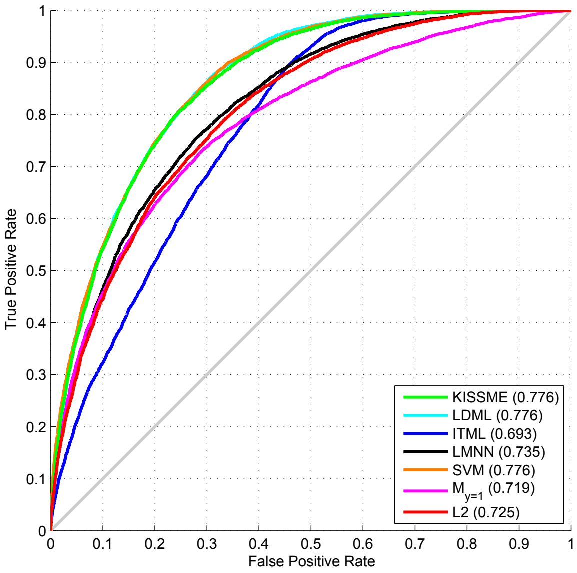


Figure 4: **Face verification results on the PubFig dataset**. For all methods we use the "high-level" description of visual face traits [13].

orders of magnitudes slower in training (see Table 1). This makes the algorithm impracticable for online or large-scale use. Interestingly, the performance of ITML drops even below the plain Euclidean distance. In Figure 2 (b) we also report the performance of the attribute features for LFW.

4.2. Person Re-Identification

The VIPeR dataset [6] consists of 632 intra-personal image pairs of two different camera views, captured outdoors. The low-resolution images (48x128px) exhibit significant variations in pose, viewpoint and also considerable changes in illumination, like highlights or shadows. Most of the example pairs contain a viewpoint change of about 90 degrees, making person re-identification very challenging. Some examples are given in Figure 3 (a). To compare our method to other approaches, we followed the evaluation protocol defined in [5, 7]. The authors split the set of 632 image pairs randomly into two sets of 316 image pairs each, one for training and one for testing, and compute the average over several runs. There is no predefined set or procedure how to obtain dissimilar pairs. Hence, we generate dissimilar pairs by randomly combining images of different persons.

To represent the images we compile a rather simple descriptor. First, we divide the images into overlapping blocks of size 8x16 and stride of 8x8. Second, to describe color cues we extract HSV and Lab histograms, each with 24 bins per channel. Third, we capture texture information by LBPs [16]. Finally, for the distance metric learning approaches we project the concatenated descriptors into a 34 dimensional subspace by PCA.

To indicate the performance of the various algorithms we report Cumulative Matching Characteristic (CMC) curves [18]. These represent the expectation of the true match being found within the first n ranks. To obtain a reasonable statistical significance, we average over 100 runs. In Figure 3 (b) we report the CMC curves for the various metric learning algorithms. Moreover, in Table 2 (b) we compare the performance of our approach in the range of the first 50 ranks to state-of-the-art methods [4, 5, 10, 23]. As can be seen, we obtain competitive results across all ranks. We outperform the other methods [5, 7, 17] even though in contrast

RANK	1	10	25	50
KISSME	**19.6%**	**62.2%**	**80.7%**	**91.8%**
LMNN	19.0%	58.1%	76.9%	89.6%
ITML	15.2%	53.3%	74.7%	88.8%
LDML	10.4%	31.3%	44.6%	60.4%
M$_{y=1}$	16.8%	50.9%	68.7%	82.0%
L2	10.6%	31.8%	44.9%	60,8%

(a)

RANK	1	10	25	50
KISSME	**20%**	**62%**	**81%**	**92%**
SDALF [5]	**20%**	50%	70%	85%
DDC [10]	19%	52%	69%	80%
PRDC [23]	16%	54%	76%	87%
KISSME*	**22%**	**68%**	**85%**	**93%**
LMNN-R* [4]	20%	**68%**	84%	**93%**

(b)

Table 2: **Person re-identification matching rates on the VIPeR dataset**. Table (a) shows the metric learning approaches (average of 100 runs) whereas (b) gives an overview of the state-of-the-art. To be comparable to LMNN-R we also report the best run (*).

to them we do not use a foreground-background segmentation. Further, we are computationally more efficient.

4.3. ToyCars

The LEAR ToyCars [15] dataset consists of 256 image crops of 14 different toy cars and trucks. The dataset exhibits changes in pose, lighting and cluttered background. The intention is to compare before unseen object instances of the known class *cars* (see Figure 5 for illustration). Thus, in testing the task is to classify if a pair of images shows the same object or not. The training set contains 7 object instances with associated 1185 similar and 7330 dissimilar image pairs. The remaining 7 object instances are in the test set. As the images differ in horizontal resolution we zero-pad them to obtain a canonical image size.

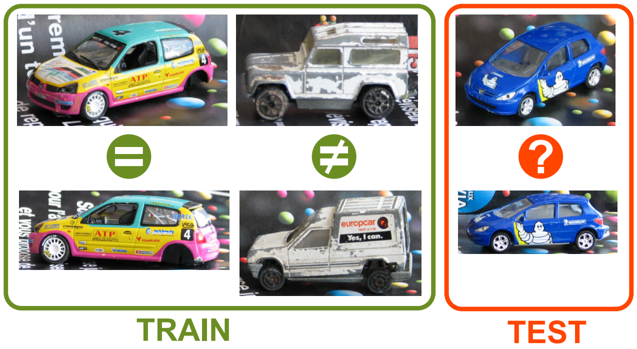


Figure 5: **LEAR ToyCars [15] dataset**. The task is to decide if two before unseen object instances of the known class cars are similar or not.

We extract a similar image representation as used in the person re-identification experiment. Therefore, the images are divided into 30x30 non-overlapping blocks. We capture color cues by HSV and Lab histograms while texture is described by LBPs [16]. The global image descriptor is a concatenation of the local ones. Using PCA the descriptor is projected onto a 50 dimensional subspace.

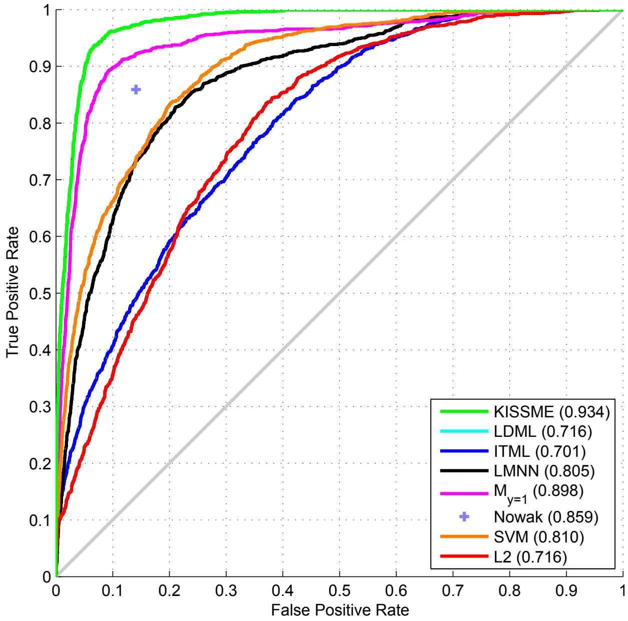


Figure 6: **ROC curves on LEAR ToyCars dataset**. Our method is able to drop error rates significantly compared to the previous work of Nowak *et al.* [15].

We conduct our experiments on this dataset in comparison to the recent approach of Nowak and Jurie [15], which builds on an ensemble of extremely randomized trees. The ensemble quantizes corresponding patch pair differences by enforcing that corresponding patches of matching pairs yield similar responses. Corresponding patches are located in a local neighborhood by NCC. In testing the similarity between an image pair is the weighted sum of corresponding patches that end up in the same leaf node.

In Figure 6 we plot ROC curves which compare our method to the work of Nowak and Jurie [15] and the related metric learning approaches. Further, we provide a baseline with a standard linear SVM [2]. Using SVM yields an

EER of 81%, already a reasonable performance. Interestingly, some of the metric learning approaches are not able to improve over the Euclidean distance. Only LMNN performs similar to the SVM. By taking the Mahlanobis distance learned form the positive pairs only we can already outperform Nowak and Jurie's approach and reach an EER of 89.8%. KISSME boosts the performance further up to 93.5%. If one considers the computation time of [15] with 17 hours (P4-3.4GHz) our approach once more shows its benefits in terms of efficiency and effectiveness.

5. Conclusion

In this work we presented our KISS method to learn a distance metric from equivalence constraints. Based on a statistical inference perspective we provide a solution that is very efficient to obtain and effective in terms of generalization performance. To show the merit of our method we conducted several experiments on various challenging large-scale benchmarks, including LFW and PubFig. On all benchmarks we are able to match or slightly outperform state-of-the-art metric learning approaches, while being orders of magnitudes faster in training. On two datasets (VIPeR, ToyCars) we even outperform approaches especially tailored to these tasks.

References

[1] L. M. Bregman. The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967. 3

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 5, 7

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proc. IEEE Intern. Conf. on Machine Learning*, 2007. 1, 2, 3, 5

[4] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Proc. Asian Conf. on Computer Vision*, 2010. 1, 6, 7

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2010. 6, 7

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recongnition, reacquisition and tracking. In *Proc. IEEE Intern. Workshop on Performance Evaluation of Tracking and Surveillance*, 2007. 2, 6

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. European Conf. on Computer Vision*, 2008. 6

[8] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. 1, 2, 3, 4, 5

[9] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *Proc. European Conf. on Computer Vision*, 2010. 1

[10] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conf. on Image Analysis*, 2011. 6, 7

[11] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2006. 1

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1, 2, 4

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. 2, 4, 5, 6

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intern. Journal of Computer Vision*, 60(2):91–110, 2004. 4

[15] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2007. 2, 7, 8

[16] T. Ojala, M. Pietikänien, and T. Mäenpä. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI*, 24(7):971–987, 2002. 6, 7

[17] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. British Machine Vision Conf.*, pages 21.1–21.11, 2010. 6

[18] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2007. 6

[19] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances NIPS*, 2006. 1, 2, 5

[20] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *Proc. IEEE Intern. Conf. on Machine Learning*, 2008. 1, 2

[21] J. Ye, Z. Zhao, and H. Liu. Adaptive distance metric learning for clustering. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2007. 1

[22] J. Yu, J. Amores, N. Sebe, and Q. Tian. Toward robust distance metric analysis for similarity estimation. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2006. 1

[23] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2011. 6, 7