Is that you? Metric learning approaches for face identification

Matthieu Guillaumin, Jakob Verbeek, Cordelia Schmid

Is that you? Metric Learning Approaches for Face Identification

Matthieu Guillaumin, Jakob Verbeek and Cordelia Schmid
LEAR, INRIA Grenoble Laboratoire Jean Kuntzmann
firstname.lastname@inria.fr

Abstract

Face identification is the problem of determining whether two face images depict the same person or not. This is difficult due to variations in scale, pose, lighting, background, expression, hairstyle, and glasses. In this paper we present two methods for learning robust distance measures: (a) a logistic discriminant approach which learns the metric from a set of labelled image pairs (LDML) and (b) a nearest neighbour approach which computes the probability for two images to belong to the same class (MkNN). We evaluate our approaches on the Labeled Faces in the Wild data set, a large and very challenging data set of faces from Yahoo! News. The evaluation protocol for this data set defines a restricted setting, where a fixed set of positive and negative image pairs is given, as well as an unrestricted one, where faces are labelled by their identity. We are the first to present results for the unrestricted setting, and show that our methods benefit from this richer training data, much more so than the current state-of-the-art method. Our results of 79.3% and 87.5% correct for the restricted and unrestricted setting respectively, significantly improve over the current state-of-the-art result of 78.5%. Confidence scores obtained for face identification can be used for many applications e.g. clustering or recognition from a single training example. We show that our learned metrics also improve performance for these tasks.

1. Introduction

Face identification is a binary classification problem over pairs of face images: we have to determine whether or not the same person is depicted in both images. More generally, visual identification refers to deciding whether or not two images depict the same object from a certain class. The confidence scores, or *a posteriori* class probabilities, for the visual identification problem can be thought of as an object-category-specific dissimilarity measure between instances of the category. Ideally it is 1 for images of different instances, and 0 for images of the same object. Importantly, scores for visual identification can also be applied for other problems such as visualisation, recognition from a single example [18], associating names and faces in images [2, 11] or video [7], or people oriented topic models [16].

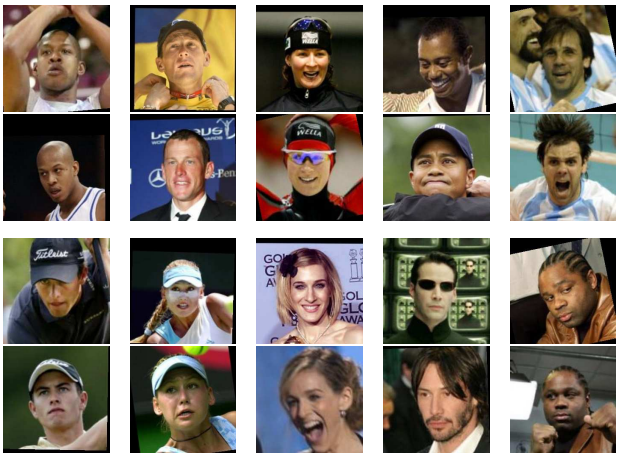


Figure 1. Several examples face pairs of the same person from the *Labeled Faces in the Wild* data set. We show pairs that were correctly (top) and incorrectly (bottom) classified with our method.

Recently there has been considerable interest for face and visual identification [5, 8, 12, 15, 20, 24]. Faces are particularly challenging due to possible variations in appearance, see for example Figure 1. Furthermore, the analysis of humans (identity, pose, actions, etc.) is an important topic in computer vision. In this paper we propose two methods for face identification based on learning Mahalanobis metrics over a given representation space. The first method, LDML, uses logistic discriminant to learn a metric from a set of labelled image pairs. Its objective is to find a metric such that positive pairs have smaller distances than negative pairs. The second method, MkNN, uses a set of labelled images, and is based on marginalising a k-nearest-neighbour (kNN) classifier for both images of a pair. The MkNN classifier computes the marginal probability that the two faces are the same person, *i.e.* marginalising over who that exactly is. For this second method we also use a learned metric, albeit one that is optimised for kNN classification [23].

Metric learning has received a lot of attention, for recent work in this area see *e.g.* [1, 6, 9, 10, 23, 25]. Most methods learn a Mahalanobis metric based on an objective function defined by means of a labelled training set, or from sets of positive (same class) and negative (different class) pairs. The difference among these methods mainly lies in their objective functions, which are designed for their specific tasks (clustering [25], kNN classification [23]). Some methods explicitly need all pairwise distances between points [10], making large scale applications (say more than 1000 data points) more difficult. Among the existing methods for learning metrics, large margin nearest neighbor (LMNN) [23] and information theoretic metric learning (ITML) [6] are state-of-the-art. Surprisingly, there is a lack of experimentation of these techniques on challenging, large, real-world data sets of human faces.

Compared to other face identification methods, experimental results show that our metric learning approaches are able to improve results significantly if more training data is available. This is, for example, not the case for the current best approach of [24], as it does not learn parameters in a discriminative manner, but based on an estimate of the covariance over all available face images.

We report experimental results on the *Labeled Faces in the Wild* (LFW) data set [14], which is the *de facto* standard dataset for face identification. It contains 13233 labelled faces of 5749 people, for 1680 people there are two or more faces. Furthermore, the data is challenging, as the faces are detected in images "in the wild", taken from *Yahoo! News*. The faces exhibit appearance variations as they occur in uncontrolled settings, including changes in scale, pose, lighting, background, hairstyle, clothing, expression, color saturation, image resolution, focus, etc. The data set comes with fixed, fully independent training and test data sets, and allows two forms of supervision. In the "restricted" setting a subset of pairs are labelled as being the same person (positive) or not (negative). In the "unrestricted" setting, one can use all available face labels, either by using them directly, or to generate larger sets of labelled pairs.

In Section 2, we describe our logistic discriminant-based metric learning method and existing state-of-the-art metrics (LMNN [23], ITML [6]), and in Section 3 our marginalised kNN method. We present the data set and experimental setup in Section 4, and the experimental results for face identification in Section 5. In the restricted setting, we find slightly improved performance, 79.3%, as compared to the current state-of-the-art result of 78.5% [24], but only our method benefits from the larger training set of the unrestricted setting, pushing the accuracy to 87.5%.

Additional experiments presented in Section 6 show that our learned metrics also lead to improvements in clustering and recognition from a single example. We present our conclusions and directions for further research in Section 7.

2. Metric Learning for Face Identification

In this section we present three methods to learn Mahalanobis metrics for visual identification. First we present two state-of-the-art methods, LMNN [23] and ITML [6]. Then we propose our method, LDML. We note $\mathbf{x}_i \in \mathbb{R}^D$ the representation of image i and y_i its class label. Image i and j form a positive pair if $y_i = y_j$, and a negative pair otherwise. The Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j), \qquad (1)$$

where $\mathbf{M} \in \mathbb{R}^{D \times D}$ is a symmetric positive definite matrix.

2.1. Large Margin Nearest Neighbour Metrics

Recently, Weinberger *et al*. introduced a method that learns a matrix $\mathbf{M}$ designed to improve results of k nearest neighbour (kNN) classification [23]. The intuition is that for each data point such a metric should make the k nearest neighbours of its own class – target neighbours – closer than points from other classes. The objective is composed of two terms. The first term minimises the distances between target neighbours, while the second term is a hinge-loss that encourages target neighbours to be at least one distance unit closer than points from other classes.

Rather than requiring pairs of images labelled positive or negative, this method needs labelled triples (i, j, l) of target neighbours (i, j) and points which should not be neighbours (i, l). It is thus not possible to use it in the restricted setting. However, triples can be formed from the labelled training data $(\mathbf{x}_i, y_i)$ in the unrestricted setting. This approach may not be optimal when thresholding the distances, but is well suited to use as base metric to define neighbours in our MkNN approach that we present in Section 3.

2.2. Information Theoretic Metric Learning

Davis *et al*. [6] have taken an information theoretic approach to optimize $\mathbf{M}$ under a wide range of possible constraints and prior knowledge on the Mahalanobis distance. This is done by regularizing the matrix $\mathbf{M}$ such that it is as close as possible to a known prior $\mathbf{M}_0$. This closeness is interpreted as a Kullbach-Leibler divergence between the two Gaussian distributions corresponding to $\mathbf{M}$ and $\mathbf{M}_0$. Typically, the other contraints will be of the form $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq u$ for positive pairs and $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq l$ for negative pairs. The trade-off between satisfying the constraints and regularization is controlled in the objective function using an additional parameter γ. The parameters $\mathbf{M}_0$, upper bound u, lower bound l and γ have to be provided, although it is also possible to resort to cross-validation techniques.

2.3. Logistic Discriminant based Metric Learning

Our proposed method is based on the idea that we would like the distance between images in positive pairs to be

smaller than the distances corresponding to negative pairs, and obtain a probabilistic estimation of whether the two images depict the same object. Using the Mahalanobis distance between two images, we model the probability p_n that pair $n = (i, j)$ is positive, *i.e.* the pair label t_n is 1, as:

$$p_n = p(y_i = y_j | \mathbf{x}_i, \mathbf{x}_j; \mathbf{M}, b) = \sigma(b - d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)), \quad (2)$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the sigmoid function and b a bias term. Interestingly for the visual identification task, the bias will directly work as a threshold value and is learned together with the metric parameters.

Note that $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$ is linear with respect to the elements of $\mathbf{M}$, and thus, when we rewrite $p_n = \sigma(b - W^\top X_n)$ where W is the vector containing the elements of $\mathbf{M}$ and X_n the entries of $(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$, the model in Eq. (2) appears as a standard linear logistic discriminant model. We use maximum log-likelihood to optimize the parameters of the model. The log-likelihood $\mathcal{L}$ can be written as:

$$\mathcal{L} = \sum_n t_n \ln p_n + (1 - t_n) \ln(1 - p_n) \quad (3)$$

$$\nabla \mathcal{L} = \sum_n (t_n - p_n) X_n, \quad (4)$$

which is known to be smooth and concave. The optimization process, using gradient ascent, is therefore simpler and faster than ITML or LMNN.

Several convex constraints can be imposed on the matrix $\mathbf{M}$, such as diagonality or positive definiteness, without losing the smoothness or concavity of the log-likelihood. In this case, the maximum likelihood estimates for the metric $\mathbf{M}$ and bias term b are obtained using the projected gradient method of [3]. Positive definiteness is required in applications where the data needs a vectorial representation. For our experiments, we did not constrain $\mathbf{M}$, as we only use the probabilities given by Eq. (2). A prior can also be added so as to obtain a MAP estimate of $\mathbf{M}$. We refer to this method as LDML, for logistic discriminant metric learning.

3. Identification with Nearest Neighbors

In the previous section we presented methods to learn Mahalanobis metrics, which are always linear transformations of an original space. With this limitation, it may be impossible to separate positive and negative pairs, as appearance variations for a single person might be non-linear and larger than the inter-person variations for a similar pose and expression. In this section, we show how kNN classification can be used for visual identification. The resulting non-linear, high-capacity classifier implicitly uses all pairs that can be generated from the labelled data.

Normally, kNN classification is used to assign single data points $\mathbf{x}_i$ to one of a fixed set of k classes associated with the training data. The probability of class c for $\mathbf{x}_i$ is

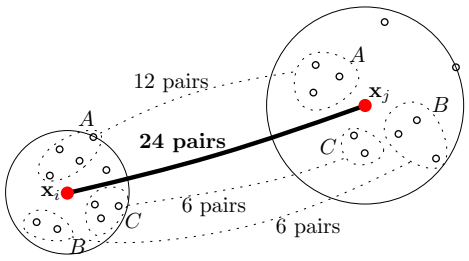


Figure 2. Schematic representation of $k = 10$ neighbourhoods for $\mathbf{x}_i$ and $\mathbf{x}_j$, and the 24 neighbour pairs (out of 100) that have the same name and contribute to the score.

$p(y_i = c|\mathbf{x}_i) = n_c^i/k$, where n_c^i is the number of neighbours of $\mathbf{x}_i$ of class c. Here, we have to predict whether a pair of images $(\mathbf{x}_i, \mathbf{x}_j)$ belongs to the same class, regardless of which class that is, and even if the class is not represented in the training data. To answer this question we compute the marginal probability that we assign $\mathbf{x}_i$ and $\mathbf{x}_j$ to the same class using a kNN classifier, which equals:

$$\begin{aligned} p(y_i = y_j | \mathbf{x}_i, \mathbf{x}_j) &= \sum_c p(y_i = c|\mathbf{x}_i) p(y_j = c|\mathbf{x}_j) \\ &= k^{-2} \sum_c n_c^i n_c^j. \end{aligned} \quad (5)$$

Alternatively, we can understand this method directly as a nearest neighbor classifier in the implicit binary labelled set of N^2 pairs. In this set, we need a measure to define neighbours of a pair. One choice to do so for a pair $(\mathbf{x}_i, \mathbf{x}_j)$ is to take all the pairs we can make using one of the k neighbours of $\mathbf{x}_i$ and one of the k neighbours of $\mathbf{x}_j$. The probability for the positive class given by this classifier for a pair is then determined by the number of positive and negative neighbour pairs, and is precisely given by Eq. (5).

Either way, the score of our Marginalized kNN (MkNN) binary classifier for a pair of images $(\mathbf{x}_i, \mathbf{x}_j)$ is based on how many positive neighbour pairs we can form from neighbours of $\mathbf{x}_i$ and $\mathbf{x}_j$. In Figure 2 we illustrate the procedure with a simple example. We expect this method to benefit from an LMNN base metric to define the neighbours, as it is designed to improve kNN classification.

Using this approach, we profit from the amount of available data and flexibly model non-linearities, at the expense of a higher computational cost at test time. It is not "local" in the sense of usual kNN classifiers or other "local learning" methods [4, 9], as MkNN measures the correspondence between two distinct local neighbourhoods. It implicitly uses all pairs we can generate from the labeled faces.

4. Data Set, Experimental Setup, and Features

The *Labeled Faces in the Wild* (LFW) data set contains 13233 face images labelled by the identity of the person [14]. In total 5749 people appear in the images, 1680

of them appear in two or more images. The faces were detected in images downloaded from *Yahoo! News* in 2002–2003, and show a big variety in pose, expression, lighting, etc. An aligned version of all faces is available, referred to as "funneled", which we use throughout our experiments.[1]

The data set comes with a division in 10 fully independent parts (folds) that can be used for cross validation experiments. The folds contain between 527 and 609 different people each, and between 1016 and 1783 faces. From all possible pairs, a small set of 300 positive and 300 negative image pairs are provided for each fold. Using only these pairs for training is referred to as the "restricted" paradigm; in this case the identity of the people in the pairs cannot be used. The "unrestricted" paradigm is used to refer to training methods that can use all available data, including the identity of the people in the images. This allows us to use the labels directly as MkNN and LMNN do, or explicitly generate a much larger number of pairs per fold (thousands in each fold). In turn, each fold is held-out to measure, on its 600 pairs from the restricted setting, performance of classifiers learned on the 9 other folds. Below, we present results following both training paradigms.

The results are reported based on the operating points of the ROC curves at equal misclassification cost (ROC-EMC). This standard performance measure slightly differs from the accuracy used to report results on the LFW website where a separate threshold is learnt for each fold. For direct comparison with the state-of-the-art in Section 5.3, we report performance using the LFW accuracy.

We have experimented with several feature sets for faces used in recent work: Local Binary Patterns (LBP) [21], and its variations proposed in [24]. For details on these descriptors we refer the reader to [24]. Following [11] we have also used SIFT descriptors [19] computed at fixed points on the face (corners of the mouth, eyes, and nose) found using a facial feature detector [7]. We compute 128 dimensional SIFT descriptors at three scales, centered on 9 points, leading to a $3 \times 9 \times 128 = 3456$ dimensional face descriptor.

5. Experimental Results on Face Identification

In Section 5.1 we present face identification results on the LFW data set in the restricted setting, and in Section 5.2 we present the first results ever published on the unrestricted setting. We compare to related work in Section 5.3.

5.1. Image Restricted Training Paradigm

We first evaluate an unsupervised baseline method; results are obtained by thresholding standard metrics. Since all our descriptors are histograms we applied the Hellinger (obtained as the L2 distance after taking the square root of the histogram values), χ^2 and L2 distances. Perhaps surprisingly, the results in Table 1 show that all descriptors and distances lead to comparable ROC-EMC of 66% to 69%. As the SIFT based descriptor performs slightly better than the others, we use it hereafter to compare methods.

Descriptor	L2	Hellinger	χ^2
LBP	67.65 ± 0.7	68.13 ± 0.7	68.33 ± 0.6
TPLBP	66.90 ± 0.4	66.82 ± 0.4	66.58 ± 0.2
FPLBP	66.52 ± 0.5	67.37 ± 0.4	67.10 ± 0.5
SIFT	$\mathbf{67.78 \pm 0.6}$	$\mathbf{68.50 \pm 0.5}$	$\mathbf{68.77 \pm 0.4}$

Table 1. ROC-EMC classification results for L2, Hellinger and χ^2 distances for different descriptors.

Method / PCA dim.		Original	Square Root
LDML		$\mathbf{76.6 \pm 0.7}$	$\mathbf{77.5 \pm 0.5}$
ITML	35	75.2 ± 0.7	75.8 ± 0.5
LDA-based		73.8 ± 0.4	74.2 ± 0.4
LDML		72.8 ± 0.6	72.8 ± 0.4
ITML	55	$\mathbf{75.6 \pm 0.6}$	$\mathbf{76.2 \pm 0.5}$
LDA-based		74.9 ± 0.8	75.3 ± 0.3
LDA-based	600	78.6 ± 0.4	79.4 ± 0.2

Table 2. ROC-EMC performances for LDML, ITML and the LDA-based method of [24] in the restricted setting (600 training pairs per fold) using SIFT. The parameters for ITML are $u = l = \gamma = 1$ and $M_0 = I$. ITML and LDML are intractable when using 600 PCA dimensions.

To apply our LDML approach we pre-process the data using PCA (separately for each fold), as otherwise the number of parameters is too large (almost 6 million). We tried several dimensionalities for the PCA projection and found that performance dropped when using more than 35 dimensions, as shown in Table 2. This is explained by the large number of parameters to estimate when using larger dimensional PCA spaces, which causes over-fitting as we only have 5400 training samples in the restricted paradigm: 600 pairs from each of the 9 folds. We also find that taking the square root of the data values gives a small improvement.

For comparison, we trained an ITML[2] [6] metric and implemented the state-of-the-art Linear Discriminant Analysis (LDA) based method [24]. The performance levels reported in Table 2 show that, using the same descriptor and the same PCA dimensions, our method can outperform the LDA-based method and ITML when over-fitting is avoided. Using more PCA dimensions increases the performance of the LDA-based method and ITML, at the expense of a larger face representation and higher training times.

Interestingly, SIFT based features yield better results than any of the descriptors reported in [24], where the best single descriptor results in 74.6% accuracy.

[1] Data set available at: http://vis-www.cs.umass.edu/lfw/

[2] Cf: http://www.cs.utexas.edu/users/pjain/itml/

Method / PCA dim.		1.000	2.000	10.000
LDML		**76.3 ± 0.6**	**77.2 ± 0.5**	**77.4 ± 0.5**
ITML	35	75.6 ± 0.7	76.2 ± 0.6	75.9 ± 0.6
LDA-based		74.0 ± 0.6	73.9 ± 0.5	73.9 ± 0.6
LDML		76.2 ± 0.4	**78.5 ± 0.5**	**80.4 ± 0.4**
ITML	55	**77.3 ± 0.5**	78.3 ± 0.4	78.4 ± 0.6
LDA-based		74.8 ± 0.6	74.8 ± 0.4	75.0 ± 0.6
LDML		71.3 ± 0.8	76.7 ± 0.8	**83.2 ± 0.4**
ITML	100	**77.6 ± 0.3**	**78.4 ± 0.5**	80.5 ± 0.5
LDA-based		76.5 ± 0.3	76.4 ± 0.5	76.4 ± 0.5
LDA-based	600	79.3 ± 0.3	79.1 ± 0.2	79.1 ± 0.3

Table 3. ROC-EMC classification results of LDML, ITML and LDA-based method in the unrestricted setting, when varying the number of training pairs per fold, and the PCA dimensionality of the SIFT descriptor. ITML and LDML are intractable with 600 PCA dimensions.

5.2. Unrestricted Training Paradigm

In the unrestricted setting more image pairs are available for training. This reduces over-fitting and allows us to use models with more parameters. Table 3 shows performance of LDML when using an increasing number of training pairs: 1000, 2000 and 10000 pairs per cross-validation fold, instead of the 600 provided in the restricted setting. As expected, we see that the models with the largest number of parameters benefit the most from an increased number of training pairs. When using a 100 dimensional PCA projection of the SIFT data, *i.e*. 5050 parameters, more than 10% increase in ROC-EMC is obtained by using 10 times more training examples (90000 in total). Interestingly, using more training data significantly increases the performance of our LDML model (up to 83.2%), but only slightly impacts ITML and does not impact the current state-of-the-art LDA-based approach at all. This is because LDA is not learning parameters in a discriminative manner, but relies on the leading eigenvectors of the covariance matrix over all available face images. These can be estimated accurately based on a limited number of training pairs.

Using the labels of the unrestricted setting, we can employ LMNN[3] and our MkNN approach. For LMNN we used a PCA projection of the data to 200 dimensions; using less dimensions gave slightly worse results, and using more dimensions gave slightly better results at the cost of much higher training times. We used 5 target neighbours to learn the LMNN metric; using between 3 and 20 target neighbours gave similar performance, other values gave slightly worse results. This resulted in a best performance of 80.5%.

We, then, applied the MkNN classifier using L2, LMNN, and LDML as base metrics. In the case of L2 and LDML as base metric, the MkNN classifier did not give as good results as the base metric. However, when using LMNN, designed for kNN classification, as a base metric, the MkNN classifier performs better when between 100 and 200 neighbours are used: 83.1% instead of 80.5%, see Figure 4. We also considered a variant where a weighted sum of the base metric and the class probability is learnt using a logistic discriminant classifier. This combination brings a small improvement over the base metric from 67.8% to 69.2% for L2, from 83.2% to 83.3% for LDML and from 80.5% to 83.5% for LMNN. Furthermore, for LMNN, this improvement is consistent over all neighbourhood sizes, as shown in Figure 3. Figure 4 shows some of the examples that were incorrectly classified using the LMNN metric, but were correctly classified using the MkNN classifier. The benefit of the MkNN classifier can be seen most for pairs with large pose and or expression changes.

[3]We used code available at: http://www.weinbergerweb.net

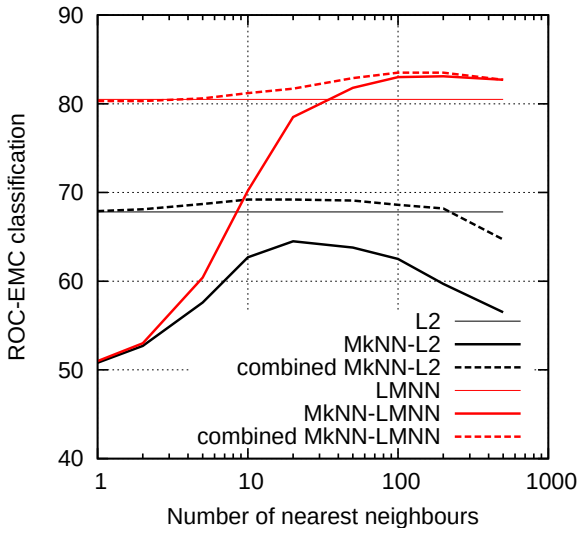


Figure 3. ROC-EMC performance using the MkNN classifier, with L2 and LMNN as base metrics.

ITML	LDA-based	LMNN	LDML	MkNN
80.5 ± 0.5	79.3 ± 0.3	80.5 ± 0.5	**83.2 ± 0.4**	**83.1 ± 0.5**

Table 4. Comparison of ROC-EMC classification results for methods in the unrestricted setting (SIFT).

Table 4 summarises the performance of the different methods in the unrestricted setting using SIFT features. We can observe that the performance of the MkNN classifier with LMNN as a base metric is comparable to the best performance we have obtained using LDML. Moreover, both our approaches outperform the state-of-the-art methods, *i.e*. LDA-based, ITML and LMNN.

5.3. Comparison to the state-of-the-art

In this section, we compare with previously published results on LFW [13, 20, 24]. We used the strict protocol

Figure 4. Examples of positive pairs correctly classified using the MkNN classifier with LMNN as a base metric, but wrongly classified using the LMNN metric alone.

to calculate the ROC curve and accuracy for our method. Note that each published result combines its own feature extraction with its own machine learning technique, making any conclusion harder to draw than in the previous sections.

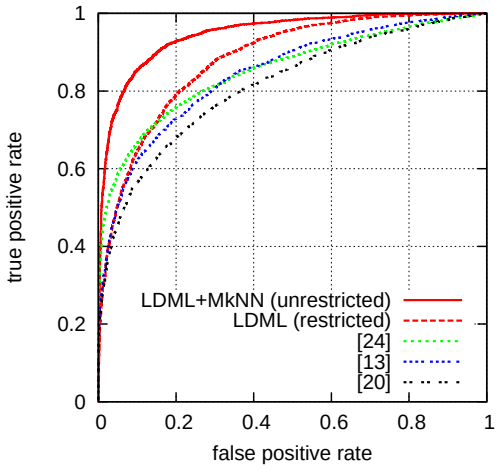
Following recent work [22, 24], we have linearly combined different scores to improve classification performance. In the restricted setting, we combine 4 descriptors (*cf.* Table 1) with the LDML metrics on the original data and its square root (8 scores). In the unrestricted setting, we combine the same inputs with the LDML, LMNN, and MkNN metrics (24 scores). The linear combinations are learnt using a logistic discriminant model for each fold independently. In the following and in Figure 5, we refer to these combined methods as LDML and LDML+MkNN.

The best result reported to date [24] attains 78.47% accuracy in the restricted setting also by combining several descriptors. LDML on the restricted setting obtains an accuracy of 79.27%. Shifting to the unrestricted setting, LDML+MkNN obtains a performance of 87.50%, which is significantly better than any previous result reported on this data set, showing the benefit of our metric learning approaches when using more training data. We observe that a combination of descriptors and metrics improves over using only one metric and one descriptor, highlighting the complementarity of our two approaches. We refer to Figure 1 for classification examples using our combined method.

6. Applications of learned face metrics

Here we show the merit of learned metrics for two applications: unsupervised clustering of face images, and face recognition from a single exemplar. We learn our metrics on 90000 pairs from 9 of the LFW folds, and apply them to faces in the held-out fold. The test fold contains 1369 faces from 601 people. In the following experiments, we focus on the 17 most frequent people (411 faces). We compare L2 and LDML on SIFT, and LDML+MkNN of Section 5.3.

Unsupervised hierarchical clustering of face images. We cluster the faces using complete-linkage hierarchical clustering. This method yields a hierachy of clusters by varying the maximum distance with which clusters can be



Method	Accuracy
Nowak, restricted [20]	73.93 ± 0.5
MERL+Nowak, restricted [13]	76.18 ± 0.6
Hybrid descriptor-based, restricted [24]	78.47 ± 0.5
LDML, restricted	$\mathbf{79.27 \pm 0.6}$
LDML+MkNN, unrestricted	$\mathbf{87.50 \pm 0.4}$

Figure 5. Comparison of our results with best results to date on the LFW data: ROC curves, and average accuracy and standard error.

merged. To compare clustering results we define a cost that reflects the labelling effort needed for a user to label the faces, *e.g.* for a personal photo album. We assume the user has two buttons: one to assign a single label to all faces in a cluster, and one to assign a label to a single face. The most efficient way to label all faces in a cluster is to first label the cluster with the name of the most frequent person, and then to correct the errors. For a cluster of N faces, the cost is $1 + (N - \max(\{n_i\}))$, where n_i denotes the number of faces of person i in the cluster. The cost to label all faces is then the sum of the costs to label the faces in each cluster. The optimal clustering has cost 17, a trivial over-clustering with a cluster for each face yields a cost of 411, while using a single cluster of all faces yields a cost of 341 as we have 71 images of the most frequent person.

In Figure 6 we show the costs as a function of the number of clusters using the L2 and LDML metrics on the SIFT data, the LDML+MkNN combination, the average for random clustering, and the minimum and maximum costs that can be obtained. Clearly, LDML yields much better clustering results than L2 for a wide range of number of clusters. For LDML the minimum cost of 109 is obtained with only 25 clusters, most of which are fairly pure. If we label the faces in each cluster by the identity of the most frequent person in that cluster then 75% of the faces are correctly classified. For the L2 metric the minimum cost is 233 for 135 clusters (92% correct but over-clustered). Combining the different descriptors with LDML leads to a de-

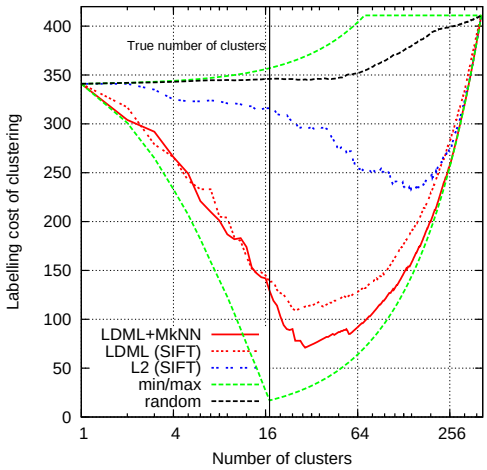


Figure 6. Labelling cost of clusterings using different metrics.

creased cost of 88 with 28 clusters (85% correct), and with LDML+MkNN the cost drops to 71 with 29 clusters (90% correct). In Figure 7 we show three example clusters from this clustering. Note that the clustering is successful despite big changes in expression, pose, and lighting.

Multi-class face recognition from single face exemplars. Here, we perform multi-class face recognition using a single, random training exemplar for each of the 17 people. We test classification accuracy on the remaining $1369 - 17 = 1352$ faces. A test face is assigned to the exemplar with the best score, or rejected if all scores are below a certain threshold. From the 1352 test faces only $411 - 17 = 394$ should be accepted as one of the 17 classes, and the remaining 958 should be rejected. We measure performance using precision at equal error rate, where the number of wrongly rejected faces equals the number of wrongly accepted faces. In Table 5 we present quantitative results showing that, as with the clustering, LDML leads to significantly better performance than L2 on the SIFT features: 39% of the accepted faces are correctly recognised, compared to only 14%. The LDML+MkNN combination boosts precision to 53%. In Figure 8 we show classification examples for LDML+MkNN.

7. Conclusion

We have introduced two new methods for visual identification: Logistic Discriminant Metric Learning (LDML), and Marginalised kNN classification (MkNN). We note that LDML can be trained from labelled *pairs* as provided in the restricted paradigm of LFW, where MkNN requires labelled training data and implicitly uses all pairs. The MkNN classifier is conceptually simple, but in practice it is computationally expensive as we need to find nearest neighbours in a large set of labelled data. This computational cost can

Figure 7. Three example clusters obtained using LDML+MkNN scores. The top two clusters are pure, and only few faces of these persons are assigned to incorrect clusters. The last cluster is typical, it contains a few faces from other people (the last 2).

Figure 8. Illustration of face recognition of 7 (out of 17) people using one training exemplar, with one person in each column. For each person we show: (a) the exemplar image, (b) a correctly recognised face of that person, (c) a non-recognised face of that person, and (d) another failure: an erroneously accepted face of another person.

Metric	L2	LDML	LDML+MkNN
Precision	14.0%	38.8%	53.3%
Faces of the 17 targets			
correctly recognised (b)	55	153	210
wrongly recognised	107	81	40
wrongly rejected (c)	232	160	144
Faces of other people			
correctly rejected	726	798	814
wrongly accepted (d)	232	160	144

Table 5. Comparison of one-exemplar classification performances. The test faces are broken down over the five possible situations. The labels (b)–(d) refer to the example images shown in Figure 8.

be alleviated by using efficient and/or approximate nearest neighbour search techniques.

LDML in combination with our descriptors yields a classification accuracy of 79.3% on the restricted setting of *Labeled Faces in the Wild* data set, where the best reported result so far was 78.5%. LDML and MkNN yield comparable accuracies on the unrestricted setting, around 83%. Remarkably, the gain when using the unrestricted setting is not observed with the current state-of-the-art method [24]. To our knowledge, we are the first to present results on the LFW data that follow and make good use of the unrestricted paradigm. Combining our methods, the accuracy is further improved to 87.5%. We also showed that metric learning leads to great improvements as compared to a simple L2 metric for applications of face similarities like clustering and recognition from a single exemplar.

Looking at the examples of failure cases of our method in Figure 1, pose changes remain one of the major challenges to be tackled in future work. Explicit modeling of invariance due to pose changes using techniques like those in [17] is an interesting option. Furthermore, we plan to apply our methods for automatic association of names and faces, and to other visual identification problems.

Acknowledgements

This work was supported by the European funded research project CLASS.

References

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahanalobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *CVPR*, 2004.

[3] D. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21(2):174–184, 1976.

[4] L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

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

[7] M. Everingham, J. Sivic, and A. Zisserman. 'Hello! My name is... Buffy' - automatic naming of characters in TV video. In *BMVC*, 2006.

[8] A. Ferencz, E. Learned-Miller, and J. Malik. Learning to locate informative features for visual identification. *IJCV*, 77:3–24, 2008.

[9] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007.

[10] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2006.

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. In *CVPR*, 2008.

[12] A. Holub, P. Moreels, and P. Perona. Unsupervised clustering for Google searches of celebrity images. In *IEEE Conference on Face and Gesture Recognition*, 2008.

[13] G. Huang, M. Jones, and E. Learned-Miller. LFW results using a combined Nowak plus MERL recognizer. In *Workshop on Faces Real-Life Images at ECCV*, 2008.

[14] G. 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.

[15] V. Jain, A. Ferencz, and E. Learned-Miller. Discriminative training of hyper-feature models for object identification. In *BMVC*, 2006.

[16] V. Jain, E. Learned-Miller, and A. McCallum. People-LDA: Anchoring topics to people using face recognition. In *ICCV*, 2007.

[17] M. Kumar, P. Torr, and A. Zisserman. An invariant large margin nearest neighbour classifier. In *ICCV*, 2007.

[18] F.-F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[20] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *CVPR*, 2007.

[21] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002.

[22] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

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

[24] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Workshop on Faces Real-Life Images at ECCV*, 2008.

[25] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2004.