Local Regularization for Multiclass Classification Facing Significant Intraclass Variations

Lior Wolf and Yoni Donner

The School of Computer Science
Tel Aviv Univerisy
Tel Aviv, Israel

Abstract. We propose a new local learning scheme that is based on the principle of decisiveness: the learned classifier is expected to exhibit large variability in the direction of the test example. We show how this principle leads to optimization functions in which the regularization term is modified, rather than the empirical loss term as in most local learning schemes. We combine this local learning method with a Canonical Correlation Analysis based classification method, which is shown to be similar to multiclass LDA. Finally, we show that the classification function can be computed efficiently by reusing the results of previous computations. In a variety of experiments on new and existing data sets, we demonstrate the effectiveness of the CCA based classification method compared to SVM and Nearest Neighbor classifiers, and show that the newly proposed local learning method improves it even further, and outperforms conventional local learning schemes.

1 Introduction

Object recognition systems, viewed as learning systems, face three major challenges: First, they are often required to discern between many objects; second, images taken under uncontrolled settings display large intraclass variation; and third, the number of training images provided is often small.

Previous attempts to overcome these challenges use prior generic knowledge on variations within objects classes [1], employ large amounts of unlabeled data (e.g., [2]), or reuse previously learned visual features [3]. Here, we propose a more generic solution, that does not assume nor benefit from the existence of prior learning stages or of an additional set of training images.

To deal with the challenge of multiple classes, we propose a Canonical Correlation Analysis (CCA) based classifier, which is a regularized version of a recently proposed method [4], and is highly related to Fisher Discriminant Analysis (LDA/FDA). We treat the other two challenges as one since large intraclass variations and limited training data both result in a training set that does not capture well the distribution of the input space. To overcome this, we propose a new local learning scheme which is based on the principle of decisiveness.

In local learning schemes, some of the training is deferred to the prediction phase, and a new classifier is trained for each new (test) example. Such schemes

have been introduced by [5] and were recently advanced and shown to be effective for modern object recognition applications [6] (see references therein for additional references to local learning methods). One key difference between our method and the previous contribution in the field is that we do not select or directly weigh the training examples by their proximity to the test point. Instead, we modify the objective function of the learning algorithm to reward components in the resulting classifier that are parallel to the test example. Thus, we encourage the classification function (before thresholding takes place) to be separated from zero.

Runtime is a major concern for local learning schemes, since a new classifier needs to be trained or adjusted for every new test example. We show how the proposed classifier can be efficiently computed by several rank-one updates to precomputed eigenvectors and eigenvalues of constant matrices, with the resulting time complexity being significantly lower than that of a full eigendecomposition. We conclude by showing the proposed methods to be effective on four varied datasets which exhibit large intraclass variations.

2 Multiclass classification via CCA

We examine the multiclass classification problem with k classes, where the goal is to construct a classifier given n training samples (x_i, y_i), with $x_i \in \mathbb{R}^m$ and $y_i \in \{1, 2, \ldots, k\}$. We assume $\sum_{i=1}^n x_i = 0$ (otherwise we center the data). Our approach is to find a transformation $T : \mathbb{R}^m \to \mathbb{R}^l$ and class vectors $v_j \in \mathbb{R}^l$ such that the transformed inputs $T(x_i)$ would be close to the class vector v_{y_i} corresponding to their class. Limiting the discussion at first to linear transformations, we represent T by a $m \times l$ matrix A such that $T(x) = A^\top x$. The formulation of the learning problem is therefore:

$$\min_{A, \{v_j\}_{j=1}^k} \sum_{i=1}^n \|A^\top x_i - v_{y_i}\|^2 \tag{1}$$

Define V to be the $k \times l$ matrix with v_j as its j'th row, so $v_j = V^\top e_j$. Also define $z_i = e_{y_i}$ where e_j is the j'th column of the identity $k \times k$ matrix I_k. Using these definitions, $v_{y_i} = V^\top z_i$ and Equation 1 becomes:

$$\min_{A,V} \sum_{i=1}^n \|A^\top x_i - V^\top z_i\|^2 \tag{2}$$

This expression can be further simplified by defining the matrices $X \in \mathbb{R}^{m \times n}$, $Z \in \mathbb{R}^{k \times n}$: $X = (x_1, x_2, \ldots, x_n)$, $Z = (z_1, z_2, \ldots, z_n)$. Equation 2 then becomes:

$$\min_{A,V} \operatorname{tr}(A^\top X X^\top A) + \operatorname{tr}(V^\top Z Z^\top V) - 2 \operatorname{tr}(A^\top X Z^\top V) \tag{3}$$

This expression is not invariant to arbitrary scaling of A and Z. Furthermore, we require the l components of the transformed vectors $A^\top x_i$ and $V^\top z_i$ to be pairwise uncorrelated since there is nothing to be gained by correlations between

them. Therefore, we add the constraints $A^\top XX^\top A = V^\top ZZ^\top V = I_l$, leading to the final problem formulation:

$$\begin{aligned} \max_{A,V} \quad & \mathrm{tr}(A^\top XZ^\top V) \\ \text{subject to } & A^\top XX^\top A = V^\top ZZ^\top V = I \end{aligned} \tag{4}$$

This problem is solved through Canonical Correlation Analysis (CCA) [7]. A simple solution involves writing the corresponding Lagrangian and setting the partial derivatives to zero, yielding the following generalized eigenproblem:

$$\begin{pmatrix} 0 & XZ^\top \\ ZX^\top & 0 \end{pmatrix} \begin{pmatrix} a_i \\ v_i \end{pmatrix} = \lambda_i \begin{pmatrix} XX^\top & 0 \\ 0 & ZZ^\top \end{pmatrix} \begin{pmatrix} a_i \\ v_i \end{pmatrix} \tag{5}$$

where λ_i, $i = 1..l$ are the leading generalized eigenvalues, a_i are the columns of A, and v_i are, as defined above, the columns of V. To classifying a new sample x, it is first transformed to $A^\top x$, and then compared to the k class vectors, i.e., the predicted class is given by $\arg\min_{1\leq j\leq k} ||A^\top x - v_j||$.

This classification scheme is readily extendable to non-linear functions that satisfy Mercer's conditions by using Kernel CCA [8,9]. Kernel CCA is also equivalent to solving a generalized eigenproblem of the form of Equation 5, so although we refer directly to linear CCA throughout this paper, our conclusions are equally valid for Kernel CCA.

In Kernel CCA, or in the linear case when $m > n$, and in many other common scenarios, the problem is ill-conditioned and regularization techniques are required [10]. For linear regression, ridge regularization is often used, as is its equivalent in CCA and Kernel CCA [8]. This involves replacing $XX^\top$ and $ZZ^\top$ in Equation 5 with $XX^\top + \eta_X I$ and $ZZ^\top + \eta_Z I$, where η_X and η_Z are regularization parameters. In the CCA case presented here, for multiclass classification, since the number of training examples n is not smaller than the number of classes k, regularization need not be used for Z and we set $\eta_Z = 0$. Also, since the X regularization is relative to the scale of the matrix $XX^\top$, we scale the regularization parameter η_X as a fraction of the largest eigenvalue of $XX^\top$.

The multiclass classification scheme via CCA presented here is equivalent to Fisher Discriminant Analysis (LDA). We provide a brief proof of this equivalence. A previous lemma was proven by Yamada et al [4] for the unregularized case.

Lemma 1. *The multiclass CCA classification method learns the same linear transformation as multiclass LDA.*

Proof. The generalized eigenvalue problem in Equation 5, with added ridge regularization, can be represented by the following two coupled equations:

$$(XX^\top + \eta I_m)^{-1} XZ^\top v = \lambda a \tag{6}$$

$$(ZZ^\top)^{-1} ZX^\top a = \lambda v \tag{7}$$

Any solution (a, v, λ) to the above system satisfies:

$$(XX^\top + \eta I_m)^{-1} XZ^\top (ZZ^\top)^{-1} ZX^\top a = (XX^\top + \eta I_m)^{-1} XZ^\top \lambda v = \lambda^2 a \quad (8)$$
$$(ZZ^\top)^{-1} ZX^\top (XX^\top + \eta I_m)^{-1} XZ^\top v = (ZZ^\top)^{-1} ZX^\top \lambda a = \lambda^2 v \quad (9)$$

Thus the columns of the matrix A are the eigenvectors corresponding to the largest eigenvalues of $(XX^\top + \eta I_m)^{-1} XZ^\top (ZZ^\top)^{-1} ZX^\top$. Examine the product $ZZ^\top = \sum_{i=n} e_{y_i} e_{y_i}^\top$. It is a $k \times k$ diagonal matrix with the number of training samples in each class (denoted N_i) along its diagonal. Therefore, $(ZZ^\top)^{-1} = \mathrm{diag}(\frac{1}{N_1}, \frac{1}{N_2}, \ldots, \frac{1}{N_k})$. Now examine $XZ^\top$: $(XZ^\top)_{i,j} = \sum_{s=1}^{n} X_{i,s} Z_{j,s} = \sum_{s:y_s=j} X_{i,s}$. Hence, the j'th column is the sum of all training samples of the class j. Denote by $\bar{X}_j$ the mean of the training samples belonging to the class j, then the j'th column of $XZ^\top$ is $N_j \bar{X}_j$. It follows that

$$XZ^\top (ZZ^\top)^{-1} ZX^\top = \sum_{j=1}^{k} \frac{N_j^2}{N_j} \bar{X}_j \bar{X}_j^\top = \sum_{j=1}^{k} N_j \bar{X}_j \bar{X}_j^\top = S_B \quad (10)$$

Where S_B is the between-class scatter matrix defined in LDA [11]. Let $S_T = XX^\top$ be the total scatter matrix S_T. $S_T = S_W + S_B$ (where S_W is LDA's within-class scatter matrix), and using S_T in LDA is equivalent to using S_W. Hence, the multiclass CCA formulation is equivalent to the eigen-decomposition of $(S_W + \eta I)^{-1} S_B$, which is the formulation of regularized multiclass LDA.

Our analysis below uses the CCA formulation; the LDA case is equivalent, with some minor modifications to the way the classification is done after the linear transformation is applied.

3 Local Learning via Regularization

The above formulation of the multiclass classification problem is independent of the test vector to be classified x. It may be the case that the learned classifier is "indifferent" to x, transforming it to a vector $A^\top x$ which has a low norm. Note that by the constraint $V^\top ZZ^\top V = I$, the norm of the class vectors v_j is $N_j^{-0.5}$ which is roughly constant for balanced data sets. This possible mismatch between the norm of the transformed example and the class vectors may significantly decrease the ability to accurately classify x. Furthermore, when the norm of $A^\top x$ is small, it is more sensitive to additive noise.

In local learning, the classifier may be different for each test sample and depends on it. In this work, we discourage classifiers that are indifferent to x, and have low $\|A^\top x\|^2$. Hence, to discourage indifference (increase decisiveness), we add a new term to the CCA problem:

$$\max_{A,V} \mathrm{tr}(A^\top XZ^\top V) + \bar{\alpha}\, \mathrm{tr}(A^\top xx^\top A)$$
$$\text{subject to} \quad A^\top XX^\top A = V^\top ZZ^\top V = I \quad (11)$$

$\mathrm{tr}(A^\top x x^\top A) = \|A^\top x\|^2$, and the added term reflects the principle of decisiveness. $\bar{\alpha}$ is a parameter corresponding to the trade-off between the correlation term and the decisiveness term. Adding ridge regularization as before to the solution of Equation 11, and setting $\alpha = \bar{\alpha}\lambda^{-1}$ gives the following generalized eigenproblem:

$$\begin{pmatrix} 0 & XZ^\top \\ ZX^\top & 0 \end{pmatrix} \begin{pmatrix} a \\ v \end{pmatrix} = \lambda \begin{pmatrix} XX^\top + \eta I - \alpha x x^\top & 0 \\ 0 & ZZ^\top \end{pmatrix} \begin{pmatrix} a \\ v \end{pmatrix} \tag{12}$$

Note that this form if similar to the CCA based multiclass classifier presented in Section 2 above, except that the ridge regularization matrix ηI is replaced by the local regularization matrix $\eta I - \alpha x x^\top$. We proceed to analyze the significance of this form of local regularization. In ridge regression, the influence of all eigenvectors is weakened uniformly by adding η to all eigenvalues before computation of the inverse. This form of regularization encourages smoothness in the learned transformation. In our version of local regularization, smoothness is still achieved by the addition of η to all eigenvalues. The smoothing effect is weakened, however, by α, in the component parallel to x. This can be seen by the representation $x x^\top = U_x \lambda_x U_x \top$ for $U_x^\top U_x = U_x U_x^\top = I$, with $\lambda_x = \mathrm{diag}(\|x\|^2, 0, \ldots, 0)$. Now $\eta I - \alpha x x^\top = U_x(\eta I - \alpha\lambda_x)U_x^\top$, and the eigenvalues of the regularization matrix are $(\eta - \alpha, \eta, \eta, \ldots, \eta)$. Hence, the component parallel to x is multiplied by $\eta - \alpha$ while all others are multiplied by η. Therefore, encouraging decisiveness by adding the term $\alpha\|A^\top x\|^2$ to the maximization goal is a form of regularization where the component parallel to x is smoothed less than the other components.

4 Efficient implementation

In this section we analyze the computational complexity of our method, and propose an efficient update algorithm that allows it to be performed in time comparable to standard CCA with ridge regularization. Our algorithm avoids fully retraining the classifier for each testing example by training it once using standard CCA with uniform ridge regularization, and reusing the results in the computation of the local classifiers.

Efficient training of a uniformly regularized multiclass CCA classifier. In the non-local case, training a multiclass CCA classifier consists of solving Equations 6 and 7, or, equivalently, Equations 8 and 9. Let $r = \min(m, k)$, and note that we assume $m \leq n$, since the rank of the data matrix is at most n, and if $m > n$ we can change basis to a more compact representation. To solve Equations 8 and 9, it is enough to find the eigenvalues and eigenvectors of a $r \times r$ square matrix. Inverting $(XX^\top + \eta I_m)^{-1}$ and $(ZZ^\top)^{-1}$ and reconstructing the full classifier (A and V) given the eigenvalues and eigenvectors of the $r \times r$ matrix above can be done in $O(m^3 + k^3)$. While this may be a reasonable effort if done once, it may become prohibitive if done repeatedly for each new test example. This, however, as we show below, is not necessary.

Representing the local learning problem as a rank-one modification. We first show the problem to be equivalent to the Singular Value Decomposition

(SVD) of a (non-symmetric) matrix, which is in turn equivalent to the eigen-decomposition of two symmetric matrices. We then prove that one of these two matrices can be represented explicitly as a rank-one update to a constant (with regards to the new test example) matrix whose eigen-decomposition is computed only once. Finally, we show how to efficiently compute the eigen-decomposition of the modified matrix, how to derive the full solution using this decomposition and how to classify the new example in time complexity much lower than that of a full SVD.

Begin with a change of variables. Let $\bar{A} = (XX^\top + \eta I_m - \alpha x x^\top)^{\frac{1}{2}} A$ and $\bar{V} = (ZZ^\top)^{\frac{1}{2}} V$. By the constraints (Equation 11, with added ridge and local regularizations), $\bar{A}$ and $\bar{V}$ satisfy $\bar{A}^\top \bar{A} = A^\top (XX^\top + \eta I_m - \alpha x x^\top) A = I$ and $\bar{V}^\top \bar{V} = V^\top Z Z^\top V = I$. Hence, the new variables are orthonormal and the CCA problem formulation (Equation 4) with added ridge regularization becomes:

$$\begin{aligned} &\max_{\bar{A},\bar{V}} \operatorname{tr}(\bar{A}^\top (XX^\top + \eta I_m - \alpha x x^\top)^{-\frac{1}{2}} X Z^\top (ZZ^\top)^{-\frac{1}{2}} \bar{V}) \\ &\text{subject to} \qquad \bar{A}^\top \bar{A} = \bar{V}^\top \bar{V} = I \end{aligned} \tag{13}$$

Define:

$$M_0 = (XX^\top + \eta I_m)^{-\frac{1}{2}} X Z^\top (ZZ^\top)^{-\frac{1}{2}} = U_0 \Sigma_0 R_0^\top \tag{14}$$

$$M = (XX^\top + \eta I_m - \alpha x x^\top)^{-\frac{1}{2}} X Z^\top (ZZ^\top)^{-\frac{1}{2}} = U \Sigma R^\top \tag{15}$$

where $U\Sigma R^\top$ is the Singular Value Decomposition (SVD) of M and similarly $U_0 \Sigma_0 R_0^\top$ for M_0. Then the maximization term of Equation 13 is $\bar{A}^\top U \Sigma R^\top \bar{V}$, which under the orthonormality constraints of Equation 13, and since we seek only l components, is maximized by $\bar{A} = U_{0|l}$ and $\bar{V} = R_{0|l}$, which are the l left and right singular vectors of M corresponding to the l largest singular values.

Since $M^\top M = R \Sigma^2 R^\top$, the right singular vectors can be found by the eigen-decomposition of the symmetric $M^\top M$. We proceed to show how $M^\top M$ can be represented explicitly as a rank-one update to $M_0^\top M_0$. Define $J_X = (XX^\top + \eta I_m)^{-1}$, then J_X is symmetric as the inverse of a symmetric matrix, and by the Sherman-Morrison formula [12],

$$\begin{aligned} &(XX^\top + \eta_X I_m - \alpha x x^\top)^{-1} = (J_X - \alpha x x^\top)^{-1} = J_X + \frac{J_X \alpha x x^\top J_X}{1 - \alpha x^\top J_X x} \\ &= J_X + \frac{\alpha}{1 - \alpha x^\top J_X x} (J_X x)(J_X x)^\top = (XX^\top + \eta_X I_m)^{-1} + \beta b b^\top \end{aligned} \tag{16}$$

where $\beta = \frac{\alpha}{1 - \alpha x^\top J_X x}$ and $b = J_X x$. β and b can both be computed using $O(m^2)$ operations, since J_X is known after being computed once. Now,

$$\begin{aligned} M^\top M &= (ZZ^\top)^{-\frac{1}{2}} Z X^\top (XX^\top + \eta I_m - \alpha x x^\top)^{-1} X Z^\top (ZZ^\top)^{-\frac{1}{2}} \\ &= (ZZ^\top)^{-\frac{1}{2}} Z X^\top \left((XX^\top + \eta I_m)^{-1} + \beta b b^\top \right) X Z^\top (ZZ^\top)^{-\frac{1}{2}} \\ &= M_0^\top M_0 + \beta (ZZ^\top)^{-\frac{1}{2}} Z X^\top b b^\top X Z^\top (ZZ^\top)^{-\frac{1}{2}} \\ &= M_0^\top M_0 + \beta c c^\top \end{aligned} \tag{17}$$

$$\tag{18}$$

where $c = (ZZ^\top)^{-\frac{1}{2}} ZX^\top b$, and again c is easily computed from b in $O(km)$ operations. Now let $w = R_0^\top \frac{c}{\|c\|}$ (so $\|w\| = 1$) and $\gamma = \beta\|c\|^2$ to arrive at the representation

$$M^\top M = R_0(\Sigma_0^2 + \gamma w w^\top)R_0^\top \tag{19}$$

It is left to show how to efficiently compute the eigen-decomposition of a rank-one update to a symmetric matrix, whose eigen-decomposition is known. This problem has been investigated by Golub [13] and Bunch et al. [14]. We propose a simple and efficient algorithm that expands on their work. We briefly state their main results, without proofs, which can be found in the original papers.

The first stage in the algorithm described in Bunch et al. [14] is deflation, transforming the problem to equivalent (and no larger) problems $S+\rho zz^\top$ satisfying that all elements of z are nonzero, and all elements of S are distinct. Then, under the conditions guaranteed by the deflation stage, the new eigenvalues can be found. The eigenvalues of $S+\rho zz^\top$ satisfying that all elements of z are nonzero and all elements of S are distinct are the roots of $f(\lambda) = 1+\rho \sum_{i=1}^{s} \frac{z_i^2}{d_i-\lambda}$, where s is the size of the deflated problem, z_i are the elements of z and d_i are the elements of the diagonal of S. [14] show an iterative algorithm with a quadratic rate of convergence, so all eigenvalues can be found using $O(s^2)$ operations, with a very small constant as shown in their experiments. Since the deflated problem is no larger than k, this stage requires $O(k^2)$ operations at most. Once the eigenvalues have been found, the eigenvectors of $\Sigma_0^2 + \gamma w w^\top$ can be computed by

$$\xi_i = \frac{(S-\lambda_i I)^{-1} z}{\|(S-\lambda_i I)^{-1} z\|} \tag{20}$$

using $O(k)$ operations for each eigenvector, and $O(k^2)$ in total to arrive at the representation

$$M^\top M = R_0 R_1 \Sigma_1 R_1^\top R_0 \tag{21}$$

Explicit evaluation of Equation 21 to find $\hat{V}$ requires multiplying $k \times k$, which should be avoided to keep the complexity $O(m^2 + k^2)$. The key observation is that we do not need to find V explicitly but only $A^\top x - v_i$ for $i = 1, 2, \ldots, k$, with v_i being the i'th class vector (Equation 1). The distances we seek are:

$$\|A^\top x - v_i\|^2 = \|A^\top x\|^2 + \|v_i\|^2 - 2v_i^\top A^\top x \tag{22}$$

with $\|v_i\|^2 = N_i$ (see Section 3). Hence, finding all exact distances can be done by computation of $x^\top AA^\top x - VA^\top x$, since $v_i^\top$ is the i'th row of V. Transforming back from $\bar{V}$ to V gives $V = (ZZ^\top)^{-\frac{1}{2}}\bar{V}$, where $(ZZ^\top)^{-\frac{1}{2}}$ needs to be computed only once. From Equations 6 and 21,

$$\begin{aligned} A^\top x &= \Sigma_1^{-1} V^\top ZX^\top (XX^\top + \eta I_m - \alpha x x^\top)^{-1} x \\ &= \Sigma_1^{-1} R_1^\top R_0^\top (ZZ^\top)^{-\frac{1}{2}} ZX^\top \left((XX^\top + \eta I_m)^{-1} + \beta bb^\top\right) x \end{aligned} \tag{23}$$

All the matrices in Equation 23 are known after the first $O(k^3+m^3)$ computation and $O(k^2+m^2)$ additional operations per test example, as we have shown

above. Hence, $A^\top x$ can be computed by a sequence of matrix-vector multiplications in time $O(k^2 + m^2)$, and similarly for

$$VA^\top x = (ZZ^\top)^{-\frac{1}{2}} R_0 R_1 A^\top x \tag{24}$$

Thus, the distances of the transformed test vector x from all class vectors can be computed in time $O(m^2 + k^2)$, which is far quicker than $O(m^3 + k^3)$ which is required by training the classifier from scratch, using a full SVD. Note that the transformation of a new vector without local regularization requires $O(ml)$ operations, and the classification itself $O(kl)$ operations. The difference between the classification times of a new test vector using local regularization, therefore, is $O(m^2 + k^2)$ compared to $O((m + k)l)$ using uniform regularization.

5 Experiments

We report results on 3 data sets: a new Dog Breed data set, the CalPhotos Mammals collection [15], and the "Labeled Faces in the Wild" face recognition data set [16]. These data sets exhibit a large amount of intraclass variation.

The experiments in all cases are similar and consist of multiclass classification. We compare the following algorithms: Nearest Neighbor, Linear All-Vs-All SVM (a.k.a "pairwise", "All-Pairs"), Multiclass CCA (the method of Section 2), and Local Multiclass CCA (Section 3). The choice of using All-Vs-All SVM is based on its simplicity and relative efficiency. A partial set of experiments verified that One-Vs-All SVM classifiers perform similarly. It is well established in the literature that the performance of other multiclass SVM schemes is largely similar [6,17]. Similar to other work in object recognition we found Gaussian-kernel SVM to be ineffective, and to perform worse than Linear SVM for every kernel parameter we tried. Evaluating the performance of non-linear versions of Multiclass CCA and Local Multiclass CCA is left for future work.

We also compare the conventional local learning scheme [5], which was developed further in [6]. In this scheme the k nearest neighbors of each test point are used to train a classifier. In our experiments we have scanned over a large range possible neighborhood sizes k to verify that this scheme does not outperform our local learning method regardless of k. Due to the computational demands of such tests, they were only performed on two out of the four data sets.

Each of the described experiments was repeated 20 times. In each repetition a new split to training and testing examples was randomized, and the same splits were used for all algorithms. Note that due to the large intraclass variation, the standard deviation of the result is typically large. Therefore, we use paired t-tests to verify that the reported results are statistically significant.

Parameter selection. The regularization parameter of the linear SVM algorithm was selected by a 5-fold cross-validation. Performance, however, is pretty stable with respect to this parameter. The regularization parameter of Multiclass CCA and Local Multiclass CCA η was fixed at 0.1 times the leading eigenvalue of $XX^\top$, a value which seems to be robust in a large variety of synthetic and real

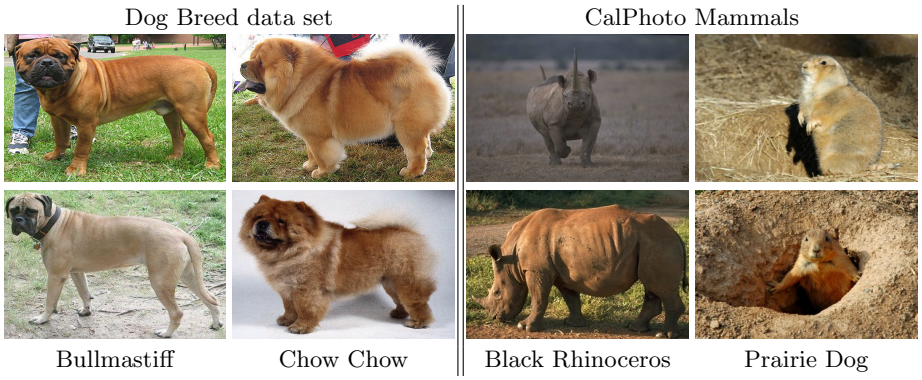


Fig. 1. Sample images from the Dog Breed and CalPhoto Mammal data sets.

data sets. The local regularization parameter β was set at 0.5η in all experiments, except for the ones done to evaluate its effect on performance.

Image representation. The visual descriptors of the images in the Dog Breed and CalPhotos Mammels data sets are computed by the Bag-of-SIFT implementation of Andrea Vendaldi [18]. This implementation uses hierarchical K-means [19] for partitioning the descriptor space. Keypoints are selected at random locations [20]. Note that the dictionary for this representation was recomputed at each run in order to avoid the use of testing data during training. Using the default parameters, this representation results in vectors of length $11,111$

The images in the face data set are represented using the Local Binary Pattern [21] image descriptor, which were adopted to face identification by [22]. An LBP is created at a particular pixel location by thresholding the 3×3 neighborhood surrounding the pixel with the central pixels intensity value, and treating the subsequent pattern as a binary number. Following [22], we set a radius of 2 and sample at the boundaries of 5 pixel blocks, and bin all patterns for which there are more than 2 transition from 0 to 1 in just one bin. LBP representations for a given image are generated by dividing an image into several windows and creating histograms of the LBPs within each window.

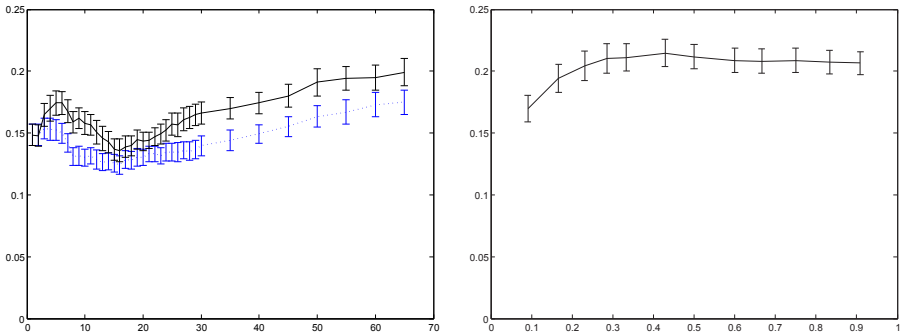
5.1 Results on individual data sets

Dog Breed images. The Dog Breed data set contains images of 34 dog species, with 4–7 photographs each, a total of 177 images. The images were collected from the internet, and as can be seen in Figure 1 are quite diverse.

Table 1 compares the classification results for a varying number of training/testing examples per breed. The results demonstrate that Local Multiclass CCA performs better than Multiclass CCA, which in turn performs better than Nearest Neighbor and SVM. Since the images vary significantly, the results exhibit a large variance. Still, all differences in the table are significant ($p < 0.01$), except for the difference between Multiclass CCA and SVM in the case of 3 training images per breed.

Table 1. Mean (± standard deviation) recognition rates (in percents) for the Dog Breed data set. Each column is for a different number of training and testing examples per breed for the 34 dog breeds.

Algorithm	1 training / 3 test	2 training / 2 test	3 training / 1 test
Nearest Neighbor	11.03 ± 1.71	14.85 ± 3.96	18.68 ± 6.35
All-Pairs Linear SVM	11.03 ± 1.71	17.50 ± 4.37	23.82 ± 6.32
Multiclass CCA	13.43 ± 3.56	19.63 ± 4.99	24.12 ± 6.92
Local Multiclass CCA	15.78 ± 3.63	21.25 ± 4.56	26.18 ± 6.39



(a) (b)

Fig. 2. Mean performance and standard deviation (normalized by $\sqrt{20}$) for additional experiments on the Dog Breed data set. (a) k-nearest neighbors based local learning. The x axis depicts k, the size of the neighberhood. Top line – the performance of the Multiclass CCA classifier, Bottom dashed line – the performance of SVM. (b) Performance for various values of the local regularization parameter. The x axis depicts the ratio of β and η.

To further understand the nature of the local learning method we performed two additional more experiments. Figure 2(a) demonstrates that the conventional local learning scheme, based on k-nearest neighbors does not seem to improve performance for any values of k. Figure 2(b) demonstrates that the performance of the Local CCA method is stable with respect to the additional parameter α.

CalPhoto Mammals. The mammal collection of the CalPhoto image repository [15] contains thousands of images. After filtering out all images for which the Latin species name does not appear and species for which there are less than 4 images, 3,740 images of 256 species remain. For each species, the images vary considerably, as can be seen in Figure 1.

In each experiment 10, 20 or 40 random species are selected. Each contributes 2 random training images and 2 test ones. Table 2 compares the classification results. Once again, Local Multiclass CCA outperforms the uniform Multiclass CCA, followed by SVM and NN. All performance differences in the table are statistically significant, except for SVM and Multiclass CCA for 40 classes.

Table 2. Mean (± standard deviation) recognition rates (percents) for the Mammals data set. Each column is for a different number of random classes per experiment. Each experiment was repeated 20 times.

Algorithm	10 classes	20 classes	40 classes
Nearest Neighbor	25.50 ± 8.57	20.25 ± 7.86	14.13 ± 3.89
All-Pairs Linear SVM	28.75 ± 10.87	25.38 ± 9.22	17.13 ± 4.20
Multiclass CCA	33.00 ± 11.63	28.75 ± 9.78	18.88 ± 4.81
Local Multiclass CCA	36.00 ± 11.19	31.87 ± 10.06	21.00 ± 5.48

Labeled Faces in the Wild. From the Labeled Faces in the Wild dataset [16], we filtered out all persons which have less than four images. 610 persons and a total of $6,733$ images remain. The images are partly aligned via funneling [23], and all images are 256×256 pixels. We only use the center 100×100 sub-image, and represent it by LBP features of a grid of non-overlapping 16 pixels blocks.

The number of persons per experiment vary from 10 to 100. For each run, 10, 20, 50 or 100 random persons and 4 random images per person are selected. 2 are used for training and 2 for testing. Table 3 compares the classification results. While the differences may seem small, they are significant ($p < 0.01$) and Local Multiclass CCA leads the performance table followed by Multiclass CCA and either NN or SVM. Additional experiments conducted for the 50 persons split show that k-nearest neighbors based local learning hurts performance for all values of k, for both SVM and Multiclass CCA.

Table 3. Mean (± STD) recognition rates (percents) for "Labeled Faces in the Wild". Columns differ in the number of random persons per experiment.

Algorithm	10 persons	20 persons	50 persons	100 persons
Nearest Neighbor	36.00 ± 12.73	25.25 ± 7.20	18.10 ± 3.77	15.27 ± 1.90
All-Pairs Linear SVM	35.00 ± 13.67	24.37 ± 5.55	18.55 ± 3.91	14.10 ± 2.39
Multiclass CCA	40.50 ± 14.68	29.25 ± 6.93	24.15 ± 5.51	20.55 ± 2.99
Local Multiclass CCA	41.25 ± 14.77	31.25 ± 6.46	25.70 ± 5.07	21.40 ± 3.02

Acknowledgments

This research is supported by the Israel Science Foundation (grants No. 1440/06, 1214/06), the Colton Foundation, and a Raymond and Beverly Sackler Career Development Chair.

References

1. Fei-Fei, L., Fergus, R., Perona, P.: A bayesian approach to unsupervised one-shot learning of object categories. In: ICCV, Nice, France (2003) 1134–1141
2. Belkin, M., Niyogi, P.: Semi-supervised learning on riemannian manifolds. Machine Learning **56** (2004) 209–239
3. Bart, E., Ullman, S.: Cross-generalization: learning novel classes from a single example by feature replacement. In: CVPR. (2005)
4. Yamada, M., Pezeshki, A., Azimi-Sadjadi, M.: Relation between kernel cca and kernel fda. In: IEEE International Joint Conference on Neural Networks. (2005)
5. Bottou, L., Vapnik, V.: Local learning algorithms. Neural Computation **4** (1992)
6. Zhang, H., Berg, A.C., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: CVPR. (2006)
7. Hotelling, H.: Relations between two sets of variates. Biometrika **28** (1936) 321–377
8. Akaho, S.: A kernel method for canonical correlation analysis. In: International Meeting of Psychometric Society. (2001)
9. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. J. Mach. Learn. Res. **4** (2003) 913–931
10. Neumaier, A.: Solving ill-conditioned and singular linear systems: A tutorial on regularization (1998)
11. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference and prediction. Springer (2001)
12. Sherman, J., Morrison, W.J.: Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. Annals of Mathematical Statistics **20** (1949) 621
13. Golub, G.: Some modified eigenvalue problems. Technical report, Stanford. (1971)
14. Bunch, J.R., Nielsen, C.P., Sorensen, D.C.: Rank-one modification of the symmetric eigenproblem. Numerische Mathematik **31** (1978) 31–48
15. CalPhotos: A database of photos of plants, animals, habitats and other natural history subjects [web application], animal–mammals collection. bscit, university of california, berkeley. (Available: `http://calphotos.berkeley.edu/cgi/img_query?query_src=photos_index&where-lifeform=Animal--Mammal`)
16. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49 (2007)
17. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research **5** (2004)
18. Vedaldi, A.: Bag of features: A simple bag of features classifier. Available: `http://vision.ucla.edu/~vedaldi/` (2007)
19. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006)
20. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: European Conference on Computer Vision, Springer (2006)
21. Ojala, T., Pietikainen, M., Harwood, D.: A comparative-study of texture measures with classification based on feature distributions. Pattern Recognition **29** (1996)
22. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: ECCV. (2004)
23. Huang, G.B., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. ICCV (2007)