Sparse Kernel Learning for Image Set Classification

Muhammad Uzair, Arif Mahmood and Ajmal Mian

Computer Science & Software Engineering
The University of Western Australia
35 Stirling Highway, Crawley, WA, Australia
muhammad.uzair@research.uwa.edu.au
{arif.mahmood,ajmal.mian}@uwa.edu.au

Abstract. No single universal image set representation can efficiently encode all types of image set variations. In the absence of expensive validation data, automatically ranking representations with respect to performance is a challenging task. We propose a sparse kernel learning algorithm for automatic selection and integration of the most discriminative subset of kernels derived from different image set representations. By optimizing a sparse linear discriminant analysis criterion, we learn a unified kernel from the linear combination of the best kernels only. Kernel discriminant analysis is then performed on the unified kernel. Experiments on four standard datasets show that the proposed algorithm outperforms current state-of-the-art image set classification and kernel learning algorithms.

1 Introduction

In image-set classification, labelled training data consists of one or more sets per class where each set contains multiple images of the same class. The test set also contains multiple instances of the same class and is assigned the label of the nearest training set by maximizing some similarity measure [1–7]. Image set classification is useful in a wide range of applications including video-based face recognition, video surveillance, person re-identification in camera networks and object categorization.

Image-set classification is often performed in two steps. First, a representation is used to encode the intra-image as well as inter-image variations within a set based on some assumptions on the set structure. In the second step, the similarity between the image-set representations is measured, usually under certain constraints such as sparsity. Classification accuracy strongly depends on the specific set representation and the underlying assumptions and constraints. Most researchers focus on finding accurate image set representations. However, no single universal set representation can efficiently encode all types of image set variations. Image set representations make assumptions about the underlying data. Some assume that the underlying set data is single mode Gaussian [8, 9, 7, 10] whereas it may be multi-modal or non-Gaussian. Others assume that an

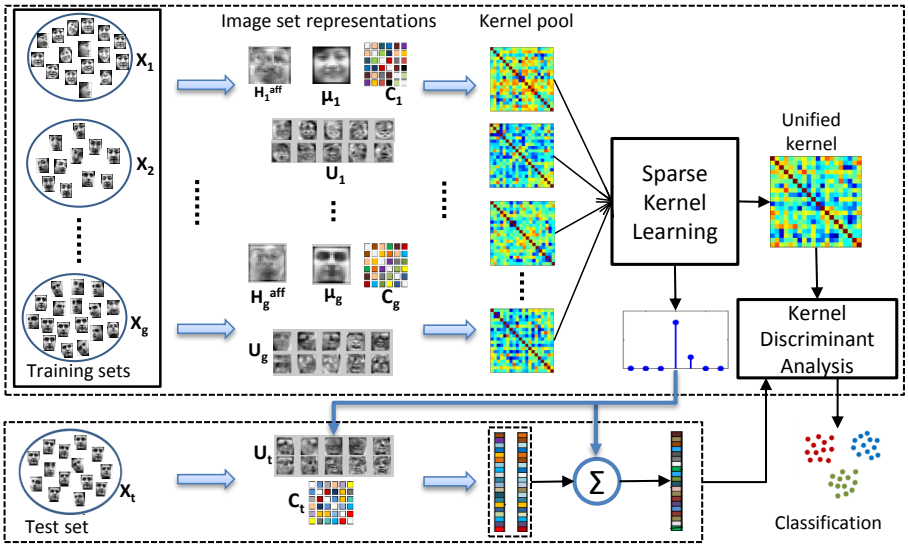
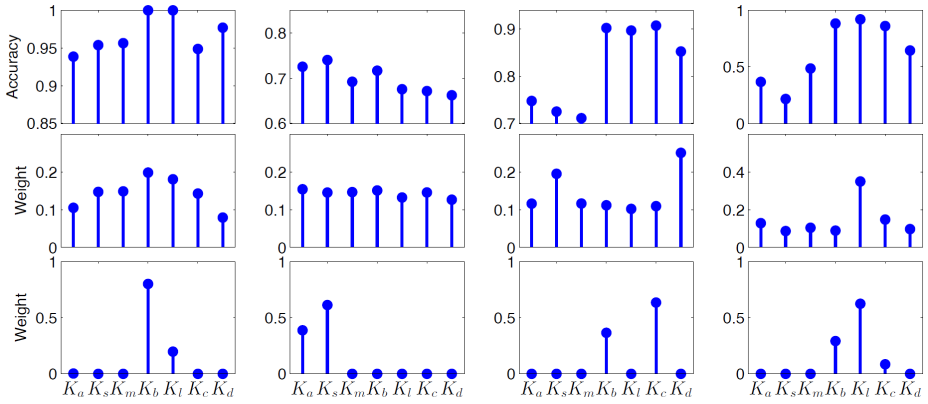


Fig. 1. Illustration of Sparse Kernel Learning. H_i^{aff}, μ_i, C_i and U_i are the affine hull, mean, covariance and subspace representations of image sets X_i. During training, SKL computes a pool of base kernels from different image set representations and automatically learns the best unified kernel from their sparse combination. In the test stage, only the kernels corresponding to the non-zero weights are computed.

image set can be represented by linear subspace bases [5, 11] whereas the actual data may lie on complex manifolds [2, 6]. Moreover, in the presence of only few images per set, the estimation of subspace bases and manifold parameters may be inaccurate. Some image set classification algorithms [3, 4] are variants of nearest neighbour whereas the image sets may overlap in some low dimensional space. Thus, no single representation performs good in all cases. In the absence of validation data, automatic selection of the most discriminative representations is a challenging problem. Moreover, there is a lack of systematic procedure for the selection and integration of efficient image set representations.

One solution to the above problem is along the lines of Multiple Kernel Learning (MKL) [12] where different types of features are expressed in terms of kernels and effectively integrated for improving classification. These have been applied to different computer vision tasks such as object categorization [13], object detection [14], multi-class object classification [15] and image set classification [1, 16]. In the case of image set classification, a pool of base kernels can be derived using different image set representations and their associated distance measures. A unified kernel can then be learned from their combination. Recently, Lin et al. [17] proposed a multiple kernel learning algorithm for dimensionality reduction (MKLDR). In MKLDR algorithm, the image data is represented by different features from which a set of base kernels is derived. The weighted combination of these base kernels are then used to learn a discriminative low dimensional subspace for classification. In MKLDR, all features are considered important and the weights assigned to different kernels do not necessarily correspond to their exact performance ratios [17]. Thus, the kernel combination using this strategy



(a) Honda (b) Youtube (c) ETH-80 (d) Cambridge

Fig. 2. Sparse kernel learning (SKL): **Top row** Individual accuracies of the kernels derived from each image set representation using their associated distance measures. Affine hull [3] (K_a), Affine hull [4] (SANP distance K_s), mean (K_m), subspace bases (K_b), Covariance (log-Euclidean kernel K_l), Covariance (Cholesky kernel K_c), Manifold [2] (MMD kernel K_d) (see Section 3) **Middle row** Weights learned by the MKLDR algorithm [17]. **Bottom row** Sparse weights learned by the proposed algorithm.

can reduce the overall classification accuracy (see Table 2) because a higher weight assigned to a poor feature degrades the quality of the overall mixture.

We propose a sparse kernel learning (SKL) algorithm that automatically learns a subset of the most discriminative base kernels derived from a pool of image set representations and their associated distance measures (see Fig. 1). Given a large number of image set kernels, our goal is to learn sparse kernel weights, without using validation data, such that the sparse combination of these kernels minimize intra-set distances and maximize inter-set distance. To the best of our knowledge, sparse kernel learning has not been formulated previously for image set classification. We impose sparsity on the kernel learning such that poor performing kernels are discarded and the final mixture is more discriminative. An additional advantage of sparsity is that only a few kernels are required to be computed at runtime. Fig. 2 shows the effectiveness of the proposed SKL algorithm. In the case of Youtube dataset, the MKLDR algorithm assigned weights to all the image set kernels which has degraded the overall accuracy. In the case of ETH-80 dataset, MKLDR assigned the highest weight to the MMD kernel K_d (derived from the manifold representation). However, the MMD kernel is not the best performer on this dataset. In contrast, MKL automatically learns a subset of the most discriminative kernels (subspace kernel K_b and Cholesky kernel K_c) while assigning zero weights to the others.

The SKL objective function is formulated as a graph embedding linear discriminant analysis criterion with ℓ_1 norm regularization. The enforcement of sparsity ensures that only the most discriminant image set kernels get non-zero weights. Once we obtain the unified kernel by the sparse linear combination

of the most discriminative image set kernels, we perform Kernel Linear Discriminant Analysis (KLDA) based classification. For experiments, we use four standard datasets and derive seven image set kernels. Our results outperform the MKLDR algorithm [17] as well as seven state-of-the-art image set classification algorithms.

2 Proposed Method

2.1 Problem Formulation

Let $G \equiv \{X_j\}_{j=1}^{g} \in \mathcal{R}^{d \times N}$ be the gallery containing g image sets and N images: $N = \sum_{j=1}^{g} n_j$, where n_j is the number of images in the j-th image set. Let $X_j = \{x_j^i\}_{i=1}^{n_j} \in \mathcal{R}^{d \times n_j}$ be the j-th image set, where $x_j^i \in \mathcal{R}^d$ is a d dimensional feature vector obtained by lexicographic ordering of the pixel elements of the i-th image in the j-th set. Instead of pixel values, the vector x_j^i may also contain feature values such as LBP or HoG features. The value of n_j may vary across image sets while the dimensionality of x_j^i remains fixed. Let c be the number of object classes and $Y = \{y_j\}_{j=1}^{g}$ be the class labels of the image sets in G. A distance matrix $d_r \in \mathcal{R}^{g \times g}$ is obtained for the gallery G such that $d_r(i,j) = d_r(X_i, X_j)$ is the distance between sets X_i and X_j using a distance measure r. Let R be the total number of distance measures each of which generating a distance matrix d_r. We convert each distance matrix d_r to a kernel matrix K_r using the Gaussian function as

$$K_r(i,j) = e^{\left(\frac{-d_r(i,j)}{\sigma_r^2}\right)}, \tag{1}$$

where σ_r is a Gaussian scale factor. The ith column of the kernel matrix K_r shows the relative position of the set X_i w.r.t. all other sets in the gallery. Therefore, we consider $K_r(i) \in \mathcal{R}^g$ a feature vector describing the set X_i. For R different distance measures, X_i has R different features descriptions. Our goal is to select a subset $L < R$ features such that when they are combined, their overall discrimination capability is maximized. For this purpose we propose to use the graph embedding linear discriminant analysis with sparsity constraints.

2.2 Sparse Kernel Learning

We represent X_i with a tensor $T_i = [K_1(i), ..., K_R(i)] \in \mathcal{R}^{g \times R}$ which is formed by concatenating all feature descriptors of X_i. For the gallery G we have g such matrices $G_t \equiv \{T_i\}_{i=1}^{g}$. For the graph embedding linear discriminant analysis the within class scatter matrix S_w and the between class scatter matrix S_b are formulated in a pairwise manner

$$S_w = \sum_{i,j=1}^{g} w_{ij} (T_i - T_j)^\top (T_i - T_j), \tag{2}$$

$$S_b = \sum_{i,j=1}^{g} \acute{w}_{ij} (T_i - T_j)^\top (T_i - T_j). \tag{3}$$

where $w_{ij} = \begin{cases} 1/n_k & \text{if } (T_i, T_j) \in c_k, \\ 0 & \text{otherwise,} \end{cases}$ and $\acute{w}_{ij} = 1/g$, n_k are the number of image sets in class c_k with label y_k.

In conventional graph embedding discriminant analysis [18], a projection matrix is learned such that the between-class similarity is minimized and the within-class similarity is maximized. In contrast, we formulate a sparse linear discriminant analysis criterion to learn an optimal linear combination of different features $K_r(i)$. We make the weight learning sparse so that the non-discriminative feature descriptors get zero weights while the more discriminative ones get high weights. Therefore, we formulate the following objective function for performing sparse discriminant analysis on S_w and S_b. The aim is to maximize the linear discriminant objective function with additional ℓ_1 and ℓ_2 norms regularizations:

$$\min_{Q} \quad \big(\text{trace}(Q^\top (S_w - S_b)Q) + \sum_j \lambda_j \|q_j\|_1 + \alpha \|Q\|_2^2\big) \quad \text{s.t.} \quad Q^\top Q = I, \tag{4}$$

where q_j is the jth column of Q, α is a constant and λ_j are the coefficients of ℓ_1 norm. Minimizing the scatter difference term means that the optimal projections Q^* should be able to minimize the within-class scatter S_w and maximize the between-class scatter S_b. The scatter difference term of the above objective function (4) is similar to the Max Margin Criterion [19] whereas the ℓ_1 norm regularization is added to ensure sparse solutions and the term $\alpha\|Q\|_2^2$ is the positive ridge penalty. The approximate sparse solutions of (4) can be obtained by rewriting the objective function as a set of Sparse PCA criteria [20]:

$$\min_{Q} \quad \|\Psi^\top D - UQ^\top D\|^2 + \sum_j \lambda_j \|q_j\|_1 + \alpha \|Q\|_2^2 \quad \text{s.t.} \quad U^\top U = I. \tag{5}$$

where Ψ and D are obtained from the SVD of $S_w - S_b$: $S_w - S_b = \Psi \Sigma \Psi^\top$ and $D = \Psi \sqrt{abs(\Sigma)} \Psi^\top$. Algorithm 1 shows the proposed method to solve the objective function in 5.

Once the termination criteria in the algorithm 1 is met, we obtain a final sparse projection matrix Q which is computed by SVD. The weight vector $\theta \in \mathcal{R}^R$ corresponds to the most dominant eigenvector in the projection matrix Q. The index i of the weight vector θ contains the weight of the feature $K_r(i)$. We ignore the sign of individual coefficients in θ by taking its absolute. Finally, θ is used to obtain a sparse linear combination of different image set kernels.

2.3 Kernel Discriminant Analysis based Classification

Since each K_r is symmetric, therefore; they can be converted to valid kernel matrices and subsequently used in a kernel based classification such as Kernel Linear Discriminant Analysis. Moreover, each K_r must be positive semidefinite to be a valid kernel matrix. This is not always guaranteed for each K_r. Therefore, we make K_r semipositive definite by simply adding a small perturbation to its

Algorithm 1 Sparse Kernel Learning Algorithm

Require: S_w and S_b from (2) and (3), C_t
Ensure: Weight vector θ
$L \Leftarrow S_w - S_b$
$L \Rightarrow \Psi \Sigma \Psi^\top$ {SVD of L}
$D \Leftarrow \Psi \sqrt{abs(\Sigma)} \Psi^\top$
$D \Rightarrow U \Lambda U^\top$ {SVD of D}
$Q^* = \mathbf{0}^{R \times R}$, $\Delta Q = 10^5$, $\epsilon = 10^{-3}, c = 0$
while $\Delta Q > \epsilon$ and $c < C_t$ **do**
$Q_{old} = Q^*$
$Q^* \equiv \min \|\Psi^\top D - U Q^\top D\|^2 + \sum_j \lambda_j \|q_j\|_1 + \alpha \|Q\|_2^2$ {Solve the Elastic Net problem}
$LQ^* \Rightarrow Q \Omega V^\top$ {SVD of LQ^*}
$U \Leftarrow QV^\top$
$\Delta Q = max(abs(Q_{old}(:) - Q^*(:)))$
$c = c + 1$
end while
$\theta \Leftarrow$ Dominant eigenvector in Q

diagonal (the absolute of its smallest non zero eigenvalue). After making all the K_r semipositive definite we can now linearly combine them in a weighted manner using θ to form a unified kernel matrix K

$$K = \sum_{r=1}^{R} \theta(r) K_r, \tag{6}$$

where θ is the sparse weight vector calculated using algorithm 1. From the theory of Reproducing Kernel Hilbert Space (RKHS) it is well known that the superposition of two valid kernels gives a new valid kernel [21]. Therefore, the proposed unified kernel K can be used with any kernel based learning algorithm to perform classification.

In this paper, we perform Kernel Linear Discriminant Analysis for classification. Having obtained K, KLDA seeks to solve the following optimization problem

$$\alpha_{opt} = \arg\max \frac{\alpha^\top K \mathcal{W} K \alpha}{\alpha^\top K K \alpha}, \tag{7}$$

where $\alpha = [\alpha_1, ..., \alpha_g]^\top$, and $\mathcal{W} \in \mathcal{R}^{g \times g}$ is a block diagonal matrix: $\mathcal{W} = diag\{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_c\}$, where $\mathcal{W}_j \in \mathcal{R}^{n_k \times n_k}$ is a matrix with all elements equal to $1/n_k$. The optimal α is given by the largest eigenvectors of the

$$(KK + \epsilon I)^{-1} (K \mathcal{W} K) \alpha = \lambda \alpha, \tag{8}$$

Note that K is often a full rank matrix however, this is not guaranteed. Therefore, a regularization term ϵ is used to ensure that KK remains invertible. By selecting the $(c-1)$ dominant eigenvectors from the solution of (8), we obtain a transformation matrix $\widehat{\alpha} = [\alpha_1, ..., \alpha_{c-1}]$. For a test image set X_t we first calculate $\mathcal{T}_t \in \mathcal{R}^{g \times R}$ where $\mathcal{T}_t(i) = K_i(G, X_t)$. The $c-1$ dimensional KLDA feature

vector $\mathcal{Y}_t$ of $\mathcal{T}_t$ in the discriminant subspace is computed as

$$\mathcal{Y}_t = \widehat{\alpha}^\top \mathcal{T}_t \theta. \tag{9}$$

Finally, to find the label of $\mathcal{Y}_t$ we use the nearest neighbour classifier in the KLDA feature space.

3 Image-set Representations and Kernels

The proposed algorithm is generic and works with any number of kernels derived from different image set representations. In this paper we consider seven image set representations and their respective set to set distance measures i.e. $\{d_r\}_{r=1}^{7}$. The distance measures are brought to the kernel domain by the Gaussian function of (1) as discussed in Section 1. The proposed SKL algorithm then learns a unified kernel as the sparse linear combination of these kernels. A brief overview of each type of image set representation and its respective kernel function is given below.

i. Affine hull kernel (K_a) [3]: An image set is represented by the affine hull model computed from the set samples. The affine hull based set-to-set distance is computed using the method in [3]. Let U_i and U_j denote the subspace bases and μ_i and μ_j denote the mean of the two image sets X_i and X_j respectively. Defining $U \equiv [U_i \quad -U_j]$, $\xi_i = \mu_i - U(U^T\mu_i)$ and $\xi_j = \mu_j - U(U^T\mu_j)$, the affine hull based image set kernel is given by

$$K_a(i,j) = e^{\left(\frac{-\|\xi_i - \xi_j\|_2}{\sigma_a^2}\right)} \tag{10}$$

ii. SANP kernel (K_s) [4]: An image set is represented by the affine hull model computed from the set samples and the samples themselves. Let U_i and U_j denote the subspace bases and μ_i and μ_j denote the mean of the two image sets X_i and X_j respectively. We compute the SANP kernel as

$$K_s(i,j) = e^{\left(\frac{(d_i+d_j)D^*}{\sigma_s^2}\right)} \tag{11}$$

where d_i and d_j are the dimensionalities of the subspaces U_i and U_j respectively (i.e. number of columns of U_i and U_j). D^* is the distance between the sparse approximated nearest points (SANP) of the two sets obtained by minimizing the following objective function [4]

$$\begin{aligned} D^* = \min_{\beta_i,\beta_j,v_i,v_j} \big(& \|\mu_i + U_i v_i - (\mu_j + U_j v_j)\|_2^2 + \omega_1(\|\mu_i + U_i v_i - X_i\beta_i\|_2^2 + \\ & \|\mu_j + U_j v_j - X_j\beta_j\|_2^2) + \omega_2\|\beta_i\|_1 + \omega_3\|\beta_j\|_1 \big) \end{aligned} \tag{12}$$

iii. Mean kernel (K_m): An image set is represented by the first order statistics i.e. the mean of the set sample. Let μ_i and μ_j denote the mean of two image sets X_i and X_j respectively. We compute the mean image set kernel as

$$K_m(i,j) = e^{(\frac{-\|\mu_i - \mu_j\|_2}{\sigma_m^2})} \tag{13}$$

iv. Subspace kernel (K_b): Let U_i and U_j denote the subspace bases of two image sets X_i and X_j respectively. We calculate the subspace based image set kernel as

$$K_b(i,j) = e^{(\frac{-\|U_i - U_j U_j^\top U_i\|_F^2}{\sigma_b^2})} \tag{14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

v. Log Euclidean kernel (K_l): An image set X is represented by its sample covariance matrix $C = XX^\top$. Note that X is first mean centred. Let C_i and C_j denote the sample covariance matrices of two image sets X_i and X_j. We compute the Log Euclidean image set kernel as

$$K_l(i,j) = e^{(\frac{-\|\log(C_i) - \log(C_j)\|_F^2}{\sigma_l^2})} \tag{15}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The logarithm of the SPD matrix C can be computed from its eigen-decomposition $C = USU^T$ by $\log(C) = U\log(S)U^T$ where $\log(S)$ is a diagonal matrix of the scaler logarithms of the eigenvalues of C.

vi. Cholesky kernel (K_c): A mean centered image set X is represented by its sample covariance matrix $C = XX^\top$. Let C_i and C_j denote the sample covariance matrices of two image sets X_i and X_j. We compute the as

$$K_c(i,j) = e^{(\frac{-\|L_i - L_j\|_F^2}{\sigma_c^2})} \tag{16}$$

where L_i is a lower triangular matrix of the Cholesky decomposition $C_i = L_i L_i^\top$.

vii. MMD kernel (K_d): An image set is represented as a collection of linear patches on a manifold. The distance between two image sets X_i and X_j is computed by using the manifold to manifold distance (MMD) method presented in [2]. Specifically each image set is first clustered into multiple linear local models. Each local model is represented by a linear subspace and the mean of the local model. The MMD is then defined as the weighted sum of the subspace distance and the mean distance between the nearest local models. The MMD distance is then kernalized using (1).

These diverse image set representations encodes different characteristics of the underlying set data based on different assumptions. For example, the mean

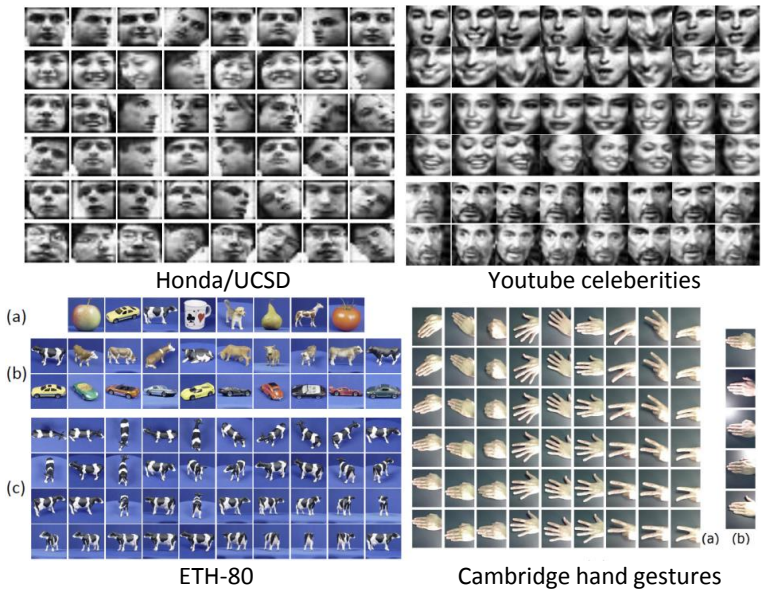


Fig. 3. Dataset Details. **HONDA/UCSD**: Each row represents images from a different image set. **Youtube celebrities:** Each row represents sample images from an image set. Two sets per subject are shown in this case. **ETH-80:** (a) Eight different object categories. (b) 10 different objects within each category. (c) Sample images from an image set of the cow category. **Cambridge Hand Gestures:** (a) Sample sequences from nine gesture classes. (b) Five different illumination conditions in the database.

and covariance based representations show the position of the image set in high dimensional space and assume the set data to be Gaussian. The mean and covariance based representations may fail easily if the set data is multimodal. Similarly, the subspace based image set representation may not work well if the actual set data lie on complex manifolds.

4 Experimental Results

We perform extensive experiments on four standard datasets capturing a wide range of operating conditions for three image set classification applications: face recognition, object categorization and hand gesture recognition.

4.1 Dataset Details

The Honda/UCSD dataset [22] contains 59 video sequences of 20 different subjects. The faces in every frame of the video sequences are automatically detected using Viola and Jones algorithm [23], resized to 20×20 grayscale images and histogram equalized (Fig. 3).

The YouTube Celebrities dataset [24] is the most challenging dataset and contains 1910 video sequences of 47 celebrities (actors, actresses and politicians)

which are collected from YouTube. Most videos are low resolution and recorded at high compression ratio, which leads to noisy and low-quality image frames. The clips contain different numbers of frames (from 8 to 400). Face image in each frame was first automatically detected by applying [23] and resized to a 20×20 (Fig. 3). We propose to compute the UoCTTI variant [25] of the histogram of oriented gradients (HOG) features using a cell size of 6 for each image. This results in a feature dimension $d = 279$ for each image. This simple and efficient pre-processing step has two advantages. Firstly, it reduces the feature dimension from 400 to 279 which significantly speeds up all the algorithms. Secondly, in our experiments we observed a significant increase in the accuracy of all the algorithms by using these features compared to using raw pixels values.

The ETH-80 dataset [26] contains images of 8 object categories where each category has 10 different objects of the same class. Each object has 41 images taken at different views which form an image set. We use 20×20 intensity images for the task of classifying an image set of an object into a known category. ETH-80 is a challenging database because it has fewer images per set, significant appearance variations across objects of the same class and larger viewing angle differences within each image set (Fig. 3).

The Cambridge Hand Gesture dataset [27] (Fig. 3) contains 900 image sequences of 9 gesture classes, which are defined by 3 primitive hand shapes and 3 primitive motions. Each class has 100 image sequences (5 different illuminations, 10 arbitrary motions, performed by 2 subjects). The recognition task involves the classification of different hand shapes as well as different hand motions at the same time. Following the experimental protocol of [28], the 100 videos of each gesture class are divided into five illumination sets (Set1, Set2, Set3, Set4 and Set5) where Set5 is chosen as the training images. The training set is further divided randomly into gallery and validation sets (10 sequences in the gallery and the other 10 sequences for validation). Since we do not use the validation set we discard it. Individual images are converted to grayscale and resized to 60×80. UoCTTI variant of HOG features with a cell size of 18 are calculated for each image resulting in a feature dimension $d = 372$. Details of all datasets used in our experiments are given in Table 1.

Table 1. Dataset details including maximum, minimum and average images per set.

Dataset	Classes	Sets/class	Min images/set	Max images/set	Avg images/set
Honda/UCSD	20	1-5	13	782	267
Youtube Celeb	47	9	8	347	150
ETH-80	8	10	41	41	41
Cambridge	9	100	37	119	71

4.2 Experimental Setup

For each dataset the important parameters to compute the image set representations are selected according to the recommendations of the original authors. For the affine hull based image set representations we preserve 98% total energy while computing the subspace bases for all the databases. For manifold representation K_d the parameters are configured as recommended in [2] for different data sets. The maximum canonical correlation is used in defining MMD. The number of connected nearest neighbours for computing geodesic distance in MMD is set to 12. For computing the Log Euclidean K_l and Cholesky kernels K_c the covariance matrix is first regularized by adding a small perturbation to its diagonal the (absolute of the smallest eigenvalue). The parameters of the MKLDR algorithm is configured according to the recommendations of the original authors [17]. For deriving the kernels, the optimal value of the Gaussian scale factor σ in (1) is selected automatically using the binary search algorithm of [17]. We set the parameter α to a small value of 10^{-6} when using the Elastic Net. The parameters λ_js can be automatically determined since the Elastic Net algorithm provides the optimal solution path of λ_js for given α [29].

For Honda dataset each subject has one image set in the gallery and the rest are used as probes. For the proposed SKL algorithm, at least two image sets per class are required in the gallery data. Therefore, when the gallery contained only one image set for a particular class, we randomly partitioned the set into two non-overlapping sub-sets. For Youtube dataset, the whole dataset is equally divided into five folds with minimal overlapping [4]. Each subject has 9 image sets. In each fold we use three image sets per class in the gallery and six image sets per class as probes. For ETH-80 dataset the gallery consists of 5 image sets per class and the remaining 5 image sets per class are used a probes. For Honda, ETH and Cambridge datasets, experiments are repeated 10-folds with different gallery probe combinations in each fold.

4.3 Results and Discussion

Table 2 summarizes our experimental results. Average recognition rates and standard deviations are reported for 10-fold experiments on Honda, ETH and Cambridge datasets and five fold experiments on the Youtube dataset.

On the Honda/UCSD dataset, the structure based image set representations perform better than the nearest sample based representations. Therefore the kernels derived from the subspace bases and the covariance representation (K_b, K_l, K_c) outperform the kernels derived from the affine hull representations (K_a, K_s, K_m) when used individually with KLDA. This is because there are enough samples available with adequate variations per set to accurately estimate the structure of the image set. The accuracy of the average kernel with KLDA is less than the maximum performing kernel K_b. This is because the lower performing kernels slightly degrade the performance of the overall mixture. Similarity, the MKLDR [17] method also uses all the kernels with different weights to compute the discriminative subspace and its performance is therefore affected by

Table 2. Comparison of average recognition rates and standard deviations (%).

Method	Kernel(s)	Honda	Youtube	ETH-80	Cambridge
KLDA	K_a	93.84±2.47	72.51±3.07	74.80±2.83	36.68 ± 1.95
	K_s	95.33±1.62	74.01±4.10	72.38±3.21	31.54 ± 1.89
	K_m	92.64±2.71	69.19±3.39	71.11±4.67	48.56 ± 1.50
	K_b	100±0.0	71.64±4.42	90.25±1.16	88.30 ± 0.08
	K_l	100±0.0	67.54±4.77	89.07±1.72	90.04± 0.05
	K_c	96.26±4.18	67.13±4.01	90.65±1.58	87.09 ± 1.65
	K_d	97.69±2.54	66.23±4.98	85.25±3.12	64.41 ± 1.25
	Avg K	97.69±1.45	72.12±3.62	87.01±5.94	87.19 ± 1.59
MKLDR [17]	All	98.71±0.18	74.08 ± 4.62	90.70±5.62	90.11 ± 1.80
Proposed SKL	Subset	**100±0**	**77.07 ± 2.01**	**94.75±0.31**	**92.43±0.04**

the poor performing kernels in this experiment. On the other hand the proposed SKL algorithm learns a sparse linear combination of only the most discriminative kernels to achieve the highest classification accuracy. Figure 2-(a) shows the weights calculated by the proposed algorithm for the Honda dataset. The proposed algorithm automatically learns high weights for the subspace based kernel K_b and the log-Euclidean kernel K_l while the other kernels gets zero weights.

On the Youtube celebrities dataset, the kernels computed from the nearest neighbour based image set representations (K_a, K_s, K_m) perform better than the kernels computed from the structure based image set representations (K_b, K_l, K_c). The reason being the useful variations in the image set data in this dataset is relatively low and the subspace or covariance structure cannot be estimated accurately. Our use of the HoG features also reduces the effects of illumination and pose variations which brings the individual samples belonging to the same classes closer. The accuracy of the MKLDR[17] is affected by the poor performing kernels. The proposed SKL algorithm achieves the highest accuracy by combing only the sample based kernels (K_a and K_s). Figure 2-(b) shows that the MKLDR algorithm assigns almost equal weights to all the representations which degrade its overall accuracy. By learning a combination of only the best subset of kernel, the proposed SKL algorithm outperform the other algorithms.

On the ETH-80 dataset, the kernels derived from the sample based representations (K_a,K_s,K_m) perform poor. For this dataset, the locations of the individual samples in the sets cannot provide discriminative information due to the large intra-set pose variations and significant intra-class object appearance differences. In this case, the structure of the image set can describe the common properties of a class more accurately. Therefore, the kernels computed from the structure based representations show more accuracy on this dataset. Figure 2-(c) shows that the propose weight learning algorithm has picked only the structure based kernels. The proposed SKL algorithm learns a combination of only the structure based kernels and hence outperforms all the others on this dataset.

On the Cambridge Hand Gestures dataset the sample based kernels K_a, K_s, K_m perform very poor when used individually with KLDA. For this dataset, the location of each individual sample cannot accommodate the hand gesture variations adequately. On the other hand the structure based kernels can capture the overall

Table 3. Comparison with existing image set classification algorithms.

Algorithm	Honda	Youtube	ETH-80	Cambridge
DCC [5]	94.87±2.24	66.75±3.47	90.25±3.06	88.31± 1.34
MMD [2]	94.87±1.16	65.12±4.36	69.72±4.01	58.06±2.71
MDA [6]	96.66±1.73	68.12±4.36	77.75±6.17	26.63±1.61
AHISD [3]	90.25±3.97	71.92±3.55	71.80±8.61	35.91±2.85
CHISD [3]	92.31±2.12	72.83±3.29	72.09±8.11	37.25±2.77
SANP [4]	94.34±1.62	74.01±3.48	72.15±8.61	30.14 ±1.35
CDL [7]	99.23±1.23	68.96±5.29	89.51±3.68	90.18±0.81
Proposed SKL	**100±0.0**	**77.07±2.01**	**94.75±0.31**	**92.43±0.04**

common properties of two gestures from the same class. Figure 2-(d) shows that the proposed SKL algorithm selectively learns higher weights for the structure based kernels for this dataset. Thus the the proposed SKL algorithm significantly outperforms the MKLDR algorithm. We also performed experiments to evaluate the performance of the proposed algorithm by setting $\lambda = 0$ and $\alpha = 0$. We noted an accuracy drop from {100, 77.07, 94.75, 92.43}% to {98.00, 73.12, 92.0, 88.44}% for Honda, Youtube, ETH-80 and Cambridge datasets respectively. This confirms that the proposed sparsity constraints indeed improve the classification accuracy.

4.4 Comparison with existing image set classification algorithms

The proposed SKL algorithm is also compared with seven state-of-the-art image set classification techniques including DCC [5], MMD [2], MDA [6], AHISD [3], CHISD [3], SANP [4] and CDL [7]. We have used the implementations from the original authors, except for MDA and CDL. For MDA, Hu's [4] implementation is used, while we have our own implementation of CDL. For a fair comparison, we follow the same protocol used previously by [3], [4], [6] and [7]. The existing image set classification algorithms consider only a single image set representation therefore the accuracies of these approaches vary for different properties of the image sets. Table 3 summarizes our results. Note that due to the use of HoG features the accuracy of the previous image set classification algorithms on the Youtube dataset has significantly increased. Also, the accuracy of AHISD and SANP algorithms is slightly lower compared to using the affine hull based kernel K_a and SANP kernel K_s used with KLDA. This is because AHISD and SANP algorithms do not perform any discriminant analysis after distance calculation, while our use of KLDA increases the inter-class similarity further. Because the proposed SKL algorithm combines only the best image set representations therefore it has shown the best accuracy on all the databases compared to the existing algorithms.

4.5 Computational time

Table 4 shows the average execution times of all algorithms for 10-fold experiments on Honda dataset using a Pentium 3.4GHz CPU with 8GB RAM and

MATLAB implementation. The computational complexity of the proposed algorithm involves the time to compute different kernel matrices plus the time of SKL and KLDA. In the training stage, the time taken to compute all the kernels is about 1100.02s while the SKL takes 0.2s. Note that in the testing phase we only compute the kernels which have non-zero weights which significantly reduces computation time compared to that of MKLDR.

Table 4. Execution times of matching one probe image set with 20 gallery image sets of the Honda/UCSD

Algorithm	Training time (sec)	Testing time (sec)
DCC [5]	0.91	0.30
MMD [2]	184.57	38.10
MDA [6]	10.55	33.00
AHISD [3]	N/A	9.10
CHISD [3]	N/A	110.10
SANP [4]	N/A	5.01
CDL [7]	1.10	0.15
MKLDR [17]	>100	56.12
Proposed SKL	>100	1.01

5 Conclusion

We proposed a sparse kernel learning (SKL) algorithm for image set classification. By optimizing a sparse linear discriminant objective function, the proposed algorithm automatically learns the most discriminative subset of kernels from a large pool. Experimental results on four standard datasets showed that the proposed SKL algorithm outperforms current state of the art image set classification algorithms. The proposed algorithm also outperformed the standard feature combination methods such as MKLDR with significant improvement in the test set matching time.

Acknowledgement. This research work was supported by ARC grants DP1096801 and DP110102399.

References

1. Lu, J., Wang, G., Moulin, P.: Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In: ICCV. (2013)
2. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: CVPR. (2008) 1–8
3. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: Computer Vsion and Pattern Recognition. (2010) 2567–2573
4. Hu, Y., Mian, A., Owens, R.: Face recognition using sparse approximated nearest points between image sets. IEEE Transactions on PAMI **34** (2012) 1992–2004
5. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE PAMI **29** (2007) 1005–1018

6. Wang, R., Chen, X.: Manifold discriminant analysis. In: Computer Vsion and Pattern Recognition. (2009) 429–436
7. Wang, Guo, H., Davis, L., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: CVPR. (2012)
8. Shakhnarovich, G., Fisher, J.W., Darrell, T.: Face recognition from long-term observations. In: European conference on Computer Vision. (2002) 851–868
9. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: CVPR. (2005)
10. Uzair, M., Mahmood, A., Mian, A., McDonald, C.: A compact discriminative representation for efficient image-set classification with application to biometric recognition. In: ICB. (2013)
11. Fukui, K., O., Y.: The kernel orthogonal mutual subspace method and its application to 3d object recognition. In: ACCV. (2007) 467–476
12. Gonen, M., Alpaydin, E.: Multiple kernel learning algorithms. JMLR (2011)
13. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: ICCV. (2007) 1–8
14. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV. (2009) 606–613
15. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV. (2009) 221–228
16. Vemulapalli, R., Pillai, J., Chellappa, R.: Kernel learning for extrinsic classification of manifold features. In: CVPR. (2013)
17. Lin, Y.Y., Liu, T.L., Fuh, C.S.: Multiple kernel learning for dimensionality reduction. IEEE Trans. on PAMI **33** (2011) 1147–1160
18. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE PAMI **29** (2007) 40–51
19. Li, X., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. Neural Networks, IEEE Transactions on **17** (2006) 157–165
20. Lai, Z., Xu, Y., Yang, J., Tang, J., Zhang, D.: Sparse tensor discriminant analysis. Image Processing, IEEE Transactions on **22** (2013) 3904–3915
21. Shawe-Taylor, J., Cristianini., N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
22. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: CVPR. (2003) I–313–I–320 vol.1
23. Viola, P., Jones, M.: Robust real-time face detection. IJCV **57** (2004) 137–154
24. Kim M., Kumar S., P.V., H., R.: Face tracking and recognition with visual constraints in real-world videos. In: CVPR. (2008)
25. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE PAMI **32** (2010) 1627–1645
26. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: CVPR. (2003) II – 409–15 vol.2
27. Kim, T.K., Wong, K.Y.K., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: Computer Vision and Pattern Recognition. (2007) 1–8
28. Kim, T.K., Cipolla, R.: Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE PAMI **31** (2009) 1415–1428
29. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of Royal Stat. Soc., Ser. B (Stat. Methodol.) **67** (2005) 301–320