Image Set Classification Using Holistic Multiple Order Statistics Features and Localized Multi-Kernel Metric Learning

Jiwen Lu[1], Gang Wang[1,2], and Pierre Moulin[3]
[1]Advanced Digital Sciences Center, Singapore
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3]Department of ECE, University of Illinois at Urbana-Champaign, IL USA
Email: jiwen.lu@adsc.com.sg; wanggang@ntu.edu.sg; moulin@ifp.uiuc.edu

Abstract

This paper presents a new approach for image set classification, where each training and testing example contains a set of image instances of an object captured from varying viewpoints or under varying illuminations. While a number of image set classification methods have been proposed in recent years, most of them model each image set as a single linear subspace or mixture of linear subspaces, which may lose some discriminative information for classification. To address this, we propose exploring multiple order statistics as features of image sets, and develop a localized multi-kernel metric learning (LMKML) algorithm to effectively combine different order statistics information for classification. Our method achieves the state-of-the-art performance on four widely used databases including the Honda/UCSD, CMU Mobo, and Youtube face datasets, and the ETH-80 object dataset.

1. Introduction

Image set classification has attracted increasing interest in computer vision and pattern recognition in recent years [1, 4, 6, 7, 10, 15, 16, 17, 20, 25, 28, 30, 34, 37, 40] due to its wide potential applications such as visual surveillance and multi-view image analysis. One representative application of image set classification is the video-based face recognition problem, where each gallery and probe face video can be considered as an image set and the characteristics of the image set are used for person identification. Different from the conventional image classification problem where each training and testing example is a single image, for image set classification, each training and testing example contains a set of image instances. Compared to a single image, an image set provides us more information to describe objects of interest. However, it is also challenging to exploit discriminative information of image sets as there are usually larger intra-class variations within a set.

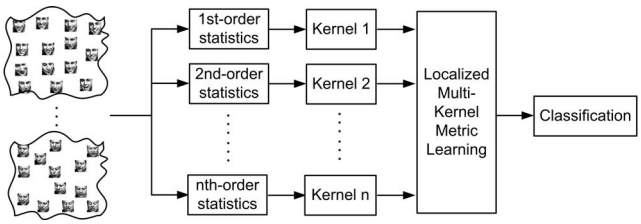


Figure 1. The basic idea of our approach. For each image set, we first compute its multiple order statistics as feature representation. For each order statistic, we compute a kernel matrix to measure the pairwise similarity of two image sets. Then, we learn a distance metric by using the localized multi-kernel metric learning (LMKML) method to combine the different order statistics. Lastly, the nearest neighbor classifier is used for classification.

There has been a number of work on image set classification over the past two decades [1, 4, 10, 16, 21, 23, 30, 34, 35, 38]. However, to our best knowledge, most existing image set classification methods usually make some prior assumptions such as singe Gaussian, Gaussian mixture models, subspace or manifold models to represent image sets. In many practical applications, these assumptions may not be held, especially when there are large and complex data variations within a set. Moreover, the models learned based on these assumptions may also lose some discriminative information for classification.

In this paper, we propose a new approach for image set classification. Given an image set, we compute its holistic multiple order statistics as features for set representation. Compared with most existing image set modeling methods [4, 16, 35], our multiple order statistics features can more robustly capture the distribution of image instances within a set in a holistic way because no parameter estimation is required. Moreover, they are also less sensitive to noise because noisy samples can be largely filtered out in the extracted statistic features. To make better use of the information extracted from different order statistics, we further

develop a localized multi-kernel metric learning (LMKML) algorithm to learn a distance metric, under which different order statistics are effectively combined and more discriminative information are exploited for classification. Experimental results on four widely used image set datasets are presented to show the efficacy of our proposed approach. The basic idea of our approach is illustrated in Figure 1.

1.1. Related work

Image Set Classification: There has been a growing interest in developing new algorithms for image set classification in recent years [1, 4, 5, 9, 10, 14, 15, 16, 21, 23, 30, 33, 35], and they can be mainly classified into two categories: parametric and nonparametric. Compared to these works, the contribution of our work is two-fold: 1) extracting multiple order statistics features to reliably represent an image set; 2) a localized multi-kernel metric learning algorithm. While [34] explored the second-order statistics of image sets for feature representation, our approach can extract more discriminative information because it considers and utilizes multiple different order statistics of image sets. We also achieve state-of-the-art performance on the image set classification problem with existing publicly available datasets.

Multiple Kernel Learning: There have been extensive work on multiple kernel learning in the literature [2, 8, 11, 13, 18, 22, 26, 29, 32, 36, 39, 41]. While many efforts have been made including classification [2, 11, 29, 36], clustering [39], transfer learning [8], and dimensionality reduction [26], little progress has been made in metric learning with multiple kernel learning. More recently, Wang *et al.* [32] proposed a metric learning method with multiple kernels by learning a universal weight vector over the whole space. Differently, our proposed LMKML algorithm learns an adaptive weight to each local region in the kernel space when learning the distance metric. Hence, our approach is complementary to existing multiple kernel learning methods.

2. Proposed Approach

Figure 1 shows the flow-chart of our proposed approach. For each image set, we first extract its multiple order statistics for set modeling. For each order statistic, we compute a kernel matrix to measure the pairwise similarity of two image sets. Then, we propose a LMKML method to learn a discriminative, localized distance metric to combine statistics information at different orders. Lastly, the nearest neighbor classifier is employed for classification. The details are introduced in the following subsections.

2.1. Set Modeling with Multiple Order Statistics

Let $X = [x_1, x_2, \cdots, x_n]$ be an image set containing n different images of an object, where $x_i \in R^d$ denotes the ith image sample. Image pixel values are used as raw features. Given each image set, we extract the following different order statistics information as features to represent the set. These multiple order statistics can reliably describe the distribution of image samples of a set, and hence can be used as image set features.

- **First-order statistics**: the mean vector m of the image set is computed, which shows the averaged position of the image set in the high dimensional space:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

- **Second-order statistics**: the covariance matrix C of the image set is computed, which represents the correlation of two individual features of each pair of samples in the image set:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - m)(x_j - m)^T \tag{2}$$

- **Third-order statistics**: the out product between the covariance matrix C and mean m of the image set is calculated, which forms a third-order tensor to measure the correlation of two individual elements of the covariance matrix and the mean vector:

$$\mathcal{T} = C \otimes m \tag{3}$$

where m is a d-dimensional vector, C is a $d \times d$ matrix, and $\mathcal{T}$ is a $d \times d \times d$ tensor, respectively. Here "$\otimes$" denotes the Kronecker product of two matrices. Note that more higher-order statistics can also be computed for each image set. However, we only consider these three in our approach because it is very expensive to compute higher-order statistics features.

Compared with previous image set representation methods, there are several advantages to model image sets with multiple order statistics information:

1. No assumption on the data distribution is required and the statistics features can be computed from an image set containing any number of samples.

2. Different order statistics information can characterize the image set from different perspectives. For example, the mean vector roughly reflects the position of the object in the high-dimensional space, and the covariance matrix represents the variance of each individual feature in the diagonal elements and measures the correlations of different features in the non-diagonal elements. Hence, these statistics features can provide complementary information to represent the image set.

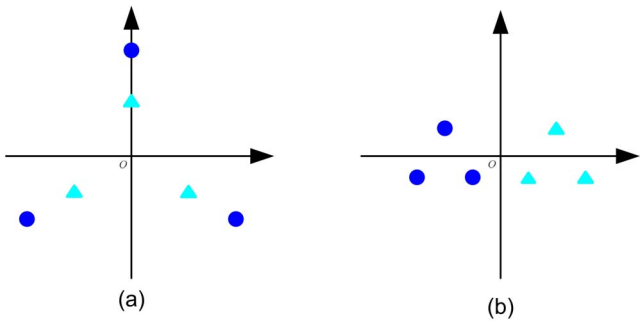


Figure 2. Illustration of the importance of different order statistics in image set classification. In this figure, the squares and triangles demote two different image sets. The first-order statistics are the same and the second-order statistics are different in (a), where the first-order statistics are different and the second-order statistics are the same in (b). Hence, we can see that different order statistics contribute different discriminative and complementary information for image set classification.

Figure 2 shows a toy example to illustrate that different order statistics contain different discriminative information for image set classification.

3. These statistic features are more robust to outlines since they are statistics of all the samples in the image set and the effect of the noisy samples can be largely alleviated, especially compared to the previous nearest sample pair based image classification methods [4, 16].

2.2. Localized Multi-Kernel Metric Learning

Having extracted multiple order statistics features, we perform classification by using the nearest neighbor classifier, which involves calculating the similarity between two image sets. We compare two statistics features in the kernel space, given the great success of kernel learning [26, 39]. This is equivalent to mapping the original statistic features to a new space, and calculating the dot product in the new space. We write the new feature for the pth statistic feature as ϕ^p, and the mapping function as $R^{d_p} \rightarrow \mathcal{F}$, where R^{d_p} is the original feature space and $\mathcal{F}$ is the mapped high-dimensional space. Though ϕ^p is usually implicit, we first consider it as an explicit feature vector for simplicity. Later, we will show any manipulation based on ϕ^p can be represented based on kernel values by using the kernel trick.

Similar to [2, 11], we assume different order statistic features can be mapped to a common high-dimensional feature space. And we aim to learn a distance metric to enforce objects from the same category to be close, and objects from different categories to be far away, in the learned metric space. Different from [2, 11] which assume the weights of different types of features (which are the different order statistic features here) are the same for all objects, we argue that weights should be data-adaptive. For example, if an image set's mean vector is discriminative, then we should assign a higher weight to it, compared to other orders. We formulate our learning problem based on this intuition as below, and call it Localized Multi-Kernel Metric Learning (LMKML).

Write $S = [S_1, S_2, \cdots, S_N]$ as the training set of N different image sets, where $S_i = [s_{i1}, s_{i2}, \cdots, s_{in_i}]$ denotes the ith image set, $1 \leq i \leq N$, and n_i is the number of samples in this image set. For each image set S_i, we compute its first-, second-, and third-order statistics m_i, C_i and $\mathcal{T}_i$, respectively. Let $X^p = [x_1^p, x_2^p, \cdots, x_N^p]$ be the pth statistic feature set of all training samples, and $x_i^p \in R^{d_p}$ denotes the pth statistic feature extracted from the ith image set S_i, where $1 \leq p \leq P$. In this work, $P = 3$ as we use three different order statistics features for image set representation. ϕ_i^p is the corresponding high-dimensional feature of x_i^p. M is the distance metric to be learned in the high-dimensional space $\mathcal{F}$. The distance between two image sets S_i and S_j under M is:

$$d(S_i, S_j) = \sum_{p=1}^{P} \eta_p(\phi_i^p)(\phi_i^p - \phi_j^p)^T M(\phi_i^p - \phi_j^p)\eta_p(\phi_j^p) \quad (4)$$

where $\eta_p(\phi_i^p)$ is a gating function to generate different positive weighting numbers for different ϕ_i^p, which will be detailed later. Because of $\eta_p(\phi_i^p)$, our learning method is "localized". It is obvious that previous global kernel weighting algorithms [2, 11] can be considered as a special case of our method, where $\eta_p(\phi_i^p)$ is the same for any ϕ_i^p.

To learn a distance metric M, we maximize inter-class variations and minimize intra-class variations, simultaneously. The objective function is formulated as:

$$\max_M J = \sum_{\substack{i,j=1 \\ (S_i,S_j)\in C^-}}^{N} \frac{d(S_i, S_j)}{N_{C^-}} - \sum_{\substack{i,j=1 \\ (S_i,S_j)\in C^+}}^{N} \frac{d(S_i, S_j)}{N_{C^+}} \quad (5)$$

where C^- and C^+ denote the inter-class and intra-class sample pairs in the training set, and N_{C^-} and N_{C^+} denote the number of pairs in these two sets, respectively.

M is symmetric and positive semidefinite. We can seek a nonsquare matrix W ($W = [w_1, w_2, \cdots, w_d]$) of size $d^{\mathcal{F}} \times d$, where $d^{\mathcal{F}}$ is the dimensionality of the high-dimensional feature space, and d is the number of basis in W, such that

$$M = WW^T \quad (6)$$

Combining Eqs. (5) and (6), we simplify J to the following form

$$J = tr\left(W^T(\frac{A_1}{N_{C^-}} - \frac{A_2}{N_{C^+}})W\right) \quad (7)$$

where

$$A_1 = \sum_{\substack{i,j=1 \\ (S_i,S_j)\in C^-}}^{N} \sum_{p=1}^{P} \eta_p(\phi_i^p)(\phi_i^p - \phi_i^p)(\phi_i^p - \phi_i^p)^T \eta_p(\phi_j^p) \quad (8)$$

$$A_2 = \sum_{\substack{i,j=1 \\ (S_i,S_j)\in C^+}}^{N} \sum_{p=1}^{P} \eta_p(\phi_i^p)(\phi_i^p - \phi_i^p)(\phi_i^p - \phi_i^p)^T \eta_p(\phi_j^p) \quad (9)$$

Generally, it is difficult or even impossible to compute A_1 and A_2 directly in the feature space $\mathcal{F}$ because the form of ϕ_i^p is usually unknown. Hence, we use the kernel trick method [3] by expressing the basis w_k as a linear combination of all the training samples in the mapped space, i.e.,

$$w_k = \sum_{i=1}^{N} u_i^k \phi_i^p \quad (10)$$

where u_i^k are the expansion coefficients. Hence,

$$\sum_{p=1}^{P} w_k^T \phi_i^p = \sum_{i=1}^{N}\sum_{p=1}^{P} u_i^k (\phi_i^p)^T \phi_i^p = \sum_{p=1}^{P} (u^k)^T K_{.i}^p \quad (11)$$

where u^k is a $N \times 1$ column vector and its ith entry is u_i^k, and $K_{.ip}$ is the ith column of the pth kernel matrix K^p. Here K^p is an $N \times N$ kernel matrix, calculated from the pth statistic feature using the RBF kernel between each pair of image set.

Then, Eqs. (7)-(9) can be rewritten as

$$J = tr\left(U^T(\frac{B_1}{N_{C^-}} - \frac{B_2}{N_{C^+}})U\right) \quad (12)$$

where

$$B_1 = \sum_{\substack{i,j=1 \\ (S_i,S_j)\in C^-}}^{N} \sum_{p=1}^{P} \eta_p(\phi_i^p)(K_{.i}^p - K_{.j}^p) (K_{.i}^p - K_{.j}^p)^T \eta_p(\phi_j^p) \quad (13)$$

$$B_2 = \sum_{\substack{i,j=1 \\ (S_i,S_j)\in C^+}}^{N} \sum_{p=1}^{P} \eta_p(\phi_i^p)(K_{.i}^p - K_{.j}^p) (K_{.i}^p - K_{.j}^p)^T \eta_p(\phi_j^p) \quad (14)$$

Now we discuss how to choose the gating function $\eta_p(\phi_i^p)$. There are a number of possible functions which could be used as the gating function. In this work, the gating function is selected as follow [11]:

$$\eta_p(\phi_i^p) = \frac{\exp(h_p^T \phi_i^p + b_p)}{\sum_{p=1}^{P} \exp(h_p^T \phi_i^p + b_p)} \quad (15)$$

where h_p and b_p are the parameters of this grating function. There are two reasons to select this gating function: 1) this function is monotonically increasing with the importance of ϕ_i^p; 2) this function can guarantees nonnegative weights and it is easy to obtain the derivatives with respect to h_p and b_p.

Since ϕ_i^p is implicit and its dimension is unknown, we express $h_p^T \phi_i^p$ as follow similar to Eq. (11):

$$h_p^T \phi_i^p = a_p^T (\phi_i^p)^T \phi_i^p = a_p^T K_{.i}^p \quad (16)$$

Then, the gating function can be written as:

$$\eta_p(\phi_i^p) = \frac{\exp(a_p^T K_{.i}^p + b_p)}{\sum_{p=1}^{P} \exp(a_p^T K_{.i}^p + b_p)} \quad (17)$$

where $a_p \in R^{N\times 1}$ and $b_p \in R^1$ are the parameters.

To our best knowledge, there is no closed-form solution to the optimization problem in Eq. (12) because we aim to learn U but have to infer a_p and b_p simultaneously. Hence, We solve this problem in an iterative manner inspired by some recent EM-like multiple kernel learning algorithms [26, 32]. The basic idea is to fix a_p and b_p, update U, and fix U, update a_p and b_p, iteratively.

We first initialize a_p and b_p with small random numbers, $1 \le p \le P$, and obtain U by solving the minimization problems in Eq. (12). We add a constraint $U^T U = I$ to restrict the scale of U such that the optimization problem in Eq. (12) with respect to W is well-posed. Then, U can be obtained by solving the following eigenvalue problem

$$(B_1 - B_2)u = \lambda u. \quad (18)$$

Let $u_1, u_2, \cdots, u_d$ be the eigenvectors of Eq. (18) corresponding to the d largest eigenvalues ordered according to $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$. An $N \times d$ transformation matrix $U = [u_1, u_2, \cdots, u_d]$ can be obtained.

Having obtained U, we use the gradient descent method to update a_p and b_p as follows:

$$a_p^{t+1} = a_p^t - \alpha \frac{\partial J}{\partial a_p} \quad (19)$$

$$b_p^{t+1} = b_p^t - \alpha \frac{\partial J}{\partial b_p} \quad (20)$$

where α is the learning rate and set as 0.000001 in our experiments.

Having updated a_p and b_p, we first re-compute the weight $\eta_p(\phi_i^p)$ in Eq. (17), and then B_1 and B_2 in Eqs. (13) and (14), respectively. Then, we update U by re-solving the eigenvalue equation in Eq. (18). We repeat this procedure until the algorithm is convergent. The proposed LMKML algorithm is summarized in **Algorithm 1**.

Algorithm 1: LMKML

Input: Training set: P $N \times N$ kernels computed from N image sets, iteration number T, feature dimension d, convergence error ϵ.

Output: Transformation matrix U and parameters a_p and b_p.

Step 1 (Initialization):
Initialize a_p^0 and b_p^0 with small random numbers.

Step 2 (Local optimization):
For $t = 1, 2, \cdots, T$, repeat
2.1. Compute B_1 and B_2 using Eqs. (13) and (14).
2.2. Solve the eigenvalue problem in Eq. (18).
2.3. Obtain $U^t = [u_1, u_2, \cdots, u_d]$.
2.4. Update a_p and b_p using Eqs. (19) and (20).
2.5. If $t > 2$, $|a_p^{t+1} - a_p^t| < \epsilon$ and $|b_p^{t+1} - b_p^t| < \epsilon$ or $|U^{t+1} - U^t| < \epsilon$, go to Step 3.

Step 3 (Output transformation matrix and parameters):
Output the matrix U and parameters a_p and b_p.

2.3. Classification

Given a testing image set X_T, we first compute its P different order statistics for feature representation, denoted as x_T^p, $1 \leq p \leq P$. Then, we calculate the distance between X_T and each training image set X_i, $1 \leq i \leq N$, as follows:

$$\begin{aligned} d(S_T, S_i) &= \sum_{p=1}^{P} \eta_p(\phi_T^p)(\phi_T^p - \phi_i^p)^T \\ &\quad WW^T(\phi_T^p - \phi_i^p)\eta_p(\phi_i^p) \\ &= \sum_{p=1}^{P} \eta_p(K_{.T}^p)(K_{.T}^p - K_{.i}^p)^T \sum_{l=1}^{d} u_l u_l^T \\ &\quad (K_{.T}^p - K_{.i}^p)^T \eta_p(K_{.i}^p) \end{aligned} \tag{21}$$

where $K_{.T}^p$ is the column vector denoting the similarity between the test image set and all the training image sets using the pth statistic feature.

Lastly, we classify the test image set x_T into class c that can minimize the distance between the test image set and all the training image sets

$$c = \arg\min_i d(S_T, S_i) \tag{22}$$

3. Experimental Results

We evaluate our proposed approach on two image set classification applications: face recognition based on image sets and set-based object categorization. The following describes the details of the experiments and results.

3.1. Datasets

Three publicly available face datasets, namely Honda/UCSD [23], CMU MoBo [12], and YouTube Celebrities [19], are used for face recognition based on image sets experiments. Each video sequence from these three datasets consists of an image set. The Honda/UCSD dataset contains 59 video sequences of 20 different subjects, and each video contains approximately 400 frames covering large variations in both out-of-plane head movement and facial expression. There are 96 video sequences of 24 subjects in the CMU MoBo dataset. For each subject, 4 video sequences are collected where each one corresponds to a different walking pattern. For each sequence, there are around 300 frames. The YouTube Celebrities dataset 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 the face detector method proposed in [31] and then resized to a 20×20 intensity image. Histogram equalization was the only pre-processing method used to alleviate illumination effect.

For object categorization, we used the ETH-80 dataset [24]. This database contains visual object images of eight different categories including apples, cars, cows, cups, dogs, horses, pears and tomatoes. For each category, there are 10 object instances and each object instance has 41 images of different views which construct an image set. The task is to recognize each image set of an object instance into a known category. Similar to previous studies [21, 34], object images were segmented from the simple background and scaled to 20×20 for classification.

3.2. Experimental Settings

To make a fair comparison with previous methods, we follow the same protocol used in [4, 16, 33, 34, 35]. On all of four datasets, we conduct experiments 10 times by randomly randomly selecting gallery/probe combinations, and compute and compare the average recognition rates of different methods. Specifically, for both the Honda and MoBo datasets, we randomly select one image set for each person as the gallery set and the remaining image sets are used for probes. For the YouTube dataset, the whole dataset is equally divided into five folds (with minimal overlapping). Each containing 9 video sequences per subject. In each fold, 3 image sets per subject are randomly selected as the gallery set and the remaining 6 are selected for probes. For the ETH-80 dataset, each category has 5 objects for gallery and the other 5 objects for probes.

3.3. Results and Analysis

Comparison with Existing Image Set Classification Methods: We compare the proposed approach with several existing image set classification methods which were proposed recently in the literature, including Discriminan-

Table 1. Average recognition rates (%) of different image set classification methods on the four datasets.

Method	Honda	MoBo	Youtube	ETH-80
DCC [21]	94.9	88.1	64.8	90.5
MMD [35]	94.9	91.7	66.7	86.1
MDA [33]	97.4	94.4	68.1	89.2
AHISD [4]	89.5	94.1	66.5	77.6
CHISD [4]	92.5	95.8	67.4	74.5
SANP [16]	93.6	96.1	68.3	80.5
CDL [34]	97.4	87.5	69.7	92.5
Our approach	**98.5**	**96.3**	**78.2**	**94.5**

t Canonical Correlation analysis (DCC) [21], Manifold-to-Manifold Distance (MMD) [35], Manifold Discriminant Analysis (MDA) [33], Affine Hull based Image Set Distance (AHISD) [4], Convex Hull based Image Set Distance (CHISD) [4], Sparse Approximated Nearest Point (SANP) [16], and Covariance Discriminative Learning (CDL) [34].

The standard implementations of all methods from the original authors were used except CDL. We carefully implemented the CDL algorithm since its code has been not publicly available. The key parameters of different methods were carefully optimized as follows: For DCC, PCA was performed to learn the linear subspace and the subspace dimensions were set as 10 to preserves 90% data energy and the corresponding 10 maximum canonical correlations were used to define the similarity of two image sets. For MMD and MDA, the parameters were configured according to [35] and [33], respectively. Specifically, the maximum canonical correlation was used in defining MMD, and the number of connected nearest neighbors for computing geodesic distance in both MMD and MDA was fixed as 12. There is no parameter setting for AHISD. For CHISD, we set the error penalty parameter to be the same as that used in [4]. For SANP, we applied the same weight parameters as in [16] for the convex optimization. For CDL, the kernel variant of LDA (KLDA) was used for discriminative learning and the regularization parameter was set the same as that used in [34]. Note that for the DCC, CDL and our proposed approach, there is a single gallery image set from each class in the Honda and MoBo datasets, we randomly divided each gallery set in these two datasets into two subsets to model the within-class variation.

Table 1 tabulates the recognition results of different image set classification methods on these four datasets. We can see that our approach performs better than the other seven compared image set classification methods, especially on the most difficult Youtube face dataset, where the improvement is significant. This is because most other compared methods require certain assumptions for image set representation and these assumptions may not hold in this challenging dataset. However, no assumption is required in our approach and hence better performance can be obtained.

Table 2. Average recognition rates (%) of different order statistics features on these four datasets.

Method	Honda	MoBo	Youtube	ETH-80
First-order	95.4	92.3	72.7	88.0
Second-order	96.5	88.9	67.5	89.5
Third-order	97.2	94.2	76.2	90.5
All-order	**98.5**	**96.3**	**78.2**	**94.5**

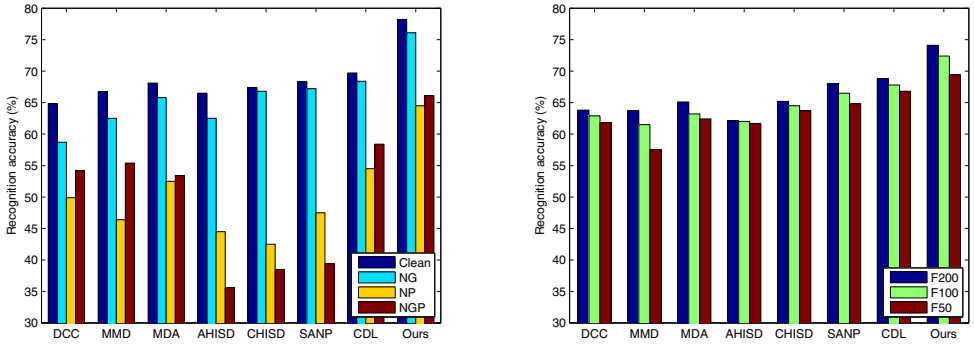
Table 3. Average recognition rates (%) of different multi-kernel metric learning methods on different datasets.

Method	Honda	MoBo	Youtube	ETH-80
GMKML	98.3	95.4	76.7	92.4
LMKML	**98.5**	**96.3**	**78.2**	**94.5**

Comparison of Different Order Statistics Features: We compare the discriminative power of different order statistics features for image set classification. For each single order statistics feature, we performed image set classification with the NN classifier. Table 2 tabulates the classification rates of different order statistics features. We can observe from this table that the third-order statistics feature achieves the best classification performance than other two order statistics features because the third-order statistics feature encodes both the first- and second-order statistics information. Meanwhile, the first- and second- order statistics are still complementary to the third-order statistics.

Localized vs. Global Multi-Kernel Metric Learning: The multi-kernel distance metric can also be learned in a global manner. To show the effect of LMKML, we assume $\eta_p(\phi_i^p)$ is the same for different x_i^p and learn a distance by using the global multi-kernel metric learning (GMKML) algorithm, where the weights of different kernels are learned and updated following the method in [27]. Table 3 tabulates the classification rates of these two methods. We can observe that our localized method can achieve better performance than the global one, which shows that learning data-specific kernel is better because it can exploit the characteristics of each data point.

Robustness Analysis: We test the robustness of our approach in case there are some noisy data in image sets or the image sets are of varying size. For the noisy data problem, we followed [4] and [34] and conducted three experiments where the gallery and/or probe sets were artificially corrupted by including one image from the other category. Similar to [34], the original clean data and three noisy datasets are called as "clean", "NG" (only gallery sets have noise data), "NP" (only probe sets have noise data), and "NGP" (both gallery and probe sets have noise data), respectively. For the varying size problem, we randomly selected a subset from each image set (both gallery and probe) and used the sub-



(a) (b)

Figure 3. Average recognition rates (%) of different image set classification methods with (a) noisy data and (b) varying data size on the Youtube dataset, respectively.

Table 4. Average recognition rates (%) of different multi-kernel metric learning methods on different datasets. For the polynomial kernel, the parameter is selected as 2.

Kernel type	Honda	MoBo	Youtube	ETH-80
Linear	98.0	96.3	77.6	93.8
Polynomial	98.0	96.0	77.8	94.0
RBF	**98.5**	**96.3**	**78.2**	**94.5**

sets for classification. We tested three cases by extracting 200, 100 and 50 frames, referred to F200, F100, and F50, respectively. In case a set contains fewer image frames, all images were used for classification. Figure 3 shows the average recognition rates of different image set classification methods on the Youtube dataset with different challenging test. From this figure, we can see that our proposed approach shows high robustness against these two challenges, with some slight performance drop. That is because we use different order statistics features as the set representation, which are robust to outlines and and the number of samples in the set. Hence, the effects of the noisy samples and varying data size can be alleviated.

Parameter Analysis: Since our approach is an iterative method, we evaluate its performance with different number of iterations. Figure 4 shows the recognition accuracy of our approach versus different number of iterations on the Youtube dataset. We can see that our proposed approach can achieve stable performance in several iterations.

Table 4 shows the recognition accuracy of our approach versus different types of kernels on different datasets. We can see that the performance of our approach is non sensitive to the kernel type selection.

Computational Time: Lastly, we compare the computational complexity of different image set classification methods using the YouTube dataset. For testing, we report the classification time for matching one probe image set with all the gallery image sets. Our hardware configuration comprises a 2.8-GHz CPU and a 10GB RAM. Table 5 shows the time spent on the training and the testing by these methods with the Matlab software. It is to be noted that training time is only required for discriminative learning methods such as DCC, MDA and our approach. We can see that the computational complexity of our approach is generally larger than the other compared methods. That is because our approach compute multiple order statistics features for image set representation, which requires more algebraic operation than other methods and hence leads to a higher computational complexity.

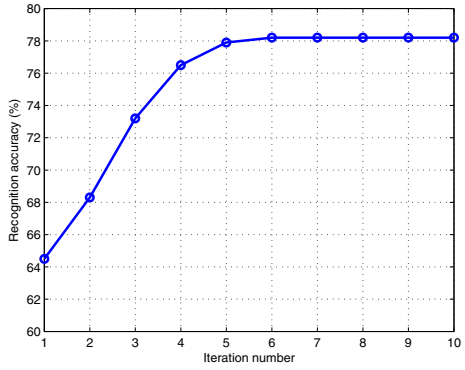


Figure 4. Average recognition rate (%) of our approach versus different number of iterations on the Youtube dataset.

4. Conclusion and Future Work

In this paper, we propose a new image set classification approach by using holistic multiple order statistics features and localized multi-kernel metric learning. The proposed approach has been evaluated on two visual classification applications: face recognition and object categorization. Experimental results on four widely used databases have shown the superiority of our approach over the state-

Table 5. Computation time (seconds) of different methods on the Youtube dataset for training and testing (classification of one image set).

Method	DCC	MMD	MDA	AHISD	CHISD	SANP	CDL	Our approach
Training	122.8	N.A.	225.0	N.A.	N.A.	N.A.	80.2	4755.8
Testing	3.8	5.4	64.8	9.2	14.5	55.6	15.6	220.3

of-the-art image set classification methods in terms of accuracy and robustness.

For future work, we are interested in designing more efficient kernel calculation method to improve the speed of our approach and exploring higher order statistics features and combine them with the features used in this work to further improve the recognition performance.

Acknowledgement

This work is partially supported by the research grant for the Human Sixth Sense Program at the Advanced Digital Sciences Center (ADSC) from the Agency for Science, Technology and Research (A*STAR) of Singapore.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, pages 581–588, 2005. 1, 2

[2] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, pages 1–8, 2004. 2, 3

[3] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000. 4

[4] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010. 1, 2, 3, 5, 6

[5] S. Chen, S. Mau, M. Harandi, C. Sanderson, A. Bigdeli, and B. Lovell. Face recognition from still images to video sequences: A local-feature-based framework. *Journal on Image and Video Processing*, 2011:11, 2011. 2

[6] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *ECCV*, pages 766–779, 2012. 1

[7] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *CVPR*, pages 2626–2633, 2012. 1

[8] L. Duan, I. Tsang, and D. Xu. Domain transfer multiple kernel learning. *PAMI*, 34(3):465–479, 2012. 2

[9] K. Fan, W. Liu, S. An, and X. Chen. Margin preserving projection for image set based face recognition. In *ICNIP*, pages 681–689, 2011. 2

[10] W. Fan and D. Yeung. Locally linear models on face appearance manifolds with application to dual-subspace based classification. In *CVPR*, pages 1384–1390, 2006. 1, 2

[11] M. Gönen and E. Alpaydin. Localized multiple kernel learning. In *ICML*, pages 352–359, 2008. 2, 3, 4

[12] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June 2001. 5

[13] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505, 2009. 2

[14] A. Hadid and M. Pietikainen. From still image to video-based face recognition: an experimental analysis. In *FG*, pages 813–818, 2004. 2

[15] M. Harandi, C. Sanderson, S. Shirazi, and B. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, pages 2705–2712, 2011. 1, 2

[16] Y. Hu, A. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011. 1, 2, 3, 5, 6

[17] Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *PAMI*, 34(10):1992–2004, 2012. 1

[18] P. Jain, B. Kulis, and K. Grauman. Fast image search for learned metrics. In *CVPR*, pages 1–8, 2008. 2

[19] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8, 2008. 5

[20] T. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *ECCV*, pages 251–262, 2006. 1

[21] T. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *PAMI*, 29(6):1005–1018, 2007. 1, 2, 5, 6

[22] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, pages 2130–2137, 2009. 2

[23] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, pages 313–320, 2003. 1, 2, 5

[24] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*. 5

[25] F. Li, Q. Dai, W. Xu, and G. Er. Weighted subspace distance and its applications to object recognition and retrieval with image sets. *IEEE Signal Processing Letters*, 16(3):227–230, 2009. 1

[26] Y. Lin, T. Liu, and C. Fuh. Multiple kernel learning for dimensionality reduction. *PAMI*, 33(6):1147–1160, 2011. 2, 3, 4

[27] J. Lu, J. Hu, X. Zhou, Y. Shang, Y. Tan, and G. Wang. Neighborhood repulsed metric learning for kinship verification. In *CVPR*. 6

[28] Y. Lui, J. Beveridge, B. Draper, and M. Kirby. Image-set matching using a geodesic distance and cohort normalization. In *FG*, pages 1–6, 2008. 1

[29] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, et al. Simplemkl. *JMLR*, 9:2491–2521, 2008. 2

[30] G. Shakhnarovich, J. Fisher, and T. Darrell. Face recognition from long-term observations. In *ECCV*, pages 361–375, 2006. 1, 2

[31] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 5

[32] J. Wang, H. Do, A. Woznica, and A. Kalousis. Metric learning with multiple kernels. In *NIPS*, pages 1–8, 2011. 2, 4

[33] R. Wang and X. Chen. Manifold Discriminant Analysis. In *CVPR*, pages 1–8, 2009. 2, 5, 6

[34] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012. 1, 2, 5, 6

[35] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, pages 1–8, 2008. 1, 2, 5, 6

[36] S. Wang, S. Jiang, Q. Huang, and Q. Tian. Multiple kernel learning with high order kernels. In *ICPR*, pages 2138–2141, 2010. 2

[37] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao. Set based discriminative ranking for recognition. In *ECCV*, pages 497–510, 2012. 1

[38] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *FG*, pages 318–323, 1998. 1

[39] B. Zhao, J. Kwok, and C. Zhang. Multiple kernel clustering. In *SIAM ICDM*, pages 638–649, 2009. 2, 3

[40] Y. Zhao, S. Xu, and Y. Jia. Discriminant clustering embedding for face recognition with image sets. In *ACCV*, pages 641–650, 2007. 1

[41] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011. 2