# Localized Multifeature Metric Learning for Image-Set-Based Face Recognition

Jiwen Lu, *Member, IEEE*, Gang Wang, *Member, IEEE*, and Pierre Moulin, *Fellow, IEEE*

***Abstract*— This paper presents a new approach to image-set-based face recognition, where each training and testing example is a set of face images captured from varying poses, illuminations, expressions, and resolutions. While a number of image set based face recognition methods have been proposed in recent years, most of them model each face image set as a single linear subspace or as the union of linear subspaces, which may lose some discriminative information for face image set representation. To address this shortcoming, we propose exploiting statistics information as feature representations for face image sets and develop a localized multikernel metric learning algorithm to effectively combine different statistics for recognition. Moreover, we propose a localized multikernel multimetric learning method to jointly learn multiple feature-specific distance metrics in the kernel spaces, one for each statistic feature, to better exploit complementary information for recognition. Our methods achieve state-of-the-art performance on four widely used video face datasets including the Honda, MoBo, YouTube Celebrities, and YouTube Face datasets.**

***Index Terms*— Face recognition, image set classification, metric learning, multikernel learning, multimetric learning.**

## I. Introduction

THERE has been a high level of interest in image set classification methods in recent years [1], [3], [4], [6], [10], [16], [19], [20], [24], [26], [29], [35], [37], [41], [49], [56], [59], [61], [63], which have a wide variety of applications in visual surveillance and multiview image analysis. One representative application is video-based face recognition, where each gallery and probe face video is considered as an image set and the characteristics of the set are used for person identification. Unlike conventional image classification, each training and testing example contains a set of image instances. Compared with a single image, an image set provides more information to describe objects of interest. However, it is also challenging to exploit discriminative information of image sets as intra-class variations are usually larger within an image set.

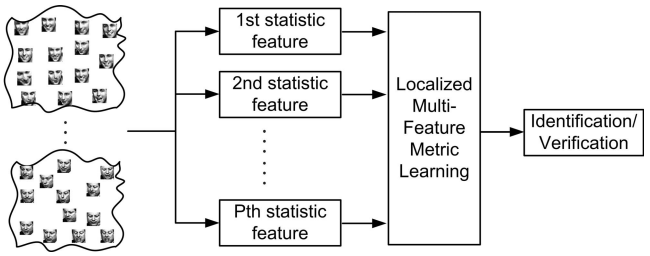



Fig. 1. Basic idea of our proposed approach. For each face image set, we first compute multiple statistics as feature representation. For each statistic, we construct a kernel matrix to measure the pairwise similarity of two face image sets. Then, we combine these $P$ statistics using a localized multifeature metric learning approach, where two kernel-based metric learning algorithms called LMKML and LMKMML were proposed, respectively. Finally, a nearest neighbor classifier is used for image-set-based face identification or face verification.

There has been substantial work on image-set-based face recognition over the past two decades [1], [3], [10], [19], [25], [28], [37], [41], [49], [50], [57]. However, most of these methods are based on prior assumptions, such as Gaussian models, Gaussian mixture models, and subspace or manifold models, to represent image sets. In many practical applications, these assumptions do not hold, especially in the presence of large and complex data variations within the face image set. Moreover, the models learned based on these assumptions may also lose some discriminative information for recognition.

In this paper, we propose a new approach to image-set-based face recognition. Fig. 1 shows the basic idea. Given a face image set, we compute multiple statistics as feature representations for the image set. Compared with other methods in [3], [19], and [50], our features robustly capture the distribution of image instances within a set because no parameter estimation is required. Moreover, our features are less sensitive to noise. To better use the information extracted from those statistics, we develop a localized multikernel metric learning (LMKML) algorithm to learn a distance metric, under which different statistic features are combined and more discriminative information is exploited for recognition. We further propose a localized multikernel multimetric learning (LMKMML) method to learn multiple feature-specific distance metrics in the kernel spaces, one distance metric

Manuscript received August 31, 2014; revised February 23, 2015; accepted March 9, 2015. Date of publication March 13, 2015; date of current version March 3, 2016. This work was supported in part by Agency for Science, Technology and Research (A*STAR), Singapore, through the Human Cyber Security Systems Program, Advanced Digital Sciences Center; in part by the Ministry of Education, Singapore, Tier 2 ARC28/14; and in part by A*STAR, Singapore, through the Science and Engineering Research Council under Grant PSF1321202099. This paper was recommended by Associate Editor Y. Keller.

J. Lu is with the Department of Automation, Tsinghua University, Beijing, 100084, China (e-mail: elujiwen@gmail.com).

G. Wang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, and also with the Advanced Digital Sciences Center, Singapore 138632 (e-mail: wanggang@ntu.edu.sg).

P. Moulin is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana—Champaign, Champaign, IL 61820 USA, and also with the Advanced Digital Sciences Center, Singapore 138632 (e-mail: moulin@ifp.uiuc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2015.2412831

for each feature, to better exploit complementary information for recognition. Experimental results on four widely used video face datasets show the effectiveness of our proposed approach.

This paper is an extended version of [34]. New contributions include the newly proposed LMKMML method, application to face verification, analysis of the proposed approach, and extensive comparisons with state-of-the-art methods in terms of both accuracy and robustness.

## II. Background

In this section, we briefly review three related topics: 1) image set based face recognition; 2) multiple kernel learning; and 3) metric learning.

### A. Image Set Based Face Recognition

Recent algorithms for image set classification can be mainly classified into parametric [1], [15], [28], [41] and nonparametric [3], [9], [16], [19], [20], [25], [48], [50] methods. Parametric methods model image sets using a parametric family of probabilistic distribution, and the Kullback–Leibler divergence between two distributions is used to measure the similarity of two image sets. Representative distributions include a single Gaussian model and a mixture of Gaussian models. However, parametric methods usually fail when the underlying distributional assumptions do not hold. To overcome these limitations, nonparametric methods have been recently proposed [3], [16], [19], [20]. They exploit geometrical information to measure the similarity of two image sets. While encouraging performance has been obtained [3], [16], [19], [20], most of these methods model each image set as a single linear subspace or as the union of linear subspaces, which may result in the loss of some discriminative information for classification. While Wang *et al.* [49] explored the use of second-order statistics of image set representation, other statistics were ignored.

### B. Multiple Kernel Learning

There has been extensive research on multiple kernel learning [2], [8], [12], [14], [21], [27], [32], [40], [47], [51], [60], [62]. The key objective is to seek an optimal combination of kernels to learn models for applications such as classification [2], [12], [40], [51], clustering [60], transfer learning [8], and dimensionality reduction [32]. However, little progress has been made in metric learning with multiple kernels. Recently, Wang *et al.* [47] proposed a multikernel metric learning method by learning a universal weight vector over the whole space. However, the characteristics of local regions in the kernel space were ignored. Moreover, most existing multiple kernel learning algorithms aim to learn a single combined kernel, which is not powerful enough to exploit the specific information of each feature. Hence, it is desirable to learn multiple distance metrics in the kernel spaces, one metric for each single feature, to jointly extract complementary information and exploit the interactions of different feature representations.

### C. Metric Learning

In recent years, a number of metric learning algorithms have been proposed in machine learning and computer vision [7], [11], [53]. Representative methods includ neighborhood component analysis [11], large-margin nearest neighbor (LMNN) [53], and information theoretic metric learning [7]. While these methods have achieved encouraging performance in applications such as face recognition [14], human activity recognition [44], person reidentification [43], [62], image retrieval [58], and visual tracking [45], [52], most of them only learn a distance metric with a single feature representation and cannot handle multiple features directly.

Lu *et al.* [33] proposed a multiview neighborhood repulsed metric learning method, which learns a latent distance metric space to combine multiple features for kinship verification. However, the weights of different features are assumed to be the same for all samples, which cannot effectively exploit the data-adaptive characteristics of the samples in classification because different features usually show different discriminative powers in different classes. Hence, it is desirable to exploit such information to learn one or multiple more discriminative distance metrics.

## III. Proposed Approach

Fig. 1 shows our proposed approach. The details are presented in Sections III-A–III-D.

### A. Face Image Set Representation

Let $X = [x_1, x_2, \ldots, x_n]$ be a face image set containing $n$ images of a subject, where $x_i \in \mathbb{R}^d$ denotes the $i$th face image sample, $1 \leq i \leq n$, $d$ is the feature dimension of each face image, which is usually set in the range [300, 1000]. Image pixel values are used as raw features. We compute the following statistics as features to represent the set.

1) *First-Order Statistic:* The sample mean vector $m$ of the image set is

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^d. \tag{1}$$

2) *Second-Order Statistic:* The sample $d \times d$ covariance matrix $C$ of the image set is

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

3) *Combined Statistic:* The Kronecker product of the covariance matrix $C$ and the mean $m$ of the image set is considered as a combined statistic feature

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

which is a $d \times d \times d$ tensor.

The mean vector $m$ roughly reflects the position of the subject in the high-dimensional face space, and the covariance matrix $C$ represents the self-variation of single feature and the correlations of different features. The combined statistic

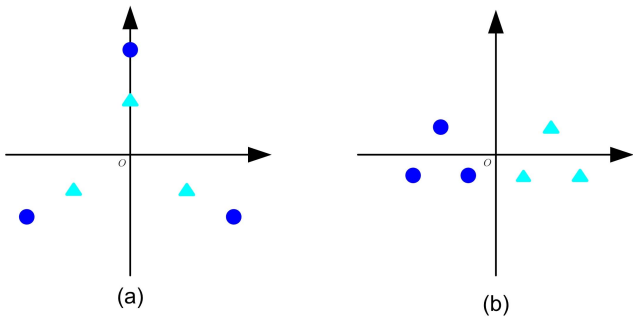



Fig. 2. Illustration of the importance of different order statistics in image set classification. In this figure, the circles and triangles demote two different image sets. (a) First-order statistics are the same and the second-order statistics are different. (b) First-order statistics are different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

is a function of $C$ and $m$ and is obtained through a kernel function.[1]

Compared with previous image set representation methods, our proposed face image set representation method presents the following key advantages.

1) No distributional or geometric assumption on the data is required.
2) The statistics can be computed from a face image set containing any number of samples.
3) Different statistics characterize the image set from different perspectives. Fig. 2 shows a toy example to illustrate that different order statistics contain different types of discriminative information for image set classification.

### *B. Localized Multikernel Metric Learning*

Having extracted a number $P$ of statistic features, we perform recognition using the nearest neighbor classifier, which involves calculating the similarity between two image sets. We use a multikernel approach and compare statistic features in the kernel space [32], [60]. This is equivalent to mapping the original statistic features to a new space and calculating the dot product in the new space. We denote the new vector for the $p$th statistic feature by $\phi^p$, and the mapping function by $\mathbb{R}^{d_p} \rightarrow \mathcal{F}$, where $\mathbb{R}^{d_p}$ is the original feature space and $\mathcal{F}$ is the new space. Though $\phi^p$ is usually implicit, we first consider it as an explicit feature vector for simplicity. Later, we will show any manipulation based on $\phi^p$ can be represented based on kernel values using the kernel trick.

Similar to [2] and [12], we assume that different statistic features can be mapped to a common high-dimensional feature space. We aim to learn a distance metric to force face image sets from the same category to be close and those from different categories to be far apart in the learned metric space. Unlike existing multikernel learning methods [2], [12] that assume the weights of different types of features (which are the different statistic features here) are the same for all classes, 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 with other features.

We formulate our LMKML problem based on this concept. Write $\mathcal{S} = [S_1, S_2, \ldots, S_N]$ as the training set of $N$ image sets, where $S_i = [s_{i1}, s_{i2}, \ldots, s_{in_i}]$ denotes the $i$th 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-order, second-order, and combined statistics $m_i$, $C_i$, and $\mathcal{T}_i$, respectively. Let $X^p = [x_1^p, x_2^p, \ldots, x_N^p]$ be the $p$th statistic feature set of all training samples and $x_i^p \in \mathbb{R}^{d_p}$ the $p$th statistic feature extracted from the $i$th image set $S_i$, where $1 \leq p \leq P$. In this paper, $P = 3$ as we use three different order statistics features for image set representation. $\phi_i^p$ is the corresponding high-dimensional feature of $x_i^p$, which for notational convenience we assume to be finite dimensional. $M$ is a matrix to be learned in the high-dimensional space $\mathcal{F}$. The similarity between two image sets $S_i$ and $S_j$ under $M$ and $\{\eta_p\}_{p=1}^P$ is defined as

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

where $\eta_p(\phi_i^p)$ is a gating function that assigns a positive weight to $\phi_i^p$, as detailed later. Because of $\eta_p(\phi_i^p)$, our learning method is localized. Clearly, previous global kernel weighting algorithms [2], [12] can be considered as a special case, where $\eta_p(\phi_i^p)$ is independent of $\phi_i^p$.

To learn the matrix $M$, we seek to simultaneously maximize inter-class variations and minimize intra-class variations. The learning criterion is

$$\max_{M,\{\eta_p\}_{1\leq p\leq P}} J = \frac{1}{N_{S^-}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^-}}^{N} d(S_i, S_j) - \frac{1}{N_{S^+}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^+}}^{N} d(S_i, S_j) \tag{5}$$

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

Denote $d^{\mathcal{F}}$ the dimensionality of the feature space. The $d^{\mathcal{F}} \times d^{\mathcal{F}}$ matrix $M$ is symmetric and positive semidefinite. We seek a matrix $W = [w_1, w_2, \ldots, w_d]$ of size $d^{\mathcal{F}} \times d$, where $d^{\mathcal{F}} \geq d$ and $d$ is the number of weight vectors in $W$, such that

$$M = WW^T. \tag{6}$$

Combining (4)–(6), we express $J$ as

$$J = \mathrm{tr}[W^T(A_1 - A_2)W] \tag{7}$$

[1]While more combined statistics could be computed from the first-order and second-order statistics, we only compute one in this paper because it is very expensive to compute such features.

where

$$A_1 = \frac{1}{N_{S^-}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^-}}^{N} \sum_{p=1}^{P} \eta_p\left(\phi_i^p\right)\left(\phi_i^p - \phi_j^p\right) \times \left(\phi_i^p - \phi_j^p\right)^T \eta_p\left(\phi_j^p\right) \tag{8}$$

$$A_2 = \frac{1}{N_{S^+}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^+}}^{N} \sum_{p=1}^{P} \eta_p\left(\phi_i^p\right)\left(\phi_i^p - \phi_j^p\right) \times \left(\phi_i^p - \phi_j^p\right)^T \eta_p\left(\phi_j^p\right) \tag{9}$$

are $d^{\mathcal{F}} \times d^{\mathcal{F}}$ matrices.

In general, it is difficult or even impossible to compute $A_1$ and $A_2$ directly in the feature space $\mathcal{F}$ because the explicit form of $\phi_i^p$ is usually unknown. Hence, we use the kernel trick by expressing the weight vector $w_k$ as a linear combination of all the training samples in the mapped space

$$w_k = \sum_{i=1}^{N} u_i^k \phi_i^p. \tag{10}$$

Hence

$$\sum_{p=1}^{P} w_k^T \phi_j^p = \sum_{i,j=1}^{N} \sum_{p=1}^{P} u_i^k \left(\phi_i^p\right)^T \phi_j^p = \sum_{p=1}^{P} (u^k)^T K_{.i}^p \tag{11}$$

where $u^k$ is a $N$-vector with $i$th entry denoted by $u_i^k$, and $K_{.i}^p$ is the $i$th column of the $p$th kernel matrix $K^p$. This is an $N \times N$ kernel matrix, calculated from the $p$th statistic feature between each pair of image sets.

Then, (5) can be expressed as

$$\max_{U,\{\eta_p\}_{1\le p\le P}} J = \mathrm{tr}[U^T (B_1 - B_2) U] \tag{12}$$

where $U = [u^1, \ldots, u^d]$ is a $N \times d$ matrix ($N < d$) and

$$B_1 = \frac{1}{N_{S^-}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^-}}^{N} \sum_{p=1}^{P} \eta_p\left(\phi_i^p\right)\left(K_{.i}^p - K_{.j}^p\right) \times \left(K_{.i}^p - K_{.j}^p\right)^T \eta_p\left(\phi_j^p\right) \tag{13}$$

$$B_2 = \frac{1}{N_{S^+}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^+}}^{N} \sum_{p=1}^{P} \eta_p\left(\phi_i^p\right)\left(K_{.i}^p - K_{.j}^p\right) \times \left(K_{.i}^p - K_{.j}^p\right)^T \eta_p\left(\phi_j^p\right) \tag{14}$$

are symmetric $N \times N$ matrices.

Now we discuss how to choose the gating functions $\eta_p(\cdot)$. As in [12], we choose

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

which is parameterized by a vector $h_p$ and a scale factor $b_p$. This gating function is monotonically increasing, nonnegative, and is easy to differentiate with respect to $h_p$ and $b_p$.

Since $\phi_i^p$ is implicit, we express $h_p^T \phi_i^p$, similarly to (11), as

$$h_p^T \phi_i^p = \sum_{i=1}^{N} a_p^T \left(\phi_i^p\right)^T \phi_i^p = \sum_{i=1}^{N} a_p^T K_{.i}^p \tag{16}$$

where $a_p \in \mathbb{R}^{N\times 1}$ and $b_p \in \mathbb{R}^1$ are the parameters to be learned. Then, the gating function can be written as

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

To the best of authors' knowledge, there is no closed-form solution to the optimization problem in (12) because we aim to learn $U$ but have to infer $a_p$ and $b_p$ simultaneously. Hence, we use an alternating optimization algorithm. The approach is to fix $a_p$ and $b_p$, update $U$, update $a_p$ and $b_p$, and so on iteratively.

We first initialize $a_p$ and $b_p$ with small random numbers, $1 \le p \le P$, and obtain $U$ by solving the optimization problem in (12). The columns of $U$ are constrained to be orthogonal. Then, $U$ can be obtained by solving the following eigenvalue problem:

$$(B_1 - B_2)u = \lambda u. \tag{18}$$

Write $U = [u_1, u_2, \ldots, u_g]$ such that the columns in $U$ are eigenvectors of (18) corresponding to the $g$ largest eigenvalues ordered according to $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_g$. Then, $U$ is the transformation matrix to be learned.

Having obtained $U$, we use the gradient descent method to update $\{a_p\}$ and $\{b_p\}$ as

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

$$b_p^{t+1} = b_p^t - \alpha \frac{\partial J}{\partial b_p} \quad \text{and} \quad 1 \le p \le P \tag{20}$$

where $\alpha$ is the learning rate, which is set to $10^{-6}$ in our experiments.

Having updated $\{a_p\}$ and $\{b_p\}$, $1 \le p \le P$, we recompute the weight $\eta_p(\phi_i^p)$ in (17), and then $B_1$ and $B_2$ in (13) and (14), respectively. Then, we update $U$ by resolving the eigenvalue equation in (18). We repeat this procedure until convergence. The proposed LMKML algorithm is summarized in Algorithm 1.

### C. Localized Multikernel Multimetric Learning

While LMKML combines multiple features using the learned distance metric, it only learns a single distance metric, which may not be powerful enough to exploit the specific information of each individual feature. This motivates us to learn feature-specific metrics. Moreover, different features of the same sample share the same identity information, and it is also necessary to exploit some common characteristic among different features. Hence, we propose a LMKMML to jointly learn feature-specific distances and sharable metrics so that complementary information can be better extracted.

**Algorithm 1** LMKML

**Input**: Training set: $P$ $N \times N$ kernels $K^p$, $1 \le p \le P$, computed from $N$ image sets, feature dimension $g$, tolerance parameter $\epsilon$.

**Output**: Transformation matrix $U$ and parameters $\{a_p\}$ and $\{b_p\}$, $1 \le p \le P$.

**Step 1 (Initialization):** Initialize $\{a_p^0\}$ and $\{b_p^0\}$, $1 \le p \le P$, with small random numbers.

**Step 2 (Local optimization):** For $t = 1, 2, \cdots, T$, repeat

**2.1**. Compute $B_1$ and $B_2$ using (13) and (14).

**2.2**. Solve the eigenvalue problem in (18), obtain $U^t = [u_1, u_2, \cdots, u_g]$.

**2.3**. For each $p$ ($1 \le p \le P$), update $a_p$ and $b_p$ using (19) and (20).

**2.4**. 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 $\{b_p\}$, $1 \le p \le P$.

Let $M_p$ be the distance metric to be learned for the $p$th feature, $1 \le p \le P$, and $M_0$ be the shared distance metric to be learned. We define the similarity between two image sets $S_i$ and $S_j$ under $\{M_p\}_{p=1}^P$, $M_0$, and $\{\eta_p\}_{p=1}^P$ as

$$d'(S_i, S_j) = \sum_{p=1}^{P} \eta_p\big(\phi_i^p\big)\big(\phi_i^p - \phi_j^p\big)^T (M_p + M_0) \times \big(\phi_i^p - \phi_j^p\big)\eta_p\big(\phi_j^p\big) \tag{21}$$

where $\eta_p(\phi_i^p)$ is a gating function that assigns a positive weight to $\phi_i^p$, as in LMKML.

Similarly, we formulate LMKMML as the following optimization problem:

$$\max_{M_0, \{M_p\}_{1\le p\le P}, \{\eta_p\}_{1\le p\le P}} H = H_1 - \lambda H_2 \tag{22}$$

where

$$H_1 = \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^-}}^{N} \frac{d'(S_i, S_j)}{N_{S^-}} - \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^+}}^{N} \frac{d'(S_i, S_j)}{N_{S^+}} \tag{23}$$

$$H_2 = \|M_0 - M\|_F^2 + \delta \sum_{p=1}^{P} \|M_p\|_F^2. \tag{24}$$

Here $\lambda$ and $\delta$ are two parameters that balance the contributions of different terms in the objective function and $M$ is the distance metric learned by LMKML.

In (22), $H_1$ aims to make the learned distance metrics discriminative, and $H_2$ models the interaction between the individual distance metric and the shared metric. Specifically, if $\delta$ is large, our LMKMML reduces to learning $P$ individual distance metrics. Otherwise, it degrades to LMKML if $\delta$ is small. Therefore, LMKMML can be considered as a special case of LMKMML if $\delta$ is set to zero, which enforces the individual distance metric to be as close as to the shared metric.

Fig. 3. From top to bottom: exemplar face images cropped from the Honda, MoBo, YTC, and YTF datasets, respectively, where images in the same row are face samples from the same person that were captured in different environments.

Since $M$ and $M_p$ are symmetric and positive semidefinite, we decompose $(M_0 + M_p)$ into $L_p L_p^T$, where $L_p \in \mathbb{R}^{d^{\mathcal{F}} \times d}$, and rewrite $H_1$ as

$$H_1 = \sum_{p=1}^{P} \mathrm{tr}\big[L_p^T \big(C_1^p - C_2^p\big) L_p\big] = \sum_{p=1}^{P} \mathrm{tr}\big(L_p^T R L_p\big) \tag{25}$$

where

$$C_1^p = \frac{1}{N_{S^-}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^-}}^{N} \eta_p\big(\phi_i^p\big)\big(\phi_i^p - \phi_j^p\big) \times \big(\phi_i^p - \phi_j^p\big)^T \eta_p\big(\phi_j^p\big) \tag{26}$$

$$C_2^p = \frac{1}{N_{S^+}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^+}}^{N} \eta_p\big(\phi_i^p\big)\big(\phi_i^p - \phi_j^p\big) \times \big(\phi_i^p - \phi_j^p\big)^T \eta_p\big(\phi_j^p\big) \tag{27}$$

$$R \triangleq \big(C_1^p - C_2^p\big) \tag{28}$$

are $d^{\mathcal{F}} \times d^{\mathcal{F}}$ matrices.

It is also difficult or even impossible to compute $C_1^p$ and $C_2^p$ directly in the feature space $\mathcal{F}$ because the explicit form of $\phi_i^p$ is usually unknown. Hence, we use the kernel trick and express $H_1$ as

$$H_1 = \sum_{p=1}^{P} \mathrm{tr}\big[U_p^T \big(D_1^p - D_2^p\big) U_p\big] \tag{29}$$

where $U_p = [u_p^1, \ldots, u_p^g]$ is an $N \times d$ matrix ($N < g$), and

$$D_1^p = \frac{1}{N_{S^-}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^-}}^{N} \eta_p\big(\phi_i^p\big)\big(K_{.i}^p - K_{.j}^p\big) \times \big(K_{.i}^p - K_{.j}^p\big)^T \eta_p\big(\phi_j^p\big) \tag{30}$$

**Algorithm 2** LMKMML

**Input**: Training set: $P$ $N \times N$ kernels $K^p$, $1 \le p \le P$, computed from $N$ image sets, feature dimension $g$, tolerance parameter $\epsilon$.

**Output**: Transformation matrix $U_0$, $U_p$, and parameters $\{a_p\}$ and $\{b_p\}$, $1 \le p \le P$.

**Step 1 (Initialization):**

**1.1**. Initialize $\{a_p^0\}$ and $\{b_p^0\}$, $1 \le p \le P$, with small random numbers.

**1.2**. Initialize $U_0 = U$, where $U$ is learned by LMKML.

**1.3**. Initialize $U_p$ with random matrices, $1 \le p \le P$, where each element is a small number.

**Step 2 (Local optimization):**

For $t = 1, 2, \cdots, T$, repeat

**2.1**. Fix $U_p$ and $U_0$, update $a_p$ and $b_p$.

**2.2**. Fix $a_p$, $b_p$ and $U_0$, update $U_p$.

**2.3**. Fix $a_p$, $b_p$ and $U_p$, update $U_0$.

**2.4**. If $t > 2$, $|a_p^{t+1} - a_p^t| < \epsilon$ and $|b_p^{t+1} - b_p^t| < \epsilon$ or $|U_p^{t+1} - U_p^t| < \epsilon$, go to Step 3.

**Step 3 (Output transformation matrix and parameters):**

Output $U_0$, $U_p$, $\{a_p\}$ and $\{b_p\}$, $1 \le p \le P$.

$$D_2^p = \frac{1}{N_{S^+}} \sum_{\substack{i,j=1 \\ (S_i,S_j)\in S^+}}^{N} \eta_p\left(\phi_i^p\right)\left(K_{.i}^p - K_{.j}^p\right) \times \left(K_{.i}^p - K_{.j}^p\right)^T \eta_p\left(\phi_j^p\right) \tag{31}$$

are $N \times N$ matrices.

In general, $H_2$ is hard to simplify because the explicit form of $M_p$ and $M_0$ is unknown. To address this, we employ an alternative method by enforcing the constraints on $U_p$ and $U_0$ so that $H_2$ can be rewritten as

$$H_2 = \|U_0 - U\|_F^2 + \delta \sum_{p=1}^{P} \|U_p\|_F^2 \tag{32}$$

where $U$ is the projection matrix of LMKML.

Now, (22) can be rewritten as

$$\max_{U_0,\{U_p\}_{1\le p\le P},\{\eta_p\}_{1\le p\le P}} H = \sum_{p=1}^{P} \mathrm{tr}\left[U_p^T\left(D_1^p - D_2^p\right)U_p\right] - \lambda\left(\|U_0 - U\|_F^2 + \delta \sum_{p=1}^{P} \|U_p\|_F^2\right). \tag{33}$$

There appears to be no closed-form solution to the optimization problem in (33). Hence, we use an alternating minimization algorithm, which is similar to that used in LMKML. The approach is to fix $a_p$, $b_p$, and $U_0$ to update $U_p$, then update $a_p$ and $b_p$ by fixing $U_0$ and $U_p$, and finally update $U_0$ by fixing $a_p$, $b_p$, and $U_p$. Gradient descent is used to update the parameters $\{a_p, b_p\}$, $U_p$, and $U_0$, where $1 \le p \le P$. The proposed LMKMML algorithm is summarized in Algorithm 2.

TABLE I

VALUE OF $N$ FOR FACE DATASETS IN OUR EXPERIMENTS

| Dataset | Honda | MoBo | YTC | YTF |
|---|---|---|---|---|
| $N$ | 40 | 48 | 141 | 141 |

TABLE II

RANK-ONE RECOGNITION RATES (%) OF DIFFERENT IMAGE-SET-BASED FACE RECOGNITION METHODS ON THE HONDA, MOBO, AND YTC DATASETS

| Method | Honda | MoBo | YTC |
|---|---|---|---|
| DCC [25] | 94.9 | 88.1 | 71.8 |
| MMD [50] | 94.9 | 91.7 | 73.7 |
| MDA [48] | 97.4 | 94.4 | 73.1 |
| AHISD [3] | 89.5 | 94.1 | 72.5 |
| CHISD [3] | 92.5 | 95.8 | 72.4 |
| SANP [19] | 93.6 | 96.1 | 73.3 |
| CDL [49] | 97.4 | 87.5 | 74.7 |
| RNP [59] | 98.5 | 96.3 | 74.3 |
| LMKML | **98.5** | **96.3** | **78.2** |
| LMKMML | **98.5** | **96.7** | **78.5** |

*1) Comparison With Multitask Large-Margin Metric Learning (MT-LMNN) [39]:* The MT-LMNN method also learns multiple metrics for classification, and has been proposed for real-world insurance data classification and speech recognition [39]. However, there are two differences between our LMKMML and MT-LMNN.

1) The weights of different metrics are learned in a localized manner in our LMKMML, and MT-LMNN learns them in a global way.
2) Our LMKMML is a kernel-based metric learning method, while MT-LMNN is a linear method.

### D. Recognition

For image-set-based face identification, given a test image set $X_T$, we first compute its $P$ statistics for feature representation, denoted by $x_T^p$, $1 \le p \le P$. Then, we calculate the similarity between $X_T$ and each training image set $X_i$ by (4) for LMKML and (21) for LMKMML. Finally, we classify the test image set $X_T$ into the class $c$ that achieves

$$c = \arg\min_i d(S_T, S_i). \tag{34}$$

Unlike face identification, the goal of face verification is to determine whether a given pair of face image sets comes from the same person or not. For face verification, the receiver operating characteristic (ROC) curve, which describes the tradeoff between false acceptance rate and true acceptance rate, is used for evaluation. The positive and negative pairs in the training set are used to compute $A_1$ and $A_2$ in (8) and (9), respectively, to learn the discriminative distance metric. Having obtained the distance metric $M$, we first compute the distance between two face image sets using (21) and normalize these distances in the range [0, 1]. Finally, the ROC curve is computed.

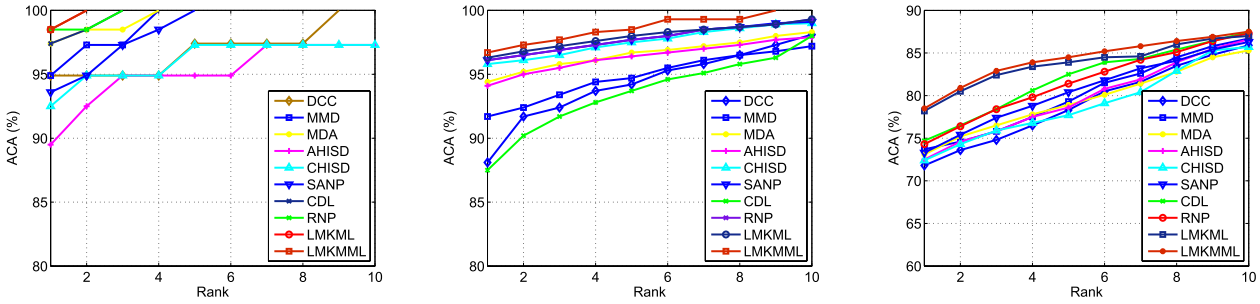



Fig. 4. Cumulative match characteristic (CMC) curves (%) of different image-set-based face recognition methods on (a) Honda, (b) MoBo, and (c) YTC datasets, respectively, where average classification accuracy (ACA) denotes the ACA.

TABLE III
COMPARISONS OF THE MEAN VERIFICATION RATE AND STANDARD ERROR (%) WITH THE STATE-OF-THE-ART RESULTS ON THE YTF DATASET UNDER THE IMAGE RESTRICTED SETTING

| Method | Accuracy |
|---|---|
| MBGS (LBP) [54] | $76.4 \pm 1.8$ |
| APEM (LBP) [30] | $77.4 \pm 1.5$ |
| APEM (fusion) [30] | $79.1 \pm 1.5$ |
| STFRD+PMML [5] | $79.5 \pm 2.5$ |
| MBGS+SVM⊖ (LBP) [55] | $79.5 \pm 2.5$ |
| VSOF+OSS (Adaboost) [36] | $79.7 \pm 1.8$ |
| Method in [38] [38] | $82.4 \pm 1.1$ |
| PHL+SILD (LBP) [22] | $80.2 \pm 1.3$ |
| DDML (LBP) [17] | $81.3 \pm 1.6$ |
| DeepFace-single [42] | $91.4 \pm 1.1$ |
| EigenPEP [31] | $84.8 \pm 1.4$ |
| LM3L [18] | $81.3 \pm 1.2$ |
| LMKML | $\mathbf{82.3 \pm 1.4}$ |
| LMKMML | $\mathbf{82.7 \pm 1.5}$ |

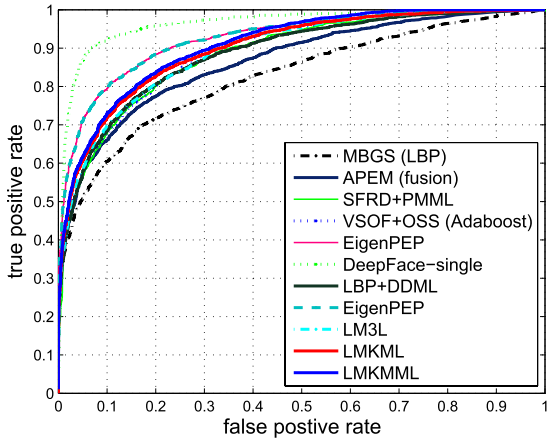



Fig. 5. Comparisons of ROC curves between our work and the state-of-the-art methods on the image restricted YTF dataset.

## IV. EXPERIMENTS

We evaluated our proposed approach on four publicly available video face databases including the Honda [28], MoBo [13], YouTube Celebrities (YTC) [23], and YouTube Face (YTF) [54] datasets. The Honda, MoBo, and YTC datasets are used to evaluate our face identification method, and the YTF dataset is selected to evaluate our face verification method.

### A. Datasets

There are 59 videos of 20 subjects in the Honda dataset. For each subject, 1–3 videos were collected, each containing around 400 frames with pose and expression variations.

There are 96 videos of 24 subjects in the Carnegie Mellon University MoBo dataset. For each subject, four video sequences were collected, each of which corresponds to a different walking pattern, and composed of about 300 frames.

The YTC dataset contains 1910 videos of 47 celebrities that were collected from YouTube. Most videos have low resolution and are highly compressed. The number of image frames in those videos in this dataset ranges from 8 to 400.

The YTF dataset contains 3425 videos of 1596 subjects that were also downloaded from YouTube. The average length of each video clip is about 180 frames. There are large variations in pose, illumination, expression, and resolution in these videos.

In the Honda, MoBo, and YTC datasets, each image frame is first automatically detected by the face detector method proposed in [46] and then resized to $20 \times 20$. For the YTF dataset, each image frame is cropped to $20 \times 20$ according to the provided eye coordinates. Hence, $d$ is set to 400 in our implementations. For each image frame in all these four datasets, we perform histogram equalization to remove illumination effects. Fig. 3 shows some cropped face images from the four datasets after resizing.

### B. Experimental Settings

To make a fair comparison with previous methods, we followed the same protocol used in [3], [19], and [48]–[50]. For the Honda, MoBo, and YTC datasets, we conducted experiments 10 times by randomly selecting gallery/probe combinations and computed the mean identification/verification rates. Specifically, we randomly selected one image set for each person as the gallery set and the remaining image sets were used for probes. For the YTC dataset,

TABLE IV
RECOGNITION/VERIFICATION RATES (%) OF DIFFERENT STATISTIC FEATURES ON THESE FOUR DATASETS

| Method | Honda | MoBo | YTC | YTF |
|---|---|---|---|---|
| First-order | 95.4 | 92.3 | 72.7 | 79.5 |
| Second-order | 96.5 | 88.9 | 67.5 | 80.5 |
| Combined statistic | 97.2 | 94.2 | 76.2 | 80.9 |
| LMKML | **98.5** | **96.3** | **78.2** | **82.3** |
| LMKMML | **98.5** | **96.7** | **78.5** | **82.7** |

TABLE V
RECOGNITION/VERIFICATION RATES (%) OF LMKML WHEN DIFFERENT COMBINATIONS OF STATISTIC FEATURES ARE USED ON THESE FOUR DATASETS

| Method | Honda | MoBo | YTC | YTF |
|---|---|---|---|---|
| Two features | 97.8 | 95.7 | 77.5 | 81.7 |
| Three features | **98.5** | **96.3** | **78.2** | **82.3** |

TABLE VI
RECOGNITION/VERIFICATION RATES (%) OF DIFFERENT MULTIKERNEL METRIC LEARNING METHODS ON DIFFERENT DATASETS

| Method | Honda | MoBo | YTC | YTF |
|---|---|---|---|---|
| GMKML | 98.3 | 95.4 | 76.7 | 81.1 |
| LMKML | **98.5** | **96.3** | **78.2** | **82.3** |

TABLE VII
RECOGNITION/VERIFICATION RATES (%) OF DIFFERENT MULTIKERNEL MULTIMETRIC LEARNING METHODS ON DIFFERENT DATASETS

| Method | Honda | MoBo | YTC | YTF |
|---|---|---|---|---|
| GMKMML | 98.3 | 95.8 | 77.5 | 81.8 |
| KMT-LMNN | 98.3 | 96.2 | 77.9 | 82.1 |
| LMKMML | **98.5** | **96.7** | **78.5** | **82.7** |

the whole dataset was equally divided into five folds (with minimal overlapping) each containing nine videos per subject. In each fold, three image sets per subject were randomly selected as the gallery set and the remaining six were selected as the probe sets. For the YTF dataset, we followed the standard evaluation protocol in [54] by evaluating our method on 5000 video pairs. Half of them were from the same person and the remaining half were from different persons. These pairs were equally divided into 10 folds and each fold contains 250 intra-personal pairs and 250 inter-personal pairs, respectively. We trained our LMKML and LMKMML on the YTC dataset and used the 10-fold cross validation strategy for face verification on the YTF dataset [54]. The value of $N$ for the datasets in our experiments is given in Table I.

### C. Results and Analysis

*1) Comparison With Existing Image Set Based Face Recognition Methods:* We conducted image-set-based face identification experiments on the Honda, MoBo, and YTC datasets and compared the proposed approach with several recently proposed methods, including discriminant canonical correlation (DCC) analysis [25], manifold-to-manifold distance (MMD) [50], manifold discriminant analysis (MDA) [48], affine-hull-based image set distance (AHISD) [3], convex-hull-based image set distance (CHISD) [3], sparse approximated nearest point (SANP) [19], covariance discriminative learning (CDL) [49], and regularized nearest points (RNPs) [59].

TABLE VIII
RECOGNITION/VERIFICATION RATES (%) OF SINGLE-METRIC AND MULTIMETRIC LEARNING METHODS ON DIFFERENT DATASETS

| Method | Honda | MoBo | YTC | YTF |
|---|---|---|---|---|
| LMKML (no regularizer) | 98.5 | 96.3 | 78.2 | 82.3 |
| LMKML (with regularizer) | 98.5 | 96.5 | 78.3 | 82.5 |
| LMKMML | **98.5** | **96.7** | **78.5** | **82.7** |

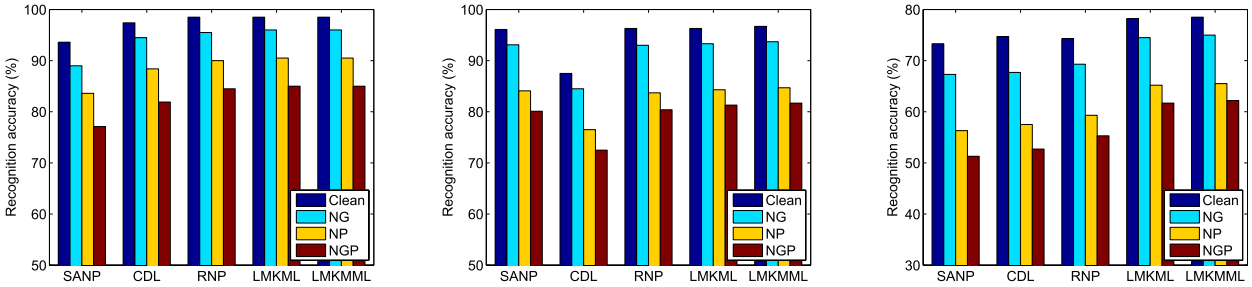
We employed the standard implementations of all these methods except CDL, which we implemented because the source code has been not released. We tuned the parameters of different methods for a fair comparison. Specifically, we applied PCA to learn a linear subspace for DCC and selected a 20-D subspace for similarity measure. For MMD and MDA, the maximum canonical correlation was used to compute MMD, and the number of nearest neighbors was set to 15. For AHISD, there was no parameter. For CHISD and SANP, we followed the same parameter settings as described in [3] and [19]. For CDL, the kernel linear discriminant analysis (KLDA) was used for discriminative learning so that it is fair to compare it with our LMKML. The regularization parameter of CDL was the same as in [49]. For RNP, there are two regularization parameters: 1) $\lambda_1$ and 2) $\lambda_2$. In our experiments, we followed the same setting in [59] and set $\lambda_1$ and $\lambda_2$ to 0.001 and 0.1, respectively. For our LMKML and LMKMML methods, the radial basis function kernel was used and the standard deviation from each statistic feature was used as the parameter to estimate the kernel values. The parameters $\lambda$ and $\delta$ were empirically set to 1 and 0.2 using the cross-validation strategy from the training set. For DCC, MDA, CDL, and our methods, since there is a single gallery image set from each class in the Honda and MoBo datasets, we randomly divided each gallery set into two subsets to model the within-class variation.

Table II tabulates the average rank-one recognition rates and Fig. 4 shows the cumulative match characteristic (CMC) curves of different methods on the Honda, MoBo, and YTC datasets, respectively.[2] Our methods outperform the other ones, especially on the most challenging YTC dataset. This is because the other 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 methods and hence better performance is obtained.

*2) Comparison With the State-of-the-Art Face Verification Methods:* We compared our approach with the state-of-the-art face verification methods on the YTF dataset.[3] These

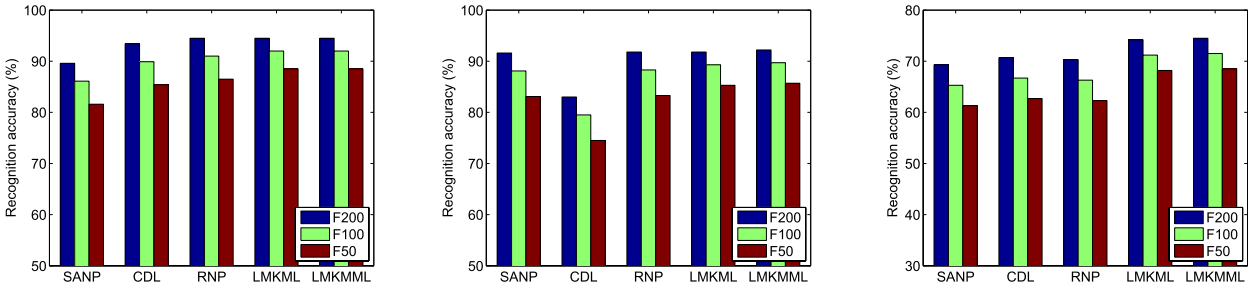
[2]The recognition performance of the other methods on the YTC dataset is much better than that reported in [34]. The reason is that we found for our detected YTC dataset that the optimal parameters of these methods are not the default ones that were recommended by the original papers.

[3]Available from: http://www.cs.tau.ac.il/~wolf/ytfaces/results.html.




(a) (b) (c)

Fig. 6. Rank-one recognition rate (%) of different image set based face recognition methods with noisy data on (a) Honda, (b) MoBo, and (c) YTC datasets, respectively.




(a) (b) (c)

Fig. 7. Rank-one recognition rate (%) of different image set classification methods with varying data size on (a) Honda, (b) MoBo, and (c) YTC datasets, respectively.

compared methods include matched background similarity (MBGS) [54], APEM [30], STFRD + PMML [5], MBGS + SVM⊖ [55], VSOF + OSS (Adaboost) [36], the method in [38], PHL + SILD [22], DDML (LBP) [17], DeepFace-single [42], EigenPEP [31], and LM3L [18]. Table III and Fig. 5 show the mean verification rate with the standard error and ROC curves on the YTF dataset, respectively. Our LMKML and LMKMML achieve competitive performance in terms of mean verification rate. While existing state-of-the-art video-based face verification methods only considered the mean information of the video, our methods exploit more statistical information. This makes it possible to better capture the relationship between different frames within the video and extract complementary information for verification.

*3) Comparison of Statistic Features:* We compared the discriminative power of different statistic features for image set classification. For each feature, we performed face recognition/verification on different datasets. Table IV tabulates the classification rates. We observe that the combined statistic feature achieves better classification performance than the other two features.

To further show the advantage of the combined statistic feature, we removed it in our LMKML and performed face recognition/verification by combining the first-order and second-order features with LMKML. Table V shows the classification rates. We observe that the combination of three features slightly outperforms that of two features in LMKML.

*4) Localized Versus Global Multikernel Metric Learning:* The multikernel distance metric can also be learned in a global manner. To show the effect of localized multikernel learning, we assume that $\eta_p(\phi_i^p)$ is constant and learn a distance using the global multikernel metric learning (GMKML) algorithm, where the weights of different kernels are learned and updated by the multiview metric learning method of [33]. Moreover, we also compare LMKML with the kernel-based multitask LMNN method [39], where the MT-LMNN method was applied in the kernel space and the weight is learned in a global way. Tables VI and VII show the recognition rates of these three methods with different learning strategies. We observe that our localized methods achieve better performance than the GMKML methods. This shows that learning a data-specific kernel is better because it exploits more geometrical information of samples in learning the distance metric(s).

*5) Single-Metric Versus Multimetric Learning:* To better show the advantage of multimetric learning over single-metric learning with multifeature representation, we also develop another baseline for LMKML by adding a regularizer on $M$ in LMKML, where the regularizer is introduced by following the same procedure in LMKMML and the only difference is we here only need to regularize one metric while LMKMML regularizes multiple metrics. Table VIII shows the recognition rates of these metric learning strategies. We observe that multimetric learning outperforms single-metric learning and the regularizer slightly improves the recognition rates on different datasets.

*6) Robustness Analysis:* We tested the robustness of our methods in case there is noise in the image sets and the image sets are of different sizes. We conducted three experiments using the same settings as in [3] and [49]: 1) the gallery

image sets were noisy; 2) the probe image sets were noisy; and 3) both the gallery and probe image sets were noisy. To make the image set noisy, we randomly selected one image from the other classes and included it in the current image set. For these three settings, we called the original and three noisy datasets clean, NG (only gallery image sets were noisy), NP (only probe sets were noisy), and NGP (both gallery and probe sets were noisy), respectively. Fig. 6 shows the mean identification rates of different image set based face recognition methods on different face datasets.

We also evaluated the performance of our approach when there are different number of frames in the image sets. We randomly selected a subset from each image set for recognition. We extracted 200, 100, and 50 frames from each image set, denoted by F200, F100, and F50, respectively. When an image set contains fewer frames, all image frames in this set were used for evaluation. Fig. 7 shows the average recognition rates of different face recognition methods on the YTC dataset.

From Figs. 6 and 7, we observe that our proposed approach shows better robustness to these two challenges. That is because we use different statistics features as the set representation, which are robust to outliers and to the number of samples in the set. Hence, the effects of the noisy samples and varying data size are alleviated.

*7) Parameter Analysis:* Fig. 8 shows recognition accuracy versus iteration number on the YTC dataset. Our iterative methods rapidly achieve stable performance.

Fig. 9 shows recognition accuracy of LMKMML as a function of $\lambda$ and $\delta$ on the YTC dataset. LMKMML achieves stable performance when $\lambda$ and $\delta$ are set as around 1.0 and 0.2, respectively.

*8) Convergence Analysis:* Fig. 10 shows the values of the objective function of LMKML and LMKMML versus iteration number on the YTC dataset. Our algorithms converge in about 30–40 iterations.

*9) Computational Time:* We compared the computational time of different algorithms on the YTC dataset. Our hardware configuration comprises a 2.8-GHz CPU and a 10-GB RAM. Table IX shows the time spent by these methods for training and testing (per face image set). It is to be noted that training time is only required for discriminative learning methods such as DCC, MDA, CDL, and our methods. We see that the computational complexity of our methods is generally larger than the other methods. That is because our methods compute multiple features for image set representation, which requires more algebraic operations than other methods and hence higher computational complexity.

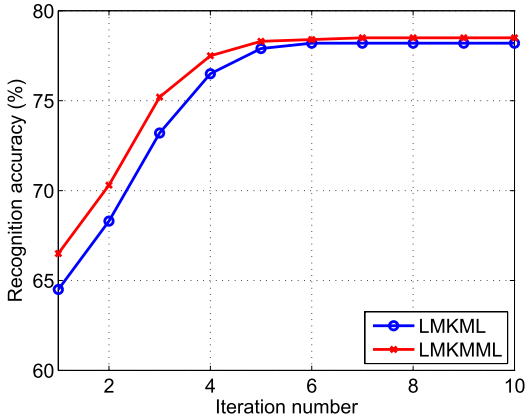



Fig. 8. Average recognition rate (%) of our methods versus iteration number on the YTC dataset.

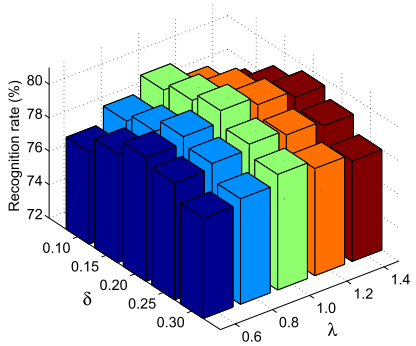



Fig. 9. Average recognition rate (%) of our LMKMML versus $\lambda$ and $\delta$ on the YTC dataset.

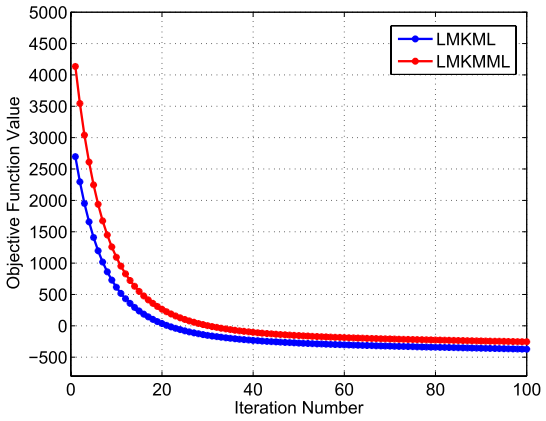



Fig. 10. Convergence curve of LMKML and LMKMML on the YTC dataset.

## D. Discussion

From the above experimental results, we make the following three observations.

1) Our proposed methods achieve better performance than the existing image-set-based face recognition methods on the Honda, MoBo, and YTC datasets, and obtain very competitive results on the YTF dataset. Compared with unsupervised methods such as MMD, AHISD, CHISD, and SANP, our methods extract discriminative information from face image sets, which is helpful to improve the recognition rate. Compared with supervised methods such as DCC, MDA, and CDL, our methods extract different statistic features from face image sets and hence exploit more complete information for classification.
2) Our methods consistently outperforms CDL on all four datasets. This is because our methods utilize different statistic features from each image set while CDL only extracts the second-order statistic feature for image set representation.

TABLE IX
COMPARISONS OF COMPUTATIONAL TIME (SECONDS) OF DIFFERENT METHODS ON THE YTC DATASET

| Method | DCC | MMD | MDA | AHISD | CHISD | SANP | CDL | LMKML | LMKMML |
|---|---|---|---|---|---|---|---|---|---|
| Training | 122.8 | N/A | 225.0 | N/A | N/A | N/A | 80.2 | 4755.8 | 5325.8 |
| Testing | 3.8 | 5.4 | 64.8 | 9.2 | 14.5 | 55.6 | 15.6 | 220.3 | 230.6 |

3) Our methods are more robust to outliers since they are statistics of all the samples in the image set and the effect of noise can be largely alleviated, especially compared with the previous nearest sample-pair-based image set classification methods such as AHISD, CHISD, and SANP.

## V. CONCLUSION

We have proposed a new image-set-based face recognition approach using multiple statistic features and localized multi-feature metric learning. Specifically, two kernel-based metric learning algorithms called LMKML and LMKMML were proposed to effectively combine multiple statistic features from face image set. The proposed approach has been evaluated on four widely used video face datasets. Experimental results show that our approach outperforms prior image-set-based face recognition methods in terms of both accuracy and robustness.

There are two interesting directions for future work.

1) The kernel computational method in this paper is time-consuming, which is one limitation of the proposed approach. It would be desirable to develop an efficient kernel approximation method to improve the kernel estimation speed, especially for combined statistic features.
2) In this paper, we applied LMKML and LMKMML for image-set-based face recognition. It would be interesting to use them for other visual analysis tasks such as visual recognition and information retrieval.

## REFERENCES


[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proc. IEEE CVPR*, Jun. 2005, pp. 581–588.

[2] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. 21st ICML*, 2004, pp. 1–8.

[3] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2567–2573.

[4] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *Proc. ECCV*, 2012, pp. 766–779.

[5] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3554–3561.

[6] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, "Image sets alignment for video-based face recognition," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2626–2633.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th ICML*, 2007, pp. 209–216.

[8] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.

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

[10] W. Fan and D.-Y. Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in *Proc. IEEE CVPR*, Jun. 2006, pp. 1384–1390.

[11] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood component analysis," in *Proc. NIPS*, 2004, pp. 2539–2544.

[12] M. Gönen and E. Alpaydin, "Localized multiple kernel learning," in *Proc. 25th ICML*, 2008, pp. 352–359.

[13] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-01-18, Jun. 2001.

[14] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE ICCV*, Sep./Oct. 2009, pp. 498–505.

[15] A. Hadid and M. Pietikainen, "From still image to video-based face recognition: An experimental analysis," in *Proc. 6th IEEE Int. Conf. FG*, May 2004, pp. 813–818.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proc. IEEE CVPR*, Jun. 2011, pp. 2705–2712.

[17] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1875–1882.

[18] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Proc. ACCV*, 2014, pp. 1–14.

[19] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *Proc. IEEE CVPR*, Jun. 2011, pp. 121–128.

[20] Y. Hu, A. S. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1992–2004, Oct. 2012.

[21] P. Jain, B. Kulis, and K. Grauman, "Fast image search for learned metrics," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.

[22] M. Kan, D. Xu, S. Shan, W. Li, and X. Chen, "Learning prototype hyperplanes for face verification in the wild," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3310–3316, Aug. 2013.

[23] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.

[24] T.-K. Kim, J. Kittler, and R. Cipolla, "Learning discriminative canonical correlations for object recognition with image sets," in *Proc. ECCV*, 2006, pp. 251–262.

[25] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[26] T.-K. Kim, J. Kittler, and R. Cipolla, "On-line learning of mutually orthogonal subspaces for face recognition by image sets," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1067–1074, Apr. 2010.

[27] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 2130–2137.

[28] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE CVPR*, Jun. 2003, pp. I-313–I-320.

[29] 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 Process. Lett.*, vol. 16, no. 3, pp. 227–230, Mar. 2009.

[30] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3499–3506.

[31] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt, "Eigen-pep for video face recognition," in *Proc. ACCV*, 2014, pp. 1–14.

[32] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.

[33] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, and G. Wang, "Neighborhood repulsed metric learning for kinship verification," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2594–2601.

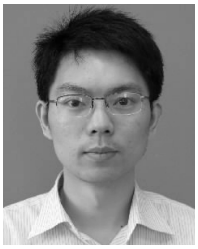
[34] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. IEEE ICCV*, Dec. 2013, pp. 329–336.

[35] Y. M. Lui, J. R. Beveridge, B. A. Draper, and M. Kirby, "Image-set matching using a geodesic distance and cohort normalization," in *Proc. 8th IEEE Int. Conf. FG*, Sep. 2008, pp. 1–6.
[36] H. Mendez-Vazquez, Y. Martinez-Diaz, and Z. Chai, "Volume structured ordinal features with background similarity measure for video face recognition," in *Proc. ICB*, Jun. 2013, pp. 1–6.
[37] A. Mian, Y. Hu, R. Hartley, and R. Owens, "Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5252–5262, Dec. 2013.
[38] H. A. B. Nguyen and W. Li, "Pose-robust representation for face verification in unconstrained videos," in *Proc. 20th IEEE ICIP*, Sep. 2013, pp. 3715–3719.
[39] S. Parameswaran and K. Q. Weinberger, "Large margin multi-task metric learning," in *Proc. NIPS*, 2010, pp. 1867–1875.
[40] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.
[41] G. Shakhnarovich, J. Fisher, and T. Darrell, "Face recognition from long-term observations," in *Proc. ECCV*, 2006, pp. 361–375.
[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1701–1708.
[43] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing KISS metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.
[44] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. ECCV*, 2008, pp. 548–561.
[45] G. Tsagkatakis and A. Savakis, "Online distance metric learning for object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1810–1821, Dec. 2011.
[46] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[47] J. Wang, H. T. Do, A. Woznica, and A. Kalousis, "Metric learning with multiple kernels," in *Proc. NIPS*, 2011, pp. 1–8.
[48] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proc. IEEE CVPR*, Jun. 2009, pp. 429–436.
[49] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2496–2503.
[50] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.
[51] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "Multiple kernel learning with high order kernels," in *Proc. 20th ICPR*, Aug. 2010, pp. 2138–2141.
[52] X. Wang, G. Hua, and T. X. Han, "Discriminative tracking by metric learning," in *Proc. ECCV*, 2010, pp. 200–214.
[53] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2005, pp. 1473–1480.
[54] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE CVPR*, Jun. 2011, pp. 529–534.
[55] L. Wolf and N. Levy, "The SVM-minus similarity score for video face recognition," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3523–3530.
[56] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao, "Set based discriminative ranking for recognition," in *Proc. ECCV*, 2012, pp. 497–510.
[57] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proc. 3rd IEEE Int. Conf. FG*, Apr. 1998, pp. 318–323.
[58] L. Yang *et al.*, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 30–44, Jan. 2010.
[59] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proc. 10th IEEE Int. Conf. FG*, Apr. 2013, pp. 1–7.
[60] B. Zhao, J. T. Kwok, and C. Zhang, "Multiple kernel clustering," in *Proc. SIAM ICDM*, 2009, pp. 638–649.
[61] Y. Zhao, S. Xu, and Y. Jia, "Discriminant clustering embedding for face recognition with image sets," in *Proc. ACCV*, 2007, pp. 641–650.
[62] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE CVPR*, Jun. 2011, pp. 649–656.
[63] P. Zhu, W. Zuo, L. Zhang, S. C.-K. Shiu, and D. Zhang, "Image set-based collaborative representation for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1120–1132, Jul. 2014.

**Jiwen Lu** (S'10–M'11) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore.

He is an Associate Professor with the Department of Automation, Tsinghua University, China. He has authored or co-authored over 100 scientific papers in his research areas, in which more than 30 papers are in IEEE Transactions journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, as well as the top-tier computer vision conferences, such as the International Conference on Computer Vision, the Conference on Computer Vision and Pattern Recognition (CVPR), and the European Conference on Computer Vision. His research interests include computer vision, pattern recognition, and machine learning.

Dr. Lu received the First Prize National Scholarship and the National Outstanding Student Award from the Ministry of Education of China in 2002 and 2003, the Best Student Paper Award from the PREMIA of Singapore in 2012, and the Top 10% Best Paper Award from MMSP in 2014. He serves as the Area Chair of the 2015 IEEE International Conference on Multimedia and Expo (ICME) and the 2015 IAPR/IEEE International Conference on Biometrics, and the Special Session Chair of the 2015 IEEE Conference on Visual Communications and Image Processing. He recently gave tutorials at conferences such as CVPR 2015, FG 2015, ACCV 2014, ICME 2014, and IJCB 2014.



**Gang Wang** (M'10) received the B.S. degree in electrical engineering from Harbin Institute of Technology, Harbin, China, in 2005 and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Illinois at Urbana—Champaign, Champaign, IL, USA, in 2010.

He is an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, and a Research Scientist with the Advanced Digital Sciences Center, Singapore. His research focuses on object recognition, scene analysis, large-scale machine learning, and deep learning. His research interests include computer vision and machine learning.

**Pierre Moulin** (F'03) received the Ph.D. degree from Washington University in St. Louis, St. Louis, MO, USA, in 1990.

He joined Bell Communications Research, Morristown, NJ, USA, as a Research Scientist. In 1996 he joined University of Illinois at Urbana—Champaign (UIUC), Champaign, IL, USA, where he is currently a Professor with the Department of Electrical and Computer Engineering, a Research Professor with the Beckman Institute and the Coordinated Science Laboratory, and an Affiliate Professor with the Department of Statistics. He was a Beckman Associate with the Center for Advanced Study, UIUC. From 2007 to 2009, he was a Sony Faculty Scholar with UIUC. His research interests include image and video processing, compression, statistical signal processing and modeling, media security, decision theory, and information theory.

Dr. Moulin received the Career Award from the National Science Foundation and the IEEE Signal Processing Society Senior Best Paper Award in 1997. He is a co-author of a paper that received the IEEE Signal Processing Society Young Author Best Paper Award with J. Liu in 2002. He has served on the Editorial Boards of IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON IMAGE PROCESSING, and PROCEEDINGS OF THE IEEE. He serves on the Editorial Boards of *Foundations and Trends in Signal Processing*. He was the Co-Founding Editor-in-Chief of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY from 2005 to 2008 and a member of the IEEE Signal Processing Society Board of Governors from 2005 to 2007, and has served the IEEE in various other capacities. He was a plenary speaker for the International Conference on Acoustics, Speech, and Signal Processing in 2006, the International Conference on Image Processing in 2011, and several other conferences. He was a Distinguished Lecturer of the IEEE Signal Processing Society from 2012 to 2013.