This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

Facial Analysis with Lie Group Kernel

Chunyan Xu, Canyi Lu, Junbin Gao, Tianjiang Wang, Shuicheng Yan

***Abstract*—To efficiently deal with the complex nonlinear variations of face images, a novel Lie group kernel is proposed in this work to address the facial analysis problems. Firstly, we present a linear dynamic model (LDM) based face representation to capture both the appearance and spatial information of the face image. Secondly, the derived linear dynamic model can be parameterized as a specially-structured upper triangular matrix, the space of which is proved to constitute a Lie group. A Lie group (LG) kernel is then designed to characterize the similarity between the linear dynamic models for any two face images and the kernel can be fed into classical kernel-based classifiers for different types of facial analysis. Finally, experimental evaluations on face recognition and head pose estimation are conducted on several challenging datasets and the results show that the proposed algorithm outperforms other facial analysis methods.**

***Index Terms*—Facial analysis, Lie group manifold, kernel learning.**

I. Introduction

Facial analysis has drawn much research interest in the past few decades. Nowadays, facial analysis techniques have been widely applied in intelligent monitoring, information security, law enforcement and human-computer interaction [1], [2], and are attracting a great deal of attention from both scientific and industrial communities.

The discrete probability distributions such as histograms have been successfully used for facial analysis, e.g. Local binary pattern [3], [4], Local derivative pattern [5] and Spatial pyramid histogram [6]. Despite its popularity, histograms have certain disadvantages such as the sensitivity to the number of bins, outliers and quantization errors. Zhou *et al.* [7] proposed a face representation algorithm based on Gaussian mixture models (GMMs) [8] for facial analysis, and employed the Kullback-Leibler divergence (KLD) between the GMM representations of images as the similarity/dissimilarity measure. Although hierarchical Gaussianization [7] can learn the hierarchical spatial information of a face image by analyzing each of these Gaussian maps, few probability distributions can model the spatial causality among image patches. For example, a different internal spatial constraint of a face image can indicate a different head pose.

In order to utilize the states transition probabilities of neighboring image patches, Li *et al*. [9] proposed an algorithm to model image patches by two dimensional hidden Markov models (HMM's). Recently, the dynamic texture has attracted the attention of many researchers as a useful tool in domains such as video synthesis, video segmentation, and video classification [10], [11]. Dynamic textures are image sequences of moving scenes that can be modeled as the output of a linear dynamical model. Under the dynamic texture model, a video is described as a sequence of images containing moving objects/scenes and showing certain stationary properties in time. Similarly, a face image can be encoded as a sequence of local patches, containing the spatial causality among patches in space.

Motivated by the recent progress in dynamic texture research, we propose a novel linear dynamic model (LDA) based face representation, which captures both the appearance and spatial structure information of the face image. The facial appearance (e.g. color, shape, texture) of a local patch is a linear function of the current spatial state vector with zero-mean Gaussian observation noise, and its corresponding spatial state is modeled as a first-order Gaussian Markov process. Moreover, face information is often reflected by local appearance, e.g., the nose, mouth and eyes. The position of a local patch may change due to the shape difference among different subjects, but not totally unconstrained. The proposed face representation naturally has the potential to model these face patches with a certain spatial relationship.

Recently, Gong *et al.* [13] represented the space of Gaussian distribution as a Lie group, which in this case is a connected Riemannnian manifold. This motivates us to parameterize each facial linear dynamic model as a specially-structured upper triangular matrix, the space of which can be identified as a Lie group [14]. It is well known that the distribution of face images, under a perceivable variation in viewpoint, illumination or facial expression, is highly nonlinear and complex. In order to efficiently deal with the complex nonlinear variations of face images, a Lie group (LG) kernel can be designed to characterize the similarity between the linear dynamic models for any two face images, and its discrimination power can be enhanced by exploiting the space structure and local information on Lie group.

In this paper, we propose an Lie group kernel based framework for analyzing face images. The whole framework consists of four steps: (1) extracting image features; (2) presenting linear dynamic model based face representation; (3) analyzing face images on the Lie group and (4) constructing a Lie group kernel. Take the face recognition as an example. As illustrated in Fig. 1, we demonstrate the framework of our algorithm. To cover a 2D image space with a 1D sequence, we evenly divide an face image into $n(n = l^2, l = 4.)$ patches, then, to produce an appropriate sequence, we scan these patches in a z-shape from the top-left of an image. Moreover, this kind of z-shape scanning, which is known as a good structure-preserving space

Corresponding author: Tianjiang Wang, tjwang@hust.edu.cn.

Email address: (xuchunyan01, canyilu) @gmail.com (C. Xu, C. Lu), jbgao@csu.edu.au (J. Gao), eleyans@nus.edu.sg (S. Yan).

C. Xu and T. Wang are with School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, P.R.China.

S. Yan and C. Lu are with Department of Electrical and Computer Engineering, National University of Singapore, 117583, Singapore.

J. Gao is with School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

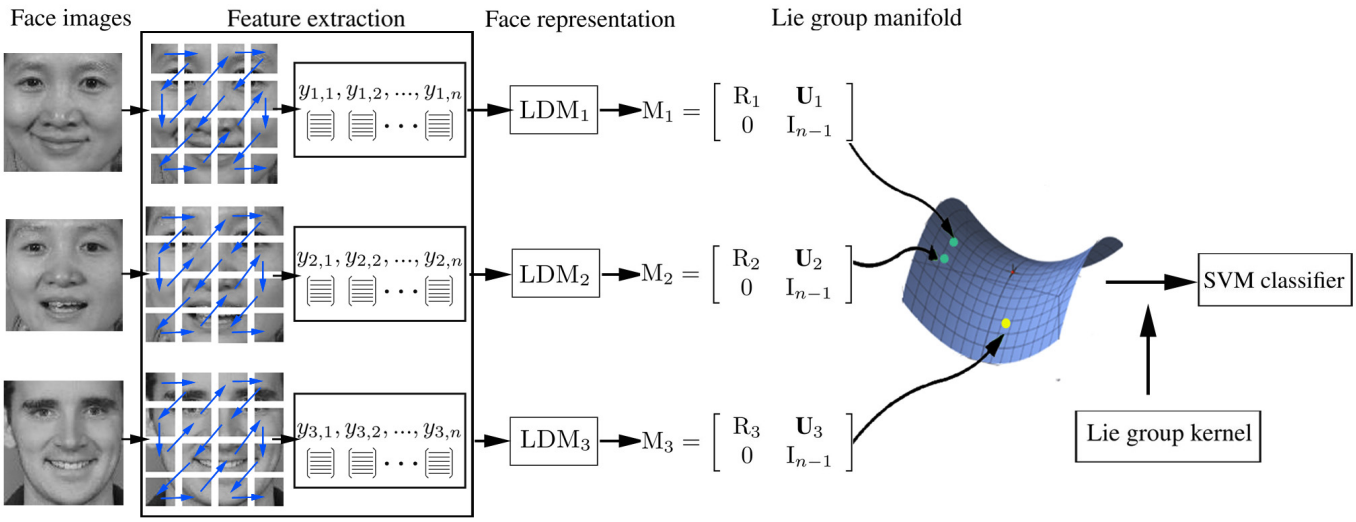


Fig. 1. The framework of facial analysis with Lie group kernel. Given several face images, which are from the FRGC V1.0 database [12], sequence facial features are extracted. The linear dynamic model (LDM) based face representation is constructed to capture both the appearance and spatial information of a face image. Then each facial linear dynamic model can be parameterized as a specially-structured upper triangular matrix, the space of which is identified as a Lie group. Finally, the SVM classifier with Lie group kernel is employed for the facial analysis problems.

filling curve algorithm employed in JPEG [15], captures both the horizontal and vertical spatial relationships among image patches. We then propose a novel approach to construct a linear dynamic model based face representation. This model can also be identified as a point on the Lie group manifold. Then an appropriate Lie group kernel suitable for Support Vector Machine (SVM) classification is developed based on a distance metric of the Lie group manifold. Finally, we conduct experiments to compare our proposed approach with some existing methods on facial analysis problems.

The remainder of this paper is organized as follows. In Section II, we will review some related work. Section III presents the LDM based face representation. In Section IV we briefly analyze the Lie group manifold and then construct a Lie group for the facial analysis problems. A Lie group kernel for SVM classification is presented in Section V, followed by experimental results in Section VI. The main findings and possible future research are summarized in Section VII.

II. Related Work

In the literature of face analysis, most previous approaches are based on global image features, including a great number of subspace-based techniques and some spatial frequency methods. Subspace-based techniques, such as principal component analysis (PCA) [16], linear discriminant analysis (LDA) [17] and independent component analysis (ICA) [18], have been popular for face analysis. These methods attempt to find a set of basis images from a training set and represent a face as a linear combination of these basis images. Many research works are also proposed to extract facial features by using spatial-frequency methods [19], which only reserve the coefficients in the low-frequency bands for face analysis.

Though global-based face representation was popular for face analysis, increasing attempts have been made to develop face analysis methods based on local features, which are considered more robust to the variations of facial expressions, illumination and low resolution. In [3], [4], [20], the local binary pattern (LBP) features, which are extracted from small regions of the face image, are adopted for face analysis. Gabor features [21], which mainly encode facial shape and appearance information, have also been used as a preprocessing stage for LBP feature extraction [4].

Recently, Lucey *et al.* [22] represented an image as a set of free patches, aiming to better employ local image features to overcome these limitations of global features. However, a human face becomes unintelligible to a human observer when the various local appearances are not in a proper spatial arrangement. Moreover, Yan *et al.* [8] introduced a local descriptor for image regression named coordinate patch, and then encoded each image as a sequence of coordinate patches. For a position $q = (q_x, q_y)^T$ within the image plane, its corresponding coordinate patch for a given image X is defined as $Q(x_q, q)^T = [f(x_q, q), q^T]$, where $f(x_q, q)$ denotes the feature vector extracted from the image patch x_q. In general the feature vector $f(x_q, q)$ is of high dimension, thus the coordinate information $q = (q_x, q_y)^T$ may be overshadowed in favor of the feature vectors.

While facial analysis can be accomplished through raw facial features like pixel values and bag of features [6], [5], it has been shown that better performance can be attained through analyzing the statistics of face images, such as probability distribution based methods [7]. A general probability distribution approach for facial analysis consists of the following three steps: (i) extracting facial features such as intensity, color, gradient, filter responses, etc., (ii) deriving the probability distribution of face images, and (iii) building a similarity metric between probability distributions to be fed into classical classifiers.

Much previous research [1] considers the facial analysis problems over certain manifolds by embedding face images onto an appropriate manifold such as a Riemannian mani-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

fold. A classification algorithm is then designed based on an appropriate similarity metric determined by the geometric properties of the manifold using the conventional statistical methods. Tuzel *et al.* [23] used the covariance matrices as feature descriptors, the space of which can be formulated as a connected Riemannian manifold. However, region convariance descriptor is an incomplete parameterized Gaussian distribution, and ignores the mean vector information of the image features. Turaga *et al.* [24] developed probability density distribution and estimation techniques that were consistent with the geometric structure of certain manifolds, for example, learning a parametric Langevin distribution on the Grassmannian manifold for each object class.

III. Linear Dynamic Model Based Face Representation

For the problem of facial analysis, face representation has many advantages when the object contains regions of distinctive details and spatial information. For example, the human face consists of distinctive local areas such as eyes, mouth and nose. The spatial constraint among image patches is relatively fixed and plays a decisive role in image recognition. Sequence image patches better capure the local sppearance feature and spatial relationships of an image, without explicitly employing any coordinate information.

Recently, Doretto *et al.* [11] proposed to treat the video clip as a sample from a linear dynamic model. The dynamic texture is a stochastic video model that treats the video as a sample from a linear dynamic model. Although it is simple, it has been shown that the linear dynamic model is surprisingly useful in domains such as video segmentation, video recogntion and video systhesis [10], [25].

Inspired by the recent progress in dynamic texture research, we propose to learn a linear dynamic model from sequence image patches of a face image. Thus a descriptive capability of the linear dynamic model can be used to characterize not only image appearance (e.g. color, shape, texture), but also the spatial causality among patches in a face image. The linear dynamic model based face representation can be shown in Fig. 2.

A. Linear Dynamic Model

The appearance of a local image patch is a realization from a stationary stochastic process with spatially invariant statistics. For the spatial constraint of a face image (spatial-varying appearance), the individual appearance of a local image patch is clearly not the independent realization from a stationary distribution, because there is an intrinsic spatial coherence in the process that needs to be captured. Therefore, the underlying assumption is that the appearance of the local image is the realization of the output of a linear dynamic model driven by white, zero-mean Gaussian noises.

For a given face image i, $Y_i = [y_{i,1}, y_{i,2}, ..., y_{i,n}] \in \mathbb{R}^{m\times n}$ is a sequence of n local patch appearances. We construct an linear dynamic model based face representation with local patch appearances Y_i of the face image i. Especially, the appearance of image patch $y_{i,j} \in \mathbb{R}^m$ is a linear function

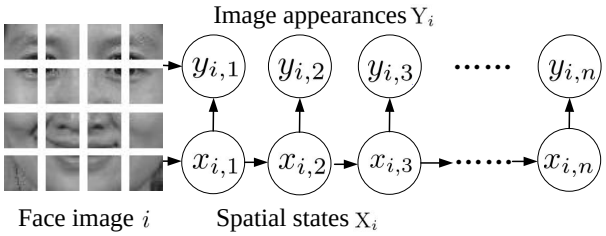


Fig. 2. An illustration of linear dynamic model based face representation. For example, we evenly split each image i into $n = 16$ local patches. $Y_i = [y_{i,1}, y_{i,2}..., y_{i,n}] \in \mathbb{R}^{m\times n}$ is a sequence of n local patch appearances. Each local patch appearance $y_{i,j}$ is modeled as a linear function of the current spatial state vector with zero-mean Gaussian observation noise. For a sequence of spatial image states (spatial-varying appearance) $X_i = [x_{i,1}, x_{i,2}, ..., x_{i,n}] \in \mathbb{R}^{k\times n}$, its linear dynamic model is modeled as a first-order Gaussian Markov process. And the example face image is from the FRGC V1.0 database [12].

of the current state vector with some observation noises, and the spatial causality relationship among image patches is represented as a state process $x_{i,j} \in \mathbb{R}^k$ with $k \leq m$. Therefore, the linear dynamic model of each face image i is:

$$\begin{cases} y_{i,j} = Cx_{i,j} + \omega_{i,j} \\ x_{i,j+1} = Ax_{i,j} + \nu_{i,j} \end{cases}, \tag{1}$$

where $C \in \mathbb{R}^{m\times k}$ is the orthonormal observation matrix, and $A \in \mathbb{R}^{k\times k}$ is the transition matrix. The state and observation noises are given by $\nu_{i,j} \sim \mathcal{N}(0, Q_i)$ and $\omega_{i,j} \sim \mathcal{N}(0, P_i)$.

For a given face image i, a sequence of $\{y_{i,j}\}_{j=1,2,...,n}$ encodes the appearance component of image patches, and the spatial causality component is encoded into the state sequence $\{x_{i,j}\}_{j=1,2,...,n}$. The hidden state is modeled as a first-order Gauss-Markov process [11], where the state at the patch $j+1$ of the face image i, $x_{i,j+1}$, is determined by the transition matrix A, the state at the image patch j, $x_{i,j}$, and the driving process $\nu_{i,j}$.

B. Parameter Estimation

The above problem we will solve can be formulated as follows: given measurements of a sample path of the process: $\{y_{i,1}, ..., y_{i,n}\}$, estimate the model parameters A, C, Q, P, a canonical realization of the process $y_{i,j}$. As described in [11], the choice of C results in a canonical realization. Then we would want the maximum-likelihood solution from finite sample, that is the argument of

$$A, C, Q, P = \arg \max_{A,C,Q,P} p(y_{i,1}, ..., y_{i,n}). \tag{2}$$

In general, the parameters of the above linear dynamic model can be learned by the maximum likelihood, e.g. N4SID [26]. Due to the possible high dimensionality of feature vectors, these algorithms are unfeasible for learning these parameters. Therefore, we intend to use the closed-form solution for the model parameters in this work. For N samples, $Y = [Y_1, Y_2, ..., Y_N] \in \mathbb{R}^{m\times(n\times N)}$ is the matrix of all image patches, and $Y = U\Sigma V^T$ with $U \in \mathbb{R}^{m\times m}$, $U^TU = I, \Sigma \in \mathbb{R}^{m\times(n\times N)}, V \in \mathbb{R}^{(n\times N)\times(n\times N)}$ and $V^TV = I$ is the singular

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

value decomposition (SVD) [27]. The unique solution can be given by

$$C = U(:, 1:k), X = \Sigma(1:k, :)V^T, \tag{3}$$

where $C \in \mathbb{R}^{m\times k}$ is the principal components of local patch appearances, $X = [X_1, ..., X_i, ..., X_N] \in \mathbb{R}^{k\times(n\times N)}$ is a matrix of image patch states estimated for all samples, and X_i is a matrix of local patch states of the face image i.

Given these state estimates, the transition matrix A is computed using the least squares estimate of the linear dependence of the spatial state [11] (assuming the spatial state random variables have zero mean),

$$A = Z_2 Z_1^{\dagger}, \tag{4}$$

where $Z_2 = [X_{1,2}^n, ..., X_{i,2}^n, ..., X_{N,2}^n]$, $Z_1 = [X_{1,1}^{n-1}, ..., X_{i,1}^{n-1}, ..., X_{N,1}^{n-1}]$, $X_{i,2}^n = [x_{i,2}, x_{i,3}, ..., x_{i,n}]$ and $X_{i,1}^{n-1} = [x_{i,1}, x_{i,2}, ..., x_{i,n-1}]$ are matrices of hidden state estimates, and $Z_1^{\dagger}$ is the pseudo-inverse of Z_1. The estimate of the covariance of the driving process is then,

$$Q_i = \frac{1}{n-1}\sum_{j=1}^{n-1} \nu_{i,j}\nu_{i,j}^T, \tag{5}$$

where $\nu_{i,j} = x_{i,j+1} - Ax_{i,j}$. Using the same method as Q_i, P_i can also be estimated from $y_{i,j}$, the observation matrix C and the state vector $x_{i,j}$.

The computational complexity of the above process is $O(m^2(n \times N) + (n \times N)^3)$, where m is the dimension of observation feature vector, n is the number of patches in a face image, N is the number of face images. For the pairwise distance between two LDMs, several attempts have been made to endow the space of linear dynamic models with a metric and probabilistic structure, such as Martin distance [28], Kullback-Liebler divergence (KLD) [29], etc. The model parameters learned as above do not lie on a linear topological space, but on a manifold. Thus we study the spatial structure of the estimated model parameters on the Lie group in the next section.

IV. Face Images on the Lie Group

In this work, we address the face analysis problems on the Lie group manifold. As the geometric properties of a manifold lead to appropriate definitions of the distance metric, most previous works formulate the facial analysis problem on certain manifolds based on appropriate geometric structures. Amari and Nagaoka [30] have stated that many important structures in information theory and statistics can be treated as structures in differential geometry by regarding a space of probabilities as a Riemannian manifold.

A manifold of n dimension is a Hausdorff topological space which has a countable base of open sets and is locally Euclidean of n dimension [1]. Riemannian manifolds are endowed with a distance measure which allows us to measure how similar two points are. Considering those general manifolds is beyond the scope of this work, and we are only interested in a particular class of Riemannian manifolds called Lie group manifold.

We shall analyze the geometric structure of the Lie group in this section, and then the derived linear dynamic model based face representation is used to construct a Lie group for the facial analysis problems.

A. Lie Group Analysis

A Lie group $\mathbf{G}$ is a smooth manifold with a group structure, in which the group operations of multiplication and inversion are smooth maps [14]. Smoothness of the smooth multiplication and inversion

$$\varphi : \mathbf{G} \times \mathbf{G} \to \mathbf{G}, \qquad \psi : \mathbf{G}^{-1} \to \mathbf{G}, \tag{6}$$

means that φ and ψ are smooth mappings of the product manifold and the inverse operation of manifold respectively.

The tangent space of the Lie group $\mathbf{G}$ to its identity element I forms a Lie algebra $\mathfrak{g}$. We can map between the Lie group and its tangent space from the identity element I using $\exp$ and $\log$ map,

$$\mathrm{m} = \log(M), \quad M = \exp(\mathrm{m}), \tag{7}$$

where $M \in \mathbf{G}$ and $\mathrm{m} \in \mathfrak{g}$ are elements of Lie group and Lie algebra, respectively. For matrix Lie groups, the exponential and logarithm maps of a matrix are given by

$$\log(M) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i}(M-I)^i, \quad \exp(\mathrm{m}) = \sum_{i=0}^{\infty}\frac{1}{i!}\mathrm{m}^i. \tag{8}$$

The distance of two points on the Lie group can be measured by the length of the curve connecting these two points. The minimum length curve between two points is called the geodesic. With the above logarithm map and the group operation, the geodesic distance [14] between two group elements can be computed as

$$D_{\mathrm{LG}}(M, M') = \| \log(M^{-1}M') \|, \tag{9}$$

where $M \in \mathbf{G}$, $M' \in \mathbf{G}$ and $\| \cdot \|$ is L_2 norm of a vector.

B. Lie Group from Facial linear dynamic model

To further analyze the linear dynamic model based face representation in Section 2, we parameterize the facial linear dynamic model as a specially-structured upper triangular matrix. The specially-structured upper triangular matrix M_i of the face image i is defined as follows.

$$M_i = \begin{bmatrix} R_i & \mathbf{U}_i \\ 0 & I_{n-1} \end{bmatrix}, \tag{10}$$

where $\mathbf{U}_i = [Ax_{i,1}, ..., Ax_{i,n-1}] \in \mathbb{R}^{k\times(n-1)}$ is the mean state vector set of the face image i. R_i is the Cholesky factorization of the positive definite covariance matrix Q_i, which means $R_i^T R_i = Q_i$. All such M_i form a Lie group $\mathbf{G}$, which can be proved by the Lie group definition.

The multiplication of any two group elements M_i and M_j is

$$\begin{aligned} M_i M_j &= \begin{bmatrix} R_i & \mathbf{U}_i \\ 0 & I_{n-1} \end{bmatrix}\begin{bmatrix} R_j & \mathbf{U}_j \\ 0 & I_{n-1} \end{bmatrix} \\ &= \begin{bmatrix} R_i R_j & R_i\mathbf{U}_j + \mathbf{U}_i \\ 0 & I_{n-1} \end{bmatrix}, \end{aligned} \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

where R_j is the Cholesky factorization of the positive definite covariance matrix Q_j. R_i and R_j are the upper triangular matrices. The multiplication of R_i and R_j, R_iR_j, is also an upper triangular matrix. We can see that the group multiplication $M_iM_j \in \mathbf{G}$ means that this is a smooth mapping φ of product manifold $\mathbf{G} \times \mathbf{G}$ into Lie group $\mathbf{G}$, $\mathbf{G} \times \mathbf{G} \to \mathbf{G}$.

And the inverse of any group element M_i is

$$M_i^{-1} = \begin{bmatrix} R_i & \mathbf{U}_i \\ 0 & I_{n-1} \end{bmatrix}^{-1} = \begin{bmatrix} R_i^{-1} & -R_i^{-1}\mathbf{U}_i \\ 0 & I_{n-1} \end{bmatrix}. \quad (12)$$

The group operations of inversion $M_i^{-1} \in \mathbf{G}$ is a smooth mapping ψ of the product manifold into the Lie group $\mathbf{G}$, $\mathbf{G}^{-1} \to \mathbf{G}$. Therefore, we can say that all such M_i form a Lie group, on which the group multiplication and inverse operation are the matrix multiplication and inverse respectively. The group operations of multiplication and inversion are smooth maps.

Then we can analyze structures of face linear dynamic models on the Lie group manifold. The geodesic length between two group elements M_i and M_j can be computed as $D_{\mathrm{LG}}(M_i, M_j) = \| \log(M_i^{-1}M_j) \|$. The geodesic is the curve with minimum length between two points on a manifold. So we can measure the distance of any two face linear dynamic models based on the defined Lie group.

V. Lie Group Kernel

As already discussed in section II, each image can be modeled as a LDM to capture both the appearance and spatial information of the face image. The above model parameters do not lie on a linear topological space, but on a Lie group. The geometric properties of Lie groups lead to an appropriate distance metric. Lie groups are endowed with a distance measure which allows us to measure how similar two points are. Our strategy here is to characterize the Lie group kernel between the linear dynamic models of two face images by exploiting the geometric properties of the Lie group.

In order to efficiently deal with the complex nonlinear variations of face images, a Lie group (LG) kernel is then designed to characterize the similarity between the linear dynamic models for any two face images and the kernel can be fed into SVM classifier for different types of facial analysis. The SVM [31] is one of examples implementing the statistical learning theory and it constructs a maximum-margin hyperplane between two classes using a set of training examples. The kernel trick is usually used in the SVM to learn a nonlinear classifier in a linear way in a high-dimensional feature space.

A kernel is actually a measure of the similarity of two points in a certain space. According to [32], any symmetric positive semi-define function, which satisfies Mercer's conditions, can be used as a kernel function in the SVM's context. With a valid distance metric, it is easy to define a kernel function. In our case, we have identified a useful distance measurement over the Lie group manifold, so it is quite natural to define an efficient kernel over the manifold.

Our Lie group kernel is defined as follows:

$$K_{\mathrm{LG}}(M_i, M_j) = \exp(-\gamma(D_{\mathrm{LG}}(M_i, M_j))), \quad (13)$$

where $D_{\mathrm{LG}}(M_i, M_j)$ is the distance metric between two face images on the Lie group manifold $\mathbf{G}$, and the parameter γ is directly related to scaling.

It is easy to prove that the newly defined Lie group kernel is a valid Mercer's kernel. First, since

$$D_{\mathrm{LG}}(M_i, M_j) = D_{\mathrm{LG}}(M_j, M_i), \quad (14)$$

we have

$$K_{\mathrm{LG}}(M_i, M_j) = K_{\mathrm{LG}}(M_j, M_i), \quad (15)$$

which means that the Lie group kernel is symmetric. K_{LG} is said to be non-negative definite if

$$\sum_{i=1}^{N}\sum_{j=1}^{N} K_{\mathrm{LG}}(M_i, M_j)c_ic_j \geq 0, \quad (16)$$

for all finite sequences of points $M_1, ..., M_N$ on the Lie group manifold and all choices of real numbers $c_1, ..., c_N$ and N denotes the number of training samples. This can be proved using the same way as [33]. Besides, computing the Lie group kernel demands for $O(N^2(n+k-1)^3)$, where $n+k-1$ is the dimension of the group elemets M_i.

VI. Experiments

To evaluate the effectiveness of our proposed method for the facial analysis problems, we systematically apply it on several face recognition and head pose estimation datasets.

A. Experimental Setups

In all our experiments, we evenly split each face image i into $n = 16$ local patches, which are used to construct the linear dynamic model based face representation. Image patches are densely sampled pixel by pixel within the corresponding image, and each patch size is set as height/$\sqrt{n}$ by width/$\sqrt{n}$ pixels. For example, if a face image is 64×64 pixels and the number n of face patches is 16, then the patch size will be 16×16 pixels. In order to produce an appropriate feature sequence, we scan these patches in a z-shape scanning, which is known as a good structure-preserving space filling curve algorithm employed in JPEG [15].

We use a feature vector to describe each patch of an face image. The GIST descriptor describes the image as a vector without detecting any interest point, and it performs well even for the low-resolution images [34]. In our below experiments, the GIST descriptor is employed to extract the $m = 512$ dimensional feature vector for each local patch. And the parameter dimension is set as $k = 40$. A Lie group kernel of the SVM classifier is then used to address the facial analysis problems on the Lie group. A one-versus-all scheme is used to tackle the multi-class problem, and the SVM training and testing are performed using the libsvm software package [31]. The parameter γ of LG kernel is selected by searching from a candidate set and report the best results.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

B. Comparing Algorithms

We compare the recognition performance of our proposed algorithm with the following SVM-based methods.

- KL Kernel and Martin kernel methods with patches. In these two kernel methods, the parameter of facial linear dynamic model is estimated the same as [11], and the derived Kullback-Liebler divergence (KLD) [10] and Martin distance [35] can be computed for any two facial linear dynamic models. The Martin kernel is then defined based on Martin distance [35], which is related to the principal angles between the subspaces of facial linear dynamic models.

$$K_{\text{Martin}}(p,q) = \exp(-\gamma(D_{\text{Martin}}(p,q)) + D_{\text{Martin}}(q,p))). \quad (17)$$

  The KL kernel [10] is defined as

$$K_{\text{KL}}(p,q) = \exp(-\gamma(D_{\text{KL}}(p,q) + D_{\text{KL}}(q,p))), \quad (18)$$

  where $D_{\text{KL}}(p,q)$ is the KL divergence between facial linear dynamic models in state spaces only for the sake of smaller computational complexity.
- Kernel methods without patches: linear kernel based SVM method and radial basis function (RBF) kernel based SVM method [31], [32].
- Baselines: Principal component analysis (PCA) [16], Linear discriminant analysis (LDA) [17], Local binary patterns (LBP) [3], [4], Gaussian kernel LDA (Gaussian KDA) [36], and Polynomial kernel LDA (Polynomial KDA)[36]. For fair comparison, we first perform dimensionality reduction with these baselines, and the dimension reduced features are then fed into linear SVM classifiers. For the PCA method, the feature dimension is set as #training-samples -1, while for LDA related methods, the feature dimension is set as #classes -1.

For all the experiments, the parameters for all the above kernels (e.g. KL kernel, Martin kernel, RBF kernel, Gaussian KDA and Polynomial KDA) are set the same as that for the LG kernel, namely searching within a predefined candidate set and reporting the best results.

C. Face Recognition Results

In our experiments, our proposed method firstly constructs LDM based face representation, then a Lie group kernel is designed to characterize the similarity between LDMs, and finally the kernel is fed into the SVM classifier for face recognition. This face recognition problem can be described as: *given a picture of a human face, to decide whether it is an example of a specific person, which may be done by comparing the face to a model for this person.*

To evaluate the effectiveness of the proposed method, we conduct experiments not only on two benchmark datasets with face images captured in controlled environments, i.e., AR [37], [38] and FRGC V1.0 [12], but also on two large-scale real-world face databases, i.e., LFW [39] and FRGC V2.0 [12]. Although LFW and FRGC V2.0 databases are designed to address the problem of *pair matching*, they can also be used to address other problems. Since our Lie group SVM kernel method cannot be applied to pair matching, we adopt a subset of the above two databases in our experiments. Specifically,

- **AR** database [37] consists of over 4,000 face images of 126 individuals. These images include front view faces with different expressions, illuminations and occlusions. In our experiments, we only use a subset of the AR face database [38]. This subset contains 1,400 face images corresponding to 100 persons (50 men and 50 women). A random subset with $l(= 3, 5, 7, 9)$ images per individual is taken with labels to form the training set, and the rest of the database is used as the testing set. The original resolution of these image faces is 165×120 pixels. Here, for computational convenience, we resize them to 66×48 pixels. We perform all processing in gray images. For each given l, we get average recognition accuracy over 20 random splits.
- **FRGC V1.0** [12] database consists of 5,658 images of 275 subjects. The facial image number for each subject varies from 6 to 48. We adopt two training/testing splitting strategies. For the first strategy, we randomly select half of the images for every single subject for model training; the rest images are used for testing. For the second strategy, we randomly choose 4 images of each subject and use the rest for testing. Then we conduct experiments with the two splittings for 20 times, respectively, and report the average classification accuracy for comparison. Since manually cropped faces are not available in FRGC database, the face images are automatically cropped and then normalized to the size of 32×32 pixels in the experiments. Some sample face images from AR database and FRGC V1.0 database are shown in Fig. 3.
- **FRGC V2.0** [12] is a large-scale face database which contains images captured in uncontrolled indoor and outdoor settings. This database provides 6 experimental protocols. Among them, Experiment 4 is considered the most challenging for image based face recognition. We use the subset (352 subjects and no less than 15 images for each subject) of Experiment 4 which have large lighting variations, aging and image blur. The selected target set contains 5,280 images, and the query set has 7,606 images. The image is normalized to 64×64 pixels. Three tests with 5, 10 and 15 target images for each subject are performed in the experiments. For each test, we get average recognition accuracy over 20 random splits.
- **LFW** [39] is a large-scale database of face images designed for unconstrained face recognition with variations in pose, illumination and expression, misalignment and occlusion, etc. Two subsets of aligned LFW [40] are used in the experiments. The image is normalized to 64×64 pixels. In subset 1 (LFW6) which consists of 311 subjects with no less than 6 samples per subject, we use the first 5 samples as training data and the remaining samples as testing data. In subset 2 (LFW11) which consists of 143 subjects with no less than 11 samples per subject, we use the first 10 samples as training data and the remaining

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

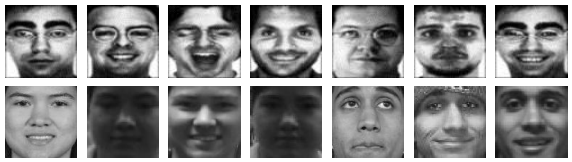


Fig. 3. Some sample face images from AR database (top row) and FRGC V1.0 database (bottom row).

Fig. 4. Some sample face images from FRGC V2.0 database(top row) and LFW database (bottom row).

samples as testing data. Some examples of FRGC V2.0 and LFW face databases are shown in Fig. 4.

For face recognition, Fig. 5, Table I and Table II present all comparison results with different kernels and algorithms. Fig. 5(a), 5(c), 5(e) and Table I show the results under different kernels (such as linear and RBF kernels with no patches, and patch-based KL kernel [10], Martin kernel [35] and our proposed LG kernel) on four face databases. As can be seen from Fig. 5(a), 5(c) and Table I, the recognition rate increases with the increasing number of training samples in general. It can be observed that our patch-based Lie group (LG) kernel algorithm outperforms the competing methods on four face databases, even on real-world face databases (FRGC V2.0 and LFW databases). The patch-based Martin kernel method performs comparably to our algorithm, while the patch-based KL kernel method performs poorly.

Moreover, compared with Martin kernel and KL kernel, our Lie group kernel algorithm leads to a more appropriate distance metric between LDMs of two face images. Thus, we can see that face recognition by exploiting the geometric properties of the Lie group manifold can improve the classification performance. Besides, the patch-based LG kernel method outperforms the other two kernel SVM methods with no patches (linear kernel and RBF kernel methods). We can then observe that face recognition with the Lie group kernel can improve the classification performance by capturing both the local appearance feature and spatial relationships of an image.

The recognition accuracy comparison of the proposed method (LG kernel) with other algorithms (namely, PCA [16], LDA [17], LBP [3], [4], Gaussian kernel LDA [36] and Polynomial kernel LDA [36]) can be seen in Fig. 5(b), 5(d), 5(f) and Table II. These results indicate that our algorithm is significantly better than other exiting methods on four face recognition databases. Moreover, the classification results are substantially better than the LBP descriptor which is a histogram of quantized LBPs pooled in a local image neighborhood. Compared with other solutions of face recognition, our proposed LG kernel employs the LG kernel by analyzing face images on the Lie group manifold and then better addresses the complex nonlinear variations of face images.

TABLE I
CLASSIFICATION ACCURACY OF DIFFERENT KERNELS ON FRGC V1.0 DATABASE.

Methods	Classification rate	
	Half:Half	4:Rest
Linear kernel + no patches	86.78%	71.97%
RBF kernel + no patches	85.75%	70.06%
KL kernel + patches	62.21%	51.43%
Martin kernel + patches	85.17%	75.47%
LG kernel + patches	**89.21%**	**79.12%**

TABLE II
CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHMS ON FRGC V1.0 DATABASE.

Methods	Classification rate	
	Half:Half	4:Rest
PCA	87.25%	56.88%
LDA	73.60%	64.42%
LBP	87.80%	71.83%
Gaussian KDA	86.63%	69.18%
Polynomial KDA	82.27%	78.48%
LG kernel + patches	**89.21%**	**79.12%**

D. Head Pose Estimation

The experiment is conducted on the Pointing'04 head pose image database [41] for the head pose estimation problem. The Pointing'04 head pose image database [41] consists of 15 sets of images. Each set contains 2 series of 93 images of the same person in different poses. The pictures in the first session are utilized as training data, and the pictures from the second session are utilized as testing data. The pose or head orientation is determined by pan and tilt angles, which vary from -90 degrees to +90 degrees. The 93 head poses include combinations of 13 pitch poses and 7 yaw poses. For this database, the image is cropped and scaled to 64×64 pixels.

To evaluate the performance of our proposed algorithm for the head pose estimation problem, we quantify each approach by mean absolute error (MAE) and classification accuracy in pitch and yaw between head poses. The MAE is defined as the average of the absolute errors between the estimated pose and the ground truth,

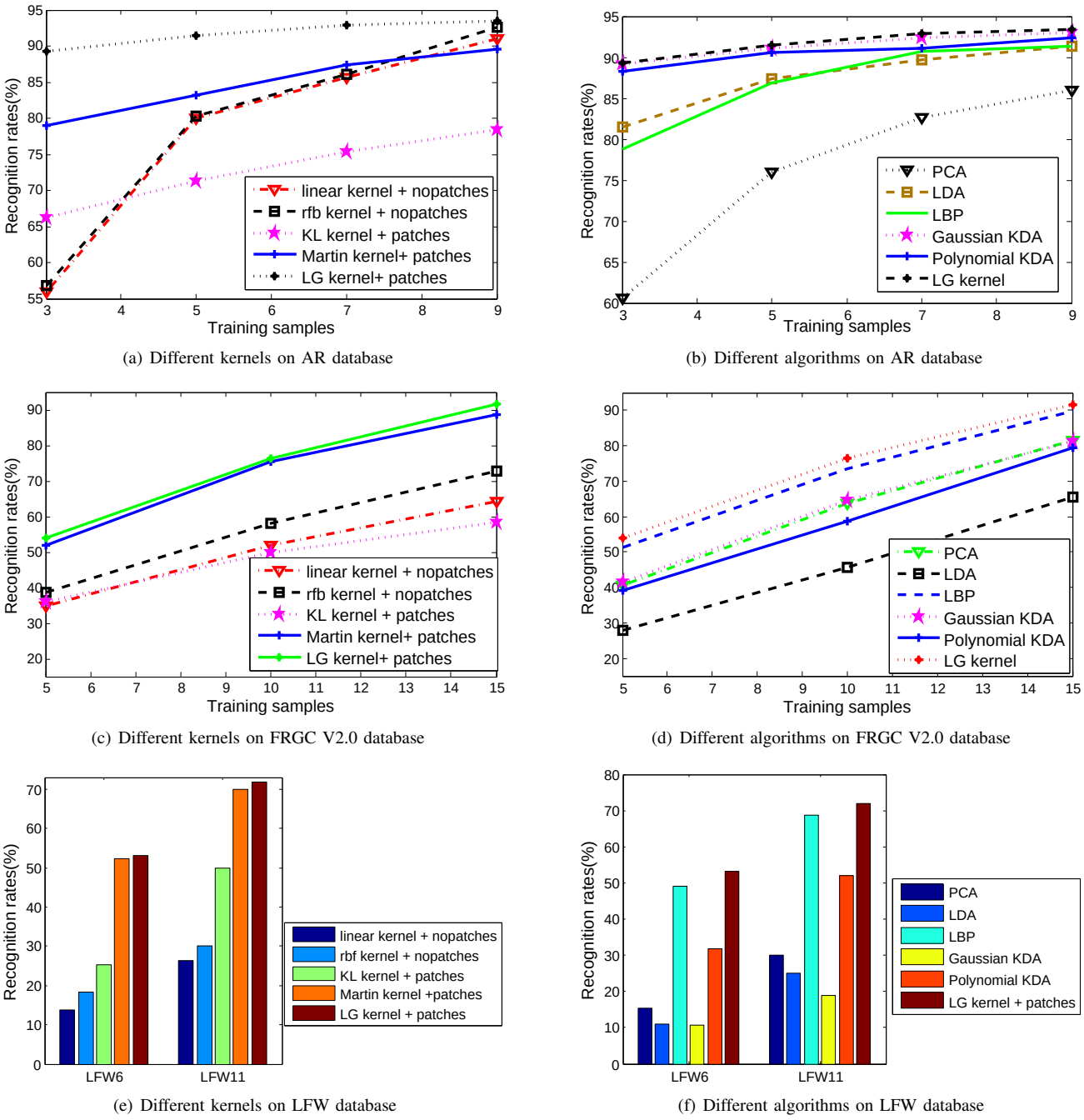
$$MAE = \frac{1}{D}\sum_{i=1}^{D}|f_i - t_i| = \frac{1}{D}\sum_{i=1}^{D} e_i, \qquad (19)$$

where t_i is the ground truth pose angle for the test image i, f_i is the corresponding pose angle of the estimated class label and D is the total number of test images. For the head pose estimation problem, MAE is computed by averaging the difference between expected pose and estimated pose for all images.

As shown in Table III, the MEA and classification accuracy of our LG kernel are the best among all kernel methods (e.g. Linear and RBF kernel SVM methods with no patches, patch-based KL kernel and Martin kernel methods) evaluated in pitch and yaw between head poses.

We compare our proposed method with some traditional head pose estimation algorithms, such as LBP, local PCA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology



(a) Different kernels on AR database

(b) Different algorithms on AR database

(c) Different kernels on FRGC V2.0 database

(d) Different algorithms on FRGC V2.0 database

(e) Different kernels on LFW database

(f) Different algorithms on LFW database

Fig. 5. Average recognition accuracy (Left) for face recognition using linear and RBF kernel SVM classifier with no patches [31], patch-based KL kernel [10], Martin kernel [35] and our LG kernel algorithms. And average recognition accuracy (Right) for face recognition using PCA [16], LDA [17], LBP [3], [4], Gaussian kernel LDA [36], Polynomial kernel LDA [36] and our LG kernel algorithms.

(LPCA) [42], [45], locality preserving projection (LPP) [42], local LDA (LLDA) [42], [45], human performance methods [41], local embed analysis (LEA) [45], high-order singular value decomposition (SVD) [45], etc. Detailed comparison results are shown in Table IV. The classification accuracy of our proposed method performs the best among all the algorithms evaluated in pitch and yaw between head poses. Moreover, the MAE of head pose estimation is substantially reduced from 9.4° (the best reported result [41]) to 5.93° for pitch head poses. For the yaw head pose, the MAE of our proposed algorithm is much better compared with the exiting methods except Stiefelhagen [44].

For the head pose estimation task, different spatial relationships of local appearances can indicate head poses. Our proposed approach obtains better performance by considering both image appearance and spatial causality among image patches with the facial linear dynamic model, and the LG kernel further captures the complex nonlinear variations of different head poses.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

TABLE III
THE MEA AND CLASSIFICATION ACCURACY OF DIFFERENT KERNELS ON POINTING'04 HEAD POSE IMAGE DATABASE, E.G. LINEAR AND RBF KERNEL SVM METHODS ARE BASED ON PIXELS, NOT PATCHES, WHILE KL KERNEL, MARTIN KERNEL AND LG KERNEL ARE BASED ON PATCHES.

Methods	MAE		Classification rate	
	Yaw	Pitch	Yaw	Pitch
linear kernel	14.23°	10.30°	51.61%	59.90%
RBF kernel	14.37°	9.24°	51.90%	60.49%
KL kernel	16.29°	11.03°	46.85%	56.21%
Martin kernel	11.57°	7.82°	57.14%	65.46 %
LG kernel	**9.78°**	**5.93°**	**62.81%**	**72.40%**

TABLE IV
THE MEA AND CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHMS ON POINTING'04 HEAD POSE IMAGE DATABASE.

Methods	MAE		Classification rate	
	Yaw	Pitch	Yaw	Pitch
Zhu$_{LLDA}$ [42]	19.1°	30.7°	-	-
Zhu$_{LLPP}$ [42]	29.2°	40.2°	-	-
Zhu$_{LPCA}$ [42]	24.5°	37.6°	-	-
Zhu$_{LPP}$ [42]	24.7°	22.6°	-	-
Zhu$_{LDA}$ [42]	25.8°	26.9°	-	-
Zhu$_{PCA}$ [42]	26.9°	35.1°	-	-
Voit [43]	12.3°	12.77°	-	-
Stiefelhagen [44]	**9.5°**	9.7°	52.0%	66.3%
Gourier [41]	10.1°	15.9°	50.0%	43.9%
Human-Per [41]	11.8°	9.4°	40.7%	59.0%
Tu$_{LEA}$ [45]	15.88°	17.44°	45.16%	50.61%
Tu$_{PCA}$ [45]	14.11°	14.98°	55.20%	57.99%
Tu$_{SVD}$ [45]	12.9°	17.97°	49.25%	54.84%
PCA	14.37°	9.24°	51.90%	60.50%
LDA	29.06°	18.75°	33.38%	41.37%
LBP	10.99°	6.78°	60.34%	70.22%
Gauss-KDA	19.71°	7.91°	50.19%	65.65%
Poly-KDA	16.60°	9.88°	48.32%	57.73%
LG kernel	9.78°	**5.93°**	**62.81%**	**72.40%**

E. Video-based face recognition

In order to further evaluate our method, we conduct an experiment on the YouTube Celebrities database [46] for the problem of video-based face recognition. As descried in section III, an image is composed of a sequence of local patches, while a face video can be described as a sequence of frames containing moving faces and showing certain stationary properties in time. Similarly, the LDM model has the potential to model the face video with a certain time-varying relationship between frames. The descriptive capability of the video LDM can be used to characterize not only the visual appearance of single frame, but also the time-varying causality among frames in a face video.

YouTube Celebrities database [46] is the largest video dataset collected for face tracking and recognition. It contains 1,910 video sequences of 47 celebrities (actors, actresses, and politicians) which are collected from YouTube. Face images are cropped from the video and these tracked faces in gray-scale format are resized to 40×40 pixels. These sequences are mostly low resolution and highly compressed. Following the same protocol as [47], we conducted five-fold cross validation experiments. In each fold, one person has 3 randomly chosen for training and six for testing.

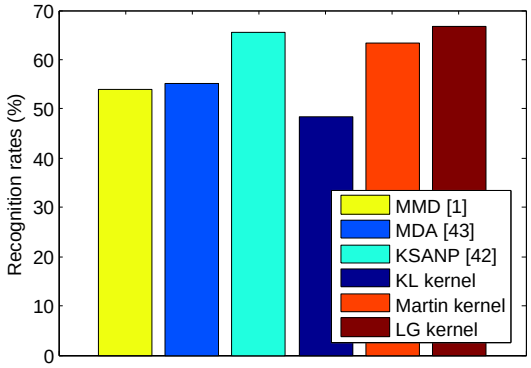


Fig. 6. Classification accuracy of different algorithms on the YouTube Celebrities database.

For the problem of video-based face recognition, we directly learn a LDM from the sequence of a face video. In our experiment, we used the cropped faces from the low quality of video frames, which makes the facial analysis more challenging. We use a local binary pattern (LBP) [3], [4] feature vector m to describe the appearance of each face frame. The dimension of the related parameters is set as $m = 2414$, $n = 8$ and $k = 40$. Besides, because the video is composed of image frames, we only compare our algorithm with Martin kernel and KL kernel for the problem of the video-based face recognition. And the average recognition accuracy of different methods are presented in Fig. 6.

Because the face videos are cropped from videos with low resolution, the recognition accuracy is lower compared to the other databases with high quality. Obviously, the LG kernel method performs better than Martin kernel and KL kernel, which shows that LG kernel also lead to a effective distance between LDMs, even with face videos in low resolution. Compared with some existing algorithms (such as Manifold Discriminant Analysis (MDA) [48], Manifold-Manifold Distance (MMD) [1], the kernel extension of the sparse approximated nearest points method (KSANP) [47]), our algorithm can achieve a better performance on the YouTube Celebrities database.

F. Algorithmic Analysis

To assess the performance of the proposed algorithm in different patch sizes, we evenly divide a face image into $n(n = 4, 9, 16, 25)$ patches. Fig. 7 depicts the effect of different patch sizes on AR database. We can see that different patch sizes can influence recognition rates, and the result of 16 $(n = 16)$ local patches is better. The recognition rate of our proposed approach is much affected by patch size, which is often set empirically. How to design a patch size robust scheme is an interesting issue for our future work.

Moreover, we give a thorough analysis of the sensitivity of our algorithm to different k values on AR database. As shown in Fig. 8, different k values have a certain impact on recognition rates, and the result of employing $k = 40$ is slightly better. From $k = 5$ to $k = 40$, the recognition rates increase with the increasing k values. However, the recognition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2365655, IEEE Transactions on Circuits and Systems for Video Technology

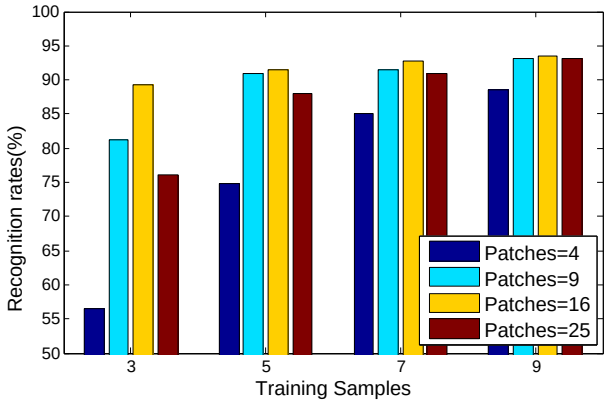


Fig. 7. Recognition rates of different (n) # of patches on AR database.

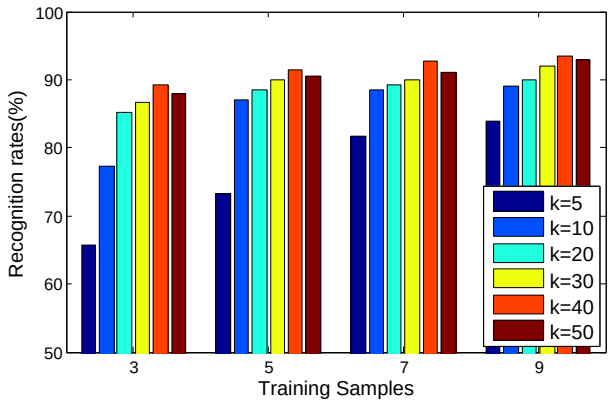


Fig. 8. Recognition rates of different k values on AR database.

rates with $k = 50$ sightly are slightly worse than the results with $k = 40$.

VII. Conclusions and Future Work

This paper presented a novel Lie group kernel for characterizing the similarity between any two face images. The linear dynamic model based face representation incorporates both image appearance and spatial information of the face. we parameterize each facial linear dynamic model as a specially-structured upper triangular matrix, the space of which can be identified as a Lie group. Then the Lie group kernel is further employed by the SVM classifier. Experiments on several face databases for face recognition and head pose estimation well demonstrated the effectiveness of the Lie group kernel for facial analysis.

There are several interesting directions for future study. The first is whether facial linear dynamic model over observation model is also useful for facial analysis, and whether it is complementary with the linear dynamic model from the state space used in this work. The second is whether this Lie group kernel is also useful for general object classification, e.g. over ImageNet database [49].

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant 61073094 and U1233119). This research was partly supported under Australian Research Council Discovery Projects funding scheme (project DP140102270).

References

[1] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao, "Manifold-manifold distance and its application to face recognition with image sets," *TIP*, vol. 21, no. 10, pp. 4466–4479, 2012. 1, 2, 4, 9

[2] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning euclidean-to-riemannian metric for point-to-set classification," June 2014. 1

[3] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *TPAMI*, vol. 28, no. 12, pp. 2037 –2041, Dec. 2006. 1, 2, 6, 7, 8, 9

[4] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition," in *ICCV*, vol. 1, Oct 2005, pp. 786–791 Vol. 1. 1, 2, 6, 7, 8, 9

[5] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor," *TIP*, vol. 19, no. 2, pp. 533–544, 2010. 1, 2

[6] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178. 1, 2

[7] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical Gaussianization for Image Classification," in *ICCV*, 2009. 1, 2

[8] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from patch-kernel," in *CVPR*, 2008, pp. 1–8. 1, 2

[9] J. Li, A. Najmi, and R. Gray, "Image classification by a two-dimensional hidden markov model," *IEEE Transactions Signal Process.*, vol. 48, no. 2, pp. 517–533, 2000. 1

[10] A. B. Chan and N. Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," in *CVPR*, 2005, pp. 846–851. 1, 3, 6, 7, 8

[11] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003. 1, 3, 4, 6

[12] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *CVPR*, vol. 1, June 2005, pp. 947–954 vol. 1. 2, 3, 6

[13] L. Gong, T. Wang, and F. Liu, "Shape of gaussians as feature descriptors," in *CVPR*, 2009, pp. 2366–2371. 1

[14] O. Tuzel, F. M. Porikli, and P. Meer, "Learning on lie groups for invariant detection and tracking," in *CVPR*, 2008, pp. 1–8. 1, 4

[15] G. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992. 2, 5

[16] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *CVPR*, 1991, pp. 586–591. 2, 6, 7, 8

[17] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *TPAMI*, vol. 19, no. 7, pp. 711–720, 1997. 2, 6, 7, 8

[18] M. Bartlett, J. R. Movellan, and T. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002. 2

[19] W. Hwang, G. Park, J. Lee, and S.-C. Kee, "Multiple face model of hybrid fourier feature for large face image set," in *CVPR*, 2006, pp. 1574–1581. 2

[20] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, no. 7, pp. 971–987, 2002. 2

[21] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von der Malsburg, "Face recognition by elastic bunch graph matching," *TPAMI*, vol. 19, no. 7, pp. 775–779, Jul 1997. 2

[22] S. Lucey and T. Chen, "A gmm parts based face representation for improved verification through relevance adaptation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, vol. 2, 2004, pp. 855–861. 2

[23] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *TPAMI*, vol. 30, no. 10, pp. 1713–1727, 2008. 3

[24] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on

stiefel and grassmann manifolds with applications in computer vision," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2008, pp. 1–8. 3

[25] G. Doretto, D. Cremers, P. Favaro, and S. Soatto, "Dynamic texture segmentation," in *Proceedings of the International Conference on Computer Vision*, vol. 2, October 2003, pp. 1236–1242. 3

[26] D. Bauer, M. Deistler, and W. Scherrer, "Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs," *Automatica*, vol. 35, no. 7, pp. 1243 – 1254, 1999. 3

[27] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed. Baltimore: Johns Hopkins University Press, 1996. 4

[28] R. Martin, "A metric for arma processes," *IEEE Transactions on Signal Processing*, vol. 48, no. 4, pp. 1164 –1170, apr 2000. 4

[29] A. B. Chan and N. Vasconcelos, "Efficient computation of the kl divergence between dynamic textures," in *Technical Report SVCL-TR-2004-02*, 2004. 4

[30] S. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. Translations of Mathematical monographs. Oxford University Press, 2000, vol. 191. 4

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011. 5, 6, 8

[32] C. Cortes and V. Vapnik, "Support-vector network," in *Machine Learning*, 1995, pp. 273–297. 5, 6

[33] M. Yang, L. Zhang, S.-K. Shiu, and D. Zhang, "Robust kernel representation with statistical local features for face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 6, pp. 900–912, 2013. 5

[34] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001. 5

[35] K. D. Cock, K. D. Cock, B. D. Moor, and B. D. Moor, "Subspace angles between linear stochastic models," in *39th IEEE Conference on Decision and Control*, 2000, pp. 1561–1566. 6, 7, 8

[36] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, 1999, pp. 41–48. 6, 7, 8

[37] A. M. Martinez and R. Benavente, "The AR Face Database," CVC, Tech. Rep., Jun. 1998. 6

[38] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *TPAMI*, vol. 31, no. 2, pp. 210–227, 2009. 6

[39] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007. 6

[40] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *ACCV*, 2010, pp. 88–97. 6

[41] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proceedings Pointing 2004, International Workshop Visual Observation of Deictic Gestures*, 2004, pp. 17–25. 7, 8, 9

[42] Z. Li, Y. Fu, J. Yuan, T. Huang, and Y. Wu, "Query driven localized linear discriminant models for head pose estimation," in *Proceedings IEEE conference on Multimedia and Expo*, 2007, pp. 1810–1813. 8, 9

[43] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *Proceedings 1st international evaluation conference Classification of events, activities and relationships*, 2007, pp. 291–298. 9

[44] R. Stiefelhagen, "Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data," *Proceedings Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, 2004. 8, 9

[45] J. Tu, Y. Fu, Y. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in *Proceedings 1st International evaluation conference Classification of events, activities and relationships*, 2007, pp. 281–290. 8, 9

[46] V. P. Minyoung Kim, Sanjiv Kumar and H. A. Rowley, "Face tracking and recognition with visual constraints in real-world videos," June 2008. 9

[47] A. S. M. Yiqun Hu and R. A. Owens, "Face recognition using sparse approximated nearest points between image sets," *TPAMI*, vol. 34, no. 10, pp. 1992–2004, 2012. 9

[48] R. Wang and X. Chen, "Manifold discriminant analysis," June 2009. 9

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255. 10

Chunyan Xu received the B.Sc. degree from Shandong Normal University, Jinan, China, in 2007 and the M.Sc. degree from Huazhong Normal University, Wuhan, China, in 2010. She is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, and also in Department of Electrical and Computer Engineering, National University of Singapore. Her research interests include computer vision, manifold learning, kernel methods, and image recognition.

Canyi Lu received the bachelor of mathematics from Fuzhou University in 2009, and the master degree in the pattern recognition and intelligent system in 2012. From August 2013, he was a phd student with the Department of Electrical and Computer Engineering at National University of Singapore. His research interests include computer vision and machine learning. His homepage is https://sites.google.com/site/canyilu.

Junbin Gao graduated from Huazhong University of Science and Technology (HUST), China in 1982 with BSc. degree in Computational Mathematics and obtained PhD from Dalian University of Technology, China in 1991. He is a Professor in Computing Science in the School of Computing and Mathematics at Charles Sturt University, Australia. He was a senior lecturer, a lecturer in Computer Science from 2001 to 2005 at University of New England, Australia. From 1982 to 2001 he was an associate lecturer, lecturer, associate professor and professor in Department of Mathematics at HUST. His main research interests include machine learning, data mining, Bayesian learning and inference, and image analysis.

Tianjiang Wang received the B. Sc. degree in computational mathematics in 1982 and the PhD degree in computer science in 1999 from Huazhong University of Science and Technology (HUST), Wuhan, China. He is currently a Professor with the School of Computer Science, Huazhong University of Science and Technology, Wuhan, China. He has finished some related projects and is the author of more than 20 related papers. His research interests include machine learning, computer vision, and data mining.

Shuicheng Yan is currently an Associate Professor at the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). Dr. Yan's research areas include machine learning, computer vision and multimedia, and he has authored/co-authored hundreds of technical papers over a wide range of research topics, with Google Scholar citation ¿14,000 times and H-index 51. He is ISI Highly-cited Researcher, 2014 and IAPR Fellow 2014. He has been serving as an associate editor of IEEE TKDE, TCSVT and ACM Transactions on Intelligent Systems and Technology (ACM TIST). He received the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM12 (Best Demo), PCM'11, ACM MM10, ICME10 and ICIMCS'09, the runner-up prize of ILSVRC'13, the winner prize of ILSVRC14 detection task, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honourable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.