An Ordered-Patch-Based Image Classification Approach on the Image Grassmannian Manifold

Chunyan Xu, Tianjiang Wang, Junbin Gao, Shougang Cao, Wenbing Tao, and Fang Liu

***Abstract*— This paper presents an ordered-patch-based image classification framework integrating the image Grassmannian manifold to address handwritten digit recognition, face recognition, and scene recognition problems. Typical image classification methods explore image appearances without considering the spatial causality among distinctive domains in an image. To address the issue, we introduce an ordered-patch-based image representation and use the autoregressive moving average (ARMA) model to characterize the representation. First, each image is encoded as a sequence of ordered patches, integrating both the local appearance information and spatial relationships of the image. Second, the sequence of these ordered patches is described by an ARMA model, which can be further identified as a point on the image Grassmannian manifold. Then, image classification can be conducted on such a manifold under this manifold representation. Furthermore, an appropriate Grassmannian kernel for support vector machine classification is developed based on a distance metric of the image Grassmannian manifold. Finally, the experiments are conducted on several image data sets to demonstrate that the proposed algorithm outperforms other existing image classification methods.**

***Index Terms*— Autoregressive moving average (ARMA) model, Grassmannian manifold, image classification, image ordered patch.**

I. Introduction

IMAGE classification is one of the most important problems for computer vision and machine learning. In addition, this covers a wide variety of application areas such as handwritten digit recognition [1]–[3], face recognition [4]–[6], scene recognition [7], [8], and even human–computer interaction [9], [10]. Most of the previous approaches for image classification are based on global image features [11], and hence are sensitive to changes in environmental conditions. In contrast to global image features, a patch-based image representation has the potential to overcome these problems [12], [13].

(a) (b)

Fig. 1. Illustration of two image representations. (a) Entire image, where each image is encoded without splitting it into patches. (b) Ordered patches. Each image is encoded as a sequence of ordered patches. Example image is from the Yale database.

Local appearance benefits image classification when the object contains regions of distinctive details. For example, the human face consists of distinctive local areas such as eyes, mouth, and nose. Unlike the free patch [12] and the coordinate patch [13], in this paper, we introduce a novel image representation called the ordered patch, which better captures the local appearance feature and spatial relationships of an image, without explicitly employing any coordinate information.

As shown in Fig. 1, to cover a 2-D image space with a 1-D sequence, we evenly divide an image into $n(n = l^2, l = 2, 3, 4$ or 5...) 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 filling curve algorithm employed in Joint Photographic Experts Group [14], captures both the horizontal and vertical spatial relationships among image patches. Z-shape scanning methods may have more widespread (strong) generalization ability than other ordering methods (such as row raster, row prime, and Hilbert order).

Ordered patches integrate both the local appearance and spatial relationships of an image, and modeling the sequence of ordered patches is equivalent to modeling these two modalities of information. Classical probability-density-based methods such as hidden Markov model (HMM) [15], Gaussian distribution, and hierarchical Gaussian distribution [13], have been used to model image appearances, whereas a few have modeled the spatial causality among image patches.

Manuscript received May 22, 2012; revised May 3, 2013 and August 27, 2013; accepted August 27, 2013. Date of publication October 10, 2013; date of current version March 10, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61073094, Grant 61073093, Grant 61371140, and Grant U1233119, and in part by Grant 9140A01060111JW0505 and Grant 51301030401. The work of J. Gao was supported by the Australian Research Council under Grant DP130100364.

C. Xu, T. Wang, S. Cao, and F. Liu are with the Intelligent and Distributed Computing Laboratory, the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China (e-mail: xuchunyan01@gmail.com; tjwang@hust.edu.cn; shougang1987@gmail.com; fang.liu@hust.edu.cn).

J. Gao is with the School of Computing and Mathematics, Charles Sturt University, Bathurst 2795, Australia (e-mail: jbgao@csu.edu.au).

W. Tao is with the Institute for Pattern Recognition and Artificial Intelligence and the National Key Laboratory of Science and Technology on Multispectral Information Processing, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: wenbingtao@gmail.com).

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/TNNLS.2013.2280752

Motivated by the recent progress in dynamic texture research [16], we model the sequence of ordered patches from each image as an autoregressive moving average (ARMA) model. Thus, the descriptive capability of the ARMA model can be used to characterize not only image appearance (e.g., color, shape, and texture), but also the spatial causality among patches in an image.

Many previous researches regard image classification, or more general object classification, as classification problems over certain manifolds [17]–[20] by embedding images/objects onto an appropriate manifold such as Riemannian. The classification algorithm is designed based on an appropriate distance metric determined by the geometric properties of the manifold with the utilization of conventional statistical methods (such as probability density estimation or discriminant analysis). In many cases, certain constraints could be specified on the parameters of the ARMA model, which may result in a nonlinear topological space. If we simply regard the model parameters as ordinary vectors from a linear space, the topological structure of the original data would be ignored or lost in this transformation. Thus, the image Grassmannian manifold is introduced to analyze the model parameters, and each image can be identified as a point on such manifold. With this representation, classification can be conducted on the image Grassmannian manifold.

In this paper, we propose an ordered-patch-based image classification framework for addressing handwritten digit recognition, face recognition, and scene recognition problems. The whole framework consists of four steps: 1) ordered-patch-based image representation; 2) building ARMA models; 3) forming the image Grassmannian manifold; and 4) constructing a Grassmannian kernel. First, each image can be represented as a sequence of ordered patches, integrating both the local appearance and spatial relationships of the image. We then propose a novel approach to construct an ARMA model from ordered patches. This model can also be identified as a point on the image Grassmannian manifold. Then, an appropriate Grassmannian kernel suitable for support vector machine (SVM) classification is developed based on a distance metric of the image Grassmannian manifold. Finally, we conduct the experiments to compare our proposed approach with some existing methods on handwritten digit recognition, face recognition, and scene recognition problems.

The rest of this paper is organized as follows. In Section II, we will review some related works. Section III presents a local descriptor for image representation called ordered patch. Section IV explains how to construct an ARMA model from these ordered patches. Section V briefly introduces the image Grassmannian manifold and then drives a distance metric. A Grassmannian kernel for SVM classification is presented in Section VI, followed by the experimental results in Section VII. Finally, Section VIII summarizes the main findings and recommends possible future directions.

II. Related Work

There have been many works proposed over the past 10 years to deal with the limitations of global features in image representation. Lucey *et al.* [12] represented an image as a set of free patches, aiming to better employ local image features for overcoming 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. Recently, Yan *et al.* [13] 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 $f(x_q, q)$ is 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. The algorithms in [12] and [13] are limited in addressing regression problems rather than classification problems. It is desirable to have a patch-based image classification framework for general visual classification tasks.

To use the states transition probabilities of neighboring image patches, Li *et al.* [15] proposed an algorithm to model image patches by 2-D HMMs. 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 [16]. The dynamic texture is a stochastic video model that treats the video as a sample from an ARMA model. A video is a sequence of images of moving scenes, which exhibit certain stationary properties over time. Similarly, an image can be encoded as a sequence of ordered patches, containing the constraint relationships among image patches in space. For a single image patch, it is a realization from a stationary stochastic process with spatially invariant statistics. For a sequence of ordered image patches (spatial-varying appearance), individual patches are clearly not independent realization from a stationary distribution, since it is very likely that there is a spatial coherence intrinsic in the process that needs to be captured. Therefore, the underlying assumption is that individual patches are realizations of the output of an ARMA model driven by an independent and identically distributed process.

Manifold learning has been widely exploited in pattern recognition, data analysis, and machine learning. Tuzel *et al.* [17] used the covariance matrices as feature descriptors, the space of which can be formulated as a connected Riemannian manifold. Riemannian manifold learning (RML) has been applied to many applications, such as Pedestrian detection [17], speech emotion classification [18], face recognition [19], and so on. Turaga *et al.* [21] 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. Gong *et al.* [22] addressed the problem of human emotion recognition based on the shape of Gaussian, which can be represented as a connected Riemannian manifold, namely a lie group. In this paper, we intend to use the image Grassmannian manifold to analyze the geometrical properties of the space of the ARMA model parameters.

III. Image as a Sequence of Ordered Patches

We introduce a novel approach of patch-based image representation named ordered patch in this section. Each image is encoded as a sequence of ordered patches.

For a given image X, the sequence of these ordered patches can be represented as

$$\{f(x_i)\}_{i=1,\ldots,n},\ f(x_i) \in \mathbf{R}^m \tag{1}$$

where n is the total number of image patches and $f(x_i)$ is the feature vector extracted from its corresponding image patch x_i. Each image in the Yale database has a pixel size of 64×64, which can be decomposed into 16 patches, each in size 16×16. For example, Fig. 1(b) shows a face image divided into 16 (4^2) patches with a given order.

The ordered-patch-based image representation, which contains both the local appearance and spatial relationships of an image, has many advantages for image classification. On the one hand, scene information is reflected by the local appearance. For example, a scene can be recognized only through local information, e.g., highway roads in a highway scene and sea in a coast scene. The human face contains different local regions, such as the nose, mouth, and eyes. On the other hand, the spatial constraint among image patches, which can be represented by the order of image patches, is relatively fixed and plays a decisive role in image classification. For the handwritten digit recognition task, different spatial relationships among image patches may produce distinctive digits. Therefore, the joint consideration of local appearance and spatial constraints among patches can be better captured by the ordered-patch-based image representation.

IV. Ordered Patch Distribution Model

For each image, the sequence of ordered patches integrates the local appearance and spatial relationships of an image. Depending on these two modalities of information, we model the sequence of ordered patches for an image. Doretto *et al.* [16] proposed to treat the video clip as a sample from an ARMA process. Although it is simple, it has been shown that the ARMA model is surprisingly useful in domains such as video synthesis, video recognition, and video segmentation. Inspired by the recent progresses in dynamic texture research [16], we propose to learn an ARMA model from the ordered patches in this section.

A. ARMA Model from Ordered Patches

For a given sequence of ordered patches, individual patches are clearly not an independent realization from a stationary distribution, since it is very likely that there is a spatial coherence structure in the image that needs to be captured. Therefore, the underlying assumption is that ordered patches are a realization of the output of an ARMA model by an independent and identically distributed process. For a given image X, we construct an ARMA model from a sequence of ordered patches $\{f(x_i)\}_{i=1,\ldots,n}$. The appearance of ordered patches $f(x_i) \in R^m$ is a linear function of the current state vector with some observation noises, and the spatial causality relationship among image patches is represented as a state process $z(x_i) \in R^n$. Therefore, each image is associated with an ARMA model determined by an unknown process

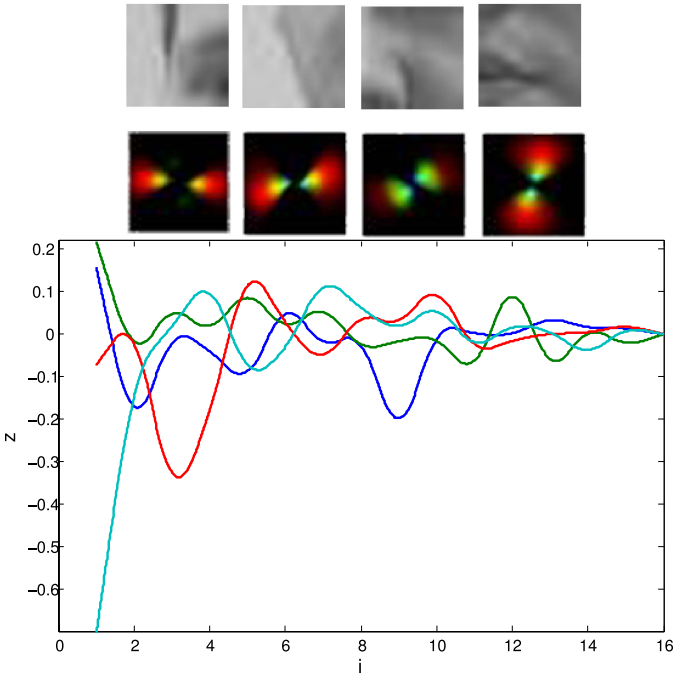


Fig. 2. Illustration of the ARMA model. Top: image patch examples from an ordered-patch sequence. Middle: appearances of the corresponding ordered-patches. Bottom: state space trajectory of the corresponding coefficients.

$$\begin{cases} f(x_i) = Cz(x_i) + \varphi(i) \\ z(x_{i+1}) = Tz(x_i) + \psi(i) \end{cases} \tag{2}$$

where $C \in R^{m\times n}$ is the observation matrix, $T \in R^{n\times n}$ is the transition matrix, and $z(x_0) \in R^n$ is the initial state vector. The state and observation noises are given by $\varphi(i) \sim N(0, P)$ and $\psi(i) \sim N(0, Q)$, respectively.

A sequence of $\{f(x_i)\}_{i=1,\ldots,n}$ encodes the appearance component of the image (ordered patches), and the spatial causality component is encoded into the state sequence $\{z(x_i)\}_{i=1,\ldots,n}$. The hidden state is modeled as a first-order Gauss–Markov process, where the state at the image patch $i+1$, $z(x_{i+1})$ is determined by the transition matrix T, the state at the image patch i, $z(x_i)$, and the driving process $\psi(i)$. Fig. 2 shows an example of a patch sequence, the corresponding appearance, and its state space coefficients.

B. Parameter Estimation

In general, the parameters of an ARMA model can be learned by maximum likelihood (ML), e.g., numerical algorithms for subspace state space system identification [23]. Due to the high dimensionality of feature vectors, these algorithms are unfeasible for learning ARMA models from ordered patches. Therefore, we intend to use the closed-form solution [16] for the model parameters in this paper. For a given image X, $F_1^n = [f(x_1), \ldots, f(x_n)] \in R^{m\times n}$ is the matrix of ordered patches, and $F_1^n = U\Sigma V^T$ is the singular

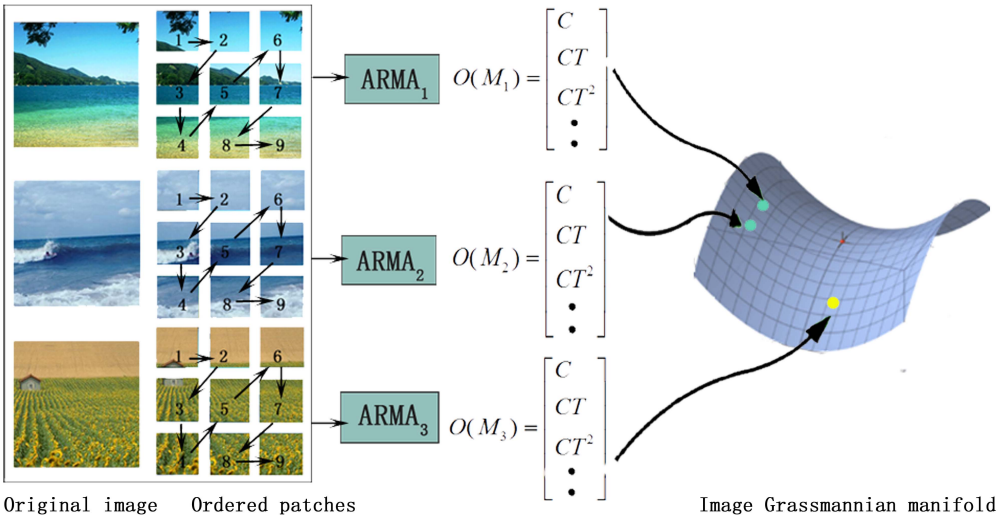


Fig. 3. Image classification framework on the image Grassmannian manifold. The example images are from the Massachusetts Institute of Technology (MIT) scene database [7].

value decomposition (SVD). Then, we define

$$\begin{aligned} \hat{C} &= U \\ \hat{Z}_1^n &= \Sigma V^T \\ \hat{T} &= \hat{Z}_2^n (\hat{Z}_1^{n-1})^{\dagger} \\ \hat{Q} &= \frac{1}{n-1} \sum_{i=1}^{n-1} \hat{v}_i \hat{v}_i^T \end{aligned} \tag{3}$$

where $\hat{Z}_1^n = [\hat{z}(x_1), \ldots, \hat{z}(x_n)]$, $\hat{Z}_1^{n-1} = [\hat{z}(x_1), \ldots, \hat{z}(x_{n-1})]$, and $\hat{Z}_2^n = [\hat{z}(x_2), \ldots, \hat{z}(x_n)]$ are the matrices of hidden state estimates, $(\hat{Z}_1^{n-1})^{\dagger}$ is the pseudoinverse of $\hat{Z}_1^{n-1}$, and $\hat{v}_i = \hat{z}(x_i) - \hat{T}\hat{z}(x_{i-1})$.

The computational complexity of the above process is $O(m^2 n + n^3)$, where m is the dimension of observation feature vector and n is the dimension of ARMA model parameters. The m^2 term dominates, since in our application $m \gg n$. For the pairwise distance between two ARMA models, several attempts have been made to endow the space of ARMA models with a metric and probabilistic structure, such as Martin distance [24], Kullback–Liebler divergence [25], and so on. The model parameters (C, T) 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 such a manifold in the following section.

V. Image Grassmannian Manifold

First, we note that the set of the estimated parameters of ARMA models defines the so-called image Grassmannian manifold. Given the manifold, an appropriate distance metric on it can be derived.

A. Image Grassmannian Manifold

The geometric properties of a manifold lead to an appropriate distance metric. In many previous works, one considers classification/recognition problems on certain manifolds based on appropriate geometric structures [21]. Amari and Nagaoka [26] 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 dimension n is a Hausdorff topological space, which has a countable base of open sets and is locally Euclidean of dimension n. Riemannian manifolds are endowed with a distance measure that allows us to measure how similar two points are. Considering those general manifolds is beyond the scope of this paper, instead we are only interested in a particular class of Riemannian manifolds called Grassmannian manifold.

As already discussed in Section IV, each image can be modeled as an ARMA model from ordered patches. The model parameters (C, T) learned as above do not lie on a linear space. To ensure that the process converges, the transition matrix T is scaled so that the largest eigenvalues lie on the unit circle. The observation matrix C is constrained to be an orthonormal matrix. Therefore, the matrix C lies on the Stiefel manifold. When n is the total number of ordered patches for a given image, the visual matrix for modeling an image can be commonly given by

$$O_n^T = \left[C^T, (CT)^T, (CT^2)^T, \ldots, (CT^n)^T\right]. \tag{4}$$

Thus, an ARMA model can be alternatively identified as a point on the Grassmannian manifold corresponding to the column space of the observation matrix [27]. We call it image Grassmannian manifold, because an image can be presented as a point on such manifold by learning an ARMA model.

Image Grassmannian analysis provides a natural way to tackle the problem of image classification. Classification on the image Grassmannian manifold can be shown in Fig. 3.

B. Distance Metric on the Image Grassmannian Manifold

The image Grassmannian manifold $G_{n,m}$ is endowed with a Riemannian structure, therefore the distance between two points on such a manifold can be measured by the geodesic, which is the length of shortest curves connecting them.

The Grassmannian distance metric defined in [27] is asymmetrical in their argument. Although it can be symmetrized using the average of the forward and backward lengths between the two points, this way doubles the complexity of computing distances. We propose a distance metric on the image Grassmannian manifold, which mitigates the complexity issue.

Suppose S is any set on the image Grassmannian manifold $G_{n,m}$. A metric on S is a distance function $d\colon S \times S \rightarrow R$, satisfying the following three properties for all $C(X_1), C(X_2), C(X_3) \in S$: 1) symmetry; 2) positivity; and 3) triangle inequality. If C is a point set on such a manifold and d is a metric on S, the pair (C, d) is called a metric space [28].

The smallest Euclidean distance between any pair of points on the image Grassmannian manifold is given by

$$\begin{aligned} &D_{\text{Grassmann}}(X_i, X_j) \\ &= \min_{M>0} \|(C(X_i) - C(X_j)M)^T (C(X_i) - C(X_j)M)\|_F \\ &= \min_{M>0} \|I - C(X_i)^T C(X_j)M \\ &\qquad - M^T C(X_j)^T C(X_i) + M^T M\|_F \end{aligned} \tag{5}$$

where $C(X_i)$, $C(X_j) \in R^{m\times n}$ are the two points on such manifold, $C(X_i)^T C(X_i) = I$, and $C(X_j)^T C(X_j) = I$. $\|\,.\,\|_F$ denotes the Frobenius matrix norm, which for an matrix A with elements A_{ij} is given by

$$\| A \|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | A_{ij} |^2 \right)^{1/2} = \left[\text{tr}(A^H A)\right]^{1/2} \tag{6}$$

where A^H is the conjugate transpose of a matrix A. Equation (5) defines a constrained optimization problem with the condition that M is positive definite. The optimization problem can be solved using the method of Lagrange multipliers, referring to [29]. If the matrix M varies over the spaces $R^{n,n}$ of all $n \times n$ matrices, the minimum of Grassmannian distance is attained at $M = C(X_j)^T C(X_i)$. Then, the Grassmannian distance can be given by

$$D_{\text{Grassmann}}(X_i, X_j) = \| I - M^T M \|_F. \tag{7}$$

To compare two ARMA models, we can compute their distance on the image Grassmannian manifold. We will use this distance metric of Grassmannian manifold in all our experiments, which has been proven to be efficient in Section VII.

VI. Grassmannian Kernel for Classification

We propose to learn a SVM classifier based on the distance metric on the image Grassmannian manifold. The SVM [30] 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 similarity of two objects/points on some space. According to [31], any symmetric positive semidefine 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 image Grassmannian manifold, so it is quite natural to define an efficient kernel over the manifold.

Our Grassmannian kernel is defined as follows:

$$K_{\text{Grassmann}}(X_i, X_j) = \exp(-\gamma\, D^2_{\text{Grassmann}}(X_i, X_j)) \tag{8}$$

where $D_{\text{Grassmann}}(X_i, X_j)$ is a kind of distance metric between two points on the image Grassmannian manifold $G_{n,m}$, and the parameter γ is directly related to scaling.

It is easy to prove that the newly defined Grassmannian kernel is a valid Mercer's kernel. First, because

$$D_{\text{Grassmann}}(X_i, X_j) = D_{\text{Grassmann}}(X_j, X_i) \tag{9}$$

thus

$$K_{\text{Grassmann}}(X_i, X_j) = K_{\text{Grassmann}}(X_j, X_i) \tag{10}$$

which means that the Grassmannian kernel is symmetric. $K_{\text{Grassmann}}$ is said to be nonnegative definite if

$$\sum_{i=1}^{D} \sum_{j=1}^{D} K_{\text{Grassmann}}(X_i, X_j) c_i c_j \geq 0 \tag{11}$$

for all finite sequences of points $X_1, \ldots, X_D$ on the image Grassmannian manifold and all choices of real numbers $c_1, ..., c_D$, and D is the number of training samples. This can be proved using the same techniques in [32] and [33]. In addition, computing the Grassmannian kernel demands for $O(D(D-1)/2n^2m)$, where m is the dimension of observation feature vector and n is the dimension of ARMA model parameters.

VII. Experiments

In this section, we systematically evaluate the effectiveness of our proposed method for handwritten digit recognition, face recognition, and scene recognition problems. Then, we give a thorough analysis of the sensitivity of our algorithm to different ordering methods and patch size. In addition, the effectiveness of the Grassmannian kernel is also discussed.

A. Experimental Setups

Each image can be encoded as a sequence of ordered patches, as already discussed in Section III. For a given image, the optimal number n of ordered patches is estimated according to the experiments. In this paper, the number n of ordered patches is four (that is 2×2) for handwritten digit recognition, 16 (that is 4×4) face recognition, and nine (that is 3×3) for scene recognition. 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 and the number n of ordered patches is 16, then the patch size will be 16×16 pixels.

For each patch of an image, we use a feature vector to describe it. The gist descriptor describes the image as a vector without detecting any interest points, and it performs well

TABLE I
RECOGNITION RATES OF DIFFERENT METHODS ACROSS VARIOUS DATABASES

Methods	Experimental Datasets	
	MNIST	USPS
Friedman$_{BSC}$[1]	91.28%	91.90%
Wohlmayr$_{ML}$[2]	83.73%	87.10%
Wohlmayr$_{CG}$[2]	91.82%	95.23%
Wohlmayr$_{CVX}$[2]	92.04%	90.10%
Pernkopf$_{DSL-BSC}$[3]	92.28%	92.40%
Ours	**93.48%**	**95.51%**

even for the low-resolution images [7]. Therefore, the gist descriptor is employed to extract the final feature vector of each patch in our experiment. A Grassmannian kernel of SVM classification is used to classify images on the image Grassmannian manifold. A one-versus-all scheme is used to learn the multiclass problem, and the parameter γ is selected using threefold cross validation over the training set. The SVM training and testing are performed using the LIBSVM software package [30].

B. Handwritten Digit Recognition Results

This experiment is conducted on Mixed National Institute of Standards and Technology (MNIST) [34] and United States Postal Service (USPS) databases for the handwritten digit recognition problem. MNIST database [34] contains 60 000 samples for training and 10 000 digits for testing. We down-sample the gray-level images by a factor of two, which results in a resolution of 14 $\times$ 14 pixels. USPS database contains 11 000 uniformly distributed handwritten digit images from zip codes of mail envelopes. The data set is split into 8000 images for training and 3000 for testing. Each digit is represented as a 16 $\times$ 16 grayscale image.

We evaluate our proposed method against some traditional handwritten digit recognition algorithms, such as Bayesian network classifier (BSC) [1], ML [2], maximum margin [2], conjugate gradient [2], and discriminative structure learning of BSC [3] methods.

Detailed comparison results are shown in Table I. Figs. 4 and 5 show per category performance for MNIST and USPS databases, respectively. In terms of classification accuracy, our proposed method outperforms all the other handwritten digit recognition algorithms.

In detail, Table I shows the classification accuracy of the proposed method and five other algorithms tested over MNIST and USPS databases. The results show that our algorithm is significantly better than the other methods. To further analyze classification performance, we also report the per category accuracy for MNIST and USPS databases in Figs. 4 and 5, respectively. As shown, the category classification rate among the digits (such as zero, one, six, and nine) is better than that for the digits two, five, and eight, probably due to more complex structure for two, five, and eight.

For the handwritten digit recognition task, different orders of local appearance information can show different digits. Our proposed approach obtains better performance by considering both image appearance and the spatial causality among image patches.

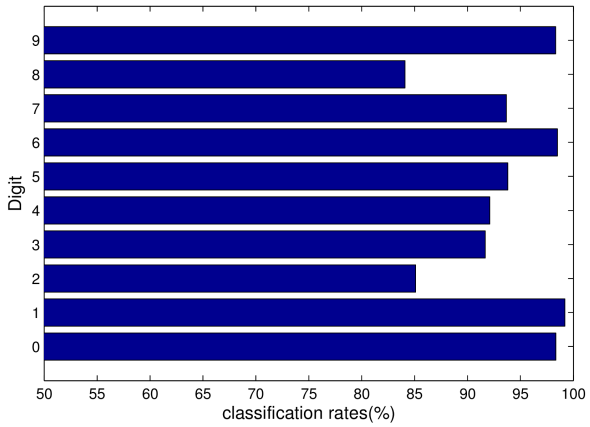


Fig. 4. Per category performance for the MNIST database.

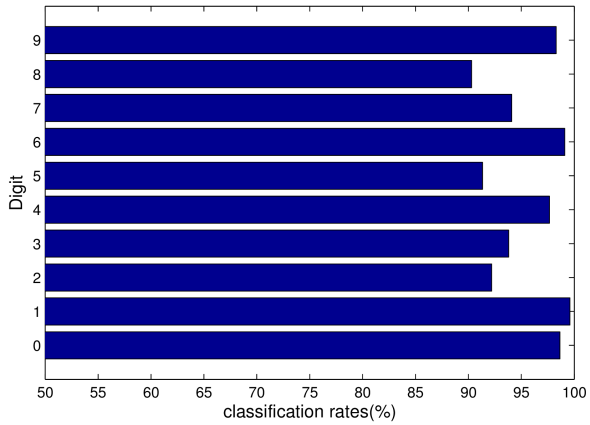


Fig. 5. Per category performance for the USPS database.

TABLE II
RECOGNITION RATES OF DIFFERENT METHODS ACROSS VARIOUS DATABASES

Methods	Experimental Datasets		
	Yale	ORL	Ext. Yale B
Wright(PCA)[36]	80%	88.1%	65.4%
Wright(LDA)[36]	87.3%	93.9%	81.3%
Wright(LPP)[36]	88.6%	88.59%	86.4%
Wright(IEM) [36]	-	96.5%	91.4%
RML [19]	85.6%	90.0%	-
Ours	**93.3%**	**96.5%**	**98.3%**

C. Face Recognition Results

For face recognition, we perform extensive experiments on three face recognition benchmarks including Yale, Olivetti-Oracle Research Lab (ORL) [28], and extended Yale B [35]. The data set splits and image scaling are identical to those described in [36] or [19].

For face recognition, Table II presents the comparison results. The first four rows show the classifier performance using the method of Principal component analysis, Linear discriminant analysis, Locality Preserving Projections and implicit elastic matching [36] respectively. The fifth row presents the results obtained using RML method [19]. As can be seen in the last row, our proposed method outperforms the other approaches on three public databases. Moreover, the classification rates are substantially better than RML.

TABLE III
CONFUSION MATRIX FOR OUR PROPOSED ALGORITHM

Truth	classified as							
	I	F	C	S	T	H	O	M
inside city	**91**	0	0	0	4	0	3	0
forest	0	**90**	0	0	0	0	5	4
coast	0	0	**87**	0	0	2	9	1
street	3	2	0	**91**	2	0	0	1
tall building	5	0	0	3	**90**	2	1	1
highway	6	1	6	1	3	**77**	4	1
open country	1	2	13	0	2	2	**76**	4
mountain	0	3	1	0	1	0	8	**88**
average accuracy: 86.2%								

TABLE IV
CONFUSION MATRIX FOR A HOLISTIC FEATURE GIST DESCRIBED IN [7]

Truth	classified as							
	I	F	C	S	T	H	O	M
inside city	**90**	0	0	3	3	1	1	0
forest	0	**91**	0	0	1	0	1	6
coast	0	0	**79**	0	0	8	12	1
street	5	1	0	**89**	1	2	1	2
tall building	9	1	0	2	**82**	0	0	5
highway	3	0	4	2	0	**87**	4	1
open country	0	3	13	2	0	5	**71**	6
mountain	0	7	2	2	1	2	5	**81**
average accuracy: 83.7%								

Thus, we can see that face recognition on the image Grassmannian manifold can improve classification performance.

D. Scene Recognition Results

First, we perform the experiment for the scene recognition problem on the publicly available MIT scene database [7]. Following the same test protocol in [7], we randomly select 100 images in each category for training and the rest of the images for testing. For the sake of conveniently constructing an ordered patch, we resize all the scenes to 255×255.

The confusion matrix of our proposed algorithm is presented in Table III. For comparison, we also show the confusion matrix given by the holistic gist descriptor proposed in [7] in Table IV. Using the holistic gist descriptor, the recognition accuracy is 83.7% on eight outdoor scene categories, which is worse than 86.2%, the accuracy given by the proposed algorithm on this database. Not surprisingly, confusion occurs between some natural scenes, which have many similarities. For example, coast and open country have the blue sky, some playing people, and so on. Our proposed method employs both the local features and spatial relationships of an image, while Oliva *et al*. [7] only extract the global features for each image and do not consider the spatial relationships of the image. The effectiveness of our proposed method is verified by the experimental results.

We also tried our algorithm on the 15-Scenes database compiled by several researchers [8], [37], [38]. This database contains 4485 images falling into 15 categories, with the number of images in each category ranging from 200 to 400. Following the same experiment setting used in [8], [37], and [38], we took 100 images per class for training and used the rest for testing.

TABLE V
CLASSIFICATION RATE (%) COMPARISON ON 15 SCENES

Algorithms	Averages accuracy
Histogram [8]	65.2
KC [37]	76.67
SPM [38]	81.4
Ours	79.3

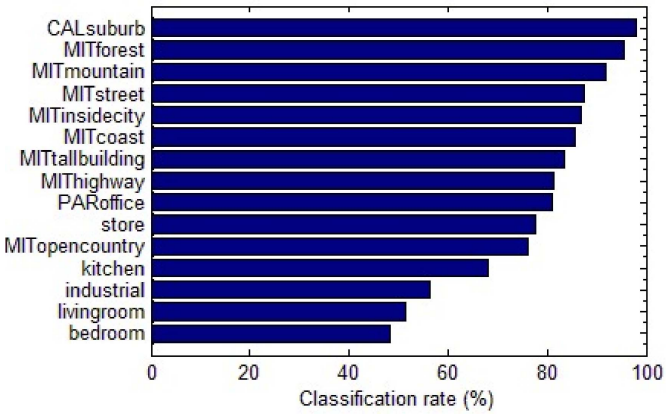


Fig. 6. Per category performance for the Scene-15 data set.

Table V shows the average recognition accuracy of our algorithm and three other methods over the 15 scene data sets. The results show that our method is significantly better than the other exiting methods (such as histogram [8] and kernel codebooks [37]) except for the spatial pyramid matching method [38]. In addition, the per category performance for the Scene-15 data set is shown in Fig. 6 from which we can see that the category performance among the outdoor scene (such as CALsuburb, MITforest, and MITmountain) is better than the indoor categories: kitchen, industrial, livingroom, and bedroom, probably due to sharing similar semantics and visual structures among the indoor scenes.

E. Algorithmic Analysis

We now give a thorough analysis of the sensitivity of our algorithm to different ordering methods, different patch sizes, and the effectiveness of the Grassmannian kernel. Then, we will discuss the relationship between its discriminative power and the parameter γ.

For the sensitivity of our algorithm to different ordering methods, we evenly divide an image into 16 (4^2) patches, and scan these patches in four different ordering methods (z-shape order [39], Hilbert order [40], row prime order, and row raster order), as shown in Fig. 7. Detailed comparison results on three face recognition databases (Yale, ORL, and extended Yale B) are presented in Table VI from which we can learn that different ordering methods have a certain impact on recognition rates. The result of z-shape scan (database) is slightly better. Scanning the patches in the z-shape captures both the horizontal and vertical spatial relationships among image patches. Hilbert order may preserve the locality, but may ignore either the horizontal or vertical spatial relationships among image patches. In addition, a row primer order can

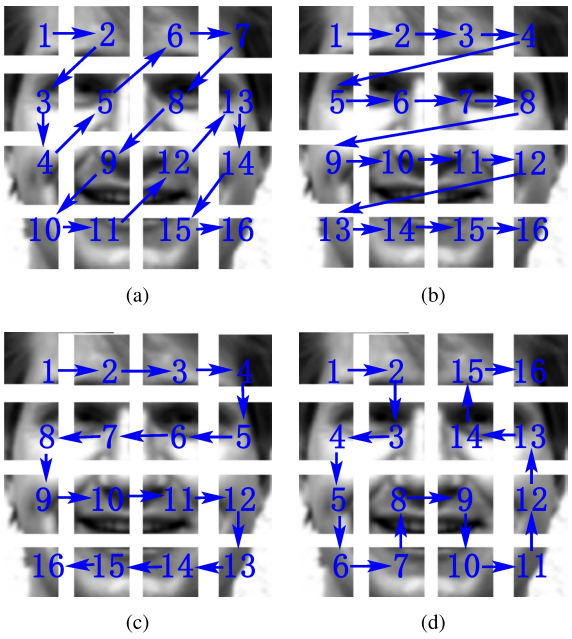


Fig. 7. Illustration of four different ordering methods. (a) Z-shape order. (b) Row (raster) order. (c) Row prime order. (d) Hilbert order.

better preserve the horizontally spatial relationships among patches, but may ignore their vertically spatial relationships. For example, as shown in Fig. 7(c), patches 1 and 8 are in vertical contiguous relationship, but the gap in the sequence of ordered patches is very large.

To assess the performance of the proposed algorithm in different patch sizes, we evenly divided an image into $n(n = l^2, l = 2, 3, 4,$ and 5) patches, and constructed an ordered sequence of image patches in the z-shape order. Table VII demonstrates the effect of different patch sizes in terms of recognition rates for three face data sets. We can see that different patch sizes can affect the recognition rates, and 16 (4^2) patches give better results.

The Grassmannian kernel performs significantly better than the Martin kernel and Kullback–Leibler (KL) kernel, as shown in Table VIII. The Martin kernel

$$K_{\text{Martin}}(X_i, X_j) = \exp(-\gamma\,(D_{\text{Martin}}(p(X_i), q(X_j)) + D_{\text{Martin}}(q(X_j)p(X_i))))$$

is defined based on Martin distance [24], which is related to the principal angles between the subspaces of the extended observability matrices of the two ARMA models ($p(X_i)$ and $q(X_j)$). The KL kernel [41] is defined as follows:

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

where $D_{\text{KL}}(p(X_i), q(X_j))$ is the KL divergence [25] between the probability distributions of the two ARMA models $p(X_i)$ and $q(X_j)$ in state spaces.

Fig. 8 shows the relationship between the discriminative power and the parameter γ on the extended Yale B database. The recognition rate is robust as long as the value of the parameter γ falls in the range of approximately from 0.001 to 0.8, however, the Grassmannian kernel with a smaller value of the parameter γ from 0.0001 to 0.001 can significantly deteriorate the recognition rates. The values γ used in this paper was selected using the cross-validation method.

TABLE VI
RECOGNITION RATES OF DIFFERENT ORDERS

Database	Row(raster)	Row Prime	Hilbert	Z-shape
Yale	91.30%	91.30%	90.70%	**93.30%**
ORL	92.50%	92.50%	92.50%	**96.50%**
Ext. YaleB	98.62%	98.56%	98.56%	**98.62%**

TABLE VII
RECOGNITION RATES OF DIFFERENT PATCH NUMBERS

Database	4 Patches	9 Patches	16 Patches	25 Patches
Yale	87.88%	91.33%	**93.30%**	89.33%
ORL	93.00%	96.00 %	**96.50%**	93.50%
Ext. YaleB	98.27%	98.60%	**98.62%**	98.50%

TABLE VIII
RECOGNITION RATES OF DIFFERENT KERNELS

Database	KL kernel	Martin kernel	Grassmannian kernel
Yale	50.00%	91.30%	**93.30%**
ORL	68.00%	96.00%	**96.50%**
Ext. YaleB	85.47%	98.56%	**98.62%**
15 Scenes	61.34%	78.52%	**79.30%**

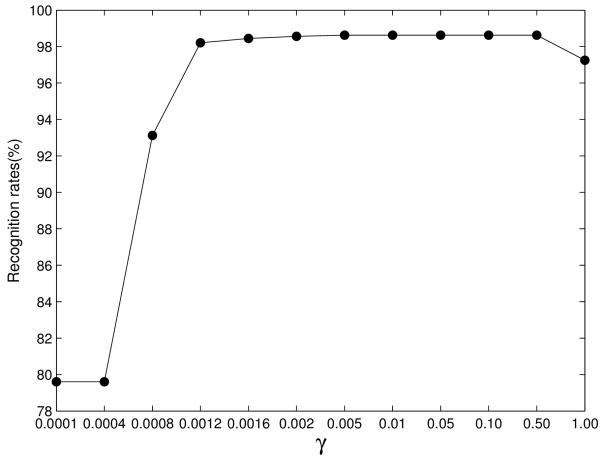


Fig. 8. Performance on the extended Yale B database for various parameter γ values of the Grassmannian kernel.

VIII. CONCLUSION

In this paper, we have studied an ordered-patch-based image representation for integrating the local appearance and spatial relationships of an image. Specifically, each image is described by an ARMA model, which can be identified as a point on Grassmannian manifold. Therefore, the image Grassmannian manifold is proposed to solve image classification problems. An appropriate Grassmannian kernel for SVM classification is also developed based on a distance metric between points on the image Grassmannian manifold. We have conducted the experiments in practical vision applications such as handwritten digit recognition, face recognition, and scene recognition.

The experimental results show that our proposed approach is more effective against the existing image classification methods.

There are several interesting issues for future work, some of which we are currently working on. For example, we aim to further develop a double kernel to improve image classification performance and investigate how to apply our proposed method for motion-blurred and noisy images.

REFERENCES

[1] D. G. N. Friedman and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, nos. 2–3, pp. 131–163, 1997.

[2] M. Wohlmayr, F. Pernkopf, and S. Tschiatschek, "Maximum margin Bayesian network classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 521–532, Mar. 2012.

[3] F. Pernkopf and J. A. Bilmes, "Efficient heuristics for discriminative structure learning of Bayesian network classifiers," *J. Mach. Learn. Res.*, vol. 11, pp. 2323–2360, Aug. 2010.

[4] D. Masip and J. Vitria, "Shared feature extraction for nearest neighbor face recognition," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 586–595, Apr. 2008.

[5] J. Yang and C. Liu, "Color image discriminant models and algorithms for face recognition," *IEEE Trans. Neural Netw.*, vol. 19, no. 12, pp. 2088–2098, Dec. 2008.

[6] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Two-stage nonnegative sparse representation for large-scale face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 35–46, Jan. 2013.

[7] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comp. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[8] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2005, pp. 524–531.

[9] M. Ursino, E. Magosso, and C. Cuppini, "Recognition of abstract objects via neural oscillators: Interaction among topological organization, associative memory and gamma band synchronization," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 316–335, Feb. 2009.

[10] P. Wang, C. Shen, N. Barnes, and H. Zheng, "Fast and robust object detection using asymmetric totally corrective boosting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 33–46, Jan. 2012.

[11] L. Gong, T. Wang, and F. Liu, "Shape of Gaussians as feature descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2366–2371.

[12] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun./Jul. 2004, pp. 855–861.

[13] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang, "Regression from patch-kernel," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[14] G. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consumer Electron.*, vol. 38, no. 1, pp. 18–34, Feb. 1992.

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

[16] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int. J. Comput. Vis.*, vol. 51, no. 2, pp. 91–109, 2003.

[17] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[18] C. Ye, J. Liu, C. Chen, M. Song, and J. Bu, "Speech emotion classification on a Riemannian manifold," in *Proc. 9th Pacific Rim Conf. Multimedia*, 2008, pp. 61–69.

[19] G. Lin and X. Mei, "Face recognition based on Riemannian manifold learning," in *Proc. IEEE ICCP*, Oct. 2011, pp. 55–59.

[20] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.

[21] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Jun. 2008, pp. 1–8.

[22] L. Gong, T. Wang, C. Wang, F. Liu, F. Zhang, and X. Yu, "Recognizing affect from non-stylized body motion using shape of Gaussian descriptors," in *Proc. ACM IEEE Conf. Symp. Appl. Comput.*, Apr. 2010, pp. 1203–1206.

[23] 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.

[24] R. Martin, "A metric for ARMA processes," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1164–1170, Apr. 2000.

[25] A. B. Chan and N. Vasconcelos, "Efficient computation of the KL divergence between dynamic textures," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 1–18.

[26] S. Amari and H. Nagaoka, *Methods of Information Geometry* (Translations of Mathematical Monographs), vol. 191. Oxford, U.K.: Oxford Univ. Press, 2000.

[27] Y. Chikuse, *Statistics on Special Manifolds*. New York, NY, USA: Springer-Verlag, Feb. 2003.

[28] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, Dec. 1994, pp. 138–142.

[29] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *Proc. 1st Int. Evaluat. Conf. Classification Events, Activit. Relationships*, 2007, pp. 291–298.

[30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[31] C. Cortes and V. Vapnik, "Support-vector network," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[32] K. Abou-Moustafa, M. Shah, F. De La Torre, and F. Ferrie, "Relaxed exponential kernels for unsupervised learning," in *Proc. 33rd Int. Conf. Pattern Recognit.*, 2011, pp. 184–195.

[33] M. Yang, L. Zhang, S.-K. Shiu, and D. Zhang, "Robust kernel representation with statistical local features for face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 900–912, Jun. 2013.

[34] Y. B. Y. LeCun, L. Bottou, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[35] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[36] J. Wright and G. Hua, "Implicit elastic matching with random projections for pose-variant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1502–1509.

[37] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 696–709.

[38] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2006, pp. 2169–2178.

[39] E. Artyomov, Y. Rivenson, G. Levi, and O. Yadid-Pecht, "Morton (Z) scan based real-time variable resolution CMOS image sensor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 947–952, Jul. 2005.

[40] D. Hilbert, "Über die stetige abbildung einer linie auf ein flächenstück," *Math. Ann.*, vol. 38, no. 3, pp. 459–460, 1891.

[41] A. B. Chan and N. Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 846–851.

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 pursuing the Ph.D. degree in the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan.

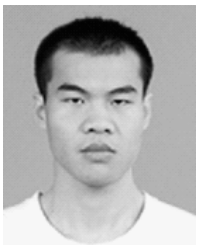
Her current research interests include computer vision, manifold learning, kernel methods, and image recognition.

Tianjiang Wang received the B.Sc. degree in computational mathematics, in 1982, and the Ph.D. degree in computer science, in 1999, from the Huazhong University of Science and Technology (HUST), Wuhan, China.

He is currently a Professor with the School of Computer Science, HUST, Wuhan, China. He has finished some related projects and is the author of more than 20 related papers. His current research interests include machine learning, computer vision, and data mining.

Junbin Gao received the B.Sc. degree in computational mathematics from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 1982 and the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 1991.

He was an Associate Lecturer, Lecturer, Associate Professor, and then Professor with the Department of Mathematics, HUST, from 1982 to 2001. From 2001 to 2005, he was a Senior Lecturer and Lecturer of computer science with the University of New England, Armidale, Australia. In July 2005, he was with the School of Information Technology, Charles Sturt University, Bathurst, Australia, as an Associate Professor of computing science. He is currently a Professor of computer science with Charles Sturt University. He is the author of over 100 papers and two books on data-based modeling, pattern recognition, and Bayesian inference. His current research interests include machine learning, kernel methods, Bayesian learning and inference, and image processing.



Shougang Cao received the B.Sc. degree in physics and the M.Sc. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2013.

His current research interests include machine learning, computer vision, and kernel methods.

Wenbing Tao received the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004.

He has been with the School of Computer Science and Technology, HUST, since 2005, where he is currently an Associate Professor. He was a Research Fellow in the Division of Mathematical Sciences, Nanyang Technological University, Nanyang, Singapore, from March 2008 to March 2009. He has published numerous papers and conference papers in the areas of image processing and object recognition. His current research interests include the areas of computer vision, image segmentation, object recognition, and tracking.

Dr. Tao serves as Reviewer for many journals, such as *International Journal of Computer Vision*, IEEE TRANSACTION ON IMAGE PROCESSING, PATTERN RECOGNITION, IMAGE VISION COMPUTING, and so on.

Fang Liu received the Ph.D. degree in computer science from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2002.

She is currently an Associate Professor with the School of Computer Science, HUST, Wuhan, China. She has finished a few related projects and is the author of more than 10 related papers. Her current research interests include machine learning, computer vision, and data mining.