Neurocomputing 74 (2011) 708–719

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# A novel robust kernel for visual learning problems

Chia-Te Liao, Shang-Hong Lai *

*Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan*

ARTICLE INFO

*Article history:*
Received 1 April 2010
Received in revised form
17 July 2010
Accepted 20 September 2010
Communicated by J. Yang
Available online 27 October 2010

*Keywords:*
Kernel-based learning
Image classification
Object recognition
Robust learning

ABSTRACT

A major challenge to appearance-based learning techniques is the robustness against data corruption and irrelevant within-class data variation. This paper presents a robust kernel for kernel-based approach to achieving better robustness on several visual learning problems. Incorporating a robust error function used in robust statistics together with a deformation invariant distance measure, the proposed kernel is shown to be insensitive to noise and robust to intra-class variations. We prove that this robust kernel satisfies the requirements for a valid kernel, so it has good properties when used with kernel-based learning machines. In the experiments, we validate the superior robustness of the proposed kernel over the state-of-the-art algorithms on several applications, including hand-written digit classification, face recognition and data visualization.

## 1. Introduction

In pattern recognition, most of the learning-based algorithms require to define a suitable similarity measure between data samples such that an objective function associated with this measure is optimized. Ideally, to make the derived model robust against irrelevant data transformations, one needs to collect a large number of training examples covering all possible data variations. However, in practice it is neither feasible to collect an enormous amount of data to cover all possible variations, nor practical to deal with the massive computational cost. Previous research suggests incorporating the prior knowledge about data variations at hand into the learning procedures so the robustness of the derived model can be enhanced. Usually one can implement this idea either by modifying the similarity measure or by generating artificial examples for learning algorithms. In the former technique, an appropriate distance measure in the pattern space can be designed such that the distance between a prototype and a pattern is not affected by deformation [1], or alternatively a properly-designed feature extractor can be used to achieve invariant data representation against the irrelevant transformations [2]. And in the later technique, *virtual examples* are generated according to a set of representative transformations for inclusion into the training dataset [3]. Hopefully, the learning machines can learn the data variations from the artificially augmented training data.

Also, robustness for a learning algorithm can be considered as the capability to handle outliers or data corruption besides the data variations. Various approaches have been proposed in the literature to handle the data corruption problem. In particular, a class of kernel-based strategies was proposed in recent years to improve the robustness of the existing learning algorithms against noise disturbance. For example, Lu et al. [4] proposed an adaptive robust kernel principal component analysis (PCA) algorithm [5] to detect and remove outliers in the kernel space. In [6] a hybrid robust kernel with a mixture of a robust error function and RBF kernels was proposed for kernel-based learning with respect to various types of noises. Barla et al. [7] presented a class of kernels by defining the similarity measure on images for some vision applications. Also, Du et al. [8] kernelized the robust fuzzy clustering algorithm for adapting arbitrary shape of clusters, where the Euclidean distance was modified to avoid the influence of noisy background data. Chen [9] integrated M-estimators into the metric of radial basis function to form a new class of robust kernels for resisting the influence of outliers. However, there was no proof for the existence of the mapped reproducing kernel Hilbert space (RKHS). In fact, one can verify that Chen's kernel does not satisfy the Cauchy–Schwarz inequality, which prevents it from being a valid kernel [10]. From the literature, we see that the behavior of these kernel-based algorithms with outliers depends on the choice of the kernel. Certain kernels, e.g. the polynomial kernel for numerical data, are known to be very sensitive to aberrant observations [11]. For the problems involving outliers, it is beneficial to use a robust kernel that is insensitive to outliers. In that case, one may say that the robustness against outliers is incorporated into the kernel design.

Due to the excellent generalization capability, support vector machine (SVM) [12] has become one of the most popular kernel machines in various fields. The generalization capability here

* Corresponding author. Tel.: +886 3 574 2958; fax: +886 3 572 3694.
*E-mail address:* lai@cs.nthu.edu.tw (S.-H. Lai).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.09.009

means the learning machine does not change significantly for data under a specific perturbation or previously unseen during the process of training. Clearly, the notion of generalization is closely related to the robustness of learning algorithms. A considerable amount of effort has been devoted on SVM to achieve better robustness from limited training data. The virtual SV method [3], for instance, is based on generating virtual examples from the current support vectors derived from the SVM training. The kernel jittering (JSV) [14] technique provides better robustness property to SVM by pre-expanding the training set into the kernel. However, because it considers all jittered forms in the kernel function, the non-symmetry of the JSV kernels can cause a problem in the training convergence. Based on a similar concept, Trafalis and Gilberta [15] investigated the training of SVM when bounded perturbation is added to the input. Rather than burdening the learning system, the Jittered Query method [16] explores a different perspective applying a set of distortions to each testing example during the query, where the responsibility for handling robustness is shifted from the training side to the query side. Different from the analytical approach, Xu et al. [17] built the linkage between the robust classification and the standard regularization scheme from the robust optimization viewpoint. The standard SVM is shown to be a special case of their robust formulation. Also, Song et al. [18] modified the formulation of SVM to solve the over-fitting problem in the presence of outliers. Debruyne et al. [11] measured the outlyingness of feature vectors and incorporated a spatial rank in the least-square SVM to obtain the robust version of RLS-SVM. Some works (e.g. [20,21]) represented examples as sets of vectors and computed similarity between the sets of vectors to achieve the invariance property.

This paper develops a robust kernel for kernel machines, such as SVM, and consequently improves their robustness in dealing with the image-related problems. The advantage of this approach is that, if a researcher proposes a particular kernel-based algorithm for some applications, one can incorporate our kernel in the proposed algorithm to obtain satisfactory classification results without worrying about its robustness. This work attempts to measure data similarities with a specially designed kernel function that is insensitive to irrelevant data transformations and noise disturbance. Generalizing the notions of robust error function and tangent distance, this work provides a novel kernel that allows us to compute similarity between data accurately and robustly for a class of kernel methods. From a theoretical point of view, we prove that the proposed kernel function satisfies the requirements for a positive definite kernel, so it can be used in a class of kernel-based learning algorithms to improve their robustness. From a practical point of view, through experiments we demonstrate that the proposed kernel when used in conjunction with various kernel methods provides superior robustness over several classical kernels on different visual learning problems, including hand-written digit classification, face recognition, and data visualization under noise disturbance.

The rest of this paper is organized as follows. In the next section, some properties of kernel functions and the related kernel methods used in the experiments are first reviewed. Subsequently, Section 3 describes the development of the proposed robust kernel and discusses its theoretic properties. In Section 4, we validate the performance of the proposed kernel on various visual learning problems. Finally, Section 5 gives the concluding remarks.

## 2. Preliminaries

In this section, we present the definition and properties concerning kernels, SVM, and kernel Fisher discriminant analysis (KFD) [22], which are needed for the development and experiments in this paper.

### 2.1. Kernel functions

The limited power of linear learning machines has been pointed out in 1960s. To make the target functions get better represented by the given attributes, one can change the data representation for the input vector $x \in \mathbb{R}^n$ with a mapping function $\Phi : \mathbb{R}^n \to F$:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \mapsto \Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_n(\mathbf{x}), \ldots), \tag{1}$$

where $F$ is the mapped space. A kernel function $k$ is employed as inner products of images under a transformation $\Phi$ of two data points $\mathbf{x}$ and $\mathbf{x'}$ in $F$:

$$k(\mathbf{x},\mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle = \sum_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}'). \tag{2}$$

A function $k$ is a kernel if and only if it satisfies the Mercer's condition [13,23] given below:

**Definition 1.** (**Positive definite kernel):** Let $X$ be the original feature space. A symmetric function $k : X \times X \to \mathbb{R}$ is a positive definite (pd) kernel if it satisfies the following condition: for any set of $m$ feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, $\mathbf{x}_i \in X$ for $i = 1,2,\ldots,m$, the corresponding Gram matrix $K$, defined by $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, is positive semi-definite; i.e. for any $c_i \in \mathbb{R}$, $i = 1,2,\ldots,m$, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j K_{ij} \geq 0. \tag{3}$$

In practice, one can simply choose a kernel satisfying the above condition without worrying about the actual form of $\Phi$. Different choices of $k$ determine the types of kernel machines that are constructed. Some standard choices include the linear kernel $k(\mathbf{x},\mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle$, the polynomial kernel $k(\mathbf{x},\mathbf{z}) = (\mathbf{x}^T\mathbf{z} + c)^d$, the sigmoid kernel $k(\mathbf{x},\mathbf{z}) = \tanh(\kappa(\mathbf{x}^T\mathbf{z}) - \theta)$, and the RBF kernel $k(\mathbf{x},\mathbf{z}) = \exp(-||\mathbf{x}\text{-}\mathbf{z}||^2/\sigma^2)$. The *kernel trick* allows the formulation of nonlinear variants of linear algorithms that can be cast in terms of dot products, and thereby the same optimization algorithm can be applied to compute a nonlinear classification function.

### 2.2. Support vector machine

Given a set of $m$ training examples $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$, where $\mathbf{x}_i \in X$ is the $i$th feature vector associated with class label $y_i \in \{1,-1\}$, SVM is trained by finding the maximal margin separating hyperplane $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, which induces the best generalization ability among all separating hyperplanes [13]. For canonical hyperplanes, the margin equals to $1/||\mathbf{w}||$. Thus, the SVM for the two-class classification problem can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{w},\xi_1,\ldots,\xi_m} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^m \xi_i \\ &\quad \text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0,\ i = 1,\ldots,m. \end{aligned} \tag{4}$$

Here, $\xi_i$ are slack variables and $C$ controls the tradeoff between loss and functional complexity. Introducing $\xi_i$, the SVM allows the separation constraints to be violated. The corresponding dual form of (4) is an equivalent quadratic programming (QP) problem:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ &\quad \text{s.t. } 0 \leq \alpha_i \leq C,\ i = 1,\ldots,m, \\ &\quad \sum_{i=1}^m y_i \alpha_i = 0, \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ is the vector of Lagrange multipliers, $\mathbf{e} \in \mathbb{R}^{m \times 1}$ is a vector of all ones, and $Q$ is an $m \times m$ matrix with entries

$Q_{ij}=y_iy_j\langle \mathbf{x}_i,\mathbf{x}_j\rangle$. In the testing phase, a vector $\mathbf{x}_{test}$ is classified by $f(\mathbf{x}_{test})=\mathrm{sgn}(\sum_{i=1}^{m}\alpha_i y_i\langle \mathbf{x}_i,\mathbf{x}_{test}\rangle+b)$. The kernel trick can be applied here because all feature vectors only occurred in dot products. Replacing the dot product $\langle \mathbf{x}_i,\mathbf{x}_j\rangle$ as the result of kernel function $k(\mathbf{x}_i,\mathbf{x}_j)$ in matrix $Q$, i.e. $Q_{ij}=y_iy_jk(\mathbf{x}_i,\mathbf{x}_j)$, SVM is capable of finding the separating hyperplane nonlinearly by computing dot products in $F$.

### 2.3. Kernel Fisher discriminant analysis

Given the dataset $S$, *Fisher's discriminant* for binary classification considers maximizing the *Rayleigh coefficient* linearly by

$$J(\mathbf{w})=\frac{\mathbf{w}^TS_B\mathbf{w}}{\mathbf{w}^TS_W\mathbf{w}}, \tag{6}$$

where $S_B=(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T$and $S_W=\sum_{c\in\{1,2\}}\sum_{\mathbf{x}\in D_c}(\mathbf{x}-\mathbf{m}_c)(\mathbf{x}-\mathbf{m}_c)^T$ are the between- and within-class scatter matrices, and $\mathbf{m}_c$ and $D_c$ denote the sample mean and the subset of data vectors of class $c$, respectively. The idea is to look for the direction $\mathbf{w}$ such that when the data vectors are projected onto it, the class centers are far apart while the spread within each class is small. The kernel trick allows solving the Fisher's discriminant nonlinearly in $F$, thereby yielding a nonlinear discriminant in the input space. Let us denote $\mathbf{a}_c=(a_1,a_2,\ldots,a_m)$, an $m$-dimensional vector with $a_i$equal to 1 if $\mathbf{x}_i$ belongs to class $c$, and 0 otherwise. Additionally, assume $\mathbf{m}_c^{\Phi}:=(1/|D_c|)K\mathbf{a}_c$, where $K$ is the Gram matrix with entries $K_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)=\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_j)\rangle$ and $|D_c|$ denotes the cardinality of $D_c$. Kernel Fisher discriminant [22] analysis thus seeks the projection that best separates data in $F$ by maximizing the criterion:

$$J(\boldsymbol{\omega})=\frac{\boldsymbol{\omega}^T\mathrm{M}\boldsymbol{\omega}}{\boldsymbol{\omega}^T\mathrm{N}\boldsymbol{\omega}}, \tag{7}$$

where $\mathrm{M}=(\mathbf{m}_1^{\Phi}-\mathbf{m}_2^{\Phi})(\mathbf{m}_1^{\Phi}-\mathbf{m}_2^{\Phi})^T$and $\mathrm{N}=KK^T-\sum_{c\in\{1,2\}}|D_c|\mathbf{m}_c^{\Phi}\mathbf{m}_c^{\Phi T}$are the $m\times m$ matrices for the between- and within-class scatters in $F$, respectively. Note that (7) can be efficiently solved by reducing it to an equivalent QP problem [23]. For a test point $\mathbf{x}_{test}$, the projection onto the discriminant $\mathbf{w}^{\Phi}$in $F$ is computed by$\langle\mathbf{w}^{\Phi},\Phi(\mathbf{x}_{test})\rangle=\sum_{i=1}^{m}\omega_ik(\mathbf{x}_i,\mathbf{x}_{test})$ for the use of subsequent procedures.

## 3. The proposed robust kernel

A kernel can be defined with a function $f$ on $X$ using a metric $d$ with the associated Gram matrix $K_{ij}:=f(d(\mathbf{x}_i,\mathbf{x}_j))$. For example, the RBF kernel, with the Gram matrix given by $K_{ij}:=\exp(-||\mathbf{x}_i-\mathbf{x}_j||^2/\sigma^2)$, has the associated metric as the *Euclidean distance*. However, the main drawback of using a square error measure $\rho$ is the lack of robustness in the presence of outliers, because a quadratic error function tends to assign an outlier a very high contribution (cf. Fig. 1(a)) to the kernel function value. This phenomenon is justified by the *influence $\psi$-function* [24], where the influence of data points increases with the size of errors without a bound. Consequently, the computed kernel function value may be dominated by few outliers that occur with low probability but contribute in a large magnitude. To make kernel approaches practically useful, we would like to tailor the kernel to meet the robust requirement such that the associated classification results are robust against irrelevant data transformations and noises/outliers. In robust statistics, researchers (see e.g. [25,26]) attempted to reduce the effect of outliers replacing the square residuals with a robust $\rho$-function of the residuals. The function $\rho(x;\sigma)$ generally looks like a square function for $x$ within a reasonably small range, and then saturates to a flat value as the residual (i.e. $x$) becomes large. The parameter $\sigma$ controls the point where the function becomes flat. The curves of the square error function [45] versus a Geman–McClure [27] $\rho$-function are shown in Fig. 1, where the Geman–McClure error function goes saturated and the influence function approaches to zero when the residual error becomes too large. The use of a robust $\rho$-function thus alleviates the unwanted influence of outliers, and it motivated us to develop a more robust kernel.

On the other hand, the ability to robustly measure similarity invariant to the irrelevant data transformations is another critical issue in the context of pattern recognition problems. Mathematically, when images of a class are subject to some transformations,

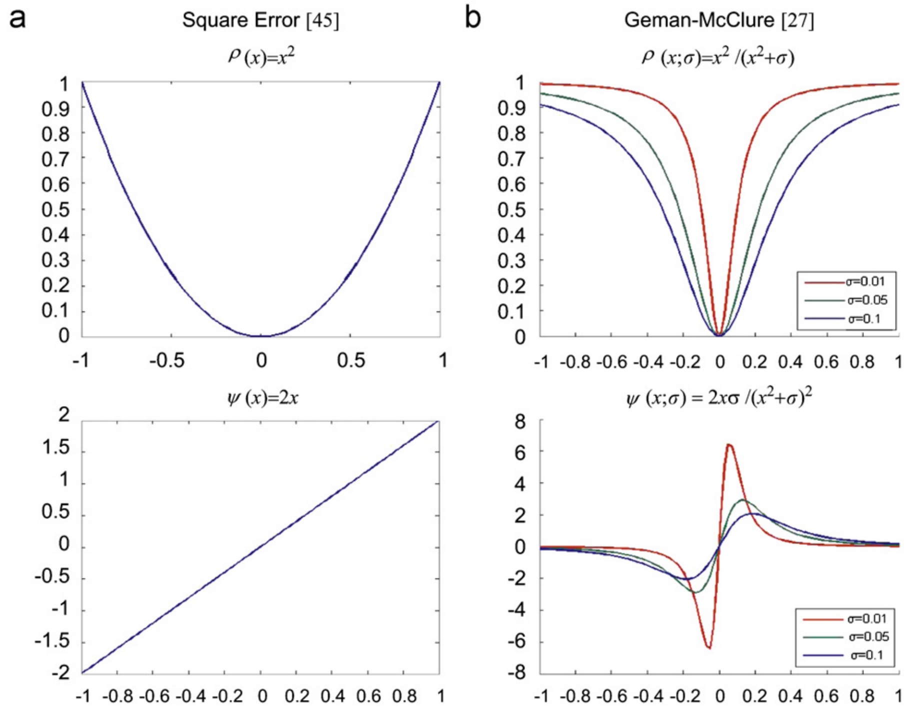


**Fig. 1.** The error functions and the corresponding influence-functions of (a) the square error function and (b) the Geman−–McClure error function with various scale $\sigma$. The influence of a datum on Geman–McClure function decreases with the size of the error, while the influence of the square error function increases linearly as shown in (a).

they span a manifold in a high-dimensional space of pixel intensities. When we compute the distance between two images, the transformations irrelevant to the expected outputs should be ignored. Hence the proper measure would be one that measures the distance between the manifolds resulting from all possible transformations of the images, rather than the Euclidean distance between the images themselves. Tangent distance [28] formalized this idea using the first-order Taylor series expansion to parameterize the image intensity function thus generalizing the metric. Precisely, assume we have the prior understanding of $r$ types of transformations known to be irrelevant to the output results. The prior knowledge about local invariance thus can be formalized as a differentiable manifold $M_{\mathbf{x}} := \{t(\mathbf{x},\mathbf{p}) | \mathbf{p} \in \mathbb{R}^r\} \subset \mathbb{R}^n$, where $t(\mathbf{x},\mathbf{p})$ is a nonlinear transformation function of image intensity that maps $\mathbf{x} \in \mathbb{R}^n$ to another point on $M_{\mathbf{x}}$ parameterized by $\mathbf{p} = (p_1,\ldots,p_r)^T \in \mathbb{R}^r$. The transformation $t(\mathbf{x},\mathbf{p})$ is assumed to be differentiable with respect to $\mathbf{p}$, and the first-order Taylor series expansion of $t(\mathbf{x},\mathbf{p})$ for an image $\mathbf{x}$ can be expressed as

$$t(\mathbf{x},\mathbf{p}) \approx t(\mathbf{x},\mathbf{0}) + \sum\nolimits_{i=1}^{r} p_i \frac{\partial t(\mathbf{x},\mathbf{0})}{\partial p_i} = \mathbf{x} + \sum\nolimits_{i=1}^{r} p_i \mathbf{t}_i, i = 1,2,\ldots,r, \tag{8}$$

where $\mathbf{t}_i$ denotes the partial differentiation of $\mathbf{x}$ with respect to the $i$th type of transformation parameterized by $p_i$. The tangent plane can be expressed by $H_{\mathbf{x}} := \{\mathbf{x} + \sum_{i=1}^{r} p_i \mathbf{t}_i | p_i \in \mathbb{R}\}$ to approximate $M_{\mathbf{x}}$ at the pattern $\mathbf{x}$. Let $\mathbf{T} = [\mathbf{t}_1\mathbf{t}_2 \cdots \mathbf{t}_r]$be an $n \times r$ matrix containing the $r$ tangent vectors for $\mathbf{x}$. For any two patterns, say $\mathbf{x}$ and $\mathbf{y}$, the tangent distance computes the shortest distance between the two tangent planes corresponding to $\mathbf{x}$ and $\mathbf{y}$ as follows [28]:

$$d_T(\mathbf{x},\mathbf{y}) = \min_{\hat{\mathbf{p}}_{\mathbf{x}},\hat{\mathbf{p}}_{\mathbf{y}}} \|\mathbf{x} + \mathbf{T}_{\mathbf{x}}\hat{\mathbf{p}}_{\mathbf{x}} - \mathbf{y} - \mathbf{T}_{\mathbf{y}}\hat{\mathbf{p}}_{\mathbf{y}}\|, \tag{9}$$

where the closed form solution is given by [29]:

$$({\mathbf{T}_{\mathbf{x}}}^T - {\mathbf{T}_{\mathbf{x}}}^T\mathbf{T}_{\mathbf{y}}({\mathbf{T}_{\mathbf{y}}}^T\mathbf{T}_{\mathbf{y}})^{-1}{\mathbf{T}_{\mathbf{y}}}^T)(\mathbf{x} - \mathbf{y}) = ({\mathbf{T}_{\mathbf{x}}}^T\mathbf{T}_{\mathbf{y}}({\mathbf{T}_{\mathbf{y}}}^T\mathbf{T}_{\mathbf{y}})^{-1}{\mathbf{T}_{\mathbf{y}}}^T\mathbf{T}_{\mathbf{x}} - {\mathbf{T}_{\mathbf{x}}}^T\mathbf{T}_{\mathbf{x}})\mathbf{p}_{\mathbf{x}}, \tag{10}$$

$$({\mathbf{T}_{\mathbf{y}}}^T - {\mathbf{T}_{\mathbf{y}}}^T\mathbf{T}_{\mathbf{x}}({\mathbf{T}_{\mathbf{x}}}^T\mathbf{T}_{\mathbf{x}})^{-1}{\mathbf{T}_{\mathbf{x}}}^T)(\mathbf{x} - \mathbf{y}) = ({\mathbf{T}_{\mathbf{y}}}^T\mathbf{T}_{\mathbf{y}} - {\mathbf{T}_{\mathbf{y}}}^T\mathbf{T}_{\mathbf{x}}({\mathbf{T}_{\mathbf{x}}}^T\mathbf{T}_{\mathbf{x}})^{-1}{\mathbf{T}_{\mathbf{x}}}^T\mathbf{T}_{\mathbf{y}})\mathbf{p}_{\mathbf{y}}. \tag{11}$$

This shortest distance can be regarded as the distance after aligning two patterns with respect to the transformation parameter vectors $\hat{\mathbf{p}}_{\mathbf{x}}, \hat{\mathbf{p}}_{\mathbf{y}}$, thus reducing the effects due to the modeled variations. Thus the problem is reduced from computing a robust similarity measure between two patterns to finding the shortest distance between two tangent planes. We give an illustrative diagram in Fig. 2 and an example image in Fig. 3 to depict using the first-order Taylor series expansion for $t(\mathbf{x},\mathbf{p})$ with respect to different parameters.

The tangent distance has, however, been shown to exhibit an important limitation of a lack of robustness to the presence of outliers. A novel robust kernel, which can be used with kernel methods, is developed by leveraging on the tangent distance and the robust $\rho$-function. Embedding the two-sided tangent distance into the residual error of the $\rho$-function, the notions of robust error function and tangent distance representation are integrated to define the proposed robust kernel function:

$$k(\mathbf{x},\mathbf{y}) = 1 - \rho(\mathbf{r}_T(\mathbf{x},\mathbf{y}),\sigma), \tag{12}$$

where

$$\mathbf{r}_T(\mathbf{x},\mathbf{y}) = \mathbf{x} + \mathbf{T}_{\mathbf{x}}\hat{\mathbf{p}}_{\mathbf{x}} - \mathbf{y} - \mathbf{T}_{\mathbf{y}}\hat{\mathbf{p}}_{\mathbf{y}}, \tag{13}$$

$\hat{\mathbf{p}}_{\mathbf{x}}$ and $\hat{\mathbf{p}}_{\mathbf{y}}$ are obtained by minimizing the distance $\|\mathbf{x} + \mathbf{T}_{\mathbf{x}}\mathbf{p}_{\mathbf{x}} - \mathbf{y} - \mathbf{T}_{\mathbf{y}}\mathbf{p}_{\mathbf{y}}\|$ and

$$\rho(\mathbf{r};\sigma) = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i^2}{r_i^2 + \sigma^2}. \tag{14}$$

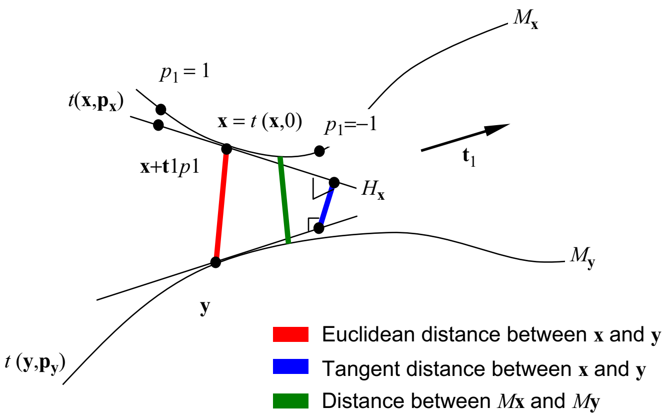

$M_{\mathbf{x}}$
$p_1 = 1$
$t(\mathbf{x},\mathbf{p}_{\mathbf{x}})$
$\mathbf{x} = t\,(\mathbf{x},0)$
$p_1 = -1$
$\mathbf{t}_1$
$\mathbf{x}+\mathbf{t}1p1$
$H_{\mathbf{x}}$
$M_{\mathbf{y}}$
$\mathbf{y}$
$t\,(\mathbf{y},\mathbf{p}_{\mathbf{y}})$
🟥 Euclidean distance between x and y
🟦 Tangent distance between x and y
🟩 Distance between Mx and My

**Fig. 2.** Illustration of the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ and the tangent distance derived by the tangent planes. The curves $M_{\mathbf{x}}$ and $M_{\mathbf{y}}$ represent the sets of points obtained by applying the chosen transformations to $\mathbf{x}$ and $\mathbf{y}$, and the lines going through are the tangent to these curves. The tangent spaces do not intersect by assuming the input space has more dimension than the number of the chosen transformations.

Here, $\mathbf{x}$ and $\mathbf{y}$ stand for two data vectors, $n$ is the vector dimension, and $\mathbf{T}_{\mathbf{x}}$ and $\mathbf{T}_{\mathbf{y}}$ denote the tangent spaces corresponding to $\mathbf{x}$ and $\mathbf{y}$, respectively. $\sigma$ is the kernel parameter corresponding to the error scale, and the $\rho$-function here is a Geman–McClure function. This kernel is designed to be concurrently robust against both irrelevant data transformations and intra-sample outliers. On the one hand, by finding the shortest distance between two tangent planes, the resulting similarity measure between two images is robust against the modeled transformations. On the other hand, using a robust $\rho$-function, the influence of an outlier far away from its reasonable appearance is largely suppressed. Compared to the previous work of robust hybrid kernel that suppresses noise from class mean [6], this kernel has the additional robustness against irrelevant data transformations using the notion of tangents.

Using the first-order Taylor series expansion in tangent distance computation, the proposed kernel is robust within a certain degree of image transformations. To illustrate the relation between the considered transformations and the resulted local robustness, an example is given by transforming a reference image $\mathbf{x}$ with different degrees and measuring the corresponding kernel values. Fig. 4(a) shows 10 typical handwritten digit images for the set of prototypes. Also another digit “7” is chosen as the reference image $\mathbf{x}$ and it was transformed by rotation of different degrees. We show some transformed reference images in Fig. 4(b). Since a kernel in general evaluates data similarity in the kernel-mapped space, in our expectation, the reference image should result in the highest kernel value when it is compared to the prototype “7” among all other digits. In Fig. 4(c), we plot the curves that represent the kernel values $k(\mathbf{x},\mathbf{y}) = 1 - \rho(\mathbf{r}_T(\mathbf{x},\mathbf{y}),\sigma)$ using the reference image $\mathbf{x}$ and prototype image $\mathbf{y}$. Note that the pattern deformation is assumed to be described by five types of image transformations as shown in Fig. 3(a). In Fig. 4(e), we also give the curves using RBF kernel for comparison.

The curve corresponding to digit “7” is close to the curve corresponding to digit “9” because these two digits are similar. If the reference image is rotated (clockwise), the RBF kernel may confuse it with other digits, such as “9” or “4”. In contrast, our kernel behaves quite stable for rotation within 20° because the embedded tangent distance computation accounts for the local image transformation. On the other hand, it is possible to simulate the effect of small image transformation not in the tangent space, such as rotation, using the tangent vectors of other types of transformations, because these image transformations are not

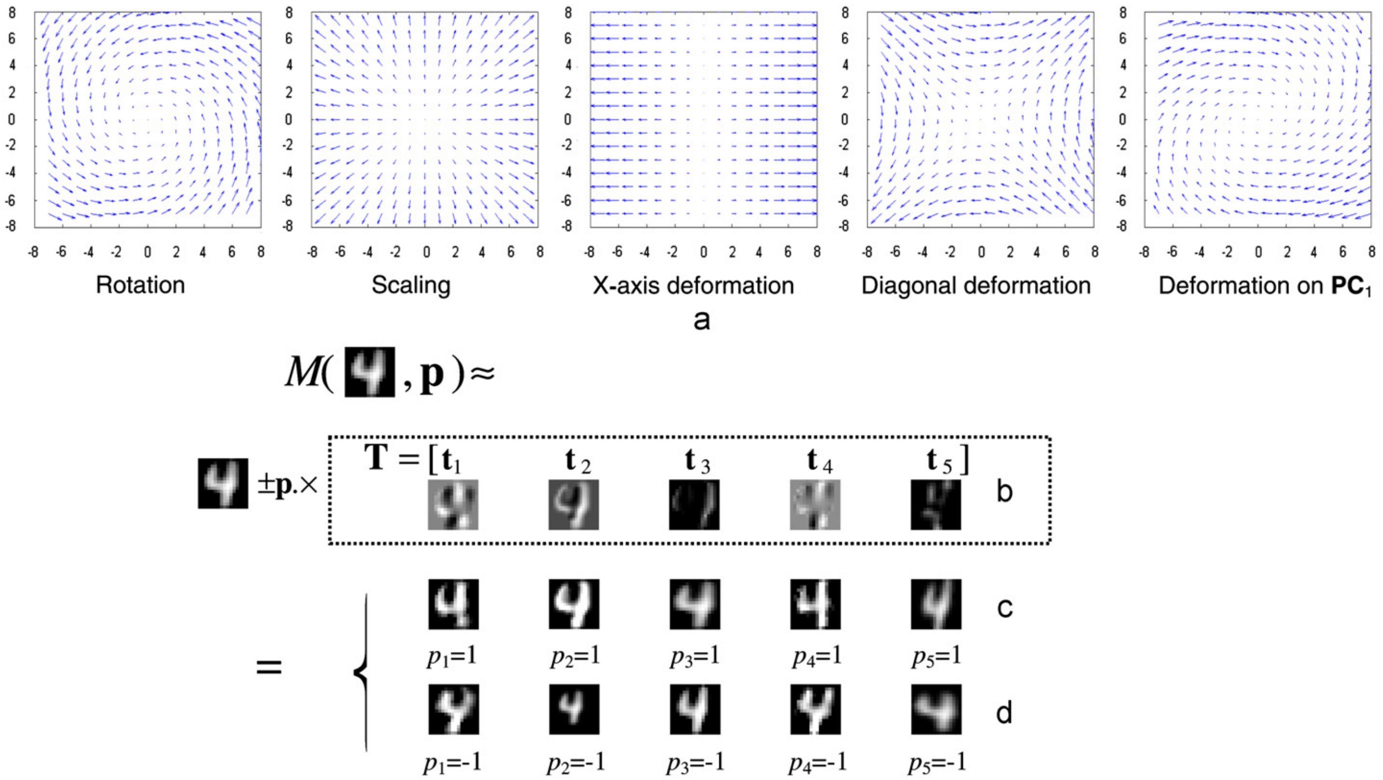


**Fig. 3.** Illustration of tangent vectors. (a) The corresponding pixel displacements shown in vector fields, (b) visualization of 5 tangent vectors, and normalized to 0–255 for representation, and (c) and (d) give the transformation results.

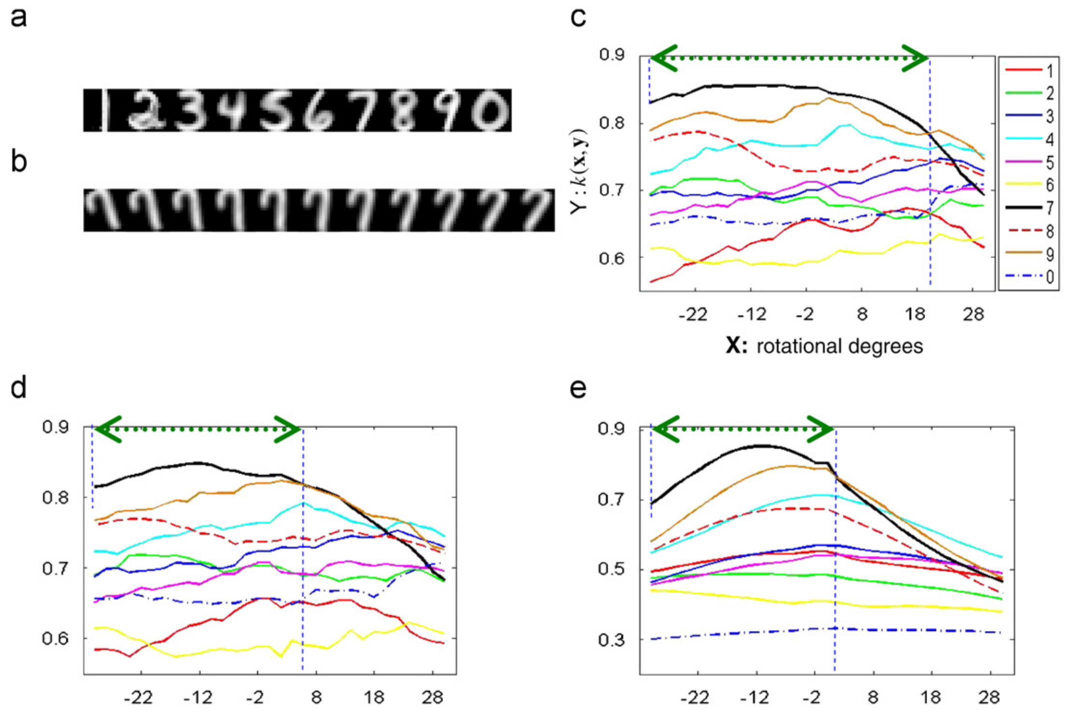


**Fig. 4.** Kernel values between (a) the prototype images and (b) the reference images using (c) the proposed kernel, (d) the proposed kernel without rotation tangent vector included into the tangent space, and (e) the RBF kernel. The original reference image is shown in the middle of (b). All images are 16 pixels in both height and width.

fully independent. We justify this statement in Fig. 4(d). Similar curves can be obtained for other transformations (scaling, skewing,…). Generally, the local robustness of the proposed kernel with respect to these transformations will lead to more accurate classification results under pattern deformation.

In constructing a new kernel, however, it is important to check the positive definiteness for its feasibility. The positive definiteness of a kernel plays an important role in kernel-based learning approach. For example, for SVM, the use of a pd kernel insures that the optimization problem (5) is convex, which assures the convergence of the algorithm towards a unique optimal solution. We next show our proposed kernel is theoretically legitimate by proving its positive definiteness. In doing so, we first review the closure properties of pd kernels [23]:

**Proposition 1.** *Let $k_1$ and $k_2$ be two positive definite kernel functions over $X \times X$ , $\beta \in \mathbb{R}^+$ , $\Phi : X \rightarrow \mathbb{R}^s$ is a mapping function with $k_3$ a positive definite kernel over $\mathbb{R}^s \times \mathbb{R}^s$ . Then the following functions are also positive definite kernels:*

$$k(\mathbf{x},\mathbf{x}') := k_1(\mathbf{x},\mathbf{x}') + k_2(\mathbf{x},\mathbf{x}') \quad (15)$$
$$k(\mathbf{x},\mathbf{x}') := \beta k_1(\mathbf{x},\mathbf{x}') \quad (16)$$
$$k(\mathbf{x},\mathbf{x}') := k_1(\mathbf{x},\mathbf{x}') \times k_2(\mathbf{x},\mathbf{x}') \quad (17)$$
$$k(\mathbf{x},\mathbf{x}') := k_3(\Phi(\mathbf{x}), \Phi(\mathbf{x}')) \quad (18)$$

Please refer to [23] for the details of the proof. Furthermore, recall that a real-valued function $f$ of a real variable is said to be positive definite if the inequality $\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j f(x_i - x_j) \geq 0$ holds for all possible combinations of the real numbers $x_i$ and $c_i$, and Schoenberg generalized this definition to $\mathbb{R}^d$so that if the inequality $\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \varphi(\mathbf{x}_i - \mathbf{x}_j) \geq 0$ holds,$\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be positive definite [33]. In other words, the $m \times m$ matrix $A$ with entries $A_{ij} = f(x_i - x_j)$ or $A_{ij} = \varphi(\mathbf{x}_i - \mathbf{x}_j)$ is a positive semi-definite matrix. Constructing and integrating the integral form, Micchelli [34] showed the *inverse multiquadric* function $F(t) = (\sigma^2 + t)^{-(1/2)}$ defined on $[0,\infty)$, $\sigma > 0$, is positive definite. Hence letting $F(t) = F(||\mathbf{x} - \mathbf{x}'||^2)$the matrix $K$ with entries $K_{ij} := (\sigma^2 + ||\mathbf{x}_i - \mathbf{x}_j||^2)^{-1/2}$ is positive semi-definite. In other words, we have the inverse multiquadric function $k_{imq}(\mathbf{x},\mathbf{x}') := (\sigma^2 + ||\mathbf{x} - \mathbf{x}'||^2)^{-1/2}$ as a pd kernel function. Based on Micchelli's results, we prove the proposed kernel is a positive definite kernel function:

**Theorem 1.** *The function $k_{robust}$ defined by*

$$k_{robust}(\mathbf{x},\mathbf{x}') := 1 - \frac{1}{n}\sum_{i=1}^{n} (x_i - x'_i)^2 / ((x_i - x'_i)^2 + \sigma^2),$$

*where $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{x}' = (x'_1, \ldots, x'_n)^T$, is a positive definite kernel.*

**Proof.** Let us define the $i$th sub-kernel function $k_{imq}^{(i)}(\mathbf{x},\mathbf{x}') := k_{imq}(\mathbf{A}_i\mathbf{x}, \mathbf{A}_i\mathbf{x}')$, where $\mathbf{A}_i$ is an $n \times n$ matrix with the $(p,q)$ entry defined by

$$(\mathbf{A}_i)_{pq} = \begin{cases} 1, & p = q = i \\ 0, & \text{otherwise} \end{cases}. \quad (19)$$

That is, $\mathbf{A}_i \in \mathbb{R}^{n \times n}$is the mapping function associated with $k_{imq}^{(i)}$, which maps the input of an inverse multiquadric by $\mathbf{A}_i\mathbf{x} = (0, \ldots, 0, x_i, 0, \ldots, 0)^T$ and hence $k_{imq}^{(i)}(\mathbf{x},\mathbf{x}') = (\sigma^2 + (x_i - x'_i)^2)^{-1/2}$. From the above proposition, $k_{imq}^{(i)}$ is a pd kernel because $k_{imq}$ is pd, and so is its square $\tilde{k}_{imq}^{(i)}(\mathbf{x},\mathbf{x}') := k_{imq}^{(i)}(\mathbf{x},\mathbf{x}') \times k_{imq}^{(i)}(\mathbf{x},\mathbf{x}')$. Because summing pd kernels or multiplying a constant $\beta \in \mathbb{R}^+$ to pd kernels still leads to pd kernels, it implies that the linear combination of all the sub-kernel functions $\tilde{k}_{imq}^{(i)}$, i.e. the proposed kernel function:

$$k_{robust} = \frac{1}{n}\sum_{i=1}^{n} \sigma^2 \tilde{k}_{mq}^{(i)},$$

is positive definite. □

Being positive definite, the proposed robust kernel is theoretically valid to be used in conjunction with the kernel-based learning techniques. In the next section, we apply it on different visual learning problems to manifest its robust performance for practical use.

## 4. Experimental results

We demonstrate the robustness of the proposed kernel on several visual learning problems, including handwritten digit classification, face recognition, and data visualization. For the digit classification and face recognition problems, we applied the robust kernel in conjunction with SVM. The SVM implementation was based on *LIBSVM* software [35]. For the digit classification problem, the proposed kernel was shown to be robust against geometric deformation and noise corruption to a certain degree. We also assessed the proposed kernel on the face recognition problem with faces under different degrees of partial occlusions. Finally, for the data visualization problem, we embedded our robust kernel into kernel Fisher discriminant (KFD) analysis and achieved more satisfactory results compared to the state-of-the-art methods.

### *4.1. Hand-written digit classification*

In this section, we apply the proposed kernel on the application of hand-written digit classification for evaluating its robustness. We begin by reporting overall trends for the relation between the data transformations and the resulted SVM classification performance, followed by demonstrating its performance under noise corruption. The USPS database [37], which contains 9298 images of size 16 × 16, was used in this experiment. For a digit $\mathbf{x}$, we assume that the corresponding manifold $M_{\mathbf{x}}$ can be described by five types of deformation, that is, $t(\mathbf{x},\mathbf{p})$ consists of five transformations to describe the local deformations. They are rotation:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{pmatrix},$$

scaling:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x + \alpha x \\ y + \alpha y \end{pmatrix},$$

and skewing along $x$-axis:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x + \alpha x \\ y \end{pmatrix},$$

diagonal direction:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x + \alpha y \\ y + \alpha x \end{pmatrix},$$

and the most dominant principal component direction. The five transformations are illustrated as vector fields in Fig. 3 taking an image of digit 4 as an example. In the experiments we simulated the tangent vectors for images letting $\theta = \pm 20^\circ$ and $\alpha = \pm 1/16$; SVM was used as the kernelized learning machine for classification.

Since SVM was originally developed as a binary classifier, the one-against-one strategy [38] was employed to extend the binary classifiers to the multi-class classification problem in our implementation. That is, we trained each binary SVM classifier, denoted by SVM$_{ij}$, with the data of the $i$th and $j$th classes, and accomplished the multi-class digit classification by applying a voting scheme. The parameter $C$ in Eq. (4) for each SVM$_{ij}$ was determined by cross validation on a separate validation set of 1000 random samples. In the following digit classification experiments, we compute the average accuracy of 11 runs with 1000 randomly selected samples for training and another 100 random samples for testing. Note that, the parameter $\sigma$ controls where the kernel function becomes saturated, and here we properly set it to 0.248 by performing cross validation [32] on the aforementioned validation set. One may alternatively determine it through optimizing a specially designed quality function. For example, Chen et al. [30] presented a unified kernel optimization framework that employs different discriminant criteria, including the local Fisher criteria (LFC), the subclass Fisher criteria (SFC), and the marginal Fisher criteria (MFC), for the optimization. Embedding our kernel function into its formulation, it can lead to a data-dependent kernel to be optimized over the combinational coefficient $\boldsymbol{\alpha}$ and $\sigma$. On the other hand, the *kernel alignment* [19,31] method provides an alternative approach, which adjusts the kernel parameter based on aligning the corresponding kernel matrix to a target kernel.

The robustness of the proposed kernel against pattern deformation was evaluated. In doing so, we used the plain digit images to train an SVM classifier, and additionally applied different types and degrees of transformations to the testing images. The representative transformations selected here were *rotation* and *skewing* on the diagonal direction. In Fig. 5, we report the average SVM recognition accuracy over 11 runs, where 1000 random samples were used in training and another 100 random samples were selected and transformed by rotation or skewing for testing. For comparison, the results of SVMs combined with other representative kernels were evaluated as well. In Fig. 5, we plot the SVM recognition accuracy using our ***Robust*** kernel, ***linear*** kernel, ***polynomial*** kernel, ***sigmoid*** kernel, and ***RBF*** kernel functions under different transformations of different degrees. It can be observed that our kernel performs robustly within a certain degree of image transformation because the image deformations of the selected types have been modeled into the tangent subspace in the proposed kernel. Although the degradation of recognition rate becomes obvious for rotation more than 20° and for skewing with magnitude $|\alpha| > 0.2$, our kernel indeed demonstrates the superior robustness against pattern deformation over the other kernels. Clearly, the tangent distance computation embedded in our kernel has played an important role when computing similarity between images with pattern deformation.

Next, we evaluate the performance of the proposed kernel when data is corrupted by different levels of noises. Three types of noises were simulated and applied to the experimental images, including: (a) additive Gaussian random noise, (b) additive salt and pepper noise, and (c) multiplicative speckle noise. Some examples of the corrupted digits are given in Fig. 6. In Fig. 7, we report the average SVM recognition accuracy over 11 runs when incorporated with our robust kernel. In Fig. 8, we report the performance of the previous robust kernel methods using SVM with four different types of robust kernels [39], which are denoted by $K^{RBF}_{dMN}$, $K^{RBF}_{d2S}$, $K^{ND}_{dMN}$, and $K^{ND}_{d2S}$, and two robust SVM methods of Jittered Query method [16] (denoted by JQ$^{k\mathrm{NN}}$) and virtual support vectors (VSV) method [3] for comparison. Note that, following [39] we use $K^{f}_{d}(\mathbf{x},\mathbf{y}) := f(d(\mathbf{x},\mathbf{y}))$to represent using metric $d$ on a particular distance based kernel $f$, where $f$ is defined as $K^{RBF}_{d}(\mathbf{x},\mathbf{y}) := \exp(-d(\mathbf{x},\mathbf{y})/2\sigma^2)$ or $K^{ND}_{d}(\mathbf{x},\mathbf{y}) := -d(\mathbf{x},\mathbf{y})^{\gamma}$. The metric $d$ is either the combination of the two square one-sided tangent distance

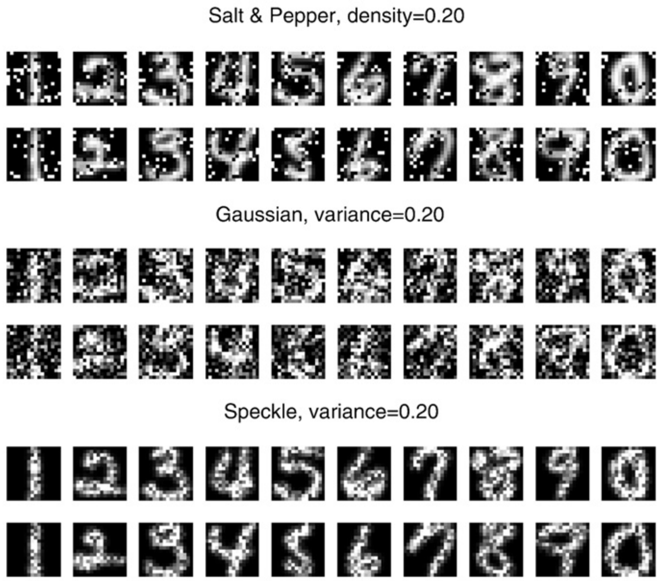


**Fig. 6.** Some example images of noise-corrupted digits in USPS database [37].

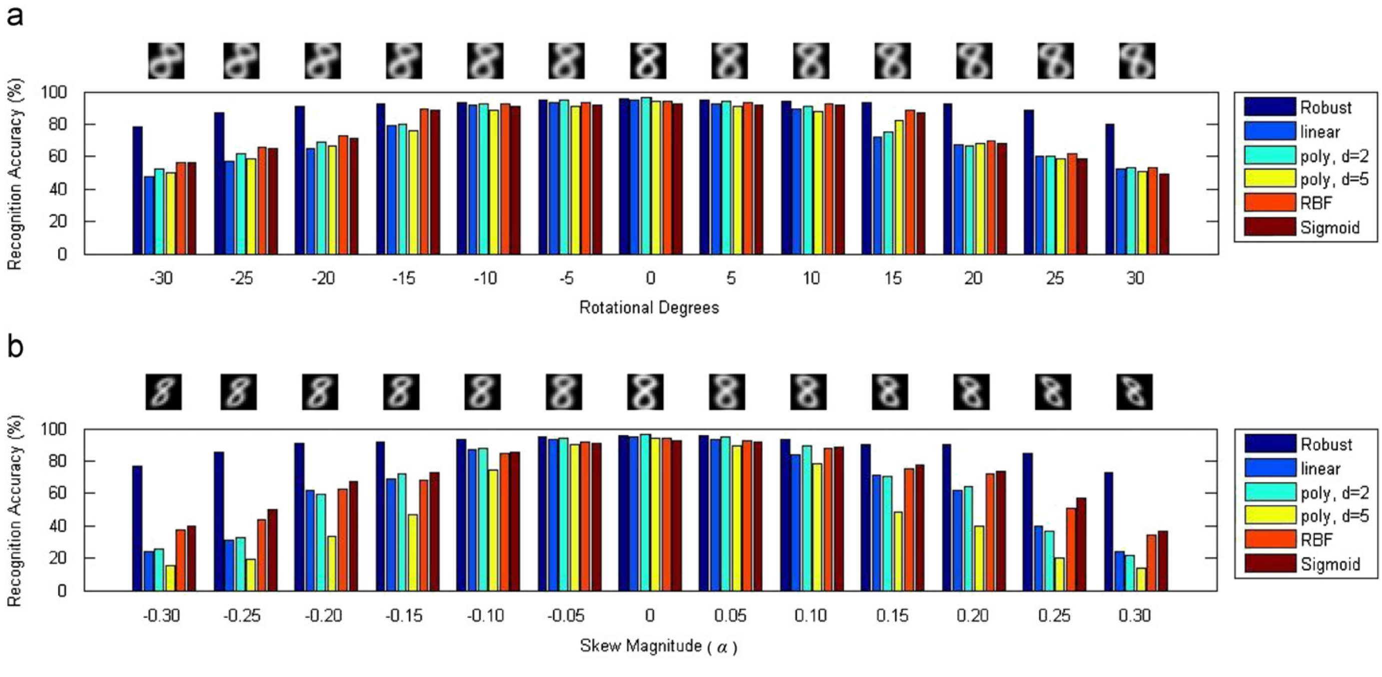


**Fig. 5.** USPS hand-written digit recognition accuracy under (a) rotation and (b) diagonal skewing deformations. The testing images were applied with the above image transformations of different degrees. Example images of digit "8" after applying the transformations of different degrees are shown in the top row.

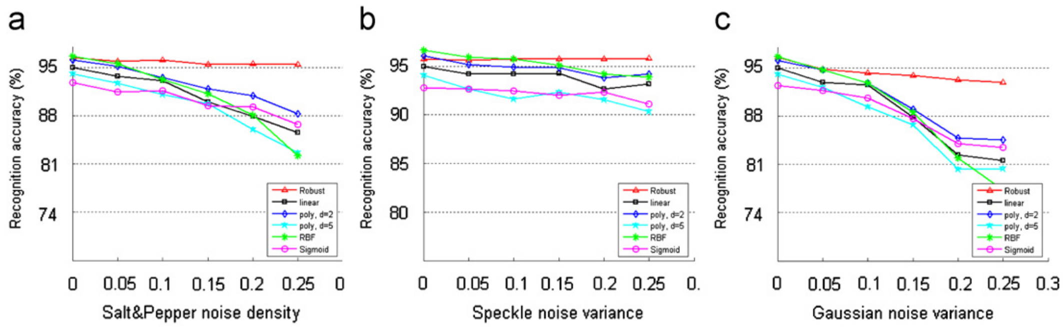


**Fig. 7.** USPS hand-written digit recognition accuracy under (a) salt & pepper noise, (b) speckle noise, and (c) Gaussian noise with various noise levels.

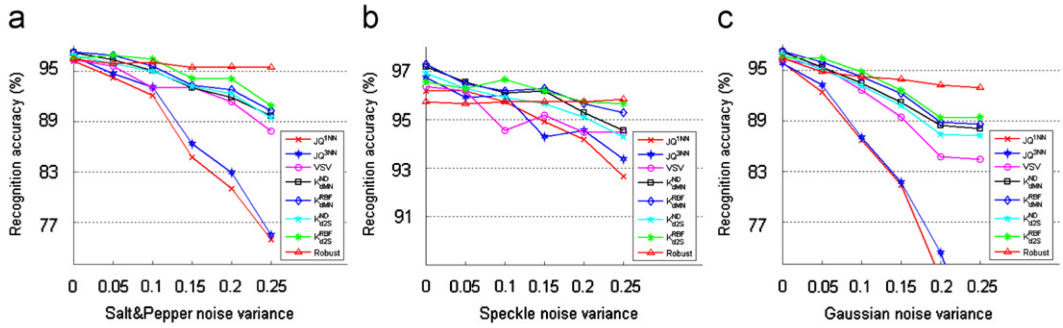


**Fig. 8.** Comparison with other methods using USPS handwritten digit dataset under (a) salt & pepper noise, (b) speckle noise, and (c) Gaussian noise with various noise levels.

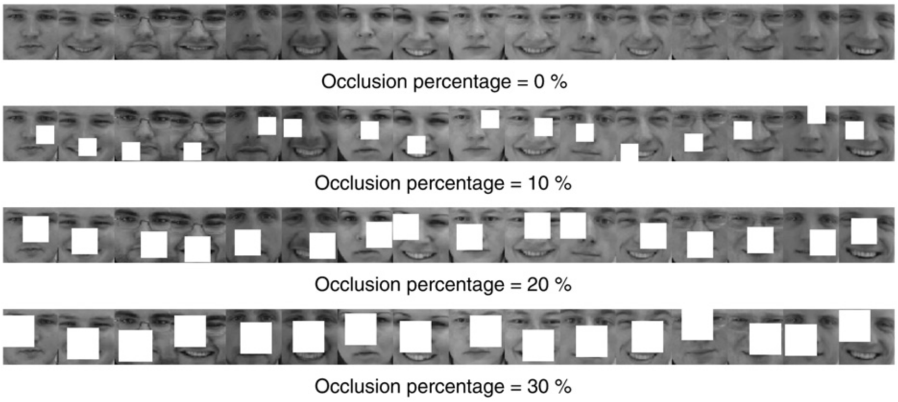
$d_{MN}(\mathbf{x},\mathbf{y}) = (0.5 \times (d_{1S}^2(\mathbf{x},\mathbf{y})+d_{1S}^2(\mathbf{y},\mathbf{x})))^{1/2}$, where $d_{1S}(\mathbf{x},\mathbf{y}) = \min_{\hat{\mathbf{p}}_\mathbf{x}} ||\mathbf{x}+\mathbf{T}_\mathbf{x}\hat{\mathbf{p}}_\mathbf{x}-\mathbf{y}||$ or the two-sided tangent distance $d_{2S}(\mathbf{x},\mathbf{y}) = \min_{\hat{\mathbf{p}}_\mathbf{x},\hat{\mathbf{p}}_\mathbf{y}} ||\mathbf{x}+\mathbf{T}_\mathbf{x}\hat{\mathbf{p}}_\mathbf{x}-\mathbf{y}-\mathbf{T}_\mathbf{y}\hat{\mathbf{p}}_\mathbf{y}||$. In the experiment, the bandwidths of the traditional RBF kernel, $K^{RBF}_{dMN}$, and $K^{RBF}_{d2S}$ kernels were determined through cross validation for each noise condition. For the parameters used in other compared robust kernels, they were set in the same way as [39] and also for those used in the VSV and JQ methods [3,16].

From Fig. 7 and 8, we see that our kernels,$K^{RBF}_{dMN}$, $K^{RBF}_{d2S}$, $K^{ND}_{dMN}$, $K^{ND}_{d2S}$, and VSV method outperform the SVMs embedded with the classical kernels. The proposed kernel outperforms the classical kernels and other types of robust kernels because it is able to handle the pattern deformation and noise/outliers concurrently, while the other methods can at most account for one of these problems. From Fig. 8, one can observe that the curve of $K^{RBF}_{d2S}$, which incorporates the RBF kernel with two-sided tangent distance, degrades faster than our kernel as the noise level increases. The difference between the Gaussian function and Geman–McClure $\rho$-function brings quite different characteristics to the resulted kernels from the robustness perspective. From the experimental results, the performance of the RBF kernel is quite sensitive to salt-and-pepper noise corruption. The JQ and VSV methods both provide better performance than the classical kernels if the noise level is not too high. However, their recognition accuracies are substantially degraded as the noise level is increased. If the SVM is used in conjunction with the proposed kernel, in contrast, the classification results are relatively stable under different noises. From the above experiments, it is thus evident that the proposed kernel is more robust than the competitive kernels under the irrelevant image transformations and noise corruption. It is also noteworthy that, even under the noise-free condition, the classification accuracy of using our robust kernel with SVM is comparable to the state-of-the-art approaches devoted to digit recognition, whose best performance is around 98% [40,41].

### 4.2. Face recognition

Next we evaluate the robustness of our kernel on the application of face recognition with partial occlusions. In this case, occluded pixels are treated as outliers rather than random noise, which may be reduced through filtering. Because of the specific structure of faces, here it is possible to estimate the parameters in the proposed kernel more preferably based on the pixel appearance in a reference face coordinate. That is, treating a face image as an $n$-dimensional vector, it is able to measure the scale parameter $\sigma_i$ of the $i$th subkernel, i.e. $\tilde{k}^{(i)}_{imq}$, from the given face images. We adopted the maximal likelihood (ML) approach to estimate the kernel parameters $\boldsymbol{\sigma} = [\sigma_1,\ldots,\sigma_n]$ for this face recognition problem because determining a large number of variables using cross validation would be computationally infeasible. The estimate of scale parameters thus corresponds to the sample variance for the $i$th pixel in the face model. Note that, different from the case of hand-written digit recognition, the tangent plane $t(\mathbf{x},\mathbf{p})$ for a face image $\mathbf{x}$ used here is to model the face image subspace for slight in-plane pose variations. Therefore we assume that $t(\mathbf{x},\mathbf{p})$ only consists of rotation, scaling, and skewing along the $x$-axis and diagonal directions in this experiment.

We used faces from FG-Net database [42] with facial expressions ranging from neutral to cheerful expression in this experiment. The experimental videos consist of 17 people, and we randomly selected 10 faces for each person to train the classifiers. In the testing phase, another 30 faces for each person exclusive of the training samples were used, where white rectangles of size 10–30% of the face region were randomly imposed on the face images for simulating partial occlusions. Some example images are given in Fig. 9. Each face image was scaled to 64 × 64 and pre-processed by histogram equalization; SVM was used as the kernelized classifier. Note that by vectorizing each image as an $n$-dimensional vector, the above-mentioned ML estimate for the parameters $\boldsymbol{\sigma} = [\sigma_1,\ldots,\sigma_n]$ here corresponds to the sample variance computed



Occlusion percentage = 0 %

Occlusion percentage = 10 %

Occlusion percentage = 20 %

Occlusion percentage = 30 %

**Fig. 9.** Examples of occluded face images from FG-Net [42].

**Table 1**
Average face recognition accuracy (%) under different levels of occlusion.

| Occlusion | Robust | Linear | Poly_$d$=2 | Poly_$d$=3 | RBF |
|---|---|---|---|---|---|
| **0%** | 99.61 | 99.57 | 97.33 | 97.41 | 97.37 |
| **5%** | 99.41 | 99.29 | 95.57 | 95.57 | 95.69 |
| **10%** | 99.10 | 97.96 | 92.90 | 93.18 | 92.71 |
| **15%** | 98.31 | 93.65 | 88.63 | 88.90 | 88.12 |
| **20%** | 97.53 | 87.37 | 83.06 | 83.41 | 83.76 |
| **25%** | 96.51 | 79.61 | 75.65 | 75.96 | 75.88 |
| **30%** | 94.55 | 74.35 | 70.43 | 71.41 | 69.76 |

from all 1700 images presented in this database. The testing procedure was repeated five times and we report the average face recognition accuracy in Table 1. With the nice robustness property of the proposed kernel, high face recognition accuracy is achieved under partial occlusions. In our experiments, we can achieve 94.55% recognition accuracy with 30% random occlusions on face images, while other kernels can only achieve less than 80% accuracy on the same dataset. When SVM is used with the classical kernels, the accuracy is significantly degraded since the occlusion pixels are considered as outliers and these kernels are prone to errors if outliers are present. That is to say, the resulted SVMs are unable to handle such occluded face images if no similar training images are included in training. Also, they cannot handle pose variations with a limited number of training examples. The nonlinear kernels perform worse than the linear kernel in this experiment, because they tend to produce stronger responses to the occlusion pixels. In addition, the linear kernel has the advantage of good generalization when the number of training data is small. In contrast, when the SVM classifier is combined with our kernel, the robustness against occlusions and small face pose variations can be significantly enhanced. As shown in the experiment, the resulted SVM handles such seriously occluded face images very well even though we did not include any occluded images in training. Clearly, it is because of the design of our kernel function that evaluates data similarity in a robust way. This experimental result thus validates the robustness of the proposed kernel for the problem of face recognition under partial occlusions.

### 4.3. Data visualization

Data visualization is an important issue in exploring intrinsic data structures and correlations in many fields. However, discovering the nonlinear data relation is often difficult especially when outliers and large data variations are involved. A similarity measure robust to irrelevant data transformations, such as pattern deformation, is thus demanded to achieve robust data visualization. To further investigate the potential of our kernel, here we apply it in conjunction with KFD analysis for data visualization.

Here, USPS hand-written digit database was again used for visualization with the same kernel parameters obtained in Section 4.1. The data visualization procedure was performed with the following procedure: first, we applied KFD in conjunction with our robust kernel to train one binary Fisher-discriminant with respect to each type of digit. In other words, for $c{=}0\sim9$, we obtained $\mathbf{w}_c^{\Phi}$ by means of $\boldsymbol{\omega}^c$ with one-against-all manner [38]. Next, for a digit image $\mathbf{x}_{test}$, its projection on $\mathbf{w}_c^{\Phi}$ was computed by $\langle \mathbf{w}_c^{\Phi},\Phi(\mathbf{x}_{test})\rangle = \sum_{i=1}^{m}\omega_i^c k(\mathbf{x}_i,\mathbf{x}_{test})$ for $c{=}0\sim9$. Concatenating all projections, each hand-written digit image was represented as a ten-dimensional vector. Finally, we applied the multidimensional scaling (MDS) [36] on all vectors to reduce the data dimensionality from ten to two for visualization.

Note that the handwritten digits of a class may involve several variations, which include all possible positions, sizes, angles, skews, writing styles, and thickness of digits. Meanwhile, the data visualization problem considered here is to map high-dimensional data into a low-dimensional space based on the global data distribution for visualization. It is therefore inevitable to lose some information during the visualization process and some details may be ignored. Consequently the projected class distributions may become strongly overlapping due to the complex within-class variations. Using the robust kernel, however, we are able to reduce the effect of the irrelevant data transformations and better preserve the details or interesting structure in the projection procedure. The images are not filtered for the data visualization, and the details or interesting structure in the original images thus is not erased using the designed robust kernel for similarity measurement.

In the experiment, 1000 randomly chosen samples from USPS database were used for visualization, and the same five tangent vectors were used to describe $t(\mathbf{x},\mathbf{p})$. In Fig. 10, we provide the visualization results using our method and other state-of-the-art approaches for comparison, including principal component analysis (PCA), linear discriminant analysis (LDA), MDS, locally linear embedding (LLE) [43] and supervised LLE (SLLE) [44], and their combinations. As shown in Fig. 10(i), the method incorporated with our robust kernel clearly demonstrates more satisfactory results in discovering the intrinsic data distributions than the state-of-the-art approaches (cf. Fig. 10(a–h)). With the KFD scheme and the proposed kernel, data instances of the same category are nicely distributed away from those of other categories, while other approaches cannot perform very well.

In order to distinguish the performance of our proposed kernel, we also embedded the classical kernels into KFD for comparison. The visualization results using KFD in conjunction with other kernels are shown in Fig. 10(j–l). It is observed that the clusters are

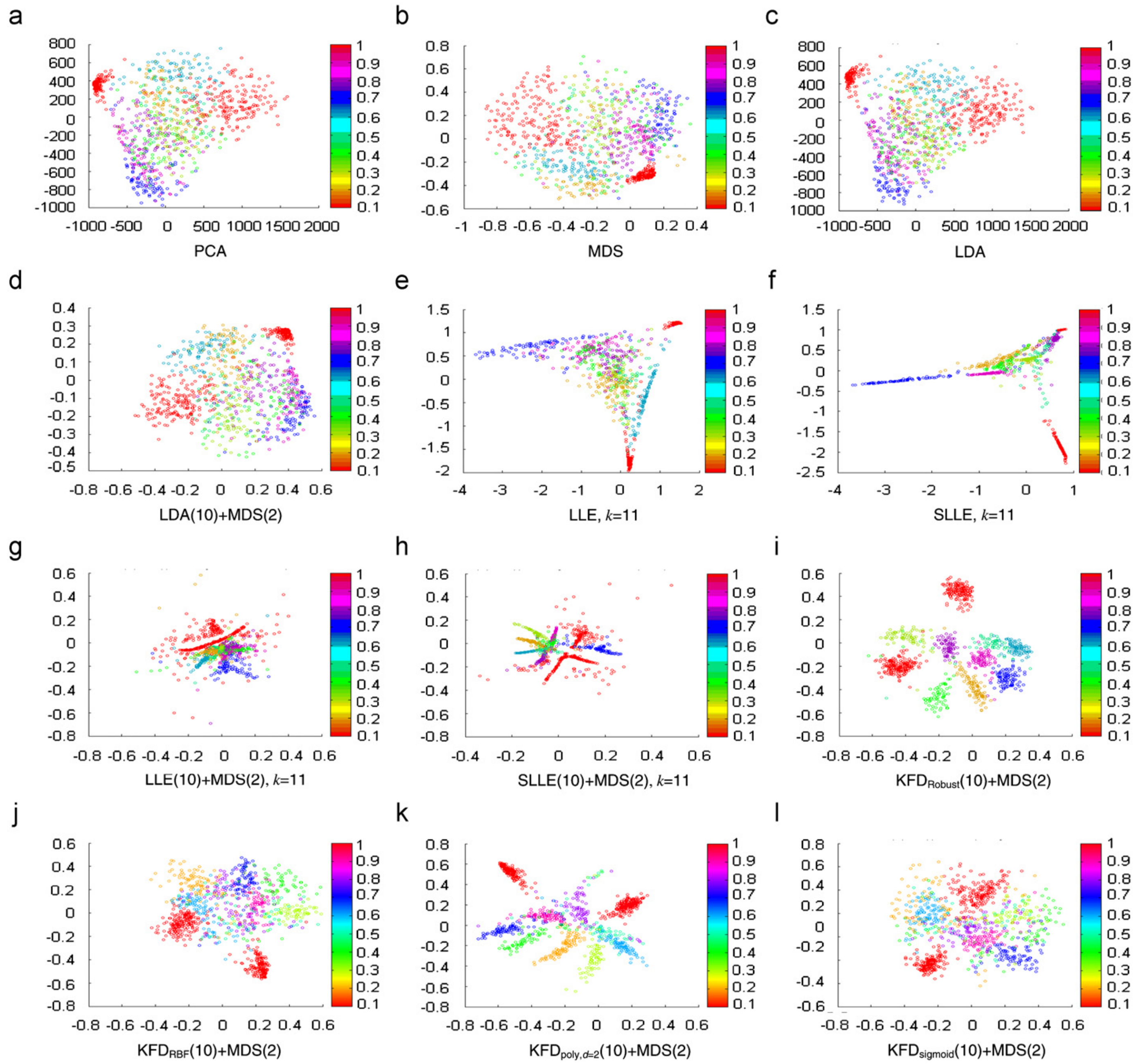


**Fig. 10.** Data dimensionality reduction results of 1000 handwritten digit images. (a–h) Results of different methods. MethodA($d_1$)+MethodB($d_2$) denotes that data dimensionality is reduced to $d_1$ by method A then to $d_2$ by method B. The result in (i) is obtained using our method, i.e. KFD+MDS, and the results (j–l) are obtained using our method embedded with different kernels in KFD.

more scattered and overlapped with each other compared to that using our robust kernel. This is because the Gram matrix derived by the proposed robust kernel has better discrimination for the handwritten digits than those using other kernels, thus leading to the superior separation over different classes after the KFD projections. To verify the robustness of our kernel, we also evaluate the performance by applying random noise to the data. The visualization results are shown in Fig. 11(a)–(f). From the figures, we can observe that even under severe noisy conditions, the KFD method embedded with our kernel still demonstrates considerable reliability in resisting noise disturbance. The data distributions under noise corruption are not much affected using the proposed robust kernel, while the data distributions obtained by other dimensionality reduction techniques may not be satisfactory for visualization.

## 5. Conclusions

In this work, we presented a novel robust kernel that integrates a robust error function and the tangent distance to make the associated visual learning more robust against random noise and pattern deformations. From a theoretical point of view, we showed the positive definiteness of the proposed robust kernel, thereby it is feasible to incorporate this kernel with various kernel-based learning algorithms. From a practical point of view, three experiments demonstrated the improved performance of the proposed robust kernel in various visual learning problems, including digit classification, face recognition, and data visualization, under different kinds of data corruption.

Finding the minimal distance between two manifolds of image transformations, we achieved a robust kernel function that is

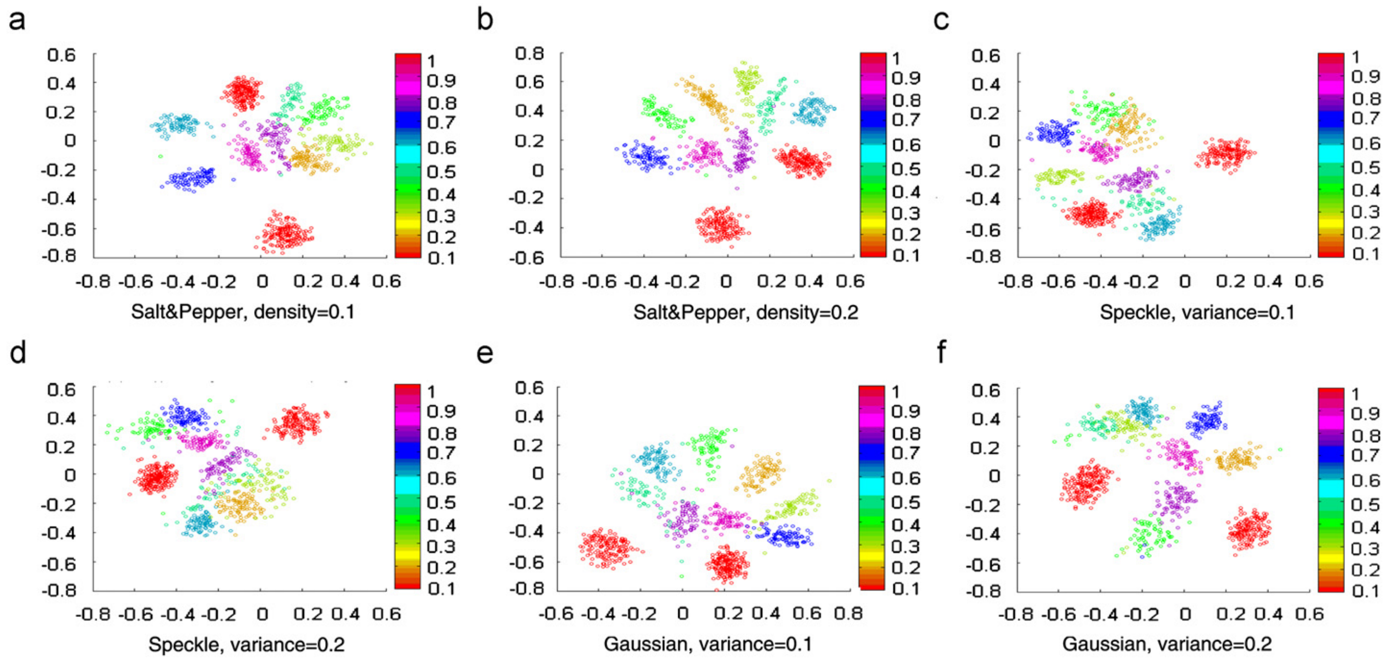


**Fig. 11.** Data visualization of USPS handwritten digit images. (a–f) KFD+MDS with the proposed kernel under various types and levels of noises.

robust against the modeled local pattern deformations. However, because the manifolds were linearly approximated in computing their distance, our formulation may not handle large variations in the transformation parameters. Finding an accurate distance between them still remains a difficult problem because there is no analytic expression for the nonlinear manifolds in general. In the future, we would like to apply the proposed robust kernel in conjunction with the other classifiers, such as RBF-networks, to see if it can improve their robustness. Also, the parameters in the proposed robust kernel can be determined using some kernel optimization techniques, such as the framework in Chen et al. [30] or the kernel alignment method [19], to further improve its performance.

## References

[1] J. Segman, J. Rubinstein, Y.Y. Zeevi, The canonical coordinates method for pattern deformation: theoretical and computational considerations, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (1992) 1171–1183.

[2] T. Vetter, T. Poggio, Image synthesis from a single example image, in: Proceedings of the European Conference on Computer Vision, Cambridge, UK, 1996, pp. 652–659.

[3] B. Schölkopf, C. Burges, V. Vapnik, Incorporating invariances in support vector learning machines, in: Proceedings of the International Conference on Artificial Neural Networks, Berlin, Germany, 1996, pp. 47–52.

[4] C. Lu, T. Zhang, R. Zhang, C. Zhang, Adaptive robust kernel PCA algorithm, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 2003, pp. VI–621-4.

[5] B. Schölkopf, A. Smola, K.-R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1999) 1299–1319.

[6] C.T. Liao, S.H. Lai, A robust kernel based on robust $\rho$-function, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Honolulu, USA, 2007, pp. II–421–II-424.

[7] A. Barla, E. Franceschi, F. Odone, A. Verri, Image kernels, in: Proceedings of the International Workshop on Pattern Recognition with Support Vector Machines, Quebec, Canada, 2002, pp. 83–96.

[8] W. Du, K. Inoue, K. Urahama, Robust kernel fuzzy clustering, Fuzzy Systems and Knowledge Discovery 1 (2005) 454–461.

[9] J.H. Chen, M-estimator based robust kernels for support vector machines, in: Proceedings of the International Conference on Pattern Recognition, Cambridge, UK, 2004, pp. 168–171.

[10] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Algorithms, Cambridge University Press, Cambridge, 2000.

[11] M. Debruyne, S. Serneels, T. Verdonck, Robustified least squares support vector classification, Journal of Chemometrics 23 (2009) 479–486.

[12] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (1995) 273–297.

[13] V. Vapnik, Statistical Learning Theory, Wiley, Chichester, 1998.

[14] D. DeCoste, B. Schölkopf, Training invariant support vector machines, Machine Learning 46 (2002) 161–190.

[15] T.B. Trafalis, R.C. Gilberta, Robust classification and regression using support vector machines, European Journal of Operational Research 173 (2006) 893–909.

[16] D. DeCoste, M.C. Burl, Distortion-invariant recognition via jittered queries, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, USA, 2000, pp. 732–737.

[17] H. Xu, C. Caramanis, S. Mannor, Robustness and regularization of support vector machines, Journal of Machine Learning Research 10 (2009) 1485–1510.

[18] Q. Song, W. Hu, W. Xie, Robust support vector machine with bullet hole image classification, IEEE Transactions on System, Man, and Cybernetics 32 (2002) 440–448.

[19] N. Cristianini, J. Kandola, A. Elisseeff, J. Shawe-Taylor, On kernel target alignment, in: Advances in Neural Information Processing Systems, vol. 14, 2002, pp. 367–373.

[20] R. Kondor, T. Jebara, A kernel between sets of vectors, in: Proceedings of the International Conference Machine Learning, Washington D.C., USA, 2003, pp. 361–368.

[21] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, Journal of Machine Learning Research 4 (2003) 913–931.

[22] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A.J. Smola, K.-R. Müller, Invariant feature extraction and classification in kernel spaces, Advances in Neural Information Processing Systems (2000) 526–532.

[23] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, 2001.

[24] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, New York, 1986.

[25] P.J. Huber, Robust Statistics, Wiley, New York, 1981.

[26] R. Maronna, D. Martin, V. Yohai, Robust Statistics—Theory and Methods, Wiley, New York, 2006.

[27] S. Geman, D.E. McClure, Statistical methods for tomographic image reconstruction, Bulletin of the International Statistical Institute 52 (1987) 5–21.

[28] P.Y. Simard, B. Victorri, Y.A. LeCun, J.S. Denker, Tangent prop—a formalism for specifying selected invariances in an adaptive network, Advances in Neural Information Processing Systems (1992) 895–903.

[29] P.Y. Simard, Y.A. LeCun, J.S. Denker, B. Victorri, Transformation invariance in pattern recognition-tangent distance and tangent propagation, Lecture Notes in Computer Science 1524 (1998) 239–274.

[30] B. Chen, H. Liu, Z. Bao, Optimizing the data-dependent kernel under a unified kernel optimization framework, Pattern Recognition 41 (2008) 2107–2119.

[31] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, M. Jordan. Learning the kernel matrix with semi-definite programming, in: Proceedings of the

International Conference on Machine Learning, Sydney, Australia, 2002, pp. 323–330.

[32] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996.

[33] J. Stewart, Positive definite functions and generalizations, a historical survey, Rocky Mountain Journal of Mathematics 6 (1976) 409–433.

[34] C.A. Micchelli, Interpolation of scattered data: distance matrices and conditionally positive functions, Constructive Approximation 2 (1986) 11–22.

[35] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines. Software, available at 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.

[36] I. Borg, P. Groenen, Modern Multidimensional Scaling, Springer-Verlag, Berlin, 1997.

[37] 〈http://www.kernel-machines.org/data.html〉.

[38] C.W. Hsu, C.J. Lin, A comparison of methods for multi-class SVMs, IEEE Transactions on Neural Networks 13 (2002) 415–425.

[39] B. Haasdonk, D. Keysers, Tangent distance kernels for support vector machines, in: Proceedings of the International Conference on Pattern Recognition, Quebec, Canada, 2002, pp. 864–868.

[40] J.X. Dong, A. Krzyzak, C.Y. Suen, Fast SVM training algorithm with decomposition on very large datasets, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 603–618.

[41] D. Keysers, R. Paredes, H. Ney, E. Vidal, Combination of tangent vectors and local representation for handwritten digit recognition, in: Proceedings of the International Workshop on Statistical Pattern Recognition, Ontario, Canada, 2002, LNCS 2396, Springer-Vertag, pp. 538–547.

[42] 〈http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html〉.

[43] L.K. Saul, S.T. Roweis, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[44] D. Ridder, O. Kouropteva, O. Okun, M. Pietikainen, R.P.W.. Duin, Supervised locally linear embedding, in: Proceedings of the International Conference on Artificial Neural Networks, Istanbul, Turkey, 2003, pp. 333–341.

[45] P. Anandan, A computational framework and an algorithm for the measurement of visual motion, International Journal of Computer Vision 2 (1989) 283–310.

**Chia-Te Liao** received his B.S. degree in computer science and information engineering from National Chi Nan University, and M.S. degree in computer science and information engineering from National Chung Cheng University, in 2001 and 2003, respectively. He is currently a Ph.D. candidate in the department of computer science at National Tsing Hua University. His research interests include pattern recognition and machine learning.

**Shang-Hong Lai** received the B.S. and M.S. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 1986, 1988 and 1995, respectively. He joined Siemens Corporate Research in Princeton, New Jersey, as a member of technical staff in 1995. Since 1999, he became a faculty member in the Department of Computer Science, National Tsing Hua University, Taiwan. He is currently a professor in the same department. In 2004, he was a visiting scholar with Princeton University. Dr. Lai's research interests include computer vision, visual computing, pattern recognition, medical imaging, and multimedia signal processing. He has authored more than 150 papers published in the related international journals and conferences. He holds ten US patents for inventions related to computer vision and medical image analysis. He has been a member of program committee of several international conferences, including CVPR, ICCV, ECCV, ACCV, ICPR, and ICME.