A NOVEL ROBUST KERNEL FOR APPLICATIONS TO IMAGES

Chia-Te Liao and Shang-Hong Lai

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

ABSTRACT

Robustness is an essential issue to computer vision and pattern recognition in developing multimedia applications. In this work, we present a robust kernel approach that is highly robust against random noises and intra-class deformations. By incorporating the robust error function used in robust statistics together with a deformation-invariant distance measure, the derived robust kernel is shown to be insensitive to the influence of outliers and robust to intra-class deformations. In the experiments, we justify our robust kernel with different kernel machines with applications to handwritten digit recognition and data visualization on the USPS database.

Index Terms—Robust kernel, robust classification, digit recognition, data visualization

1. INTRODUCTION

In many machine learning applications, most of the learning algorithms are guaranteed to learn the optimal classifier if the amount of data is infinite. However, in real situations, it is impractical to gather infinite samples covering all data variations. Taking the example of handwritten digit recognition, it needs to present all possible variations of a category in the prototype set, which includes all possible positions, sizes, angles, skews, writing styles, thickness of digits. In addition, the acquired data is often corrupted by outliers or noise, which also distracts the learning method from learning correct information. Thereby, many methods have been proposed to enhance the robustness of the existing learning algorithms against irrelevant data transformation. Some made use of an invariant distance measure constructed in such a way that the distance between a prototype and a pattern is not affected by the irrelevant transformation. For example, Huttenlocher [1] et al. developed a method based on Hausdorff distance to robustly computing similarities between images. Some made efforts for data representations by designing feature extractors that are minimally infected by the irrelevant transformation, such as those using SIFT descriptor for object recognition [2].

Numerous approaches were proposed recently to improve the robustness of learning algorithms based on the kernel methods. For example, Lu et al. [4] proposed to detect and remove outliers by kernelizing Principal Component Analysis (KPCA). Liao and Lai [6] presented a hybrid robust kernel with a mixture of a ρ-function and an RBF kernel to relieve noise effects for the learning algorithms. There, kernel trick hinges on a suitable representation of the patterns and a similarity measure. In this paper, we tailor a robust kernel for image applications. We introduce a new kernel for the kernel methods by integrating the robust error function together with a transformation invariant distance measure. The proposed kernel can be used in formulating the nonlinear variants of linear algorithms (e.g. SVM, PCA) that can be cast in terms of dot products, and thus enhance their robustness. In the experiments, by incorporating with support vector machines (SVM) and linear discriminant analysis (LDA), the derived kernel demonstrated the superior robustness to the conventional kernels in resisting noise corruption and image deformation.

The rest of this paper is organized as follows. In section 2, we present a brief review of kernels. Subsequently, section 3 provides a detailed description of the proposed robust kernel. In section 4, the robustness of the proposed kernel is justified by some computer vision applications. Finally, discussions and remarks are presented in section 5.

2. REVIEW OF KERNELS

The limited power of linear learning machines has been pointed out in [3]. To make target functions better-represented by the given attributes, one can change the data representations via a set of mapping functions. The data representation can be changed via a set of mapping functions Φ:

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

A kernel function k is employed as inner products of images under a transformation Φ of two data points **x** and **x'** in F:

$$k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle = \sum_{i=1}^{n} \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')\,. \tag{2}$$

A function k is a kernel if and only if it is symmetric and the associated kernel matrix (i.e. the Gram matrix) formed by subsets of input space X are non-negative; Mercer’s theorem gives the conditions for k to be a kernel. Different choices of k determine the type of learning machines that is constructed, where conventional choices include polynomial

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

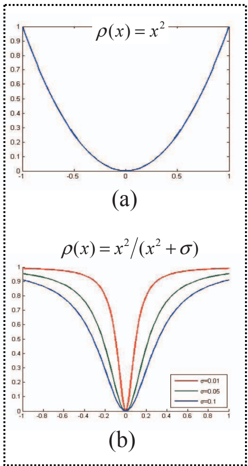


Fig. 1: (a) The squared error function **(b)** the Geman-McClure error function with different σ.

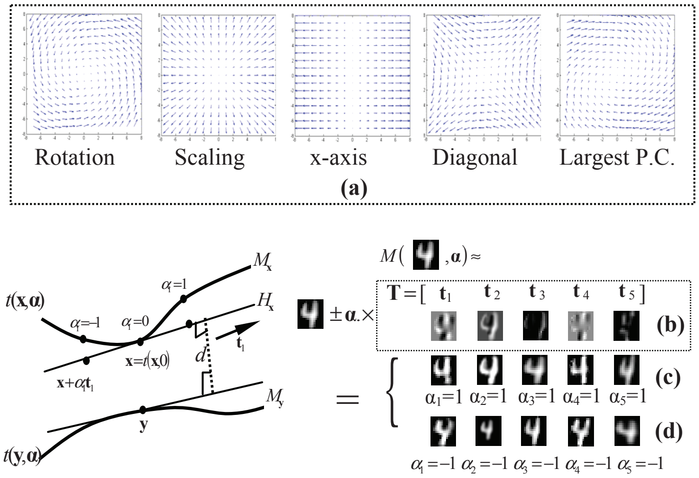


Fig. 2: Illustration of the 5 example tangent vectors. **(a)** The corresponding pixel displacements shown in vector field, **(b)** visualization of 5 tangent vectors, and normalized to 0-255 for representation, and **(c)&(d)** give the transformation results.

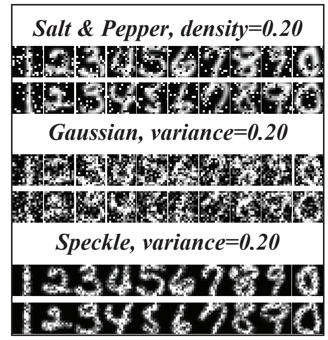


Fig. 3: Some example images of noise-corrupted digits in USPS database for Salt & Pepper noise, Gaussian noise and speckle noise.

kernel $k(\mathbf{x},\mathbf{y})=(\mathbf{x}^T\mathbf{y}+c)^p$, sigmoid kernel $k(\mathbf{x},\mathbf{y})=\tanh(\kappa(\mathbf{x}^T\mathbf{y})-\theta)$, and RBF kernel $k(\mathbf{x},\mathbf{y})=\exp\left(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2\right)$. By embedding kernels into the linear learning machines, many successful frameworks in various applications have been developed.

3. THE PROPOSED ROBUST KERNEL

This paper is focused on establishing a new robust kernel for image-related applications. It concurrently uses a robust ρ-function as well as the notion of tangent planes to map data to another feature space. In the mapped space, it is insensitive to pattern deformation and noise influence is suppressed. In the following, we detail the proposed robust kernel.

Given a dataset X, a kernel can be defined with a function f on X using a metric d, and the Gram matrix K is given by $K_{ij} := f(d(\mathbf{x}_i,\mathbf{x}_j))$. For example, the RBF kernel is given by $k(\mathbf{x},\mathbf{y})=\exp\left(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2\right)$, where σ is a kernel parameter and the associated metric is the *Euclidean distance*. However, one drawback of using the squared error function is known to lack of robustness in the presence of outliers. As shown in Figure 1(a), the squared error measure used in the RBF kernel grows quadratically, thus it can be strongly influenced by a small number of outlier components. Researchers in robust statistics have proposed to employ different ρ-functions to replace the squared error function, thus alleviating the influence of outliers. Figure 1(b) illustrates the Geman-McClure ρ-function [7], where the influence of outlier data points is reduced by setting the error of an outlier component saturated.

We integrate the notion of tangent distance [5] into the Geman-McClure error function in our kernel design so make it concurrently robust to geometric deformation. Suppose there are r different types of transformation in X, such as translation and rotation. In a sense, we can align two patterns (say, **x** and **y**) via finding the shortest distance between two tangent planes corresponding to **x** and **y**. The first-order Taylor expansion is used to approximate r deformation at a pattern x':

$$M(\mathbf{x}',\boldsymbol{\alpha})=M(\mathbf{x}',\mathbf{0})+\sum_{i=1}^{r}\alpha_i\frac{\partial M(\mathbf{x}',0)}{\partial\alpha_i}+O(\alpha_i^2)\ , \quad (3)$$

$$\approx \mathbf{x}'+\sum_{i=1}^{r}\alpha_i\mathbf{t}_i,\ i=1,2,\ldots,\ r$$

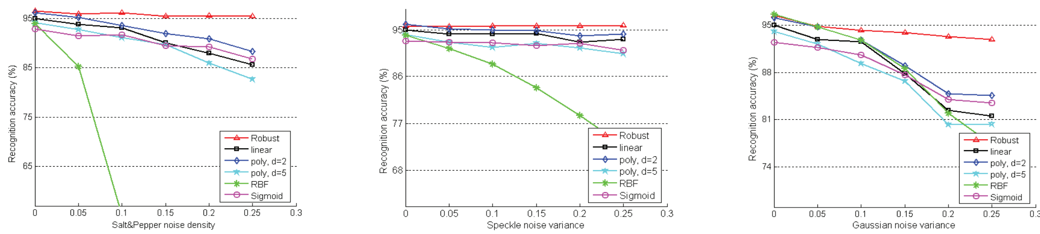
where $M_i(\mathbf{x}',\alpha_i)$ represents the image described by $\mathbf{x}'$ with i-th type of transformation parameterized by α_i, and $M_i(\mathbf{x}',0)=\mathbf{x}'$. Let $\mathbf{T}=[\mathbf{t}_1\ \mathbf{t}_2\ \ldots\ \mathbf{t}_r]$ be a r-by-d matrix containing the r tangent vectors for $\mathbf{x}'$. In this manner, the surface of r possible transformation of a pattern is approximated by its tangent plane at the pattern. Thus, the problem is reduced from computing the similarity measure between two patterns to finding the shortest distance between two tangent planes. We illustrate the idea of tangent distance in Fig. 2.

By embedding the two-sided tangent distance [5] into the residual error part of the ρ-function, our robust kernel is defined as follows:

$$K(\mathbf{x},\mathbf{y})=1-\rho\left(\mathbf{r}_T(\mathbf{x},\mathbf{y});\sigma\right), \quad (4)$$

where

$$\mathbf{r}_T(\mathbf{x},\mathbf{y})=\mathbf{x}+\mathbf{T}_x\hat{\boldsymbol{\alpha}}_x-\mathbf{y}-\mathbf{T}_y\hat{\boldsymbol{\alpha}}_y \quad (5)$$



(a) **(b)** **(c)**

Fig. 4: Comparison of other methods using USPS handwritten digit dataset **(a)** Salt & Pepper noise **(b)** Speckle noise, and **(c)** Gaussian noise with various noise density and variance

with $\hat{\boldsymbol{\alpha}}_x$ and $\hat{\boldsymbol{\alpha}}_y$ obtained by minimizing the distance $\left\|\mathbf{x}+\mathbf{T}_x\boldsymbol{\alpha}_x-\mathbf{y}-\mathbf{T}_y\boldsymbol{\alpha}_y\right\|$, and

$$\rho(\mathbf{r};\sigma)=\frac{1}{n}\sum_{i=1}^{n}\frac{r_i^2}{r_i^2+\sigma^2}\cdot \tag{6}$$

Here, $\mathbf{x}$ and $\mathbf{y}$ denote two data vectors, n is the vector dimension, and T_x and T_y denote the tangent spaces corresponding to $\mathbf{x}$ and $\mathbf{y}$ respectively. σ is the kernel parameter. Note that the above ρ-function we applied here is the Geman and McClure function.

4. EXPERIMENTAL RESULTS

4.1. Hand-Written Digits Recognition

In the experiment, we apply the proposed robust kernel for handwritten digit recognition. The USPS hand-written digits database [8] was applied here and three types of noises are simulated for evaluating the robustness of the proposed kernel. The applied types of noise are additive Gaussian noise, additive salt and pepper noise, and the multiplicative speckle noise. Some examples of the corrupted digits images are shown in Figure 3. In the experiment, we assume five digit transformation (i.e. r=5) to describe the pattern deformation, which are rotation, scaling, and skewing the digits on the directions of x-axis, diagonal direction, and the largest principal component from the training data. These five transformations are depicted in Figure 2. Here, we choose SVM as the kernelized linear machine for its reported good generalization ability.

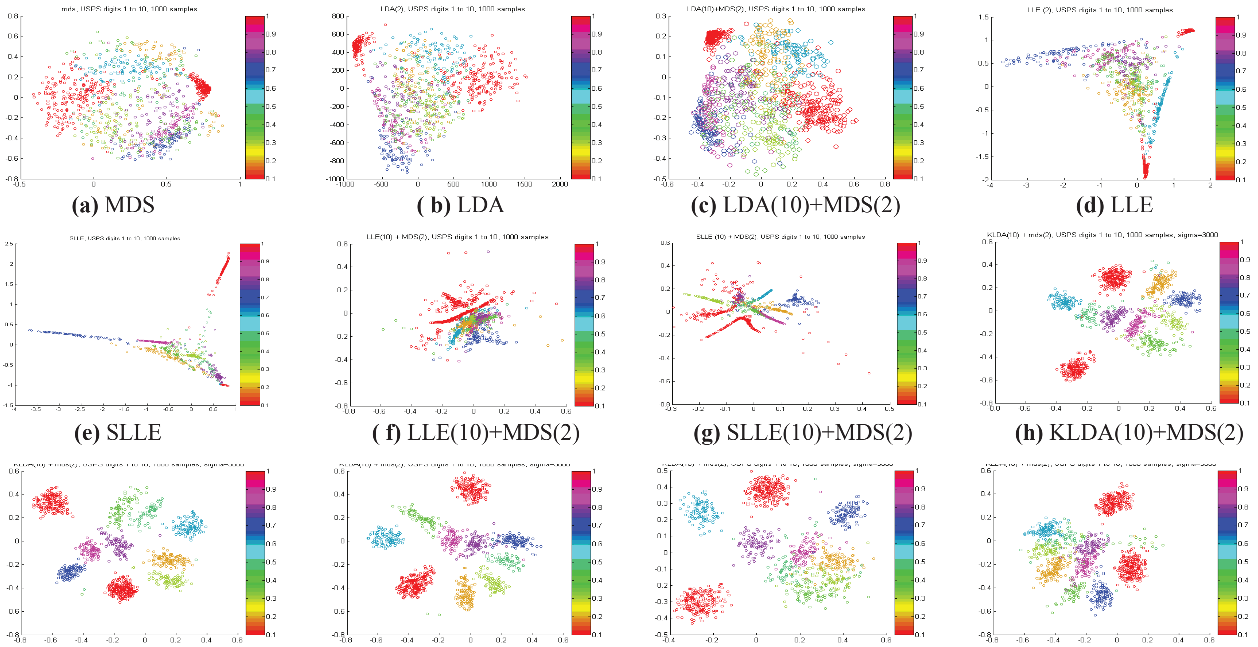
For computational efficiency, in the experiments we randomly chose 1000 samples for SVM training and randomly pick another 100 samples for testing using our proposed kernel. The recognition accuracy was obtained by averaging 11 repetitions. The performance of several frequently-used kernels is evaluated as well for comparison. In Figure 4, ***Robust*** denotes that SVM incorporates with our robust kernel, ***linear*** indicates SVM is with linear kernel $k(\mathbf{x},\mathbf{y})=(\mathbf{x}\bullet\mathbf{y})$, ***poly d*** stands for the polynomial kernel with degree d, ***sigmoid*** for the sigmoid kernel, and ***RBF*** is the RBF kernel. From the figure, we see the digit recognition accuracy is degraded by different degrees when the traditional kernels are used with SVM. On the contrary, when SVM is used with the proposed robust kernel, the classification accuracy becomes stable, and noise corruption only has little effect in the recognition results. Also, we notice that, even in a noise-free condition, the proposed robust kernel still performs slightly better than those general kernels. These results demonstrate the superiority of the proposed kernel by means of robustness against noise over several frequently-used kernels.

4.2. Data Visualization

Solving classification problems typically involves analyzing a large amount of high-dimensional data. However, the high dimensionality of data often complicates the problem, and leads to *the curse of dimensionality*. From the view point of data representation, discovering the compact representation of high dimensional data often helps in many fields. Several data dimensionality reduction frameworks are proposed such as PCA, LDA, and those finding the low-dimensional embeddings such as Multidimensional Scaling (MDS) and Locally Linear Embedding (LLE). However, errors caused by outliers are often inevitable when discovering the nonlinear data relations in a high dimensional space. To investigate the potential of our proposed kernel, here, we apply it to kernelize LDA (KLDA) for better data representation in low dimension.

The set of hand-written digits in Figure 5 is visualized by the procedures: First, with respect to each digit type, we used KLDA to seek a binary Fisher-Discriminant in the mapped space by one-against-others approach. Each hand-written digit image can be represented as a ten-dimensional vector by concatenating ten projections. Next, we apply MDS to reduce the dimensionality from ten to two for visualizing these vectors.

In the experiments, 1000 randomly-selected samples from USPS database are used for visualization. We compared our method with different data dimensionality reduction methods of MDS, LDA, LLE, Supervised LLE (SLLE), and their combinations. As shown in Figure 5, we see the data



(a) MDS **(b)** LDA **(c)** LDA(10)+MDS(2) **(d)** LLE

(e) SLLE **(f)** LLE(10)+MDS(2) **(g)** SLLE(10)+MDS(2) **(h)** KLDA(10)+MDS(2)

(i) Salt&Pepper, density=0.1 **(j)** Salt&Pepper, density=0.2 **(k)** Speckle, variance=0.1 **(l)** Speckle, variance=0.2

Fig. 5: Data dimensionality reduction results by 1000 handwritten digit images. **(a-g)** Results of different methods, where *MA*(d_1)+*MB*(d_2) denotes data dimensionality is reduced to d_1 by Method A then to d_2 by Method B **(h)** our method **(i-l)** under various degree of noise effects

instances after applying our method are nicely distributed according to their categories. From the perspective of data dimensionality reduction (reduced from 16x16 to 2), the method applying our robust kernel has shown outstanding capability for clearly separating digits according to their categories. From the perspective of data visualization, our method clearly reveals the relationship of digits category by category. On the contrary, for other methods the exhibition is not satisfactory. We further added different types of noises to the dataset to verify the robustness of the proposed method. The results are shown in Figure 5(**i-l**). With the capability to robustly measure data similarity, our algorithm retains consistency against noise disturbance. The proposed robust kernel method thus provides good capability in understanding the intrinsic data structures even under a severe noisy condition. Compared to the traditional methods the experiments confirmed the superior performance of the robust kernel in data visualization application.

5. CONCLUSIONS

In this paper, we presented a novel robust kernel that integrates a robust error function and a transformation invariant distance measure, thus making the kernel-based classifier or the dimension-reduction method highly robust against noise and irrelevant data transformation. In the experiments, the robustness is justified on applications in pattern recognition and data visualization under various noise disturbances. In the future, we plan to investigate the efficiency of our robust kernel on other applications, such as face recognition with occlusions and image-denoising by kernelized PCA.

ACKLEDGEMENTS

This work was partially supported by National Science Council, Taiwan, under the grant NSC 97-2218-E-007 -011.

REFERENCES

[1] Huttenlocher, D., Klanderman, G. and Rucklidge, W.: Comparing Images Using the Hausdorff Distance, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9, (1993) 850–863

[2] Mori, G., Belongie, S. and Malik, J.: Efficient Shape Matching Using Shape Contexts, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 27, No. 11, (2005) 1832- 1837.

[3] Scholkopf, B. and Smola, A.: Learning with Kernels, MIT Press (2002)

[4] Lu, C., Zhang, T., Zhang, R., and Zhang, C.: Adaptive Robust Kernel PCA Algorithm, Proc. of ICASSP (2003) 621-624

[5] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri.: Transformation Invariance in Pattern Recognition-Tangent Distance and Tangent Propagation, In *LNCS, 1524*, pp. 239–274. Springer, 1998

[6] Liao, C.T. and Lai, S.-H.: A Robust Kernel Based on Robust ρ-Function, Proc. ICASSP(2007) CD-ROM

[7] Geman, S. and McClure, D. E.: Statistical Methods for Tomographic Image Reconstruction, Bulletin of the Int. Statistical Institute, Vol. 52 (1987) 5-21

[8] http://www.kernel-machines.org/data.html