# A ROBUST KERNEL BASED ON ROBUST ρ-FUNCTION

*Chia-Te Liao and Shang-Hong Lai*

Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan 300
{dr928317, lai}@cs.nthu.edu.tw

## ABSTRACT

Noise-resistance capability is a very important issue to signal processing systems as well as machine learning applications. In this work, we present a new kernel that is highly robust against outliers and random noises. By incorporating a robust ρ-function into the distance metric, the derived robust kernel was shown to be very insensitive to the influence of outlier elements. In the experiments, we show that the proposed kernel brought significant improvement to the Support Vector Machines (SVM) classifier in face recognition accuracy and outperformed several traditional kernels for corrupted data. We also applied our kernel to the kernel Principal Component Analysis (PCA) and evaluate the efficiency in recovering contaminated face images. Experiments show our robust kernel also brings benefits in noise-reduction applications.

***Index Terms*—** Kernels, robustness, SVM, PCA

## 1. INTRODUCTION

Kernel methods are known to represent complex (nonlinear) relationships in the feature space, and have been employed to provide linear strategies with non-linear capabilities by computing data similarities in a more representative feature space [1]. Researchers have developed new kernels that are dedicated to specific applications, such as the string kernels for text categorization [2], or the pyramid kernel for image classification [3]. Yet, since most of learning strategies are to synthesize solutions from given training examples, they suffer drawbacks to be sensitive to the corrupted data, either in training or testing. Taking face recognition for example, facial images are often corrupted by the low-quality image acquisition equipments with various kinds of noise involved. In the meantime, corrupted face images are very possible to become the critical instances during the training procedure, so they are usually very difficult to be correctly classified. In this work, motivated by the robust error function used in robust statistics, we developed a new kernel which is very robust to resist different high-level noise corruption in data. Briefly, our proposed kernel is a weighted linear combination of two functions: one is a robust ρ-function borrowed from robust statistics and the other is the Radial Basis Function (RBF) kernel function. The proposed kernel is shown to satisfy the Mercer's condition. Being a kernel, it can be used in conjunction with several kinds of linear learning strategies and provides them with the nonlinear abilities, including SVM, PCA, as well as Linear Discriminant Analysis (LDA).

The rest of this paper is organized as follows: First we briefly reviewed the backgrounds of kernel tricks in section 2. Subsequently, section 3 provides a detailed description of the proposed robust kernel based on a robust error function. Some experimental results on face recognition from noisy images are shown in section 4 to give a significant improvement of robustness against different types of noises by using our kernel. In addition, we also show the results of applying this robust kernel to the kernel PCA for image noise reduction. Finally, we conclude this paper with some remarks and discussions in section 5.

## 2. BACKGROUND OF KERNELS

The limited power of linear learning machines was pointed out in 1960s [4]. In the real applications, we are frequently not able to directly express the target function simply by a linear combination of the given attributes. A common strategy in machine learning society is to change the data representation by

$$\mathbf{x} = (x_1, x_2, ..., x_d) \mapsto \Phi(\mathbf{x}) = \left(\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), ..., \Phi_n(\mathbf{x})\right). \quad (1)$$

Kernel-based learning algorithms are based on the idea of projecting the feature vector $\mathbf{x}$ from the original space $X \subseteq R^d$ to a Reproducing Kernel Hilbert Space (RKHS) H. The computational power of linear learning machines is thus increased with computing linear relationships in the projected, much higher dimensional feature space. Having the mapping function $\Phi$, data point $\mathbf{x}$ is embedded into H with the mapping $\Phi(\mathbf{x})$ Therefore, we can measure the data similarity with the kernel function $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$, where $K : X \times X \to R$ . Mercer's theorem provides the characterization that $K$ is a kernel, if and only if the Gram matrix

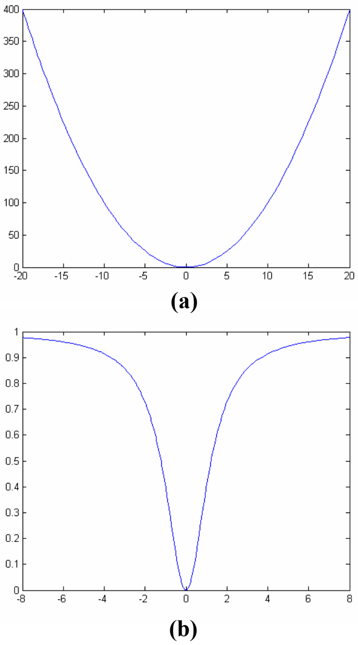



**Figure 1:** **(a)** The squared error function and **(b)** the robust Geman-McClure error function [6].

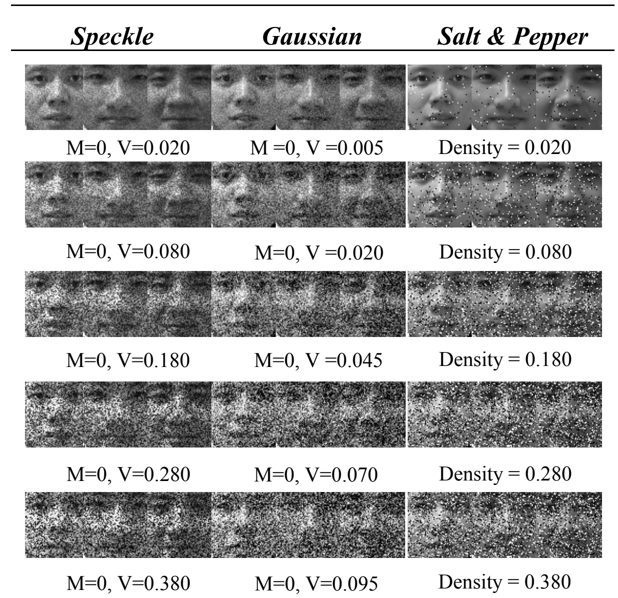



**Figure 2:** Some example of noise-corrupted face images used in our experiments. M and V stand for mean and variance, respectively.

$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ is non-negative. Note that the linear learning machines, such as PCA and LDA, can be included into H since the associated feature vectors can be computed with inner products. The most well-known example that the kernel tricks are employed in conjunction with linear strategies is SVM. By looking for the optimal separating hyperplane (OSH) in kernel space, SVM obtains the maximal generalization capability and achieves many successful results in different fields.

## 3. PROPOSED ROBUST KERNEL

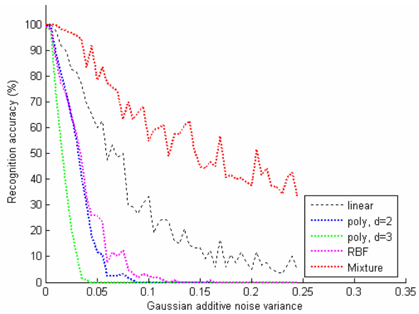
We describe the details of the proposed robust kernel in this section. For a given RBF kernel function $f$ and a set of training data $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_l, y_l)\} \subseteq (X \times Y)^l$, we can define a Gram matrix $K$ given by $K_{ij} := f(d(\mathbf{x}_i, \mathbf{x}_j))$, where $d$ is a metric on $X$. The above kernel function is well-known since it forms the core of a radial basis function network. A popular example is the RBF kernel with the metric defined by the inner product $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle}$ together with the exponential function $f$, i.e. $K_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)$. Given the assumption that none of all pairs of two training sample $\mathbf{x}_i$ and $\mathbf{x}_j$ are the same, the RBF kernel can span an infinite dimensional feature space H since there is no restriction on the total number of training samples. In other words, the hypothesis space function class F has an infinite VC dimension. Also, since the Gram matrix $K$ is of full rank and every mapped vector $\Phi(\mathbf{x}_i)$ in H is linearly independent with the other mapped feature vectors [5], it turns out that this kind of kernel is very sensitive to the high-level noise corruption in the data. In Figure 1, seeing that the sum of squared differences in RBF function is strongly influenced by large differences between corresponding feature vector components, therefore outliers or large noises can easily dominate the kernel function value evaluation. To alleviate the influence, researchers in robust statistics proposed to replace the square error function by a robust ρ-function, which is more tolerant to the outlier and noise in the data. This concept inspires our development of robust kernel based on the robust ρ-function [6], which makes the influence of outlier elements saturated. Figure 1 depicts the square error function and the Geman-McClure ρ-function [6]. The proposed robust kernel function is given by the following equation:

$$K_{robust}(\mathbf{x}_i, \mathbf{x}_j) := K_\rho(\mathbf{x}_i, \mathbf{x}_j) + \alpha K_0(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

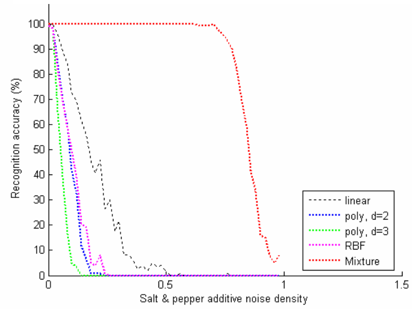
where

$$K_\rho(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\sum_{k=1}^{m} (\mathbf{x}_i^k - \mathbf{x}_j^k)^2 \Big/ \left((\mathbf{x}_i^k - \mathbf{x}_j^k)^2 + 2\sigma^2\right)\right), \tag{3}$$

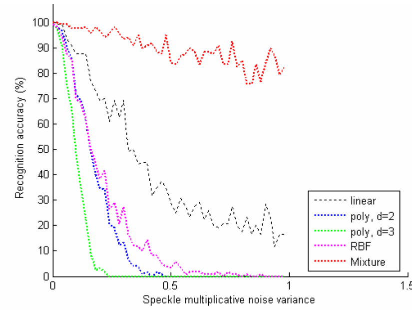
$$K_0(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right). \tag{4}$$



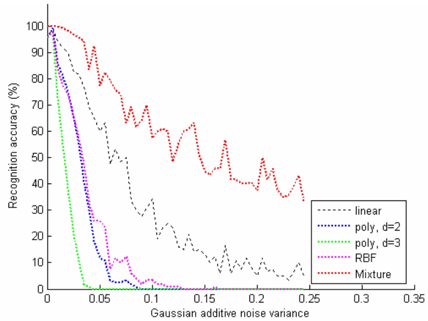

**(a)** C-SVM, additive Gaussian noises



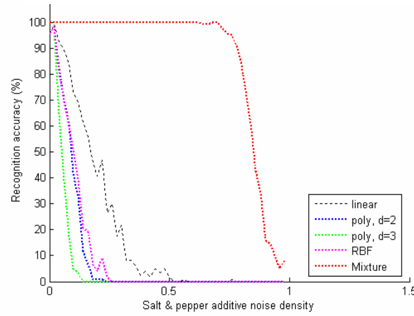

**(b)** C-SVM, additive salt and pepper noises



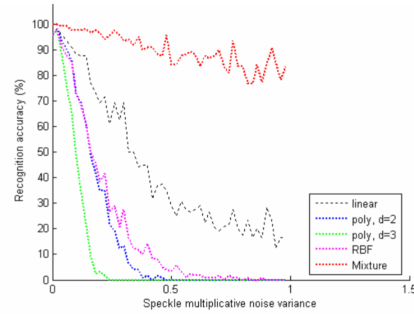

**(c)** C-SVM, multiplicative speckle noises




**(d)** ν-SVM, additive Gaussian noises




**(e)** ν-SVM, additive salt and pepper noises




**(f)** ν-SVM, multiplicative speckle noises

**Figure 3:** Face recognition accuracy by using **(a-c)** C-SVM and **(d-f)** ν-SVM with various types of kernels under different level noises.

**Figure 4:** Kernel PCA noise reduction with our robust kernel. The $1^{st}$, $3^{rd}$ and $5^{th}$ rows are samples contaminated by Gaussian noise with variance ranging from 0.001 to 0.010, while the $2^{nd}$, $4^{th}$ and $6^{th}$ rows are the recovered images after kernel PCA.

Here, $m$ is the dimension of the feature vector and $\alpha$ is a weighting parameter to compromise between the two functions $K_\rho$ and $K_0$. Note that $K_\rho$ is derived from the robust ρ-function to achieve robust comparison between feature components and $K_0$ is the Gaussian RBF kernel included to make the composite function satisfy Mercer's condition. The mixture weight $\alpha$ between these $K_\rho$ and $K_0$ is selected to satisfy the Mercer's condition; let the minimal eigenvalues for the Gram matrices $K_\rho$ and $K_0$ be denoted by $\lambda_{\min}(K_\rho)$ and $\lambda_{\min}(K_0)$, respectively. Then, the minimal eigenvalue of $K_{robust}$ is bounded by:

$$\lambda_{\min}(K_{robust}) \geq \lambda_{\min}(K_\rho) + \alpha\lambda_{\min}(K_0). \tag{5}$$

Since the Gram matrix for the Gaussian RBF kernel is positive-definite, we have $\lambda_{\min}(K_0) > 0$. By setting $\alpha$ to be $-\lambda_{\min}(K_\rho)/\lambda_{\min}(K_0)$, we can assure $K_{robust}$ to be non-negative and satisfy the Mercer's condition.

## 4. EXPERIMENTAL RESULTS

### 4.1. Face recognition with SVM

To validate the noise-resistance capability of the proposed kernel, in the experiments, we simulated several types of noises on face images and evaluated the face recognition accuracy. The performance of the proposed robust kernel is compared with several types of traditional kernels by using SVM as the classifier. The applied noise types are additive Gaussian noise, additive salt and pepper noise, and the multiplicative speckle noise. For the Gaussian noise, the original face image I was perturbed by $I' = I + N(0, v)$ where *N* is a normal distribution random variable with zero mean and variance *v*. The variance is ranged from [0, 0.25] in the experiments. For the case of salt and pepper noise, on and off pixels were randomly added into the original image *I* with density *d* sampled between [0, 0.98]. Next face images are corrupted by speckle multiplicative noises by $I' = I + n \times I$, where *n* is a uniform distributed random variable with zero mean and variance [0, 0.98]. Pixels here were scaled into [0, 1]; noise corruptions are applied to both training and testing images. Some examples of the noise-corrupted face images of different noise levels are shown in Figure 2.

The results of face recognition accuracy by using SVM classifier together with various kernels are shown in Figure 3. Note that there are totally 20 categories (persons) in the face database with 6 instances each category. In the figures, we experimented on two forms of SVM classifiers, i.e. C-SVM and ν-SVM, in conjunction with several kernels, including the proposed robust kernel and some standard kernels, for face recognition. The label ***linear*** denotes SVM classifier with linear kernel is used, ***poly d=n*** stand for the polynomial kernel with degree n, ***RBF*** denotes the Gaussian RBF kernel, while ***Mixture*** denotes our proposed robust kernel respectively. None of instances of test face images ever present in the training set. We also notice that the weighting coefficient α is mostly extremely small, which implies the second part of $K_0$ is insignificant. In the experiments our proposed robust kernel always provides the best face recognition accuracy compared to other traditional kernels in both C-SVM and ν-SVM classifiers.

### 4.2. Noise-reduction

The basic idea of kernel PCA is to nonlinearly map the data into the RKHS F and then perform linear PCA in F; it is equivalent to doing nonlinear PCA in the original input space. Here, we applied our robust kernel to kernel PCA for the image-denoising problem. Some related works can be found in [7]. In this experiment, we used facial images of 67 persons with 3 different illuminations, taken from CMUPIE database. Due to the memory limitation in computing kernel matrix, images were normalized to 32 x 32. In addition to the original illumination changes, in the experiments, we further added random Gaussian noises with variance [0.001-0.010] to the face images; and there are totally 2010 faces with kernel size 2010 x 2010. By solving the pre-image problem [7], the contaminated face images were recovered with better quality. We demonstrated some resulting images in Figure 4, and we can see the images after our kernel PCA noise reduction are much clearer, which means our kernel is quite insensitive to noise effects.

## 5. CONCLUSIONS

In this paper, we developed a new and robust kernel. The proposed robust kernel function is based on using the robust ρ-function to alleviate the influence of outliers and high noises. For face recognition applications, the proposed robust kernel provided much more accurate recognition results compared to other traditional kernels in conjunction with the SVM classifier under very noisy environment. For image-denoising problem, by applying our noise-insensitive robust kernel to the kernel PCA, we can recover more clear images from noisy face images. In the future, it would be interesting to further investigate the applications of the proposed robust kernel on other problems under different noisy environments, especially to incorporate it with other learning algorithms.

## 6. REFERENCES


[1] Scholkopf, B. and Smola, A., *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[2] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” *Proceedings of the European Conference on Machine Learning*, Berlin Germany, pp. 137–142, 1998.

[3] K. Grauman, T. Darrell, “The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features,” *Proceedings of IEEE International Conference on Computer Vision*, Beijing, China, pp. 1458-1465, 2005.

[4] Minsky, M. L. and Papert, S. A., *Perceptrons*, MIT Press, Cambridge, MA, 1969.

[5] C. A. Micchelli, “Algebraic Aspects of Interpolation,” *Proceedings of Symposium in Applied Mathematics*, pp. 81-102, 1986.

[6] S. German and D. E. McClure, “Statistical Methods for Tomographic Image Reconstruction,” *Bulletin of the International Statistical Institute.* LII-4:5, 1987.

[7] J. T. Kwok and I. W. Tsang, "The Pre-Image Problem in Kernel Methods," *IEEE Transactions on Neural Networks,* VOL. 15, NO. 6, pp. 1517-1525, 2004.