KERNEL WIENER FILTER WITH DISTANCE CONSTRAINT

[1,2]*Makoto Yamada and* [3]*Mahmood R. Azimi-Sadjadi*

[1]Transportation Systems Division, Hitachi Ltd., Ichige 1070, Hitachinaka, Ibaraki, 312-8506, Japan
[2]Department of Statistical Science, The Graduate University for Advanced Studies,
4-6-7 Mimami-Azabu, Minato-ku, Tokyo, 106-8569 Japan
[3]Dept. of Electrical and Computer Engr., Colorado State University, Fort Collins, CO, 80523, U.S.A
email: myamada@ism.ac.jp

ABSTRACT

In this paper, we introduce a non-iterative nonlinear kernel Wiener filtering method using kernel Canonical Correlation Analysis (CCA) framework. This approach is based upon the theory of reproducing kernel Hilbert spaces. A method is proposed to find approximate Wiener filtered signal in the original signal space by solving an optimization problem in the higher dimensional space. Unlike the conventional iterative approaches which rely on nonlinear optimization problem, our proposed method directly finds the pre-image using distance constraints in the higher mapped domain. The signal estimation and reconstruction capability of the new method is demonstrated and benchmarked on the United States Postal Service (USPS) digits database. Moreover, a comparison with the conventional kernel Wiener filter is presented.

Keywords: Kernel Wiener Filter, Canonical Correlation Analysis, Nonlinear Signal Estimation, Distance constraint

1. INTRODUCTION

In the kernel Wiener filter, the filtering is applied in the higher dimensional mapped space where the original signals are typically unknown. To solve this problem, one may look for the signal in the original signal space, referred to as "pre-image" [1, 2], by minimizing the estimation error in the higher dimensional space. In [1], the problem of kernel Wiener filter using kernel Canonical Correlation Analysis (CCA) framework was addressed. The reduced-rank Wiener filter problem in the higher dimensional mapped domain was solved by using the *kernel trick*. A method was proposed to find approximate Wiener filtered signal in the original space by solving an optimization problem in the higher dimensional space. The mean squared error (MSE) between the mapped data and the reconstructed data using kernel Half CCA (HCCA), which is important for reduced-rank estimation, was minimized with respect to a pre-image. However, the results in [1] on image restoration and reconstruction showed that this solution for pre-image typically leads to poor reconstruction and signal restoration though the noise outside the signal/image is significantly reduced. Moreover, the approach have the convergence problem, since the final solution is depended on the iterative approach.

In an attempt to alleviate this problem, we follow the approach proposed in [3]. In this reference, the authors find the pre-images of kernel Principal Component Analysis (PCA) using a simple linear algebra-based approach that does not suffer from convergence and local minima problems. The Half-CCA (HCCA) method in [1] is adopted here in order to obtain reduced rank kernel Wiener filtered pre-images, which correspond to the mapped Wiener filtered signals in the two-channel CCA framework. The pre-images are obtained non-iteratively using a simple singular value decomposition (SVD)-based approach.

In the following sections, we will first derive the kernel version of the HCCA for reduced rank kernel Wiener filtering. The problem of finding the kernel Wiener filter pre-image is then cast into a high dimensional optimization problem, and the solution of which is implicitly obtained in the lower dimensional signal space by utilizing linear algebra. The kernel Wiener filter results are experimentally compared to the conventional kernel Wiener filter on United States Postal Service (USPS [1]) digits data set.

2. KERNEL HCCA AND RELATION TO KERNEL WIENER FILTER

Let $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^p$ be two random vectors and $\boldsymbol{\phi}(\cdot) : \mathbb{R}^m \longrightarrow \mathbb{R}^{m'}$ and $\boldsymbol{\psi}(\cdot) : \mathbb{R}^p \longrightarrow \mathbb{R}^{p'}$ be the corresponding nonlinear mapping functions that map $\mathbf{x}$ and $\mathbf{y}$ into the higher dimensional space with $m \ll m'$ and $p \ll p'$ where $m' \leq p'$. Thus, the mapped vectors are given by $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{m'}$ and $\boldsymbol{\psi}(\mathbf{y}) \in \mathbb{R}^{p'}$, respectively. Now, let us assume that $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{y})$ are zero mean vectors with covariance matrices $\mathrm{R}_{\phi\phi} \in \mathbb{R}^{m' \times m'}$ and $\mathrm{R}_{\psi\psi} \in \mathbb{R}^{p' \times p'}$ and cross-covariance matrix $\mathrm{R}_{\phi\psi} = \mathrm{R}_{\psi\phi}^T \in \mathbb{R}^{m' \times p'}$. The mapping matrices of HCCA are obtained by finding an SVD of the half coherence matrix

[1]The United States Postal Service data set is downloadable at "http://www.kernel-machines.org/"

$\mathrm{C} = \mathrm{R}_{\phi\psi}\mathrm{R}_{\psi\psi}^{-T/2} \in \mathbb{R}^{m' \times p'}$ [1]:

$$\begin{aligned} \mathrm{C} = \mathrm{F}\Lambda\mathrm{G}^T \quad \text{or} \quad \mathrm{F}^T\mathrm{C}\mathrm{G} = \Lambda, \\ \text{with} \quad \mathrm{F}^T\mathrm{F} = \mathrm{I}, \;\; \mathrm{G}^T\mathrm{G} = \mathrm{I}, \\ \Lambda = \left[\begin{array}{cc} \Lambda_{m'} & \mathbf{0} \end{array} \right]; \quad \Lambda_{m'} = diag[\lambda_1, \ldots, \lambda_{m'}] \end{aligned} \tag{1}$$

where $\mathrm{F} \in \mathbb{R}^{m' \times m'}$ and $\mathrm{G} \in \mathbb{R}^{p' \times p'}$ are the orthogonal canonical mapping matrices that map $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{y})$ to their half canonical coordinates and $\Lambda \in \mathbb{R}^{m' \times p'}$ is the canonical correlation matrix with elements that measure the coherence between the individual half canonical coordinates.

It has been shown [1] that the canonical coordinate mappings and the corresponding canonical correlation matrix can be found by solving a coupled generalized eigenvalue problem,

$$\mathrm{R}_{\phi\psi}\mathrm{D}_{\psi} = \mathrm{D}_{\phi}\Lambda \tag{2}$$

$$\mathrm{R}_{\psi\phi}\mathrm{D}_{\phi} = \mathrm{R}_{\psi\psi}\mathrm{D}_{\psi}\Lambda^T \tag{3}$$

where $\mathrm{D}_{\phi} = \mathrm{F}$ and $\mathrm{D}_{\psi} = \mathrm{R}_{\psi\psi}^{-T/2}\mathrm{G}$ are the relevant HCCA mapping matrices.

In practice, however, the covariance matrices are estimated from samples of the data. Assume that the sample data matrices $\mathrm{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathrm{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ are observed where $\mathbf{x}_i$, $\mathbf{y}_i$ $i \in [1,\ n]$ are the ith realizations of the two channel processes. Now, define the mapped sample matrices Φ and Ψ as

$$\Phi = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)] \in \mathbb{R}^{m' \times n} \tag{4}$$

$$\Psi = [\boldsymbol{\psi}(\mathbf{y}_1), \ldots, \boldsymbol{\psi}(\mathbf{y}_n)] \in \mathbb{R}^{p' \times n} \tag{5}$$

Also, let $k_{\phi}(\cdot,\cdot)$ and $k_{\psi}(\cdot,\cdot)$ be the inner products Mercer kernels [2] in the implicit spaces $\mathbb{R}^{m'}$ and $\mathbb{R}^{p'}$:

$$k_{\phi}(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^T\phi(\mathbf{y}) \tag{6}$$

$$k_{\psi}(\mathbf{x},\mathbf{y}) = \boldsymbol{\psi}(\mathbf{x})^T\boldsymbol{\psi}(\mathbf{y}) \tag{7}$$

The kernel Gram matrices associated with (6) and (7) are given by

$$\mathrm{K}_{\phi} = \Phi^T\Phi = [k_{\phi}(\mathbf{x}_i,\mathbf{x}_j)] \in \mathbb{R}^{n \times n} \tag{8}$$

$$\mathrm{K}_{\psi} = \Psi^T\Psi = [k_{\psi}(\mathbf{y}_i,\mathbf{y}_j)] \in \mathbb{R}^{n \times n} \tag{9}$$

The kernel Gram matrices K_{ϕ} and K_{ψ} are non-negative definite and have the same rank as the column rank of their corresponding mapped sample data matrices [4].

Using the sample covariance matrices instead of the theoretical ones and utilizing the non-singular kernel Gram matrices (i.e. Gaussian kernel), the generalized eigenvalue problem in (2) and (3) can be cast into the kernelized eigenvalue problem [1] as,

$$\frac{1}{n}\mathrm{K}_{\phi}\hat{\mathrm{D}}_{\phi} = \hat{\mathrm{D}}_{\phi}\Lambda\Lambda^T \tag{10}$$

$$\frac{1}{n}\mathrm{K}_{\phi}\mathrm{K}_{\psi}\hat{\mathrm{D}}_{\psi} = \mathrm{K}_{\psi}\hat{\mathrm{D}}_{\psi}\Lambda^T\Lambda \tag{11}$$

where $\hat{\mathrm{D}}_{\phi}$ and $\hat{\mathrm{D}}_{\psi}$ satisfy $\mathrm{D}_{\phi} = \Phi\hat{\mathrm{D}}_{\phi}$ and $\mathrm{D}_{\psi} = \Psi\hat{\mathrm{D}}_{\psi}$. Note that now the generalized (kernelized version) eigenvalue problem of (10) and (11) for $\hat{\mathrm{D}}_{\phi}$ and $\hat{\mathrm{D}}_{\psi}$ only depend on the kernel Gram matrices K_{ϕ} and K_{ψ}. Consequently, $\hat{\mathrm{D}}_{\phi}$ and $\hat{\mathrm{D}}_{\psi}$ are implicitly obtained without computing the mapped data matrices Φ and Ψ.

Remark:

Thus far, we have assumed that the data in the higher mapped domain have zero mean. In practice, this assumption may not be valid, and hence one needs to zero mean the data. In this case, the following modifications are required.

$$\hat{\mathrm{K}}_{\phi} = \mathrm{J}\mathrm{K}_{\phi}\mathrm{J} \quad \text{and} \quad \hat{\mathrm{K}}_{\psi} = \mathrm{J}\mathrm{K}_{\psi}\mathrm{J} \tag{12}$$

where $\mathrm{J} = (\mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \in \mathbb{R}^{n \times n}$ is called centering matrix, and $\mathbf{1} = [1, \cdots, 1]^T \in \mathbb{R}^n$ is the one vector.

Half canonical coordinates are optimal for computing the reduced rank kernel Wiener filter of $\boldsymbol{\phi}(\mathbf{x})$ from $\boldsymbol{\psi}(\mathbf{y})$, when the objective of estimation is to minimize MSE of the two-channels [5]. The rank-r ($r \leq n$) estimate of $\boldsymbol{\phi}(\mathbf{x})$ from $\boldsymbol{\psi}(\mathbf{y})$ is given as $\boldsymbol{\phi}_r(\hat{\mathbf{x}}) = \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$ [5]. The rank-r estimator H_r and its corresponding error covariance matrix of the filtering error $\mathbf{e}_r = \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}_r(\hat{\mathbf{x}}) = \boldsymbol{\phi}(\mathbf{x}) - \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$ are

$$\mathrm{H}_r = \mathrm{F}_r\Lambda_r\mathrm{G}_r^T\mathrm{R}_{\psi\psi}^{-1/2} \tag{13}$$

$$= \mathrm{D}_{\phi,r}\Lambda_r\mathrm{D}_{\psi,r}^T \tag{14}$$

$$\mathrm{R}_{ee}(r) = \mathrm{R}_{\phi\phi} - \mathrm{F}_r\Lambda_r\Lambda_r^T\mathrm{F}_r^T \tag{15}$$

$$\Lambda_r = diag[\lambda_1, \ldots, \lambda_r] \tag{16}$$

where $\mathrm{F}_r \in \mathbb{R}^{m' \times r}$ and $\mathrm{G}_r \in \mathbb{R}^{p' \times r}$ have the leading r eigenvectors of F and G, $\Lambda_r \in \mathbb{R}^{r \times r}$ contains the first r largest singular values of Λ, and $\mathrm{D}_{\phi,r}$ and $\mathrm{D}_{\psi,r}$ have the first r columns of D_{ϕ} and D_{ψ}.

3. FINDING THE PRE-IMAGE BASED ON DISTANCE CONSTRAINT

In [3], the authors proposed a multi-dimensional scaling (MDS) idea that attempts to find the pre-images that exactly satisfy the input-space distance constraints without any iterative computation. This method is based upon simple linear algebra and offers much better reconstruction and signal/image estimation. Here, we utilize this method to find the reduced rank kernel Wiener filter solutions.

The Euclidean distances between $\mathbf{x}$ and $\hat{\mathbf{x}}$ and between $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\phi}(\hat{\mathbf{x}})$ are given by $d(\mathbf{x},\hat{\mathbf{x}})$ and $d(\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\hat{\mathbf{x}}))$, respectively. Since the distances between the pre-image and training samples i.e $d(\boldsymbol{\phi}(\mathbf{x}_i),\boldsymbol{\phi}(\hat{\mathbf{x}}))\ \ \forall\ i$ are typically related to input space distances $d(\mathbf{x}_i,\hat{\mathbf{x}})$, it is possible to embed the distance structure around a pre-image in the input space to that in the feature space [3].

Let us consider the relationship between the higher mapped distance and the distance in the input space. For a Gaussian

kernel case i.e. $k(\mathbf{x},\mathbf{y}) = \exp(\frac{-||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2})$ where σ is an arbitrary parameter, a distance in feature space is given by

$$\begin{aligned} d(\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\hat{\mathbf{x}})) &= (\boldsymbol{\phi}(\mathbf{x})-\boldsymbol{\phi}(\hat{\mathbf{x}}))^T(\boldsymbol{\phi}(\mathbf{x})-\boldsymbol{\phi}(\hat{\mathbf{x}})) \\ &= 2-2\exp(\frac{-||\mathbf{x}-\hat{\mathbf{x}}||^2}{2\sigma^2}) \end{aligned} \quad (17)$$

and hence, the distance in the input space is written in terms of that in the higher dimensional space i.e.

$$d(\mathbf{x},\hat{\mathbf{x}}) = -2\sigma^2 \log\{\frac{1}{2}\big(2-d(\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\hat{\mathbf{x}}))\big)\} \quad (18)$$

Now, the rank-r Wiener filter estimate of $\boldsymbol{\phi}_r(\hat{\mathbf{x}}) \in \mathbb{R}^{m'}$ from $\boldsymbol{\psi}(\mathbf{y}) \in \mathbb{R}^{p'}$, is given [1] by $\boldsymbol{\phi}(\hat{\mathbf{x}}) = \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$. Moreover, we assume that the rank-r estimate is close to the real solution. Thus, the distance in the higher dimensional mapped space is rewritten as

$$\begin{aligned} d(\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\hat{\mathbf{x}})) &= (\boldsymbol{\phi}(\mathbf{x})-\mathrm{H}_r\boldsymbol{\psi}(\mathbf{y}))^T(\boldsymbol{\phi}(\mathbf{x})-\mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})) \\ &= 1-2\alpha+\tau \end{aligned} \quad (19)$$

where $\alpha = \boldsymbol{\phi}(\mathbf{x})^T\mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$ and $\tau = \boldsymbol{\psi}(\mathbf{y})^T\mathrm{H}_r^T\mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$. The second term in (19) is written as

$$\begin{aligned} \alpha &= \boldsymbol{\phi}(\mathbf{x})^T\mathrm{D}_{\phi,r}\Lambda_r\mathrm{D}_{\psi,r}^T\boldsymbol{\psi}(\mathbf{y}) && (20) \\ &= k_\phi(\mathrm{X},\mathbf{x})^T\hat{\mathrm{D}}_{\phi,r}\Lambda_r\hat{\mathrm{D}}_{\psi,r}^T k_{\boldsymbol{\psi}}(\mathrm{Y},\mathbf{y}) && (21) \end{aligned}$$

For τ in (19), using the orthogonal property $\mathrm{D}_{\phi,r}^T\mathrm{D}_{\phi,r} = \mathrm{I}_r$, $\mathrm{H}_r^T\mathrm{H}_r$ is represented by

$$\begin{aligned} \mathrm{H}_r^T\mathrm{H}_r &= (\mathrm{D}_{\phi,r}\Lambda_r\mathrm{D}_{\psi,r})^T(\mathrm{D}_{\phi,r}\Lambda_r\mathrm{D}_{\psi,r}) \\ &= \mathrm{D}_{\phi,r}^T\Lambda_r^2\mathrm{D}_{\phi,r} \end{aligned} \quad (22)$$

and hence τ is given by

$$\tau = k_\psi(\mathrm{Y},\mathbf{y})^T\hat{\mathrm{D}}_{\psi,r}\Lambda_r^2\hat{\mathrm{D}}_{\psi,r}^T k_{\boldsymbol{\psi}}(\mathrm{Y},\mathbf{y}) \quad (23)$$

The main idea in [3], [6] is to exploit only a limited number of nearest data samples to estimate the pre-image. More specifically, we define the vector $\mathbf{d}_i^2 = [d(\boldsymbol{\phi}(\mathbf{x}_i),\boldsymbol{\phi}(\hat{\mathbf{x}}))]_i$ where $\mathbf{x}_i$ is one of the $\tilde{n}$ closest data points to the pre-image. Now, using these $\tilde{n}$ neighbors that are selected by (19) we form $\tilde{\mathrm{X}} = \{\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_{\tilde{n}}\}$ and compute its SVD after the mean centering operation. That is, compute SVD of $m \times \tilde{n}$ matrix $\tilde{\mathrm{X}}\mathrm{J}$

$$\tilde{\mathrm{X}}\mathrm{J} = \mathrm{U}\Sigma\mathrm{V}^T = \mathrm{U}\mathrm{Z} \quad (24)$$

where J is a centering matrix defined before, U is an orthogonal matrix, and $\mathrm{Z} = \Sigma\mathrm{V}^T$ contains columns that are the projections of $\mathbf{x}_i$ onto the columns of U. Then, the approximate pre-image can be obtained [3] as,

$$\hat{\mathbf{z}} = -\frac{1}{2}\Sigma^{-1}\mathrm{V}^T(\mathbf{d}^2-\mathbf{d}_0^2) \quad (25)$$

where $\mathbf{d}^2 = [d(\boldsymbol{\phi}(\tilde{\mathbf{x}}_1),\boldsymbol{\phi}(\hat{\mathbf{x}})),\ldots,d(\boldsymbol{\phi}(\tilde{\mathbf{x}}_{\tilde{n}}),\boldsymbol{\phi}(\hat{\mathbf{x}}))]^T \in \mathbb{R}^{\tilde{n}}$ is the distance from the $\tilde{n}$ nearest neighbor data samples and $\mathbf{d}_0^2 = [||\mathbf{z}_1||^2,\ldots,||\mathbf{z}_{\tilde{n}}||^2]^T \in \mathbb{R}^{\tilde{n}}$ is the distance from the origin in the input space. Transforming back to the original coordinate system in the input space gives the pre-image $\hat{\mathbf{x}}$

$$\hat{\mathbf{x}} = \mathrm{U}\hat{\mathbf{z}} + \boldsymbol{\mu}_x \quad (26)$$

where $\boldsymbol{\mu}_x$ is the mean of $\tilde{\mathrm{X}}$.

4. RESULTS OF KERNEL WIENER FILTER FOR IMAGE RESTORATION / RETRIEVAL

In this section, the simple USPS data set is used to validate our proposed methods. The USPS data set is 256-dimensional handwritten digits (0 to 9) in the range of $[-1\ ,1]$. We used a Gaussian kernel $k_\phi(\mathbf{x},\mathbf{y}) = k_\psi(\mathbf{x},\mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2})$. The $\mathbf{x}$-channel of kernel HCCA consists of clean USPS images, and the $\mathbf{y}$-channel is the noise corrupted image (additive noise) $\mathbf{y} = \mathbf{x} + \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a white Gaussian noise vector with statistics $(0,\sigma_\eta^2)$. Figures 1 and 2 show the original 100 clean digit images and the corresponding noisy images corrupted by additive white Gaussian noise with SNR=1dB, respectively.

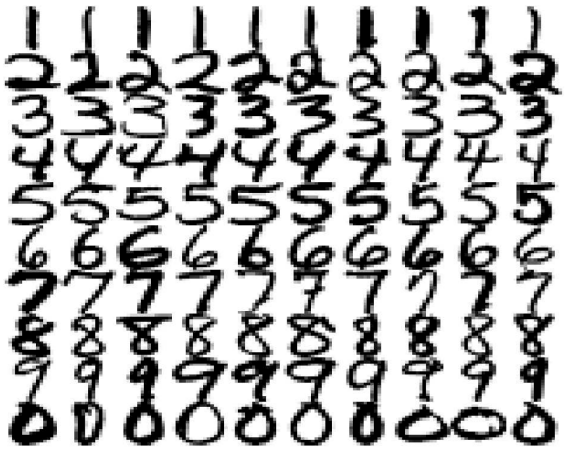

Fig. 1. Original images

We randomly chose 1000 samples for the training and 100 samples for testing. The pre-image $\hat{\mathbf{x}}$ is computed using (26) without any iterative computation as opposed to the conventional approach. Figures 3 and 4 show the reduced rank kernel Wiener filtering results using the proposed method in this paper and the standard kernel Wiener filter, respectively. The parameters are chosen as: $\sigma^2 = 0.7$, $\tilde{n} = 20$, and $r = 300$ and the conventional kernel Wiener filter with parameters $\sigma^2 = 0.053$ and $r = 300$. Comparing these results, it is evident that the proposed method performed much better than the conventional one in terms of its restoration capability. Clearly, the proposed method does not suffer from convergence to local minima problems, which leads to erroneous restoration and

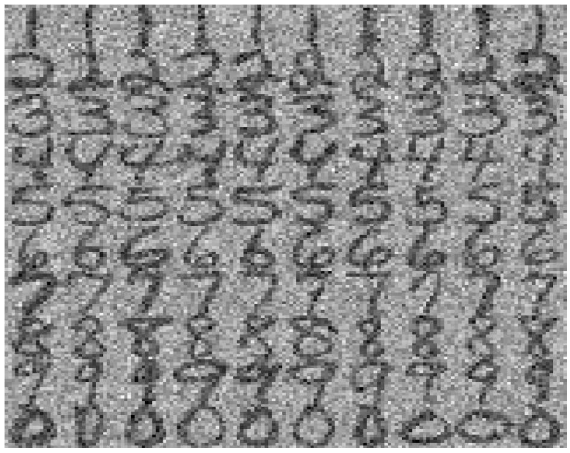

Fig. 2. Noisy Images (SNR = 1dB).

reconstruction as shown in Figure 4 results. Moreover, the average squared estimation error for the proposed and conventional methods are 5.541 and 7.095, respectively. Therefore, these results show the promise of the SVD-based kernel Wiener filter for signal/image reconstruction and restoration.

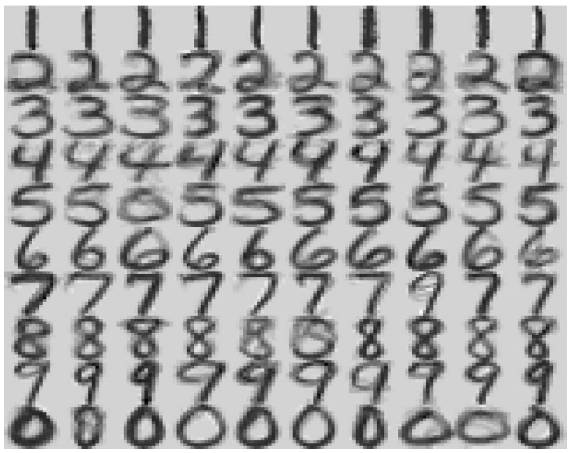

Fig. 3. Linear algebra based kernel Wiener filter with $\sigma^2 = 0.7$, $\tilde{n} = 20$, and $r = 300$.

5. CONCLUSION

An SVD-based kernel Wiener filter using HCCA framework is introduced for nonlinear signal estimation and reconstruction. Kernel HCCA was first extended and its relation to the reduced rank kernel Wiener filtering was demonstrated. The solution to the reduced-rank Kernel Wiener filter was then obtained by solving the higher dimensional optimization problem in the original signal space to find Wiener filtered pre-images. To avoid the iterative computation, kernel Wiener filter is obtained by embedding the distance information into the higher dimensional space, and finding pre-images only utilizing linear algebra. The signal estimation and reconstruction capability of the proposed method was demonstrated and benchmarked against the conventional kernel Wiener filter on the USPS data set.

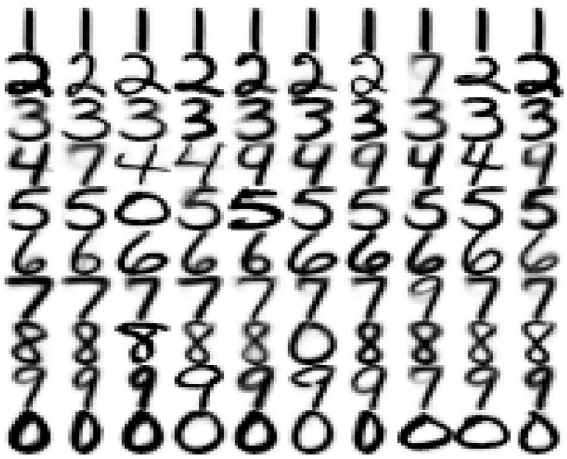

Fig. 4. Conventional kernel Wiener filter with $\sigma^2 = 0.053$ and $r = 300$.

6. ACKNOWLEDGMENTS

We thank to Yukihiko Yamashita for useful comments and discussions.

7. REFERENCES

[1] M. Yamada and M R. Azimi-Sadjadi, "Kernel Wiener Filter using Canonical Correlation Analysis Framework," *Proc. IEEE Workshop on Statistical Signal Processing (SSP'05)*, pp.320-325, Bordeaux, France, July 2005.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA: MIT Press, 2002

[3] J. Kwok and I. Tsang, "The pre-image problem in kernel methods," *IEEE Trans. Neural Networks,* vol. 15, no. 6, pp. 1517-1525, Nov. 2004

[4] A. Pezeshki, M. R. Azimi-Sadjadi, and L. L. Scharf, "Kernel-based canonical coordinate decomposition of two-channel nonlinear maps," *Proc. 2004 Int. Joint Conf. Neural Networks (IJCNN2004),* pp. 3019-3024, July 2004.

[5] L. L. Scharf, *Statistical Signal Processing.* MA:Addison-Wesley, 1991, pp. 330-331.

[6] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000