KERNEL WIENER FILTER USING CANONICAL CORRELATION ANALYSIS FRAMEWORK

Makoto Yamada and Mahmood R. Azimi-Sadjadi

Department of Electrical and Computer Engineering
Colorado State University
Fort Collins, CO 80523
email: {makoto, azimi}@engr.colostate.edu

ABSTRACT

This paper addresses the problem of kernel Wiener filter using kernel Canonical Correlation Analysis (CCA) framework. We solve the Wiener filter problem in the higher dimensional mapped domain using the *kernel trick*. A method is proposed to find approximate Wiener filtered signal in the original space by solving an optimization problem in higher dimensional space. The final form of kernel Wiener filter that relates to kernel Gram matrices, corresponds to the mean shift procedure or weighted nearest neighbor retrieval. The signal estimation and reconstruction capability of the kernel Wiener filter is demonstrated on the United States Postal Service (USPS) digits database. Moreover, a comparison between the linear Wiener filter and reduced-rank kernel Wiener filter is also presented.
Keywords: Kernel Wiener Filter, CCA, Mean Shift Procedure, Pre-image.

1. INTRODUCTION

Canonical Correlation Analysis (CCA) was introduced by Hotelling [1] and further developed by Anderson [2] for the analysis of linear dependence between two data channels. CCA decomposes the linear dependence between the original channels into the linear dependence between the canonical coordinates of the channels, where this linear dependence is easily determined by the corresponding canonical correlations. Another canonical coordinate system is obtained by Half CCA (HCCA) [3, 4] in which one of the channels is not whitened as opposed to CCA where both channels are whitened. The coordinate system given by HCCA is important for reduced-rank estimation where the objective of estimation is to minimize the mean squared error (MSE) [4].

CCA or HCCA work efficiently if the original data set is linearly representable. However, they may not be appropriate representations when the data set is nonlinearly separable. Thus, nonlinear extension of CCA or HCCA is required which may be accomplished by using kernel, leading to kernel CCA or HCCA. The basic idea of kernel method is to map the data in the input space to higher dimensional space via some nonlinear mapping functions, and apply linear operations in that space. This has been applied to many problems such as Principal Component Analysis (PCA) [5], and Fisher's Discriminant Analysis (FDA) [5]. In [6], CCA is extended to the kernel case in order to measure coherence between the high order attributes of the original two channel data. However, when the dimensions of the non-linearly mapped data become larger than the number of samples in the two-channel data sets, the canonical correlations no longer represent the coherence between the mapped data [4, 6]. This limits the usefulness of the original kernel CCA in practical applications. Nonetheless, this is not a problem with kernel HCCA.

The relation between Wiener filter and CCA or HCCA is well-known [4, 7, 8], while the relationship of their kernel versions has not been rigorously addressed. In [9], kernel Wiener filter is proposed using Taylor expansion up to the first order of noisy channel to reduce the computations and provide a realistic solution.

Since kernel Wiener filter is applied to the higher dimensional space, the mapped signals are usually unknown. Thus, one needs to look for the signal (or image) in the original space by minimizing the error between the original signal and filtered signal in the higher dimensional space. This is referred to as "pre-image" [10, 11]. One approach for finding the pre-image is through the noise reduction procedure using kernel PCA [10, 11]. In this reference, the error between the mapped data and the reconstructed data using kernel PCA was minimized, and the optimization problem was implicitly solved with respect to a pre-image.

In this paper, we follow HCCA and the optimization framework to obtain Wiener filtered pre-images which correspond to the mapped Wiener filtered signal and minimize MSE of the two-channels. Kernel Wiener filter is expressed in terms of the kernel Gram matrices and the pre-image is iteratively computed. Kernel Wiener filter is found to be

equivalent to the mean shift procedure or the boundary optimization method [12]. We show that it is also equivalent to the weighted nearest neighbor retrieval.

In the sequel, we will first derive the kernel versions of HCCA and then show reduced rank kernel Wiener filter. Then, the problem of finding the kernel Wiener filter preimage is cast in a high dimensional optimization problem the solution of which is obtained in the lower dimensional signal space. The link between the mean-shift procedure or the weighted nearest neighbor retrieval is also established. Kernel Wiener filter estimation and reconstruction are experimentally verified by using the USPS [1] digits data set.

2. KERNEL HCCA AND RELATION TO KERNEL WIENER FILTER

Kernel CCA or HCCA originated from the idea of the nonlinear kernel-based information processing methods [13, 14]. The idea is that the vectors are first mapped into the higher dimensional space, and then CCA or HCCA is applied in that space. However, when the data is mapped into the higher dimensional space, the process becomes computationally intractable. To overcome this problem, the so called *kernel trick* is utilized [13, 14], where all the operations are performed in the lower dimensional space and the kernel functions are computed directly from the original data channels. Among the widely used kernel functions are Gaussian and polynomial kernels [7].

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 $\boldsymbol{\psi}(\mathbf{y}) \in \mathbb{R}^{p'}$ with mean vectors $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\mu}_\psi$, respectively, and composite covariance matrix,

$$E\left[\begin{pmatrix} \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\mu}_\phi \\ \boldsymbol{\psi}(\mathbf{y}) - \boldsymbol{\mu}_\psi \end{pmatrix} \begin{pmatrix} \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\mu}_\phi \\ \boldsymbol{\psi}(\mathbf{y}) - \boldsymbol{\mu}_\psi \end{pmatrix}^T \right] = \begin{bmatrix} \mathrm{R}_{\phi\phi} & \mathrm{R}_{\phi\psi} \\ \mathrm{R}_{\psi\phi} & \mathrm{R}_{\psi\psi} \end{bmatrix} \quad (1)$$

where $\mathrm{R}_{\phi\phi} \in \mathbb{R}^{m' \times m'}$ and $\mathrm{R}_{\psi\psi} \in \mathbb{R}^{p' \times p'}$ are covariance matrices of $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{y})$ and $\mathrm{R}_{\phi\psi} = \mathrm{R}_{\psi\phi}^T \in \mathbb{R}^{m' \times p'}$ is the cross-covariance matrix between $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{y})$. The half canonical coordinates [4, 6] of $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{y})$ are then generated using

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathrm{F}^T & 0 \\ 0 & \mathrm{G}^T \mathrm{R}_{\psi\psi}^{-1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\mu}_\phi \\ \boldsymbol{\psi}(\mathbf{y}) - \boldsymbol{\mu}_\psi \end{bmatrix}, \quad (2)$$

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

where $\mathrm{R}_{\psi\psi}^{1/2}\mathrm{R}_{\psi\psi}^{T/2} = \mathrm{R}_{\psi\psi}$, and $\mathrm{F} \in \mathbb{R}^{m' \times m'}$ and $\mathrm{G} \in \mathbb{R}^{p' \times p'}$ are the orthogonal matrices in the singular value decomposition (SVD) of the half coherence matrix $\mathrm{C} = \mathrm{R}_{\phi\psi}\mathrm{R}_{\psi\psi}^{-T/2} \in \mathbb{R}^{m' \times p'}$:

$$\begin{aligned} \mathrm{C} = \mathrm{F}\Lambda\mathrm{G}^T \quad \text{or} \quad \mathrm{F}^T\mathrm{C}\mathrm{G} = \Lambda, \\ \text{with} \quad \mathrm{F}^T\mathrm{F} = \mathrm{I}, \ \ \mathrm{G}^T\mathrm{G} = \mathrm{I}, \\ \Lambda = \begin{bmatrix} \Lambda_{m'} & \mathbf{0} \end{bmatrix}; \quad \Lambda_{m'} = diag[\lambda_1, \ldots, \lambda_{m'}] \end{aligned} \quad (3)$$

where $\Lambda \in \mathbb{R}^{m' \times p'}$ is the canonical correlation matrix which measures the cross-correlations between the canonical coordinates $\mathbf{u}$ and $\mathbf{v}$ i.e $E[\mathbf{u}\mathbf{v}^T] = \Lambda$ [8]. Correspondingly, $\mathrm{D}_\phi^T = \mathrm{F}^T$ and $\mathrm{D}_\psi^T = \mathrm{G}^T\mathrm{R}_{\psi\psi}^{-1/2}$ are the mapping matrices that transform the data to their half canonical coordinates i.e. $\mathbf{u} = \mathrm{D}_\phi^T(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\mu}_\phi)$ and $\mathbf{v} = \mathrm{D}_\psi^T(\boldsymbol{\psi}(\mathbf{y}) - \boldsymbol{\mu}_\psi)$, where $\mathrm{R}_{uu} = \mathrm{D}_\phi^T\mathrm{R}_{\phi\phi}\mathrm{D}_\phi$, $\mathrm{R}_{vv} = \mathrm{I}$, and $\mathrm{R}_{uv} = E[\mathbf{u}\mathbf{v}^T] = \mathrm{D}_\phi^T\mathrm{R}_{\phi\psi}\mathrm{D}_\psi = \Lambda$.

Now, using the half coherence matrix $\mathrm{C} = \mathrm{R}_{\phi\psi}\mathrm{R}_{\psi\psi}^{-1/2} = \mathrm{F}\Lambda\mathrm{G}^T$, post-multiplying C by G and using the orthogonal property i.e $\mathrm{G}^T\mathrm{G} = \mathrm{G}\mathrm{G}^T = \mathrm{I}$ and the definitions of D_ϕ and D_ψ yields,

$$\mathrm{R}_{\phi\psi}\mathrm{D}_\psi = \mathrm{D}_\phi\Lambda \quad (4)$$

Also, pre-multiplying C by F^T, post-multiplying it by $\mathrm{R}_{\psi\psi}^{1/2}$, and using the orthogonal property of F gives

$$\mathrm{R}_{\psi\phi}\mathrm{D}_\phi = \mathrm{R}_{\psi\psi}\mathrm{D}_\psi\Lambda^T \quad (5)$$

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. Define the mapped sample matrices Φ and Ψ as

$$\Phi = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)] \in \mathbb{R}^{m' \times n} \quad (6)$$

$$\Psi = [\boldsymbol{\psi}(\mathbf{y}_1), \ldots, \boldsymbol{\psi}(\mathbf{y}_n)] \in \mathbb{R}^{p' \times n} \quad (7)$$

Also, the sample covariance matrices are given by

$$\hat{\mathrm{R}}_{\phi\phi} = \frac{1}{n}\Phi\mathrm{P}_{\mathbf{1}}^{\perp}\Phi^T \quad (8)$$

$$\hat{\mathrm{R}}_{\phi\psi} = \frac{1}{n}\Phi\mathrm{P}_{\mathbf{1}}^{\perp}\Psi^T \quad (9)$$

$$\hat{\mathrm{R}}_{\psi\psi} = \frac{1}{n}\Psi\mathrm{P}_{\mathbf{1}}^{\perp}\Psi^T \quad (10)$$

where $\mathrm{P}_{\mathbf{1}}^{\perp} = \mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ with $\mathbf{1} = [1, \ldots, 1] \in \mathbb{R}^n$ is a centering matrix. We can replace the covariance matrices $\mathrm{R}_{\phi\phi}$, $\mathrm{R}_{\phi\psi}$, and $\mathrm{R}_{\psi\psi}$ in (4) and (5) by their sample covariance matrices without any loss of generality. Thus, we get

$$\frac{1}{n}\Phi\mathrm{P}_{\mathbf{1}}^{\perp}\Psi^T\mathrm{D}_\psi = \mathrm{D}_\phi\Lambda \quad (11)$$

$$\frac{1}{n}\Psi\mathrm{P}_{\mathbf{1}}^{\perp}\Phi^T\mathrm{D}_\phi = \frac{1}{n}\Psi\mathrm{P}_{\mathbf{1}}^{\perp}\Psi^T\mathrm{D}_\psi\Lambda^T \quad (12)$$

Let $k_\phi(\cdot,\cdot)$ and $k_\psi(\cdot,\cdot)$ define the inner products Mercer kernels [5] in the implicit spaces $\mathbb{R}^{m'}$ and $\mathbb{R}^{p'}$:

$$k_\phi(\mathbf{x},\mathbf{y}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{y}) \quad (13)$$

$$k_\psi(\mathbf{x},\mathbf{y}) = \boldsymbol{\psi}(\mathbf{x})^T\boldsymbol{\psi}(\mathbf{y}) \quad (14)$$

The kernel Gram matrices associated with (13) and (14) are given by

$$\mathrm{K}_\phi = \Phi^T\Phi = [k_\phi(\mathbf{x}_i,\mathbf{x}_j)] \in \mathbb{R}^{n\times n} \quad (15)$$

$$\mathrm{K}_\psi = \Psi^T\Psi = [k_\psi(\mathbf{y}_i,\mathbf{y}_j)] \in \mathbb{R}^{n\times n} \quad (16)$$

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 [7].

From (11) and $\mathrm{D}_\psi = (\frac{1}{n}\Psi\mathrm{P}_{\mathbf{1}}^{\perp}\Psi^T)^{-T/2}\mathrm{G}$, it is obvious that Span$\{\mathrm{D}_\phi\}$ = Span$\{\Phi\}$ and Span$\{\mathrm{D}_\psi\}$ = Span$\{\Psi\}$ i.e. $\mathrm{D}_\phi = \Phi\hat{\mathrm{D}}_\phi$ and $\mathrm{D}_\psi = \Psi\hat{\mathrm{D}}_\psi$. Pre-multiplying (11) and (12) by Φ^T and Ψ, respectively, and replacing $\mathrm{D}_\psi = \Phi\hat{\mathrm{D}}_\phi$ and $\mathrm{D}_\psi = \Psi\hat{\mathrm{D}}_\psi$, and utilizing the kernel Gram matrices yields

$$\frac{1}{n}\mathrm{K}_\phi\mathrm{P}_{\mathbf{1}}^{\perp}\mathrm{K}_\psi\hat{\mathrm{D}}_\psi = \mathrm{K}_\phi\hat{\mathrm{D}}_\phi\Lambda \quad (17)$$

$$\mathrm{K}_\psi\mathrm{P}_{\mathbf{1}}^{\perp}\mathrm{K}_\phi\hat{\mathrm{D}}_\phi = \mathrm{K}_\psi\mathrm{P}_{\mathbf{1}}^{\perp}\mathrm{K}_\psi\hat{\mathrm{D}}_\psi\Lambda^T \quad (18)$$

If the kernel Gram matrices are all non-singular e.g. for Gaussian kernel case [6], then (17) and (18) can be rewritten as,

$$\frac{1}{n}\mathrm{P}_{\mathbf{1}}^{\perp}\mathrm{K}_\phi\hat{\mathrm{D}}_\phi = \hat{\mathrm{D}}_\phi\Lambda\Lambda^T \quad (19)$$

$$\frac{1}{n}\mathrm{K}_\phi\mathrm{P}_{\mathbf{1}}^{\perp}\mathrm{K}_\psi\hat{\mathrm{D}}_\psi = \mathrm{K}_\psi\hat{\mathrm{D}}_\psi\Lambda^T\Lambda \quad (20)$$

The generalized eigenvalue problem of (19) and (20) 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 higher mapped data matrices Φ and Ψ.

Half canonical coordinates are optimal for computing the reduced rank kernel Wiener filter of $\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\mu}_\phi$ from $\boldsymbol{\psi}(\mathbf{y}) - \boldsymbol{\mu}_\psi$, when the objective of estimation is to minimize MSE of the two-channels [3]. Let us assume that the estimated mean vector is almost the same as the true mean vector i.e $\boldsymbol{\mu}_\phi = \mathrm{H}_r\boldsymbol{\mu}_\psi$, then the rank-$r$ ($r \le n$) estimate of $\boldsymbol{\phi}(\mathbf{x})$ from $\boldsymbol{\psi}(\mathbf{y})$ is given as $\boldsymbol{\phi}_r(\hat{\mathbf{x}}) = \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$ [3, 7]. 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

$$\begin{aligned}\mathrm{H}_r &= \mathrm{F}\hat{\Lambda}\mathrm{G}^T\mathrm{R}_{\psi\psi}^{-1/2} \\ &= \mathrm{D}_\phi\hat{\Lambda}\mathrm{D}_\psi^T \quad (21)\\ \mathrm{R}_{ee}(r) &= \mathrm{R}_{\phi\phi} - \mathrm{F}\hat{\Lambda}\hat{\Lambda}^T\mathrm{F}^T \quad (22)\\ \hat{\Lambda} &= \begin{bmatrix}\Lambda_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} \quad (23)\\ \Lambda_r &= diag[\lambda_1,\ldots,\lambda_r] \quad (24)\end{aligned}$$

where Λ_r contains the first r largest singular values of Λ.

Alternatively, the covariance matrix of the filtering error $\mathrm{R}_{ee}(r)$ may be rewritten as [4]

$$\begin{aligned}\mathrm{R}_{ee}(r) &= \mathrm{R}_{ee} + (\mathrm{H}-\mathrm{H}_r)\mathrm{R}_{\psi\psi}(\mathrm{H}-\mathrm{H}_r)^T \\ &= \mathrm{R}_{ee} + (\mathrm{C}-\mathrm{C}_r)(\mathrm{C}-\mathrm{C}_r)^T \quad (25)\end{aligned}$$

where $\mathrm{R}_{ee} = E[(\boldsymbol{\phi}(\mathbf{x}) - \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y}))(\boldsymbol{\phi}(\mathbf{x}) - \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y}))^T] = \mathrm{R}_{\phi\phi} - \mathrm{R}_{\phi\psi}\mathrm{R}_{\psi\psi}^{-1}\mathrm{R}_{\psi\phi}$ is the error covariance matrix in estimating $\boldsymbol{\phi}(\mathbf{x})$ from $\boldsymbol{\psi}(\mathbf{y})$ using the full-rank Wiener filter $\mathrm{H} = \mathrm{R}_{\phi\psi}\mathrm{R}_{\psi\psi}^{-1}$, and $\mathrm{C}_r = \mathrm{F}\hat{\Lambda}\mathrm{G}^T$. Applying the trace operator to (25), the optimal value of MSE of the rank-r kernel Wiener filter H_r is

$$\begin{aligned}e_r^2 &= tr\{\mathrm{R}_{ee} + (\mathrm{C}-\mathrm{C}_r)(\mathrm{C}-\mathrm{C}_r)^T\} \\ &= tr\{\mathrm{R}_{ee}\} + tr\{\Lambda\Lambda^T - \hat{\Lambda}\hat{\Lambda}^T\} \\ &= e^2 + \sum_{i=r+1}^{m'}\lambda_i^2 = e^2 + \xi_r \quad (26)\end{aligned}$$

where $e^2 = tr\{\mathrm{R}_{ee}\}$ is the MSE of the full-rank kernel Wiener filter estimator H, and $\xi_r = \sum_{i=r+1}^{m'}\lambda_i^2$ is the excess MSE due to rank reduction.

3. KERNEL HALF CCA AND RECONSTRUCTION WITH GAUSSIAN KERNELS

In actual situations, we are interested in a reconstruction in the input space rather than in the higher dimensional mapped domain, since the filtered signals in the higher mapped domain are basically not observable. The approach here attempts to solve this problem by computing a pre-image vector $\hat{\mathbf{x}}$ that satisfies $\boldsymbol{\phi}_r(\hat{\mathbf{x}}) = \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$, and H_r is the rth reduced-rank kernel Wiener filter. When the solution for vector $\mathbf{x}$ exists in the original feature space, this equation holds. However, the vector $\mathbf{x}$ that satisfies this condition does not always exist, and it need not be unique either [10, 11]. Thus, we need to estimate $\hat{\mathbf{x}}$ that will be a good approximation to $\mathbf{x}$ in the input space.

Now, if the estimated vector has no exact pre-image $\hat{\mathbf{x}}$, one can find an approximation by minimizing the cost function,

$$J = \| \boldsymbol{\phi}_r(\hat{\mathbf{x}}) - \mathrm{H}_r\boldsymbol{\psi}(\mathbf{y}) \|^2 \quad (27)$$

Alternatively, we can write

$$J = \alpha - 2\boldsymbol{\phi}_r(\hat{\mathbf{x}})^T\mathrm{H}_r\boldsymbol{\psi}(\mathbf{y}) + \tau \quad (28)$$

where $\alpha = \boldsymbol{\phi}_r(\hat{\mathbf{x}})^T\boldsymbol{\phi}_r(\hat{\mathbf{x}})$ and $\tau = \boldsymbol{\psi}(\mathbf{y})^T\mathrm{H}_r^T\mathrm{H}_r\boldsymbol{\psi}(\mathbf{y})$. Note that α is a constant scalar for Gaussian kernels and τ is independent of $\hat{\mathbf{x}}$. Substituting (21) into (28) yields,

$$J = \alpha - 2\boldsymbol{\phi}_r(\hat{\mathbf{x}})^T\mathrm{D}_\phi\hat{\Lambda}\mathrm{D}_\psi^T\boldsymbol{\psi}(\mathbf{y}) + \tau \quad (29)$$

Thus, we can rewrite (29) as,

$$J = \alpha - 2k_{\phi,r}^T(\mathrm{X}, \hat{\mathbf{x}})\boldsymbol{\gamma} + \tau \quad (30)$$

where

$$\begin{aligned} \boldsymbol{\gamma} &= \hat{\mathrm{D}}_\phi \hat{\Lambda} \hat{\mathrm{D}}_\psi^T k_\psi(\mathrm{Y}, \mathbf{y}) & (31) \\ k_{\phi,r}(\mathrm{X}, \hat{\mathbf{x}}) &= \Phi^T \boldsymbol{\phi}_r(\hat{\mathbf{x}}) \\ &= [k_\phi(\mathbf{x}_1, \hat{\mathbf{x}}), \ldots, k_\phi(\mathbf{x}_n, \hat{\mathbf{x}})]^T & (32) \\ k_\psi(\mathrm{Y}, \mathbf{y}) &= \Psi^T \boldsymbol{\psi}(\mathbf{y}) \\ &= [k_\phi(\mathbf{y}_1, \mathbf{y}), \ldots, k_\phi(\mathbf{y}_n, \mathbf{y})]^T & (33) \end{aligned}$$

All of these operations are performed by utilizing the *kernel trick*.

To minimize J, we differentiate (30) with respect to $\hat{\mathbf{x}}$, and set the result to 0. For a Gaussian kernel i.e. $k_{\phi,r}(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{c})$, we get,

$$\begin{aligned} \frac{\partial J}{\partial \hat{\mathbf{x}}} &= \boldsymbol{\gamma}^T \frac{\partial k_{\phi,r}(\mathrm{X}, \hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}} \\ &= \sum_{i=1}^{n} \gamma_i \exp(-\frac{\| \mathbf{x}_i - \hat{\mathbf{x}} \|^2}{c})(\mathbf{x}_i - \hat{\mathbf{x}}) = 0 \quad (34) \end{aligned}$$

which yields the solution for $\hat{\mathbf{x}}$ as

$$\hat{\mathbf{x}} = \frac{\sum_{i=1}^{n} \gamma_i \exp(-\frac{\|\mathbf{x}_i - \hat{\mathbf{x}}\|^2}{c})\mathbf{x}_i}{\sum_{i=1}^{n} \gamma_i \exp(-\frac{\|\mathbf{x}_i - \hat{\mathbf{x}}\|^2}{c})} \quad (35)$$

Note that, this equation can be solved iteratively using

$$\hat{\mathbf{x}}^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_i \exp(-\frac{\|\mathbf{x}_i - \hat{\mathbf{x}}^{(t)}\|^2}{c})\mathbf{x}_i}{\sum_{i=1}^{n} \gamma_i \exp(-\frac{\|\mathbf{x}_i - \hat{\mathbf{x}}^{(t)}\|^2}{c})} \quad (36)$$

where γ_i's are obtained from (30) and t is the index for iterations. It can be shown that (36) is equivalent to the mean shift procedure [12]. If we define the following conditional probability

$$P(\mathbf{x}_i | \hat{\mathbf{x}}^{(t)}, \boldsymbol{\gamma}) = \frac{\gamma_i \exp(-\frac{\|\mathbf{x}_i - \hat{\mathbf{x}}^{(t)}\|^2}{c})}{\sum_{j=1}^{n} \gamma_j \exp(-\frac{\|\mathbf{x}_j - \hat{\mathbf{x}}^{(t)}\|^2}{c})} \quad (37)$$

then (36) can be rewritten as

$$\hat{\mathbf{x}}^{(t+1)} = \sum_{i=1}^{n} P(\mathbf{x}_i | \hat{\mathbf{x}}^{(t)}, \boldsymbol{\gamma})\mathbf{x}_i \quad (38)$$

which is nothing but the weighted nearest neighbor retrieval [15]. Figure 1 shows the graphical interpretation of kernel Wiener filter. As can be seen, the training sample closest to $\hat{\mathbf{x}}^{(t)}$ influences the retrieval and restoration process the most.

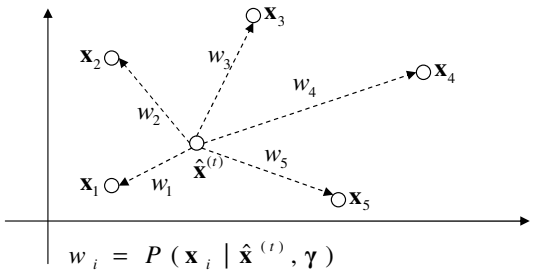

Fig. 1. Weighted Nearest Neighbor Retrieval.

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 Gaussian kernel of the form $k_\phi(\mathbf{x}, \mathbf{y}) = k_\psi(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2})$ with several choices for parameter σ. 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 2 and 3 show the original 100 clean digit images and the corresponding noisy images corrupted by additive white Gaussian noise with SNR=1dB, respectively.

We randomly chose 1000 samples for the training and 100 samples for testing. The pre-image $\hat{\mathbf{x}}$ is calculated iteratively using (36). The initial conditions, $\hat{\mathbf{x}}^{(0)}$, is set to noisy data $\mathbf{y}$, since kernel Wiener filter converges to different solutions $\hat{\mathbf{x}}$ when the initial condition $\hat{\mathbf{x}}^{(0)}$ is not close to the centroid of the expected class. The optimal solution is generated within 30 iterations.

Fig. 2. Original images

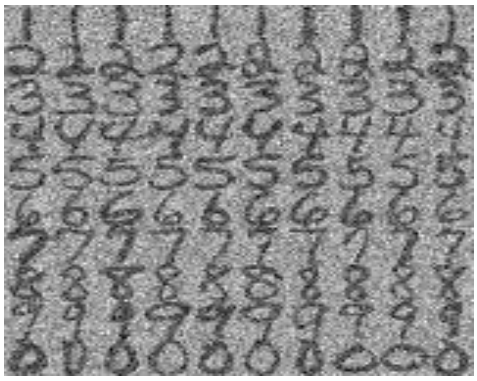

Fig. 3. Noisy Images (SNR = 1dB).

Figures 4 shows the result of the filtering using reduced-rank kernel Wiener filter for $\sigma = 3.92$ with rank $r = 600$. For this choice of rank r, the excess MSE is found to be $\xi_r = 0.2424$. As can be seen, the filter restores the original images successfully. However, interestingly some of the filtered digits have been changed to the different numbers i.e. not correctly retrieved. This is due to the fact that kernel Wiener filter not only restores the images but also plays the role of a weighted nearest neighbor retrieval, as mentioned before. That is, the training sample that yields the highest conditional probability in (37) will be retrieved after the denoising process.

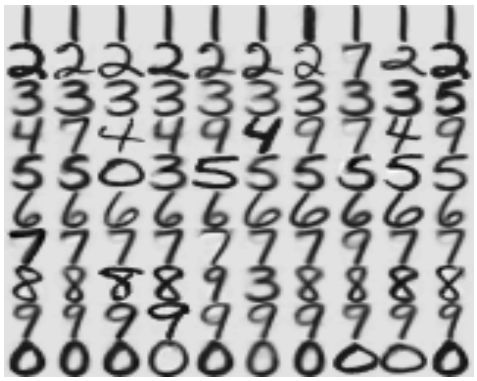

Fig. 4. Kernel Wiener filter with CCA ($\sigma = 3.92, r = 600$).

Figure 5 shows the filtered images by reduced-rank kernel Wiener filter with rank $r = 100$. As can be seen, the filtered results are more mixed together when compared to the results for $r = 600$. This implies that smaller ranks increase the incident of wrong retrieval while the estimation quality is still very good. In this case, the excess MSE, $\xi_r = 0.7480$, which is much larger than that for rank $r = 600$ case indicating poorer quality of the filtered images. Comparing Figures 4 and 5, it is obvious that rank rich kernel Wiener filter provided much better restoration with less noticeable smearing artifacts. Therefore, increasing the rank of the reduced-rank kernel Wiener filter clearly refines the filtered images.

Fig. 5. Kernel Wiener filter with CCA ($\sigma = 3.92, rank = 100$).

Figure 6 shows the full-rank linear Wiener filtered image. Comparing Figures 4 and 6, it is evident that the kernel Wiener filter performed better than the linear Wiener filter in terms of restoring the degraded images. Nonetheless, it also results in wrong retrieval in some cases due to variations in the training and testing data samples and the fact that kernel Wiener filter also performs some type of recall.

Fig. 6. Full-rank Wiener Filtered images.

5. CONCLUSION

Kernel Wiener filter using HCCA framework is addressed in this paper. Kernel HCCA was first extended and its relation to reduced rank kernel Wiener filter was demonstrated. The solution to the reduced-rank Kernel Wiener filter is obtained by solving the higher dimensional optimization problem to find Wiener filtered pre-images. The pre-image is obtained iteratively in the lower dimensional space. Additionally, the results in this paper showed that using this kernel HCCA to implement kernel Wiener filter not only leads to signal/image estimation but also performs some type of signal/image retrieval. Relation of the proposed method to weighted nearest neighbor retrieval or the mean-shift procedure was also established and experimentally confirmed on the USPS database.

6. REFERENCES

[1] H. Hotelling, "Relation between two sets of variates," *Biometrika*, vol. 28 pp. 321-377, 1936

[2] T. W. Anderson, *An Introduction to multivariate statistical analysis*, New York: Wiley, 2003.

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

[4] A. Pezeshki, "Two-channel signal processing in canonical coordinates," Ph.D. Dissertation, Colorado State University, Fort Collins, CO, Oct. 2004.

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

[6] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, "Empirical canonical coordinate decompositions in subspaces for two-channel linear and nonlinear maps," *IEEE Trans. on Signal Processing*submitted sept.2004, under revision.

[7] 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.

[8] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate and capacity," *IEEE Trans. on Signal Processing*, vol. 48, pp. 824-831, March 2000.

[9] Y. Washizawa and Y. Yamashita, "Kernel Wiener Filter", *Workshop on Information-Based Induction Sciences (IBIS)*, Nov 11 - 12. 2003 (Japanese).

[10] S. Mika, B. Schölkopf, A. Smola, K.R.Müller, M. Scholz, and G. Rätsch, J. Weston, "Kernel PCA and denoising in feature spaces," *Advances in Neural Information Processing Systems 11*, pages 536 – 542, Cambridge, MA, 1999. MIT Press.

[11] G. Bakir, J. Weston, and B. Schölkopf, "Learning to Find Pre-Images," *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.

[12] M. Fashing and C. Tomasi, "Mean shift is a Bound Optimization", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, No.3, March 2005

[13] V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995

[14] K. Müller, S. Mika, G.Rätsch, K. Tsuda, and B. Schölkopf, "An Introduction to Kernel-Based Learing Algorithms," *IEEE Trans. on Neural Networks*, Vol.12,No.2,pp.181-201,2001

[15] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (2nd ed), New York: Wiley, 2001.