Nearest Neighbor Classifier Generalization through Spatially Constrained Filters

Simon Lucey[a,b,*], Ahmed Bilal Ashraf[c]

[a] *Commonwealth Science and Industrial Research Organization (CSIRO), Australia*
[b] *Queensland University of Technology (QUT), Australia*
[c] *Carnegie Mellon University, USA*

Abstract

It is widely understood that the performance of the nearest neighbor (NN) rule is dependent on: (i) the way distances are computed between different examples, and (ii) the type of feature representation used. Linear filters are often used in computer vision as a pre-processing step, to extract useful feature representations. In this paper we demonstrate an equivalence between (i) and (ii) for NN tasks involving weighted Euclidean distances. Specifically, we demonstrate how the application of a bank of linear filters can be re-interpreted, in the form of a symmetric weighting matrix, as a manipulation of how distances are computed between different examples for NN classification. Further, we argue that filters fulfill the role of encoding local spatial constraints into this weighting matrix. We then demonstrate how these constraints can dramatically increase the generalization capability of canonical distance metric learning techniques in the presence of unseen illumination and viewpoint change.

Keywords: nearest neighbor classification, distance metric learning, filter learning

1. Introduction

The nearest neighbor (NN) rule [1] is one of the oldest and simplest classification methods used in learning and vision. The NN rule classifies an unlabeled example by the label of its nearest neighbors in the training set. The classical NN rule is based on evaluating the Euclidean distance between data points. Euclidean distance, however, does not leverage any statistical patterns that might be estimated from a large training set. Consequently, a number of researchers have demonstrated that NN classification can be greatly improved by learning, specifically for weighted Euclidean distances, an appropriate distance metric from labeled examples [2, 3, 4, 5, 6, 7]. The success of this approach has spawned the discipline of *distance metric learning* that strives to learn a weighted Euclidean metric that optimizes a particular learning criterion. Notable distance metric learning techniques of this form include canonical principal component analysis [8], linear discriminant analysis [8], and more recently non-parametric discriminant analysis [9] and large margin nearest neighbor [7] classification.

Applying an ensemble of linear filter banks has proved advantageous [10, 11, 12] as a pre-processing step to extract useful feature representations for NN classification problems in computer vision. Unfortunately, these approaches are heavily reliant on heuristics like: (i) the choice of filter class (e.g., Gabor, log-Gabor, Haar, edge, etc.), and (ii) the number of filters from that class. Attempts to answer these questions have often been based previously on heuristics or qualitative biological motivations.

In this paper we argue that most of these questions can be largely circumvented if we re-interpret the application of these filters as a manipulation of the distance metric (i.e. the weighting matrix) in the Fourier domain. Our approach centers upon a hitherto overlooked equivalence between weighted Euclidean NN classification of images preprocessed with linear filters and distance metric learning. The main contributions of this paper are,

- Demonstrating that because the role of filters can be re-interpreted as a Fourier weighting matrix, the number and type of filters is inherently am-

[*] Corresponding author is Simon Lucey.
Email addresses: simon.lucey@csiro.au (Simon Lucey), bilal@cmu.edu (Ahmed Bilal Ashraf)

biguous with respect to linear NN classification, whereas the choice of weighting matrix is always unique. This insight greatly reduces the guess work/heuristics in filter selection and allows one to explore the role of distance metric learning in filter learning. (Section 3)

- We propose a method for learning this unique Fourier weighting matrix through the augmentation of canonical eigenvector methods for distance metric learning. (Section 4)
- The introduction of spatial constraints, like those seen in conventional filters, which we argue theoretically and demonstrate empirically are essential to good generalization performance. (Section 5)

The most important contribution of the work in this paper, however, is the remarkable generalization performance obtained through our approach for situations when the train image set is considerably different in appearance to the test image set (e.g. different viewpoints for the train and test sets). In these situations state of the art approaches to distance metric learning, such as large margin nearest neighbor (LMNN) classifiers [7], start to fail whereas our approach exhibits impressive invariance and superior performance. Further, we demonstrate our approach also outperforms biologically motivated Gabor filter banks in this generalization task.

Relation to previous work.: It should be noted that there has been some previous work performed in literature on the topic of learning filters for vision tasks such as object alignment and classification [13, 14, 15]. The paper closest in spirit to our own, can be found in the seminal work of Kumar et al. [15] concerning discriminant analysis using Volterra kernels (specifically when a first order approximation of the kernel is employed). Like our approach, the authors present an approach that applies discriminant analysis to sub-patches in an image, rather than the whole image itself. Their approach exhibits impressive empirical performance for a number of NN facial identity tasks compared to current state of the art. Although similar conceptually, our work differs substantially to the work of Kumar et al. [15]. First, our work is centrally motivated by the connection between filters as a distance metric in the Fourier domain. Unlike [15] we provide theoretical insight into the uniqueness of filters with respect to linear NN classification. Finally, our work is focused on investigating the role of filters with spatial constraints in encoding generalizations into classifiers whereas [15] concentrated solely on classification performance.

Notation: Images/signals in this paper shall always be expressed in vector form (e.g., $\mathbf{x}$), where vectors are always represented in lower-case bold. Matrices are always expressed in upper-case bold (e.g., $\mathbf{A}$). A $\hat{}$ applied to any vector denotes the DFT of a vectorized image/signal such that $\hat{\mathbf{x}} \leftarrow \mathbf{F}\mathbf{x}$, where $\mathbf{F}$ is the $N \times N$ matrix of complex basis vectors for mapping to the Fourier domain for any N dimensional vectorized image/signal. We present our work in this paper in a manner that is agnostic about the nature of the signal (i.e. whether the signal is $1D$, $2D$, $3D$, etc.) as a $\mathbf{F}$ can always be formed and its role expressed in vector form. In practice, however, since we are working with images it should be assumed that $\mathbf{F}$ refers generally to a vectorized 2D-DFT. We have chosen to employ a Fourier representation in this paper due to its particularly useful ability to represent convolutions as a Hadamard product in the Fourier domain. Additionally, we take advantage of the fact that $\text{diag}(\hat{\mathbf{g}})\hat{\mathbf{x}} = \hat{\mathbf{g}} \circ \hat{\mathbf{x}}$, where $\circ$ represents the Hadamard product, and diag() is an operator that transforms a N dimensional vector into a $N \times N$ dimensional diagonal matrix. The role of filter $\hat{\mathbf{g}}$ or image/signal $\hat{\mathbf{x}}$ can be interchanged with this property. Any transpose operator ^{T} on a complex vector or matrix in this paper additionally takes the complex conjugate in a similar fashion to the Hermitian adjoint [16].

2. Distance Metric Learning

As noted by [7] the NN classification rule is quite sensitive to the type of metric used. Here we shall assume a weighted Euclidean distance,

$$|| \mathbf{V}(\mathbf{x}_i - \mathbf{x}_j) ||^2 \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two input vectors between which we are calculating distances, and $\mathbf{V}$ is a projection matrix. Equation 1 can also be written as,

$$\mathcal{D}_{\mathbf{Q}}(\mathbf{x}_i, \mathbf{x}_j) = || \mathbf{x}_i - \mathbf{x}_j ||^2_{\mathbf{Q}} \tag{2}$$

where,

$$\mathbf{Q} = \mathbf{V}^T\mathbf{V} \ . \tag{3}$$

Throughout this paper we use the notation $|| \mathbf{x} ||^2_{\mathbf{Q}}$ to represent the quadratic term $\mathbf{x}^T\mathbf{Q}\mathbf{x}$. Any matrix $\mathbf{Q}$ formed in this way from a real-valued matrix $\mathbf{V}$ is guaranteed to be positive semidefinite (i.e. to have no negative eigenvalues).

Learning Q: There are an abundance of approaches in literature [7] on how to best learn $\mathbf{Q}$ for NN rule classification. Eigenvector methods are perhaps the most popular in literature. Notable examples include principal component analysis [8], linear discriminant analysis [8], relevant vector analysis [5] and non-parametric discriminant analysis [9]. These methods attempt to learn $\mathbf{Q}$ indirectly through the estimation of $\mathbf{V}$, and differ largely in how they use labeled or unlabeled data. Direct methods, such as Mahalanobis metric clustering [6] or large margin nearest neighbor (LMNN) classifiers [7] attempt to learn $\mathbf{Q}$ directly through convex optimization.

3. Filters as Distance Transforms

Let $\mathbf{x}$ represent a N dimensional vectorized image that is passed through a bank of M linear filters such that $\mathbf{g}_m$ represents the vectorized impulse response of the m^{th} filter. One can obtain the N dimensional vector response,

$$\hat{\mathbf{r}}_m = \hat{\mathbf{x}} \circ \hat{\mathbf{g}}_m \tag{4}$$

where $\hat{\mathbf{g}}$, $\hat{\mathbf{x}}$ and $\hat{\mathbf{r}}$ are the vectorized complex 2D discrete Fourier transforms (DFT) [16] of the vectorized real images $\mathbf{g}$, $\mathbf{x}$ and $\mathbf{r}$ respectively. In the common case where $\mathbf{x}$ is larger than the filter image $\mathbf{g}$, zeros can be padded to ensure it is the same size as $\mathbf{x}$. We should note that the operation in Equation 4 can be equivalently accomplished purely in the image (spatial) domain through the use of efficient 2D convolution operators, however, we have chosen to employ a Fourier representation due to its particularly useful ability to represent a 2D convolution as a Hadamard product in the Fourier domain.

One can now obtain an over-complete representation $\mathbf{z}$ of $\mathbf{x}$ based on the concatenation of filter output responses,

$$\mathbf{z} = [\mathbf{r}_1^T, \ldots, \mathbf{r}_M^T]^T \tag{5}$$

where M is the number of filters in the bank.

Theorem 3.1. *One can always express the Euclidean distance between two over-complete representations $\mathbf{z}_i$ and $\mathbf{z}_j$, derived from multiple filters $\{\hat{\mathbf{g}}_m\}_{m=1}^M$ and the raw images $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively as,*

$$\mathcal{D}_{\mathbf{I}}(\mathbf{z}_i, \mathbf{z}_j) =\| \hat{\mathbf{h}} \circ (\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j) \|^2 \tag{6}$$

where $\hat{\mathbf{h}}$ is a single filter.

Proof. $\mathcal{D}_{\mathbf{I}}(\mathbf{z}_i, \mathbf{z}_j)$[1] can be expressed in the frequency domain by using Parseval's relation [16]. Parseval's relation states that the energy content of any signal is preserved as we move from the spatial to the Fourier space. As a result one can express,

$$\begin{aligned}\mathcal{D}_{\mathbf{I}}(\mathbf{z}_i, \mathbf{z}_j) &= \sum_{m=1}^{M} \| \hat{\mathbf{g}}_m \circ (\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j) \|^2 \\ &= \sum_{m=1}^{M} \| (\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j) \|_{\mathbf{S}}^2\end{aligned} \tag{7}$$

where,

$$\mathbf{S} = \sum_{i=1}^{M} \text{diag}\{\hat{\mathbf{g}}_i\}^T \text{diag}\{\hat{\mathbf{g}}_i\} \tag{8}$$

is a diagonal weighting matrix in the Fourier domain. From the perspective of computing distances, applying a bank of M linear filters $\{\mathbf{g}_i\}_{i=1}^M$ is equivalent to applying a single filter $\hat{\mathbf{h}}$ such that $\mathbf{S} = \text{diag}\{\hat{\mathbf{h}}\}^T \text{diag}\{\hat{\mathbf{h}}\}$. One can see that, unlike the filter banks $\{\mathbf{g}_i\}_{i=1}^M$, $\mathbf{S}$ is unique and can always be represented by a single filter $\hat{\mathbf{h}}$.

Equation 7 shows that the step of pre-processing images by passing them through a bank of linear filters corresponds to weighting the distance between images in the Fourier space as specified by the matrix $\mathbf{S}$ given in Equation 8. In Equation 7 we can replace the operation of taking the Fourier transform by pre-multiplying the signals with a matrix $\mathbf{F}$ containing the vectorized $2D$ Fourier basis. We may then write the distance in Equation 6 as follows.

$$\begin{aligned}\mathcal{D}_{\mathbf{I}}(\mathbf{z}_i, \mathbf{z}_j) &= \mathcal{D}_{\mathbf{Q}}(\mathbf{x}_i, \mathbf{x}_j) \\ &= \| \mathbf{x}_i - \mathbf{x}_j \|_{\mathbf{Q}}^2\end{aligned} \tag{9}$$

where

$$\mathbf{Q} = \mathbf{F}^T \mathbf{S} \mathbf{F} \ . \tag{10}$$

4. Filter Metric Learning

In Section 3, we have shown that from the perspective of classification, the effect of linear filters is to manipulate the distance metric with a weighting matrix $\mathbf{Q} = \mathbf{F}^T\mathbf{S}\mathbf{F}$, where $\mathbf{S} = \text{diag}\{\hat{\mathbf{h}}\}^T \text{diag}\{\hat{\mathbf{h}}\}$.

Eigenvector methods: In this section we will demonstrate how eigenvector methods for distance metric learning can be adapted to learning filters. Eigenvector methods for distance metric learning canonically attempt to maximize the objective function,

$$\arg\max_{\hat{\mathbf{h}}} \frac{\hat{\mathbf{h}}^T \mathbf{C}_1 \hat{\mathbf{h}}}{\hat{\mathbf{h}}^T \mathbf{C}_2 \hat{\mathbf{h}}} \text{ subject to } \hat{\mathbf{h}}^T \hat{\mathbf{h}} = 1 \tag{11}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are symmetric scatter matrices encoding the particular requirements on $\hat{\mathbf{h}}$ (e.g., preserving energy, discriminating between classes, etc.). The vector $\hat{\mathbf{h}}$ can be found efficiently and deterministically by finding the leading eigenvector of $\mathbf{C}_2^{-1}\mathbf{C}_1$. Traditionally a scatter matrix can always be expressed in the generic form $\mathbf{C} = \sum_{\mathbf{x}_k \in C} \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T$ where $C = \{\hat{\mathbf{x}}_k\}_{k=1}^K$.

[1] $\mathbf{I}$ indicates an $MN \times MN$ identity matrix indicating an unweighted Euclidean distance.

Unfortunately, the vector $\hat{\mathbf{h}}$ estimated from Equation 11 using traditional scatter matrices cannot be considered a filter as nothing about convolutions has been encoded into the objective function. In fact, it should be noted one can find $\hat{\mathbf{h}}$ in the Fourier or spatial domains with the exact same result due to Parseval's relation (i.e. inner products are equal in the spatial and Fourier domains).

Theorem 4.1. *A modified filter scatter matrix* $\mathbf{C}^*$ *can always be formed from the traditional scatter matrix* $\mathbf{C}$ *where,*

$$\mathbf{C}^* = \sum_{\mathbf{x}_k \in C} diag(\hat{\mathbf{x}}_k) diag(\hat{\mathbf{x}}_k)^T \tag{12}$$

such that the solution to $\hat{\mathbf{h}}$ *in Equation 11 using the modified scatter matrices is equivalent to solving,*

$$\arg\max_{\hat{\mathbf{h}}} \frac{\sum_{C_1 \in \hat{\mathbf{x}}_k} \| \hat{\mathbf{h}} \circ \hat{\mathbf{x}}_k \|^2}{\sum_{C_2 \in \hat{\mathbf{x}}_j} \| \hat{\mathbf{h}} \circ \hat{\mathbf{x}}_j \|^2} \tag{13}$$

which is the filter that maximizes the objective function across all shifts.

Proof. If we pass all the images $C = \{\hat{\mathbf{x}}_k\}_{k=1}^K$, stemming from scatter matrix $\mathbf{C}$, through a filter specified by the impulse response $\hat{\mathbf{h}}$, then the variance in the filtered space may be written as,

$$\sigma^2_{\hat{\mathbf{h}}} = \sum_{\mathbf{x}_k \in C} \| \hat{\mathbf{h}} \circ \hat{\mathbf{x}}_k \|^2 \quad . \tag{14}$$

As demonstrated earlier, we may replace the Hadamard product with a matrix product by using the diag(.) operator as follows,

$$\sigma^2_{\hat{\mathbf{h}}} = \sum_{\mathbf{x}_k \in C} \| \mathrm{diag}(\hat{\mathbf{x}}_k) \hat{\mathbf{h}} \|^2 \tag{15}$$

which can be simplified as follows,

$$\begin{aligned} \sigma^2_{\hat{\mathbf{h}}} &= \hat{\mathbf{h}}^T (\sum_{\mathbf{x}_k \in C} \mathrm{diag}(\hat{\mathbf{x}}_k)^T \mathrm{diag}(\hat{\mathbf{x}}_k)) \hat{\mathbf{h}} \\ &= \hat{\mathbf{h}}^T \mathbf{C}^* \hat{\mathbf{h}} \end{aligned} \tag{16}$$

where $\mathbf{C}^*$ has been defined previously in Equation 12. Thus if $\hat{\mathbf{h}}$ is a filter, then $\hat{\mathbf{h}}^T \mathbf{C}^* \hat{\mathbf{h}}$ represents the data variance in the filtered space, justifying the employment of the modified scatter matrix $\mathbf{C}^*$ within Equation 11. It is this result that enables the co-option of eigen decomposition methods for learning filters.

Multiple eigenvectors: One may note in Equation 11 that we are only solving for a single eigenvector $\hat{\mathbf{h}}$, rather than multiple eigenvectors (which is often done in traditional eigenvector methods like PCA and LDA). The choice of solving for a single filter/eigenvector stems directlty from the previous result in Theorem 3.1 stating that: (i) multiple filters are not unique, and (ii) an equivalent distance measure can always be obtained using a single filter. This result has also been confirmed empirically in the results section (Figure **??**) of this paper.

5. Spatial Constraints

A fundamental difference between the learned filter $\hat{\mathbf{h}}$ in Equation 11 and traditional filters (e.g., Gabor, log-Gabor, Haar, edge, etc.) lies in their spatial support region. For the case of traditional filters, the support region is typically quite small (e.g., 5 × 5, 8 × 8, etc.) in comparison to the images upon which they are applied. No such spatial constraint is enforced on the filter in Equation 11. As a consequence $\hat{\mathbf{h}}$ is the same size as the images from which it is learned. This large spatial support has some unwanted properties, as we will discuss in the latter sections, with respect to generalization (i.e. how tuned is $\mathbf{h}$ to the training data). In this section we will present an approach for constraining this support region using existing distance metric learning techniques.

Enforcing spatial constraints: As in Section 4 we shall concern ourselves with eigenvector methods for distance metric learning, taking the view that augmentations on this canonical form can be applied to other metric learning techniques with nominal effort (as previously discussed in Section **??**). Adapting Equation 11 one can enforce a smaller spatial support region through the application of the following constraints,

$$\begin{aligned} \arg\max_{\mathbf{h}} \quad & \frac{\mathbf{h}^T (\mathbf{F}^T \mathbf{C}_1^* \mathbf{F}) \mathbf{h}}{\mathbf{h}^T (\mathbf{F}^T \mathbf{C}_2^* \mathbf{F}) \mathbf{h}} \\ \text{subject to} \quad & \mathbf{h}^T \mathbf{h} = 1 \\ & \mathbf{h}(i) = 0 \ \ \forall \ \ i \notin \Omega_{h \times w} \end{aligned} \tag{17}$$

where $\Omega_{h \times w}$ represents the set of indices stemming from a $h \times w$ local neighborhood of positions at the center of the vectorized $2D$ filter $\mathbf{h}$, $\mathbf{C}_1^*$ and $\mathbf{C}_2^*$ are the augmented scatter matrices defined by the training images. Other than the additional spatial constraints, one should note that Equation 17 differs from Equation 11 in that we are solving for $\mathbf{h}$ spatially instead of $\hat{\mathbf{h}}$ in the Fourier domain. This has to be done so we can enforce our new spatial constraints, and is accomplished by employing the matrix $\mathbf{F}$ containing the vectorized $2D$ Fourier basis into the objective function.

Problem: The original objective function in Equation 11 is solved efficiently through eigen-decomposition. However, the introduction of the spatial constraints in Equation 17 changes our objective function into a quadratic fractional programming problem with equality constraints. This shift, unfortunately, greatly complicates the optimization task. The sheer number of constraints is especially problematic as employing training images of dimensionality N would result in $N - hw$ constraints where h and w are the height and width of the desired spatial support region defined by $\Omega_{h\times w}$.

Theorem 5.1. *Equation 17 can be re-written as a tractable eigen-decomposition task,*

$$\arg\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T(\mathbf{A}^T\mathbf{C}_1^*\mathbf{A})\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T(\mathbf{A}^T\mathbf{C}_2^*\mathbf{A})\boldsymbol{\alpha}} \quad \textit{subject to} \quad \boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1 \tag{18}$$

where $\mathbf{A}$ *is a submatrix of the Fourier basis matrix* $\mathbf{F}$ *whose rows lie in the desired spatial support region* $\Omega_{h\times w}$. *As a consequence, instead of solving for* $\mathbf{h} \in \mathcal{R}^N$ *which has the same dimensionality as the training images, the result of Equation 18 is* $\boldsymbol{\alpha} \in \mathcal{R}^{hw}$ *will instead have the same dimensionality as the desired support region.*

Proof. Equation 17 can be re-written in the following form,

$$\begin{aligned} \arg\max_{\boldsymbol{\alpha},\boldsymbol{\xi}} \quad & \frac{\boldsymbol{\alpha}^T(\mathbf{A}^T\mathbf{C}_1^*\mathbf{A})\boldsymbol{\alpha} + \boldsymbol{\xi}^T(\boldsymbol{\Psi}^T\mathbf{C}_1^*\boldsymbol{\Psi})\boldsymbol{\xi}}{\boldsymbol{\alpha}^T(\mathbf{A}^T\mathbf{C}_2^*\mathbf{A})\boldsymbol{\alpha} + \boldsymbol{\xi}^T(\boldsymbol{\Psi}^T\mathbf{C}_2^*\boldsymbol{\Psi})\boldsymbol{\xi}} \\ \text{subject to} \quad & \boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1 \\ & \boldsymbol{\xi} = \mathbf{0} \end{aligned} \tag{19}$$

where $\boldsymbol{\alpha} \in \mathcal{R}^{hw}$ and $\boldsymbol{\xi} \in \mathcal{R}^{N-hw}$ are sub-vectors of $\mathbf{h} \in \mathcal{R}^N$, where $\mathbf{h}$ is the original filter being solved in Equation 17. The vector $\boldsymbol{\alpha}$ contains the elements of $\mathbf{h}$ relating to the indices in $\Omega_{h\times w}$. Similarly, the vector $\boldsymbol{\xi}$ contains the elements of $\mathbf{h}$ relating to the indices not in $\Omega_{h\times w}$. The matrices $\mathbf{A} \in \mathcal{R}^{hw\times N}$ and $\boldsymbol{\Psi} \in \mathcal{R}^{(N-hw)\times N}$ are submatrices of the Fourier matrix $\mathbf{F} \in \mathcal{R}^{N\times N}$. In a similar fashion to $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$, $\mathbf{A}$ and $\boldsymbol{\Psi}$ contain the rows of $\mathbf{F}$ relating to the indices in, and not in $\Omega_{h\times w}$ respectively. As a consequence of the constraint $\boldsymbol{\xi} = \mathbf{0}$ we can now obtain Equation 18 as the second term of Equation 19 has to be zero.

6. Experiments

Our aim in these experiments was to evaluate the *generalization* properties of NN[2] classification for a number of distance metric methods. To this end we conducted experiments to evaluate the generalization performance of NN identity classification for: (i) matched, and (ii) mismatched viewpoints of faces. The train set, from which the weight matrix $\mathbf{Q}$ is learned, is always from a single viewpoint. In our experiments it is only the test set where the viewpoint is varied. To emphasize this point, when the $\mathbf{Q}$ matrix is learned it has knowledge of a single and fixed viewpoint of the face. It is only during testing that other viewpoints of the face are presented. For all our experiments in this paper the train set stems from the frontal (0^o) viewpoint of the face.

MultiPIE: The MultiPIE face dataset [17] was used in all our experiments, as it is considered one of the largest and most comprehensive of its kind. It consists of images of 346 subjects. Each subject has been photographed by illuminating the face from 20 different illuminations, and the images have been captured from 15 viewpoints. Moreover, the subjects were asked to elicit facial expressions corresponding to 5 high level emotional states – neutral, happy, disgust, surprise, and pain. In our experiments all images were registered using hand labeled eye coordinates, with the face area then cropped to give a 140 × 80 image (irrespective of viewpoint). In all experiments in this paper subject identities are different in the train and test sets.

Distance metrics considered: We chose to compare our proposed method against canonical distance metric learning methods for learning $\mathbf{Q}$ such as PCA and LDA [8]. State of the art methods for distance metric learning were also considered, namely the popular LMNN classifier of Weinberger [7]. A biologically motivated $\mathbf{Q}$ was also considered through the employment of a bank of Gabor filters. The $\mathbf{Q}$ matrix was obtained in this instance through the application of Equations 8 and 10 discussed earlier in this paper. For the PCA, LDA, LMNN and Gabor filter derived $\mathbf{Q}$ matrices application specific parameters (such as the number of eigenvectors, number of filter banks, etc.) were tuned using cross-validation. For brevity, these specific implementation details are omitted but can be found in our experiments which are available online[3].

For our own proposed approach we employed the modified scatter matrices, as discussed in Section 4, stemming from the traditional scatter matrices used in LDA. As discussed in Section 4 only the first eigenvector is used to solve for the filter. Unlike, traditional LDA and LMNN methods on images, no initial PCA step was

[2] For all the experiments in this paper we shall be performing NN classification for $k = 1$, we refer to this as simply NN classification.

[3] During the double blind review process we have omitted this hyperlink.

employed in our approach to reduce dimensionality and avoid overfitting issues. Instead, the inclusion of spatial constraints, as discussed in Section 5, automatically handles this issue.

Illumination variation: Figure 1 depicts results for identity classification for a subset of the MultiPIE dataset in the presence of 20 different illumination conditions (fixed neutral expression). As expected for the matched viewpoint experiments in (a) (where the training and testing images both stem from the frontal (0^o) viewpoint) that LMNN, LDA and our approach obtain close to 100% classification accuracy. Far more interestingly, however, are the results in (b) which demonstrate substantial performance deterioration for all canonical distance metric learning methods (i.e., PCA, LDA and LMNN).

Contrastingly, our learned filter remains virtually unchanged still achieving close to 100% classification performance. A similar effect can be noticed for the Gabor filter metric approach, which obtains virtually identical classification performance in (a) and (b) (with Gabor filters also outperforming LMNN for the mismatched scenario). The effect of filter support size, for our learned filter, was also investigated on this illumination subset of MultiPIE. Figure 2 depicts identification results as a result of spatial support. This result strongly demonstrates the importance of limiting spatial support in order to reduce overfitting and encourage generalization. Experiments in this instance were carried out on matched viewpoints. A visualization of the filters, for varying spatial support, can be found in Figure 3. An interesting thing to note in Figure 3 is the highly synthetic nature of the learned filters, compared to canonical filters such as Gabor which are smoothly varying.

Expression variation: Similar experimental results can be found in Figure 4 for a subset of the MultiPIE dataset where 5 different expression conditions are presented (fixed frontal illumination). For these experiments in (a), our learned filter does not fair as well compared to LDA and LMNN. This is in contrast to the results for (a) seen in Figure 1 for illumination variation. This result, however, is largely expected as expression variation is much more spatially specific than illumination variation making it harder for a spatially invariant filter to offer invariance. Interestingly, however, in (b), which is of primary interest in this paper, our learned filter still outperforms LDA and LMNN.

7. Discussion

An obvious question that stems from the results in Figures 1 and 4 is: *why do distance metric learning techniques like PCA, LDA and even LMNN behave so poorly in mismatched conditions, whereas filter metrics such as Gabor and our own learned filter exhibit such promising invariance?*

A partial answer to this question can be found if we obtain a visualization of a candidate eigenvector method to distance metric learning using canonical and filter modified (with spatial constraints) scatter matrices. Eigenvectors were obtained for both viewpoints used in testing. For this discussion we chose one of the simplest forms of eigenvector method distance metric learning, namely PCA. In Figure 5 we depict canonical PCA using the first 5 eigenvectors of the illumination variation subset of MultiPIE dataset for both frontal (0^o) and profile (45^o) sets. Inspecting these eigenvectors, one can clearly see that (a) stems from the frontal view, and (b) stems from the profile view. In Figure 5 we depict filter PCA eigenvectors (estimated using a modified scatter matrix and an 8×8 spatial support). For this visualization we chose to keep all the eigenvectors, instead of keeping the first principal eigenvector. Unlike 5, it is very difficult to ascertain in Figure 6 which eigenvectors stem from which view. Instead, the eigenvectors from both viewpoints (a) and (b) look very similar.

Even though traditional PCA performed poorly compared to LDA and LMNN in Figures 1 and 4, the visualizations in Figures 5 and 6 goes some way to explaining why non-filter distance metric learning methods perform poorly. Traditional distance metric methods like PCA, LDA and even LMNN are spatial specific (i.e., images should all roughly follow the same spatial configuration). The distance metric learning framework presented in this paper, circumvents this problem, by learning distances (through the application of filters) that are approximately spatially invariant. This is largely why our proposed approach works so well on unseen viewpoints.

8. Conclusions

In this paper we demonstrate how the application of an ensemble of linear filter banks, as a pre-processing step, before NN classification can be re-interpreted as a manipulation of the distance metric (i.e. the weighting matrix). As a result we also demonstrate how canonical distance metric learning techniques can be augmented to learn filters, and take much of the guess

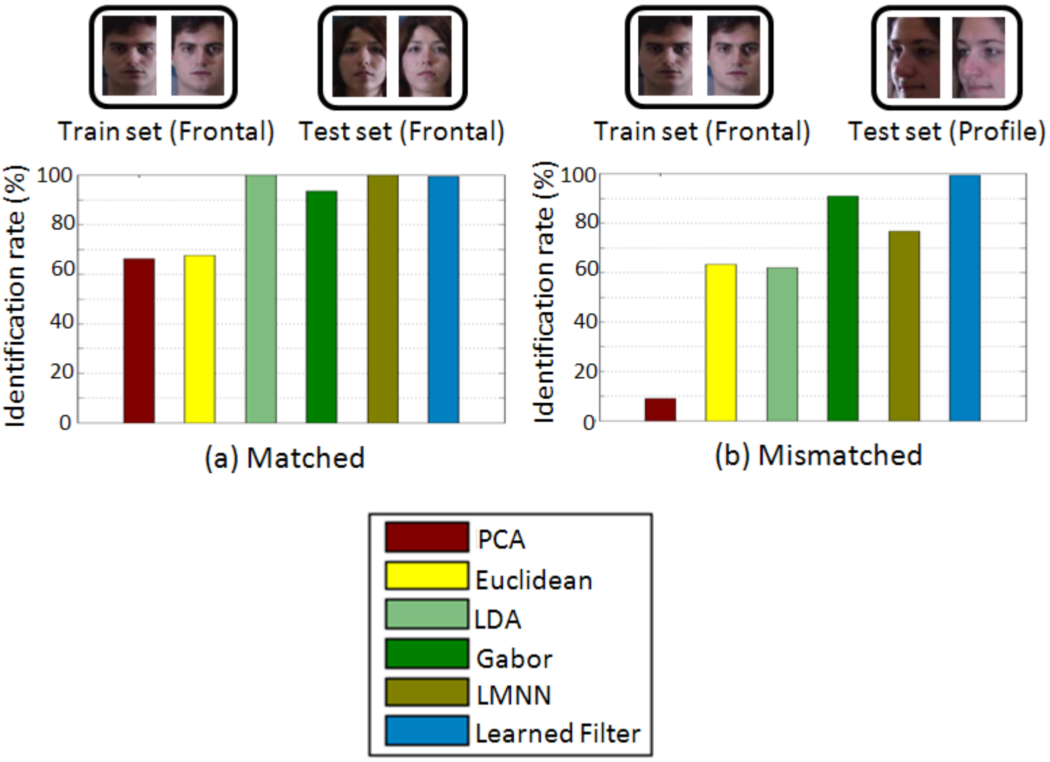

Figure 1: Comparison of distance metric methods for: (a) matched, and (b) mismatched viewpoints in the presence of illumination variation. For both (a) and (b) the train set images employed a frontal viewpoint (0^o), whereas the test set employed a viewpoint for (a) 0^o, and (b) 45^o. For (b), which is of central interest in this paper, our approach outperforms both LMNN classifiers and biologically motivated Gabor filter banks.

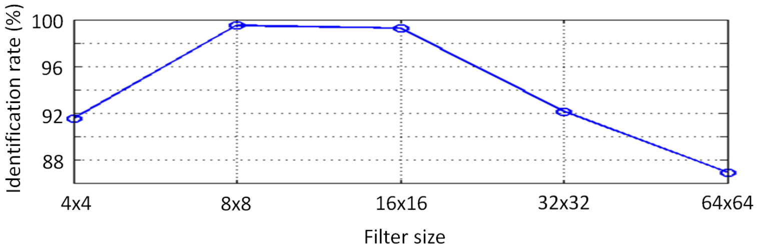

Figure 2: Identification results as a function of filter support size. Empirically, we found a filter size of 8×8 gave the best performance (99.54%).

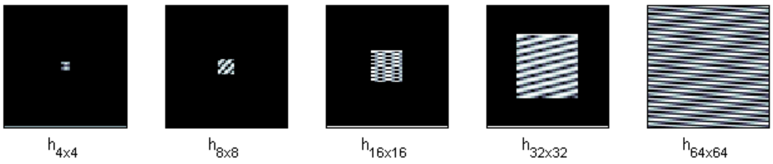

Figure 3: Visualization of our learned filters for varying spatial support.

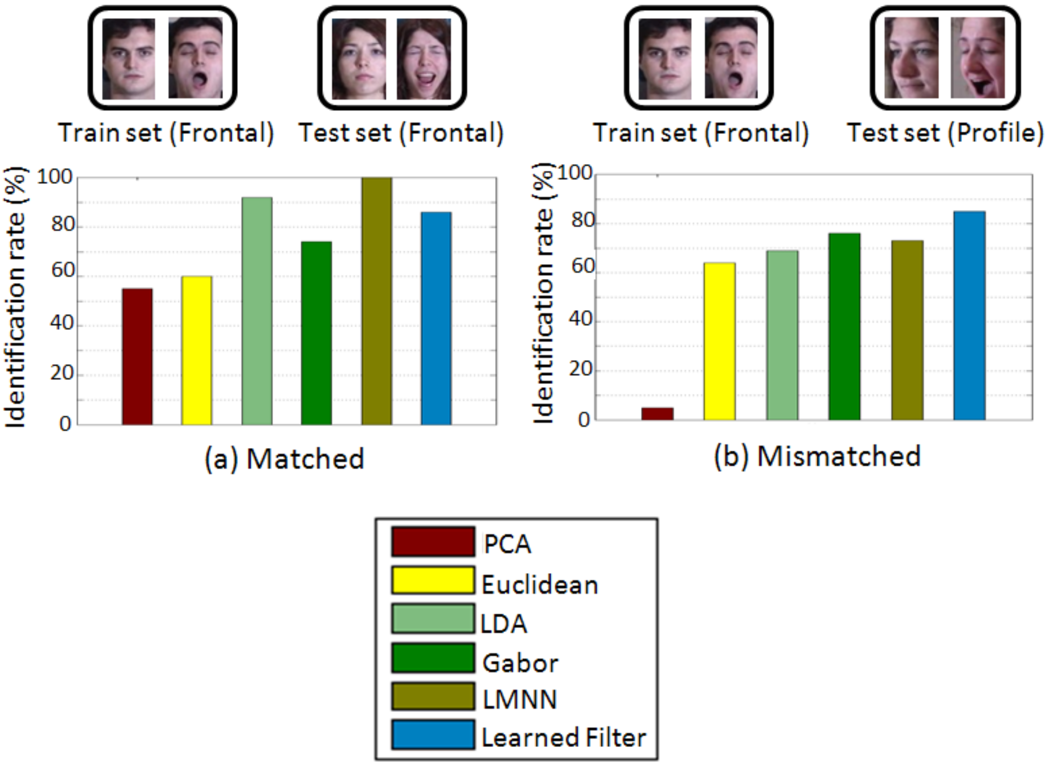

Figure 4: Comparison of distance metric methods for: (a) matched, and (b) mismatched viewpoints in the presence of expression variation. For both (a) and (b) the train set images employed a frontal viewpoint (0^o), whereas the test set employed a viewpoint for (a) 0^o, and (b) 45^o. Similarly to Figure 1 for the results in (b), which is of central interest in this paper, our approach outperforms both LMNN classifier and biologically motivated Gabor filter banks distance metrics.

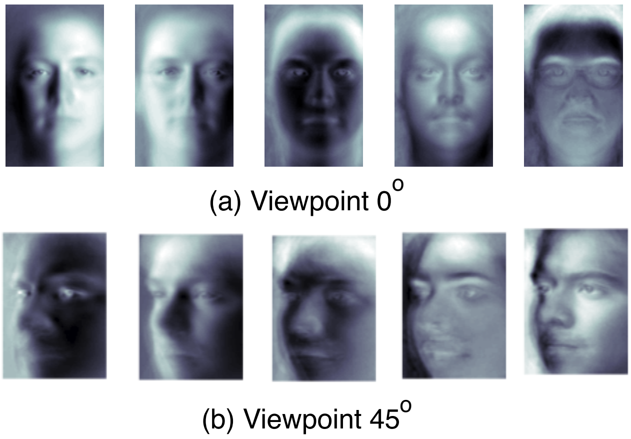

Figure 5: Visualization of the first 5 eigenvectors stemming from PCA using traditional scatter matrices for the (a) matched test set viewpoint (0^o), and the (b) mismatched test set viewpoint (45^o).

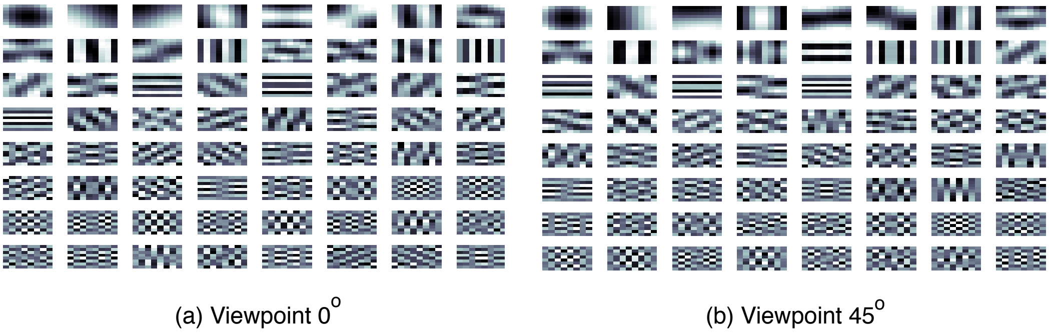

Figure 6: Visualization of eigenvectors stemming from PCA using filter modified scatter matrices with a spatial support of 8×8 pixels. Eigenvectors are presented for (a) matched test set viewpoint (0^o), and the (b) mismatched test set viewpoint (45^o).

work/heuristics out of selecting filters for a specific vision task. Finally, we demonstrated the useful generalization properties of our filters for classification tasks including illumination and expression variation under unseen viewpoints outperforming biologically motivated filters (i.e. Gabor).

References

[1] T. Cover, P. Hart, Nearest neighbor pattern classification., in: In IEEE Transactions in Information Theory, 1967, pp. 21 – 27.

[2] S. Chopra, R. Hadsell, Y. LeCun, Learning a similiarty metric discriminatively, with application to face verification, in: CVPR, 2005, pp. 349 – 356.

[3] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighbourhood components analysis, in: NIPS, MIT Press, Cambridge, MA, 2005, pp. 513 – 520.

[4] S. Shalev-Shwartz, T. Singer, A. Y. Ng, Online and batch learning of pseudo-metrics, in: ICML, 2004, pp. 94 – 101.

[5] N. Shental, T. Hertz, D. Weinshall, M. Pavel, Adjustment learning and relevant component analysis, in: ECCV, 2002.

[6] E. P. Xing, A. Y. Ng, M. I. Jordan, S. Russell, Distance metric learning, with application to clustering with side-information, in: NIPS, Vol. 14, 2002, pp. 521–528.

[7] K. Q. Weinberger, J. Blitzer, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: In NIPS, MIT Press, 2006.

[8] K. Fukunaga, Introduction to statistical pattern recognition, Academic Press, 1990.

[9] Z. Li, D. Lin, X. Tang, Nonparametric discriminant analysis for face recognition, PAMI 31 (4) (2009) 755–761.

[10] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition, IEEE Trans. Image Processing 11 (4) (2002) 467–476.

[11] A. Yilmaz, M. Gokmen, Eigenhill vs. eigenface and eigenedge, Pattern Recognition 34 (1) (2001) 181–184.

[12] H. Bischof, H. Wildenauer, A. Leonardis, Illumination insensitive recognition using eigenspaces, Computer Vision and Image Understanding 95 (1) (2004) 86 – 104.

[13] F. De la Torre, A. Collet, J. F. Cohn, T. Kanade, Filtered component analysis to increase robustness to local minima in appearance models, in: CVPR, 2007.

[14] B. V. K. Vijaya Kumar, A. Mahalanobis, R. D. Juday, Correlation pattern recognition, Cambridge University Press, 2005.

[15] R. Kumar, A. Banerjee, B. C. Vemuri, Volterrafaces: Discriminant analysis using volterra kernels, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[16] A. V. Oppenheim, A. S. Willsky, Signals & Systems, 2nd Edition, Prentice Hall, 1996.

[17] R. Gross, J. S. Baker, I. Matthews, T. Kanade, Multi-PIE, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2008.