Ubiquitously Supervised Subspace Learning

Jianchao Yang, *Student Member, IEEE*, Shuicheng Yan, *Member, IEEE*, and Thomas S. Huang, *Life Fellow, IEEE*

***Abstract*—In this paper, our contributions to the subspace learning problem are two-fold. We first justify that most popular subspace learning algorithms, unsupervised or supervised, can be unitedly explained as instances of a ubiquitously supervised prototype. They all essentially minimize the intraclass compactness and at the same time maximize the interclass separability, yet with specialized labeling approaches, such as ground truth, self-labeling, neighborhood propagation, and local subspace approximation. Then, enlightened by this ubiquitously supervised philosophy, we present two categories of novel algorithms for subspace learning, namely, misalignment-robust and semi-supervised subspace learning. The first category is tailored to computer vision applications for improving algorithmic robustness to image misalignments, including image translation, rotation and scaling. The second category naturally integrates the label information from both ground truth and other approaches for unsupervised algorithms. Extensive face recognition experiments on the CMU PIE and FRGC ver1.0 databases demonstrate that the misalignment-robust version algorithms consistently bring encouraging accuracy improvements over the counterparts without considering image misalignments, and also show the advantages of semi-supervised subspace learning over only supervised or unsupervised scheme.**

***Index Terms*—Dimensionality reduction, image misalignment, semi-supervised subspace learning, supervised subspace learning, unsupervised subspace learning.**

I. Introduction

DIMENSIONALITY reduction techniques [12] are significant for both data representation and classification in many computer vision applications. These algorithms can be roughly divided into three categories. The first category is unsupervised, which includes the pioneering work Principal Component Analysis (PCA) [11], and also involves most manifold learning algorithms, such as ISOMAP [17], Locally Linear Embedding (LLE) [14], as well as its linear extension Neighborhood Preserving Embedding (NPE) [9], [6], and Laplacian Eigenmaps (LE) [2] with its linear extension Locality Preserving Projections (LPP) [10]. The second category is supervised, and it utilizes the class label information for pursuing efficient representation for classification. Among them, the most popular ones are Linear Discriminant Analysis (LDA) [7] and its variants, including Nonparametric Discriminant Analysis (NDA) [8], Marginal Fisher Analysis (MFA) [19] and Local Discriminant Embedding (LDE) [4]. The third category is semi-supervised [3], [16] which addresses the problem on how to utilize unlabeled data for promoting supervised algorithms. A survey of semi-supervised learning is referred to [21]. In this work, we focus on linear dimensionality reduction, namely subspace learning, techniques owing to their simplicity and effectiveness. Despite of the variety of many popular subspace learning algorithms, Yan *et al.* [19] claimed that they can be mathematically unified within a general framework, called *Graph Embedding*. This framework derives a low-dimensional feature space which preserves the adjacency relationship among sample pairs in addition to constraints from scale normalization or a penalty graph [19].

In this paper, beyond the commonness in mathematical formulation [19] shared by different subspace learning algorithms, we claim that most popular subspace learning algorithms, no matter supervised or unsupervised, can be unitedly explained as instances of a ubiquitously supervised prototype, yet with specialized labeling approaches. The prototype pursues a low-dimensional subspace which minimizes the intraclass compactness and at the same time maximizes the interclass separability. This unified prototype provides new perspectives to understand many popular subspace learning algorithms. For example, PCA is a specific LDA, which regards each sample as a unique class; NPE is also supervised by propagating the label of each sample to the nearest sample in the subspace constructed by its k-nearest neighbors.

The proposed ubiquitous prototype is for general problems; when specific to computer vision applications, this prototype may suffer from the image misalignment issue, which is generally encountered in applications related with images. Here, misalignment means that the image cropping process is not always perfect, and there may exist translations in horizontal and vertical axes, rotation and scaling compared to the ideal cropping rectangle. In Fig. 1, we demonstrate the affection of image misalignments to PCA subspace. The results show that the small misalignments of the images may cause great changes to the subspaces, and the reconstructed images of the misaligned images from the well-aligned subspace are blurred. There were some attempts to deal with the misalignment issue within the context of unsupervised algorithms. Shashua [22] explored the possibility of using manifold pursuit to cope with image misalignments for PCA, and Frey *et al.* [23] proposed to learn a transformation invariant component model (including PCA and Independent Component Analysis) in a generative framework. However, these algorithms are unsupervised and, therefore, generally not as good as supervised algorithms for classification purpose. In this paper, we present a general

Manuscript received August 17, 2008. Current version published January 09, 2009. This work was supported in part by the U.S. Government VACE program and in part by AcRF Tier-1 Grant of R-263-000-464-112, Singapore. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Elad.

J. Yang and T. S. Huang are with the Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: jyang29@uiuc.edu; huang@ifp.uiuc.edu).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore (e-mail: eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.2009415

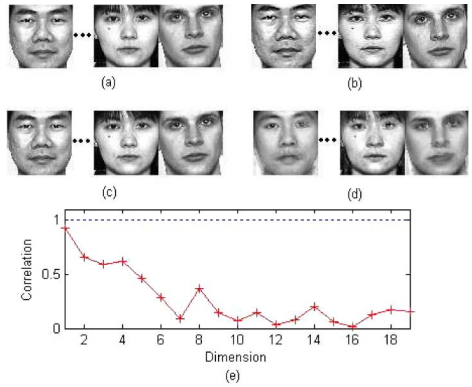

Fig. 1. Affection of image misalignments to PCA subspaces. (a) 20 well-aligned training images of size 64 by 64 pixels. (b) 20 slightly misaligned images with scaling less than 1 ± 0.05 and translation less than ± 2 pixels. (c) Reconstructed images for (a) using the PCA space learned from well-aligned training images. (d) Reconstructed images for (b) using the PCA space learned from well-aligned training images. (e) Correlation between the corresponding principal components of the two PCA spaces trained on well-aligned and misaligned images, which shows that two PCA spaces differ greatly in term of component-to-component similarity. Note that the dashed line is used to indicate the locations with correlation of 1.

solution for promoting algorithmic robustness for general subspace learning. First, we propose a general regularization term which characterizes the robustness of a subspace to the above four types of image misalignments. Then we justify that this objective function essentially characterizes the intraclass compactness, and, hence, this regularization term can be integrated into the ubiquitously supervised prototype, yielding the general formulation for misalignment-robust subspace learning. In addition, this ubiquitous prototype elicits us to develop the general formulation for semi-supervised subspace learning, which integrates the labels from both ground truth and the labeling approaches used by unsupervised learning algorithms.

The rest of the paper is organized as follows. Section II introduce the ubiquitously supervised prototype for most popular subspace learning algorithms, followed by the detailed justifications. The general formulations for misalignment-robust and semi-supervised subspace learning are introduced in Section III. Experimental results are presented in Section IV, and we conclude this paper in Section V.

II. Ubiquitously Supervised Subspace Learning

Here, we assume that the training data are given as $X = [x_1, x_2, \ldots, x_N]$ where $x_i \in \mathbb{R}^m$ and N is the total number of samples. For a supervised classification problem, like face recognition, the training data can be rearranged as $X^c = [x_0^c, x_1^c, \ldots, x_{n_c}^c], c = 1, 2, \ldots, N_c$, where n_c is the number of training samples for the cth class, x_0^c is the mean of the samples belonging to the cth class, namely $x_0^c = \sum_{i=1}^{n_c} x_i^c / n_c$, and we have $N = \sum_{c=1}^{N_c} n_c$. For ease of representation, we add a special class $X^{N_c+1} = [x_0^{N_c+1}]$ where $x_0^{N_c+1} = \bar{x}$, which is the mean of all training data, and, thus, we have $n_{N_c+1} = 0$. In practice, dimensionality reduction is in great demand owing to the fact that the effective information for classification often lies within a lower dimensional feature space.

A simple but effective way for dimensionality reduction is to find a column linearly independent matrix $P = [p_1, p_2, \ldots, p_d] \in \mathbb{R}^{m \times d}$, $\|p_k\| = 1, k = 1, 2, \ldots, d$, for transforming the original high-dimensional datum x into a low-dimensional form $y \in \mathbb{R}^d$ (usually $d \ll m$) as

$$y = P^T x. \tag{1}$$

Many algorithms, supervised or unsupervised, have been proposed for pursuing such a P. In this paper, we justify that most popular subspace learning algorithms can be unitedly explained as instances of a ubiquitously supervised prototype, and the diversities of these algorithms come from the labeling approaches, namely how to get class labels. In the following, we first introduce the common supervised subspace learning prototype, and then demonstrate in detail how the supervised and unsupervised subspace learning algorithms can be explained by this prototype along with specialized labeling approaches.

A. Ubiquitously Supervised Prototype

Two criterions can be used for measuring the potential classification capability of a low-dimensional feature space [19], namely, intraclass compactness and interclass separability.

Intraclass Compactness: The intraclass compactness is characterized by the weighted sum of the distances between sample pairs of the same class. Denote the nonnegative matrix $W^c \in \mathbb{R}^{(n_c+1)\times(n_c+1)}$ as the weight matrix measuring the importance of the data pairs within the cth class for characterizing intraclass compactness, and then the intraclass compactness can be measured by

$$\begin{aligned}\hat{S}_w(P) &= \sum_{c=1}^{N_c} \sum_{i=0}^{n_c} \sum_{j=0}^{n_c} \left\| P^T x_i^c - P^T x_j^c \right\|^2 W_{ij}^c \\ &= \mathrm{Tr}(P^T S_w P)\end{aligned} \tag{2}$$

where $S_w = \sum_{c=1}^{N_c} \sum_{i=0}^{n_c} \sum_{j=0}^{n_c} W_{ij}^c (x_i^c - x_j^c)(x_i^c - x_j^c)^T$ and $\mathrm{Tr}(\cdot)$ means the trace of a square matrix.

Interclass Separability: The interclass separability is characterized by the weighted sum of the distances between sample pairs of different classes. The larger is the sum, the greater is the potential classification capability. Note that the distances may also be measured between the samples and means. Denote the nonnegative matrix $W^{c_1,c_2} \in \mathbb{R}^{(n_{c_1}+1)\times(n_{c_2}+1)}$ as the weights between the data denoted as X^{c_1} and X^{c_2} for measuring interclass separability, and then the interclass separability can be formulated as

$$\begin{aligned}\hat{S}_b(P) &= \sum_{c_1 \neq c_2} \sum_{i=0}^{n_{c_1}} \sum_{j=0}^{n_{c_2}} \left\| P^T x_i^{c_1} - P^T x_j^{c_2} \right\|^2 W_{ij}^{c_1,c_2} \\ &= \mathrm{Tr}(P^T S_b P)\end{aligned} \tag{3}$$

where $S_b = \sum_{c_1 \neq c_2} \sum_{i=0}^{n_{c_1}} \sum_{j=0}^{n_{c_2}} W_{ij}^{c_1,c_2} (x_i^{c_1} - x_j^{c_2})(x_i^{c_1} - x_j^{c_2})^T$.

To obtain a low-dimensional feature space that is good for classification, it is desirable to minimize the intraclass compactness, namely $\min_P Tr(P^T S_w P)$, and at the same time maximize the interclass separability, namely $\max_P \mathrm{Tr}(P^T S_b P)$.

To simultaneously achieve the above two objectives, we can solve the following optimization problem [7]

$$P^* = \arg\min_P \frac{\mathrm{Tr}(P^T S_w P)}{\mathrm{Tr}(P^T S_b P)}. \tag{4}$$

In the following subsections, we will introduce in detail that most popular subspace learning algorithms can be justified as instances of the above prototype with specialized labeling approaches.

B. Supervised Algorithms

1) Linear Discriminant Analysis (LDA) [1]: **Objective function:** LDA searches for the directions that are the most effective for discrimination, by minimizing the ratio between the intraclass and interclass scatters, namely

$$\min_P \frac{\sum_{c=1}^{N_c}\sum_{i=1}^{n_c}\left\|P^T x_i^c - P^T x_0^c\right\|^2}{\sum_{c=1}^{N_c} n_c \left\|P^T x_0^c - P^T x_0^{N_c+1}\right\|^2}. \tag{5}$$

Justification: It is easy to verify that the numerator only includes the distance between samples and the corresponding means of the same class, while the denominator consists only the weighted distances between means of difference classes (here the total mean is considered to be a special class as mentioned above). Thus, $W^{c,c^{N+1}} = n_c, c = 1, 2, \ldots, N$ and $W^{c,c'} = 0$ for $c \neq c', c \leq N$ and $c' \leq N$. Hence, LDA is an instance of the ubiquitously supervised prototype as demonstrated in (2)–(4). The labels used are directly from the ground truths.

2) Marginal Fisher Analysis (MFA) [19]: **Objective function:** By removing the underlying assumption of LDA, that is, the samples of each class follow Gaussian distribution, MFA provides a new way to measure the intraclass compactness and interclass separability in a nonparametric way [19]

$$\min_P \frac{\sum_{c=1}^{N_c}\sum_{j\in N_{k_1}^c(i) \text{ or } i\in N_{k_1}^c(j)}\left\|P^T x_i^c - P^T x_j^c\right\|^2}{\sum_c \sum_{(i,j)\in P_{k_2}(c) \text{ or } (j,i)\in P_{k_2}(c')}\left\|P^T x_i^c - P^T x_j^{c'}\right\|^2} \tag{6}$$

where $N_{k_1}^c(i)$ indicates the index set of the k_1 nearest neighbors of the sample x_i^c within the cth class, and $P_{k_2}(c)$ is a set of data pairs that are the k_2 nearest pairs among the set $\{(x_i^c, x_j^{c'}), c \neq c', \forall\, i, j, c'\}$.

Justification: In the objective function of MFA, the numerator is the sum of distances between the data pairs, one of which is among the k_1 nearest neighbors of another datum within the same class; while the denominator is the sum of the distances between the marginal sample pairs from different classes. Hence, MFA is also a special case of (2)–(4); and similar to LDA, the class labels are from ground truths.

C. Unsupervised Algorithms

1) Principal Component Analysis (PCA) [11]: **Objective function:** PCA searches for the projection directions which can best reconstruct the original data along with the data mean; equivalently, these projection directions yield the maximal variations for the training data, and the optmization problem is

$$\max_P \sum_{i=1}^{N} ||P^T(x_i - \bar{x})||^2 = \min_P \frac{1}{\sum_{i=1}^{N} ||P^T(x_i - \bar{x})||^2}. \tag{7}$$

Justification: As seen from the above formula, the objective function of PCA is in a very similar form as the proposed supervised prototype except for the constant in the numerator. Does there exist certain relationship between PCA and supervised algorithm like LDA? Here, we justify that PCA is essentially a specific LDA with *self-labeling* approach according to the following explanations. First, each data is considered to constitute a unique class, and then the labeled datum is denoted as $X^c = [x_0^c, x_1^c], c = 1, 2, \ldots, N$, where $x_0^c = x_1^c = x_c$; consequently, we have the mean data $X^{N+1} = [\bar{x}]$. Based on this explanation, the (2) equals to zero, and the maximization of (3) equals to minimize the objective function of PCA in (7) with the assumption that $W^{c,c^{N+1}} = n_c = 1, c = 1, 2, \ldots, N$ and $W^{c,c'} = 0$ for $c \neq c'$, $c \leq N$ and $c' \leq N$. Therefore, PCA is also a specific instance of the ubiquitously supervised prototype.

2) Locality Preserving Projections (LPP) [10]: **Objective function:** LPP finds an embedding that preserves local information and best detects the essential manifold structure of the data set. Its objective function is

$$\arg\min_P \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\|P^T x_i - P^T x_j\|^2 W_{ij}}{\sum_{i=1}^{N} D_{ii}\|P^T x_i - P^T \bar{x}\|^2} \tag{8}$$

where the weight matrix W is defined as $W_{ij} = \exp\{-\|x_i - x_j\|^2/t\}$ if $i \in N_k(j)$ or $j \in N_k(i)$; 0, otherwise. $N_k(i)$ means the indexes of the k nearest neighbors of sample x_i and t is a parameter for defining similarity. D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. Note that the denominator is $\sum_{i=1}^{N} D_{ii}\|P^T x_i\|^2$ instead of $\sum_{i=1}^{N} D_{ii}\|P^T x_i - P^T \bar{x}\|^2$ in [10], and here we refine it for two reasons: 1) it naturally satisfies the requirement of zero mean as in [2], the precursor of LPP; and 2) the online available code of [10] practically removes the mean for all the samples in the preprocessing PCA step, and, hence, the implementation follows our formulation of LPP here.

Justification: Besides acting as a popular manifold learning algorithm, LPP can also be explained a specific multilabeled supervised subspace learning algorithm by propagating labels via the neighboring relationship. More specifically speaking, first, each datum is considered to constitute a unique class, then the label of each sample is propagated to its k nearest neighbors, and consequently each sample has multiple labels. For each class, $X^c = [x_0^c, x_1^c, \ldots, x_{n_c}^c], c = 1, 2, \ldots, N$, where x_1^c is the concerned sample and $x_j^c, j = 2, 3, \ldots, n_c$, are the neighbors of x_1^c or the samples whose neighbors include x_1^c. An illustration of this label propagation process in LPP is shown in Fig. 2. With the multilabel interpretation, the numerator of the objective function characterizes the sum of the weighted (with heat kernel function) distance between a sample and its k nearest

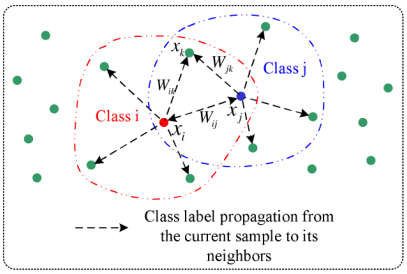

Fig. 2. Illustration of label propagation within neighborhood for LPP.

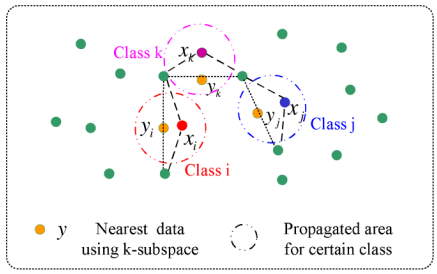

Fig. 3. Illustration of label propagation by k-subspace approximation for NPE.

samples, while the denominator characterizes the interclass separability as in (3) (note that the D_{ii} acts similarly as n_c in LDA).

3) Neighborhood Preserving Embedding (NPE) [9]: **Objective function:** NPE is the linear extension of LLE [14], and it seeks a low-dimensional feature space which preserves the locality reconstruction relationship. The optimization problem is defined as

$$\min_P \frac{\sum_{i=1}^N \left\| P^T x_i - \sum_{j\in N_k(i)} P^T W_{ij} x_j \right\|^2}{\sum_{i=1}^N \|P^T x_i - P^T \bar{x}\|^2} \tag{9}$$

where the matrix W is computed through minimizing $\sum_{i=1}^N \|x_i - \sum_{j\in N_k(i)} W_{ij} x_j\|^2$, s.t. $\sum_j W_{ij} = 1$. The definition of $N_k(i)$ is the same as in LPP. Similar to LPP, the reasons to subtract the total mean in the denominator are two-fold: 1) it naturally satisfies the requirement of zero mean as in [14], the precursor of NPE; and 2) the online available code of [9] practically removes the mean for all the samples in the preprocessing PCA step, and the implementation follows our formulation here.

Justification: Denote $y_i = \sum_{j=1}^N W_{ij} x_j$, then the objective function of NPE can be rewritten as

$$\min_P \frac{\sum_{i=1}^N \|P^T x_i - P^T y_i\|^2}{\sum_{i=1}^N \|P^T x_i - P^T \bar{x}\|^2}. \tag{10}$$

Note that $S_i = \{\sum_{j\in N_k(i)} W_{ij} x_j;\ \sum_{j\in N_k(i)} W_{ij} = 1\}$ constitutes a subspace plus an offset, and we call it k-subspace here.

NPE can be interpreted in the supervised manner as follows. First, each datum is considered to constitute a unique class, and then the label of each sample is propagated to the nearest datum within the k-subspace S_i constructed by its k nearest samples. The resulting data set for the cth class is $X^c = [x_0^c, x_c, y_c]$, and $X^{N+1} = [\bar{x}]$. Then the numerator is the sum of distances between samples of the same class while the denominator is the sum of distances between samples and the total mean. Therefore, NPE is also an instance of ubiquitously supervised prototype by propagating label with k-subspace approximation. An illustration of this explanation is displayed in Fig. 3.

D. Discussions

As described above, most popular supervised or unsupervised subspace learning algorithms can be explained in a supervised manner. Here, we would like to highlight some aspects of this unified supervised understanding.

1) Instead of a unified mathematical formulation of graph embedding as demonstrated in [19], we concern a unified learning type in this work, and justify that most popular subspace learning algorithms are essentially supervised except for different labeling approaches.
2) Unlike the work of Yan *et al.* [19], our ubiquitously supervised prototype provides a supervised perspective to understand unsupervised subspace learning algorithms; and these unsupervised algorithms are equivalent to the specific labeling preprocessing approaches followed by instantiations of the ubiquitously supervised prototype.
3) The unified explanation of both supervised and unsupervised learning algorithms generalized the concept of intraclass compactness and interclass separability with flexible labeling approaches, which allows one to formulate misalignment-robust subspace learning and semi-supervised subspace learning in the following section.

III. Two General Formulations Extended From Ubiquitous Prototype

The above ubiquitously supervised prototype and its instances are generally proposed based on the assumption that all those features are strictly aligned, namely, each dimension of the feature has specific meaning. However, in many computer vision applications, misalignment is a common issue. For example, in face recognition and hand written character recognition, the cropping process is often not perfect. Subspace learning algorithms often suffer from such misalignment issue, which may severely degrade the algorithmic generalization capability, since the image misalignment may make the distribution of the testing data greatly different from that of the training data. Although there exist many methods to deal with this problem such as bags of features or convolutional neural network, we focus on subspace learning algorithms. In this section, we provide a general solution for improving the robustness and the generalization capability of the ubiquitously supervised prototype to the image misalignment issues encountered in vision problems. In addition, we will introduce a general formulation to semi-supervised subspace learning by naturally integrating the label information from both ground truth and the labeling approaches used by unsupervised learning algorithms.

A. Misalignment-Robust Subspace Learning

Image alignment is critical for extracting stable and robust features in many computer vision applications such as face recogni-

tion; however, the misalignment issues, including translations in horizontal and vertical axes, scaling and rotation, are inevitable especially for practical systems. Current subspace learning algorithms mostly are sensitive to image misalignments as shown in Fig. 1, yet few works have been devoted to explicitly solving this problem. In this section, we provide a general solution to this problem for all the subspace learning algorithms that can be explained by the ubiquitously supervised prototype.

1) Misalignment-Robust Regularization Term: To achieve the misalignment robustness, there are several possible solutions. One is to re-align the images with certain generative models, and Tu's work [18] belongs to this type. Another is to seek a subspace which is robust to image misalignment. Our solution belongs to this type. Since image misalignment distribution is often roughly predictable, we define the misalignment-robust regularization term in a probabilistic manner as

$$\sum_{i=1}^{N} \int_{T,r,\theta} \|P^T x_i - P^T x_i(T,r,\theta)\|^2 p(T,r,\theta)\, dT\, dr\, d\theta \tag{11}$$

where $x_i(T,r,\theta)$ is the transformed image of x_i from the misalignment parameters, $T = [T_x, T_y]$ are the translation variables, and we assume they distribute within $[-2,2] \times [-2,2]$ on pixel level; r is the scaling factor distributed within [0.9, 1.1] and θ is the rotation angle distributed within $[-5^\circ, 5^\circ]$;[1] $p(T,r,\theta)$ is the probability distribution function defined on T, r, and θ. We assume that T_x, T_y, r and θ are independent; then $p(T,r,\theta) = p(T_x)p(T_y)p(r)p(\theta)$. We assume that each parameter roughly follows Gaussian distribution yet only validate with the areas mentioned above.

Equation (11) can be computed out as

$$\begin{aligned}
&\hat{S}_m(P) \\
&= \sum_{i=1}^{N} \int_{T,r,\theta} \|P^T x_i - P^T x_i(T,r,\theta)\|^2 \\
&\quad \times p(T,r,\theta)\, dT\, dr\, d\theta \\
&= \sum_{i=1}^{N} \int_{T,r,\theta} \mathrm{Tr}((P^T x_i - P^T x_i(T,r,\theta))(P^T x_i \\
&\quad - P^T x_i(T,r,\theta))^T) p(T,r,\theta)\, dT\, dr\, d\theta \\
&= \sum_{i=1}^{N} \int_{T,r,\theta} \mathrm{Tr}(P^T (x_i - x_i(T,r,\theta))(x_i - x_i(T,r,\theta))^T P) \\
&\quad \times p(T_x)p(T_y)p(r)p(\theta)\, dx\, dy\, dr\, d\theta \\
&= \mathrm{Tr}\left(P^T \left(\sum_{i=1}^{N} \int_{T,r,\theta} (x_i - x_i(T,r,\theta))(x_i \right.\right. \\
&\quad \left.\left. - x_i(T,r,\theta))^T p(T_x)p(T_y)p(r)p(\theta)\, dx\, dy\, dr\, d\theta \right) P \right) \\
&= \mathrm{Tr}(P^T S_m P).
\end{aligned}$$

[1]These parameters are set according to the statistics of the misalignments based on our own face alignment algorithm similar to ASM [5].

Using the above definition, we basically exploit all data samples with reasonable image misalignment for each data point. The term $\hat{S}_m(P)$ can be considered as characterizing the intraclass compactness if we label the synthesized misaligned images with the same label as the original one.

2) General Misalignment-Robust Subspace Learning: Based on the ubiquitously supervised prototype in (4) and the misalignment-robust regularization term in (12), we have the general formulation for misalignment-robust subspace learning as

$$\begin{aligned}
&\min_P \frac{\mathrm{Tr}(P^T S_w P) + \lambda \mathrm{Tr}(P^T S_m P)}{\mathrm{Tr}(P^T S_b P)} \\
&\qquad = \min_P \frac{\mathrm{Tr}(P^T (S_w + \lambda S_m) P)}{\mathrm{Tr}(P^T S_b P)}.
\end{aligned} \tag{12}$$

The integration of the two terms can more truthfully characterize the intraclass compactness for the misaligned data, which brings robustness to general subspace learning algorithms.

3) Implementation Details: In the implementation, we utilize some strategies for promoting computational efficiency of (11). We use triangle distribution to approximate the gaussian distribution, and for the integral of r and θ, we use the sampling approach for approximating the integral. In our experiments, the computation cost for S_m is several seconds for each image with unoptimized Matlab code on a PC with 2.8-HZ CPU and 1-G memory.

The objective function in (12) is nonlinear, and commonly there does not exist closed form solution. In this work, as conventionally [7], we transform the objective function in another form

$$\max_P \mathrm{Tr}[(P^T (S_w + \lambda S_m) P))^{-1} (P^T S_b P)] \tag{13}$$

and the generalized eigenvalue decomposition method is used for deriving the projection matrix P [7].

4) Discussions: One naive way of dealing with image misalignments might be to add more synthesized samples to the training set by varying the parameters of translation, rotation and scaling. Adding more synthesized samples to the training set might help to some extent, but our solution is more general and elegant:

1) Generally, directly adding synthesized samples is applicable only for supervised subspace learning, where the synthesized samples can be labeled as the same class with the original datum from which they are generated. However, for unsupervised subspace learning, the class label is not explicitly used, so misalignment robustness cannot be well achieved in these algorithms. This idea is illustrated using PCA as in Fig. 1. The synthesized samples added will be far away from the original datum according to the criteria of PCA, which is inconsistent with the target of misalignment robustness. In contrast, our algorithm can label the misaligned samples with the original one and find a projection that make the misaligned data as compact as possible while at the same time keeps to the criteria of the original PCA.
2) Our misalignment-robust regularization term utilizes the information from all possible misalignments by integral, while the way of directly adding synthesized samples is

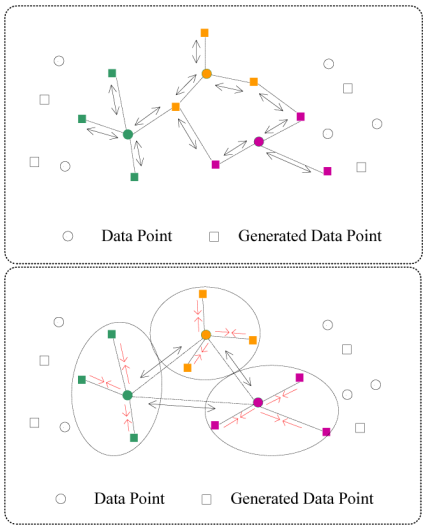

Fig. 4. Latent difference among PCAs for the cases. Left: Directly adding synthesized samples. Right: Misalignment-robust regularization term. Note that, for (a), the samples synthesized from a datum are expected to be far away from the datum after PCA, while for (b), the samples virtually sampled around a datum are expected to be close to the datum after misalignment-robust PCA.

limited by the number of synthesized samples that can be added. For examples, for a database with 10 000 training samples as in our experiments, the computational cost will be prohibitively high for algorithms like MFA, even we only add 100 synthesized samples for each datum. The computational cost of MFA mainly lies in the construction of the intrinsic and penal graphs, and adding only 100 synthesized samples for each datum will increase the computational cost and memory requirement for the graphs by 10^4 times. In our framework, the matrix S_w is computed in the same way as in the case without considering image misalignments, and the matrix S_m is computed from the integral process in (12).

3) Our algorithm is more flexible with the parameter λ before the misalignment robust regularization term. This parameter can be adjusted to the incoming data. For example, if the misalignment for the incoming data is severe, λ should be larger to penalize the effects of the misalignment.

B. General Formulation for Semi-Supervised Subspace Learning

Semi-supervised learning recently attracted much attention, and was widely used for regression and classification problems [21]. The main idea of semi-supervised learning is to utilize unlabeled data for improving the classification and generalization capability on the testing data. Fig. 5 illustrates the potential of the unlabeled data in dimensionality reduction for classification. Commonly the unlabeled data is utilized as an extra regularization term in the objective function of traditional supervised learning algorithms. In this work, motivated by the unified supervised explanations for most popular subspace learning algorithms, we present a general formulation for semi-supervised subspace learning, and the objective functions of unsupervised manifold learning algorithms, such as LPP and NPE, are used

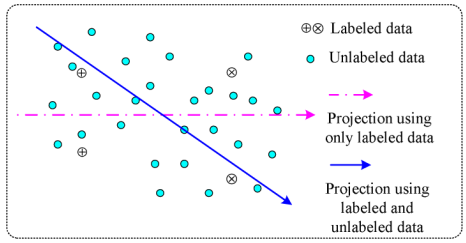

Fig. 5. Unlabeled data is useful for promoting classification capability.

as the regularization terms for supervised subspace learning algorithms. That is

$$\min_P \frac{\gamma \mathrm{Tr}(P^T S_w P) + (1-\gamma)\mathrm{Tr}\left(P^T S_w^r P\right)}{\mathrm{Tr}(P^T S_b P)} = \min_P \frac{\mathrm{Tr}(P^T \tilde{S}_w P)}{\mathrm{Tr}(P^T S_b P)} \quad (14)$$

where $\mathrm{Tr}(P^T S_w^r P)$ comes from the intraclass compactness terms of LPP or NPE, $\gamma \in (0,1)$ is a coefficient to balance the supervised term and manifold regularization term, and $\tilde{S}_w = \gamma S_w + (1-\gamma) S_w^r$. Note that the matrix S_w and S_b are computed from the labeled data as in (2) and (3) given the weight matrices or graphs, while the term S_w^r is computed from both labeled and unlabeled data as in (2). The intraclass matrix $\tilde{S}_w$ reflects the variance for both labeled and unlabeled data. The advantages of the semi-supervised learning algorithm will be validated by experiments as demonstrated in Section IV.

It is worthy to highlight that, compared to other specific algorithms for semi-supervised learning, our contribution is to offer a unified framework and solution, and the recent proposed algorithm, Semi-Supervised Discriminant Analysis (SDA) [24], is a special case of this framework. The SDA algorithm is motivated from using regularization for LDA when the labeled training set is too small, and its objective function is

$$\max_P \frac{\mathrm{Tr}(P^T S_b P)}{\mathrm{Tr}(P^T S_w P) + \alpha J(P)} \quad (15)$$

where $J(P)$ is the regularizer on the projection matrix P. In [24], the authors used the Laplacian Graph [2] to compute $J(P)$ as

$$J(P) = \mathrm{Tr}(P^T X L X^T P) = \mathrm{Tr}\left(P^T S_w^r P\right) \quad (16)$$

where L is the Laplacian matrix [25]. If we choose LDA graph to compute S_w and S_b, and choose Laplacian graph to compute S_w^r in our general formulation (14), it is exactly SDA.

IV. Experiment Results

To evaluate the effectiveness of the proposed general framework for misalignment-robust subspace learning, we systematically compare the original algorithms with their misalignment-robust counterparts for PCA, LDA, MFA, and NPE. Two popular face databases CMU PIE [15] and FRGC Ver1.0 [13] are used to for the comparison. Also, the advantages of general semi-supervised subspace learning over only supervised or unsupervised scheme are verified by integrating the MFA and NPE algorithms. In all the experiments, the face recognition is

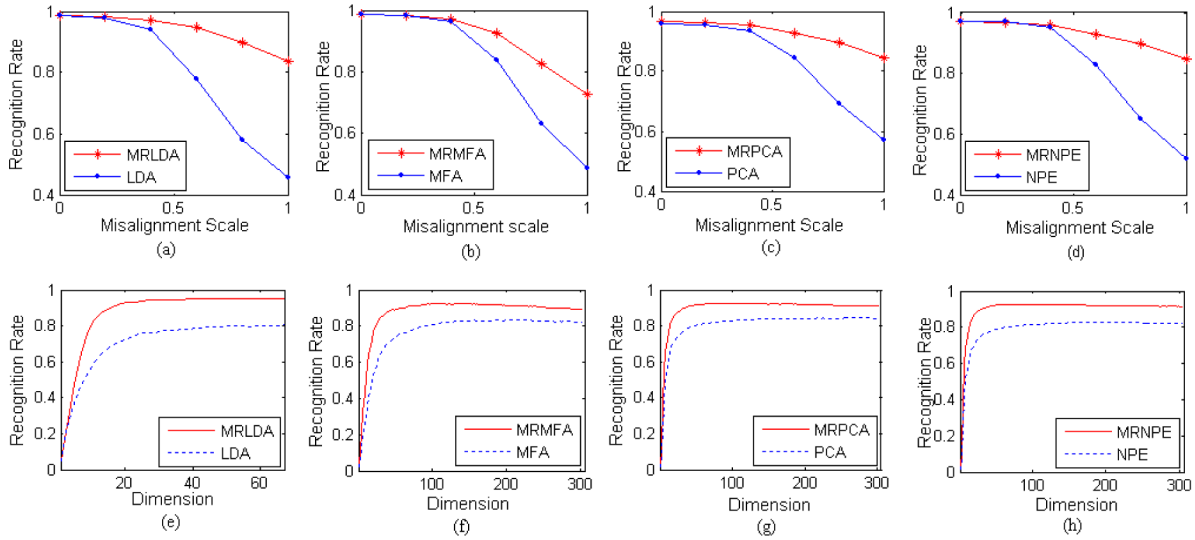

Fig. 6. Comparison between the original version and misalignment-robust counterparts of LDA, MFA, PCA, and NPE on the CMU PIE database with well-aligned training set. Top: Face recognition accuracy versus misalignment scale. Down: Face recognition accuracy versus feature dimension at the misalignment scale of 0.6.

conducted by using Nearest Neighbor as the final classifier for simplicity.

A. Data Set Preparation

The CMU PIE (Pose, Illumination and Expression) database contains 41 368 images of 68 people under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiment, the data set contains five near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. So, there are 170 images for each individual and more than 10 000 images are used for the experiments. The images are well aligned by manually marking the positions of two eyes, and finally cropped to the size of 32-by-32 pixels.

FRGC Ver1.0 is a database from FRVT2006 evaluation, and contains 275 persons and totally 5658 images of different resolutions. In this work, we implemented a face alignment algorithm similar to Active Shape Models [5] to automatically align the faces, and then crop the faces to the size of 32-by-32 pixels based on the derived eye positions. Hence, unlike CMU PIE database, there exist misalignments in FRGC Ver1.0.

B. Misalignment Robust Subspace Learning

In this section, we will introduce the experimental results on two databases in two scenarios, and then discuss the parameter selection problem of λ.

1) Inhomogeneous Misalignment Scenario: CMU PIE Database: For the CMU PIE database, we randomly select 100 images of each person for model training, and the left 70 images for testing. As the images are well-aligned in the CMU PIE database, we synthesize different scales of misalignments for testing data. The scale 1 corresponds to the largest possible misalignment with up to ± 2 pixels in translations, $\pm 2^{\circ}$ rotation, and the scaling between $[0.95\ 1.05]$. Two sets of experimental results of the algorithms LDA, MFA, PCA, and NPE, as well as their misalignment-robust counterparts are displayed in Fig. 6. The first set (a), (b), (c), and (d) demonstrate the influence of misalignments to different algorithms, and the second set (e), (f), (g), and (h) show the algorithmic performance on different feature dimension in the case of testing data with misalignment of scale 0.6. The results demonstrate that our proposed algorithms can handle subspace mismatch issue caused by misalignments much better than the counterparts without considering the misalignment issue. Note that in this work, we do not focus on evaluating which subspace learning algorithm is the best, instead we evaluate the algorithmic robustness to image misalignments; hence, we only implement the most popular subspace learning algorithms for comparison. For all the experiments, we first conduct PCA to reduce the feature to a dimension by retaining 98% of the energy, and then compare all the algorithms on this dimensionality reduced feature space. For NPE, the nearest neighbor number is set as 5, and the nearest neighbor numbers for within-class and between-class are set as 5 and 20 for MFA in all the experiments.

2) Homogeneous Misalignment Scenario: FRGC Ver1.0: For the FRGC Ver1.0 database, we randomly select half of the images of each person for model training, and the left half for testing. These images are automatically aligned and cropped, and, hence, there exist misalignments for both training and testing data. The comparison results of the original version and the misalignment-robust counterparts of LDA, MFA, PCA, and NPE are displayed in Fig. 7. As we can see, our misalignment robust algorithms all outperform the original ones. That means, our algorithms can learn a better subspace that is robust to image misalignments.

3) Parameter Selection: In all the above experiments, we need to choose the parameter λ as in (12). Generally speaking, the best λ should be roughly proportional to the misalignment degree. In Fig. 8, we show the influence of the weight λ on face recognition accuracy for three algorithms we have talked about on the two databases. As we can see, overall the misalignment algorithms are not very sensitive to the parameter; in a large

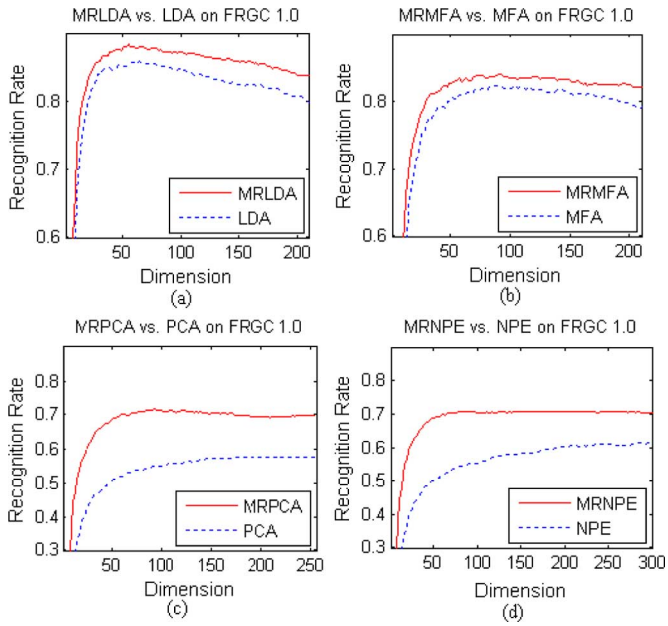

Fig. 7. Comparison of the original version and misalignment-robust counterpart of PCA, NPE, LDA, and MFA on the misaligned FRGC Ver1.0 database.

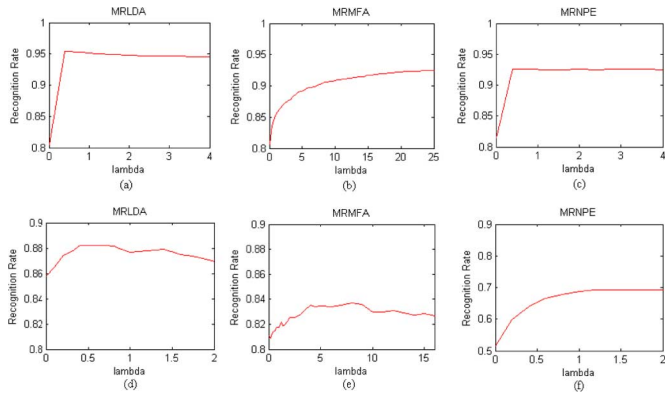

Fig. 8. Affection of the weight parameter λ for misalignment-robust algorithms. Figures (a), (b), and (c) are results on CMU PIE with 0.6 scale misalignment, and (d), (e), and (f) are results on FRGC 1.0.

range, these algorithms are much better than the corresponding algorithms without the regularization term $(\lambda = 0)$. However, different algorithms are different in sensitivity to the parameter: MFA is the least sensitive algorithm to λ. In real applications, λ depends on the incoming data or the accuracy of the alignment algorithm and should be chosen empirically.

C. Effectiveness of Semi-Supervised Subspace Learning

We take as an example the $\mathrm{MFA}+\mathrm{NPE}$ for semi-supervised subspace learning. For the CMU PIE database, the first 20 images of each person are used for model training, and the 80 images thereafter are used as unlabeled data for semi-supervised learning and final classification.[2] For FRGC Ver1.0 database, we select 2 images for those subjects with only 6 images, and one quarter of the images for others, as the training data, and the rest are used as unlabeled data for semi-supervised subspace learning and final classification. The comparison of MFA, NPE, and the semi-supervised algorithm $\mathrm{MFA}+\mathrm{NPE}$ are demonstrated in Fig. 9, and the results show that the our semi-supervised learning brings encouraging accuracy improvements over MFA and NPE alone.

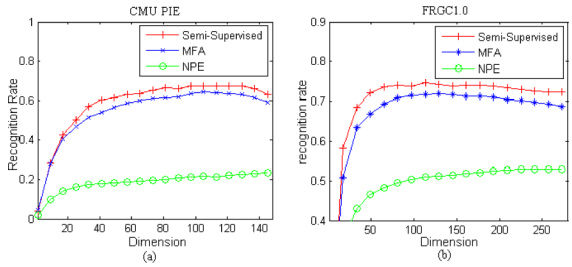

Fig. 9. Semi-Supervised subspace learning using MFA and NPE on CMU PIE and FRGC Ver1.0 databases.

V. Conclusion

In this paper, we proposed a ubiquitously supervised prototype which unitedly explains most popular subspace learning algorithms as its instances. Then this prototype was further enhanced for computer vision applications to obtain the robustness to image misalignments, and consequently the potential algorithmic generalization capability was promoted. To the best of our knowledge, it is the first work dedicated to tackling and providing general solution to the image misalignment problem encountered by most subspace learning algorithms in computer vision applications. In addition, a general formulation for semi-supervised subspace learning was presented by naturally integrating the labels from both ground truth and the labeling approaches used by unsupervised learning algorithms.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Adv. Neural Inf. Process. Syst.*, pp. 585–591, Dec. 2001.

[3] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn. J.*, vol. 56, pp. 209–239, 2004.

[4] H. Chen, H. Chang, and T. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, vol. 2, pp. 846–852.

[5] T. Cootes, G. Edwards, and C. Taylor, "Comparing active shape models with active appearance models," in *Proc. Brit. Machine Vision Conf.*, 1999, vol. 1, pp. 173–182.

[6] Y. Fu and T. Huang, "Graph embedded analysis for head pose estimation," in *Proc. 7th Int. Conf. Automatic Face and Gesture Recognition*, Apr. 2006, pp. 3–8.

[7] K. Fukunnaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1991.

[8] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1983, vol. 5, no. 6, pp. 671–678.

[2]We did not use all the data for semi-supervised learning due to the impractical computational cost for constructing the graph for NPE.

[9] X. He, D. Cai, S. Yan, and H. Zhang, "Neighborhood preserving embedding," in *Proc. 10th Int. Conf. Computer Vision*, Oct. 2005, vol. 2, pp. 1208–1213.
[10] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[11] I. Joliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.
[12] A. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[13] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 947–954.
[14] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 22, pp. 2323–2326, 2000.
[15] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," in *Proc. Eur. Conf. Computer Vision*, 2003, vol. 25, pp. 1615–1618.
[16] V. Sindhwani, P. Niyogi, M. Belkin, and S. Keerthi, "Linear manifold regularization for large scale semi-supervised learning," in *Proc. ICML Workshop on Learning with Partially Classified Training Data*, 2005, pp. 80–83.
[17] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2322, Dec. 2000.
[18] J. Tu, A. Ivanovic, X. Xu, L. Fei-Fei, and T. Huang, "Variational shift invariant probabilistic PCA for face recognition," in *Proc. 18th Int. Conf. Pattern Recognition*, 2006, vol. 3, pp. 548–551.
[19] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[20] J. Ye, R. Janardan, C. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 982–994, Aug. 2004.
[21] X. Zhu, Semi-Supervised Learning Literature Survey, Dept. Comput. Sci., Univ. Wisconsin, Tech. Rep. 1530, 2005.
[22] A. Shashua, A. Levin, and S. Avidan, "Manifold pursuit: A new approach to appearance based recognition," presented at the Int. Conf. Pattern Recognition, 2002.
[23] B. Frey and N. Jojic, "Transformed component analysis: Joint estimation of spatial transformations and image components," presented at the IEEE Int. Conf. Computer Vision, 1999.
[24] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," presented at the IEEE Int. Conf. Computer Vision, 2007.
[25] F. R. K. Chung, "Spectral graph theory," presented at the Reg. Conf. Ser. Mathematics, 1997.

Jianchao Yang (S'08) received the B.E. degree from the Department of Electronics Engineering and Information Science, University of Science and Technology of China (USTC), China, in 2006. He is currently pursuing the Ph.D. degree at the University of Illinois at Urbana-Champaign (UIUC), Urbana.

Since Fall 2006, he has been with the Department of Electrical and Computer Engineering, UIUC. He is currently working with Prof. T. S. Huang on his Ph.D. degree. His research interests include image processing, computer vision, and machine learning.

Shuicheng Yan (M'06) received the B.S. and Ph.D. degrees from the Applied Mathematics Department, School of Mathematical Sciences, Peking University, China, in 1999 and 2004, respectively.

His research interests include computer vision and machine learning. Currently, he is an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore.

Thomas S. Huang (LF'01) received his B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., and the M.S. and D.Sc. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the faculty of the Department of Electrical Engineering, MIT, from 1963 to 1973, and on the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University, West Lafayette, IN, from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, Urbana, where he is now the William L. Everitt Distinguished Professor of Electlrical and Computer Engineering, Research Professor in the Coordinated Science Laboratory, Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology, and Co-Chair of the Institute's major research theme Human Computer Intelligent Interaction. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 20 books, and over 500 papers in network theory, digital filtering, image processing, and computer vision.

Dr. Huang is a member of the National Academy of Engineering; a Foreign Member of the Chinese Academies of Engineering and Sciences; a Fellow of the International Association of Pattern Recognition and the Optical Society of American; and has received a Guggenheim Fellowship, an A.V. Humboldt Foundation Senior U.S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis". In 2001, he received the IEEE Jack S. Kilby Medal. In 2002, he received the King-Sun Fu Prize, International Association of Pattern Recognition, and the Pan Wen-Yuan Outstanding Research Award. In 2005, he received the Okawa Prize. In 2006, he was named by IS&T and SPIE as the Electronic Imaging Scientist of the year. He is a Founding Editor of the *International Journal Computer Vision, Graphics, and Image Processing* and Editor of the Springer Series in Information Sciences, published by Springer Verlag.