Locality Versus Globality: Query-Driven Localized Linear Models for Facial Image Computing

Yun Fu, *Member, IEEE*, Zhu Li, *Senior Member, IEEE*, Junsong Yuan, *Student Member, IEEE*, Ying Wu, *Senior Member, IEEE*, and Thomas S. Huang, *Life Fellow, IEEE*

***Abstract*—Conventional subspace learning or recent feature extraction methods consider *globality* as the key criterion to design discriminative algorithms for image classification. We demonstrate in this paper that applying the *local* manner in sample space, feature space, and learning space via linear subspace learning can sufficiently boost the discriminating power, as measured by discriminating power coefficient (DPC). The proposed solution achieves good classification accuracy gains and shows computationally efficient. Particularly, we approximate the global nonlinearity through a multimodal localized piecewise subspace learning framework, in which three locality criteria can work individually or jointly for any new subspace learning algorithm design. It turns out that most existing subspace learning methods can be unified in such a common framework embodying either the global or local learning manner. On the other hand, we address the problem of numerical difficulty in the large-size pattern classification case, where many local variations cannot be adequately handled by a single global model. By localizing the modeling, the classification error rate estimation is also localized and thus it appears to be more robust and flexible for the model selection among different model candidates. As a new algorithm design based on the proposed framework, the query-driven locally adaptive (QDLA) mixture-of-experts model for robust face recognition and head pose estimation is presented. Experiments demonstrate the local approach to be effective, robust, and fast for large size, multiclass, and multivariance data sets.**

***Index Terms*—Discriminating power coefficient (DPC), face recognition, globality, head pose estimation, human-centered computing (HCC), locality, mixture-of-experts model, subspace learning.**

Manuscript received August 27, 2007; revised November 21, 2007 and January 11, 2008. First published September 16, 2008; current version published November 26, 2008. This work was supported in part by the U.S. Government VACE program and in part by the National Science Foundation (NSF) under Grant CCF 04-26627. The views and conclusions are those of the authors, not of the U.S. Government or its Agencies. This paper was recommended by W. Zhu.

Y. Fu is with the BBN Technologies, Cambridge, MA 02138 USA (e-mail: yfu@bbn.com).

Z. Li is with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: zhu.li@ieee.org).

J. Yuan and Y. Wu are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: j-yuan@u.northwestern.edu; yingwu@ece.northwestern.edu).

T. S. Huang is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801 USA (e-mail: huang@ifp.uiuc.edu).

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/TCSVT.2008.2004933

I. Introduction

HUMAN-CENTERED COMPUTING (HCC), a recently emerging field, tightly connects various methodologies that are applied to computing system designs and implementations with personal, social and cultural contexts of human activities for human machine interaction [26]. As a key human identity, human face in multimedia data, such as images and videos, has attracted much attention and plays essential roles in supporting human activity analysis.

The difficulty of appearance-based face analysis, such as recognition [14] and pose estimation [6], stems from the difficulty in statistical modeling of face images under pose, scale, expression, occlusion, and illumination variations. Conventional global linear models [16] under the Gaussian-assumption [17], such as principal components analysis (PCA) [20] and linear discriminant analysis (LDA) [27], have been found to be effective in either an unsupervised or supervised manner. Especially, the incorporation of the labeled data improves the performance by finding subspaces where discriminative features are preserved, while nondiscriminative features are dropped. However, for the real-world system design, such kind of linear models in a global manner are fundamentally limited because of the complexity in local variations and the nonlinearity of the underlying face manifold structure. To further enhance the recognition accuracy, better subspace learning methods with higher discriminative ability need to be developed considering local modeling.

In recent years, advances have focused on nonlinear methods (e.g., kernel based method [22], [23] or manifold learning [2], [5]) and graph modeling method (e.g., GE [24]). By applying a nonlinear kernel mapping or a locally preserved graph embedding, the discriminating power of the feature representation can be significantly improved. However, a complex numerical problem arises because the covariance modeling in kernel method or the neighborhood graph in the graph modeling method is typically of $N \times N$ dimension, where N is the number of the labeled training samples. When N is very large, the solution can be unstable and the operation on the $N \times N$ matrix is computationally impractical. Thus, more efficient methods are desired. Moreover, the manifold learning solutions are nonlinear functions depending on the training data, and cannot directly handle unknown test data. For example, it is difficult to embed a new query point into the learned nonlinear manifold without recalculating the embedding with the whole data set. The linearization of the nonlinear manifold learning or graph modeling methods, e.g., locality preserving projections

(LPP) [3] and locally embedded analysis (LEA) [5], partially solves this problem by finding a linear subspace embodying the graph embedding objective. Even though the solutions have better performance than pure Euclidean metric based approaches like PCA and LDA, they are still global and linear solutions. When the sample space is large, the subspace/metric learning performance decreases, since the large data set defeats the discriminating power of the locality preserved graph embedding model. Furthermore, a single global model that tries to capture all the variations in the sample space has serious limitations. To fit the large size training data, a kernel mapping original data to a much higher dimensional space is required, which sometimes results in a complex decision boundary with a poor generalization ability and often involves time-consuming optimization or computation of pairwise distances.

A. Our Work

To tackle the mentioned problems, we present new feature extraction algorithms based on a localized piecewise linear subspace learning framework. The basic idea is to let the subspace learning procedure embody local structure and characteristic, which has been demonstrated to effectively boost the discriminating power of the feature extraction [5], [7]–[9] in several aspects. The SVM-KNN in [25] is the most related work to our ideas. However, since it is an non-trivial issue to find informative nearest neighbors in high-dimensional spaces [31], a more comprehensive model is expected for the discriminant analysis. We therefore present a general framework, which can unify many of existing algorithms and formulate new algorithms to improve the discriminating power. In addition, it embodies much more generalized novel properties and can cover the particular case in [25]. This framework is composed of two main levels in a top-down manner. In the first level, which is the overview of the basic learning scheme, it assumes that most existing subspace learning methods can be categorized into two classes, globality learning or locality learning [34]–[38]. In the second level, the first-level strategy is specified in the feature space, sample space, and learning space. We therefore define the following concepts to provide basic learning criteria for our framework.

- Feature-globality (FG) or feature-locality (FL): FG takes each training image as a single feature with each pixel being a dimension of the feature vector/matrix. FL selects local parts or local patches in the global feature space to build multiple models. The final decision is based on the voting of both local and global models via appropriate model fusion.
- Sample-globality (SG) or sample-locality (SL): SG, like conventional methods, applies all training data points to build the global model. SL partitions the sample space or searches local neighborhoods of a query to build linear models in local data sets.
- Learning-globality (LG) or learning-locality (LL): In a graph embedding view, LG constructs a globally connected graph to measure the data affinity for learning the low-dimensional representation. LL removes the unimportant edges to construct a partially connected graph embodying local connectivity.

It turns out that the above three local criteria can work individually or jointly for any new designs of subspace learning algorithms. Most existing subspace learning methods can be unified in such a common framework with either a global or local learning manner. To measure the performance of those criteria, we also introduce the model discriminating power coefficient (DPC), which provides a quantitative evaluation to compare global and local learning strategies. We will prove in both theory and experiment that local learning can improve the DPC of global learning in a valid range and thereby significantly enhance the classification accuracy. By applying the locality in subspace learning, the numerical difficulty problem for large database is naturally addressed, where many local variations cannot be adequately handled by a single global model. By localizing the modeling, the classification error rate estimation is also localized and thus it appears to be more robust and flexible for model selection among different model candidates.

Based on the framework, we developed a mixture-of-experts model of the query-driven locally adaptive (QDLA) method that can work with multiple appearance models for face recognition and head pose estimation. A single QDLA model is achieved as follows. A local neighborhood of the querying (unknown) face [25] is first identified from labeled faces, in a sample-locality manner. Depending on the number and quality of the labeled samples in the neighborhood, a linear model for classification is created, as well as the resulting classification error rate. The mixture-of-experts model is achieved by building multiple QDLA models according to feature-locality and voting the final result with corresponding QDLA models based on the error rates. Specifically, we build multiple appearance models of different parts and scales of faces to improve the face recognition performance. A piecewise linear subspace learning method in a learning-locality manner is applied in the QDLA model to map out the global nonlinear structure for head pose estimation. The local sensitive hash (LSH) [1], [13] based fast querying strategy is also introduced to deal with the high-dimensional nearest neighbors (NN) search problem. The proposed mixture-of-experts model is a general framework for classification and applicable to large size, multiclass, and multivariance face recognition data sets.

The paper is organized into the following sections. In Section II, we discuss the globality in linear models and introduce the concept of DPC. In Section III, we amply present the idea of the three types of locality in linear models and discuss how to build the joint local models. Extensive experiments and evaluations are presented in Section IV to support our theory. We draw the conclusion and point out future directions at the end of the paper.

II. Globality in Linear Models

Learning a model which can be applied globally in the sense of feature, learning algorithm and sample set is a common practice in the literature. Examples include Eigenface and Fisherface in face recognition, in which the data distribution is assumed to be Gaussian. A general idea to apply globality in subspace learning can be explained as follows. Take the whole face image as the original feature and train the entire given data to learn a

linear projection, which can map the original data to a low-dimensional subspace embodying the discriminative property. In this section, we will first highlight the objective of applications for our proposed ideas.

A. Objective for HCC Problems

Since human identity and activity are basic key issues in HCC, we mainly investigate the face recognition [14] and head pose estimation [6] problems in this paper. For those problems, typically a training data set of m subjects, with n samples in total, that characterized by identity and pose, $\{L_i = [a_i, b_i]\}_{i=1}^{n}$, is given as aligned and cropped $w \times h$ image luminance data, $\{X_i\}_{i=1}^{n}$ are the vectorized image data, where $X_i \in \mathbb{R}^{w \times h}$. A different set of test data with $\tilde{m}$ subjects and $\tilde{n}$ samples in total is also given, denoted as $\{\tilde{X}_i\}_{i=1}^{\tilde{n}}$, where $\tilde{X}_i \in \mathbb{R}^{w \times h}$. Both face recognition and pose estimation consist of a subspace learning (discriminative feature extraction) phase, where the objective is to find a d-dimensional subspace with bases $\mathbf{p}_i \in \mathbb{R}^D$, and therefore to obtain a projection $P \in \mathbb{R}^{d \times D}$, where $P = [\mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_d]^T$ and $D = w \times h$, such that classifiers like SVM or nearest neighbor [15] achieve the highest accuracy. Here, the projection matrix P is used to characterize this subspace where subject or pose variations are captured.

B. Model Solutions for Sample-Globality

Conventional model solutions for sample-globality are often based upon the assumption that the training data are drawn from the same underlying distribution as the test data. PCA and LDA can represent the two specific cases of unsupervised and supervised globality [15], [16]. The unsupervised globality takes the entire training data as a whole and maximizes the data variance in the projected linear space

$$\mathbf{p} = \arg\max_{\mathbf{p}, \|\mathbf{p}\|=1} \mathbf{p}^T S \mathbf{p}, \quad S = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \tag{1}$$

where $\bar{X}$ is the mean vector of $\{X_i\}_{i=1}^{n}$. The supervised globality takes the advantage of label L_i by projecting between-class appearances far-apart while keeping within-class appearances closer

$$\begin{aligned} \mathbf{p} &= \arg\max_{\mathbf{p}, \|\mathbf{p}\|=1} \frac{\mathbf{p}^T S_{\mathrm{b}} \mathbf{p}}{\mathbf{p}^T S_{\mathrm{w}} \mathbf{p}} \\ S_{\mathrm{b}} &= \sum_{c=1}^{m} n_c (\bar{X}_c - \bar{X})(\bar{X}_c - \bar{X})^T \\ S_{\mathrm{w}} &= \sum_{c=1}^{m} \sum_{\mathcal{F}_L(X_i)=c} (X_i - \bar{X}_c)(X_i - \bar{X}_c)^T \end{aligned} \tag{2}$$

where $\bar{X}_c$ is the mean vector of n_c samples in class c. Since S, S_{b}, and S_{w} are functions of all X_i, the global subspace is a function of the entire training data.

C. Discriminating Power Coefficient

Most existing linear subspace learning methods can find a good graph embedding interpretation [4], [24] in a common framework, where $\mathcal{G} = \{\{X_i\}_{i=1}^{n}, W\}$ denotes an undirected weighted graph with a vertex set $\{X_i\}_{i=1}^{n}$ and similarity matrix W. The general objective is to preserve data points adjacent relationship. In particular, the intrinsic graph of PCA connects all the data pairs with equal weights and is constrained by scale normalization on the projection vector, while the intrinsic graph of LDA connects all the data pairs with the same class labels and the weights are in inverse proportion to the sample size of the corresponding class. The intrinsic graph of PCA is also used as the penalty (between-class) graph of LDA.

In order to reflect the discriminating power of the model to characterize inter and intra-class point relationships, it is necessary to measure the tradeoffs between the complexity of the embedded graph, $\mathcal{G}(\{X_i\}_{i=1}^{n})$, and the expressive power of the model, P. A graph $\mathcal{G}(\{X_i\}_{i=1}^{n})$ is composed of its vertices $V(\{X_i\}_{i=1}^{n})$ and edge set $E(\{X_i\}_{i=1}^{n})$.

Definition: Discriminatating Power Coefficient (DPC)

Given the training data set $\{X_i\}_{i=1}^{n}$ and its graph embedding $\mathcal{G}(\{X_i\}_{i=1}^{n})$, the model DPC of a linear model $P \in \mathbb{R}^{d \times D}$ is defined as the ratio between the number of free variables in the model P and number of edges involved in $\mathcal{G}(\{X_i\}_{i=1}^{n})$

$$\mathrm{DPC}(X, P) = \frac{d \times D}{|E(X)|}. \tag{3}$$

A relatively large DPC value indicates a large discriminating power of the model. It is clear to see that DPC is directly proportional to both d and D and inversely proportional to $|E(X)|$. Given fixed $|E(X)|$, a larger d indicates more available discriminative subspace bases while a larger D provides more local details of feature patterns to enhance DPC. As $|E(X)|$ grows for a given model, the DPC decreases. For example, when $d \times D$ is fixed, we have $|E(X)|_{\mathrm{PCA}} > |E(X)|_{\mathrm{LDA}}$ and therefore $\mathrm{DPC}(X, P_{\mathrm{PCA}}) < \mathrm{DPC}(X, P_{\mathrm{LDA}})$. We derive that the $|E(X)|$ for unsupervised and supervised globality can be calculated as

$$|E(X)|_{\mathrm{Un-Globality}} = \binom{n}{2} \tag{4}$$

$$|E(X)|_{\mathrm{Su-Globality}} = \sum_{c=1}^{m} \binom{n_c}{2}. \tag{5}$$

Notice that a single model P contains $d \times D$ variables to characterize the subspace, where lies the manifold, spanned by n data points. For large size problems such as face recognition or head pose estimation, as n grows, the number of edges in an affinity graph of both unsupervised and supervised globality is in power growth. We expect that the DPC has a limited valid range since there are obvious inaccurate indications for very small $|E(X)|$ in the case of statistical insufficiency. In another word, the larger the DPC, the better the model. But, the DPC cannot be unlimitedly large due to trivial $|E(X)|$ values. We will discuss this issue in the experiment section.

III. Locality in Linear Models

Given a training data set $\{X_i\}_{i=1}^{n}$, the data size $D = w \times h$ is fixed. To boost the DPC of the resulting model, an intuitive way is to either enlarge d or reduce $|E(X)|$ in (3). The subspace size d can be enhanced by feature-locality, in which locality in the feature space can provide chances to gain more discriminative

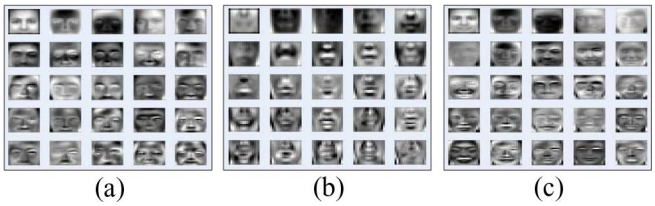


Fig. 1. Feature-locality versus globality. Part-based face modeling for feature-locality using FERET database. (a) Upper part face model (18 × 16). (b) Lower part face model (14 × 18). (c) Full face model (21 × 28).

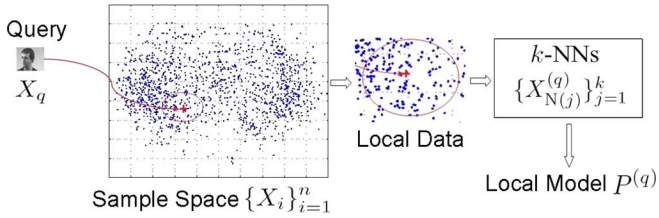


Fig. 2. Query-driven sample-locality model flowchart.

bases for subspace learning. The number of embedded graph edges $|E(X)|$ can be reduced by sample-locality, in which the size-reduced subgraph is achieved by reducing the number of vertices in the original graph. Moreover, note that d and $|E(X)|$ are often correlated. Learning locality can achieve higher DPC by changing the two parameters at the same time using a particular graph construction strategy.

A. Feature-Locality

Feature locality is the key in part-based and patch-based models [11], [12], [32]. The basic idea is to consider not only the original whole image features, but also parts or local patches extracted from the images simultaneously [7], [33]. This approach offers the advantage to enrich the feature space because the parts can be modeled more or less independently. Thus, the subspace size d can be enhanced by combining both global and local features to gain more discriminating power reflected by the DPC. For example, Fig. 1 shows the part-based face models on the FERET database. Since the upper and lower parts of human face appearance have less dependency, we can model the two separately and combine them with the full model. It turns out that we have three subspaces in total for an integrated model which provides larger d than any single subspace.

B. Sample-Locality

Instead of improving the DPC by directly reducing the edges of a global graph $\mathcal{G}$, sample-locality achieves the higher DPC by reducing the number of vertices in $\mathcal{G}$. A feasible way to realize this as suggested by [10] is to partition the training data into a hierarchical structure via kd-tree. For each data subset corresponding to a sub tree, we can compute its model via unsupervised or supervised globality. As a result, there is a set of linear models with a hierarchical structure that we need to deal with. However, this strategy brings a difficulty to select the right model or hierarchical levels that offer the best discriminating power, especially when a query point lies on the boundaries of the kd-tree partitions.

1) Model Solutions for Sample-Locality: To solve this problem, instead of building a model P for each data partition node in the kd-tree, a query point driven local neighborhood based model is computed. The first step for the modeling is to search for the nearest neighbors of the query datum in the feature space. By considering only the local distribution of the high-dimensional data, such a local model is more discriminative than the global one. For a given query $X_q \in \mathbb{R}^{w\times h}$, we find its k-NNs $\{X_{\mathrm{N}(j)}^{(q)}\}_{j=1}^k$ in the original sample space, where $k \ll n$. The unsupervised sample-locality takes $\{X_{\mathrm{N}(j)}^{(q)}\}_{j=1}^k$ as a whole and maximizes the data variance in the projected and localized sample space

$$\begin{aligned}\mathbf{p}^{(q)} &= \arg\max_{\mathbf{p},\|\mathbf{p}\|=1} \mathbf{p}^T S^{(q)} \mathbf{p} \\ S^{(q)} &= \sum_{i=1}^{k} \left(X_i^{(q)} - \bar{X}^{(q)}\right)\left(X_i^{(q)} - \bar{X}^{(q)}\right)^T \end{aligned} \tag{6}$$

where $X_i^{(q)} \in \{X_{\mathrm{N}(j)}^{(q)}\}_{j=1}^k$ for $i = 1,\ldots,k$, and $\bar{X}^{(q)}$ is the mean of $\{X_{\mathrm{N}(j)}^{(q)}\}_{j=1}^k$. The supervised model of this query point for sample-locality is formulated as

$$\begin{aligned}\mathbf{p}^{(q)} &= \arg\max_{\mathbf{p},\|\mathbf{p}\|=1} \frac{\mathbf{p}^T S_{\mathrm{b}}^{(q)} \mathbf{p}}{\mathbf{p}^T S_{\mathrm{w}}^{(q)} \mathbf{p}} \\ S_{\mathrm{b}}^{(q)} &= \sum_{c=1,n_c'\geqslant n_0}^{m'} n_c' \left(\bar{X}_c^{(q)} - \bar{X}^{(q)}\right)\left(\bar{X}_c^{(q)} - \bar{X}^{(q)}\right)^T \\ S_{\mathrm{w}}^{(q)} &= \sum_{c=1,n_c'\geqslant n_0}^{m'} \sum_{\mathcal{F}_L\left(X_i^{(q)}\right)=c} \left(X_i^{(q)} - \bar{X}_c^{(q)}\right) \\ &\quad \times \left(X_i^{(q)} - \bar{X}_c^{(q)}\right)^T \end{aligned} \tag{7}$$

where m' $(m' < m)$ is the number of classes in $\{X_{\mathrm{N}(j)}^{(q)}\}_{j=1}^k$. n_c' is the number of samples in class c. n_0 is the minimum number of samples required in each class. This is used to remove trivial points with limited impact of graph structure. Since $S^{(q)}$, $S_{\mathrm{b}}^{(q)}$, and $S_{\mathrm{w}}^{(q)}$ are functions of X_q and $X_i^{(q)}$, the learned subspaces are functions of both query and training data in the localized sample space. Fig. 2 shows the flowchart for the query-driven sample-locality model.

2) DPC for Sample-Locality: The DPC of the sample-locality model is calculated as

$$\mathrm{DPC}\left(X^{(q)}, P^{(q)}\right) = \frac{d \times D}{\left|E\left(X^{(q)}\right)\right|} \tag{8}$$

where the number of local graph edges, $|E(X^{(q)})|$, for unsupervised and supervised sample-locality are given by

$$\left|E\left(X^{(q)}\right)\right|_{\mathrm{Un-S-Locality}} = \binom{k}{2} \tag{9}$$

$$\left|E\left(X^{(q)}\right)\right|_{\mathrm{Su-S-Locality}} = \sum_{c=1}^{m'} \binom{n'_c}{2}. \tag{10}$$

We have $|E(X^{(q)})|_{\mathrm{Un-S-Locality}} < |E(X)|_{\mathrm{Un-Globality}}$ and $|E(X^{(q)})|_{\mathrm{Su-S-Locality}} < |E(X)|_{\mathrm{Su-Globality}}$, since $k \ll n$ and $m' < m$. Hence, the sample-locality offers better discriminating power than globality in the sense that the model is well adapted to the local data and the DPC can be tuned to achieve better performance.

3) Fast Nearest Neighbors Querying: The exhaustive linear scan method to search for the NNs is computationally expensive for large databases $(O(n))$. Note for every query datum, we need to perform such an NN query to build the local model. One speed-up solution is to take advantage of the data set spatial distribution in $\mathbb{R}^{w\times h}$. From the database point of view, NN search can be treated as a querying problem. In order to speed up the NN search, many index structures have been well studied in the database community, such as kd-tree and R-tree. Nevertheless, almost all of such index-based search methods suffer from the problem of curse-of-dimensionality. When compared with exhaustive linear scan search, such index-based search methods normally cannot perform better in the high-dimensional space. We instead use LSH [1], [13] to speed up the NN query. The local model can then be built based on the NNs of the query. To overcome the curse-of-dimensionality, LSH provides a randomized solution for the high-dimensional NN search problem. Instead of searching for the exact k-NN, LSH tends to search for the approximate NN which is defined as ϵ-nearest neighbor search (ϵ-NNS)—For the data set $\{X_i\}_{i=1}^n$, we preprocess it to efficiently search for the approximate NNs of any given query $X_q \in \mathbb{R}^{w\times h}$, that is, to find $X_p \in \{X_i\}_{i=1}^n$, such that $\mathrm{dist}(X_q, X_p) \leq (1+\epsilon)\mathrm{dist}(X_q, \{X_i\}_{i=1}^n)$. Here, $\mathrm{dist}(X_q, \{X_i\}_{i=1}^n)$ denotes the distance of X_q to its closest neighbor in $\{X_i\}_{i=1}^n$ and ϵ is a predefined parameter.

C. Learning-Locality

Learning-locality changes both d and $|E(X)|$ to enhance the DPC. This task can be achieved by constructing a particular graph embedded model $\mathcal{G} = \{\{X_i\}_{i=1}^n, W\}$ with effective weights designed to fit the learning-locality criterion. One criterion is the local connectivity, which is to remove graph edges from the global graph via k-NN search, ϵ-thresholding, or from the ground truth.

1) Model Solutions for Learning-Locality: Given training data $\{X_i\}_{i=1}^n$, the weights W can be either calculated in closed-form or set with empirical values after constructing the graph. Locally embedded analysis (LEA) [5] and locality preserving projections (LPP) [3] can represent the two specific cases.

LEA constructs the graph and learns the weights in a neighborhood-preserving manner by exploiting the local symmetries of linear reconstructions [2]. The linear subspace learning on the entire training data is formulated as

$$w_j^{(i)} = \underset{\substack{w_j^{(i)} \\ \sum_j w_j^{(i)}=1}}{\arg\min} \sum_{i=1}^{n} \left\| X_i - \sum_{j=1}^{k} w_j^{(i)} X_{\mathrm{N}(j)}^{(i)} \right\|^2 \tag{11}$$

$$\mathbf{p} = \arg\min_{\mathbf{p}} \sum_{i=1}^{n} \left\| \mathbf{p}^T X_i - \sum_{j=1}^{k} w_j^{(i)} \mathbf{p}^T X_{\mathrm{N}(j)}^{(i)} \right\|^2,$$
$$\text{s.t. } \mathbf{p}^T X_i X_i^T \mathbf{p} + \left(\sum_{j=1}^{k} w_j^{(i)} \mathbf{p}^T X_{\mathrm{N}(j)}^{(i)}\right) \times \left(\sum_{j=1}^{k} w_j^{(i)} \mathbf{p}^T X_{\mathrm{N}(j)}^{(i)}\right)^T = 1 \tag{12}$$

where $\{X_{\mathrm{N}(j)}^{(i)}\}_{j=1}^k$ are the k-NNs of X_i, and $\{w_j^{(i)}\}_{j=1}^k$ are the graph weights.

LPP learns the subspace that preserves the essential manifold structure by measuring the local neighbor distance information. The graph weights are set empirically, $w_{ij} = \exp(-\|X_i - X_j\|^2/t)$ when X_i and X_j are the k-NNs of each other, otherwise $w_{ij} = 0$. Here, t is a tunable parameter. The linear subspace learning on the entire training data is formulated as

$$\mathbf{p} = \underset{\substack{\mathbf{p} \\ \mathbf{p}^T X_i X_i^T \mathbf{p}=1}}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{n} (\mathbf{p}^T X_i - \mathbf{p}^T X_j)^2 w_{ij}. \tag{13}$$

2) DPC for Learning-Locality: The above two cases indeed construct the same graph with different weights if we fix the parameter k. Then we have the same number of graph edges, $|E(X,k)|_{\mathrm{L-Locality}} = n \times k/2$. So the DPC of the model solution for learning-locality is given by

$$\mathrm{DPC}(X,k) = \frac{2 \times d \times D}{n \times k}. \tag{14}$$

Note that the number of graph edges for learning-locality only grows linearly, which can achieve higher DPC than globality. From some simple derivations, we have $|E(X,k)|_{\mathrm{L-Locality}} < |E(X)|_{\mathrm{Su-Globality}}$ when $k < n-1$. By introducing the parameter k and enhancing the value of d, learning-locality expands the learned subspace with more discriminative bases, which leads to the enhanced DPC. For example, supervised globality like LDA in an m-class problem can only learn a subspace with up to $m-1$ dimensions. But the learning-locality like LEA can learn a subspace with the dimension up to n, which is obviously larger than $m-1$.

D. Joint Local Models

The more effective way for designing discriminative learning algorithms is to integrate different local concepts. Since the feature, sample, and learning localities are from less dependent aspects of local definition, it is straightforward to combine some or all of them to construct new models. For example, we can first localize the k-NNs of a query datum with sample-locality; then apply learning-locality in the local data set to learn a piecewise linear subspace; finally label the test datum with a neighborhood-based classifier. More specifically, we develop the framework of combining feature-locality and sample-locality together, and design a mixture-of-experts model of QDLA Fisher faces [27] for face recognition applications.

1) Feature-Locality + Sample-Locality: For the general image classification problem, we can design the joint model

of feature-locality and sample-locality. Assume we have a known set of object images $\mathcal{I} = \{I_i\}_{i=1}^{n}$ with class label set $\mathcal{L} = \{L_j\}_{j=1}^{m}$. Now for an unknown test image sample $\tilde{I}$, we need to select a label $\tilde{L}$ from $\mathcal{L}$ to identify $\tilde{I}$. Our joint locality algorithm for a single appearance model is as follows.

- Identify a local neighbor $\tilde{I}_q$ of the query (unknown) image $\tilde{I}$ from known labeled images.
- Pick k-NNs in the training data $\mathcal{I}$ for $\tilde{I}_q$.
- In case the k-NNs have the same label L_k, then label $\tilde{I}$ with L_k and exit; otherwise, depending on the number and quality of the labeled samples in the neighborhood, create a local linear model for $\tilde{I}_q$, along with the locality-model classification error rate.
- Classify $\tilde{I}$ with the local model that has the lowest training error.

The joint model of feature-locality and sample-locality is summarized as follows.

- Build multiple appearance models with variation in parts, patches and resolution of object images.
- Run the above joint locality algorithm on each appearance model respectively.
- Vote the final classification decision with individual local models and the global model based on error rates.

As an instantiation of this model design, we introduce the following mixture-of-experts model of query-driven locally adaptive fisher faces.

2) Mixture-of-Experts Model of Query-Driven Locally Adaptive Fisher Faces: As an example in the face recognition scenario, the mixture-of-experts model of QDLA fisher faces is a joint model of feature-locality and sample-locality. The framework of mixture-of-experts model of QDLA-Fisher faces for face recognition is shown in Fig. 3. For the global appearance based face model, we take the cropped whole face image as the learning feature. We apply three different appearance models with labeled face data, that is, to obtain $\{I_i^{(j)} | i = 1, 2, \ldots, n; j = 1, 2, 3.\}$ from original face image data $\mathcal{I}$, where j indicates different appearance models. An example for the basis functions of upper, lower, and full FERET face models are visualized in Fig. 1. The parts can be overlapped to capture the appearance in different areas and resolutions. In the recognition phase, for an unknown face $\tilde{I}$, and its projection $I_P^{(j)}$ in the jth model space, we apply LSH to identify its local neighborhood with $\mathrm{N}(I_P^{(j)}) = \{I_i^{(j)} | \mathrm{s.t.} \| I_i^{(j)} - I_P^{(j)} \| < r^{(j)}\}$ and build the local fisher model for each appearance feature set, i.e., the distribution of each class of faces contained in $\mathrm{N}(I_P^{(j)})$. Here $r^{(j)}$ denotes the local neighborhood radius. For the jth Fisher model, its training error rate e_j is also computed and recorded, depending on $\mathrm{N}(I_P^{(j)})$. The query face is then classified by combining these local models.

After we do face recognition with each local model, the final recognition result is based on the error rate e_j of each local model. In the preferred embodiment, we select the local model with the minimum training error rate and treat its recognition result as our final result. If there is a tie in error rates, the preference is given in the order of full face model, upper face model and lower face model. We also define two parameters to tune the local model. One is the default local Fisher model classification error e_0 and the other is the minimum local sample ratio $s_{\min}$. During the query stage, if the ratio between the number of local samples within the range of the local neighborhood radius and the number of total training samples is lower than $s_{\min}$, we consider the local Fisher model is insufficient for training. In this case, we substitute the feature-locality model with the feature-globality model and adopt the error rate e_j to be the default global Fisher model classification error rate.

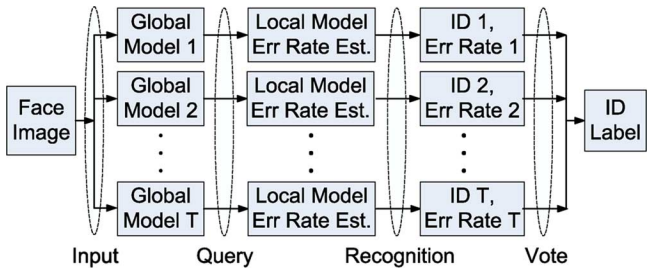


Fig. 3. Joint local model. Framework of mixture-of-experts model of QDLA-Fisher faces for face recognition.

IV. Experiments and Evaluations

We evaluate the proposed methods with extensive experiments on several most popular data sets for the HCC applications.

A. Data Sets and Methods

We use following standard face data sets in the experiments. Sample images of those face data sets are shown in Fig. 4.

- *FERET Database* [18]. The original database, containing 1209 subjects, was released in 2001 and consists of 14051 greyscale images. The images, in a resolution of 256 × 384, were taken with head views ranging from frontal to left and right profiles. We select 2550 near frontal face images, crop and resize the images to the size of 21 × 28 according to the eye locations and pupil distance.
- *ORL Database* [19]. This database contains 40 subjects with ten grayscale face images for each. The 400 images, in a resolution of 92 × 112, were taken at different times, varying lighting, facial expressions (open/closed eyes, smiling/not smiling) and accessories (glasses/no glasses), showing whole frontal and slight tilt of the head. We crop and resize the images to the size of 21 × 28 according to the eye locations and pupil distance.
- *YALE Face Database* [27]. This database contains 165 grayscale images of 15 individuals. The 11 images for each subject, in a resolution of 320 × 243, were taken in the conditions of center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. The images are cropped and resized to 32 × 32.
- *UMIST Face Database* [30]. This database contains 564 grayscale face images for 20 subjects. Each subject has 19 to 36 images, in a resolution of 220 × 220, in various angles from left profile to frontal view. The images are cropped and resized to 28 × 34.

Fig. 4. Data sets. (a) Pointing'04 head-pose image database. The left part shows all the subjects in the database. The right part shows all the head pose orientations of one subject. (b) Face image samples in the ORL, YALE, UMIST, and UIUC-IFP-Y internal databases.

TABLE I
LIST OF DIFFERENT METHODS

Abbreviation	Method
Un-Globality	Unsupervised Globality
Su-Globality	Supervised Globality
Un-L-Locality	Unsupervised Learning-Locality
Su-L-Locality	Supervised Learning-Locality
Un-S-Locality	Unsupervised Sample-Locality
Su-S-Locality	Supervised Sample-Locality
Su-S+L-Locality	Supervised Sample+Learning-Locality
Su-F+S-Locality	Supervised Feature+Sample-Locality

- *UIUC-IFP-Y Internal Face Database* [5]. This is an internal face database which contains 3520 grayscale images taken from video sequences for 22 subjects, with 160 images for each. Each cropped image has a resolution of 40×40, with large variations in facial expression, illumination, pose and occlusion.
- *Pointing'04 Head-Pose Image Database* [28], [29]. This database consists of 15 sets of images for 15 subjects, wearing glasses or not and having various skin colors. Each set contains two series of 93 images of the same person at different poses. The first series is used for training, and the second is for test. The pose or head orientation is determined by the pan and tilt angles, which vary from $-90°$ to $+90°$. Fig. 4(a) shows the Pointing'04 head-pose images.

We evaluate our proposed framework by comparing the different methods listed in Table I.

B. Learning-Locality versus Globality

Extensive face recognition experiments are performed to compare the discriminating power between learning-locality and globality. Both unsupervised and supervised algorithms that cover most cases of globality are evaluated. The algorithm examples corresponding to unsupervised globality, supervised globality, and learning-locality are PCA, LDA, and LEA, respectively. The four data sets ORL, YALE, UMIST, and UIUC-IFP-Y are adopted for the experiments. We choose the Euclidean distance and the nearest neighbor classifier in all the recognition experiments and assume that the gallery set [17] of each experiment is the same as the training set. We randomly select 3, 6, and 25% images of each subject for training on ORL, YALE, and UMIST, respectively and the rest of 7, 5, and 75% for test. In total, the training data size of the three databases are 120, 90, and 145, respectively, while the test data size are 280, 75, and 419. In a brute-force manner, the recognition rates of PCA, LDA, and LEA on each case of dimensionality reduction are all calculated. To generalize the performance, we repeat the data set partition 100 times on each data set. Fig. 5(a)–(c) show the average recognition rates of the 100-times run for each method against the dimension of the subspace. Table II summarizes the lowest average error rates of unsupervised globality, supervised globality, and learning-locality on the ORL, YALE, and UMIST respectively. It turns out that the learning-locality consistently outperforms the globality with lowest error rates of 9.7%, 7.7%, and 4.3% on the three data sets respectively. The significant improvement of learning-locality has no extra sacrifice on the dimensionality since its reduced dimensions for the optimal cases are comparable to or even lower than those of globality.

We choose the UIUC-IFP-Y data set to evaluate the generalization ability of learning-locality. In this experiment, the gallery set, training set and test set are all different. The 20, 10, and 130 images of each subject are randomly selected for training, gallery and test, respectively. For all 20 subjects, we have 440, 220, and 2860 images in total for each data sets. Fig. 5(d) shows the average recognition rates of the 100-times run for each method against the dimension of the subspace. It turns out from Table II that the lowest error rates and the corresponding reduced dimensions are (32.7%, 31.2%, 10.7%) and (145, 21, 32) respectively for the three methods. These results still indicate that the learning-locality consistently and significantly outperforms the globality.

C. Sample-Locality and Learning-Locality versus Globality

We compare the sample-locality and learning-locality with globality in dealing with the head pose estimation problem on the Pointing'04 head-pose database. In this scenario, we take each head pose orientation as one class, so we have $m = 93$ and $n_c = 15$. The first series of each set is used for training, and the second for test. For each method in the comparison, we learn the linear subspace via globality, locality, or joint metrics, and estimate test head poses via the NN classifier.

The error rates for pan and tilt angle estimations for $D = 16 \times 16$ and $D = 32 \times 32$ are shown in Table III. We use

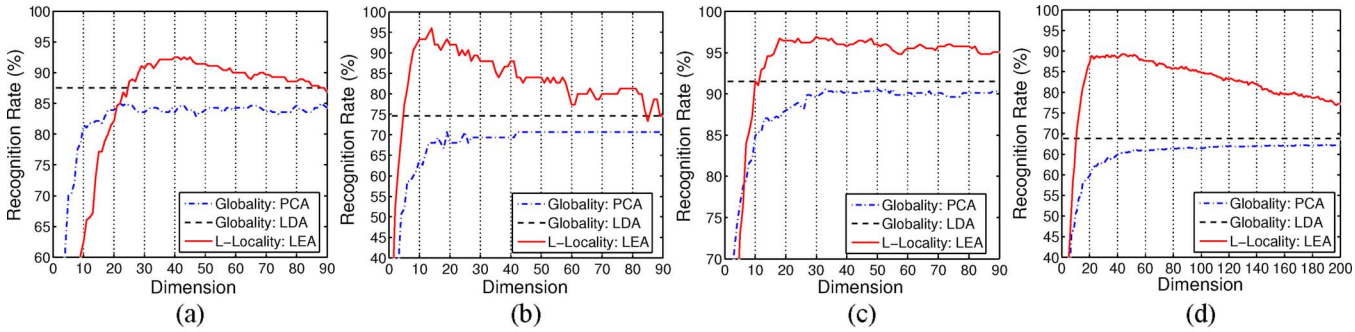


Fig. 5. Learning-locality versus unsupervised/supervised globality with face recognition evaluations on the ORL, YALE, UMIST and UIUC-IFP-Y data sets. The algorithm examples representing unsupervised globality, supervised globality, and learning-locality are PCA, LDA, and LEA, respectively. (a) ORL. (b) YALE. (c) UMIST. (d) UIUC-IFP-Y.

TABLE II
FACE RECOGNITION PERFORMANCE FOR UNSUPERVISED GLOBALITY, SUPERVISED GLOBALITY, AND LEARNING-LOCALITY ON THE ORL, YALE, UMIST, AND UIUC-IFP-Y DATA SETS

Method	ORL		YALE		UMIST		UIUC-IFP-Y	
	Error	Dim.	Error	Dim.	Error	Dim.	Error	Dim.
Unsupervised Globality	14.4%	88	25.8%	42	10.7%	33	32.7%	145
Supervised Globality	12.8%	39	23.6%	14	9.1%	19	31.2%	21
Learning-Locality	**9.7%**	40	**7.7%**	14	**4.3%**	18	**10.7%**	32

TABLE III
HEAD POSE ESTIMATION ERROR RATES ON THE POINTING'04 HEAD-POSE DATABASE

Method	Pan		Tilt	
	$(D = 16^2)$	$(D = 32^2)$	$(D = 16^2)$	$(D = 32^2)$
Un-Globality	33.5%	26.9%	44.3%	35.1%
Su-Globality	30.1%	25.8%	33.3%	26.9%
Un-L-Locality	67.7%	63.4%	76.3%	61.3%
Su-L-Locality	30.1%	24.7%	31.2%	**22.6%**
Un-S-Locality	25.2%	24.5%	37.8%	37.6%
Su-S-Locality	**20.4%**	**19.1%**	**30.7%**	30.7%

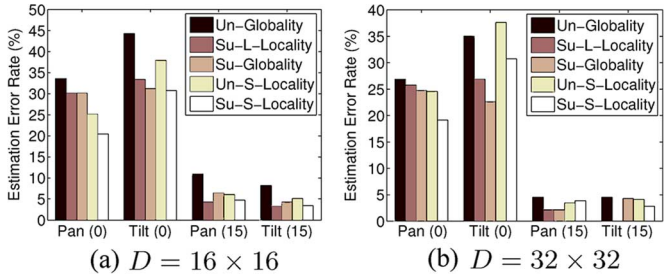


Fig. 6. Sample-locality versus unsupervised/supervised globality with head pose estimation on the Pointing'04 head-pose database. The algorithm examples representing unsupervised globality, supervised globality, and learning-locality are PCA, LDA, and LPP, respectively. For sample-locality, we have $k = 32$ for the k-NNs local search. $(0) \longleftrightarrow 0°$ estimation tolerance; $(15) \longleftrightarrow 15°$ estimation tolerance. (a) $d_{\mathrm{Un-Globality}} = 16$, $d_{\mathrm{Su-Globality}} = 6$, (b) $d_{\mathrm{Un-Globality}} = 32$, $d_{\mathrm{Su-Globality}} = 16$.

the popular algorithms PCA, LDA and LPP to represent unsupervised globality, supervised globality, and unsupervised/supervised locality respectively. The evaluation has 0° estimation error tolerance. For sample-locality, we have $k = 32$ for the k-NNs local search. We can see that all kinds of local metrics outperform the globality metric. Supervised sample-locality performs the best overall, and achieves the best results in three out of four cases, followed by another supervised learning-locality. Supervision is important since supervised methods consistently perform better than unsupervised methods. Unsupervised learning-locality does not perform well in this scenario. However, the unsupervised sample-locality mitigates the lack of labeling information by localization, and rather surprisingly, performs well and is comparable to supervised globality and supervised learning-locality. This observation indicates the benefit brought by sample-locality.

A more extensive comparison on this evaluation is summarized in Fig. 6, which also shows the case of 15° estimation tolerance. For Fig. 6(a), we set $D = 16 \times 16$, $d_{\mathrm{Un-Globality}} = 16$, and $d_{\mathrm{Su-Globality}} = 6$, while for Fig. 6(b), $D = 32 \times 32$, $d_{\mathrm{Un-Globality}} = 32$, and $d_{\mathrm{Su-Globality}} = 16$. We observe that supervised sample-locality and supervised learning-locality both outperform globality in most cases.

D. Feature-Locality + Sample-Locality versus Globality

We train each single upper, lower, and full model of QDLA-Fisher faces with the FERET images. The model sizes are defined by 18×16, 14×18, and 21×28, respectively. Fig. 1 shows the basis functions of the QDLA-Fisher model created by the FERET data. Since the ORL face data set has specific labels for each subject with the 10 poses, we partition the data set into 6 subsets with different training and test samples (some difficult recognition cases). Table V shows the description of the data set partitions for the 6 recognition experiments. The model basis functions and model parameter $r^{(j)}$ are fixed for each single QDLA-Fisher model. The average face recognition accuracy for the six different data set partitions under the FERET basis is plotted in Fig. 7. We can see that the mixture-of-experts model of QDLA-Fisher outperforms all the single QDLA-Fisher model under the 6 data set partitions. The full QDLA-Fisher model performs better than the upper and lower models, but worse than the mixture-of-experts model of QDLA-Fisher

TABLE IV
RECOGNITION ACCURACY (%) FOR THE ORL DATABASE WITH QDLA-FISHER FACES OF MIXTURE-OF-EXPERTS MODEL BASED ON FERET BASIS FUNCTIONS. (UPPER, LOWER, AND FULL ⟷ SINGLE MODELS OF QDLA-FISHER FACES; PCA ⟷ EIGENFACE RECOGNITION; QDLA ⟷ MIXTURE-OF-EXPERTS MODEL OF QDLA-FISHER FACES; TIME ⟷ COMPUTATION TIME NEEDED FOR EACH RECOGNITION ATTEMPT; m ⟷ # OF SUBJECTS; n_c ⟷ # OF TRAINING IMAGES PER SUBJECT; $\tilde{n}_c$ ⟷ # OF TEST IMAGES PER SUBJECT; $s_{\min}$ ⟷ MINIMUM LOCAL SAMPLE RATIO.)

Cases	$m=30$	$n_c=8$	$\tilde{n}_c=2$	$s_{\min}=0.3$			$m=30$	$n_c=8$	$\tilde{n}_c=2$	$s_{\min}=0.2$		
Test#.	Upper	Lower	Full	PCA	QDLA	Time(s)	Upper	Lower	Full	PCA	QDLA	Time(s)
1	0.8167	0.6667	0.9333	0.8333	0.9310	0.1325	0.8333	0.6833	0.9333	0.8333	0.9310	0.1187
2	0.8500	0.7000	0.9000	0.8500	0.9286	0.1315	0.8833	0.6833	0.9000	0.8500	0.9643	0.1096
3	0.9000	0.6667	0.9500	0.9333	0.9492	0.1169	0.8833	0.6667	0.9667	0.9333	0.9661	0.1198
4	0.8833	0.5833	0.9500	0.8500	0.9828	0.1253	0.8500	0.5833	0.9500	0.8500	0.9828	0.1125
5	0.8500	0.5333	0.9000	0.8167	0.9298	0.1268	0.8333	0.5167	0.9000	0.8167	0.9298	0.1013
6	0.7833	0.5833	0.9000	0.7667	0.8947	0.1190	0.7833	0.6000	0.9000	0.7667	0.8772	0.1086
7	0.8833	0.5833	0.9500	0.8500	0.9828	0.1219	0.8500	0.5833	0.9500	0.8500	0.9828	0.1013
8	0.8167	0.5500	0.9000	0.8000	0.8966	0.1224	0.8167	0.5500	0.9000	0.8000	0.8966	0.1172
9	0.8500	0.7000	0.9000	0.8500	0.9286	0.1219	0.8833	0.6833	0.9000	0.8500	0.9643	0.0979
10	0.7833	0.6167	0.9333	0.7833	0.9310	0.1237	0.8167	0.5833	0.9167	0.7833	0.9310	0.1096
Mean	0.8417	0.6183	0.9217	0.8333	**0.9355**	0.1242	0.8433	0.6133	0.9217	0.8333	**0.9426**	0.1097
Var.	0.0017	0.0037	0.0006	0.0022	0.0009	0.0000	0.0011	0.0038	0.0007	0.0022	0.0013	0.0001
Cases	$m=40$	$n_c=6$	$\tilde{n}_c=4$	$s_{\min}=0.2$			$m=40$	$n_c=8$	$\tilde{n}_c=2$	$s_{\min}=0.2$		
Test#.	Upper	Lower	Full	PCA	QDLA	Time(s)	Upper	Lower	Full	PCA	QDLA	Time(s)
1	0.8438	0.5250	0.8750	0.8375	0.9456	0.2338	0.8375	0.4875	0.8750	0.8000	0.9333	0.2544
2	0.8000	0.5250	0.8688	0.8187	0.9392	0.2363	0.8875	0.6250	0.9125	0.8500	0.9605	0.2597
3	0.8688	0.4750	0.8500	0.7875	0.9007	0.2342	0.8625	0.4875	0.8875	0.7750	0.9221	0.2628
4	0.8562	0.5313	0.8750	0.8000	0.9145	0.2357	0.9125	0.5625	0.9500	0.8625	0.9744	0.2310
5	0.8125	0.5188	0.8250	0.7500	0.8919	0.2300	0.8000	0.5125	0.8250	0.8000	0.8919	0.2327
6	0.8187	0.5313	0.8562	0.8125	0.9067	0.2319	0.9375	0.5500	0.9125	0.8375	0.9733	0.2195
7	0.7937	0.5000	0.8625	0.7500	0.9139	0.2309	0.8625	0.5125	0.9375	0.8250	0.9615	0.2351
8	0.8125	0.5188	0.8250	0.7500	0.8919	0.2290	0.8750	0.5750	0.9500	0.8250	0.9620	0.2293
9	0.9063	0.5500	0.9125	0.8625	0.9608	0.2376	0.8875	0.5500	0.8875	0.8875	0.9342	0.2697
10	0.9063	0.5500	0.8938	0.8063	0.9346	0.2362	0.8125	0.5125	0.8875	0.8000	0.9221	0.2484
Mean	0.8419	0.5225	0.8644	0.7975	**0.9200**	0.2336	0.8675	0.5375	0.9025	0.8263	**0.9435**	0.2443
Var.	0.0017	0.0005	0.0008	0.0015	0.0006	0.0000	0.0018	0.0019	0.0015	0.0012	0.0007	0.0003

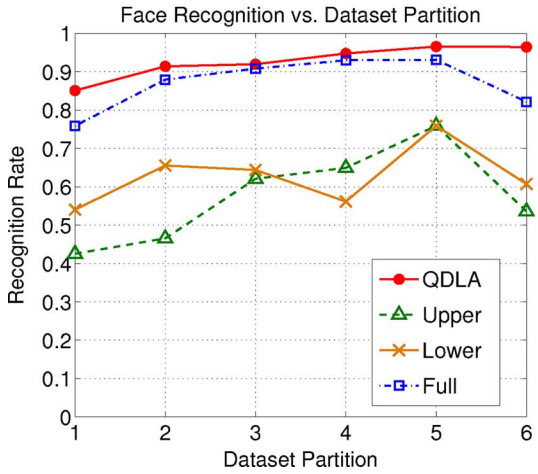


Fig. 7. Average ORL face recognition accuracy for the 6 different data set partitions under the same model parameter settings.

TABLE V
DESCRIPTION OF THE DATA SET PARTITIONS FOR FIG. 7

Cases	Sub. ID	Test ID	r^{Upper}	r^{Lower}	r^{Full}
1	6:10, 16:40	3, 7, 8	400	600	800
2	6:10, 16:40	3, 8	400	600	800
3	11:40	5, 8, 10	400	600	800
4	11:40	5, 10	400	600	800
5	11:40	10	400	600	800
6	1:20, 31:40	5	400	600	800

faces. To further demonstrate the generalization and robustness properties of our method, we design the following more specific recognition experiments.

The recognition performances for the ORL database with the mixture-of-experts model of QDLA-Fisher faces based on the FERET basis functions are shown in Table IV. We randomly partition the 10 images of each subject by either 6 labeled and 4 un-labeled, or 8 labeled and 2 un-labeled faces. We set up 10 sets of 160 and 80 recognition attempts in total for the four cases of test. We set the local neighborhood radius $r^{(j)}$ to 420, 400, and 860 for the single upper, lower, and full models of QDLA-Fisher faces respectively in the upper two cases of Table IV, and set $r^{(j)}$ to 640, 600, and 1000 in the lower two cases. The default NN classification error rate is set to 0.005 and the minimum local sample ratio is set to 0.2 or 0.3. The experiments show positive and encouraging results since the recognition and query modules are accurate, fast, and robust. We can observe that the recognition performance of the QDLA-Fisher mixture-of-experts model is better than any single QDLA-Fisher model and the global PCA Eigenface [20] recognition. It also has very low recognition rate variance of the random tests since the localization of models and adaptive multiple modes of classification make the error estimation robust. The computation time needed for each recognition is only $0.23 \sim 0.24s$, shown in Table IV, on a 2.0-GHz Pentium CPU and 512 MB RAM PC with an un-optimized Matlab 6.0 implementation. The NN query is also computationally efficient, costing $0.005 \sim 0.02s$ for each query (depending on the value of radius $r^{(j)}$).

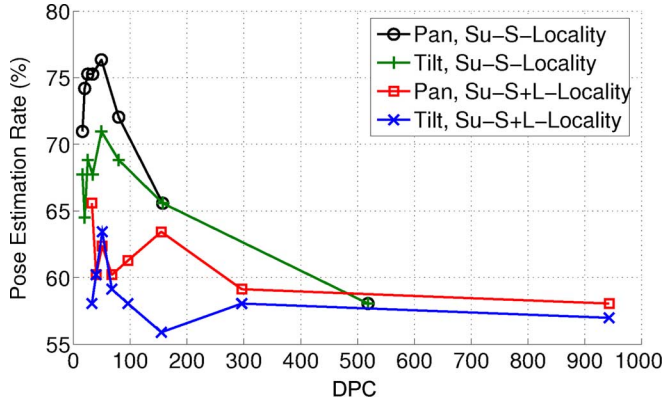


Fig. 8. Accuracy rate versus DPC for head pose estimation.

E. Discussions

To investigate the discriminative metric performance changing with the DPC measurement, we set up experiments for supervised sample-locality (localized LDA) and supervised sample-locality+learning-locality (localized LPP). For each query point, the local neighborhood size k and subspace dimension d are selected to compute the local metrics. The supervised sample-locality metric based head pose estimation rate and its DPC are plotted in Fig. 8 for $d = 32$. In Fig. 8, the estimation performance falls off as the DPC decreases beyond certain points (around 50 or 150, etc.). We can see that the DPC is only valid during the limited value range [0, 200]. It is not guaranteed to get effective discriminating classifications when the DPC is large.

The DPC captures the empirical tradeoffs between the data complexity, as indicated by the number of edges among data points, and the model complexity, as indicated by the number of free variables in the linear models. In Fig. 8, the local neighborhood size is changed from 15, 30, 45, ... to 120. When the local neighborhood size is small, the model does not capture the data variation well, the metric learnt does not generalize well, and the recognition performance is not good. When the local neighborhood size is very large, the data variation is well captured, but the model power is also overwhelmed, where the overall trend of estimation performance decreases with the DPC increasing. A good tradeoff seems to be at a DPC value between 50 and 80, where the data variations are well captured by the models and the models achieve good recognition performance. This phenomenon conforms to our theoretical discussions in Section II. Such statement is well demonstrated in Fig. 8 for both Su-S-Locality and Su-S+L-Locality cases. Note that here the DPC value may not be generalized. This particular example is used to demonstrate the tradeoff towards any problem and model on the DPC concept.

The computational complexity of the local metric for the head pose estimation experiment is summarized in Table VI, with various subspace dimension d and size of neighborhood k. The computational cost is low in the locality framework. The average speed of pan/tilt angle estimation with a supervised sample-locality metric is about 7 to 10 estimations per second, with un-optimized Matlab 6.0 code running on a 2.0 GHz Pentium CPU and 512 MB RAM PC. The joint model of sample-locality and learning-locality does not introduce more computations and even saves the computational cost in most cases.

TABLE VI
COMPARISON ON COMPUTATIONAL COMPLEXITY (SECOND PER ESTIMATION) FOR HEAD POSE ESTIMATION

Method	$k = 30$	$k = 60$	$k = 90$
Su-S-Locality ($d = 16$)	0.105	0.132	0.121
Su-S-Locality ($d = 32$)	0.145	0.146	0.176
Su-S+L-Locality ($d = 16$)	0.094	0.122	0.104
Su-S+L-Locality ($d = 32$)	0.132	0.116	0.144

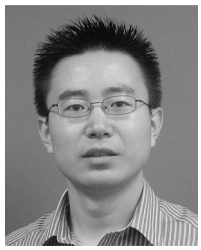
V. CONCLUSION

In this paper, we have discussed the basic issue in the linear subspace learning field—which criterion, globality or locality, is more effective to enhance the discriminating power of feature extraction? Measured by the discriminating power coefficient, we demonstrate in both theory and experiment that applying the local manner in sample, feature, and learning spaces via linear subspace learning can improve the discriminating performance of globality and thereby significantly enhance the classification accuracy. We also address the problem of numerical difficulty in the scenario of large size recognition, where many local variations cannot be adequately handled by a single global model. By localizing the modeling, more discriminative bases and multiple models are exploited to contribute to the multimodal piecewise subspace learning framework. This new strategy appears to be more robust and flexible for model selection among different model candidates. In the extensive experiments and evaluations, the proposed locality learning framework is demonstrated to be effective, robust, and fast for large size, multiclass, and multivariance data sets. In the future work, we will try to find out the criterion that can help us decide which types of localities should be applied to what kinds of scenarios. More attentions will be on designing advanced local metric modeling and judicious use of multiple models [21] for real-world applications in HCC.

REFERENCES

[1] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distribution," in *Proc. SCG'04*, 2004, pp. 253–262.
[2] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.
[3] X. He and P. Niyogi, "Locality preserving projections," presented at the NIPS'03, 2003.
[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[5] Y. Fu and T. S. Huang, "Locally linear embedded eigenspace analysis," [Online]. Available: www.ifp.uiuc.edu/~yunfu2/papers/LEA-Yun05.pdf, IFP-TR, UIUC, 2005
[6] Y. Fu and T. S. Huang, "Graph embedded analysis for head pose estimation," in *IEEE Conf. FG'06*, Southampton, U.K., 2006, pp. 3–8.
[7] Y. Fu, J. Yuan, Z. Li, T. S. Huang, and Y. Wu, "Query-driven locally adaptive Fisher faces and expert-model for face recognition," in *IEEE Conf. ICIP'07*, 2007, pp. 141–144.
[8] Y. Fu, Z. Li, T. S. Huang, and A. K. Katsaggelos, "Locally adaptive subspace and similarity metric learning for visual data clustering and retrieval," *Comp. Vis. Image Understand.*, vol. 110, no. 3, pp. 390–402, 2008.
[9] Z. Li, Y. Fu, J. Yuan, T. S. Huang, and Y. Wu, "Query driven localized linear discriminant models for head pose estimation," in *Proc. IEEE Conf. ICME'07*, 2007, pp. 1810–1813.

[10] Z. Li, L. Gao, and A. K. Katsaggelos, "Locally embedded linear subspaces for efficient video indexing and retrieval," in *Proc. IEEE Conf. ICME'06*, 2006, pp. 1765–1768.
[11] M. Liu, S. Yan, Y. Fu, and T. S. Huang, "Flexible X-Y patches for face recognition," in *Proc. IEEE Conf. ICASSP'08*, 2008, pp. 2113–2116.
[12] G.-D. Guo and C. Dyer, "Patch-based image correlation with rapid filtering," presented at the IEEE Conf. CVPR'07, 2nd Beyond Patches Workshop, Minneapolis, MN, Jun. 18–23, 2007.
[13] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VLDB'99*, 1999, pp. 518–529.
[14] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*. New York: Springer-Verlag, 2005.
[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2002.
[16] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, 2001.
[17] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, 2000.
[18] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, 2000.
[19] F. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Conf. WACV'94*, 1994, pp. 138–142.
[20] M. A. Turk and A. P. Pentland, "Face recognition using Eigenfaces," in *Proc. IEEE Conf. CVPR'91*, 1991, pp. 586–591.
[21] R. Xiao, W. J. Li, Y. D. Tian, and X. O. Tang, "Joint boosting feature selection for robust face recognition," in *Proc. IEEE Conf. CVPR'06*, 2006, pp. 1415–1422.
[22] M. H. Yang, "Face recognition using kernel methods," presented at the NIPS'01, 2001.
[23] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proc. ICML'04*, 2004, pp. 369–376.
[24] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, 2007.
[25] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Conf. CVPR'06*, 2006, pp. 2126–2136.
[26] A. Jaimes, N. Sebe, and D. Gatica-Perez, "Human-centered computing: A multimedia perspective," in *Proc. ACM Conf. Multimedia*, 2006, pp. 855–864.
[27] P. N. Belhumeur, J. P. Hepanha, and D. J. Kriegman, "Eigenfaces versus Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.
[28] J. Letissier and N. Gourier, "The pointing'04 data sets," presented at the IEEE ICPR Pointing'04 Workshop, 2004.
[29] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," presented at the IEEE ICPR Pointing'04 Workshop, 2004.
[30] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," *Face Recognition: From Theory to Applications*, ser. NATO ASI Series F, vol. 163, Computer & Systems Sciences, pp. 446–456, 1998.
[31] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," in *Proc. 7th Int. Conf. Database Theory*, 1999, pp. 217–235.
[32] B. Heisele and V. Blanz, "Morphable models for training a component-based face recognition system," in *Face Processing, Advanced Modeling and Methods*, W. Zhao and R. Chellapa, Eds. New York: Elsevier, 2006, pp. 439–462.
[33] Y. Gao, Y. Wang, X. Feng, and X. Zhou, "Face recognition using most discriminative local and global features," in *Proc. IEEE Conf. ICPR'06*, 2006, vol. 1, pp. 351–354.
[34] K. Huang, H. Yang, I. King, and M. R. Lyu, "Local learning versus Global learning: An introduction to maxi-min margin machine," in *Support Vector Machines: Theory and Applications*. New York: Springer-Verlag, 2005, vol. 177/2005, Studies in Fuzziness and Soft Computing, pp. 113–132.
[35] V. de Silva and J. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *Proc. NIPS'02*, 2002, pp. 705–712.
[36] T.-K. Kim and J. Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 318–327, Mar. 2005.
[37] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1493–1516, 1997.
[38] A. Weingessel and K. Hornik, "Local PCA algorithms," *IEEE Trans. Neural Netw.*, vol. 11, no. 6, pp. 1242–1250, 2000.



Yun Fu (S'07-M'08) received the B.Eng. degree in information engineering from the School of Electronic and Information Engineering, Xi'an Jiaotong University, China, in 2001; the M.Eng. degree in pattern recognition and intelligence systems from the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, in 2004; the M.S. degree in statistics from the Department of Statistics, University of Illinois at Urbana-Champaign (UIUC), in 2007; and the Ph.D degree in Electrical and Computer Engineering (ECE), ECE department, UIUC, in 2008.

From 2001 to 2004, he was a research assistant at the Institute of Artificial Intelligence and Robotics at XJTU. From 2004 to now, he is a graduate fellow and research assistant at the Beckman Institute for Advanced Science and Technology, ECE department and Coordinated Science Laboratory at UIUC. He was a research intern with Mitsubishi Electric Research Laboratories, Cambridge, MA, in summer 2005; with Multimedia Research Lab of Motorola Labs, Schaumburg, IL, in summer 2006. He jointed BBN Technologies, Cambridge, MA, as a Scientist in 2008 to build and lead the computer vision and machine learning team. His research interests include machine learning, human computer interaction, image processing, multimedia and computer vision.

Dr. Fu is the recipient of the 2002 Rockwell Automation Master of Science Award, two Edison Cups of the 2002 GE Fund Edison Cup Technology Innovation Competition, the 2003 HP Silver Medal and Science Scholarship, the 2007 Chinese Government Award for Outstanding Self-financed Students Abroad, the 2007 DoCoMo USA Labs Innovative Paper Award (IEEE ICIP'07 best paper award), the 2007-2008 Beckman Graduate Fellowship, and the 2008 M. E. Van Valkenburg Graduate Research Award. He is a life member of Institute of Mathematical Statistics (IMS) and 2007-2008 Beckman Graduate Fellow.

Zhu Li (SM'07) received the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, in 2004.

He was with the Multimedia Research Lab (MRL), Motorola Labs, 2000–2008, where he was a Principal Staff Research Engineer. He joined faculty of Dept of Computing, Hong Kong Polytechnic University in 2008. His research interests include video coding and communication, game theory and optimization decomposition techniques in multimedia streaming and networking, manifold modeling and machine learning in biometrics, multimedia analysis, retrieval and mining. He has 10 issued or pending patents, 30 publications in book chapters, journals and conference proceedings in these areas.

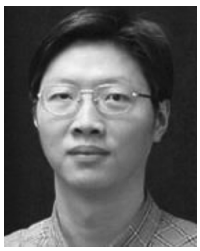
Dr. Li received the Best Poster Paper Award at IEEE Int'l Conf. on Multimedia & Expo (ICME), Toronto, 2006, and the DoCoMo Labs Innovative Paper Award (Best Paper) at IEEE Int'l Conf. on Image Processing (ICIP), San Antonio, 2007.

Junsong Yuan (S'06) received the B.S. degree in communication engineering from the Special Program for the Gifted Young of Huazhong University of Science and Technology, Wuhan, China, in 2002, and the M.Eng. degree from the National University of Singapore, Singapore, in 2005. He is currently pursuing the Ph.D. degree in electrical engineering and computer science at Northwestern University, Evanston, IL.

During the summer of 2008, 2007 and 2006, he was a research intern with the Communication and

Collaboration Systems group, Microsoft Research, Redmond, WA, Kodak Research Labs, Rochester, NY, and Motorola Labs, Schaumburg, IL, respectively. From 2003 to 2004, he was a research assistant in the Institute for Infocomm Research in Singapore. His research interests include computer vision, multimedia data mining, and statistical machine learning.

Mr. Yuan was awarded the National Outstanding Student and the Hu-Chunan fellowship by the Ministry of Education of China in 2001.



Ying Wu (SM'06) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1994, the M.S. degree from Tsinghua University, Beijing, China, in 1997, and the Ph.D.degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001.

From 1997 to 2001, he was a Research Assistant at the Beckman Institute for Advanced Science and Technology at UIUC. During summer 1999 and 2000, he was a research intern with Microsoft Research, Redmond, WA. In 2001, he joined the Department of Electrical and Computer Engineering at Northwestern University, Evanston, IL, as an Assistant Professor. He is currently an Associate Professor of Electrical Engineering and Computer Science at Northwestern University. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction.

Dr. Wu serves as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, SPIE *Journal of Electronic Imaging*, and IAPR *Journal of Machine Vision and Applications*. He received the Robert T. Chien Award at UIUC in 2001, and the NSF CAREER award in 2003.

Thomas S. Huang (S'61–M'63–SM'7-6-F'79–LF'01) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, China, and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was at the faculty of the Department of Electrical Engineering at MIT from 1963 to 1973, and at the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at 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 Science, and a Fellow of the International Association of Pattern Recognition and the Optical Society of America, and has received a Guggenheim Fellowship, an A. von Humboldt Foundation Senior US 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. He is a Founding Editor of the *International Journal of Computer Vision, Graphics, and Image Processing* and Editor of the *Springer Series in Information Sciences*, published by Springer.