# Synchronized Submanifold Embedding for Person-Independent Pose Estimation and Beyond

Shuicheng Yan, *Member, IEEE*, Huan Wang, Yun Fu, Jun Yan, Xiaoou Tang, *Fellow, IEEE*, and Thomas S. Huang

***Abstract*—Precise 3-D head pose estimation plays a significant role in developing human-computer interfaces and practical face recognition systems. This task is challenging due to the particular appearance variations caused by pose changes for a certain subject. In this paper, the pose data space is considered as a union of submanifolds which characterize different subjects, instead of a single continuous manifold as conventionally regarded. A novel manifold embedding algorithm dually supervised by both identity and pose information, called *snchronized submanifold embedding* (SSE), is proposed for *person-independent* precise 3-D pose estimation, which means that the testing subject may not appear in the model training stage. First, the submanifold of a certain subject is approximated as a set of simplexes constructed using neighboring samples. Then, these simplexized submanifolds from different subjects are embedded by synchronizing the locally propagated poses within the simplexes and at the same time maximizing the intrasubmanifold variances. Finally, the pose of a new datum is estimated as the propagated pose of the nearest point within the simplex constructed by its nearest neighbors in the dimensionality reduced feature space. The experiments on the 3-D pose estimation database, CHIL data for CLEAR07 evaluation, and the extended application for age estimation on FG-NET aging database, demonstrate the superiority of SSE over conventional regression algorithms as well as unsupervised manifold learning algorithms.**

***Index Terms*—Age estimation, manifold learning, simplex, subspace learning, 3-D pose estimation.**

## I. Introduction

A face image encodes a variety of useful information, such as identity [29], emotion [25], and head pose [1], which are significant for developing practical and humanoid computer vision systems. The problems of identity verification and emotion recognition have been extensively studied as conventional multiclass pattern recognition problems in the computer vision literature. Many commercial systems have been developed for human identity verification. However, the research on head pose estimation, especially for precise 3-D head pose estimation, is still far from mature due to the underlying difficulties and challenges. First, the database and ground truth are much more difficult to obtain than the identity and emotion information. Second, the style of pose variation is personalized, and greatly depends on the 3-D geometry of the human head. Finally, the pose labels are of real values, and, hence, the pose estimation problem is essentially a regression problem rather than a multiclass pattern recognition problem.

Current research [5], [9], [16], [27] on appearance based head pose estimation can be roughly divided into three categories. The first category [15], [16] formulates pose estimation as a conventional multiclass pattern recognition problem, and only rough pose information is inferred from these algorithms. The second category takes pose estimation as a regression problem, and nonlinear regression algorithms, e.g., Neural Network [5], are used for learning the mapping from the original appearance features to the pose label. The last category assumes that the pose data lie on or nearly on a low-dimensional manifold, and manifold embedding techniques [6], [8], [9], [12], [18], [19] are utilized for learning a more effective representation for pose estimation. In this work, we address the challenging problem of person-independent precise 3-D head pose estimation, instead of the rough discrete pose estimation in the pan direction as done conventionally, and, hence, the algorithms like linear discriminant analysis [10] from the first category are inapplicable in our scenario. To effectively exploit the underlying geometry structure information of the pose data space as well as the available identity and pose information, our solution is pursued within the third category, but many algorithms within this category, e.g., [18], are not suitable for the task we concern in this paper since they were proposed with the underlying assumption that the testing subject appears in the model training set.

In this paper, we present a dually supervised manifold embedding algorithm for person-independent precise 3-D head pose estimation motivated from the following observations: 1) the pose sample data are often from multiple subjects, and distributed on distinctive submanifolds of different subjects instead of a single continuous manifold assumed by most conventional manifold learning [4], [22], [26] algorithms, such as ISOMAP [23], Locally Linear Embedding (LLE) [21], and Laplacian Eigenmaps [3]; 2) these submanifolds commonly share similar geometric shapes as shown in Fig. 1; and 3) a

Manuscript received September 08, 2007; revised April 04, 2008. Current version published December 12, 2008. Part of this manuscript was submitted as evaluation report to CLEAR07. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dan Schonfeld.

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

H. Wang is with the Department of Computer Science, Yale University, New Haven, CT 06511 USA (e-mail: huan.wang@yale.edu).

Y. Fu and T. S. Huang are with the Beckman Institute, University of Illinois, at Urbana-Champaign, Urbana, IL 61820 USA (e-mail: yunfu2@ifp.uiuc.edu; huang@ifp.uiuc.edu).

J. Yan is with the Microsoft Research Asia, Beijing, China, 100080 (e-mail: junyan@microsoft.com).

X. Tang is with the Department of Information Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong, and also with Microsoft Research Asia, Beijing, China, 100080 (e-mail: xtang@ie.cuhk.edu.hk; xitang@microsoft.com).

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.2008.2006400

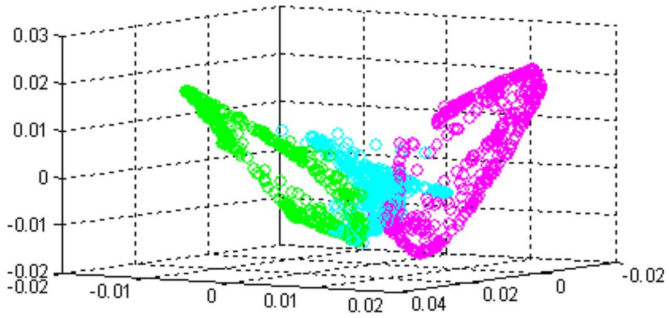


Fig. 1. Three-dimesional embedding of the pose data from three subjects. The data are shown to lie on three distinctive submanifolds instead of a single continuous manifold.

desirable representation for 3-D head pose estimation should be person-independent, namely the model trained on training data has good generalization capability on data from unknown subject.

Our proposed manifold embedding algorithm is dually supervised by both identity and pose information. More specifically speaking, first, the submanifold of each subject is approximated as a set of simplexes [17] constructed using neighboring samples, and the pose label is further propagated within all the simplexes by using the generalized barycentric coordinates [17]. Then these submanifolds are synchronized by seeking the counterpart point of each sample within the simplexes of a different subject, and consequently the synchronized submanifold embedding is formulated to minimize the distances between these aligned point pairs and at the same time maximize the intra-submanifold variance. Finally, for a new datum, a simplex is constructed using its nearest neighbors measured in the dimensionality reduced feature space, and then its pose is estimated as the propagated pose of the nearest point within the simplex.

Here we highlight some aspects of our proposed algorithm, referred to as *Synchronized Submanifold Embedding* (SSE), by comparing with conventional manifold learning algorithms for head pose estimation.

1) *What are the advantages of SSE over unsupervised manifold learning algorithms?* In the past decade, unsupervised manifold learning techniques have attracted much attention for both theoretical research and practical applications. Among them, ISOMAP [23], LLE [21], and Laplacian Eigenmaps [3] are the most popular ones. Most of these algorithms are unsupervised, and, hence, the derived low-dimensional representation is not guaranteed to be optimal for classification or regression problems. SSE sufficiently utilizes both the pose label information and the identity information to alleviate the difference of data from different subjects yet with similar poses, and, hence, it has the potential of yielding a more robust representation for person-independent precise 3-D head pose estimation.

2) *Why not to use conventional supervised manifold learning algorithms?* Manifold learning was previously explored in a supervised manner [20] by considering the labeling information for computing the local distances or similarities. That is, the distance computed on features is replaced by the product or weighted sum of the distances computed on features and on pose labels. However, for this type of supervised algorithm, the local distance or similarity is often dominated by the label information, and the derived representation does not essentially reflect the original manifold structure, and, hence, they are more like conventional general supervised learning algorithms than manifold learning algorithms. Moreover, conventional supervised manifold learning algorithms, such as Supervised LLE [20], were designed for multiclass classification problems and, hence, inapplicable for precise 3-D head pose estimation. SSE instead aligns the submanifolds with the propagated pose labels and within each submanifold, the manifold information is sufficiently retained. Hence, SSE is supervised and also follows the essence of manifold learning. Finally, SSE is dually supervised by both pose labels and identity information, while instead conventional supervised manifold learning algorithms utilize only one type of information.

The rest of the paper is organized as follows. Section II introduces the motivations from conventional manifold learning algorithms, followed by the formulation of synchronized submanifold embedding. Experimental results on precise 3-D head pose estimation, and the extended application of age estimation, are demonstrated in Section III. We conclude this paper in Section IV.

## II. Synchronized Submanifold Embedding for Person-Independent Pose Estimation

Here, we assume that the training sample data are given as $X^c = [x_1^c, x_2^c, \cdots, x_{n_c}^c]$, where $x_i^c \in \mathbb{R}^m$, $i = 1, 2, \cdots, n_c$, and $c = 1, 2, \cdots, N_c$. $n_c$ is the number of training samples for the $c$th subject, $N_c$ is the number of subjects, and we have $N = \sum_{c=1}^{N_c} n_c$ samples in total. Correspondingly, the pose labels are presented as $\Theta^c = [\theta_1^c, \theta_2^c, \cdots, \theta_{n_c}^c]$, $c = 1, 2, \cdots, N_c$, where $\theta_i^c \in \mathbb{R}^3$, $i = 1, 2, \cdots, n_c$ and three values of $\theta_i^c$ are the pan, tilt, and yaw angles of the sample $x_i^c$. For ease of presentation, we denote the concatenated sample data as $X = [x_1, x_2, \cdots, x_N]$ and the concatenated label matrix as $\Theta = [\theta_1, \theta_2, \cdots, \theta_N]$.

### A. Motivations

Recent work [6], [9], [19] demonstrated the effectiveness of manifold learning techniques for head pose estimation. The high-dimensional pose data are assumed to lie on or nearly on a low-dimensional continuous manifold, and the manifold learning techniques such as LLE and Laplacian Eignmaps, or their linear extensions [9], [12], are used for manifold embedding. Then the Nearest Neighbor criterion [10] or other simple linear regression approach is used for final head pose estimation.

Though there were some attempts [20] to develop supervised manifold learning algorithms for multiclass classification problems, most manifold learning algorithms run in an unsupervised manner for regression problems like precise 3-D head pose estimation. Our work presented in this paper is motivated by the observation that both *identity* and *pose* information are mostly available in the model training stage and they are useful for developing effective person-independent precise head pose estimation algorithm. More specifically speaking, it is commonly believed [10] that for regression or classification problems, the label information can greatly improve algorithmic performance compared with the unsupervised algorithms which utilize only original feature information. Besides the pose label information, the identity information is valuable for *person-independent* head pose estimation. On the one hand, the pose data often come from multiple subjects, and lie on separated distinctive submanifolds; hence, the assumption that the data lie on a single continuous manifold cannot be satisfied in this scenario. On the other hand, the submanifolds of different subjects often share similar geometric structures as shown in Fig. 1, and the algorithmic person-independence and generalization capabilities can be further promoted by synchronizing the pose labels on different submanifolds.

To sufficiently utilize both the pose label information and the identity information, we provide as follows a dually supervised algorithm, called synchronized submanifold embedding, to seek an effective representation for person-independent precise 3-D head pose estimation.

### *B. Synchronized Submanifold Embedding*

As shown in Fig. 1, the pose image data of a certain subject constitute a continuous submanifold. To obtain a person-independent representation for 3-D head pose estimation, it is natural to learn a low-dimensional subspace by synchronizing these submanifolds, such that the samples from different subjects yet with similar poses will be projected to similar low-dimensional representations.

Before formally describing our solution to learn such a subspace, we review some terminologies on simplex [17] and generalized barycentric coordinates.

A *$k$-simplex* is a $k$-dimensional analogue of a triangle. Specifically, a $k$-simplex is the convex hull of a set of $(k+1)$ affinely independent pointd[1] in some Euclidean space of dimension $k$ or higher. Mathematically speaking, denote the vertices as $Z = [z_0, z_1, \cdots, z_k]$, and then the $k$-simplex is expressed as

$$\mathcal{S}_k = \left\{ \sum_{j=0}^{k} t_j z_j : \sum_{j=0}^{k} t_j = 1, t_j \geq 0 \right\}. \qquad (1)$$

The coordinates $[t_0, t_1, \cdots, t_k]$ in $\mathcal{S}_k(Z)$ are called the generalized barycentric coordinates, which are the generalization of barycentric coordinates. An illustration of simplexes is shown in Fig. 2.

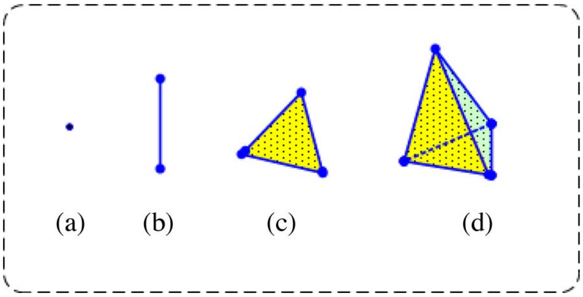


Fig. 2. Illustration of simplexes: (a) 0-simplex, (b) 1-simplex, (c) 2-simplex, and (d) 3-simplex.

*1) Submanifold Simplexization:* As the head pose label can be of any real value within [0 360), it is often difficult to obtain images with exactly the same poses yet from different subjects. Hence, the submanifolds cannot be directly aligned based on these discrete training samples, and traditional supervised subspace learning algorithms like Linear Discriminant Analysis [2] cannot be used for the task we concern in this paper.

In this work, we present an approach to transform the labeled discrete samples on a submanifold into a set of continuous simplexes with propagated pose labels. For each sample datum $x_i^c$, the $k$-nearest neighbors of the same subject measured by pose label distance are used to construct a $k$-simplex as

$$\mathcal{S}_k\left(x_i^c\right) = \left\{ \sum_{j=0}^{k} t_j x_{i_j}^c : \sum_{j=0}^{k} t_j = 1, t_j \geq 0 \right\} \qquad (2)$$

where $\{x_{i_j}^c, j = 1, 2, \cdots, k\}$ is the $k$ nearest neighbors of sample $x_i^c$ within the same submanifold and $x_{i_0}^c = x_i^c$.

Motivated by the work of LLE [21], we assume in this work that the nonnegative linear reconstruction relationship within the $k$-simplex $\mathcal{S}_k(x_i^c)$ can be bidirectionally transformed between features and pose labels. That is, for a point within $\mathcal{S}_k(x_i^c)$, denoted as $y_k^t(x_i^c) = \sum_{j=0}^{k} t_j x_{i_j}^c$, its pose label can be propagated from the poses of vertices using the same corresponding generalized barycentric coordinate vector $t$ as

$$\theta_k^t\left(x_i^c\right) = \sum_{j=0}^{k} t_j \theta_{i_j}^c. \qquad (3)$$

Note that the bidirectional propagation of the generalized barycentric coordinates between features and labels is assumed only within a local neighborhood like the $k$-simplex around a certain sample, which is in accord with the general locally linear assumption of a manifold [21].

In this way, beyond a set of discrete samples, each submanifold is expressed as a set of labeled continuous simplexes, and then for each datum $x_i^c$, it has the potential to find a counterpart point with the same pose within the simplexes of any other subject. Consequently, these submanifolds of different subjects can be synchronized by aligning these data pairs.

*2) Submanifold Embedding by Pose Synchronization:* As described above, we aim to pursue a low-dimensional representation such that the submanifolds of different subjects are aligned according to the precise pose labels. For each sample $x_i^c$, the point within the reconstructed simplexes of the $c'$th subject $(c' \neq c)$ and with the most similar pose is calculated in two steps. First, the generalized barycentric coordinates of this point is computed as

$$(\tilde{o}, \tilde{t}) = \arg\min_{o,t} \left\| \theta_i^c - \theta_k^t\left(x_o^{c'}\right) \right\|^2 \qquad (4)$$

then the corresponding datum and label are derived as

$$y\left(x_i^c, c'\right) = \sum_{j=0}^{k} \tilde{t}_j x_{\tilde{o}_j}^{c'} \qquad (5)$$

[1] In this work, the affinely independent property is assumed for the $k$ nearest neighbors of a datum, which is commonly satisfied since $k$ is small in our experiments.

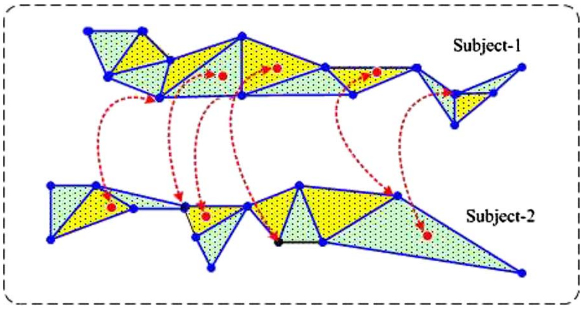


Fig. 3. Illustration of submanifold synchronization by simplicization. Note that to facilitate display, we utilize the 2-simplex for demonstration and the Euclidian distance in the 2-D plane does not exactly reflect true distance between sample pair. The blue points represent the training samples, and the red points represent the corresponding synthesized points in distinctive submanifolds with the same poses. The dashed bidirectional lines connect the point pairs with the same poses.

$$\theta\left(x_i^c, c'\right) = \sum_{j=0}^{k} \tilde{t}_j \theta_{o_j}^{c'}. \tag{6}$$

*Remark:* For a given $o$, the task becomes a standard quadratic optimization problem

$$\min_t \left\| \theta_i^c - \sum_{j=0}^{k} t_j \theta_{o_j}^{c'} \right\|^2, \ st. \ \sum_{j=0}^{k} t_j = 1, t_j \geq 0 \tag{7}$$

which can be solved by general optimization tool, such as the *quadprog* function in Matlab.

There are serval ways to derive a low-dimensional representation for synchronizing these submanifolds, and in this paper, we utilize the linear projection approach, namely, the manifold embedding is achieved by seeking a projection matrix $W \in \mathbb{R}^{m \times d}$ (usually $d \ll m$) and

$$y_i = W^T x_i \tag{8}$$

where $y_i \in \mathbb{R}^d$ is the low-dimensional representation of sample $x_i$.

On the one hand, the projection matrix $W$ should minimize the distances between each sample to its nearest neighbor (measured by pose label distance) within the simplexes of any other subject. Namely, it should minimize

$$\hat{S}_{syn}(W) = \sum_{c=1}^{N_c} \sum_{i=1}^{n_c} \sum_{c' \neq c} \left\| W^T x_i^c - W^T y\left(x_i^c, c'\right) \right\|^2 I\left(x_i^c, c'\right) \tag{9}$$

where the indicator function $I(x_i^c, c') = 1$, if $\|\theta_i^c - \theta(x_i^c, c')\| \leq \varepsilon$; 0, otherwise. $\varepsilon$ is a threshold to determine whether to synchronize the point pairs, and in this work, $\varepsilon$ is set as 2 for the pose estimation problem and as 1 for age estimation in the extended application.

On the other hand, to promote the separability of different poses, it is desirable to maximize the distances between different sample pairs, namely

$$\hat{S}_{sep}(W) = \sum_{c=1}^{N_c} \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} \left\| W^T x_i^c - W^T x_j^c \right\|^2. \tag{10}$$

To achieve these dual objectives, the projection matrix $W$ is derived as

$$\arg\max_W \frac{\hat{S}_{sep}(W)}{\hat{S}_{syn}(W)} = \arg\max_W \frac{Tr(W^T S_1 W)}{Tr(W^T S_2 W)} \tag{11}$$

where

$$S_2 = \sum_{c=1}^{N_c} \sum_{i=1}^{n_c} \sum_{c' \neq c} \left(x_i^c - y\left(x_i^c, c'\right)\right)\left(x_i^c - y\left(x_i^c, c'\right)\right)^T I\left(x_i^c, c'\right)$$
$$S_1 = \sum_{c=1}^{N_c} \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} \left(x_i^c - x_j^c\right)\left(x_i^c - x_j^c\right)^T. \tag{12}$$

The objective function in the optimization problem (11) is nonlinear and commonly there is no closed form solution. Usually, it is transformed into another more attractive form as $\arg\max_W Tr[(W^T S_2 W)^{-1}(W^T S_1 W)$ and solved with the generalized eigenvalue decomposition method as

$$S_1 w_i = \lambda_i S_2 w_i \tag{13}$$

where the vector $w_i$ is the eigenvector corresponding to the $i$th largest eigenvalue $\lambda_i$, and it constitutes the $i$th column vector of the projection matrix $W$.

### C. Pose Estimation by Local Simplex Propagation

After we obtain the projection matrix $W$, the sample data are all transformed into the low-dimensional feature space as in (8), and then all the training samples are denoted as $Y = [y_1, y_2, \cdots, y_N]$.

For a new datum $x$, first, we also transform it into the low-dimensional feature space as $y = W^T x$. Then, we search for its nearest point within the simplex constructed using its $(k+1)$ nearest neighbors in the low-dimensional feature space, namely, search for the generalized barycentric coordinate vector $\tilde{t}$ by optimizing

$$\min_t \left\| y - \sum_{j=0}^{k} t_j y_{i_j} \right\|^2, \ \text{s.t.} \ \sum_{j=0}^{k} t_j = 1, t_j \geq 0 \tag{14}$$

where $y_{i_j}, j = 0, 1, \cdots, k$, are the $(k+1)$ nearest samples of $y$. Then, the label of the new datum is predicted by propagating the generalized barycentric coordinates to the labels of the vertices of the constructed $k$-simplex

$$\theta_x = \sum_{j=0}^{k} \tilde{t}_j \theta_{i_j}. \tag{15}$$

## III. Experiments

In this section, we systematically evaluate the effectiveness of our proposed algorithm, synchronized submanifold embedding (SSE), for person-independent precise 3-D head pose estimation. We use the latest precise 3-D head pose estimation database, CHIL data, from the CLEAR07 evaluation [30] for the experiments. To further demonstrate the generality of SSE in person-independent estimation, we evaluate our algorithm on the age estimation problem and the popular aging database FG-NET [31]. For comparison, Principal Components Analysis (PCA) [13], [24] and Locally Embedded Analysis (LEA) [9] are implemented. As mentioned beforehand, conventional supervised subspace learning algorithm LDA cannot be directly applied for the 3-D pose estimation problem, and, hence, we did not implement for comparison. The pose estimation from PCA and LEA is also based on local simplex propagation as in SSE. Also, the popular regression algorithms Neural Network (NN) [5] and Quadratic Models (QM) [14] are implemented for comparison in both pose and age estimation.

### A. Person-Independent Precise 3-D Head Pose Estimation

*1) Data Set: CHIL Data for CLEAR07 Evaluation:* The CLEAR workshop [30] is an international effort to evaluate systems that are designed to recognize events, activities, and their relationships in interaction scenarios. In this work, we use the latest pose database, CHIL data, in the CLEAR07 evaluation, and this database is intended for precise 3-D head pose estimation.

In the CHIL data, observations from four cameras that are placed in a room's upper corners are obtained for each subject. This data set includes 15 different persons standing in the middle of the room, rotating their heads towards all possible directions while wearing a magnetic motion sensor (Flock of Birds) in order to obtain their ground truth head orientations. The task is to estimate the head orientations with respect to the room's coordinate system, thus to obtain a joint estimate from all four views to achieve a hypothesis more robust than estimating from just one single camera. Some sample data are displayed in Fig. 4, and the four images in each column are from the same subject and captured by four cameras simultaneously. Precise 3-D pose estimation in this scenario is very difficult due to the fact that the images are in a very low resolution and also noisy.

In our experiments, we use the same experimental configuration as designed by the evaluation committees. For training, ten videos, including annotations of the head bounding boxes and the original ground truth information about the true head pose, are provided. For evaluation, five videos along with the head bounding box annotations are provided. The ground truth information is used for scoring. People appearing in the training set do not appear in the evaluation set. Since manual annotations of the head bounding box only occur at every fifth frame of the videos, only hypotheses corresponding to these time stamps are going to be scored [30]. Finally, the training set contains 5348 pose samples (each sample consists of four images captured by four different cameras) from the 10 subjects, and the testing set contains 2402 pose samples. Each image is cropped and scaled to size 40-by-40, and then gray-level values of all the four images are concatenated as the feature vector for each pose sample.

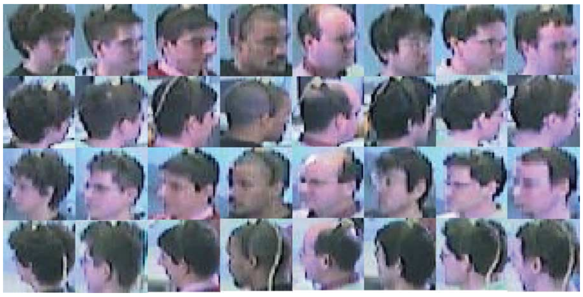


Fig. 4. Cropped sample images in the CHIL data for CLEAR07 evaluation. Note that each column contains four images of the same subject captured by the four cameras.

For all the experiments, we conduct PCA and reduce the feature dimension to 300, and then all the other algorithms are performed on the dimensionality reduced feature space. The Mean Absolute Error (MAE) [14] is used for accuracy evaluation.

*2) Embedding Visualization and Divergency:* The algorithms PCA, LEA, and SSE all provide a linear embedding of the manifold/submanifold from the original feature space to a low-dimensional feature space. As described before, the person-independent property is critical for algorithmic generalization to unknown testing subjects.

In this subsection, we evaluate the person-independent characteristic of the low-dimensional feature space derived from the training set of CHIL data. The left plot in Fig. 5 displays the $3-D$ distribution of samples from different subjects at different poses in the derived feature space from SSE, and we can observe that the samples of different subjects yet at similar poses gather together in the feature space, which coincides with the target of our SSE algorithm.

The right plot in Fig. 5 shows the divergency of the dimensionality reduced samples around 10 poses compared between PCA, LEA, and SSE. For computing the divergency, the feature dimension is set as 3, and the divergency is defined as the standard deviation of the nine samples around certain pose, normalized by the standard deviation of all the samples in the training set. The results shows that the divergency based on the submanifold embedding from SSE is much smaller than those based on the manifold embeddings from PCA and LEA. The low divergency ensures a good generalization capability of SSE on the testing data.

*3) Precise 3-D Head Pose Estimation Results:* By following the experimental configuration for the CLEAR07 evaluation, we evaluated the performance of the algorithms PCA, LEA, QM, NN, and SSE. We also implemented a supervised version of LEA by extending the algorithm in [20], and the corresponding algorithm is referred to as Supervised LEA (SLEA). The detailed results are shown in Fig. 6 and Table I. From these results,[2] we can have the following observations.

[2] Our reported results here are better than what we reported in [30], because we further refined the algorithmic parameters. The NN results reported in [30] are slightly better that what we reported here because they used extra features besides image intensities.

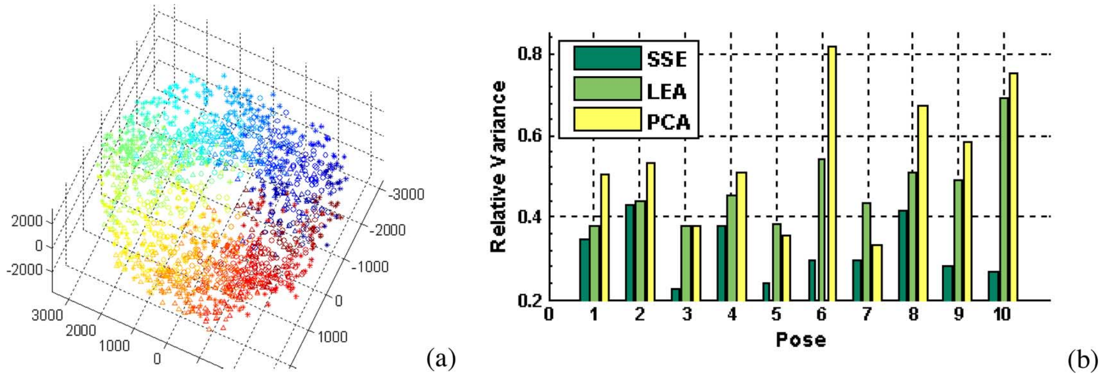


Fig. 5. Embedding visualization and divergency evaluation: (a) the 3-D distribution of the samples from different subjects and at different poses; and (b) the divergency of samples around certain poses for the algorithms PCA, LEA, and SSE. Note that in plot (a), the shapes of the samples reflect different subjects and the colors reflect difference poses, and only three subjects are shown for the ease of display; and the horizontal axes are the indexes of the ten randomly selected poses in plot (b).

TABLE I
MEAN ABSOLUTE ERRORS OF THE ALGORITHMS PCA, LEA/SLEA, QM, NN, AND SSE ON THE CHIL DATA OF THE CLEAR07 EVALUATION. NOTE THAT THE OPTIMAL PARAMETERS USED FOR DIFFERENT SUBJECTS ARE DIFFERENT, AND THE TOTAL AVERAGE IS NOT THE WEIGHTED AVERAGE OF THE RESULTS FROM THE FIVE SUBJECTS

| | Subject-1 | Subject-2 | Subject-3 | Subject-4 | Subject-5 | Total Average |
|---|---|---|---|---|---|---|
| Pan-PCA | 8.54 | 8.19 | 6.91 | 4.53 | 4.78 | 6.94 |
| Pan-LEA | 7.60 | 8.77 | 6.33 | 4.50 | 4.51 | 6.72/6.68 (SLEA) |
| Pan-QM | 61.31 | 55.37 | 41.27 | 33.45 | 38.12 | 46.78 |
| Pan-NN | 64.17 | 45.83 | 40.27 | 39.80 | 41.09 | 46.82 |
| Pan-SSE | 8.45 | 7.27 | 6.22 | 4.33 | 3.94 | **6.60** |
| Tilt-PCA | 8.49 | 5.97 | 11.59 | 5.25 | 12.53 | 10.86 |
| Tilt-LEA | 7.88 | 5.74 | 12.29 | 5.29 | 12.23 | 10.87/10.86 (SLEA) |
| Tilt-QM | 7.97 | 8.10 | 17.38 | 8.05 | 23.48 | 11.83 |
| Tilt-NN | 14.42 | 15.99 | 17.85 | 11.48 | 16.37 | 15.10 |
| Tilt-SSE | 8.61 | 6.28 | 9.08 | 4.92 | 9.64 | **8.25** |
| Roll-PCA | 4.66 | 2.59 | 4.20 | 2.86 | 3.30 | 4.01 |
| Roll-LEA | 5.41 | 2.59 | 4.06 | 2.90 | 2.91 | 4.07/3.97 (SLEA) |
| Roll-QM | 7.68 | 5.83 | 10.33 | 6.61 | 7.01 | 7.55 |
| Roll-NN | 11.24 | 10.07 | 12.23 | 12.04 | 11.79 | 11.44 |
| Roll-SSE | 5.55 | 2.22 | 3.72 | 2.38 | 2.34 | **3.42** |

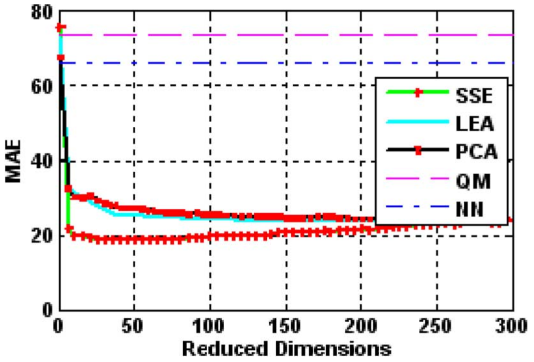


Fig. 6. Sum of total average MAEs for PCA, LEA, and SSE on different feature dimensions for precise 3-D pose estimation on the CHIL data in the CLEAR07 evaluation. Note that the results of QM and NN are expressed as lines in the figure, and MAE is the sum of the MAEs for three different directions.

1) SSE consistently achieves lower MAE than PCA, LEA, QM, and NN for both individual subject evaluation and overall evaluation.
2) NN and QM perform badly in this experiment, which should come from the fact that the training subjects and the testing subjects are different, and NN as well as QM lack enough generalization capability, since they did not explicitly pursue the person-independence.
3) The performance of LEA is better than that of PCA, which validates the effectiveness of exploiting manifold structure of the data space for pose estimation [9]. SLEA is generally better than LEA.

### B. *Beyond: Person-Independent Age Estimation*

Besides precise 3-D head pose estimation, our proposed algorithm SSE can be used for any regression problems containing images from different subjects. The general idea of SSE is to employ the identity information for pursuing person-independent representation. Here we take the age estimation problem as an example to demonstrate its potential applications in other domains.

*1) Data Set: FG-NET Aging Database:* The FG-NET aging database [31] is used in our experiments. It contains 1002 face images of 82 subjects with ages ranging from 0 to 69, and each subject has multiple images of different ages as shown in Fig. 7.

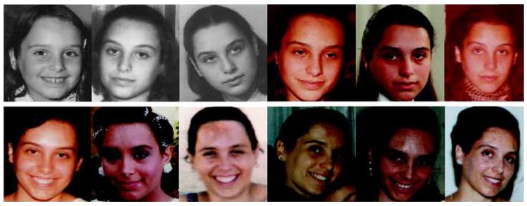


Fig. 7. Cropped samples from the FG-NET aging database. Note that all these images are from the same subject yet of different ages.

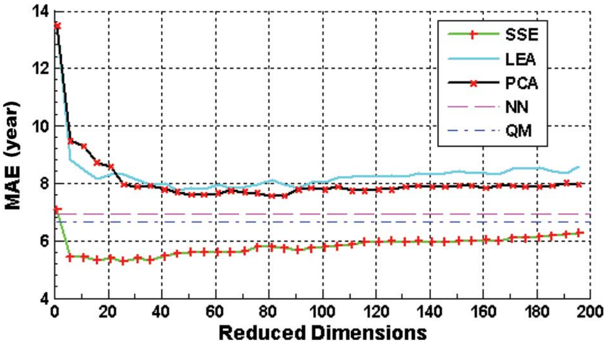


Fig. 8. Mean Absolute Errors of PCA, LEA, and SSE on different feature dimensions for age estimation on the FG-NET aging database. Note that the results of QM and NN are expressed as lines in the figure.

TABLE II
MEAN AVERAGE ERRORS OF THE ALGORITHMS PCA, NPE, QM, NN, AND SSE ON THE FG-NET DATABASE WITH THE LEAVE-ONE-PERSON-OUT STRATEGY

| Group (Sample Number) | *PCA* | *LEA* | *QM* | *NN* | *SSE* |
|---|---|---|---|---|---|
| Age 0- 9 (371) | 3.66 | 3.89 | 5.67 | 5.25 | **2.06** |
| Age 10-19 (339) | 4.81 | 4.85 | 5.54 | 5.24 | **3.26** |
| Age 20-29 (144) | 8.95 | 8.67 | 5.92 | **5.85** | 6.03 |
| Age 30-39 (79) | 15.00 | 13.02 | 10.27 | 11.29 | **9.53** |
| Age 40-49 (46) | 19.07 | 19.46 | 12.24 | 16.48 | **11.17** |
| Age 50-59 (15) | 18.67 | 26.13 | 18.60 | 28.80 | **16.00** |
| Age 60-69 (8) | 36.25 | 39.00 | 28.00 | 39.50 | **26.88** |
| Average | 7.49 | 7.65 | 6.70 | 6.95 | **5.21** |

The first 200 appearance parameters of Active Appearance Models [7] are extracted based on the provided 68 key facial points [14], and used as input features for age estimation. For detailed information on shape, texture, and appearance parameters, please refer to [7]. The Leave-One-Person-Out (LOPO) strategy is used to evaluate the performance of difference algorithm, and the Mean Absolute Error is again used for measuring accuracy as in the pose estimation experiments.

*2) Age Estimation Results:* Detailed experimental results are shown in Fig. 8 and Table II. The experimental results again validate the effectiveness of SSE over the PCA and LEA algorithms in estimation accuracy. The QM and NN algorithms work well in this experiments, and they perform better than both PCA and LEA. Despite the nonlinear property of the QM and NN algorithms, their performances are not as good as SSE which takes into account both the age information and the subject identity information, and promotes the generalization capability on the testing data.

## IV. DISCUSSIONS

In this paper, we presented a framework for precise 3-D head pose estimation by seeking effective submanifold embedding with the guidance of both pose and subject identity information. First, the submanifolds of different subjects are simplexized such that they can be synchronized according to the pose labels propagated within the simplexes. Then, submanifold embedding is derived by aligning the pose distribution within different submanifolds, and finally the pose label of a new datum is predicted as the propagated pose of the nearest point within the simplex constructed using its nearest neighbors in the derived low-dimensional feature space. The effectiveness of the proposed algorithm was validated by the experiments on the latest 3-D head pose estimation database, CHIL data for CLEAR07 evaluation, and its extended application for age estimation on the popular FG-NET aging database. One future research direction on this work is to develop dually supervised manifold embedding algorithms which can benefit both subject identification and the estimation of pose or age information simultaneously; and another direction is to extend this method for pose-independent face recognition, which is a dual problem of person-independent pose estimation discussed in this paper. Also, there often exists uncertainty/noise in the obtained pose label, so we are planning to study the regression problem with uncertain labels in our future work.

## REFERENCES

[1] S. Ba and J. Dobez, "A probabilistic framework for joint head tracking and pose estimation," in *Proc. Int. Conf. Pattern Recognition*, 2004, vol. 4, pp. 264–267.

[2] 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.

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

[4] C. Bregler and S. Omohundro, "Nonlinear image interpolation using manifold learning," *Adv. Neural Inf. Process. Syst.*, pp. 973–980, 1995.

[5] L. Brown and Y. Tian, "Comparative study of coarse head pose estimation," in *Proc. IEEE Workshop on Motion and Video Computing*, 2002, pp. 125–130.

[6] L. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang, "Head pose estimation using fisher manifold learning," in *Proc. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 203–207.

[7] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[8] A. Elgammal and C. Lee, "Separating style and content on a nonlinear manifold," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 1, pp. 478–485.

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

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

[11] 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.

[12] N. Hu, W. Huang, and S. Ranganath, "Head pose estimation by nonlinear embedding and mapping," in *Proc. IEEE Int. Conf. Image Processing*, 2005, pp. 342–345.

[13] I. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[14] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 621–628, Jan. 2004.

[15] S. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang, "Kernel machine based learning for multi-view face detection and pose estimation," in *Proc. Int. Conf. Computer Vision*, 2001, vol. 2, pp. 674–679.

[16] S. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," *IEEE Trans. Image Process.*, vol. 14, no. 6, pp. 705–712, Jun. 2005.

[17] J. Munkres, *Elements of Algebraic Topology*. New York: Perseus, 1993.

[18] H. Murase and S. Nayar, "Parametric eigenspace representation for visual learning and recognition," *Proc. SPIE*, vol. 2031, pp. 378–391, 1993.

[19] B. Raytchev, I. Yoda, and K. Sakaue, "Head pose estimation by nonlinear manifold learning," in *Proc. 17th Int. Conf. Pattern Recognition*, 2004, vol. 4, pp. 1051–4651.

[20] D. Ritte, O. Kouropteva, O. Okun, M. Pietikainen, and R. Duin, "Supervised locally linear embedding," in *Proc. Artificial Neural Networks and Neural Information*, 2003, pp. 333–341.

[21] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 22, pp. 2323–2326, Dec. 2000.

[22] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, 2003.

[23] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 22, pp. 2319–2323, 2000.

[24] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 13, pp. 71–86, 1991.

[25] H. Wang and N. Ahuja, "Facial expression decomposition," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, vol. 2, pp. 958–965.

[26] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 988–995.

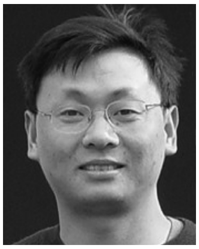
[27] M. Wenzel and W. Schiffmann, "Head pose estimation of partially occluded faces," in *Proc. 2nd Canad. Conf. Computer and Robot Vision*, 2005, pp. 353–360.

[28] S. Yan, D. Xu, B. Zhang, H. 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, Jan. 2007.

[29] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, pp. 399–458, 2003.

[30] [Online]. Available: http://isl.ira.uka.de/clear07/?The_Evaluation

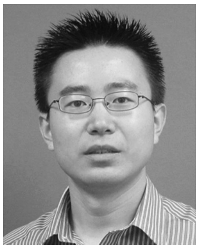
[31] The fg-Net Aging Database [Online]. Available: http://sting.cycollege.ac.cy/~alanitis/fgnetaging/index.htm



**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 with the Department of Electrical and Computer Engineering, National University of Singapore.

**Huan Wang** received the B.Eng. degree from Zhejiang University and the M.Phil. degree from The Chinese University of Hong Kong, both in the major of information engineering. He is currently pursuing the Ph.D degree at the Computer Science Department, Yale University, New Haven, CT. His research interests include artificial intelligence, machine learning, and computer vision.



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

From 2001 to 2004, he was a Research Assistant at the Institute of Artificial Intelligence and Robotics (AI&R), XJTU. From 2004 to 2007, he was a Research Assistant at the Beckman Institute for Advanced Science and Technology, UIUC. He was a Research Intern with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, Summer 2005, and with the Multimedia Research Lab of Motorola Labs, Schaumburg, IL, in Summer 2006. He jointed BBN Technologies, Cambridge, MA, as a Scientist in 2008. 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 for Excellent Chinese Student, the 2007 Chinese Government Award for Outstanding Self-financed Students Abroad, the 2007 DoCoMo USA Labs Innovative Paper Award (IEEE ICIP 2007 best paper award), the 2007–2008 Beckman Graduate Fellowship, and the 2008 M. E. Van Valkenburg Graduate Research Award. He is a student member of Institute of Mathematical Statistics (IMS) and a Beckman Graduate Fellow.

**Jun Yan** received the Ph.D. degree in digital signal processing and pattern recognition from the Department of Information Science, School of Mathematical Science, Peking University, China, in 2006.

He was a Research Associate at CBI, HMS, Harvard, Cambridge, MA, in 2005. Currently, he is an Associate Researcher with Microsoft Research Asia.

**Xiaoou Tang** (S'93–M'96–SM'02–F'08) received the B.S. degree from the University of Science and Technology of China, Hefei, in 1990, the M.S. degree from the University of Rochester, Rochester, NY, in 1991, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996.

He is a Professor and the Director of the Multimedia Lab, Department of Information Engineering, Chinese University of Hong Kong. He is also the group manager of the Visual Computing Group at the Microsoft Research Asia. His research interests include computer vision, pattern recognition, and video processing

Dr. Tang is a local chair of the IEEE International Conference on Computer Vision (ICCV) 2005, an area chair of ICCV'07, a program chair of ICCV'09, and a general chair of the ICCV International Workshop on Analysis and Modeling of Faces and Gestures 2005. He is a Guest Editor of the Special Issue on Underwater Image and Video Processing of the IEEE JOURNAL OF OCEANIC ENGINEERING and the Special Issue on Image- and Video-Based Biometrics of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

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

He was on the Faculty of the Department of Electrical Engineering at 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 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 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 Sciences, and a Fellow of the International Association of Pattern Recognition and the Optical Society of American. He 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.