Recognize High Resolution Faces: From Macrocosm to Microcosm

Dahua Lin
Dept. of Information Engineering
The Chinese University of Hong Kong
dhlin4@ie.cuhk.edu.hk

Xiaoou Tang
Microsoft Research Asia
Beijing, China
xitang@microsoft.com

Abstract

Human faces manifest distinct structures and characteristics when observed in different scales. Traditional face recognition techniques mainly rely on low-resolution face images, leading to the lost of significant information contained in the microscopic traits. In this paper, we introduce a multilayer framework for high resolution face recognition exploiting features in multiple scales. Each face image is factorized into four layers: global appearance, facial organs, skins, and irregular details. We employ Multilevel PCA followed by Regularized LDA to model global appearance and facial organs. However, the description of skin texture and irregular details, for which conventional vector representation are not suitable, brings forth the need of developing novel representations. To address the issue, Discriminative Multiscale Texton Features and SIFT-Activated Pictorial Structure are proposed to describe skin and subtle details respectively. To effectively combine the information conveyed by all layers, we further design an metric fusion algorithm adaptively placing emphasis onto the highly confident layers. Through systematic experiments, we identify different roles played by the layers and convincingly show that by utilizing their complementarities, our framework achieves remarkable performance improvement.

1 Introduction

Human face is an architecture consisting of distinct structures and characteristics when observed in different scales. As illustrated in figure 1, from macrocosm to microcosm, the views in different scales show greatly different pictures: a macroscopic view presents a basic configuration of facial organs, such as eyes, nose and mouth; while a microscopic perspective leads us to subtle details. Motivated by the belief that informations contained in these scales have their special roles respectively, we approach the face recognition problem by utilizing the features in all perceptible scales.

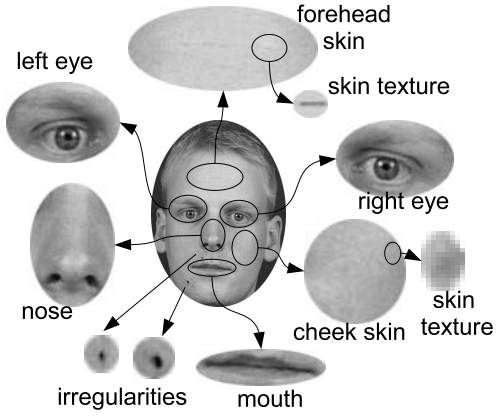

Figure 1. The Decomposition of Face: A face is composed of organs and skin, the skin is composed of organized textures.

We first give a brief review of current progress in face recognition techniques. The well-known Eigenfaces[14] introduces PCA to face representation, which effectively alleviates the high dimensionality difficulties in appearance-based approaches, and thus makes subspace methods increasingly popular. Thereafter, a number of dimension reduction algorithms are proposed, in which LDA[18] and its improved versions[25][4][13][26][27] are among the most successful. To cope with more complex intra-class variations, efforts have been devoted to their nonlinear extensions[7][19], and improved face descriptors including shape-texture-decomposition[23] and Gabor wavelet features[5].

Despite the success achieved by these methods, they still suffer from drastic performance degradation in the situations of remarkable environmental change. A fundamental limitation originates from that most of current approaches are based on low-resolution images. In these images, only a global appearance of the faces can be seen, while subtle details are blurred, thus important information embedded in these micro-structures are lost. Recent advancement

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

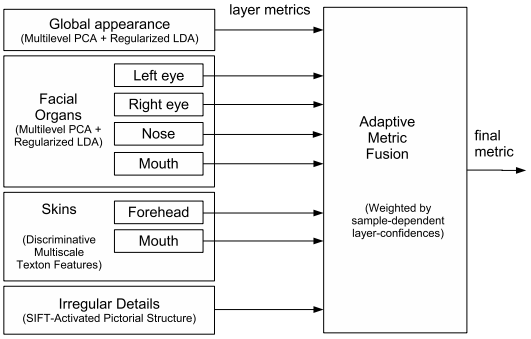

Figure 2. The multilayer framework.

of high-definition cameras and high-performance computers makes high-resolution face analysis possible and consequently brings us a great opportunity to break the limitation by making use of the microscopic features.

In this paper, we develop a multilayer framework for high resolution face recognition with information in multiple scales utilized. In the framework, each face is decomposed into 4 layers: global appearance, facial organs (eyes, nose, and mouth), skins (forehead skin, and cheek skin), and irregular details. We propose different feature description schemes for these layers, specially tailored to their different characteristics. They are listed below.

Global appearance & Facial Organs The global appearance and main organs of faces are highly structured, hence, a subspace model is appropriate. We first adopt Multilevel PCA for dimension reduction. Then Regularized LDA is proposed to transform the features to a discriminative subspace.
Skins Skin does not possess a general spatial structure, instead, it is textured. A typical skin is formed by repetition of texture units (textons). We propose the Discriminative Multiscale Texton Features: firstly a common skin texton dictionary is established, based on which, we describe the texture by texton distribution with soft histogram. R-LDA ensues to reduce the irrelevant variations.
Irregular details They have neither a regular appearance nor a spatial order. However, for each person, existence of some inherent irregular details is a strong personal distinction. We proposes SIFT-Activated Pictorial Structure, which unifies SIFT[6] for detecting and describing local interest regions and an elastic graph[17] for modelling their spatial configuration. Thereby, we can effectively express the distribution of irregular details on faces, and compute the similarity in a probabilistic manner.

Furthermore, we develop an adaptive metric fusion algorithm to combine the features in all layers. It constructs the final decision in a probabilistic way based on layer-confidences and emphasizes the highly distinctive layers. The integrated framework is shown in figure 2.

The major contributions of our framework lie in the following aspects. First, compared to most current face recognition techniques based on low resolution images (or downsampled version of high resolution images), our work gives a comprehensive solution to true-sense high resolution face recognition. Second, we take the lead in making use of microscopic structures, including skin textures and irregular details. Novel representations tailored to their particular natures are proposed, which are completely different from the conventional models based on fixed-dimensional vector space. To our best knowledge, both the Multiscale texton features and SIFT-activated pictorial structure are new methodologies in the field. They open up the microcosmic realm for face recognition. Third, we devise a flexible metric-fusion method with the layers adaptively weighted for different persons and under different conditions. Fourth, through systematic evaluations of different layers as well as the whole framework, we identify different roles played by these layers and present a new insight into face recognition.

2 Layer Models

2.1 Global Appearance and Facial Organs

The global appearance and facial organs are highly structured, thus they can be well represented by a fixed-length-vector, and it is reasonable to assume that the sample vectors reside on a subspace of much lower dimension. Hence, we learn subspace models for the global appearance and four facial organs (left and right eyes, nose, and mouth) respectively. For a high resolution image, the prohibitively high dimension (over 50000) of appearance-based representation renders traditional dimension reduction methods infeasible. To tackle this difficulty, the multilevel PCA strategy is adopted: we first partition the target area into sub-regions, so that each sub-region contains about $1000 \sim 2000$ pixels. Then PCA models are trained and applied to these sub-regions respectively. Finally, a high-level PCA is learned on the stacked vectors concatenating principal components for all sub-regions, in order to attain a more compact representation. For images in a higher resolution, such a PCA-hierarchy can be easily extended to more levels.

However, the main target of PCA is for compression, thus it may not lead to optimal discrimination. To extract the discriminant features, LDA[18] is further applied to the top-level principal components. It pursues a linear transform $\mathbf{W}$ maximizing the ratio of between-class scattering to within-class scattering.

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|}. \tag{1}$$

Suppose the training set has n images from C different persons, denoted by $\{(I_1, c_1), (I_2, c_2), \ldots, (I_n, c_n)\}$, where c_i is the label of the i-th sample. Denote the principal-component-vector for I_i by $\mathbf{x}_i$, then the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter matrix $\mathbf{S}_w$ are re-

spectively defined as

$$\mathbf{S}_b = \frac{1}{n}\sum_{k=1}^{C} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T, \quad (2)$$

$$\mathbf{S}_w = \frac{1}{n}\sum_{k=1}\sum_{i:c_i=k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T, \quad (3)$$

where $\mathbf{m}_k$ is the mean vector of the k-th person, and $\mathbf{m}$ is the overall mean. It is known that LDA tend to suffer from the singularity of $\mathbf{S}_w$ in a high dimensional space. A series of improved LDA algorithms are proposed to address the difficulty, including PCA+LDA[25], Enhanced Fisher Model[4], Unified Subspace Analysis[26], and Null-space LDA[13]. They resolve the problem at the expense of losing information in either principal subspace or null space of $\mathbf{S}_w$. Inspired by the works in[27][2], we derive the *Regularized LDA* as follows:

1. Perform eigen-decomposition on $\mathbf{S}_w$ as $\mathbf{S}_w = \mathbf{U}\mathbf{\Lambda}_w\mathbf{U}^T$. Here $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ are diagonal matrix of eigenvalues in descending order, $\mathbf{U}$ is orthogonal matrix composed of corresponding eigenvectors;
2. Determine the dimension or principal subspace of $\mathbf{S}_w$, denoted by r, so that 95% of the intra-class variation energy is preserved in the principal subspace Ω_p;
3. Divide $\mathbf{U}$ into $[\mathbf{U}_p,\ \mathbf{U}_c]$, which are the basis of Ω_p and Ω_c (the orthogonal complement of Ω_p);
4. Compute $\rho = \frac{1}{d-r}\sum_{i=r+1}^{d}\lambda_i$;
5. Compute the regularized whitening transform: $\mathbf{T}_1 = [\mathbf{U}_p\mathbf{\Lambda}_p^{-\frac{1}{2}},\ \rho^{-\frac{1}{2}}\mathbf{U}_r]$, then $\mathbf{K}_b = \mathbf{T}_1^T\mathbf{S}_b\mathbf{T}_1$;
6. Apply PCA on $\mathbf{K}_b$ to get $\mathbf{T}_2$, finally get $\mathbf{W} = \mathbf{T}_1\mathbf{T}_2$.

Compared with other LDA-based methods, there are two improvements: 1) Both the principal subspace and its complement are retained in step.5, thus no information would be lost in the whitening stage. Different from dual-space LDA[27], in which the two complementary subspaces are separately treated in ensuing steps, in our method they are combined before the second-stage diagonalization, thus can be utilized more effectively. 2) Our method has its theoretical root in statistics: it is equivalent to approximating $\mathbf{S}_w$ by a regularized nonsingular matrix $\widetilde{\mathbf{S}}_w$, and it can be proved[2] that such an approximation is optimal in sense of minimizing the K-L divergence between the Gaussian distributions induced by covariance matrices $\mathbf{S}_w$ and $\widetilde{\mathbf{S}}_w$. With the Regularized LDA model learned, we can compute the discriminant feature vectors by $\mathbf{y} = \mathbf{W}^T\mathbf{x}$. Hence, for two faces represented by $\mathbf{x}_1$ and $\mathbf{x}_2$, the distance in this layer is given by $d = ||\mathbf{W}^T\mathbf{x}_1 - \mathbf{W}^T\mathbf{x}_2||$.

2.2 Skin Textures

The face skin is a textured surface consisting of repetitive texture units, which we call *textons*. Such a textured nature prefers a texture-based description. There are generally three families of texture models. A typical family of methods are based on filter bands, they apply FIR filters to the images. Merely relying on power spectrum, they lacks ability of describing local spatial relation. Later, MRF is brought to the domain[9][15], which is inherently suitable for modeling local spatial structures. However, the size of MRF model increases exponentially with the neighborhood size, which confines its capability in capturing large-scale patterns. Zhu et. al propose the the "FRAME"[22] to unify these two families. But the Gibbs sampling involved makes it too slow for practical applications. Since the concept *texton*[24][16][21] becomes popular in computer vision, it has achieved great success. The key rationale is to characterize the textures by texton-distribution.

We establish the skin texture representation based on texton-distribution. As illustrated figure 3, the procedure given as follows comprises three stages: Filtering, Dictionary Building, and Discriminant Learning.

Training Phase: Learning Models

a. Apply Gabor-filters (5 scales and 8 orientations) to extract a set of 40-dim response vectors for all pixels in skin region.

b. The response vectors over all pixels in skin regions in all training images are accumulated to form a vector pool.

c. Progressively build the texton-dictionary: continue randomly sampling from all response vectors, if the sampled is close to some cluster, then put it to the cluster and update the center*(texton)*, otherwise, create a new cluster centered at new sample. The process continues until no new textons are found and all cluster centers become stable.

d. For each image, categorize its skin pixels to the K-texton clusters, then get the K-bin histograms by counting vectors classified to each cluster. Normalize the histogram so that the sum of scalars in the bins equals 1. This histogram characterizes the texture of an image.

e. Regard these histograms as K-dim vectors, and apply Regularized LDA on them, to solve the transform matrix $\mathbf{W}_h$.

Testing Phase: Extracting features for a new image

a. Filter the skin region with Gabor filter bands to acquire response vectors.

b. Categorize the compressed vectors on all pixels to clusters and compute the K-bin histogram.

c. Transform the histogram to discriminant vector by $\mathbf{W}_h$.

Compared to conventional texton-based methods, our method has two improvements specially made for skin modeling. **First**, motivated by the observation that structure of skin textures in a high-resolution image is consisting of multiple levels, we obtain the texton description by the inherently-scalable Gabor wavelet. **Second**, because there are tremendous amount of pixels in a set of high resolution images, conventional clustering methods such as K-Means are infeasible. We develop an efficient scheme to build texton dictionary: instead of exhausting all samples, we progressively sample from the patch pool until the clusters are

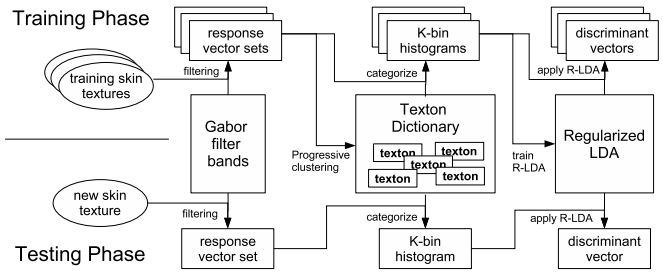

Figure 3. Illustration of Discriminant Multiscale Texture Feature

stable. Thereby, much less samples are required. **Third**, observing that the texton histogram encodes not only essential features but also variations cause by illumination change, we learn regularized LDA to extract the intrinsic information, and reduce the interference. The features extracted in this way is called *Discriminant Multiscale Texture Feature*.

2.3 Irregular Details

Irregular details refer to the subtle yet distinctive local marks on a face image, such as naevus. Face recognition can substantially benefit from the use of the irregular details owning to their two significant properties: 1) The special arrangement of irregular details constitute a strong distinction of a person; 2) The local details are insensitive to illumination change. Realizing that there exist a vast number of irregular details caused by temporary contamination, we restrict our model to only utilizing the local details satisfying the following conditions: **1) distinctiveness**: they have special local pattern and do not resemble other parts of the faces such as skin texture. **2) stability**: they should stably occur in nearly all images of a person, and can be repeatedly detected.

Recently, a trend of using local feature based description arises in the field of generic object recognition, which comes in three lines: 1) the approach matching two sets of local features[6]; 2) the approach using statistics of local features without correspondence[3][8]; 3) the statistical models formulating both the local features and their interrelations. Representative works include Constellation models[20][1], Graphical models[10], and Pictorial structure[17]. The third family is superior to the other two in that it utilizes the spatial relation of the local features.

Inspired by these works, we consider two aspects of the irregular details: local appearance and spatial configuration, as shown in fig.4. Assume that a face image has m local interest regions, we represent them by a size-variable set, denoted by $\mathcal{S} = \{(\mathbf{a}^{(l)}, \mathbf{x}^{(l)}, \theta^{(l)}, s^{(l)})\}_{l=1}^{m}$. Here, $\mathbf{a}^{(l)}$ describes the shape-free local appearance of the l-th interest region; $\mathbf{x}^{(l)}$ is the position of region center; $\theta^{(l)}$ is its principal orientation; and $s^{(l)}$ is the scale of the region. A person is represented by center values of these quantities, i.e. $\mathcal{M} = \{(\bar{\mathbf{a}}^{(l)}, \bar{\mathbf{x}}^{(l)}, \bar{\theta}^{(l)}, \bar{s}^{(l)})\}_{l=1}^{m}$. Firstly, we derive a probabilistic formulation modeling the local characteristics. With the assumption that they independently satisfy normal distributions, and that the correspondence have been established between a face $\mathcal{S}$ and a personal model $\mathcal{M}$, we have

$$\begin{aligned} p^{(lc)}(\mathcal{S}|\mathcal{M}) &= \prod_{l=1}^{m} p(\mathbf{a}^{(l)}|\bar{\mathbf{a}}^{(l)}; \mathbf{\Sigma}_a) p(\mathbf{x}^{(l)}|\bar{\mathbf{x}}^{(l)}; \mathbf{\Sigma}_x) \\ &\quad p(\theta^{(l)}|\bar{\theta}^{(l)}; \sigma_\theta) p(s^{(l)}|\bar{s}^{(l)}; \sigma_s); \end{aligned} \tag{4}$$

$$p(\mathbf{a}^{(l)}|\bar{\mathbf{a}}^{(l)}; \mathbf{\Sigma}_a) \propto \exp\left(-\frac{(\mathbf{a}^{(l)} - \bar{\mathbf{a}}^{(l)})^T \mathbf{\Sigma}_a^{-1} (\mathbf{a}^{(l)} - \bar{\mathbf{a}}^{(l)})}{2}\right),$$

$$p(\mathbf{x}^{(l)}|\bar{\mathbf{x}}^{(l)}; \mathbf{\Sigma}_x) \propto \exp\left(-\frac{(\mathbf{x}^{(l)} - \bar{\mathbf{x}}^{(l)})^T \mathbf{\Sigma}_x^{-1} (\mathbf{x}^{(l)} - \bar{\mathbf{x}}^{(l)})}{2}\right),$$

$$p(\theta^{(l)}|\bar{\theta}_l; \sigma_\theta) \propto \exp\left(-\frac{1}{2}\frac{(\theta^{(l)} - \bar{\theta}^{(l)})^2}{\sigma_\theta^2}\right),$$

$$p(s^{(l)}|\bar{s}^{(l)}; \sigma_s) \propto \exp\left(-\frac{1}{2}\frac{(s^{(l)} - \bar{s}^{(l)})^2}{\sigma_s^2}\right).$$

Moreover, the spatial configuration of the interest regions also play an important role in describing a face. We formulate the spatial configuration with an Elastic Graph, in which the pairwise distances are assumed to satisfy Gaussian distribution:

$$p^{(sc)}(\mathcal{S}|\mathcal{M}) \propto \prod_{l_1, l_2} \exp\left(\frac{-(||\mathbf{x}^{(l_1)} - \mathbf{x}^{(l_2)}|| - ||\bar{\mathbf{x}}^{(l_1)} - \bar{\mathbf{x}}^{(l_2)}||)^2}{2\sigma_{sc}^2}\right). \tag{5}$$

With the formulation above, the model for the k-th person can be learned by Maximum Likelihood(ML) incorporating both local characteristics and spatial configuration,

$$\mathcal{M}_k = \underset{\mathcal{M}}{\operatorname{argmax}} \prod_{j=1}^{n_k} p^{(lc)}(\mathcal{S}_{kj}|\mathcal{M}) p^{(sc)}(\mathcal{S}_{kj}|\mathcal{M}). \tag{6}$$

Taking logarithm of Eq.(6), we see that it is in essence an energy minimization problem with the energy given by

$$E_k = \sum_{j=1}^{n_k} E(\mathcal{S}_{kj}|\mathcal{M}_k), \tag{7}$$

$$E(\mathcal{S}_{kj}|\mathcal{M}_k) = \sum_{l=1}^{m} El^{(l)}(\mathcal{S}_{kj}|\mathcal{M}_k) + \sum_{l=1}^{m}\sum_{l=1}^{m} Es^{(l_1,l_2)}(\mathcal{S}_{kj}|\mathcal{M}_k). \tag{8}$$

Here El and Es respectively reflect the fidelity of local characteristics and spatial configurations, given by

$$\begin{aligned} El^{(l)}(\mathcal{S}|\mathcal{M}) &= (\mathbf{a}^{(l)} - \bar{\mathbf{a}}^{(l)})^T \mathbf{\Sigma}_a^{-1} (\mathbf{a}^{(l)} - \bar{\mathbf{a}}^{(l)}) \\ &+ (\mathbf{x}^{(l)} - \bar{\mathbf{x}}^{(l)})^T \mathbf{\Sigma}_x^{-1} (\mathbf{x}^{(l)} - \bar{\mathbf{x}}^{(l)}) \\ &+ \sigma_\theta^{-2} (\theta^{(l)} - \bar{\theta}^{(l)})^2 + \sigma_s^{-2} (s^{(l)} - \bar{s}^{(l)})^2, \end{aligned} \tag{9}$$

$$Es^{(l_1,l_2)}(\mathcal{S}|\mathcal{M}) = \sigma_{sc}^{-2}(||\mathbf{x}^{(l_1)} - \mathbf{x}^{(l_2)}|| - ||\bar{\mathbf{x}}^{(l_1)} - \bar{\mathbf{x}}^{(l_2)}||)^2. \quad (10)$$

Solving $\mathcal{M}_k$ is a convex quadratic programming problem guaranteeing global optima. Note that in our formulation, all personal models share the same covariances and variances, which is justified by two reasons: 1) The number of samples of each person is small, and is insufficient to give a reliable estimation of the covariances. 2) The variations are mainly caused by external condition changes. Due to the limited samples for each person, they may not capture all types of variations. Hence, it is proper to pool all regions in all samples to robustly evaluate the covariances and variances. The rationale is similar to that of LDA, where $\mathbf{S}_w$ is calculated by pooling the variations of all classes.

Under the formulation, we first need to select a set of interest regions satisfying the aforementioned requirements and give their descriptions. Mikolajczyk et. al have given a comparative study on state-of-the-art interest region detectors and local descriptors[12], and shown that Scale Invariant Feature Transform (SIFT)[6] outperformed the others in terms of repeatability. In this paper, we use the Hessian-Laplace detector (the detector used in SIFT) for interest region detection, which localizes the scale-space maxima of the DoG and is claimed to achieve scale-invariant detection[6]. The detection process would evaluates the position $\mathbf{x}$, scale s, and principal direction θ of every interest region. Then for each region, we employ SIFT to give a local description $\mathbf{a}$. It is a 128-bin histogram characterizing the local gradient distribution in 16 relative locations and 8 relative orientations. Such a description is scale- and rotation-invariant and highly distinctive. Based on SIFT, we devise the following scheme to select the valid interest regions:

1. Use SIFT to detect interest regions as candidates and give local descriptions on them.

2. For each person, perform agglomerative clustering to cluster the candidate regions detected on all his images:

a. Initialize by considering each region as a cluster;

b. Progressively combine two clusters with the condition that every pair of regions in a cluster should be similar and close enough, i.e. the differences of local appearances, positions, orientations and scales should be below some thresholds;

c. Repeat the process until no more combinations found.

3. Refine clusters by reassigning each region to the best cluster.

4. Filtering: retain the clusters meeting following constraints:

a. (Stability) the regions in the cluser should appear in no less than 90% of the samples of the person;

b. (Distinctiveness) the region are not similar to any skin textons, and its entropy should be higher than some threshold.

5. If a cluster contains more than one regions from the same image, only keep the one closest to the cluster center.

After the interest region is selected, we can learn the personal models for each person as follows:

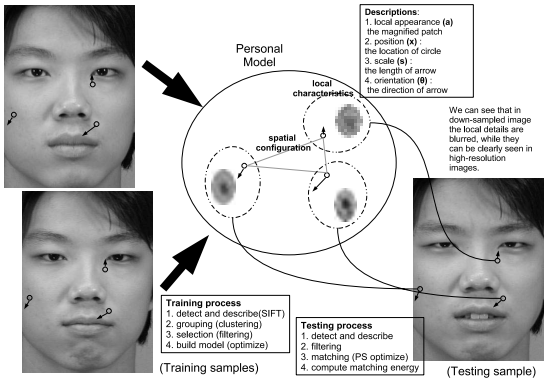

Figure 4. The SIFT-Activated Pictorial Structure.

1. For each person, set m to be the number of selected clusters. If $m > 0$, we continue the following steps; otherwise it indicates that the irregular models cannot be built, then we will not use irregular details for this person.

2. For l-th cluster, compute the means for local descriptions.

3. Compute the covariances and variances for local descriptions for all clusters of all persons.

4. Estimate $\mathbf{\Sigma}_a$, $\mathbf{\Sigma}_x$, σ_θ, σ_s by averaging all the computed covariances and variances on local appearances, positions, orientations, and scales respectively.

5. Estimate σ_{sc} by computing the variances of distances in each pair of regions and avaerging them.

6. With the covariances and variances estimated, we can learn the personal models $\mathcal{M}_1, \ldots, \mathcal{M}_C$ by minimizing E_k with convex quadratic programming.

To match a learned model $\mathcal{M}$ to a new image, that is to determine the positions, scales, and orientations for each model-region on the new image. It can be implemented by

$$\widehat{\mathcal{S}} = \underset{\mathcal{S}}{\operatorname{argmax}}\, E(\mathcal{S}|\mathcal{M}). \quad (11)$$

The formulation given in Eq.(6) is an extension of Pictorial Structure(PS)[17]. In [17], a dynamic programming method is introduced for matching a PS model to a new image, which can give an approximately global optimal solution by first pruning the elastic graph to a minimum spanning tree and then searching the optimal configurations by forward-backward tracing. Targeting our problem, we make two important improvements to the matching algorithm.

First, the algorithmic complexity is $O(nh^2)$, where h is the number of possible positions for each region. It is typical to impose a grid over the image as candidate positions, incurring large computational cost. To address the difficulty, we combine SIFT detector with PS model. Only the SIFT-detected candidates, instead of dense grid points, are activated in the searching process, thereby h is

drastically reduced and thus the efficiency is increased.

Second, it is possible that for some regions, there may be no correspondence detected in the input image. As to this issue, we preset a threshold El_{max}. For a region, if the matching energies between it and all detected regions are higher than El_{max}, it is considered to have no correspondences. In this case, we just set $El^{(l)} = El_{max}$. By imposing an upper bound on the local matching energies, the model is robustly accommodated to the situation without suitable correspondences. The process of matching a person to a new image is given below

1. For an input image, use SIFT to detect interest regions, and give local description.
2. Filter out the regions not meeting distinctiveness condition.
3. Apply the dynamic programming procedure[17] with our El_{max} upper-bound to establish region correspondence.
4. Use Eq.(8) to compute the matching energy. $(\frac{E(\mathcal{S}|\mathcal{M})}{m})^{\frac{1}{2}}$ is considered as the model-sample-distance.

Through the integration of SIFT and Pictorial Structure formulation, we develop a novel and effective paradigm to model the irregular details, called *SIFT-Activated Pictorial Structure (SAPS)*.

3 Adaptive Metric Fusion

As shown in figure 2, our framework consists of totally $L = 8$ channels in all the layers, which respectively evaluate distance metrics based on their special models. We develop an adaptive metric fusion algorithm to give final decisions by integrating all layers. Let $d_k^{(l)}(\mathbf{x})$ denote the distance between a sample $\mathbf{x}$ and the k-th person.[1] With a simplified assumption that the samples satisfy an isotropic normal distribution in the discriminant space, we get

$$\mathbf{x} \sim \mathcal{N}(\mathbf{m},\ \sigma^2\mathbf{I}) \Rightarrow d(\mathbf{x};\mathbf{m}) \sim \sqrt{\frac{2}{\pi\sigma^2}} \exp\left(-\frac{d^2}{2\sigma^2}\right). \tag{12}$$

Let $\mathcal{L}_l$ denote the l-th layer, $d_k^{(l)}(\mathbf{x})$ denote the distance between the sample $\mathbf{x}$ and the k-th person in $\mathcal{L}_l$. Suppose each layer serves as an "expert", and independently makes decision. The decisions from $\mathcal{L}_l$ are expressed by posteriori based on distances

$$p(k|d_k^{(l)};\mathcal{L}_l) = \frac{P(k)p(d_k^{(l)}|k;\mathcal{L}_l)}{\sum_{q=1}^{C} P(q)p(d_q^{(l)}|q;\mathcal{L}_l)}, \tag{13}$$

$$\text{here,} \quad p(d_k^{(l)}|k;\mathcal{L}_l) = \sqrt{\frac{2}{\pi\sigma_l^2}} \exp\left(-\frac{d_k^{(l)2}}{2\sigma_l^2}\right). \tag{14}$$

[1]For the first 3 layers, d is the distance between the sample vector and the center vector of the k-th person in the discriminant space; while for the 4-th layer, it is the sqrt root of the average matching energy.

We just assume all classes share equal priors: $P(k) = \frac{1}{C},\ k = 1,\ldots,C$, while σ_l is the standard deviation of samples in the l-th layer. We then introduce a concept called *confidence* to reflect the expert's certainty on its decision, denoted by $c(l)$. Hence, the combined decision is

$$p(k|\{\mathcal{L}\}_{l=1}^{L}) = \sum_{l=1}^{L} c(l) \cdot p(k|d_k^{(l)};\mathcal{L}_l). \tag{15}$$

A straightforward way to define the confidences is setting $c(l)$ to be the average correct rate obtained in $\mathcal{L}_l$. A separate evaluation set is used to give an objective assessment on the correct rates. Carefully examining the faces, we find that different faces possess different distinctions. For the faces with special eyes, we should emphasize the eye-channel; while for the faces with conspicuous naevus, the irregular-detail-layer should be heavily weighted. Motivated by the rationale, we derive the adaptive confidence, given by

$$c(l) = \frac{\log C - H(k|\mathcal{L}_l)}{\log C}, \tag{16}$$

here the $H(k|\mathcal{L}_l)$ is the entropy of classes conditioned on distances evaluated in $\mathcal{L}_l$, defined by

$$H(k|\mathcal{L}_l) = \sum_{k=1}^{C} p(k|d_k^{(l)};\mathcal{L}_l) \log p(k|d_k^{(l)};\mathcal{L}_l). \tag{17}$$

We see that when the distances to all classes are equal, the expert completely fail to make judgment; then $H(k|\mathcal{L}_l) = \log C$, i.e. $c(l) = 0$; while if the expert evaluates that $p(k|\mathcal{L}_l) = 1$, it definitely make sure that the sample belongs to the k-th class, then $H(k|\mathcal{L}_l) = 0$, i.e. $c(l) = 1$. By this way, emphasis is placed on the layers that can clearly distinguish between the classes. Furthermore, according to the Fano's inequality in information theory, Eq.(16) gives an upper bound on the average correct rate.

4 Experiments

To systematically investigate the layer models and the whole framework, we conduct experiments on two high resolution face databases. The first one is XM2VTS[11]. It consists of 1180 images of size 720×576 pixels, which are from 295 persons with each person having 4 samples captured in different sections. For each person, we take the first two sections for training, while other two sections for testing. Thus, we have 590 training samples, and 590 probe samples. The other database is collected by our lab, which comprises 1600 images from 200 persons. For convenience, we call this high resolution database *HRDB*. The images are of size 1024×768, and captured with obvious illumination and expression change. For each person, 4 images are used

	XM2VTS						HRDB					
Part	PCA	LDA	NDA	EFM	GDA	R-LDA	PCA	LDA	NDA	EFM	GDA	R-LDA
global	0.836	0.883	0.871	0.907	0.897	**0.942**	0.548	0.841	0.896	0.868	0.884	**0.935**
lefteye	0.527	0.556	0.353	0.564	0.614	**0.636**	0.561	0.843	0.703	0.870	0.854	**0.901**
righteye	0.515	0.553	0.393	0.554	0.534	**0.614**	0.573	0.836	0.688	0.856	0.863	**0.896**
nose	0.290	0.320	0.112	0.324	0.320	**0.369**	0.260	0.423	0.300	0.449	0.433	**0.493**
mouth	0.397	0.431	0.241	0.436	0.449	**0.485**	0.194	0.448	0.248	0.469	0.399	**0.499**

Table 1. **Correct rates on global appearance layer and facial organs layer.**

	XM2VTS					HRDB				
Combination	vote	dist.sum	prob.sum	FCMF	Ada.MF	vote	dist.sum	prob.sum	FCMF	Ada.MF
all organs	0.707	0.763	0.756	0.775	**0.841**	0.873	0.906	0.914	0.933	**0.959**
global+organs	0.863	0.878	0.892	0.910	**0.953**	0.919	0.930	0.935	0.945	**0.955**
skins(cheek+forehead)	0.380	0.547	0.612	0.620	**0.669**	0.351	0.560	0.521	0.538	**0.573**
organs+skins	0.780	0.859	0.864	0.883	**0.917**	0.889	0.940	0.954	0.963	**0.970**
global+organs+skins	0.908	0.951	0.933	0.959	**0.985**	0.936	0.946	0.951	0.955	**0.976**
global+irregu.	0.946	0.951	0.953	0.956	**0.963**	0.941	0.943	0.953	0.959	**0.962**
organs+skins+irregu.	0.797	0.890	0.905	0.922	**0.969**	0.904	0.904	0.943	0.960	**0.960**
all layers	0.929	0.978	0.978	0.986	**0.990**	0.959	0.944	0.954	0.964	**0.986**

Table 2. **Correct rates on combinations of layers. Five fusion schemes are compared. From left to right, they are majority voting, summing of distances, sum of probabilities of Eq.(13), confidence-fixed metric fusion, adaptive metric fusion.**

for training, while the other 4 are for testing. Note that in both databases, the training set and testing set are disjoint.

To reduce the effect of affine photometric transform, every face image is first normalized so that the mean intensity value is shifted to zero, and the standard deviation of pixel values is unified to 1. The 4 layers are respectively constructed by applying different masks to extract the target region as illustrated in fig.1. Note that the irregular-detail-layer is based on whole face region with the facial organs removed. The global appearance is down-sampled from the initial image by a factor of 4, while all other layers are in the original resolution. A two-level PCA hierarchy is applied to the global appearance for dimension reduction, while standard PCA is used for facial organs. In each PCA model, 95% of the variation energy is preserved.

Test Models of Global Appearance and Facial Organs. We first test the global appearance layer and the facial organs layer, where representative algorithms including PCA[14], LDA[18], Kernel LDA(GDA)[7][19], Nullspace LDA(NDA)[13], Enhanced Fisher Model(EFM)[4], and our regularized LDA(R-LDA) are compared. The parameters are determined by cross-validation, and the best results obtained are shown in table 1[2]. From the results we observe that: **(1)** Global appearance outperforms the layers based on individual organs by a large margin, which indicates that it contains most discriminant information. **(2)** As to the organs, eyes lead to the best performance. Compared to other parts, they convey more information. The accuracy of mouth-channel is not very good, just slightly better than nose, which is due to its significant variation across different images of a person. **(3)** The proposed Regularized LDA shows very good accuracy and robustness, and consistently surpasses other algorithms.

[2]In the table, Dict.size is the number of textons in dictionary, Hist.cr and Disc.cr are the correct rates acquired by direcly using χ^2-distance on original histograms and using L^2-distance on the discriminant vectors.

Database	Part	Dict.size	Hist.cr	Disc.cr
XM2VTS	cheek	11	0.234	**0.417**
	forehead	19	0.227	**0.349**
HRDB	cheek	13	0.060	**0.349**
	forehead	20	0.120	**0.385**

Table 3. **Results of Skin Texture Models**

Test Skin Texture Models. We use texton-based approach to model the skins. By employing the progressive sampling strategy, only about 3×10^5 patches are needed to build a stable texton dictionary. Considering that there are totally 4×10^7 patches available in the training pool, it really leads to a great computational saving. The results are shown in table 3. We can see from the results that: **(1)** The texton dictionary is small, containing about $10 \sim 20$ textons, which confirms the rationale that the skins are formed by repetition of a small set of units. **(2)** Directly comparing texton-histograms yields a very poor accuracy. It is due to two reasons: on one hand the skin appearance is seriously affected by illumination change; on the other hand the skins of different person are similar. **(3)** The accuracies achieved

merely based on skin are far beyond random-guess, which tells us that the skin texture indeed contains important discriminant information.

Test Irregular Details Models. For XM2VTS and HRDB, there are respectively 38% and 56% of the persons have stable irregular details. However, merely based on matching the irregular details, we can achieve very high accuracy (98.5% and 99.2%) on these subset of persons.

Test Layer Combinations. Finally, we employ different schemes to combine the channels to see their compound performances. From the results given in table 2, we have the following observations: **(1)** The combination of all facial organs results in a higher accuracy than any individual performance (given in table 1). This reflects the complementarities between different organs. **(2)** The combination of two types of skins also notably increases the performance. It is somewhat surprising that only the skin models can lead to accuracies up to about 60%. This sufficiently manifests the fact that the skin texture embeds discriminant information. Such a performance is also owning to the effectiveness of R-LDA and the fusion scheme. **(3)** By incorporating other layers with the global layer, the performance is remarkably enhanced. When all 4 layers are integrated, the performances attain nearly perfect: 99.0% and 98.6%. The results convincingly reveal the complementary nature of these layers. **(4)** The adaptive fusion scheme consistently exhibits better performance than other schemes, owning to its adaptability and flexibility.

5 Conclusion

The paper establishes a novel and effective paradigm for high resolution face recognition, incorporating new representations for modeling skin texture and irregular details and a flexible fusion scheme. The systematic experiments on the framework lead to interesting views into the complementarities of the layers, which hopefully inspires a rethinking in the face recognition methodologies.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region. The work was conducted at the Chinese University of Hong Kong.

References

[1] A.Holub and P.Perona. A discriminative framework for modelling object classes. In *Proc of CVPR'05*, 2005.

[2] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19(7):696–710, 1997.

[3] B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, 2000.

[4] C. Liu and H. Wechsler. Enhanced fisher linear discriminant models for face recognition. In *Proc. of CVPR'98*, 1998.

[5] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Imag. Proc.*, 11(4), 2002.

[6] D.G. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[7] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

[8] G.Csurka, C.R.Dance, L.Fan, J.Willamowski, and C.Bray. Visual categorization with bags of keypoints. In *Proc of ECCV'04*, 2004.

[9] H.Deng and D.A.Clausi. Gaussian mrf rotation-invariant features for image classification. *IEEE Trans'PAMI*, 26(7):951–955, 2004.

[10] J.Winn and N.Jojic. Locus: Learning object classes with unsupervised segmentation. In *Proc of ICCV'05*, 2005.

[11] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *2nd Conf. on AVBPA*, 1999.

[12] K.Mikolajczyk and C.Schmid. A performance evaluation of local descriptors. *IEEE Trans'PAMI*, 27(10):1615–1630, 2005.

[13] L. Chen, H. Liao, J. Lin, M. Ko, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem of lda. *Pattern Recognition*, 33(10), 2000.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuro.*, 3(1):71–86, 1991.

[15] M.Varma and A.Zisserman. Texture classification: Are filter banks necessary? In *Proc of CVPR'03*, 2003.

[16] M.Varma and A.Zisserman. A statistical approach to texture classification from single images. *IJCV*, 2005.

[17] P.F.Felzenszwalb and D.P.Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[18] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. PAMI*, 19(7):711–720, 1997.

[19] Q. Liu and R. Huang. Face recognition using kernel based fisher discriminant analysis. In *Proc. of FGR'02*, 2002.

[20] R.Fergus, P.Perona, and A.Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc of CVPR'03*, 2003.

[21] S. Lazebnik, C.Schmid, and J.Ponce. A sparse texture representation using local affine regions. *IEEE Trans. PAMI*, 27(8):1265–1278, 2005.

[22] S.C.Zhu, Y.Wu, and D.Mumford. Filters, random fields and maximum entropy(frame): Towards a unified theory for texture modeling. *IJCV*, 27(2):107–126, 1998.

[23] T.F.Cootes, G.J. Edwards, and C.J.Taylor. Active appearance models. *IEEE Trans. PAMI*, 23(6):681–685, 2001.

[24] T.Leung and J.Malic. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[25] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *Proc. of FGR'98*, 1998.

[26] X. Wang and X. Tang. Unified subspace analysis for face recognition. In *Proc. of ICCV'03*, 2003.

[27] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proc. of CVPR'04*, 2004.