Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Automatic 3D face recognition from depth and intensity Gabor features☆

Chenghua Xu[a], Stan Li[a], Tieniu Tan[a,*], Long Quan[b]

[a]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing 100080, PR China
[b]Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 29 June 2008
Received in revised form 14 November 2008
Accepted 5 January 2009

Keywords:
3D face recognition
Feature sampling
Depth Gabor images
Feature selection
Fusion

ABSTRACT

As is well known, traditional 2D face recognition based on optical (intensity or color) images faces many challenges, such as illumination, expression, and pose variation. In fact, the human face generates not only 2D texture information but also 3D shape information. In this paper, we investigate what contributions depth and intensity information makes to face recognition when expression and pose variations are taken into account, and we propose a novel system for combining depth and intensity information to improve face recognition systems. In our system, local features described by Gabor wavelets are extracted from depth and intensity images, which are obtained from 3D data after fine alignment. Then a novel hierarchical selecting scheme embedded in linear discriminant analysis (LDA) and AdaBoost learning is proposed to select the most effective and most robust features and to construct a strong classifier. Experiments are performed on the CASIA 3D face database and the FRGC V2.0 database, two data sets with complex variations, including expressions, poses and long time lapses between two scans. Experimental results demonstrate the promising performance of the proposed method. In our system, all processes are performed automatically, thus providing a prototype of automatic face recognition combining depth and intensity information.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Face recognition is one of the most active research areas in the study of pattern recognition and computer vision. Over the past several decades, most work has focused on two-dimensional images. Due to the complexity of face recognition, it is still difficult to develop a robust automatic face recognition system. The difficulties mainly include the complex variations from many aspects, such as poses, expressions, illumination, aging, and subordinates. Of these problems, pose variations and illumination variations commonly influence the accuracy of 2D face recognition systems. Many studies [1] have sought to resolve 2D face recognition. According to evaluations of commercially available and mature prototyped face recognition systems provided by face recognition vendor tests (FRVT) [2], the recognition results under the unconstrained condition are not satisfactory.

To develop a robust face recognition system, additional information has been considered. Two typical solutions are to use infrared images [3] and to use 3D images [4,7]. Infrared images are robust to changes in environmental lighting, but they are too sensitive to changes in environmental temperature. Thus, its use is still limited. Another solution is to utilize 3D information. With the development of 3D capturing equipment, it has become faster and easier to obtain 3D shape and 2D texture information to represent a real 3D face. Currently, most equipment based on active stereo vision is robust to illumination variations. Thus, the obtained 3D shape represents the actual information irrespective of lighting. Moreover, complete transformations between different 3D images can be computed in the 3D space. This efficiently removes the transformation out of the image plane, which is very difficult in 2D face recognition. Recently some research results have illustrated that 3D data have more advantages than traditional 2D data [4]. Using 3D data is considered to be a promising way to improve the robustness and accuracy of recognition systems.

Currently, 3D face recognition is attracting attention. Actually, face recognition using 3D information can solve some problems in 2D face recognition. Although there exist some difficulties in 3D face recognition, such as coping with expression variations, the inconvenience of information capture and large computational costs, these problems have been the focus of recent research.

☆This work is supported by the NSFC/RGC Joint Research Scheme N_HKUST602/05, the National Basic Research Program of China (Grant No. 2004CB318100), the National Natural Science Foundation of China (Grant No. 60736018, 60702024, 60723005), and the National Hi-Tech Research and Development Program of China (Grant No. 2006AA01Z193, 2007AA01Z162).

* Corresponding author. Tel.: +86 10 8262 1145; fax: +86 10 6255 1993.
E-mail address: tnt@nlpr.ia.ac.cn (T. Tan).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.01.001

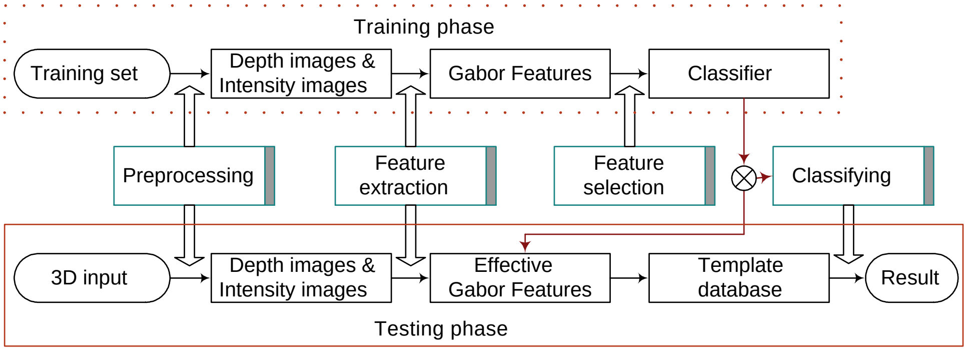

Fig. 1. Overview of our proposed framework.

1.1. Overview of approach

In this paper, we investigate how depth and intensity information contributes to face recognition when expression and pose variations are taken into account and we build a robust and accurate face system by selecting and fusing the most effective depth and intensity features. All processes included in our training and test phases are fully automated.

As in typical biometric systems, our system includes two phases: the training phase and the testing phase as illustrated in Fig. 1. The training process is performed once on a predefined training set to learn an effective classifier. It includes three important procedures: preprocessing, feature extraction and feature selection. Their descriptions are as follows:

(1) Preprocessing: By translating and rotating one input 3D image to align one reference 3D image, face poses and changed positions between the face and the equipment are normalized. This step is fully automatic. According to the aligned images, we obtain the normalized depth images and intensity images.
(2) Feature representation: Robust feature representation is very important to the whole system. It is expected that these features are invariant to rotation, scale, and illumination. The existing work usually uses raw depth and intensity features. In our systems, we extract local features to describe the individual information using Gabor filters with multiple scales and multiple orientations.
(3) Feature selection: As is commonly a problem, using multiple channels and multiple types of local features results in a much higher dimensional features space. A large number of local features can be produced with varying parameters in the position, scale, and orientation of the filters. Some of these features are effective and important for the classification task, whereas others may not be so. A uniform use of them or a manual selection has been proven ineffective. In this phase, we propose a systematic hierarchical framework for selecting the most effective features and constructing the classifier. This framework combines the linear discriminant analysis (LDA) and AdaBoost learning to fuse 2D and 3D Gabor features at the "feature" level.

During the testing phase, an input is classified according to the learned classifier in the training phase. By preprocessing the input, we obtain its normalized depth image and intensity image. This procedure is the same as that in the training phase. The feature extraction is a little different from that of the training phase. In this phase, we only compute the features that are present in the learned classifier, instead of all the features. Subsequently, the template database contains the corresponding features. In this way, we can largely reduce the spatial and computational resources. Finally, we classify the input characterized by extracted features into one category using the learned classifier.

1.2. Contributions of this paper

In this paper, we propose a new scheme to combine depth and intensity features to overcome problems due to expression and pose variations and build a robust and automatic face recognition system. The main contributions of this paper are the following:

- To reduce the large dimensions of Gabor features, we propose a hierarchical selecting scheme for selecting effective features and constructing the classifier. In this scheme, the combination of LDA and AdaBoost learning distinctly improves the learning efficiency.
- We propose an LDA-based algorithm to determine the optimal sub-sampling region in the Gabor images. This scheme has better performance than the existing PCA-based algorithm [26].
- We use AdaBoost learning to select the most effective Gabor features and boost them into a stronger classifier. This is one attempt to apply statistical learning to fuse 2D and 3D face recognition at the "feature" level.

The rest of this paper is organized as follows. Section 2 gives a concise review of existing works. Section 3 introduces the preprocessing procedure, which is very important to robust recognition. Section 4 describes the Gabor features of depth images and intensity images. Section 5 describes the proposed hierarchical selecting scheme for selecting effective features and constructing the classifier. Section 6 reports the experimental results and gives some comparisons with existing methods. Finally, Section 7 summarizes this paper.

2. Related work

Face recognition based on 3D information is not a new topic. Studies have been conducted since the end of the last century [31,32]. The representative methods use features extracted from the facial surface to characterize an individual. Lin et al. [11] extracted semi-local summation invariant features in a rectangular region surrounding the nose of a 3D facial depth map. Then, the similarity between them was computed to determine whether they belonged to the same person. Al-Osaimi et al. [12] integrated local and global geometrical cues in a single compact representation for 3D face recognition. The global cues provide geometrical coherence for the local cues resulting in the descriptiveness of the unified representation. Each local tensor field is integrated with every global field in a 2D histogram which is indexed by a local field in one dimension and a global field in the other dimension. Finally, PCA coefficients of the 2D histograms are concatenated into a single feature vector.

In another method, the human face can also be considered as a 3D surface, and the global difference between two surfaces provides the distinguishability between faces. Beumier et al. [33] constructed some central and lateral profiles to represent an individual and proposed two methods of surface matching and central/lateral profiles to compare two instances. They obtained the matching value by minimizing the distance of the profiles. In more recent work, Medioni et al. [34] built a complete and automatic system to perform face authentication by modelling 3D faces using stereo vision and analyzing the distance map between gallery and probe models. Although they obtained promising verification results, this method had large computational costs. Lu et al. [15] used the hybrid iterative closest point (ICP) algorithm to align the reference model with the scanned data and adopted registration errors to distinguish between different faces. Chang et al. [8] divided the whole facial surface into sub-regions. The rigid regions around the nose area were matched and combined to perform the recognition. Following the idea of dividing the sub-regions, Faltemier et al. [9] introduced a system for 3D face recognition based on the fusion of results from a group of regions that had been independently matched. Their experimental results demonstrated that using 28 small regions on the face led to the highest level of 3D face recognition. Russ et al. [10] presented an approach enabling a good alignment of 3D face point clouds while preserving face size information. They scaled the generic model when aligning the input face to the generic model. Kakadiaris et al. [16] used an annotated face model (AFM) to fit the changed face surface and then obtained the deformation image by the fitted model. A multi-stage alignment algorithm and the advanced wavelet analysis resulted in robust performance.

All of these studies illustrated the feasibility of 3D face recognition. However, perhaps due to the limitation of data, the above schemes only use the shape features of facial surfaces while ignoring the texture information.

Face recognition combining 3D shape and 2D intensity/color information is a developing topic. The combination of 2D and 3D information provides an opportunity to improve face recognition performance. Wang et al. [5] described facial feature points by using Gabor filter responses in a 2D domain and point signatures in a 3D domain. Then, classification was done by support vector machines with a decision directed acyclic graph. Tsalakanidou et al. [14] constructed embedded hidden Markov models (EHMM) classifiers based on depth and intensity information and then used fusion rules to combine the matching scores. In [7], Chang et al. evaluated the recognition scheme with different combinations of 2D and 3D information and showed that the combination of 2D and 3D information was most effective in characterizing an individual. Cook et al. [38] proposed a new method combining intensity and range images that was insensitive to expression variation based on log-Gabor templates. By breaking a single image into 75 semi-independent observations, the reliance of the algorithm upon any particular part of the face is relaxed, allowing robustness in the presence of occlusions, distortions and facial expressions. Bronstein et al. [17] represented a facial surface based on geometric invariants to isometric deformations and realized multi-modal recognition by integrating flattened textures and canonical images, which was robust to some expression variations. Mian et al. [18] handled the expression problem using a fusion scheme in which three kinds of methods, spherical face representation (SFR), scale-invariant feature transform (SIFT)-based matching and a modified ICP were combined to achieve the final result. Their results showed the potential of appearance-based methods to solve the expression problem in 3D face recognition. In the Face Recognition Grand Challenge Experiment Workshop [19], the work by Hüsken et al. [36] contributed to the fusion of 2D and 3D information. They performed hierarchical graph matching on 2D and 3D images and used the fusion rule of weighted sums to combine the matching scores. Maurer et al. [37] analyzed the performance of the Geometrix ActiveID Biometric Identity System, which fused shape and texture information. Their results showed that the combination of 2D and 3D information had promising results for face recognition.

In the above existing methods, some distinct problems remain unresolved:

- Some works [36,7] have been done to compare 2D face recognition and 3D face recognition. However, some details are ignored on how depth and intensity information contributes to recognizing faces with expression and pose variations.
- Only the fusion at the "decision" level is considered in the existing methods. It is well known that the fusion of the "feature" level based on statistical learning has obtained a promising result in 2D face recognition. It is very reasonable to use it in the fusion of 2D and 3D face recognition.

In this work, we attempt to address these remaining problems in face recognition.

3. Preprocessing

It is assumed in this paper that one face is described by one 3D point cloud captured by one 3D laser scanner (typically a Mintor VIVID 900/910 series sensor in our work). Each point cloud consists of thousands of points in the 3D space. These discrete points approximately describe the face surface. In the CASIA 3D face database, each point is described with 3D spatial coordinates and corresponding RGB color coordinates. In the FRGC V2.0 database, a separate image is used to describe the corresponding color information. In this section, we describe how the original 3D data are preprocessed. That is, we exactly register the data and then obtain the normalized depth and intensity images. This work prepares for the feature extraction and representation in the next section. Our preprocessing includes two main steps, registration of 3D face surfaces and acquisition of depth and intensity images.

3.1. Nose tip detection and registration

Different from 2D optical images, the nose is the most distinct feature in facial 3D data. We have proposed a robust method to localize the nose tip, which is described in detail in our previous work [21]. This algorithm utilizes two local surface properties, i.e., local surface features and local statistical features. It is fully automated, robust to noisy and incomplete input data, immune to rotation and translation and suitable to different resolutions. According to the experiments in the databases of the CASIA 3D face database and the FRGC V2.0 database, the correct detection rate is over 99%.

The registration process was described in our previous work [20], and here we only summarize it as follows. We select a front 3D image with neutral expression as the fixed model, and all the other 3D images are rotated and translated to align with it. Based on the detected nose tip, all the range images are translated and coarsely aligned together. Then, we use the ICP algorithm [22] to further register them. The facial image usually contains local deformation, such as expression variations. Strictly, the transformation among different scans is non-rigid. During registration, we only consider the region around the nose tip where the surface is robust to expression variations. Thus, the registration can avoid the unwanted influences of the mouth and the jaw.

3.2. Depth and intensity images

Depth and intensity images are obtained from registered 3D data. The 3D data that we are processing have the same subject size and

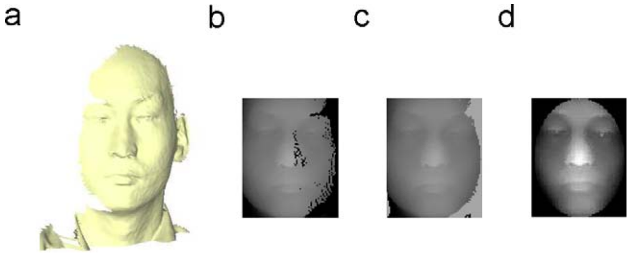

Fig. 2. Acquisition and preprocessing of the depth image: (a) registered 3D data; (b) normalized depth image; (c) depth image after removing noise and filling holes; (d) depth image after using the elliptical mask.

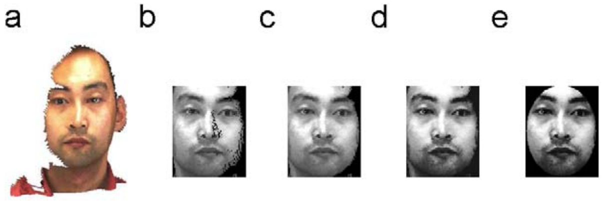

Fig. 3. Acquisition and preprocessing of the intensity image: (a) registered 3D data with texture; (b) normalized intensity image; (c) intensity image after removing noise and patching holes; (d) intensity image after histogram equalization; (e) intensity image after using the elliptical mask.

each 3D point has both 3D spatial coordinates and RGB color information. We use a 160 mm × 128 mm rectangle centered at the detected nose tip to crop the original 3D data. Then, the cropped region is converted to a depth image (see Fig. 2(b)) and an intensity image (see Fig. 3(b)) with 100 × 80 pixels. However, due to the quality of original 3D data, the depth and intensity images we use usually contain much noise, such as holes and outliers. We can obtain clear images by the following processes.

The preprocessing of depth images includes noise removal and hole filling. We use the following scheme to remove the outliers. For each pixel, the mean is computed for the 5 × 5 sub-window around it. If the difference between this pixel and the mean is larger than a preset threshold, this pixel is considered as an outlier and is removed. We fill the holes by linear interpolation of the neighboring pixels. The result is shown in Fig. 2(c).

The intensity images are processed using a similar method as on the depth images. Fig. 3(c) shows the processed image. As is well known, variation in the lighting strongly influences the presentation of the intensity images. In addition, histogram equalization is used to reduce the influence of the illumination variation (see Fig. 3(d)).

Human faces have elliptical shapes. Most works on face recognition use an elliptical template and ignore the regions outside the template that contain many uncertain factors. Here, we also use this kind of elliptical mask. The masked depth and intensity images are shown in Figs. 2(d) and (e), respectively.

Intensity images can also be obtained from 2D color images in the FRGC V2.0 database. Although this scheme can deal with holes and noise in intensity images, there are two difficulties. The first is that rectifying pose variations out of the image plane is difficult. The second is that detecting feature points in 2D images is computationally expensive. Thus, we do not use this scheme in our work.

4. Feature representation

In this paper, we use 2D Gabor filters of depth and intensity images to characterize an individual. The Gabor wavelets represent the properties of spatial localization, orientation selectivity, spatial frequency selectivity and quadrature phase relationships, and they have been experimentally verified to be a good approximation to the response of cortical neurons [35]. The representation of faces using Gabor wavelet has been successfully used in 2D face recognition [23]. This representation of an image describes the facial characteristics of both the spatial frequency and spatial relations. The 2D Gabor wavelets can be defined as follows [24]:

$$\Psi(z) = \frac{k_{\mu,\nu}^2}{\sigma^2} \exp\left(-\frac{k_{\mu,\nu}^2 z^2}{2\sigma^2}\right)\left[\exp(ik_{\mu,\nu}z) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{1}$$

where $z=(x,y)$, and μ and ν define the orientation and scale of the Gabor wavelets, respectively. $k_{\mu,\nu}$ is defined as follows:

$$k_{\mu,\nu} = k_\nu e^{i\Phi_\mu} \tag{2}$$

where $k_\nu = k_{max}/f^\nu$ and $\Phi_\mu = \pi\mu/8$. k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain.

The Gabor kernels in Eq. (1) are self-similar since they can be generated from the mother wavelet by scaling and rotation via the wave vector $k_{\mu,\nu}$. More scales or rotations can increase the dependencies of neighbor samples. We use Gabor kernels with five different scales, $\nu \in \{0,\dots,4\}$, and eight orientations, $\mu \in \{0,\dots,7\}$, with the parameters $\sigma = 2\pi$, $k_{max} = \pi/2$ and $f=\sqrt{2}$. The number of scales and orientations is selected to represent the facial characteristics of spatial locality and orientation selectivity.

The Gabor representation of an image, called the Gabor image, is the convolution of the image with the Gabor kernels as defined by Eq. (1). For each image pixel, we have two Gabor parts: the real part and the imaginary part. We transform these two parts into two kinds of Gabor features: Gabor magnitude features and Gabor phase features. In this paper, we use magnitude features to represent the facial Gabor features. Because the Gabor transformation strongly responds to edges [26], we first perform Gabor transformation and then use the facial mask. Figs. 4 and 5 show the Gabor magnitude features of depth and intensity under different scales and orientations.

Compared with the intensity Gabor images, the depth Gabor images are smoother. This is easy to understand since the value of the pixels in the depth image changes less than does the value in the intensity image. The smoother depth Gabor image can reduce the influence of noise, but it cannot describe the facial features in detail. This is why we perform face recognition by combining depth and intensity information. Further, we show that Gabor representation outperforms depth or intensity information in characterizing an individual in Section 6.

5. Feature and classifier learning

The dimensionality of the Gabor features is extremely large since Gabor filters with multiple scales and orientations are adopted. In this work, the size of the depth image or intensity image is 100 × 80 and Gabor representation with five scales and eight orientations is considered. Thus, the dimensionality of the Gabor feature vector will reach 640,000 (100 × 80 × 8 × 5 × 2). Though the elliptical mask is considered, the dimensionality is still 475,840 (5948 × 8 × 5 × 2). It is difficult to select the effective features using common learning schemes, such as AdaBoost learning [29], because a huge account of memory is required.

We develop a hierarchical selecting scheme for dimensionality reduction and for building a strong classifier as shown in Fig. 6. This scheme includes four stages. First, a new sub-sampling method is proposed to determine the optimal sampling positions in each Gabor image based on LDA [28]. Second, for the optimal sampling positions of each Gabor image, AdaBoost learning [29] is applied to select the effective features (30–100 features). Third, using AdaBoost

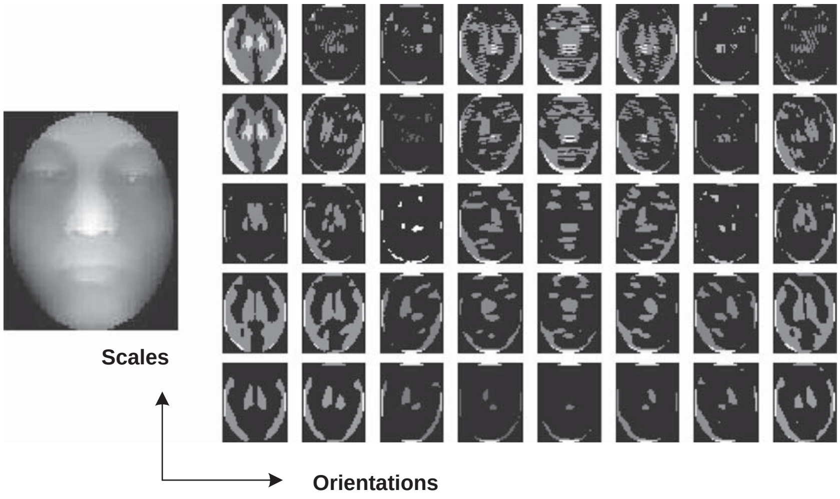

Fig. 4. The magnitudes of the Gabor representations of a depth image.

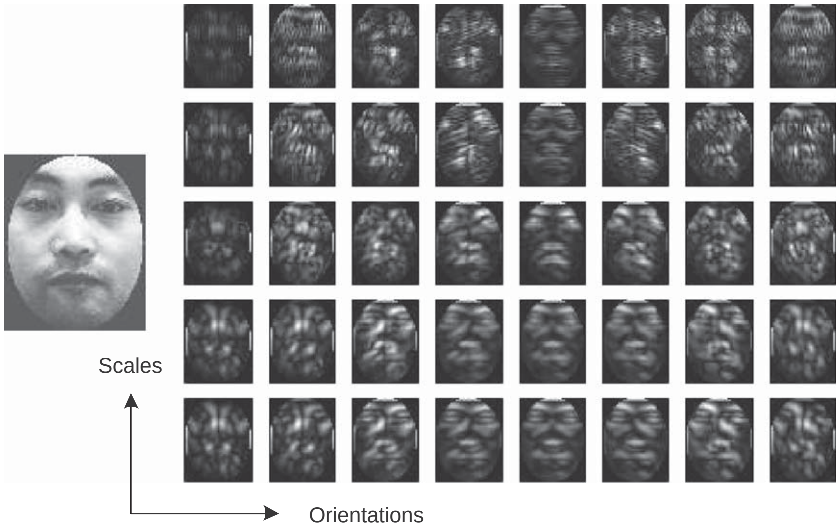

Fig. 5. The magnitudes of Gabor representations of an intensity image.

learning again, the dimensionality of the depth Gabor images is reduced to 206, and the dimensionality of the intensity Gabor images is reduced to 201. Finally, the AdaBoost algorithm with the cascade structure [29] is used to select the most effective features and construct the cascaded classifier. Three main techniques, i.e., optimal sub-sampling, feature learning and classifier learning, are described in the following three sections.

5.1. Optimal sub-sampling using LDA

The usual method for sub-sampling is to consider the sub-windows uniformly distributed in an image [25]. However, different facial regions differ in importance to face recognition, and regular sampling will lead to the loss of some important discriminant information. Liu et al. [26] proposed a scheme based on principle component analysis (PCA) for selecting the optimal sampling positions in the Gabor images. This scheme only considers the dimensionality reduction under the criterion of minimizing the global energy loss and ignores between-class and within-class discrimination. Here, we develop an optimal sub-sampling scheme based on LDA [28], which minimizes the within-class distance when maximizing the between-class distance. The detailed algorithm is described as follows.

Gabor images under different scales and orientations should have the different sub-sampling positions. Here, we construct them one by one. We first consider all the Gabor images under one scale and one orientation in the training set. One family of vectors with 5948 dimensions (after using the elliptical mask) are generated in the training set. The optimal discriminant vectors constructing the LDA subspace are computed by solving the following criterion in the standard LDA algorithm [28]:

$$W^* = \operatorname{argmax}(J(W)) = \frac{W^T S_B W}{W^T S_W W} \tag{3}$$

where S_B and S_W are the between-class and within-class scatter matrices, respectively.

Following the solution in [28], we obtain N optimal discrimination vectors using the training set in the CASIA 3D face database

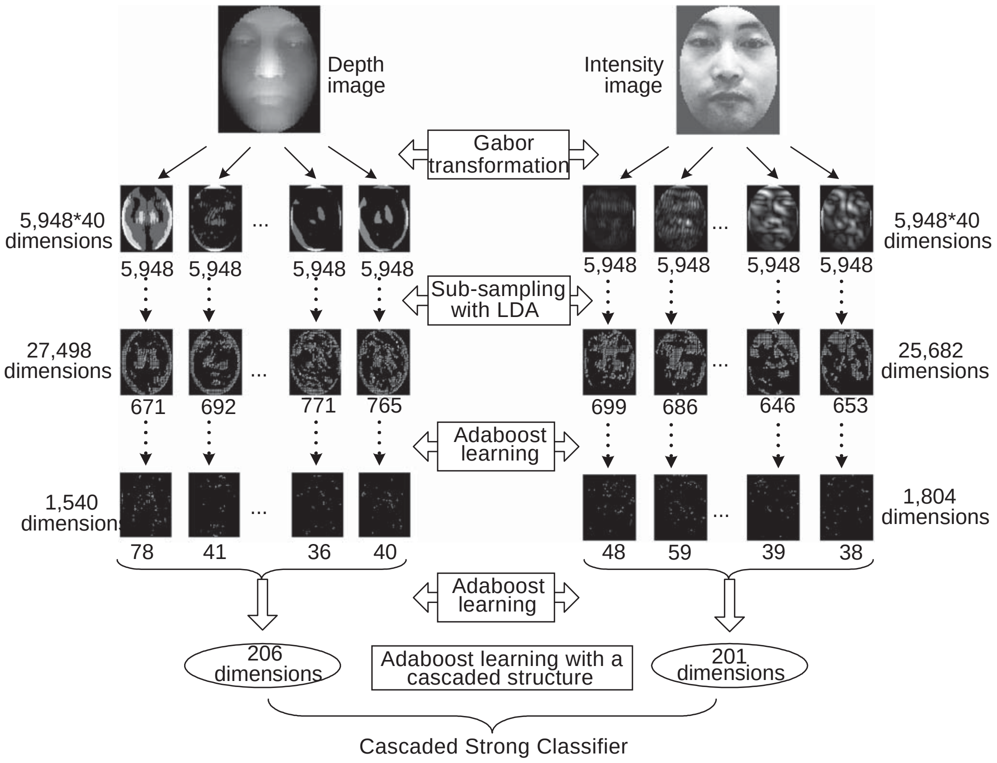

Fig. 6. Feature selection and classifier construction.

(see Section 6.1). In our training set, there are 23 subjects. The effective number of discrimination vectors is equal to the number of subjects minus one, i.e., $N = 22$. The dimensionality of these vectors is equal to the number of pixels of images with mask, i.e., 5948 in this work.

These discrimination vectors were used to generate the LDA subspace in previous studies [28]. Here, we use them to determine the optimal sub-sampling position. One summation vector, V, is generated by summing the absolute values of all of these vectors, that is,

$$V = \left(\sum_{k=1}^{k=N} |v_{k1}|, \sum_{k=1}^{k=N} |v_{k2}|, \ldots, \sum_{k=1}^{k=N} |v_{km}| \right) \tag{4}$$

where N is the number of vectors and m is the dimensionality of the vectors. The magnitude of V at a particular position represents the corresponding variations among the training set, which also reflects the corresponding importance in distinguishing the faces.

A preset percentage of the points with the largest magnitudes is considered as the optimal sub-sampling position. According to [26], the recognition accuracy is reasonable when the preset percentage is 25%, we use this threshold. Fig. 7(b) shows the optimal sub-sampling positions at this percentage.

The selected positions are usually very close and a sampling scheme is further used to reduce the redundancy. Here, we use an $n \times n$ sampling scheme, i.e., in an $n \times n$ sub-window, only an optimal sub-sampling position is reserved. Also according to [26], the smaller the sub-window, the better the recognition performance. Here, we adopt $n = 2$, and the final template is shown in Fig. 7(c).

We also calculate the optimal sub-sampling position using PCA [25] (see Fig. 7(a)). By comparing Fig. 7(a) with Fig. 7(b), we can see that the latter contains fewer optimal sub-sampling positions around the mouth than the former contains. This is very reasonable since the mouth is prone to expression variations, resulting in an increase of within-class distance.

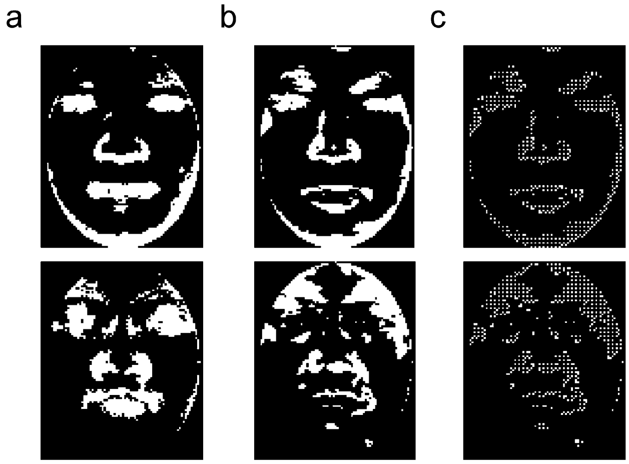

Fig. 7. Optimal sub-sampling positions of depth (top row) and intensity (bottom row) Gabor images under one scale and orientation: (a) from PCA; (b) from LDA; (c) 2×2 sampling after LDA. The white points are the optimal sub-sampling positions.

By observing the distribution of the optimal sub-sampling positions in Fig. 7, we can also find the differences between the depth Gabor images and the intensity Gabor images. The former (the top row) contains more sampling points around the edge of the elliptical mask than the latter (the bottom row). This shows that the margin data from depth images are important to distinguishing faces. This is easy to understand since different people have different facial sizes and facial surfaces separate largely in this region. However, in the corresponding region of the intensity images, the discrimination is weak.

For the Gabor images under the other scales and orientations, the corresponding templates are computed according to the above method. The dimensionality of each Gabor image is reduced to between 600 and 800.

5.2. Feature learning

Using the templates in the last section, the dimensionality of each Gabor image is still large. In this section, we introduce the method to further select the most effective features using AdaBoost learning [29]. AdaBoost learning can efficiently solve three fundamental problems [29]: (1) selecting effective features from redundant feature space, (2) constructing weak classifiers using selected features and (3) building a strong classifier. Here, we apply it to select the most effective features.

AdaBoost learning essentially works for a two-class classification problem. While face recognition is a multi-class problem, we convert it into one of the two classes using the representation of intra-personal vs. extra-personal classes, following [30]. Intra-personal examples are obtained by using differences in images of the same person, whereas extra-personal examples are obtained by using differences in images of different people. In the training set of the CASIA 3D face database, there are 23 subjects and 33 instances of each subject, and 12,144 intra-personal examples and 275,517 extra-personal examples are generated.

The AdaBoost learning algorithm [29] is implemented in the training set of the CASIA 3D face database. Here it is used in two sequential steps. First, from the depth Gabor images under each scale and orientation, the effective features in this scale and orientation are selected; from the intensity Gabor images under each scale and orientation, the effective features in this scale and orientation are also selected. Second, from selected features in depth Gabor images under all scales and orientations, the effective depth Gabor features are further selected; in the same way, the effective intensity Gabor features are learned from the intensity Gabor images. The selected features in the second step are more important than these in the first step in characterizing an individual. After these two steps, 206 depth features and 201 intensity Gabor features are reserved.

5.3. Classifier learning

After effective depth Gabor features and intensity Gabor features are selected, the cascaded AdaBoost learning procedure is used to create a strong classifier with a cascading structure. In this procedure, the input is intra-personal and extra-personal examples with 407 dimensions (206 selected depth and 201 intensity Gabor features), and the output is a cascaded strong classifier with 18 layers and 275 features.

During recognition stage, the feature vector of one probe sample is generated by extracting the corresponding features shown in the final cascaded classifier, and its difference with each gallery example forms the difference vector, x. For each vector, x, the i th layer of the cascaded classifier returns the similarity measure, S_i. The larger this similarity value, the more this sample belongs to the intra-personal space. If $S_i < 0$, the i th layer rejects the sample. Using the similarities from the multiple layers, we can obtain its total similarity:

$$S = \sum_{i=1}^{L} S_i \tag{5}$$

where L is the number of layers and S_i is the similarity value from the i th layer. Thus, we can obtain the sample's similarity with each gallery example. Then, the nearest neighbor scheme is used to decide which class the test sample belongs to.

6. Experiments

We validate our proposed scheme and compare with existing methods on the CASIA 3D face database and the FRGC V2.0 database.

6.1. Databases

We use two 3D face databases to test our proposed algorithm: one is the CASIA 3D face database; the other is FRGC V2.0 database [13]. The former was collected in our lab during August and September 2004 using a non-contact 3D digitizer, Minolta VIVID 910, working on Fast Mode. This database contains 123 subjects, with each subject having 33 images. The total number of range images is 4059. During the data collection, we considered not only the separate variations in expressions, poses and illumination, but also combined variations in expressions with the lighting from the right side and poses with a smiling expression. Some examples are shown in Figs. 8 and 9. This database contains complex variations that are challenging to any algorithm.

The database of 4059 images is divided into three subsets, that is, the training set, the gallery set and the probe set. The training set contains 759 images, corresponding to the last 23 of the 123 subjects, 33 images for each subject. The gallery set contains 100 images from the first images of the other 100 subjects (under the condition of front view, office lighting, and neutral expression). The other 3200 images from the above 100 subjects are used as the probe set.

The probe set is further divided into seven subsets:

- IV (400 images): illumination variations, including top, bottom, left and right lighting.
- EV (500 images): expression variations, including smile, laugh, anger, surprise and eyes closed.
- EVI (500 images): expression variations under the lighting from the right side.
- PVS (700 images): small pose variations, including views of front, left/right 20–30°, up/down 20–30° and tilt left/right 20–30°.
- PVL (200 images): large pose variations, including views of left/right 50–60°.
- PVSS (700 images): small pose variations with smiling.
- PVSL (200 images): large pose variations with smiling.

The second database is the FRGC V2.0 database [13]. There were three sessions: spring 2003, fall 2003 and spring 2004 based on acquisition time. In each session, one range image and one color image were recorded by a Minolta VIVID 900/910 series sensor working in the fine mode every one or two weeks. Thus, the final database contains 4950 records (one record contains one range image and its corresponding color image) from 557 people. The FRGC explicitly specifies that spring 2003 (943 scans, known as the FRGC V1.0 database) be used for training and the remaining two sets (4007 scans) be used for validation.

These two databases represent the most challenging data sets currently available. In this paper, they are used in different ways: the first is for evaluating the recognition performance and

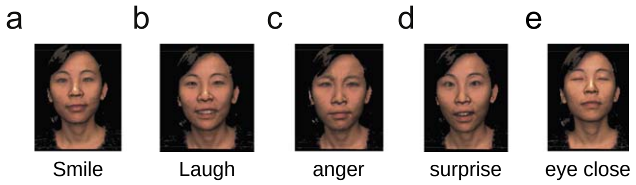

Fig. 8. Expression variations in our 3D face database.

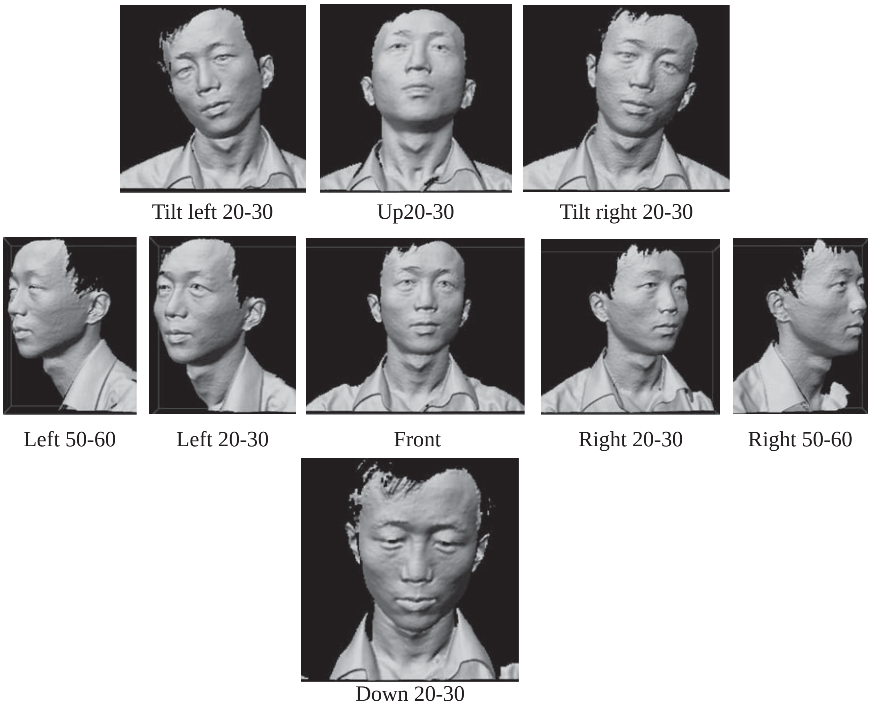

Fig. 9. Pose variations in our 3D face database.

comparing the different schemes, and the second is for evaluating the verification performance and comparing the different methods in recent publications. According to the method in Section 3, we preprocess all the data in the CASIA 3D face database and the FRGC V2.0 database and obtain the normalized depth images and intensity images. From these normalized images, we extract different features to represent an individual.

Table 1
Rank-one recognition accuracy in the CASIA 3D face database (100 subjects) using PCA (%).

Datasets	IV	EV	EVI	PVS	PVL	PVSS	PVSL
Depth	97.0	72.2	74.0	90.7	50.0	81.0	46.5
Intensity	96.0	81.6	84.2	69.9	49.5	68.1	48.5
Depth Gabor	98.3	74.4	75.8	91.4	51.5	82.4	49.0
Intensity Gabor	96.5	85.4	91.2	75.3	65.5	77.6	61.5

6.2. Depth vs. intensity

Depth information and intensity information are commonly used in the previous work. Chang et al. [7] compared their importance in characterizing an individual and concluded that the recognition performance was similar when using a single intensity image or a single depth image. They drew this conclusion under the limited condition of the front view, the smile expression and simple lighting changes. Moreover, the intensity images and the depth images are obtained using different sensors. Here, we test the recognition performance in the CASIA 3D face database and draw a different conclusion.

In our first experiment, raw depth images and raw intensity images are used to characterize the subjects, respectively. The PCA algorithm [27] is used for reducing dimensionality since it is widely applied in 2D and 3D face recognition. The similarity measure is Euclidean distance. The training set described in the last section is used to create the face subspace for recognition. The gallery set and the probe sets are defined in the above. The rank-one recognition accuracy in different probe subsets is listed in the top two rows in Table 1.

In our second experiment, we use the Gabor features of depth images and intensity images to describe an individual. The classifier is constructed by the same method as above and the rank-one recognition is shown in the bottom two rows in Table 1.

From Table 1, we can draw the following conclusions:

- Depth and intensity Gabor features outperform raw depth and intensity information, respectively. On each probe set, depth Gabor features have higher recognition accuracy than raw depth information; intensity Gabor features also have higher recognition accuracy than raw intensity information. This is the reason that we adopt the Gabor features.
- Under expression variations, raw intensity information is more robust than raw depth information; Gabor features of intensity images are also more robust than Gabor features of depth images. This is illustrated by the recognition accuracy in EV and EVI prove sets.
- Under small pose variations (PVS and PVSS probe sets), raw depth information and Gabor features of depth images are more robust than raw intensity information and Gabor features of intensity images, respectively.
- Large pose variations (left/right 50–60°, PVL and PVSL probe sets) degrade the recognition accuracy to a large extent. Large poses not only result in the serious shortage of some profile information, but also influence the accuracy of registration.
- Because the illumination variation is small in this database, it only slightly worsens the performance. Under the single variation of illumination (IV subset), depth information is a little better than intensity information.

Conclusions (2) and (3) also show that the complementarity of depth information and intensity information is useful to improve the recognition performance. It also gives a reasonable explanation for the conclusion in [7,14].

Table 2
Rank-one recognition accuracy in the CASIA 3D face database (100 subjects) using fusion at the "decision" level (%).

Datasets	IV	EV	EVI	PVS	PVL	PVSS	PVSL
Depth + intensity	97.0	84.4	88.9	85.2	65.0	82.9	60.0
Gabor features	97.8	86.4	90.4	89.0	70.5	85.6	64.5

Table 3
Rank-one recognition accuracy in the CASIA 3D face database (100 subjects) using fusion at the "feature" level (%).

Datasets	IV	EV	EVI	PVS	PVL	PVSS	PVSL
Depth + intensity	97.0	84.6	89.0	85.6	86.0	82.9	73.5
Gabor features	98.3	90.0	93.3	91.0	91.0	87.9	79.0

6.3. Fusion vs. learning

There are two ways to combine the features with different properties to describe an individual. They may happen at the "decision" level and the "feature" level. Here, we compare their recognition performance when depth and intensity Gabor features are combined. These experiments are performed on the CASIA 3D face database.

For the fusion at the "decision" level, multiple single classifiers are constructed using different kinds of features and then the obtained scores are combined using fusion rules. Using this framework, Chang et al. [7] fuse the depth and intensity information and get promising results. In one of our experiments, we repeat their algorithm and use the PCA to build two single classifiers based on depth and intensity features, respectively. The training set described in Section 6.1 is used to obtain the PCA subspace. The weighted sum rule, following [7], is used to fuse the scores from different classifiers. Table 2 shows the rank-one recognition accuracy in the different probe sets. In the first row, depth information and intensity information are combined to describe the individual. In the second row, depth and intensity Gabor features are combined. We also used other fusion rules, such as max, min, product and mean rules, but they had worse performance than the weighted sum rule [7].

For the fusion at the "feature" level, the features with different properties are combined to construct one classifier. The proposed method in this work selects the effective features using one hierarchical learning framework and realizes the combination at the "feature" level. In one experiment, we select effective features from the raw depth and intensity images and build the strong classifier directly using AdaBoost learning. The rank-one recognition accuracy is shown in the first row of Table 3. The other experiment works on Gabor features. The proposed scheme (see Fig. 6) is used to select the effective Gabor features and construct the strong classifier. The rank-one recognition accuracy is shown in the second row of Table 3. Fig. 10 shows the cumulative match score (CMS) curves in terms of two single classifiers based on Gabor features, fusion at the "decision" level and fusion at the "feature" level. The training set, test set and probe set are the same as in Section 6.1 in both experiments.

From the results in Fig. 10 and Tables 2 and 3, we can clearly see that AdaBoost learning is slightly better than the fusion scheme in terms of recognition accuracy of both raw data and Gabor representation. Especially under the large pose variations (PVL and PVSL probe subsets), fusion based on AdaBoost learning has a distinct improvement. This shows that AdaBoost learning utilizes the more distinguishing features than does the fusion at the "decision" level.

Comparing the results in Table 2 with those in Table 1, we can see that the fusion at the "decision" level does not always increase the recognition accuracy and that AdaBoost learning always outperforms the single classifier using raw and Gabor features, respectively. The fusion at the "feature" level based on AdaBoost learning is more robust than the fusion at the "decision" level.

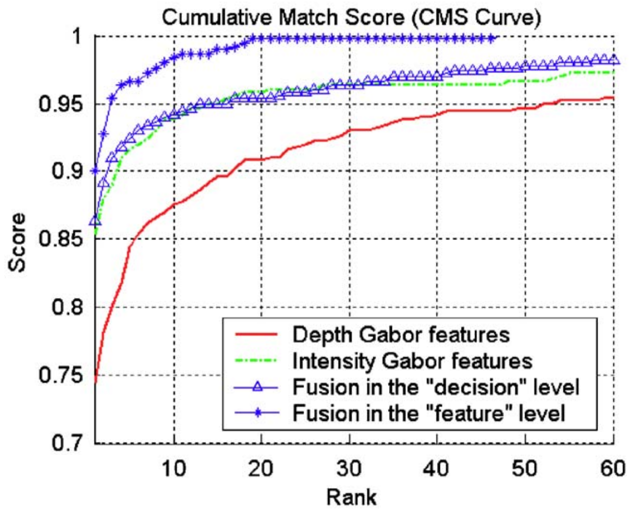

Fig. 10. CMS curves in the EV probe subset.

Table 4
Verification rate at 0.1% FAR based on ROC III of the FRGC V2.0 database.

Methods	Verification rate at 0.1% FAR (%)
Kakadiaris et al. [16]	97.0
Faltemier et al. [9]	94.9
Our method	95.3

6.4. Comparison with other algorithms

To demonstrate the performance of our algorithm, we perform comparative experiments that have been used in recent publications based on the FRGC V2.0 database [13]. In these experiments, we retrain our classifier by our method in Section 5 using the combined data of the training set in the 3D face database and the FRGC V2.0 database. From the following experiments, we can see that the accuracy of our algorithm is similar to that in the recent publications. Moreover, our algorithm is very efficient.

Our first experiment is "Experiment 3" described in the FRGC program [13]. We give the verification performances based on the protocol of ROC III. In this experiment, the gallery images come from one semester and the probe entries come from the following semester. This ensures that the time sequence between gallery and probe is maintained. Our algorithm achieves a verification rate of 95.3% at a false acceptance rate (FAR) of 0.1% as shown in Table 4.

In Table 4, we can see that the performance of our algorithm is slightly better than that in [9], but not better than that in [16]. Kakadiaris et al. [16] presented an algorithm for 3D face recognition that used an AFM to create a unique representation of a human face. Their algorithm fit the AFM to the face data after alignment and then the deformation data were processed with two different wavelet transforms (Pyramid and Haar) to extract a signature of the participant. Recognition was performed by fusing the individual scores from the Pyramid and Haar wavelets. They reported the best performance of 97.0% verification at a 0.1% FAR. They used a complex approach that required an AFM, which took about 15 seconds from the raw scanner data to the final features on a typical PC (3.0 GHz P4, 1 GByte RAM). The result of our algorithm is similar to theirs. However, relative to Kakadiaris' algorithm, we use a simpler method,

Table 5
Verification rate at 0.1% FAR using the FRGC v2.0 database.

Methods	Neutral vs. All (%)
Maurer et al. [37]	95.8
Hüsken et al. [36]	97.3
Mian et al. [18]	99.3
Our method	97.5

which only takes less than 5 seconds in our PC with 3.0 GHz CPU and 1 GByte RAM.

Our fusion method of 2D and 3D information is at the "feature" level. Here, we compare our method to other methods using multimodal information [37,36,18]. In this experiment, we separate the "neutral" set from the validation data of the FRGC V2.0 database based on the subject's expression. We calculate the verification rate at 0.1% FAR in the model of "neutral vs. all". The result and the stated performance in [37,36,18] are listed in Table 5.

From Table 5, we can see that our method outperforms the methods in [37,36]. Mian et al. [18] developed a complex method. They used 3D SFR and SIFT to create a rejection classifier to reduce the overall processing time after detecting the nose, performing pose correction and normalization in both 2D and 3D. They segmented the 3D images into two regions (nose and eye/forehead) and matched them independently by the modified ICP method. They reported verification of 99.3% at 0.1% FAR based on a neutral gallery and all images. The main reason that their scheme outperforms our method is that they use two regions (nose and eye/forehead) to independently match the face while we only use the nose region to register the 3D data. Their rejection classifier and region matching embedded in the recognition process result in a large computing cost.

In our fusion scheme, we use AdaBoost learning to select the most effective features from the depth and intensity Gabor images and boost them into a stronger classifier. Most of the selected features come from the regions insensitive to expression variations as shown in Fig. 6. Thus, we have superior performance in the above experiments. We also see that our registration method is simple relative to other studies [16,18]. In our future work, we will consider a better preprocessing method to improve the performance of the full whole system.

7. Conclusions

This paper uses depth and intensity Gabor images to construct a robust classifier for face recognition. Since the dimensionality of Gabor features of depth and intensity images is extremely high, we propose a novel hierarchical selecting scheme embedded with LDA and AdaBoost learning for dimensionality reduction and building the effective classifier. By analyzing our experimental results and comparing the existing methods in 3D face databases with the complex variations, we illustrate the promising performance of the proposed scheme and draw the following significant conclusions:

- Intensity information is more robust than depth information under expression variations; depth information is more robust than intensity information under pose variations. Thus, their combination is helpful in improving recognition performance.
- Gabor features from depth and intensity images outperform the raw depth and intensity information, respectively.
- The fusion at the "feature" level (based on AdaBoost learning) outperforms the fusion at the "decision" level when combining the features with different properties. In particular, AdaBoost learning largely improves the verification performance.

The fully automatic implementation in this work also provides a promising way to build a robust recognition system integrating depth and intensity information.

References

[1] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys (CSUR) Archive 35 (4) (2003) 399–458.
[2] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, J.M. Bone, FRVT 2002: Evaluation Report, Technical Report NISTIR 6965, March 2003 ⟨http://www.frvt.org/FRVT2002/documents.htm⟩.
[3] S.Z. Li, R. Chu, S. Liao, L. Zhang, Illumination invariant face recognition using near-infrared images, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4) (2007) 627–639.
[4] K. Bowyer, K. Chang, P. Flynn, A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition, Computer Vision and Image Understanding 101 (1) (2006) 1–15.
[5] Y. Wang, C. Chua, Y. Ho, Facial feature detection and face recognition from 2D and 3D images, Pattern Recognition Letters 23 (2002) 1191–1202.
[7] K.I. Chang, K.W. Bowyer, P.J. Flynn, An evaluation of multi-model 2D+3D biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (4) (2005) 619–624.
[8] K.I. Chang, K.W. Bowyer, P.J. Flynn, Multiple nose region matching for 3D face recognition under varying facial expression, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1695–1700.
[9] T.C. Faltemier, K.W. Bowyer, P.J. Flynn, A region ensemble for 3-D face recognition, IEEE Transactions on Information Forensics and Security 3 (1) (2008) 62–73.
[10] T. Russ, C. Boehnen, T. Peters, 3D face recognition using 3D alignment for PCA, in: Proceedings of the CVPR'06, 2006, pp. 1391–1398.
[11] W. Lin, K. Wong, N. Boston, Y. Hu, Fusion of summation invariants in 3D human face recognition, in: Proceedings of the CVPR'06, 2006, pp. 1369–1376.
[12] F.R. Al-Osaimi, M. Bennamouna, A. Miana, Integration of local and global geometrical cues for 3D face recognition, Pattern Recognition 41 (3) (2008) 1030–1040.
[13] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, et al., Overview of the face recognition grand challenge, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 947–954.
[14] F. Tsalakanidou, S. Malassiotis, M.G. Strintzis, Face localization and authentication using color and depth images, IEEE Transactions on Image Processing 14 (2) (2005) 152–168.
[15] X. Lu, A.K. Jain, D. Colbry, Matching 2.5D face scans to 3D models, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 31–43.
[16] I.A. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, Y. Lu, N. Karampatziakis, T. Theoharis, 3D face recognition in the presence of facial expressions: an annotated deformable model approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4) (2007) 640–649.
[17] A.M. Bronstein, M.M. Bronstein, R. Kimmel, Expression-invariant representations of faces, IEEE Transactions on Image Processing 16 (1) (2007) 188–197.
[18] A.S. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2D–3D hybrid approach to automatic face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (11) (2007) 1927–1943.
[19] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, et al., Overview of the face recognition grand challenge, in: Proceedings of the CVPR, 2005, pp. 947–954.
[20] C. Xu, Y. Wang, T. Tan, L. Quan, 3D face recognition based on G-H shape variation, Lecture Notes in Computer Science, vol. 3338, Springer, Berlin, 2004, pp. 233–243.
[21] C. Xu, Y. Wang, T. Tan, L. Quan, A robust method for detecting nose on 3D point cloud, Pattern Recognition Letters 27 (13) (2006) 1487–1497.
[22] P.J. Besl, N.D. Mckay, A method for registration of 3-D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 239–256.
[23] L. Wiskott, J. Fellous, N. Kruger, C.V. Malsburg, Face recognition by elastic bunch graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 775–779.
[24] T.S. Lee, Image representation using 2D Gabor wavelets, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (10) (1996) 959–971.
[25] C.J. Liu, H. Wechsler, Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition, IEEE Transactions on Image Processing 11 (4) (2002) 467–476.
[26] D.H. Liu, K.M. Lam, L.S. Shen, Optimal sampling of Gabor features for face recognition, Pattern Recognition Letters 25 (2004) 267–276.
[27] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[28] P.N. Belhumur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[29] P. Viola, M. Jones, Robust real-time object detection, in: Proceedings of the 2nd International Workshop on Statistical Computational Theories of Vision, 2001.
[30] B. Moghaddam, A. Pentland, Beyond Euclidean eigenspaces: Bayesian matching for vision recognition, in: Face Recognition: From Theories to Applications, ISBN 3-540-64410-5, 1998, p. 921.
[31] G.G. Gordon, Face recognition based on depth and curvature features, in: Proceedings of the CVPR'92, 1992, pp. 108–110.
[32] Y. Yacoob, L.S. Davis, Labeling of human face components from range data, in: CVGIP: Image Understanding, vol. 60(2), 1994, pp. 168–178.

[33] C. Beumier, M. Acheroy, Automatic 3D face authentication, Image and Vision Computing 18 (4) (2000) 315–321.

[34] G. Medioni, R. Waupotitsch, Face modeling and recognition in 3-D, in: Proceedings of the AMFG'03, 2003, pp. 232–233.

[35] J. Jones, L. Palmer, An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex, Journal of Neurophysiology (1987) 1233–1258.

[36] M. Hüsken, M. Brauckmann, S. Gehlen, C. Malsburg, Strategies and benefits of fusion of 2D and 3D face recognition, in: IEEE Workshop on Face Recognition Grand Challenge Experiments, 2005, pp. 174–181.

[37] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, G. Medioni, Performance of geometrix *ActiveID*^{TM} 3D face recognition engine on the FRGC data, in: IEEE Workshop on Face Recognition Grand Challenge Experiments, 2005, pp. 154–160.

[38] J. Cook, C. Mccool, V. Chandran, S. Sridharan, Combined 2D/3D face recognition using log-Gabor templates, in: IEEE Conference on Video and Signal Based Surveillance, 2006, pp. 83–90.

About the Author—CHENGHUA XU received the BS degree in Industry Automation in 1999 and the MS degree in Control Theory and Engineering in 2001 from Northeastern University. Then he obtained the PhD degree in 2005 in Institute of Automation, Chinese Academy of Sciences (CASIA). After that, he worked in NEC China labs as Associate Researcher. His current interests include computer vision, pattern recognition and computer graphics.

About the Author—STAN Z. LI received the PhD degree from Surrey University, UK. He is currently a professor at National Laboratory of Pattern Recognition (NLPR), the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA); and co-director of Joint Laboratory for Intelligent Surveillance and Identification in Civil Aviation (CASIA-CAUC). He worked at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore.

About the Author—TIENIU TAN graduated with a BSc from Xi'an Jiaotong University in 1984, and obtained his MSc (in 1986) and PhD (in 1989) degrees from Imperial College of Science, Technology and Medicine, London, UK. Prior to his return to China in 1998, he worked at the University of Reading, UK, as Research Fellow, Senior Research Fellow and Lecturer. He currently serves as the President of the Institute of Automation as well as the Director of the NLPR. He leads the Intelligent Recognition and Digital Security Group of the NLPR. His current research focuses on the visual surveillance and monitoring of dynamic scenes, personal identification based on multiple biometric features such as face, iris, fingerprint, handwriting and gait, and watermarking of digital multimedia data such as digital static images and digital video.

About the Author—LONG QUAN received the PhD degree in Computer Science from INPL, France, in 1989, and the Habilitation in Computer Science from INPG, France, in 1997. Before joining the Computer Science Department at the Hong Kong University of Science and Technology (HKUST) in 2001, he has been a French CNRS senior research scientist at INRIA in Grenoble since 1990. His research interests focus on vision geometry, 3D reconstruction, image-based modeling, structure from motion, and image-based rendering.