Gender and Age-Group Classification Based on the Integration of Multiple Classifiers with Various Image Features

画像情報を利用した 複数識別器統合による性別と年齢層の識別

July, 2007

Department of Computer Science
Waseda University

Kazuya Ueki

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Over the last 20 years or so, face recognition has become a popular area of research in computer vision [44] [51] [52] [55]. Face recognition technologies are growing more sophisticated and becoming a big part of our lives. There are a large number of commercial, security, and forensic applications that require the use of face recognition technologies.

We have been researching and developing an automatic gender and age-group classification system by extracting human features from images. Automatic gender classification has received substantial attention from researchers for the last 15 years. One of the potential benefits of gender and age-group classification systems is that the demographic data can be used for market research purposes. Classifying personal features, such as gender and age-group, of customers shopping in convenience stores or department stores using in-store cameras will enable these stores to provide the customers with personalized services. These stores will also have a marketing advantage using such detailed customer information. However, when there are so many customers coming into a store at the same time, the shop clerk cannot acknowledge and process information on all of them, especially with regards to

customers who merely browse and do not have direct contact with the clerk. Therefore, the automatic gender and age-group classification systems using cameras that are already attached for security purposes are desirable. Moreover, a user-friendly human-machine interface (HMI) will become promising using such technique.

Another advantage is being able to retrieve images of people using search engines. Considering the enormous number of images, it is impossible to manually label all of the contents. One basic filter is to determine the gender or age-group of the person in the image.

Another example is cigarette or alcohol vending machines equipped with security cameras and emergency buzzers, they can prevent children from buying cigarettes or alcohol by identifying age-groups. Because of these strong advantages, there is ongoing research on the analysis of the facial images taken with the camera and the classification of gender and age-group.

Despite the high level of current interest in gender and age-group classification systems, there is still no system which is able to work accurately in a real world environment. There are many underlying causes that make it difficult. In general, current face recognition systems encounter difficulties with large facial appearance variations due to head pose, illumination, and expression changes, especially when used in practical applications. Additionally, gender and age-group classification systems are plagued by many other problems. The first drawback is the resolution of the image. It is difficult to acquire high-resolution face images in the real life environment. If a gender and age-group classification system for market research purposes is considered, surveillance cameras are used in order to detect faces and recognize gender and age-groups. Facial images that are captured by surveillance cameras usually have a very low-resolution, which significantly limits the performance of gender and age-group classification system. Face verification

systems can use clear high-resolution images, but surveillance systems can not be expected to show people's faces clearly. The second drawback is that there are fundamental difficulties in classifying gender and age-group - even humans cannot recognize a person's age correctly. In the age-group classification systems, as there would be some people who would look younger but are actually older than their appearance, and vice versa, it is almost impossible or infeasible to achieve 100% accuracy. The third drawback is that no large-scale database has been constructed covering a wide range of age-groups. Therefore, there has been no attempt to analyze the data in the narrower range of age-groups.

In order to solve these problems, many images were taken under different lighting conditions and a large database was established, including a wide range of age-groups. With this database, we will be able to focus on new classification methods; subdividing the age categories into small ranges such as 5 or 10-years. We will then find a better dimensional reduction algorithm to reduce illumination changes. Moreover, we will use different features, create many classifiers and integrate them to reduce errors. Here, two ways of integration are considered. First, we will use not only the facial area but also other information, for instance, hair and neck etc. Second, in order to obtain variations, we will change feature extraction methods, even from a single image.

1.2 Gender Classification: Survey

Over the last 15 years, gender classification from facial images has been one of the most actively researched topics in pattern recognition. A successful gender classification system has many potential applications such as user-friendly human-machine interfaces, multimodal interaction on multimedia terminals, efficiency in demographic data collection, and automatic customer

analysis in convenience stores, department stores, and shopping malls. A number of studies have been conducted to classify gender from facial images. Early gender classification approaches can fall into one of two categories: (i) geometry-based approaches, which are based on geometric features such as face width, mouth size, distances, etc., and (ii) appearance-based approaches, which find the decision boundary between male and female from training images without extracting any geometric features.

Geometry-based methods are based on geometric features such as face width, mouth size, distances, etc. Brunelli et al. [5] used 16 geometric features (pupil to eyebrow separation, nose width, mouth width, etc.) as the input to two competing HyperBF networks, one from male and the other from female. Burton et al. [6] extracted 73 points from 179 (91 males and 88 females) frontal views of faces and used discriminant analysis to classify gender using point-to-point distances. Fellous et al. [15] used 22 horizontal and vertical facial measurements with 109 images for training and experimented with 57 test images. These methods provided an average error rate of more than 10%.

Appearance-based methods find the decision boundary between male and female from training images by training classifiers such as neural network (NN), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) classifier. Cottrell et al. [2] proposed a face, emotion, and gender recognition method using neural networks, "EMPATH". Their study used 64x64 pixel images containing 20 individuals (10 male and 10 female subjects). Golomb et al. [3] trained a fully connected two-layer network, "SexNet", to classify gender from 30x30 facial images using 90 images (45 male and 45 female images). Yen et al. [7] investigated representations developed by different types of networks (PDP, RBF) using images from a large database of 1,400 faces. Similar to the above methods, Tamura et al. [11] used a multi-layer neural

network to classify gender from face images of multiple resolutions (32x32, 16x16 and 8x8). Abdi et al. [8] used PCA-based image representations with RBF networks and perceptron networks. A very favorable classification rate was achieved by a perceptron classifier trained with PCA-based features using 160 facial images (80 males and 80 females). The above methods achieved an error rate of around 10%.

Gutta et al. [18] proposed hybrid classification architectures for gender and ethnic classification of human faces and showed feasibility using a collection of 3,006 facial images corresponding to 1,009 subjects from the FERET database [35]. This hybrid approach consists of an ensemble of RBF networks and inductive decision trees (DT). The best average error rate of their experiments was 4%. Gutta et al. [33] further proposed a mixture of experts consisting of ensembles of RBFs and reported an error rate of 4%. Moghaddam et al. [40] investigated the use of nonlinear SVMs for gender classification. They used low-resolution thumbnails (21x12) processed from 1,755 images from the FERET face database and compared to traditional pattern classifiers such as linear, quadratic, Fisher linear discriminant, nearest-neighbor, RBF classifiers, and large ensemble-RBF networks. They reported an error rate of 3.4% using Gaussian RBF kernel. Furthermore, the difference in classification performance between low-resolution thumbnails (21x12) and high-resolution images (84x84) was only 1%. Sun et al. [41] demonstrated that Genetic Algorithms (GA) could select good subsets of features in order to reduce the classification error. In their study, four different classifiers were compared: a Bayes classifier, a neural network classifier, a SVM classifier and a classifier based on LDA. Walawalkar et al. [45] presented a multi-modal gender classification using SVMs for both audio and visual cues.

F. H. C. Tivive et al. [61] used a class of convolutional neural networks for gender classification. These networks are built upon the concepts of local

receptive field processing and weight sharing, which makes them more tolerant to distortions and variations in two dimensional shapes. Tested on two separate data sets, the proposed networks achieve better classification accuracy than the conventional feedforward multilayer perceptron networks. On the FERET benchmark dataset, the proposed convolutional neural networks achieve a classification rate of 97.1%.

S. Baluja et al. [64] presented a method based on AdaBoost using a low resolution grayscale picture of a face. They matched the performance obtained with SVMs. However, the classification was achieved with a fraction of the computational expense; the classifiers are 1-2 orders of magnitude faster (approximately 50 times) than SVMs.

In their research, facial information alone was used, and the performance limits seemed to be approached. In order to exceed the classification limits, we integrate the multiple methods to take full advantage of each approach. More precisely, multiple classifiers can be generated by training multiple sets of samples that are produced from different feature vectors, and can improve the accuracy of the classification. Firstly, we use not only facial parts but also other information, especially ties and *décolletages* (clothes with low-cut necklines) are focused on. If someone wears a tie, they are more likely to be a man. Also, if someone wears a *décolletage*, they are more likely to be a woman. Secondly, we try to extract different information even from a single source. For instance, bits of information from monochrome, color and edge images are extracted, and we then integrate the results of those extractions.

1.3 Age-group Classification: Survey

Although a person's age is one of the important factors for face recognition, only a few researchers had paid attention and applied age to the task of face recognition. It is worth noting that aging is becoming known as one of

major difficulties in this area. As the difference in age between the training face and testing face grows, the performance of most algorithms significantly degrades. Recently, researchers have developed many methods to handle the aging problem.

Burt et al. [9] investigated the process of aging using face composites from different age-groups and caricature algorithms. They generated average faces for different age-groups, using images of subjects with ages between 20 and 62 years. According to their experimental evaluation, in most cases, the perceived age of the blended images was consistent with the actual age of the subjects used for generating each composite, showing that age information for each age-group was retained through the process of blending. O'Toole et al. [16] [24] used three-dimensional facial information for building a parametric 3D face model. They used a caricature algorithm in order to exaggerate or deemphasize distinctive 3D facial features; in the resulting caricatures, the perceived age was increased or decreased according to the exaggeration level, suggesting that 3D distinctive facial features were emphasized in older face. Choi [25] used PCA and 3-D face shape model to extract the age change components from 3-D facial images, and then added the age change components to test image to synthesize the facial images at different ages. Lanitis et al. [26] [31] [32] [42] proposed a face recognition system robust to age variation They built a face model and an age function to isolate age change.

Wang et al. [63] proposed an automatic age simulation method for robust face recognition. Their experiments showed that the recognition rate was satisfied with age simulation.

Compared with gender classification, very few attempts have been made at age-group classification [27] [38] due to the following three reasons: (i) increase in the number of classes, (ii) inaccuracy even by human evaluations,

and (iii) no database with a large data set. However, shop owners in Japan have potential needs of age-group classification systems especially in convenience stores or shopping malls for market research purposes. They would like to record the customer's demographic data such as gender and age-group. Actually, at convenience stores in Japan, when a purchase is made, the shop clerk at the cash register punches a key to input the customer's gender and estimated age-group in 5 or 10-year increments to collect customers' demographic data.

Kwon et al. [27] presented a theory that had only been implemented to classify input images into one of three age-groups: babies, young adults, and senior adults. Horng et al. [38] proposed an age classification system based on facial features to classify a facial image into one of four age-groups: babies (0-2), young adults (3-39), middle-aged adults (40-59), and old adults (60-). However, both of them were only based on rough classification, while the marketing use requires more precise age estimation, for instance, 5 or 10-year-range age-group estimation.

Kalamani et al. [62] applied Fuzzy Latice Neural (FLN) model to age classification system. They defined three winkle featurs; winkle density, wrinkle depth and average skin variance.

In this dissertation, we introduce new age-group classification algorithms called 2DLDA and 2DHDLA in order to improve the classification rates, where a large data set was created and age-groups are subdivided into smaller age-groups such as 5 or 10-year range age-groups. Appearance-based approaches are adopted as a feature extraction method. This is commonly used for real-world applications such as face recognition and gender classification systems for the reason of practicality. The appearance-based approaches find the decision boundary from training images without extracting any geometric features, whereas the geometry-based approaches need high-resolution

images in order to extract the precise positions of facial features such as eyebrows, nose, wrinkles, etc. Additionally, two directions of two-dimensional algorithms, R-2DPCA, C-2DPCA, R-2DLDA and C-2DLDA, are presented and they are integrated. We also look at age-group classification from a different point of view. More precisely, perceived ages given by observers are used and considered in our experiments.

1.4 Thesis Outline

This dissertation is organized into 8 chapters. Chapter 1 introduces the background of our research, and the survey of previous work in gender and age-group classification. Chapter 2 presents a detailed review of our new large-scale database (WIT-DB), which includes more than 5,000 Japanese subjects (approximately 2,500 females and about 3,000 males). This database is used in all experiments. This database is appropriate for practical real-world applications. Chapter 3 describes gender classification methods developed from the integration of facial, hairstyle, and clothing images. Experimental results are also described. In Chapter 4, we propose gender classification methods based on integration of multiple classifiers using different features of facial and neck images. The goal of this chapter is to push back the boundaries of the traditional techniques and to get the best performance. In Chapter 5, new two-dimensional projection methods called 2DLDA and 2DHLDA are introduced to achieve better performance in age-group classification. We compare our proposed two-dimensional methods to conventional methods. In Chapter 6, perceived ages instead of actual ages are considered. The performances based on actual age data and perceived age data are compared. We also show which data (actual-age based data or perceived-age based data) should be used and how we can improve the class separability and classification rates. Chapter 7 shows age-group classification methods based on multiple two-

dimensional projection algorithms. We divide two-dimensional projections into two different directions (row-direction and column-direction), called R-2DPCA and C-2DPCA, R-2DLDA and C-2DLDA. We present the detailed experiments using normalization and fusion techniques. Chapter 8 summarizes the proposed work and presents future directions related to this work.

Chapter 2

WIT-DB (Waseda human-computer Interaction Technology - DataBase)

Along with the development of face recognition algorithms, a comparatively large number of face databases have been collected and used for training face recognition algorithms and testing the performance of those. The representative publicly available databases are FERET (USA) [35], XM2VTS (UK) [36], AR Face DB (USA) [22] [39], CMU Pose, Illumination, and Expression (PIE) Database (USA) [47]. Recently face database for Asian, such as CAS-PEAL (Chinese face database) [53] and Korean Face Database (KFDB) [48], have been constructed. There is only one Japanese database (HOIP database), which includes 300 subjects (150 males and 150 females). However, in these databases, the number of data in one age-group is not sufficient or even there is no actual age data provided. In terms of person's features, the person's neck and clothes are substantial factors to discriminate gender and age-groups, but there is no huge database that contains a large number of color images showing people's necks and clothes. For these reasons, considering market research applications in Japan, these databases are not adequate to recognize people's gender and age-groups. Thus, first of all, we developed

Table 2.1: Overview of the number of subjects depending on recording conditions, gender and age-groups in WIT-DB

gender	age-group	recording conditions							total
		1	2	3	4	5	6	7	
females	0-8	7	0	0	28	71	0	41	147
	9-11	2	0	0	20	78	0	41	141
	12-14	0	2	0	7	29	0	44	82
	15-19	0	147	57	82	9	0	56	351
	20-24	4	4	0	33	10	25	78	154
	25-29	8	0	0	5	5	112	67	197
	30-34	11	0	0	4	11	50	116	192
	35-39	7	1	0	10	101	30	81	230
	40-44	1	1	0	18	104	5	130	259
	45-49	0	5	3	23	19	3	145	198
	50-54	1	1	0	9	4	2	146	163
	55-59	0	1	0	2	0	0	145	148
	60-	0	0	0	4	3	1	146	154
males	0-8	9	0	0	14	96	0	36	155
	9-11	5	0	0	19	148	0	23	195
	12-14	0	1	0	15	48	0	56	120
	15-19	0	218	162	166	3	0	42	591
	20-24	5	11	20	96	41	41	10	224
	25-29	9	2	5	11	6	243	10	286
	30-34	7	1	1	7	3	168	12	199
	35-39	10	1	2	11	8	213	18	263
	40-44	0	1	0	13	14	77	86	191
	45-49	7	2	0	13	12	59	95	188
	50-54	3	0	4	15	0	30	115	167
	55-59	0	0	0	3	1	17	146	167
	60-	3	0	0	8	0	3	144	158
total		99	399	254	636	824	1,079	2,029	5,320

new large-scale database called WIT-DB (Waseda human-computer Interaction Technology - DataBase) for gender and age-group classification system.

We will briefly introduce WIT-DB below. WIT-DB has been collected at Waseda University and NEC Soft, Ltd since 2003. It contains images of 5,320 different Japanese subjects (2,416 females and 2,904 males). The images were

recorded with video cameras in 7 different recording conditions, and then digitized. Table 2.1 gives the number of subjects in each recording condition, gender and age-group in WIT-DB. In some recording conditions, the images systematically sampled a large number of poses and illumination conditions. The pose angle varies from $+90°$ to full frontal and on to $-90°$, but frontal images are mainly used in our experiments. Figure 2.1 shows sample images from WIT-DB, which include the person's neck and clothes. The faces were sometimes illuminated by dominant light sources. The resulting changes in facial expression are typically subtle, often switching between "neutral" and "smiling." The way of recording, image resolution and the size of the images were also different depending on the recording conditions. For example, in some environments the subjects were naturally walked, while in other conditions they were seated on a stool and instructed to maintain a constant head position (although slight movements were unavoidable). Therefore, all images are cropped and rectified according to the manually located eye and mouth positions in the training and testing phases. Figure 2.2 shows examples of cropped images, which are 64x64 pixels. Input images were rotated so that the eyes were perfectly aligned horizontally and the distance between the eyes and mouth scaled to 20 pixels.

Figure 2.1: Sample images from WIT-DB

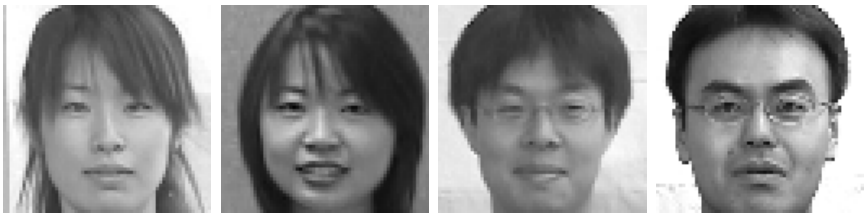


Figure 2.2: Cropped sample images from WIT-DB(64x64)

Chapter 3

Gender Classification based on the Integration of Facial, Hairstyle, and Clothing Images

3.1 Introduction

In this chapter, we present a method of gender classification by integrating facial, hairstyle, and clothing images. Initially, input images are separated into facial, hairstyle and clothing areas. Then we adopt the Principal-Component-Analysis based feature extraction and Gaussian-Mixture-Model-based likelihood calculation on each classification category. The classification results are then integrated into a single score using some known prior data based on the Bayes' rule. Experimental results showed that our integration strategy significantly reduced error rate in gender classification compared with the conventional facial only approach.

In Section 3.2, we describe our gender classification method. In Section 3.3, we provide a method of gender classification using the facial images. In Section 3.4, we focus on the method using hairstyle images. In Section 3.5 and 3.6, we focus on ties and *décolletages* (clothes with low-cut necklines), which are two clothing characteristics that differentiate gender. In Section 3.7, we describe the framework that integrates information concerning facial,

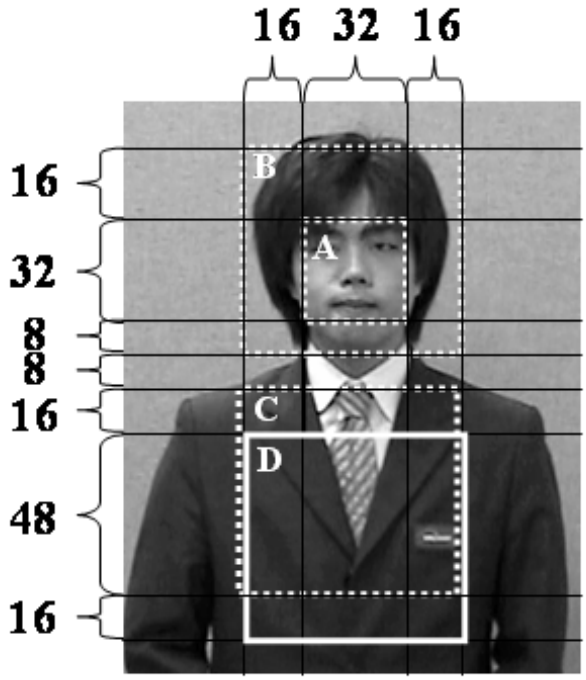


Figure 3.1: Example input image

hairstyle, and clothing images. Conclusion and future works are given in Section 3.8 and 3.9.

3.2 Our Gender Classification Method

3.2.1 Database

In this chapter, 7,432 images from WIT-DB are used for testing. These images contain the entire upper body. An example image is shown in Figure 3.1. Area A (small area of the face) and Area B (large area of the face) are used to classify gender by extracting facial and hairstyle information respectively. Area C is used to detect the person's tie, and Area D is to detect the person's *décolletage*.

3.2.2 Outline of Our Method

We will attempt to use four different features specifically face (Area A), hairstyle (Area B), tie (Area C) and *décolletage* (Area D) rather than only face, and integrate them in order to achieve better performance. Simply speaking, four types of features are classified as shown below:

1. Classify gender (male/female) directly using Area A (facial features).
2. Classify gender (male/female) directly using Area B (hairstyle features).
3. Classify tie or non-tie using Area C (tie area).
4. Classify *décolletage* or non-*décolletage* using Area D (*décolletage* area).

If a tie is found in Area C, this person is much more likely to be a man than a woman. In a similar way, if a *décolletage* is found in Area D, this person is more likely to be a woman.

Broadly, our proposed method consists of four steps; (1) feature extraction, (2) dimensionality reduction, (3) classification and (4) integration. The detailed steps are described in Figure 3.2.

In the first step, a discriminative feature from each area is extracted. We then project the high dimensional image data to a low-dimensional PCA subspace. We omit the explanation about PCA since it will be described in Section 5.2.2. Classification is then performed using the Gaussian mixture model (GMM), which will be explained in the next section. Finally, Bayes' rule is used for the process of integration.

3.2.3 Gaussian Mixture Model

Some of the classification techniques available are Support Vector Machine (SVM) [10] [23] and Neural Networks (NN). We chose to use another technique, called Gaussian Mixture Model (GMM) [1]. We constructed female GMM using female data and male GMM from male data, and made classifications by comparing the output likelihoods from each model. The advantage of using GMM is that it can automatically express many different faces using mixture distribution, such as whether a person wears glasses or not and whether a person's mustache is heavy or not.

Gaussian mixture model (GMM) is a mixture of several Gaussian distributions and can therefore represent different subclasses inside one class. The Gaussian probability density function in one dimension is a bell shaped curve defined by two parameters, mean μ and variance σ^2. In the d-dimensional space it is defined in a matrix form as

$$P(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})], \tag{3.1}$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ the covariance matrix. The probability density function is defined as a weighted sum of Gaussians

$$P(\boldsymbol{x}; \theta) = \sum_{m=1}^{M} \alpha_m P_m(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{3.2}$$

where α_m is the weight of the component m, $0 < \alpha_m < 1$ for all components, and $\sum_{m=1}^{M} \alpha_m = 1$. The parameter list

$$\theta = \{\alpha_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \alpha_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\} \tag{3.3}$$

defines a particular Gaussian mixture probability density function.

3.3 Gender Classification using Facial Images

3.3.1 Feature Extraction

In this section, upper body images (Area A of Figure 3.1) are used in order to classify gender. These images include a person's eyebrows, eyes, nose, and mouth. These images are 32x32 pixels, and is converted to 256-level grayscale. Since images will be captured at low resolution, it is difficult to extract the detailed features of a face. Thus, features were taken out from approximately 11,656 facial images including 4,389 female images and 7,267 male images by compressing dimensions using Principal Component Analysis (PCA). When cumulative proportion of PCA was set to 80%, 1,024 dimensions were compressed into 36 dimensions.

3.3.2 Experiments using Facial Images

We constructed female GMM using approximately 2,800 female images and male GMM using approximately 4,600 male images on a 36-dimensional space (compressed using PCA). Accuracy of gender classification was evaluated by using 2,397 female images and 5,035 male images as inputs to female GMM and male GMM. The number of Gaussians is set to 10 for both males and females. The output likelihood values from both GMMs are compared for classifications. Table 3.1 gives the result of gender classification using the facial images. The error rate in gender classification for facial images is 10.4%.

Table 3.1: Gender classification result using facial images

Method	# Sample	# Error	Error rate
Female	2,397	373	15.6%
Male	5,035	400	7.9%
Total	7,432	773	10.4%

3.4 Gender Classification using Hairstyle Images

3.4.1 Feature Extraction

Hairstyle is considered to be one of the most distinguishable features in gender classification. We created hairstyle image from Area B of Figure 3.1. The hairstyle area is extracted based on color information. This image is resized to 32x32 pixels, and is converted to 5-level grayscale. The hairstyle feature extraction method is performed by carrying out PCAs of the entire set of hairstyle images as was done on the facial images. The number of images is 11,726 including 4,433 female images and 7,293 male images. Cumulative proportion is 80%, and reduced dimension is 31.

3.4.2 Experiments using Hairstyle Images

We constructed female GMM using approximately 2,800 female images and male GMM using approximately 4,700 male images on a 31-dimensional space (compressed using PCA). We constructed female GMM and male GMM, and made classifications by comparing the output likelihoods from each model. 2,397 female images and 5,035 male images were used as inputs to female GMM and male GMM. The number of Gaussian is set to 5 for both males and females. The output likelihood values from both GMMs are compared for classifications. Table 3.2 gives the result of gender classification for hairstyle images.

Table 3.2: Gender classification result using hairstyle images

Method	# Sample	# Error	Error rate
Female	2,397	381	15.9%
Male	5,035	607	12.1%
Total	7,432	988	13.3%

3.5 Tie/non-tie Classification

3.5.1 Feature Extraction

We created tie/non-tie images by applying Laplacian filtering to Area C from Figure 3.1. This image is resized to 24x24 pixels, and is converted to 256-level grayscale. We extracted features using PCA with 7,577 edge images including 1,212 tie images and 6,365 non-tie images. Cumulative proportion is 60%, and reduced dimension is 57.

3.5.2 Experiments in Tie/non-tie Classification

We constructed tie GMM using approximately 800 tie edge images and non-tie GMM using approximately 4,100 non-tie edge images on a 57-dimensional space (compressed by PCA). We constructed tie GMM and non-tie GMM, and made classifications by comparing the output likelihoods from each model. 1,212 tie images and 6,220 non-tie images were used as inputs to tie GMM and non-tie GMM. The number of Gaussian for tie GMM is set to 1, and 5 for non-tie GMM. Table 3.3 gives the result of tie/non-tie classification.

Table 3.3: Tie/non-tie classification result using edge images

Method	# Sample	# Error	Error rate
Tie	1,212	91	7.5%
Non-tie	6,220	95	1.5%
Total	7,432	186	2.5%

3.6 *Décolletage*/non-*décolletage* Classification

3.6.1 Feature Extraction

We next created *décolletage*/non-*décolletage* images. What is different about this image from the tie images is the extraction of skin area from the image. We created skin images from Area D of Figure 3.1. This image is resized to 24x24 pixels, and is converted to 256-level grayscale. The skin area is extracted based on the person's facial skin color information. We extracted features using PCA with approximately 7,577 skin images (approximately 210 *décolletage* images and 7,367 non-*décolletage* images). Cumulative proportion is 30%, and reduced dimension is 12.

3.6.2 Experiments in *Décolletage*/non-*décolletage* Classification

We constructed *décolletage* GMM using approximately 300 *décolletage* skin images and non-*décolletage* GMM using approximately 500 non-*décolletage* skin images on a 12-dimensional space (compressed by PCA). Classification is done as was with other images described in the previous sections.

207 *décolletage* images and 7,225 non-*décolletage* images were used as inputs to *décolletage* GMM and non-*décolletage* GMM respectively. The number of Gaussian for *décolletage* GMM is set to 1, and 5 for non-*décolletage* GMM. Table 3.4 gives the result of *décolletage*/non-*décolletage* classification for skin images.

Table 3.4: *Décolletage*/non-*décolletage* classification result using skin images

Method	# Sample	# Error	Error rate
Décolletage	207	43	20.8%
Non-*décolletage*	7,225	417	5.8%
Total	7,432	460	6.2%

3.7 Multiple Information Integration

3.7.1 Integration Method

We denote the face feature extraction data as x_F, the hairstyle feature extraction data as x_H, the tie/non-tie data as x_T, the *décolletage*/non-*décolletage* data as x_D. Considering each data independent, the ratio of the female probability $\Pr[F|x_F, x_H, x_T, x_D]$ and the male probability $\Pr[M|x_F, x_H, x_T, x_D]$ is calculated as follows:

$$
\begin{aligned}
&\Pr[F|x_F, x_H, x_T, x_D] : \Pr[M|x_F, x_H, x_T, x_D] \\
&= \Pr[F] \times (\Pr[x_F|F])^{n_F} \times (\Pr[x_H|F])^{n_H} \\
&\quad \times (\Pr[tie|F] \cdot \Pr[x_t|tie] + \Pr[\overline{tie}|F] \cdot \Pr[x_t|\overline{tie}])^{n_T} \\
&\quad \times (\Pr[dec|F] \cdot \Pr[x_d|dec] + \Pr[\overline{dec}|F] \cdot \Pr[x_d|\overline{dec}])^{n_D} \\
&: \Pr[M] \times (\Pr[x_F|M])^{n_F} \times (\Pr[x_H|M])^{n_H} \\
&\quad \times (\Pr[tie|M] \cdot \Pr[x_t|tie] + \Pr[\overline{tie}|M] \cdot \Pr[x_t|\overline{tie}])^{n_T} \\
&\quad \times (\Pr[dec|M] \cdot \Pr[x_d|dec] + \Pr[\overline{dec}|M] \cdot \Pr[x_d|\overline{dec}])^{n_D}, \qquad (3.4)
\end{aligned}
$$

which is calculated using the following prior probability:

$$
\begin{aligned}
&\Pr[F] = \Pr[M] = 0.500, \\
&\Pr[tie|F] = 0.005, \Pr[tie|F] = 0.995, \\
&\Pr[dec|F] = 0.050, \Pr[dec|F] = 0.950, \\
&\Pr[tie|M] = 0.100, \Pr[tie|M] = 0.900, \\
&\Pr[dec|M] = 0.010, \Pr[dec|M] = 0.990, \qquad (3.5)
\end{aligned}
$$

and n_F, n_H, n_T, n_D are weight parameters to compensate for differences between each likelihood. n_F, n_H are set to 1, and n_T, n_D are set to 5. These prior probabilities can be calculated by using the statistics taken from a real life environment.

3.7.2 Experimental Result Based on the Likelihood-based Integration

An integrated result is shown in Table 3.5. Facial, hairstyle, tie and *décolletage* information all seem to affect the classification result to some degree.

Table 3.5: Gender classification result using integrated information: F, H, T, and D represent face, hairstyle, tie, and *décolletage* respectively.

	method of integration	# Sample	# Error	Error rate
Female	F	2,397	373	15.6%
	F + H		305	12.7%
	F + H + T + D		297	12.4%
Male	F	5,035	400	7.9%
	F + H		308	6.1%
	F + H + T + D		282	5.6%
Total	F	7,432	773	10.4%
	F + H		613	8.2%
	F + H + T + D		579	7.8%

3.8 Conclusion

This chapter proposed a method of gender classification by integrating information from different parts of a single image. By integrating the likelihoods of the hairstyle and clothing, we were able to reduce 25.1% of false classifications made by the conventional, facial only approach. Experimental results show that classifying extracted images of the face, the hairstyle, and the clothing individually is effective in gender classification.

3.9 Future Works

This work involved images from only the frontal view, but we are planning to incorporate images from various angles. Moreover, We plan to adapt the integration theory mentioned in Section 3.7.1 to multi-frame images (movies). Furthermore, we plan to use the physical and clothing information in order to recognize not only gender and age-group but also occupation-type (such as corporate employee or student).

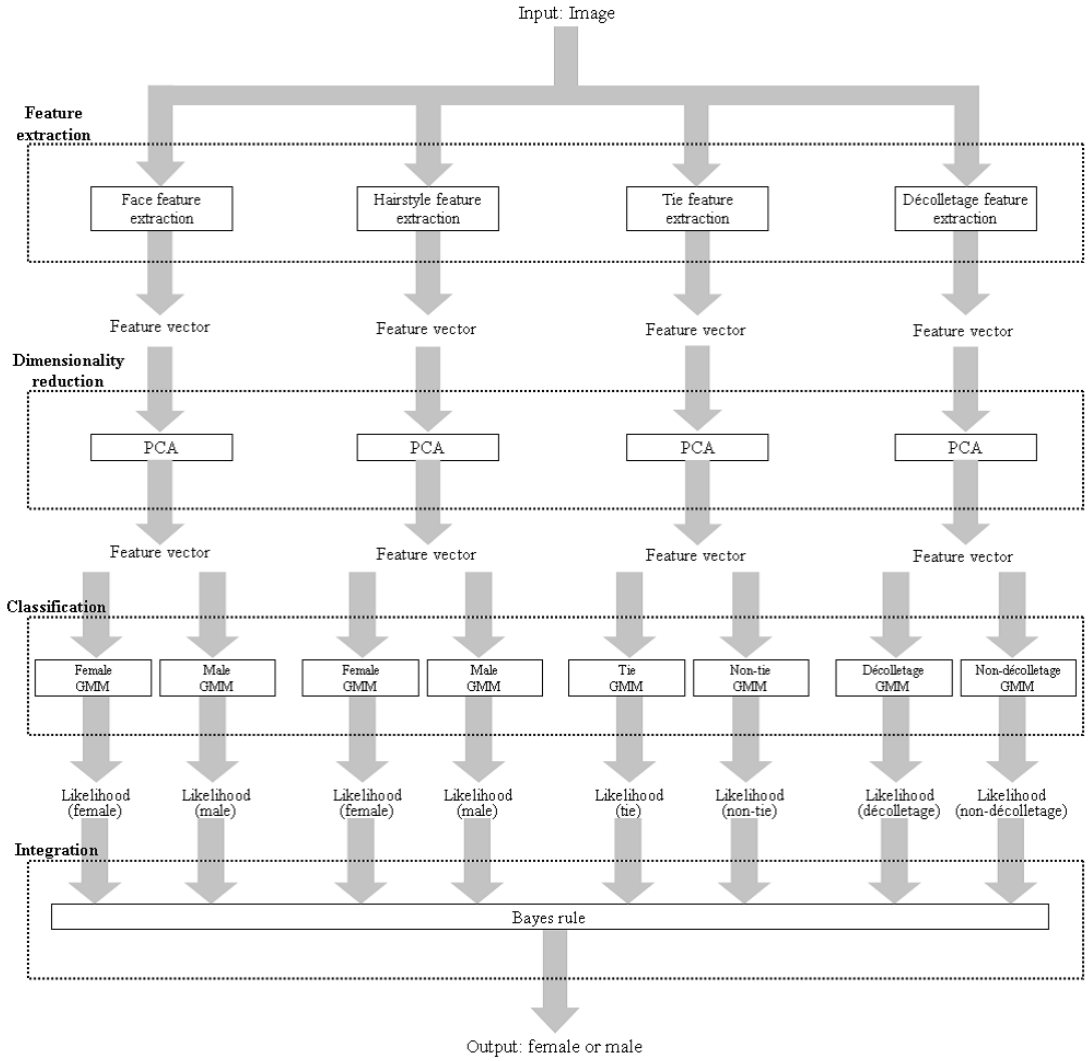


Figure 3.2: Our gender classification scheme

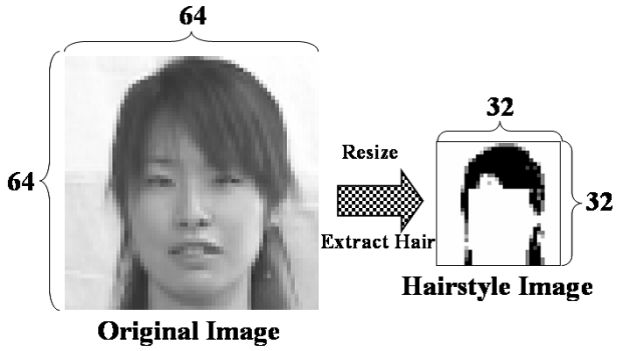


Figure 3.3: Example hairstyle image

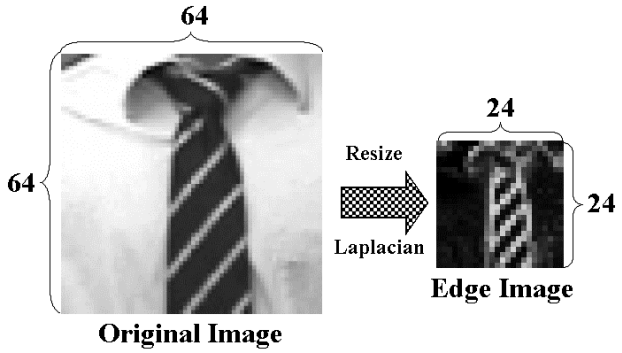


Figure 3.4: Example edge image for tie/non-tie classification

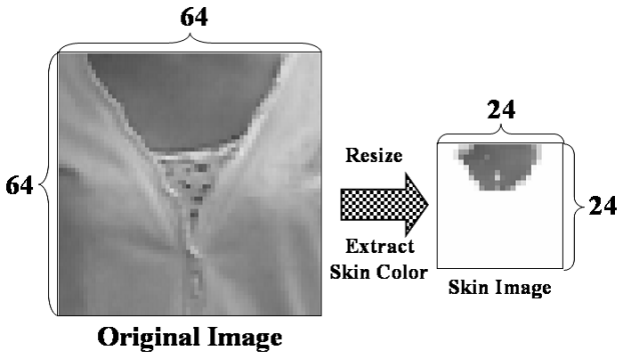


Figure 3.5: Example skin image for *décolletage*/non-*décolletage* classification

Chapter 4

Gender Classification based on Integration of Multiple Classifiers Using Different Features of Facial and Neck Images

4.1 Introduction

In this chapter, we will try to reduce as many errors in gender classification as possible. The best performance has been achieved by using SVM classifiers in [40] [41], especially SVMs with RBF kernel function are the most effective classifiers in gender classification. However, only a single classifier was eventually used to classify gender in these studies. In addition, previous research in other areas has shown that classifier combination has improved the recognition accuracy rather than single classifier approaches [20]. For these reasons, we employ SVM with a kernel function using facial monochrome images as our baseline, and try to improve the performance by integrating multiple classifiers, which have different characteristics. As for the different characteristics, we use not only facial region, but also neck region, and extract different features from each region. The different characteristics are

used because the different region and features include other important contributing factors in gender classification such as skin color, neck size, jaw line, neckline, color of clothing, and types of clothing etc. One of the best points about integrating other information is that different types of error patterns made by other information could be more likely to help reduce the errors. Another point is that if a classifier on facial monochrome images does not have confidence, other classifiers could compensate for it. Hence, in our experiments, we separate images into facial and neck regions and monochrome, color, and edge images individually from the facial region and the neck region are extracted.

We describe SVM in Section 4.2, and our proposed framework in Section 4.3. Experimental results for each component are discussed in Section 4.4, and integration results are given in Section 4.5. Conclusions are presented in Section 4.6.

4.2 Support Vector Machines (SVMs)

We investigated SVMs [10] [19] [23] for gender classification. SVM is one of the most successful classification techniques in pattern recognition. The basic idea of SVMs is to find the optimal linear hyperplane such that the expected classification error for unseen test samples is minimized. For constructing non-linear decision functions, SVMs can map the input data from input space into a high-dimensional feature space using kernel functions. Thus, SVMs with kernel techniques have been used in various applications such as classification and regression.

Given a labeled set of l training samples

$$(\boldsymbol{x}_1, y_1) \cdots (\boldsymbol{x}_l, y_l), \boldsymbol{x}_i \in \Re^n, y_i \in \{-1, +1\}. \tag{4.1}$$

y_i is the associated label that shows which class x_i belongs to. We assume that the distribution of these two classes is such that they are linearly separable, i.e. a hyperplane separating the two classes exists as follow:

$$(\boldsymbol{w} \cdot \boldsymbol{x}) + b = 0, \tag{4.2}$$

where $\boldsymbol{w}$ is the classifier's parameter vector, and b is a bias term. An optimal Lagrange multiplier α_i^*, and an optimal bias term b^* are computed by solving the quadratic programming problem, and the discriminant hyperplane is defined as:

$$f(\boldsymbol{x}) = \text{sign}(\sum_{i=1}^{l} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b^*), \tag{4.3}$$

where $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is a kernel function and the sign of $f(\boldsymbol{x})$ indicates the membership of $\boldsymbol{x}$. Possible choices of kernel functions include polynomial, Gaussian, and sigmoidal. For this study, the Gaussian kernel:

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}), \tag{4.4}$$

was chosen since it was empirically observed to perform better than others. In our experiment, we use the distances from the hyperplane as features for the second classifiers:

$$d = \sum_{i=1}^{l} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b^*. \tag{4.5}$$

4.3 Overview of the Proposed Approach

First, input images for gender classification are normalized to account for geometry and illumination changes. The region used in our experiment is

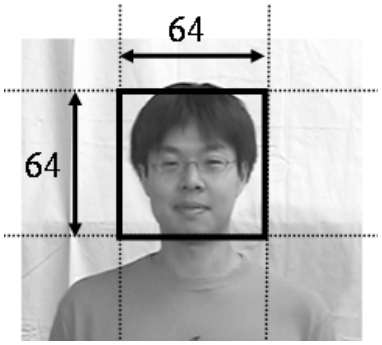


Figure 4.1: Facial region

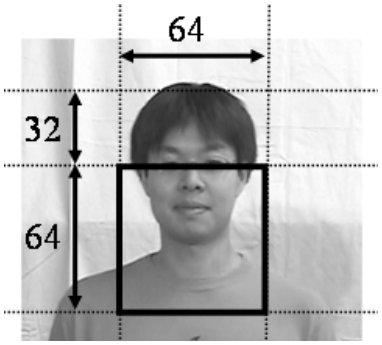


Figure 4.2: Neck region

a facial region of 64x64 pixels, as shown in Figure 4.1, and neck region of 64x64 pixels, as shown in Figure 4.2.

Our methodology for gender classification is shown in Figure 4.3. We extracted facial and neck regions, and converted each image to monochrome, color, and edge images. The feature vector was extracted from the intensity value of each image pixel. $\boldsymbol{x}_{\mathrm{M}}^{(\mathrm{F})}$ represents the feature vector from the monochrome facial image, $\boldsymbol{x}_{\mathrm{C}}^{(\mathrm{F})}$ represents the feature vector from the color facial image, and $\boldsymbol{x}_{\mathrm{E}}^{(\mathrm{F})}$ represents the feature vector from the edge facial image. In the same way, $\boldsymbol{x}_{\mathrm{M}}^{(\mathrm{N})}$, $\boldsymbol{x}_{\mathrm{C}}^{(\mathrm{N})}$ and $\boldsymbol{x}_{\mathrm{E}}^{(\mathrm{N})}$ represent the feature vector from monochrome, color, and edge neck images respectively. We used SVMs as gender classifiers, $\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{F})}$, $\mathrm{SVM}_{\mathrm{C}}^{(\mathrm{F})}$, $\mathrm{SVM}_{\mathrm{E}}^{(\mathrm{F})}$, $\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{N})}$, $\mathrm{SVM}_{\mathrm{C}}^{(\mathrm{N})}$, and $\mathrm{SVM}_{\mathrm{E}}^{(\mathrm{N})}$ which were induced from six feature vectors, $\boldsymbol{x}_{\mathrm{M}}^{(\mathrm{F})}$, $\boldsymbol{x}_{\mathrm{C}}^{(\mathrm{F})}$, $\boldsymbol{x}_{\mathrm{E}}^{(\mathrm{F})}$, $\boldsymbol{x}_{\mathrm{M}}^{(\mathrm{N})}$, $\boldsymbol{x}_{\mathrm{C}}^{(\mathrm{N})}$, and $\boldsymbol{x}_{\mathrm{E}}^{(\mathrm{N})}$. The performance of each SVM was evaluated using 2-fold cross validation: half of the data was used for training and the remaining half for tests, and by repeating the same process by reversing, all the data were evaluated. We used the distance from the hyperplane to $\boldsymbol{x}$ in equation (4.5) in order to integrate the six types of information. The distances were denoted by $d_{\mathrm{M}}^{(\mathrm{F})}$, $d_{\mathrm{C}}^{(\mathrm{F})}$, $d_{\mathrm{E}}^{(\mathrm{F})}$, $d_{\mathrm{M}}^{(\mathrm{N})}$, $d_{\mathrm{C}}^{(\mathrm{N})}$, and $d_{\mathrm{E}}^{(\mathrm{N})}$. The distances were set to zero in the absence of neck images. As for the integration of six types of information, we tried three different methods: (1) distance summation based integration,

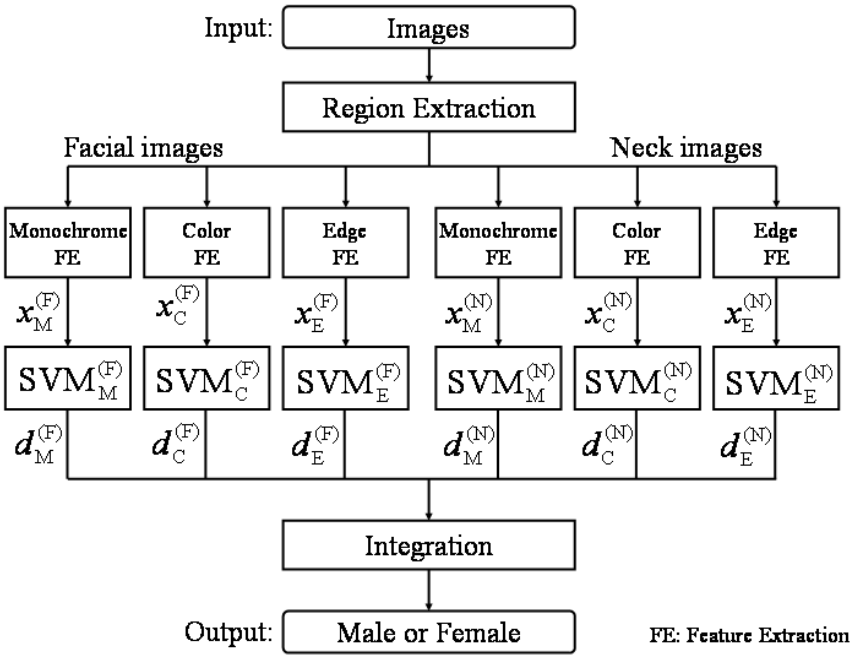


Figure 4.3: Our gender classification scheme

(2) GMM (Gaussian Mixture Model) based integration, and (3) SVM based integration. The reason we chose a distance summation method is that the problem we try to solve is not so complicated. In addition, we are trying to use a different type of classifier; a model based classifier, for comparison purposes. GMM is chosen because we suppose GMM expresses many different faces automatically by using a mixture distribution, such as whether a person wears glasses or not and whether a person's facial hair is thick or not. Finally, SVM based integration is done because SVM is one of the most powerful machine learning techniques to solve a binary classification problem.

4.4 Experiments

In this section, we briefly introduce our database, and experimental results for each component are discussed. Moreover, we combined the results from

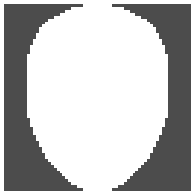


Figure 4.4: A mask image

(a) (b)

Figure 4.5: Examples of facial images: (a)Male (b)Female

monochrome facial images and the results from other images such as facial color images, neck monochrome images, etc. to improve classification accuracy more than the facial monochrome based approaches in [40] [41].

4.4.1 Database

We use most of frontal view images in WIT-DB (totally 26,569 images – 14,392 male and 12,177 female images). These images were taken from subjects in a wide variety of lighting conditions and ages, from 3 to 85 years. Figure 4.1 and Figure 4.2 show examples of the facial and neck images used in our study. There are 25,430 neck images which make up 95% of all images in our database. The remaining images (5%) do not include a neck region because the region is cut off due to the corner of the image.

4.4.2 Monochrome Facial Image Based Gender Classifier

An oval mask shown in Figure 4.4 was applied to eliminate background, and an oval region for the face, shown in Figure 4.5, was used for gender classification. This mask image was chosen since it reflected the best classification accuracy. The resulting image has 2,500 pixels. Histogram equalization was done to each image to reduce the influence of different lighting conditions (side light, bright light, very bright spots in the image, and very low light) in WIT-DB. Table 4.1 shows experimental results.

Table 4.1: Error rate comparisons between different methods. When facial region was used, the number of female, male and total samples was 12,177, 14,392 and 26,569 respectively. When neck region was used, the number of female, male and total samples was 11,622, 13,808 and 25,430 respectively. The distance summation based integration was carried out.

Method	Female		Male		Total	
	# Errors	Error rates	# Errors	Error rates	# Errors	Error rates
$SVM_M^{(F)}$	509	4.18%	660	4.59%	1,169	4.40%
$SVM_C^{(F)}$	860	7.06%	1,045	7.26%	1,905	7.17%
$SVM_E^{(F)}$	1,441	11.83%	922	6.41%	2,363	8.89%
$SVM_M^{(N)}$	1,066	9.17%	645	4.67%	1,711	6.73%
$SVM_C^{(N)}$	1,026	8.83%	1,645	10.91%	2,671	10.50%
$SVM_E^{(N)}$	1,747	15.03%	1,082	7.84%	2,829	11.12%
$SVM_M^{(F)} + SVM_C^{(F)}$	398	3.27%	593	4.12%	991	3.73%
$SVM_M^{(F)} + SVM_E^{(F)}$	506	4.16%	551	3.83%	1,057	3.98%
$SVM_M^{(F)} + SVM_M^{(N)}$	508	4.17%	480	3.34%	988	3.72%
$SVM_M^{(F)} + SVM_C^{(N)}$	414	3.40%	587	4.08%	1,001	3.77%
$SVM_M^{(F)} + SVM_E^{(N)}$	501	4.11%	522	3.63%	1,023	3.85%

We discovered the two reasons why the integration of different features were effective. The first reason is that most falsely classified data is located near the SVM hyperplane, and most correctly classified data is relatively

Table 4.2: Relationship between the distance from the hyperplane and the error rate.

Distance	# Samples	# Errors	Error rates
$0.5 \leq d_{\mathrm{M}}^{(\mathrm{F})}$	24,326	507	2.08%
$1.0 \leq d_{\mathrm{M}}^{(\mathrm{F})}$	21,188	186	0.88%

far from the hyperplane. We found that within a distance of 1.0 from the hyperplane there were 983 errors, which were 81.4% of all errors. Table 4.2 shows that there were 24,326 and 21,188 samples beyond a distance of 0.5 and 1.0 respectively from the hyperplane, and these data contained only a small number of errors (2.08% and 0.88%). This is because during the training process, even if falsely classified data exists in the wrong place, SVM classifier put the falsely classified data closer to the hyperplane to minimize the cost function. The second reason is that the different types of error patterns are made by other classifiers based on the different features. In this case, the correctly classified data is more likely to exist far from the SVM hyperplane and compensate the error made by the falsely classified data near the hyperplane. Table 4.3 shows that the number of correct data and the ratio of correct data with color facial images, edge facial images, monochrome neck images, color neck images, and edge neck images in 983 errors from facial monochrome images. These two reasons provide enough motivation for incorporating results from different classifiers to decrease error rates of gender classification.

4.4.3 Color Facial Image Based Gender Classifier

The RGB representation of color images is not suitable for images containing a wide variety of lighting conditions, because the (R, G, B) represents not only color but also luminance. Luminance varies across a person's face due to ambient lighting. Therefore, each (R, G, B) pixel in the image was

Table 4.3: The number of correct data and the ratio of correct data with color facial images, edge facial images, monochrome neck images, color neck images, and edge neck images in 983 errors using facial monochrome images. These 983 errors exist within a distance of 1.0 from the hyperplane.

Method	# Correct data	Ratio of correct data
$\mathrm{SVM}_{\mathrm{C}}^{(\mathrm{F})}$	623	63.4%
$\mathrm{SVM}_{\mathrm{E}}^{(\mathrm{F})}$	584	59.4%
$\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{N})}$	605	61.5%
$\mathrm{SVM}_{\mathrm{C}}^{(\mathrm{N})}$	585	59.5%
$\mathrm{SVM}_{\mathrm{E}}^{(\mathrm{N})}$	596	60.6%

transformed into chromaticity space shown below:

$$r = R/(R+G+B), \tag{4.6}$$

$$g = G/(R+G+B), \tag{4.7}$$

$$b = B/(R+G+B), \tag{4.8}$$

and they were used for classification. We used the same mask images for color facial images. Experimental results for color facial images are shown in Table 4.1. Table 4.1 also showed that, although the performance of color facial images was less powerful than that of monochrome facial images, adding the distances $d_{\mathrm{M}}^{(\mathrm{F})}$ and $d_{\mathrm{C}}^{(\mathrm{F})}$ from the optimal hyperplane of $\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{F})}$ and $\mathrm{SVM}_{\mathrm{C}}^{(\mathrm{F})}$ was very helpful in reducing conventional errors. By integrating color facial images, we obtained the relative error reduction of 15.2%. Here, the relative error reduction (RER) [%] is calculated as follows:

$$RER = \frac{E_o - E_i}{E_o} \times 100, \tag{4.9}$$

(a) (b)

Figure 4.6: Examples of edge images: (a)Male (b)Female

where E_o is the number of errors in the original method, and E_i is the number of errors in the improved method. This means we were able to reduce 15.2% of the erroneous classification made by the facial image only approach, signifying the effectiveness of integration between monochrome and color images.

4.4.4 Edge Facial Image Based Gender Classifier

We prepared edge images, as shown in Figure 4.6, by using a Laplacian filter in order to effectively extract wrinkle, mustache, hair and the outline of the face from facial images. Edge images were not processed using mask images as monochrome and color images. All pixels (4,096 pixels) were used as the inputs due to the use of hair contour. Classification results obtained by the use of extracted features from edge images are shown in Table 4.1. As far as we can see, the performance of edge images fell short of classification using monochrome and color images, but we recognized the improvements in classification by integrating monochrome and edge information. Table 4.1 also shows the integration results. By integrating edge images, we were able to obtain the relative error reduction of 9.6% compared with the baseline (facial image only approach).

(a) (b)

Figure 4.7: Examples of neck images: (a)Male (b)Female

(a) (b)

Figure 4.8: Examples of edge neck images: (a)Male (b)Female

4.4.5 Neck Image Based Gender Classification

Most of the conventional studies for gender classification have focused on the use of a facial region, however, little attention has been given to other regions. Factors such as neck size, jaw line, neckline, color of clothing, and types of clothing such as suits and skirts can also be used to classify gender. Concerning clothes color, there is limited information for its use in gender classification, nevertheless we can infer information about gender. Using these types of information is informative for gender classification if it is not deliberately made to deceive the system. In the previous chapter, we proposed a method of integration by using a tie and *décolletage* (clothes with low-cut neckline) and validated our approach experimentally. Thus in this study, we focused our attention on the neck region which includes neck size, jaw line, neckline, and color of clothing, in addition to tie and *décolletage*, and tried to classify gender directly. We used monochrome, color, and edge images from the neck region in the same way as we used the face region. Sample images are shown in Figure 4.7 and Figure 4.8. SVM classifiers $\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{N})}$,

$\mathrm{SVM}_{\mathrm{C}}^{(\mathrm{N})}$ and $\mathrm{SVM}_{\mathrm{E}}^{(\mathrm{N})}$ were used for classification, and the results are shown in Table 4.1. In addition, results concerning the integration of monochrome facial images are also shown in Table 4.1. Results show that various kinds of information contribute to the elimination of classification error rates over monochrome facial only approach.

4.5 Integration of Facial Information and Neck Information

4.5.1 Integration Using Six Types of Information

As discussed in the previous section, it was found that all facial and neck features were significant in gender classification, therefore we integrated all information that were extracted.

As for the integration, we experimented three different approaches: (1) distance summation based integration, (2) GMM based integration, and (3) SVM based integration.

The first, **distance summation based integration**, was to add the distances from the hyperplane as shown below:

$$y = \mathrm{sign}(d_{\mathrm{total}}), \tag{4.10}$$

where

$$d_{\mathrm{total}} = d_{\mathrm{M}}^{(\mathrm{F})} + d_{\mathrm{C}}^{(\mathrm{F})} + d_{\mathrm{E}}^{(\mathrm{F})} + d_{\mathrm{M}}^{(\mathrm{N})} + d_{\mathrm{C}}^{(\mathrm{N})} + d_{\mathrm{E}}^{(\mathrm{N})}. \tag{4.11}$$

The second, **GMM based integration**, was to use the GMM classifier. Gaussian models from each gender are estimated in 6 dimensional feature space by using the EM Algorithm. The Gaussian model is a type of probability distribution model. The d-dimensional normal distribution density

function is defined as

$$P(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})], \tag{4.12}$$

where the vector $\boldsymbol{\mu}$ is the mean of the normal distribution, and the matrix $\boldsymbol{\Sigma}$ is the variance-covariance matrix. The gender classification using test images is done based on likelihoods from each gender model, and we make classifications by comparing the output likelihoods from each model. GMM classifier with 3 mixture components was used for our experiments.

The last, **SVM based integration**, to use the SVM classifier using a 6-dimensional vector as shown below:

$$\boldsymbol{d} = (d_{\mathrm{M}}^{(\mathrm{F})}, d_{\mathrm{C}}^{(\mathrm{F})}, d_{\mathrm{E}}^{(\mathrm{F})}, d_{\mathrm{M}}^{(\mathrm{N})}, d_{\mathrm{C}}^{(\mathrm{N})}, d_{\mathrm{E}}^{(\mathrm{N})}). \tag{4.13}$$

SVM classifier with a Gaussian kernel function was used for our experiments.

Experimental results are shown in Figure 4.9. We achieved very good results; the relative error reduction was 27.5% using GMM or SVM classifiers.

Falsely classified data were at a small distance from the hyperplane of SVM, so using color facial images, including lipstick color and makeup color, and neck images, including jaw line, neckline, and color of clothing, were able to decrease our errors. Monochrome and color facial images did not show long hair on women because of the oval region, but edge facial images and neck images included hair region and helped improve the accuracy. The reason why our method can improve the accuracy is that the results from different classifiers have different types of error patterns as shown Table 4.3 and other classifiers can compensate for it. On the other hand, most images classified erroneously even after the integration were extremely difficult to

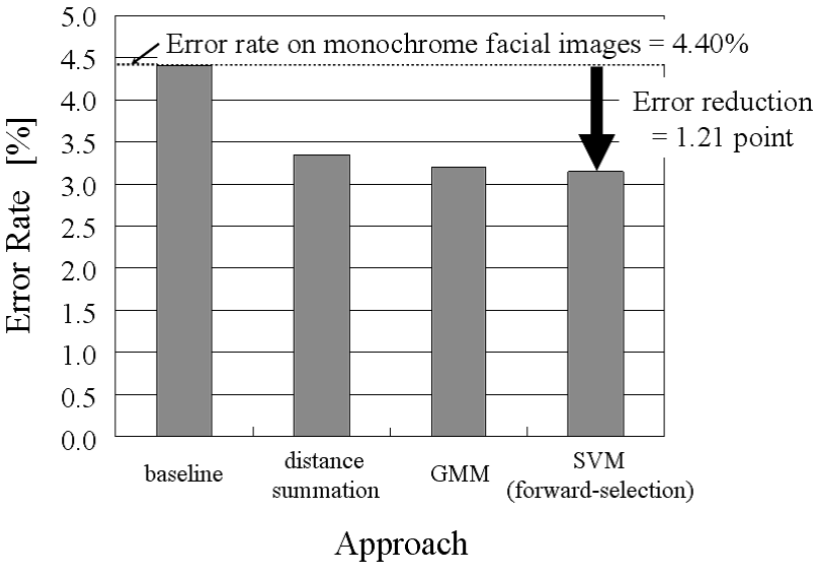


Figure 4.9: Error rate comparisons of integration methods using facial and neck information

classify even by the human eye.

In addition, we used a combination of both face and neck region (96×64) as shown in Figure 4.2, and constructed $\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{F+N})}$ as a gender classifier in order to confirm the effectiveness of integration. Experiments showed an error rate of 5.45%, which was worse than the baseline.

4.5.2 Investigation on the Contribution to the Classification Accuracy Using Forward-Selection

We next conducted experiments in determining the contribution of each piece of information to the process using forward-selection, a method of finding the best combination of variables by sequentially incrementing the number of variables used. We used the SVM based integration method, which produced the best classification accuracy. Six different types of information were used in this process, as shown in Table 4.4. The contribution of each classifier

Table 4.4: Order of contribution of each information using forward-selection. The number of female, male and total samples was 12,177, 14,392 and 26,569 respectively. The SVM based integration was carried out.

(a) Error rates

Order	Method	Female		Male		Total	
		# Errors	Error rates	# Errors	Error rates	# Errors	Error rates
1	$\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{F})}$	509	4.18%	660	4.18%	1,169	4.40%
2	$1 + \mathrm{SVM}_{\mathrm{M}}^{(\mathrm{N})}$	476	3.91%	506	3.52%	982	3.70%
3	$2 + \mathrm{SVM}_{\mathrm{C}}^{(\mathrm{F})}$	402	3.30%	472	3.28%	874	3.29%
4	$3 + \mathrm{SVM}_{\mathrm{C}}^{(\mathrm{N})}$	372	3.05%	480	3.34%	852	3.21%
5	$4 + \mathrm{SVM}_{\mathrm{E}}^{(\mathrm{N})}$	380	3.12%	457	3.18%	837	3.15%
6	$5 + \mathrm{SVM}_{\mathrm{E}}^{(\mathrm{F})}$	387	3.18%	460	3.20%	847	3.19%

(b) Error reduction

Order	Method	Error reduction	Relative error reduction
1	$\mathrm{SVM}_{\mathrm{M}}^{(\mathrm{F})}$	Baseline	Baseline
2	$1 + \mathrm{SVM}_{\mathrm{M}}^{(\mathrm{N})}$	0.70 point	16.0%
3	$2 + \mathrm{SVM}_{\mathrm{C}}^{(\mathrm{F})}$	1.11 point	25.2%
4	$3 + \mathrm{SVM}_{\mathrm{C}}^{(\mathrm{N})}$	1.19 point	27.1%
5	$4 + \mathrm{SVM}_{\mathrm{E}}^{(\mathrm{N})}$	1.25 point	28.4%
6	$5 + \mathrm{SVM}_{\mathrm{E}}^{(\mathrm{F})}$	1.21 point	27.5%

did not necessarily depend on each classifier's accuracy, as shown in Table 4.1. The individual accuracy of classifiers, as shown in Table 4.1, is not sufficient by itself to determine which one should be integrated next nor whether we should integrate further classifiers or not. Without any experiments, it is almost impossible to discover whether or not independent classifiers are obtainable, which can compensate for each other. If the database, image region or feature extraction is changed, we need to experiment and select features using the forward-selection procedure to make sure that we can obtain useful information. In this chapter, we deliberately produced such variations

using neck region, color and edge images in addition to monochrome ones. Since we can extract facial color from color facial images and wrinkles from edge facial images, the combination of information from these two types of images is thought to be effective to classify gender. Neck region images can also be considered effective for gender classification. (Additionally, we can also extract gender-specific features such as the size of a neck, the color of a cloth, and shapes of clothes – ties, *décolletages*, etc.) The most favorable classification resulted when five types of information, excluding facial edge images, were used, and it showed a 28.4% relative reduction in error over a baseline on the monochrome facial image approach.

4.6 Conclusions

In this chapter, we proposed a method of gender classification using facial images and neck images to decrease error rates made by the conventional approach. Concerning feature extraction, we used not only monochrome images but also color and edge images, and integrated the individual results. Experimental results show that using multiple classifiers and integrating different effective features, which have different error patterns, can reduce errors, because most falsely classified data were close to the SVM hyperplane. Experimental results also show a 28.4% relative reduction in error over the baseline, and our approach is significantly better than the single classifier approach based on monochrome facial images, which is considered as marginal performance.

Chapter 5

New Projection Methods for Age-Group Classification

5.1 Introduction

This chapter presents feature extraction methods for age-group classification. Two types of approaches can be considered to classify age-groups: (i) geometry-based and (ii) appearance-based approaches. As seen in Section 1.3, most previous research in age-group classification applied to geometry-based approach. The geometry-based approach is robust against pose and orientation changes. However, the real-world applications are considered, it is difficult to locate facial features due to several corruptions such as illumination, noise and occlusion. For this reason, the appearance-based approach will be used as in most face recognition systems. However, features should be robust against various lighting conditions, and feature extraction should be processed in real-time. Therefore, we would like to focus on the dimensionality reduction, which can reduce computational cost and lighting variations, and can improve the separability in age-groups.

Generally, three appearance-based statistical methods, namely Principal Component Analysis (PCA) [4] [8] [14], Independent Component Analysis (ICA) [43] and Linear Discriminant Analysis (LDA) [14] [39], are widely

used for face recognition, because a two dimensional image has huge dimensionality and recognition would be computationally inefficient. Let us now briefly introduce these conventional projection methods. PCA finds a set of representative projection vectors, which has the largest variance among training data, such that the projected samples retain the most information about original samples. The most representative vectors are the eigenvectors corresponding to the largest eigenvalue of the covariance matrix. ICA captures both second and higher-order statistics and projects the input data onto the basis vectors that are as statistically independent as possible, while PCA deals with second-order statistics. LDA uses the class information and finds a set of vectors that Fisher discriminant criterion. It simultaneously maximizes the between-class scatter while minimizing the within-class scatter in the projective feature vector space. While PCA and ICA can be called unsupervised learning techniques, LDA is supervised learning technique because it needs class information for each image in the training process. LDA can enhance class separability of all sample images for classification purposes, but whenever the number of samples is less than the dimensionality of the samples, the scatter matrix may become singular, and the execution of LDA may encounter the so-called Small Sample Size Problem (S3 Problem), therefore transformation matrix can not be computed. The S3 Problem is often encountered when we use facial images in face recognition because of the high dimensionality. For example, a 64x64 monochrome image implies a feature space of 4,096 dimensions, therefore more than 4,096 samples are necessary to calculate the LDA projection matrix.

Due to the S3 problem, before LDA can be applied to reduce dimensionality, PCA is commonly used for dimensionality reduction: PCA+LDA [12] [14]. However, PCA step may extract nuisance dimensions, such as lighting conditions, and degenerate classification accuracy because of discarding

important discriminative dimensions. Therefore, direct LDA (DLDA) methods, which can accept a small number of high-dimensional data and optimizes Fisher's criterion directly without dimensionality reduction steps, were proposed to solve the S3 problem [34] [37].

Incidentally, Heteroscedastic Linear Discriminant Analysis (HLDA) [17] is becoming popular in state-of-the-art speech recognition systems. HLDA can separate the original feature space into two independent subspaces; useful dimensions and nuisance dimensions. However, calculating the transformation matrix is difficult due to high dimensionality and extreme sparseness of the data.

In this chapter, new two-dimensional algorithms, two-dimensional linear discriminant analysis (2DLDA) and two-dimensional heteroscedastic linear discriminant analysis (2DHLDA) are developed in order to find the most discriminant projection vectors for age-group classification. These methods overcome the singularity problem implicitly (S3 Problem of LDA), and reduce the influence of different lighting variations.

We briefly introduce conventional projection methods in Section 5.2, and we describe our proposed framework in Section 5.3. Experimental results for our method and conventional methods are discussed in Section 5.4, and conclusions are presented in Section 5.5.

5.2 Review of Previous Approaches

As facial images have very high dimensionality, we need efficient feature extraction before classification in order to robustly classify age-groups under various lighting conditions with speed. In this section, conventional projection methods based on an appearance-based approach are introduced. Many algorithms have been developed for the projection from high dimensional facial space to the lower dimensional space, namely PCA, LDA, and 2DPCA,

on which our new two-dimensional approaches (2DLDA and 2DHLDA) are based.

5.2.1 Formulation

Generally, a two dimensional image of size $h \times w$ pixels can be viewed as a vector in a high dimensional space. The easiest way to create a vector from an array is to concatenate its columns, thus getting a vector $\boldsymbol{x}$, where $\boldsymbol{x} \in \Re^d$ and $d = h \times w$.

Let the training set of n face images be $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. For the moment, we focus on linear dimensionality reduction, i.e., using a $d \times \tilde{d}$ transformation matrix $\boldsymbol{W}$, $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{W}. \tag{5.1}$$

5.2.2 Principal Component Analysis (PCA)

PCA [4] [8] [14] is a classical feature extraction widely used in the area of face recognition to reduce the dimensionality. PCA seeks to find the vectors that best describe the data in terms of reproducibility, however these vectors may not include the effective information for classification, and may eliminate discriminative information.

PCA aims to find the eigenvalues of the covariance matrix $\boldsymbol{C}$,

$$\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \tag{5.2}$$

where $\bar{\boldsymbol{x}}$ denotes the average of $\boldsymbol{x}_i$. One of the drawbacks of adapting PCA to face recognition is that the performance would be deleteriously affected by illumination changes.

5.2.3 Linear Discriminant Analysis (LDA)

LDA [14] [39] is a well-known technique for finding a set of projecting vector $\boldsymbol{W}_{\text{LDA}}$ best discriminating different classes. The within-class scatter matrix $\boldsymbol{S}_w$ and the between-class scatter matrix $\boldsymbol{S}_b$ are defined as below:

$$\boldsymbol{S}_w = \frac{1}{n}\sum_{j=1}^{c}\sum_{\boldsymbol{x}\in c_j}(\boldsymbol{x}-\bar{\boldsymbol{x}}_j)(\boldsymbol{x}-\bar{\boldsymbol{x}}_j)^T, \tag{5.3}$$

$$\boldsymbol{S}_b = \frac{1}{n}\sum_{j=1}^{c} n_i(\bar{\boldsymbol{x}}_j-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_j-\bar{\boldsymbol{x}})^T, \tag{5.4}$$

where n_i denotes the number of samples in class $c_j (j = 1, 2, \cdots, c)$, and n denotes the total number of samples. $\bar{\boldsymbol{x}}_j$ denotes the average of samples in c_j-th class, $\bar{\boldsymbol{x}}$ denotes the average of all training samples. One way to find the transformation matrix $\boldsymbol{W}_{\text{LDA}}$ is to use Fisher's criterion. It can be achieved by maximizing the ratio as shown in equation (5.5),

$$J(\boldsymbol{W}_{\text{LDA}}) = \frac{\boldsymbol{W}_{\text{LDA}}^T \boldsymbol{S}_b \boldsymbol{W}_{\text{LDA}}}{\boldsymbol{W}_{\text{LDA}}^T \boldsymbol{S}_w \boldsymbol{W}_{\text{LDA}}}. \tag{5.5}$$

$\boldsymbol{W}_{\text{LDA}}$ can be constructed by the set of largest eigenvalues of $\boldsymbol{S}_b\boldsymbol{S}_w^{-1}$. The maximum value of discriminative space is $c-1$, where c denotes the number of classes, since the rank of $\boldsymbol{S}_b$ is $c-1$.

5.2.4 Heteroscedastic Linear Discriminant Analysis (HLDA)

HLDA [17] is an extension of the simpler LDA method. Both LDA and HLDA try to find the best linear discriminant, but they differ in the underlying assumptions. LDA simplifies most practical problems by assuming the covariance matrices to be equal for all classes. The HLDA projection matrix,

$\boldsymbol{W}_{\text{HLDA}}$, for a $d-$dimensional feature space, $\boldsymbol{x}$, may be written as

$$\boldsymbol{y} = \boldsymbol{W}_{\text{HLDA}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{W}_p \\ \boldsymbol{W}_{d-p} \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{y}_p \\ \boldsymbol{y}_{d-p} \end{bmatrix}. \tag{5.6}$$

where the top p dimensions, $\boldsymbol{y}_p$, are deemed to be those dimensions that contain discriminatory information, the useful dimensions, and the final $(d-p)$-dimensions, $\boldsymbol{y}_{d-p}$, contain no useful information, the nuisance dimensions. HLDA transforms are trained using maximum likelihood (ML) estimation and the EM algorithm.

$$\boldsymbol{\mu}_j = \begin{bmatrix} \boldsymbol{\mu}_{j,1} \\ \vdots \\ \boldsymbol{\mu}_{j,p} \\ \boldsymbol{\mu}_{0,p+1} \\ \vdots \\ \boldsymbol{\mu}_{0,n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_j^p \\ \boldsymbol{\mu}_0^{(d-p)} \end{bmatrix}, \tag{5.7}$$

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_j^p & 0 \\ 0 & \boldsymbol{\Sigma}_0^{(d-p)} \end{bmatrix}. \tag{5.8}$$

Here, $\boldsymbol{\mu}_0^{(d-p)}$ is common to all the class means, and the $\boldsymbol{\mu}_j^p$ are different for each class. The $\boldsymbol{\Sigma}_j$ have also been partitioned in the corresponding manner, such that $\boldsymbol{\Sigma}_0^{(d-p)}$ is common for all the classes, whereas $\boldsymbol{\Sigma}_j^p$ are different for different classes.

$$P(\boldsymbol{x}_i) = \frac{|\boldsymbol{W}|}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{g(i)}|}} \exp\left(-\frac{(\boldsymbol{y}_i - \boldsymbol{\mu}_{g(i)})^T \boldsymbol{\Sigma}_{g(i)}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{g(i)})}{2} \right) \tag{5.9}$$

where $\boldsymbol{x}_i$ belongs to the group $g(i)$. In order to get the best estimator for $\boldsymbol{W}$, the log-likelihood of the data $L = \sum_{i=1}^{n} \log P(\boldsymbol{x}_i)$ under the linear transformation and under the constrained Gaussian model assumption for each class

is formed as

$$\log L = -\frac{1}{2}\sum_{i=1}^{n}\{(\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{\mu}_{g(i)})^T\Sigma_{g(i)}^{-1}(\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{\mu}_{g(i)}) + \log((2\pi)^d|\boldsymbol{\Sigma}_{g(i)}|)\} + n\log|\boldsymbol{W}|. \quad (5.10)$$

The above likelihood function can now be maximized with respect to its parameters. A straight-forward maximization with respect to various parameters can be a time consuming task. However the task can be considerably simplified by first calculating the optimal values of the mean and variance parameters in terms of the linear transformation $\boldsymbol{W}$. Differentiating the likelihood equation with respect to the parameter $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ and finding the point where the partial derivatives are zero, gives us the mean and variance estimates:

$$\hat{\boldsymbol{\mu}}_j^p = \boldsymbol{W}_p^T\bar{\boldsymbol{x}}_j, j = 1, \cdots, c, \quad (5.11)$$

$$\hat{\boldsymbol{\mu}}_0^{(d-p)} = \boldsymbol{W}_{n-p}^T\bar{\boldsymbol{x}}, \quad (5.12)$$

$$\hat{\boldsymbol{\Sigma}}_j^p = \boldsymbol{W}_p^T\boldsymbol{C}_j\boldsymbol{W}_p, j = 1, \cdots, c, \quad (5.13)$$

$$\hat{\boldsymbol{\Sigma}}_0^{(d-p)} = \boldsymbol{W}_{n-p}^T\boldsymbol{C}\boldsymbol{W}_{n-p}, \quad (5.14)$$

where

$$\bar{\boldsymbol{x}}_j = \frac{1}{n_j}\sum_{g(i)=j}\boldsymbol{x_i}, \quad (5.15)$$

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x_i}, \quad (5.16)$$

$$\boldsymbol{C}_j = \frac{1}{n_j}\sum_{g(i)=j}(\boldsymbol{x}_i - \bar{\boldsymbol{x}_j})(\boldsymbol{x}_i - \bar{\boldsymbol{x}_j})^T, \quad (5.17)$$

$$\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T. \quad (5.18)$$

Note that the $\boldsymbol{\mu}_j, j = 1, \cdots, c$ can be calculated if $\boldsymbol{W}$ is known, and $\boldsymbol{\Sigma}_j, j = 1, \cdots, c$ can be calculated if $\boldsymbol{\mu}_j, j = 1, \cdots, c$ and $\boldsymbol{W}$ are known, We obtain the likelihood of the data ($L(\boldsymbol{W}; \{\boldsymbol{x}_i\})$) in terms of $\boldsymbol{W}$ by substituting the values of the optimized $\boldsymbol{\mu}_j, j = 1, \cdots, c$ and $\boldsymbol{\Sigma}_j, j = 1, \cdots, c$ in equation (5.10). We can simplify ($L(\boldsymbol{W}; \{\boldsymbol{x}_i\})$) and then maximized with respect to $\boldsymbol{W}$ to give

$$\hat{\boldsymbol{W}} = \arg\max\Bigg\{ -\frac{n}{2}\log|\boldsymbol{W}_{d-p}^T \boldsymbol{C} \boldsymbol{W}_{d-p}| - \sum_{j=1}^{c} \frac{n_j}{2}\log|\boldsymbol{W}_p^T \boldsymbol{C}_j \boldsymbol{W}_p| + n\log|\boldsymbol{W}|\Bigg\}, \quad (5.19)$$

where $\hat{\boldsymbol{W}}$ is the estimate of the parameter $\boldsymbol{W}$. At this point one may choose to use only the first p columns of $\hat{\boldsymbol{W}}$ to obtain the best discriminating projection under the Gaussian model assumption.

5.2.5 2-Dimensional Principal Component Analysis (2DPCA)

Let $X_i \in \Re^{h \times w}$, for $i = 1, \cdots, n$ be the n images in the training dataset. As opposed to standard PCA, 2DPCA [49] is based on two-dimensional image matrices rather than one-dimensional vectors, and obtains higher recognition accuracy than PCA.

2DPCA projects an $h \times w$ random image matrix $\boldsymbol{X}$ onto $w \times \tilde{w}$ transformation matrix $\boldsymbol{W}_{\text{2DPCA}}$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_{\text{2DPCA}}, \quad (5.20)$$

where $\boldsymbol{Y}$ denotes a $h \times \tilde{w}$ feature matrix. Transformation matrix $\boldsymbol{W}_{\text{2DPCA}}$ is calculated by solving the maximization problem of $J(\boldsymbol{W}_{\text{2DPCA}})$:

$$J(\boldsymbol{W}_{\text{2DPCA}}) = tr(\tilde{\boldsymbol{C}}), \quad (5.21)$$

where $\tilde{\boldsymbol{C}}$ denotes the covariance matrix of the projected training data, and $tr(\tilde{\boldsymbol{C}})$ denotes the trace of $\tilde{\boldsymbol{C}}$. This maximization problem is equivalent to solving the eigenvalue problem of image covariance (scatter) matrix $\boldsymbol{G}$:

$$\boldsymbol{G} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}}). \tag{5.22}$$

5.2.6 2DPCA+PCA

2DPCA needs more coefficients to represent an image, and the dimension of the 2DPCA feature vector is always much higher. Thus, PCA is used for further dimensional reduction after 2DPCA, i.e. 2DPCA+PCA. They indicate that the performance of 2DPCA+PCA is still better than that of PCA only for the same dimensionality.

5.3 Proposed Method

5.3.1 2-Dimensional Linear Discriminant Analysis (2DLDA)

2DLDA, based on 2DPCA and LDA, is proposed in this section. The main difference between classical LDA and 2DLDA is data representation. LDA works with vectorized representations of data, while 2DLDA works with data in matrix representation. The within-class scatter matrix $\boldsymbol{S}_w$ and the between-class scatter matrix $\boldsymbol{S}_b$ are defined in 2DLDA as well as in LDA:

$$\boldsymbol{S}_w = \frac{1}{n}\sum_{j=1}^{c}\sum_{\boldsymbol{X}\in c_j}(\boldsymbol{X} - \bar{\boldsymbol{X}}_j)(\boldsymbol{X} - \bar{\boldsymbol{X}}_j)^T, \tag{5.23}$$

$$\boldsymbol{S}_b = \frac{1}{n}\sum_{j=1}^{c} n_j(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})^T, \tag{5.24}$$

where $\boldsymbol{X}$ denotes an $h \times w$ random image matrix, $\bar{\boldsymbol{X}}$ denotes the average of all training samples $\boldsymbol{X}_i$, and $\bar{\boldsymbol{X}}_j$ denotes the average of samples in c_j-th class. A $w \times \tilde{d}$ transformation matrix $\boldsymbol{W}_{\mathrm{2DLDA}}$ is calculated by equation (5.5). Projecting $\boldsymbol{X}$ onto $\boldsymbol{W}_{\mathrm{2DLDA}}$ yields an $h \times \tilde{d}$ feature matrix $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{2DLDA}}$.

5.3.2 2DLDA+LDA

In this section, two-phased approach 2DLDA+LDA is employed: 2DLDA is done for the first dimensionality reduction step such as from 32×32 to 32×10, and LDA is used for the second dimensionality reduction such as from 320 dimensions to less than 10 dimensions.

One advantage of the 2DLDA+LDA approach is that it can solve the S3 problem, since the transformation matrix is computable using a smaller amount of data as compared to LDA. For example, in the case of 32×32 monochrome images, when we derive 6 component vectors (32×6 features in total) using 2DLDA and the dimensionality is reduced from 192 to 10 using LDA, 192 images are sufficient, whereas more than 1,024 images are necessary using LDA only approach.

5.3.3 The Other 2DLDAs

After 2004, the other 2DLDAs [50] [56] [57] [58] [59] [60] were proposed independently at around the same time. They all treated the image data not as vectors but as matrices.

Ye's 2DLDA [50] aims to find the optimal transformation matrices $\boldsymbol{L}_k$ and $\boldsymbol{R}_k$ such that the class structure of the original high-dimensional space is preserved in the low-dimensional space. The algorithm is given in Table 5.1. They showed 2DLDA and the combination of 2DLDA and LDA, namely 2DLDA+LDA, where the dimension by 2DLDA is further reduced by LDA, is competitive with classical LDA in terms of classification accuracy.

Table 5.1: The procedure of Ye's 2DLDA; $\bar{\boldsymbol{X}}_j$ denotes the average of samples in class c_j

Ye's 2DLDA [50]	
input:	training data $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$, test data $\boldsymbol{X}$, reduced row dimension $\tilde{w}$, reduced column dimension $\tilde{h}$, the number of iterations k_{max}
output:	projected training data $\{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n\}$, projected test data $\boldsymbol{Y}$
Step 1:	$\boldsymbol{R}_0 = (I_{\tilde{w}}, 0)^T, k = 1$
Step 2:	Compute $\boldsymbol{S}_w^R = \sum_{j=1}^{n_c} \sum_{\boldsymbol{X}_i \in c_j} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)\boldsymbol{R}_{k-1}\boldsymbol{R}_{k-1}^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T$, $\boldsymbol{S}_b^R = \sum_{j=1}^{n_c} n_j(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})\boldsymbol{R}_{k-1}\boldsymbol{R}_{k-1}^T(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})^T$
Step 3:	Compute the first $\tilde{h}$ eigenvectors $\{\phi_l^L\}_{l=1}^{\tilde{h}}$ of $(\boldsymbol{S}_w^R)^{-1}\boldsymbol{S}_b^R$
Step 4:	$\boldsymbol{L}_k = [\phi_l^L, \cdots, \phi_{\tilde{h}}^L]$
Step 5:	Compute $\boldsymbol{S}_w^L = \sum_{j=1}^{n_c} \sum_{\boldsymbol{X}_i \in c_j} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T\boldsymbol{L}_{k-1}^T\boldsymbol{L}_{k-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)$, $\boldsymbol{S}_b^R = \sum_{j=1}^{n_c} n_j(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})^T\boldsymbol{L}_{k-1}^T\boldsymbol{L}_{k-1}(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})$
Step 6:	Compute the first $\tilde{w}$ eigenvectors $\{\phi_l^R\}_{l=1}^{\tilde{w}}$ of $(\boldsymbol{S}_w^L)^{-1}\boldsymbol{S}_b^L$
Step 7:	$\boldsymbol{R}_k = [\phi_l^R, \cdots, \phi_{\tilde{w}}^R]$
Step 8:	If $k < k_{max}$, $k = k + 1$ and go to Step 2, else output $\boldsymbol{Y} = \boldsymbol{L}_k^T\boldsymbol{X}\boldsymbol{R}_k$

Li et al. [56], Xiong et al. [57] and Chen et al. [60] proposed the exact same algorithm as ours described in Section 5.3.1. They named this algorithm 2D-LDA, two-dimensional Fisher discriminant analysis (2DFDA) and MatFLDA, respectively.

Yang et al. [58] also developed 2DLDA. However they applied projections twice: the first one is in a horizontal direction and the second is in a vertical direction. Specifically, given image $\boldsymbol{X}$, we obtain its feature matrix $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{R}$. Then, we transpose $\boldsymbol{Y}$ and determine the transformation matrix $\boldsymbol{L}$ using the same manner. Projecting $\boldsymbol{Y}^T$ onto $\boldsymbol{L}$, we obtain $\boldsymbol{Z}^T = \boldsymbol{Y}^T\boldsymbol{L}$. The resulting feature matrix is $\boldsymbol{Z} = \boldsymbol{L}^T\boldsymbol{Y}$.

Kong et al. [59] proposed 2D Fisher Discriminant Analysis (2D-FDA) containing Unilateral 2D Fisher Discriminant Analysis (U2D-FDA) and Bilateral 2D Fisher Discriminant Analysis (B2D-FDA). They considered a left-multiplying U2D-FDA

$$\boldsymbol{Y} = \boldsymbol{L}^T \boldsymbol{X} \tag{5.25}$$

and right-multiplying U2D-FDA.

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{R} \tag{5.26}$$

After performing the left- and right-multiplying U2D-FDA, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ were obtained for each image: They were combined together for recognition. The steps for recognition were as follows: firstly $\boldsymbol{Y}$ and $\boldsymbol{Z}$ were transformed into 1D vectors for each image, then PCA was applied onto these vectors. Finally, two shorter vectors were combined into one vector for classification.

5.3.4 2DHLDA

HLDA [17], viewed as a generalization of LDA, tries to find the best linear discriminant, and removes the restriction that all the within-class covariance matrices are the same. We propose 2DHLDA, which is an extension of the 2DLDA method and assumes the covariance matrices to be different for all classes as well as in the case of HLDA. The 2DHLDA projection matrix $\boldsymbol{W}_{\text{2DHLDA}}$ is written as

$$\boldsymbol{Y} = \boldsymbol{W}_{\text{2DHLDA}}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{W}_p \\ \boldsymbol{W}_{d-p} \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} \boldsymbol{Y}_p \\ \boldsymbol{Y}_{d-p} \end{bmatrix}. \tag{5.27}$$

where $\boldsymbol{W}_p$ is a matrix consisting of the first p of matrix $\boldsymbol{W}_{\text{2DHLDA}}$ and $\boldsymbol{W}_{d-p}$ consists of the remaining $d-p$ rows, the top p dimensions contain discrim-

inatory information, the useful dimensions, and the final $d-p$ dimensions contain the nuisance dimensions.

The optimal transformation matrix is calculated by maximizing the log-likelihood Gaussian function in [17]. The final solution can be obtained as

$$\hat{\boldsymbol{W}} = \arg\max\Bigg\{ -\frac{n}{2}\log|\boldsymbol{W}_{d-p}^T\boldsymbol{C}\boldsymbol{W}_{d-p}| - \sum_{j=1}^{c}\frac{n_j}{2}\log|\boldsymbol{W}_p^T\boldsymbol{C}_j\boldsymbol{W}_p| + n\log|\boldsymbol{W}|\Bigg\}, \quad (5.28)$$

where

$$\boldsymbol{C}_j = \frac{1}{n_j}\sum_{g(i)=j}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T, \quad (5.29)$$

$$\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T, \quad (5.30)$$

$$\bar{\boldsymbol{X}}_j = \frac{1}{n_j}\sum_{g(i)=j}\boldsymbol{X}_{\boldsymbol{i}}, \quad (5.31)$$

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{\boldsymbol{i}}. \quad (5.32)$$

5.3.5 2DHLDA+LDA

A two-phased approach 2DHLDA+LDA is also employed: 2DHLDA is done for the first dimensionality reduction step from 32×32 to 32×10, and LDA is used for the second dimensionality reduction from 320 dimensions to 4 dimensions.

Figure 5.1: Sample images from WIT-DB

5.4 Experiments for Age-group Classification

5.4.1 Database

In this chapter, 26,222 images (14,214 male and 12,008 female images) are used for testing. Figure 5.1 shows an example of the facial image used in our study. The image size used here is a facial region of 32x32 pixels. We separated age-groups into 11 classes, and our goal is to classify 11-class age-groups with high accuracy. Table 5.2 shows the number of data in each age-group class. These age-groups are based on actual age, and not appearance age. As we mentioned in the previous chapter, these images were taken from subjects in a wide variety of lighting conditions and age-groups, from 3 to 85 years of age, because our research is motivated by real-world application that must be robust against lighting variations.

5.4.2 Outline of Experiments

12,008 female data and 14,214 male data are individually treated in this experiment, and the performance of age-group classification is evaluated using 2-fold cross validation in each gender: approximately 6,000 female images are used for training and the remaining images for tests, and by repeating the same process by reversing, all the data are evaluated (same approach for male data).

We use the new proposed methods, 2DLDA+LDA and 2DHLDA+LDA, in order to reduce dimensionality for feature extraction, and also use PCA

Table 5.2: The number of images used in this chapter.

class	age-group	# of female images	# of male images
1	3-14	1,749	2,211
2	15-19	2,060	3,205
3	20-24	768	1,184
4	25-29	810	1,105
5	30-34	866	748
6	35-39	1,070	1,057
7	40-44	1,257	960
8	45-49	1,000	974
9	50-54	845	885
10	55-59	758	922
11	60-85	825	963
total		12,008	14,214

and LDA for comparison purposes. Training data and test data are plotted in low-dimensional space to verify the position of each age-group data.

After the dimensionality reduction, we construct Gaussian models in low-dimensional space, and make classifications by comparing the output likelihoods from each model. Eleven Gaussian models from each 11 age-groups are estimated in a low-dimensional 2DLDA+LDA (or 2DHLDA+LDA) feature space. The Gaussian model is a type of probability distribution model. The d-dimensional normal distribution density function is defined as

$$P(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})], \tag{5.33}$$

where the vector $\boldsymbol{\mu}$ is the mean of the normal distribution, and the matrix $\boldsymbol{\Sigma}$ is the variance-covariance matrix. Gaussian components are estimated using the Expectation Maximization (EM) algorithm. The age-group classification using test images is done based on likelihoods from each age-group model after the process of 2DLDA+LDA (or 2DHLDA+LDA) projection.

The likelihood scores for each class are computed and the class with the highest likelihood is chosen. For the reason of the difficulty in age-group classification, the classification rate in the **10-year range**, which includes the contiguous class with the higher likelihood, and in the **15-year range**, which includes both contiguous classes, are observed.

5.4.3 A Projection Example

To confirm the age-group classification ability of PCA, LDA, and 2DLDA+LDA, projected training data and test data are plotted: 1st dimension (x-axis), 2nd dimension (y-axis), 3rd dimension (x-axis), 4th dimension (y-axis), and so on.

In the case of using 2DLDA+LDA (32×32 to 32×10 by 2DLDA), Figure 5.2 shows the first 6 dimension of projected training data (on the left) and test data (on the right); from the top graph to the bottom graph, x-axis and y-axis represent 1st and 2nd, 3rd and 4th, and 5th and 6th dimensions. Training data is separated in 5th and 6th dimensions to some degree, whereas separable classes are difficult to find in the test data. Test data seems to be separable in the first 4 dimensions. For the reason of separability, the first 4 dimensions are used in the experiment in addition to using all the dimensions. Furthermore, the point to observe is that the graphs in the first 2 dimensions line up in order of age-groups from top-left to middle-left. Graphs in the first 2 dimensions using LDA are quite similar to ones in 2DLDA+LDA. On the other hand, graphs in the PCA feature space are different from ones in the 2DLDA+LDA feature space: it seems to be difficult to classify age-groups using not only the test data, but also the training data.

5.4.4 Experimental Results

Figure 5.3, 5.4 and 5.5 show classification accuracy using PCA, LDA, 2DLDA+LDA and 2DHLDA+LDA, when age-groups are in the 5-year, 10-year, and 15-year range respectively. PCA dimension is set to 4, 10 and 50, LDA dimension is set to 4, 10. In 2DLDA+LDA, 2DLDA dimension is set to 32×10 and LDA dimension is set to 4 and 10. In 2DHLDA+LDA, 2DHLDA dimension is set to 32×10 and LDA dimension is set to 4.

Experimental results verify high efficiency of our approach, 2DLDA+LDA and 2DHLDA+LDA. Figure 5.2 shows that in LDA and 2DLDA+LDA, using the first 4 dimensions is better suited for making Gaussian models than by using 10 dimensions. In every method, classification rates in male are higher than the ones in female, and the result shows classifying age-group using female images to be difficult.

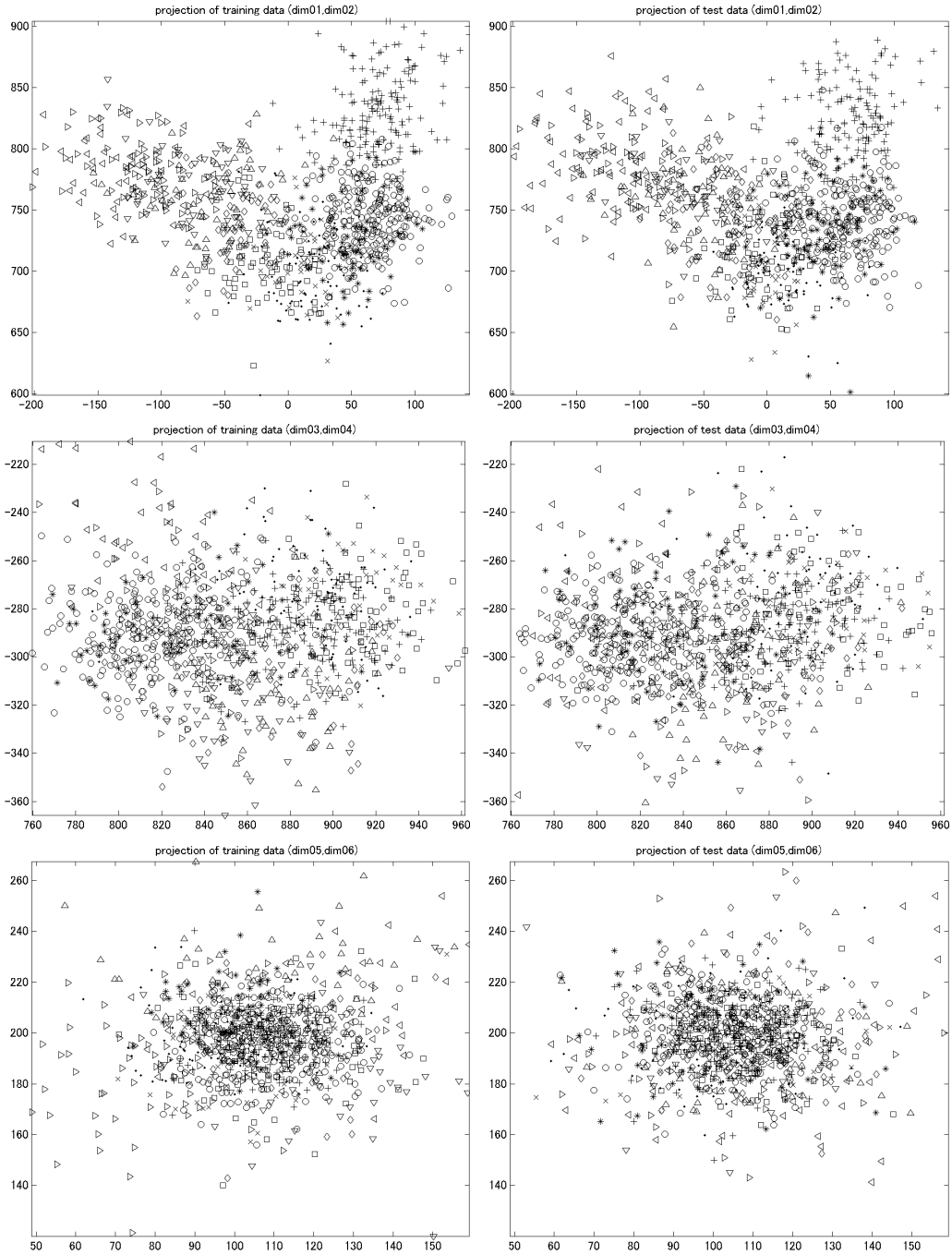
Moreover, Figure 5.6 shows classification accuracy using 2DLDA+LDA, which is superior to other methods. The row denotes the class based on the actual age, and the column denotes the class based on experiment, and the class with the higher classification rate is darker in color. In terms of younger age-groups (under 19) and older age-groups (over 50), classification rates are higher in each gender, however as for age-groups between 20 and 49, especially in females, classification decreases as shown in the figure.

5.5 Conclusion

In this chapter, we proposed two-phased approaches (2DLDA+LDA and 2DHLDA+LDA) for age-group classification using facial images under various lighting conditions. Our approach does not require PCA, which extracts lighting condition variations, and solves the S3 problem under a small amount of samples. Additionally our experiments showed that 2DLDA+LDA

and 2DHLDA approaches are superior to that using only LDA under a large amount of samples. Then, effective feature extraction, which is not for lighting condition variation but for the age-group classification, was achieved. In addition, highly discriminative Gaussian model classifiers made by using only the first 4 dimensions were effective for classification.

For future work, we plan to solve some familiar problems, such as viewing orientation, partial occlusion of facial features and facial expression. Moreover, we plan to evaluate the classification rates using images under unknown lighting conditions.



+:class 1, ◯:class 2, ∗:class 3, ●:class 4, ×:class 5, □:class 6,
◊:class 7, △:class 8, ▽:class 9, ▷:class 10, ◁:class 11

Figure 5.2: The projected data using 2DLDA+LDA (male).

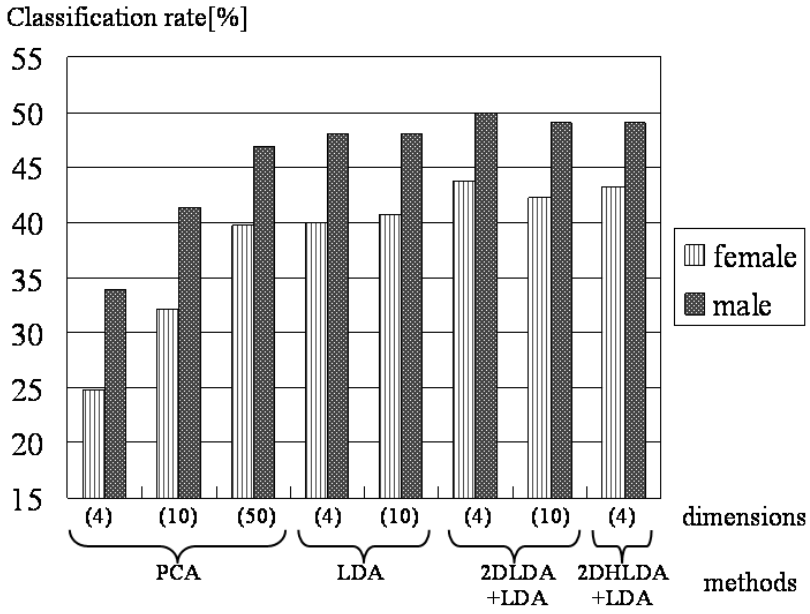


Figure 5.3: The age-group classification rates (within the 5-year range) based on different projection methods.

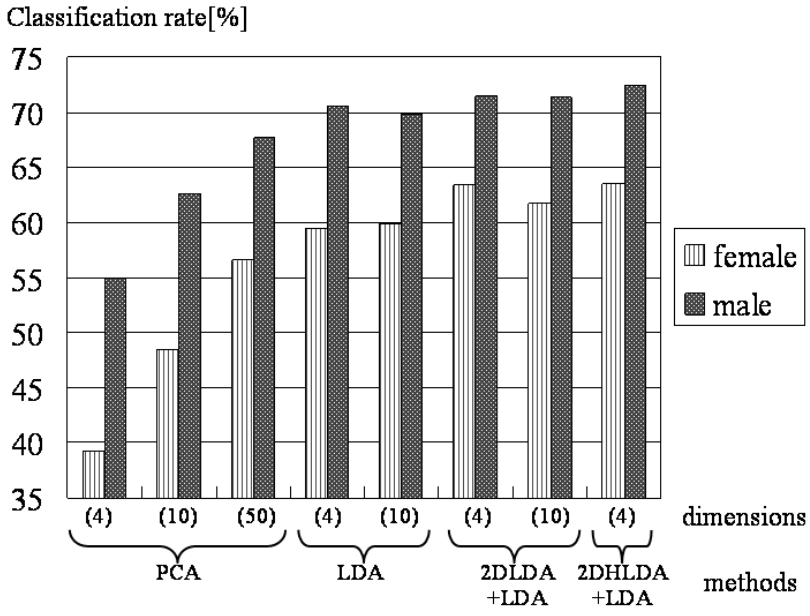


Figure 5.4: The age-group classification rates (within the 10-year range) based on different projection methods.

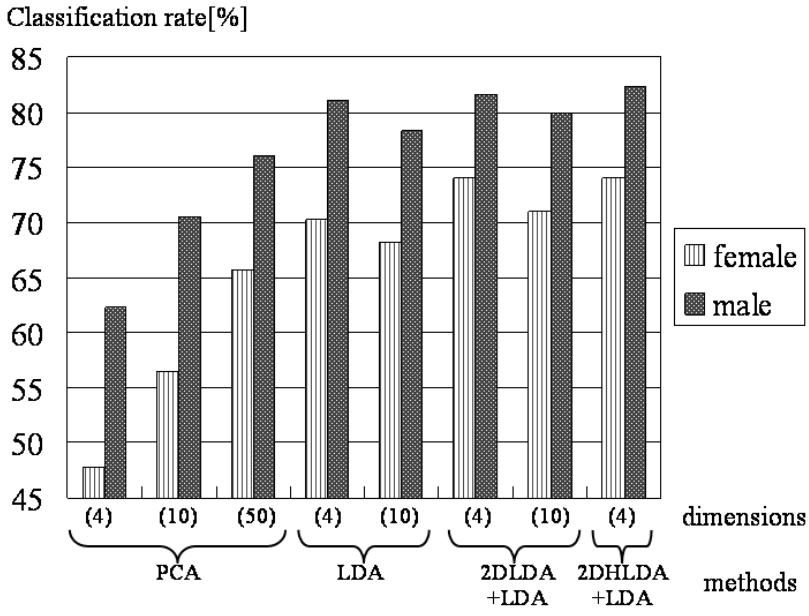


Figure 5.5: The age-group classification rates (within the 15-year range) based on different projection methods.

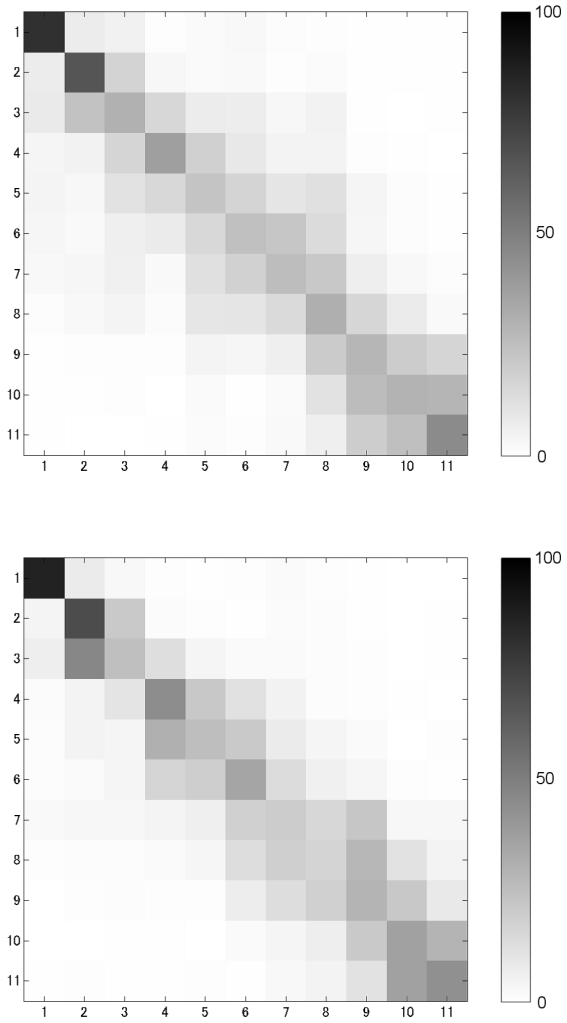


Figure 5.6: The confusion matrix of age-group classification rates based on 2DLDA+LDA. (top: female data, bottom: male data)

Chapter 6

Comparison of Age-Group Classification Performance Using Actual-age-based Data and Perceived-age-based Data

6.1 Abstract

From the difficulties in age-group classification described in the previous chapter, data inconsistency and contradictions between classes will be especially focused on here.

When the previously described pattern recognition techniques are used, one of the difficulties in classifying age-groups is that the training data includes inconsistency and contradictions. For instance, in the case of distinguishing between people in their 20's and 30's, adding people who look older than 29 to a 20's-based model and people who look younger than 30 to a 30's-based model makes classifiers confused and incapable of properly classifying subjects.

In order to solve this problem, we used perceived ages judged by human observers. Two types of experiments for age-group classification based on actual age and perceived age were performed using facial images.

Firstly, system performance and the performance of human observers are compared, to investigate whether the system can obtain equivalent accuracy rates to human. Secondly, perceived-age-based training data are used instead of actual-age-based training data to decrease the data contradictions between classes as much as possible. Then, we determine which training data (actual-age-based or perceived-age-based) can achieve the best class separability and make it easier to classify age-groups.

The remainder of this chapter is organized as follows: In Section 6.2, our database and the definition of the perceived age are described. Age-group classification experiments using actual-age-based data and perceived-age-based data are developed in Section 6.3. Finally, discussion and conclusions are presented in Section 6.4.

6.2 Evaluation Methods

6.2.1 Giving a Perceived Age

We asked six people to look at the sample images and give a perceived age in order to evaluate performance using not only the actual age but also the perceived age. Several facial images from various angles for one subject are observed, and one-year range of perceived ages are assigned not to images but to subjects. Image samples used for judging include mainly the facial region, but some samples also include the upper body region as shown in Figure 2.1. Table 6.1 shows the number of images sorted based on actual and perceived ages.

6.2.2 Age-group Classification Algorithm

Our age-group classification scheme is shown in Figure 6.1. The size of each facial image used in this chapter is 64x64 pixels, with 256 grey levels per

Table 6.1: The number of images based on actual and perceived ages

(a) Females

Class no.	Age-group	Actual age		Perceived age	
		The # of person	The # of samples	The # of person	The # of samples
1	3-14	364	1,746	351	1,675
2	15-19	349	2,060	318	1,869
3	20-24	152	763	176	942
4	25-29	197	813	188	772
5	30-34	192	866	254	1,153
6	35-39	227	1,070	259	1,208
7	40-44	256	1,253	224	1,096
8	45-49	198	1,000	168	867
9	50-54	163	845	164	802
10	55-59	148	758	159	862
11	60-85	154	825	139	753
Total		2,400	11,999	2,400	11,999

(b) Males

Class no.	Age-group	Actual age		Perceived age	
		The # of person	The # of samples	The # of person	The # of samples
1	3-14	461	2,205	432	2,063
2	15-19	589	3,205	538	2,874
3	20-24	224	1,184	241	1,382
4	25-29	286	1,105	231	936
5	30-34	198	748	262	1,020
6	35-39	262	1,057	253	1,018
7	40-44	191	960	223	1,066
8	45-49	187	974	204	1,025
9	50-54	167	885	191	1,017
10	55-59	166	922	141	781
11	60-85	157	963	172	1,026
Total		2,888	14,208	2,888	14,208

pixel, as shown in Figure 2.2. First of all, histogram equalization is applied to all the images in order to eliminate illumination changes. In terms of

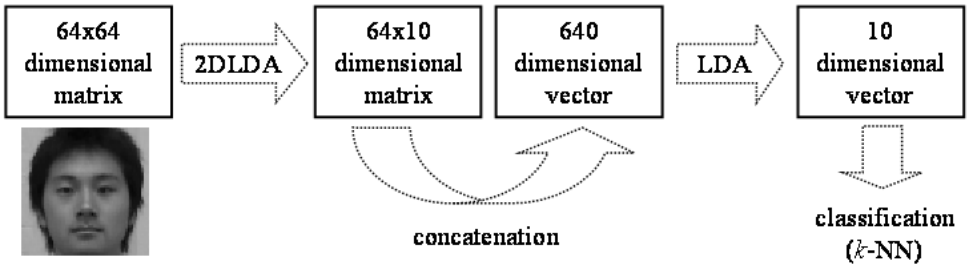


Figure 6.1: Age-group classification scheme

feature extraction (projection onto a low dimensional space), 2DLDA+LDA, which showed better performance in the previous chapter, is used. The first projection 2DLDA reduces the dimensionality from a 64x64 matrix to a 64x10 matrix. After concatenating the resulting matrix column by column (or row by row), we are able to obtain a 640 dimensional vector. The second projection LDA also reduces the dimensionality from 640 to 10. Finally, k-Nearest Neighbor classifiers (kNN) are used for the classification with $k = 100$.

6.2.3 Average Error Distance

In addition to classification rates for test data $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$, the **average error distance** is defined as

$$d_{jk} = \frac{1}{n} \sum_{i=1}^{n} |j - k| \tag{6.1}$$

and evaluated to measure the degree of the error. The variable j denotes the class index based on actual age, and k denotes the class index determined by experiments. The farther the age-group class chosen, the farther the distance is. For example, when a person in their early 20's (20 - 24) is misclassified, being classified as a person in their late 40's (45 - 49), this is more serious

than the same subject being classified as a person in their late 20's (25 - 29), so that the average error distance is larger. Not only the error rates, but also the degree of the errors can be calculated by using these distances.

6.3 Age-group Classification Experiments

6.3.1 Classification Performance Comparison between System Evaluations and Human Evaluations

We analyze the classification accuracy achieved by six observers in order to check if the system performance is comparable to human evaluations. Classification rates in every age-group are shown in Figure 6.2, average classification rates and average error distances over the whole range of age-groups are shown in Table. 6.2.

Table 6.2: Classification rates and average error distances between classes using all image samples

	Females		Males	
	Accuracy rates	Average error distances	Accuracy rates	Average error distances
Observer-1	47.54	0.8392	54.95	0.6080
Observer-2	44.17	0.8154	49.45	0.6898
Observer-3	48.04	0.7238	57.76	0.5222
Observer-4	46.29	0.8029	51.87	0.6766
Observer-5	50.00	0.7433	58.62	0.5419
Observer-6	44.54	0.8142	52.35	0.6243
Our system	42.20	1.2473	48.53	0.8934

Figure 6.2 shows that our system performance and human performance have higher rates when subjects are below 19 or over 60 years old, but most accuracy rates are less than 50% in the rest of the age-groups. Table 6.2 also shows that recognizing a female's age-group is much more difficult than recognizing a male's age-group in both cases of system and human evaluations.

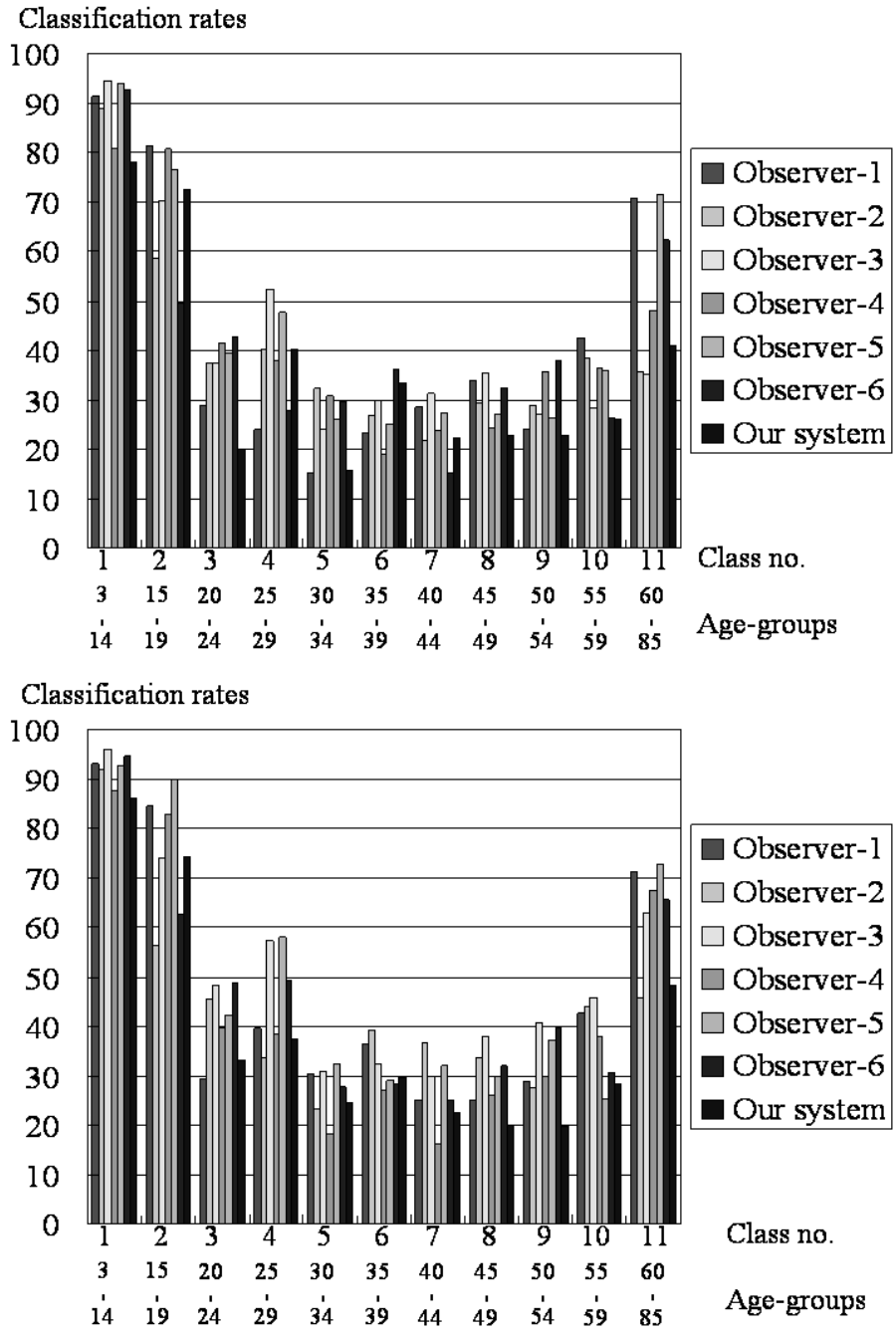


Figure 6.2: Classification rates in each age-group (accuracy based on human eyes and system) using female data (top) and male data (bottom)

From the analysis of whole experiments, our system performance cannot achieve higher accuracy than human performance, but can get relatively equivalent accuracy. However, experimental results on average error distances show that our system performance is in fact lower, and includes obvious errors that do not appear in human evaluations. This is because the human observers have at least two advantages. One of these advantages is that they have the opportunity to view not only facial images but also the upper body. Additionally, various angles of the face can be observed when the perceived ages are given, whereas only one facial image is used to judge age-groups in our system. In order to solve this problem, we expect that using the parts of upper body such as clothes, in addition to faces, will improve our performance as we proposed in the previous chapter.

6.3.2 Accuracy Comparison between Actual-Age-Based Training Data and Perceived-Age-Based Training Data

We construct perceived-age-based training data instead of actual-age-based training data and compare the results of the two methods in order to reduce the unbalanced data and improve separability between classes. Average perceived ages given by six observers are used in our experiment as shown below.

(Evaluation-1) Training data and test data are configured with actual ages

(Evaluation-2) Training data and test data are configured with perceived ages

(Evaluation-3) Training data is configured with perceived ages, and test data is configured with actual ages

(Evaluation-4) Training data is configured with actual ages, and test data is configured with perceived ages

Table 6.3: Classification rates and average errors between classes when training data and test data are based on actual age and perceived age

(a) Females

	Accuracy rates [%]			Average
	5-year range	10-year range	15-year range	error distances
Evaluation-1	42.20	60.51	69.34	1.2473
Evaluation-2	39.36	61.21	71.21	1.1716
Evaluation-3	38.60	59.81	70.66	1.2157
Evaluation-4	38.04	57.56	67.85	1.2807

(b) Males

	Accuracy rates [%]			Average
	5-year range	10-year range	15-year range	error distances
Evaluation-1	48.53	71.27	80.10	0.8934
Evaluation-2	49.51	74.11	82.88	0.8220
Evaluation-3	47.89	72.14	80.47	0.8969
Evaluation-4	47.68	70.78	80.37	0.8846

(c) Totals

	Accuracy rates [%]			Average
	5-year range	10-year range	15-year range	error distances
Evaluation-1	45.63	66.34	75.17	1.0554
Evaluation-2	44.87	68.21	77.54	0.9821
Evaluation-3	43.64	66.49	75.98	1.0429
Evaluation-4	43.26	64.73	74.64	1.0659

Evaluation-1 is a normal actual-age-based experiment. Evaluation-2 is a perceived-age-based experiment, and better clustering and separability can be expected using perceived ages compared to actual-age-based methods. Evaluation-3 is expected to give equal separability to evaluation-2, and is used in situations where there are many data samples, but the subjects' ages are unknown. Evaluation-4 cannot be used in every situation. However, it is useful in cases where there are too many sample images and their perceived ages are needed.

Evaluations are carried out using the accuracy rate and the average error distance in every age-group. As for accuracy in every age-group, k-nearest neighbor classifiers are used and three ranges of accuracy rates are computed: the classification rate in the 5-year range, which only includes a class with the maximum number of nearest neighbors, in the 10-year range, which includes the contiguous class with the larger number of nearest neighbors, and in the 15 year range, which includes both contiguous classes.

Experimental results are shown in Table 6.3. These four evaluation methods have almost the same performance in terms of classification accuracy, but one that uses perceived ages can reduce the average error distances and have good separability, whereas one that uses actual ages cannot.

6.4 Discussion and Conclusions

In this chapter, two types of evaluations were carried out using perceived age based on observations by six people. Firstly, we compared the system performance to the person's performance using actual age. Secondly, we compared actual-age-based classifiers to perceived-age-based classifiers in order to reduce the number of contradictory data samples between classes and to enhance class separability.

Our experiments showed that system performance was about as high as that which was achieved by observers in terms of classification accuracy, but still had room for improvement in terms of average error distances; there were obvious errors in our system that did not appear in human evaluations. Our experiments also showed that better clustering could be performed and average error distances could be reduced using perceived-age-based data instead of actual-age-based data. These results show that using only consistent data samples gives better performances in cases of 2DLDA or LDA projection. In real-world installations such as convenience stores, shop clerks' cus-

tomer analyses at cash registers heavily rely on the customer's perceived age, since their actual ages are unknown. We found that it was better to construct training data not by using actual ages but by using perceived ages. Moreover, when there are many training samples without actual ages given, constructing the training data based on perceived age is generally successful and gives an approximately equivalent performance to that based on actual age. In addition, we confirmed that the results differed only slightly when there were many training samples based on actual age, so therefore giving perceived age was not necessarily required.

Chapter 7

Age-group Classification based on Multiple Two-Dimensional Feature Extraction Algorithms

7.1 Introduction

In this chapter, we focus on the fusion of multiple two-dimensional feature extraction algorithms to improve the age-group classification accuracy.

As for the appearance-based approaches in face recognition, 2DPCA [49] and 2DLDA [50] have been proposed in recent years, and these have shown better performances than conventional PCA or LDA [12] [14]. In this chapter, these two-dimensional approaches are adopted for the age-group classification. There are several variations to use the two-dimensional approaches. However, all of the previous approaches only focused on the use of one variation and had no interest for the combination of these variations. We found that the different two-dimensional approaches gave different types of errors even by extracting from one sample image. For all these reasons, a fusion-based age-group classification method, which combines multiple classifiers using different two-dimensional feature extraction methods from one sample image, is proposed in order to reduce the error rates in classifying age-groups. This idea is based on the theory that fusion of multiple data can achieve bet-

ter performance when they can overcome the shortcomings of each other. It has been shown theoretically that a more accurate classifier can be obtained by combining multiple classifiers [20]. These information fusion techniques are usually used in the multimodal systems [28] [46], for example, recognizing a person by using facial image and voice, recognizing gender using facial image and body size information, and so on. However, even for the single source systems which use only one source originally, if we can prepare multiple classifiers with variations, which have different types of errors, we can improve the performance by combining them.

Bagging, the other classifier combination approach, makes variations in classifiers by giving perturbation in the training dataset of each classifier [13]. Boosting introduces weight for each training sample and makes variations by changing the weight set of each classifier [29]. In this chapter, we try to make variations in classifiers by changing the analysis methods applied to the source, especially in terms of two-dimensionality of facial data analysis.

In our approach, firstly, the row and column directions of two-dimensional projections (2DPCA or 2DLDA) have been done, and PCA or LDA is used for further dimensional reduction after the first two-dimensional projection, i.e. 2DPCA+PCA and 2DLDA+LDA. (Here, the notation "A+B" represents "apply method A then method B".) Multiple classifiers from each direction are made in lower dimensional feature space. Two types of normalization methods and four types of fusion methods are used for integrating different kinds of information. Multiple scores from multiple classifiers using 2DPCA+PCA and 2DLDA+LDA are also combined to achieve better classification accuracy. Finally, experiments are conducted to compare the performance of the proposed method with existing methods and to evaluate the effectiveness of our method.

The rest of this chapter is organized as follows: Section 7.2 reviews the

two-dimensional algorithms; Section 7.3 presents our proposed method; our experiments are introduced in Section 7.4; We conclude in Section 7.5.

7.2 Two-dimensional Algorithms

There are many ways to reduce the dimensionality. In recent years, 2DPCA [49] and 2DLDA [50] became prevalent because of the better performance compared to conventional PCA or LDA. The nonlinear feature extraction methods, such as Kernel Principal Component Analysis (KPCA) [21] and Kernel Fisher Discriminant Analysis (KFDA) [30] are also used. However, one of the drawbacks of kernel methods is too time consuming not only in the training process, but also in the classification process especially for a large data set without the sparse algorithms. For these reasons, in this section, we focus on multiple two-dimensional projections in the row and the column directions by using 2DPCA and 2DLDA.

Here, 2DPCA and 2DLDA, which are used in our experiments, are briefly introduced.

7.2.1 Row-based 2DPCA (R-2DPCA) and Column-based 2DPCA (C-2DPCA)

Yang's 2DPCA [49], referred as row-based 2DPCA in this chapter, uses row-directional base vectors. That means the Yang's 2DPCA is antisymmetric in the treatment of the rows and the columns of images. Thus, we prepare another 2DPCA, column-based 2DPCA which uses column-directional base vectors. We expect this slight difference produces different error tendencies between two classifiers with respective 2DPCA, and also expect the combination of them improves the performance.

In Yang's 2DPCA shown in Table 7.1, the eigenvectors of $\boldsymbol{C}^{(r)}$ only reflect the information between rows of images, in other words, 2DPCA only

works in the column directional reduction. We call this method "Row-based 2DPCA (R-2DPCA)". We also prepare "Column-based (C-2DPCA)", which uses the column direction of images. These methods are shown in Table 7.1 and 7.2. Here, $\bar{\boldsymbol{X}}$ denotes the average of all training samples, and $\bar{\boldsymbol{X}}_j$ denotes the average of samples in class c_j. In all our experiments, PCA is used for further dimensional reduction after 2DPCA, i.e. 2DPCA+PCA. R-2DPCA+PCA and C-2DPCA+PCA will be noted as method 1 and 2, respectively, and used them in the following chapter.

Table 7.1: The procedure of Row-based 2DPCA (R-2DPCA) (J. Yang's 2DPCA [49])

Method 1: Row-based 2DPCA (R-2DPCA)	
input:	training data $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$, test data $\boldsymbol{X}$, reduced row dimension $\tilde{w}$
output:	projected training data $\{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n\}$, projected test data $\boldsymbol{Y}$
Step 1:	Compute image covariance matrix $\boldsymbol{C}^{(r)} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)$
Step 2:	Compute the first $\tilde{w}$ eigenvectors $\{\phi_l^{(r)}\}_{l=1}^{\tilde{w}}$ of $\boldsymbol{C}^{(r)}$, $\boldsymbol{W}_{\text{2DPCA}}^{(r)} \leftarrow [\phi_1^{(r)}, \cdots, \phi_{\tilde{w}}^{(r)}]$
Step 3:	Project test image $\boldsymbol{X}$ onto $\boldsymbol{W}_{\text{2DPCA}}^{(r)}$ yields an h by $\tilde{w}$ matrix $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_{\text{2DPCA}}^{(r)}$

7.2.2 Row-based 2DLDA (R-2DLDA) and Column-based 2DLDA (C-2DLDA)

Ye's 2DLDA [50] shown in Table 5.1 used the projections in the row and column directions at the same time, whereas we prepare two different 2DLDA, row-based 2DLDA and column-based 2DLDA, in which the treatment of rows and columns of images are antisymmetric, in order to produce different error tendencies.

Table 7.2: The procedure of Column-based 2DPCA (C-2DPCA)

Method 2: Column-based 2DPCA (C-2DPCA)	
input:	training data $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$,
	test data $\boldsymbol{X}$,
	reduced column dimension $\tilde{h}$
output:	projected training data $\{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n\}$,
	projected test data $\boldsymbol{Y}$
Step 1:	Compute image covariance matrix
	$\boldsymbol{C}^{(c)} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T$
Step 2:	Compute the first $\tilde{h}$ eigenvectors $\{\phi_l^{(c)}\}_{l=1}^{\tilde{h}}$ of $\boldsymbol{C}^{(c)}$,
	$\boldsymbol{W}_{\text{2DPCA}}^{(c)} \leftarrow [\phi_1^{(c)}, \cdots, \phi_{\tilde{h}}^{(c)}]$
Step 3:	Project test image $\boldsymbol{X}$ onto $\boldsymbol{W}_{\text{2DPCA}}^{(c)}$ yields
	an $\tilde{h}$ by w matrix $\boldsymbol{Y} = \boldsymbol{W}_{\text{2DPCA}}^{(c)}\boldsymbol{X}$

One disadvantage of Ye's 2DLDA is the difficulty of finding the optimal $\boldsymbol{L}_k$ and $\boldsymbol{R}_k$ simultaneously. They derive an iterative algorithm. The optimal $\boldsymbol{L}_k$ is computed for a fixed $\boldsymbol{R}_k$, then best $\boldsymbol{R}_k$ is computed for a fixed $\boldsymbol{L}_k$ and this procedure is repeated for a series of rounds.

On the contrary, we define Row-based 2DLDA (R-2DLDA), Column-based 2DLDA (C-2DLDA) as shown in Table 7.3 and 7.4. In order to improve the classification accuracy, we project images into the row and the column directions separately, and construct classifiers in each direction. In all our experiments, LDA is used for further dimensional reduction after 2DLDA, i.e. 2DLDA+LDA. R-2DLDA+LDA and C-2DLDA+LDA will be noted as method 3 and 4, respectively, and used them in the following chapter.

7.3 Our Proposed Method

We propose a new method for age-group classification based on one facial image. The most important part of our strategy is to extract different types of features in consideration of fusion by using R-2DPCA, C-2DPCA, R-2DLDA

Table 7.3: The procedure of Row-based 2DLDA (R-2DLDA)

Method 3: Row-based 2DLDA (R-2DLDA)	
input:	training data $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$,
	test data $\boldsymbol{X}$, reduced row dimension $\tilde{w}$
output:	projected training data $\{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n\}$,
	projected test data $\boldsymbol{Y}$
Step 1:	Compute image within-class scatter matrix
	$\boldsymbol{S}_w^{(r)} = \frac{1}{n}\sum_{j=1}^{n_c}\sum_{\boldsymbol{X}_i \in c_j}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)$,
	and image between-class scatter matrix
	$\boldsymbol{S}_b^{(r)} = \frac{1}{n}\sum_{j=1}^{n_c} n_j(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})^T(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})$
Step 2:	Compute the first $\tilde{w}$ eigenvectors $\{\phi_l^{(r)}\}_{l=1}^{\tilde{w}}$ of $(\boldsymbol{S}_w^{(r)})^{-1}\boldsymbol{S}_b^{(r)}$,
	$\boldsymbol{W}_{\text{2DLDA}}^{(r)} \leftarrow [\phi_1^{(r)}, \cdots, \phi_{\tilde{w}}^{(r)}]$
Step 3:	Project test image $\boldsymbol{X}$ onto $\boldsymbol{W}_{\text{2DLDA}}^{(r)}$ yields
	an h by $\tilde{w}$ matrix $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_{\text{2DLDA}}^{(r)}$

Table 7.4: The procedure of Column-based 2DLDA (C-2DLDA)

Method 4: Column-based 2DLDA (C-2DLDA)	
input:	training data $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$,
	test data $\boldsymbol{X}$, reduced column dimension $\tilde{h}$
output:	projected training data $\{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n\}$,
	projected test data $\boldsymbol{Y}$
Step 1:	Compute image within-class scatter matrix
	$\boldsymbol{S}_w^{(c)} = \frac{1}{n}\sum_{j=1}^{n_c}\sum_{\boldsymbol{X}_i \in c_j}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_j)^T$,
	and image between-class scatter matrix
	$\boldsymbol{S}_b^{(c)} = \frac{1}{n}\sum_{j=1}^{n_c} n_j(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})^T$
Step 2:	Compute the first $\tilde{h}$ eigenvectors $\{\phi_l^{(c)}\}_{l=1}^{\tilde{h}}$ of $(\boldsymbol{S}_w^{(c)})^{-1}\boldsymbol{S}_b^{(c)}$,
	$\boldsymbol{W}_{\text{2DLDA}}^{(c)} \leftarrow [\phi_1^{(c)}, \cdots, \phi_{\tilde{h}}^{(c)}]$
Step 3:	Project test image $\boldsymbol{X}$ onto $\boldsymbol{W}_{\text{2DLDA}}^{(c)}$ yields
	an $\tilde{h}$ by w matrix $\boldsymbol{Y} = \boldsymbol{W}_{\text{2DLDA}}^{(c)}\boldsymbol{X}$

and C-2DLDA from one facial image.

7.3.1 Our Strategy

Our age-group classification scheme is shown in Figure 7.1. First of all, R-2DPCA, C-2DPCA, R-2DLDA and C-2DLDA are used in the first dimension reduction step respectively. As the dimensionality of the input data space increases, it becomes exponentially more difficult to find global optima for the parameter space to fit models. This is well known as the curse of dimensionality. Our experiment actually showed that 2DPCA (2DLDA) without PCA (LDA) produced worse classification accuracy. Therefore, PCA and LDA are now used for further dimensional reduction after a two-dimensional reduction step. The number of dimensions are experimentally chosen based on the classification accuracy in each projection method. After reducing the dimensions, for each gender, Gaussian models, which are shown in equation (5.33), are constructed on four different feature spaces such as R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA spaces.

In the testing phase, the likelihoods for each class are computed by the log posterior probability;

$$s(c_j|\boldsymbol{x_i}) = \log \Pr[c_j|\boldsymbol{x_i}] = \log \frac{\Pr[\boldsymbol{x_i}|c_j]}{\sum_{i=1}^{n_c} \Pr[\boldsymbol{x_i}|c_j]}, \tag{7.1}$$

and the class with the highest likelihood is chosen by comparing the output likelihoods from each class.

The next step is to integrate the likelihoods to get a higher accuracy than that of each individual classifier. The premise of fusion is that different classifiers or features can overcome the drawbacks of each other. In theory, integrating different data or classifiers can achieve better performance when they are independent of each other or they can overcome the shortcomings of each other. Since the likelihoods in the different feature spaces are het-

erogeneous, the likelihoods need to be normalized to be comparable to each other before combining them. The most commonly used normalization and fusion techniques are mentioned in Section 7.3.2 and 7.3.3, respectively.

7.3.2 Normalization Techniques

Two different normalization methods are exploited to transform heterogeneous scores into a common domain.

Min-max: The simplest normalization technique is the Min-max normalization. Min-max normalization is best suited for the case where the bound (maximum and minimum values) of the scores produced by a matcher are known. Given a set of output scores $\{s_j\}, j = 1, \cdots, n_c$, the normalized scores are given by

$$s_{new} = \frac{s_{old} - \min_j s_j}{\max_j s_j - \min_j s_j}. \tag{7.2}$$

Min-max normalization retains the original distribution of scores except for a scaling factor and transforms all the scores into a common range $[0, 1]$.

Z-score: The most commonly used score normalization technique is the z-score that is calculated using the arithmetic mean and standard deviation of the given data. This scheme can be expected to perform well if prior knowledge about the average score and the score variations of the matcher is available. This is calculated using the mean μ and the standard deviation of the scores σ from each model. Normalized scores are given by

$$s_{new} = \frac{s_{old} - \mu}{\sigma}. \tag{7.3}$$

7.3.3 Fusion Techniques

We have investigated four types of integration techniques such as sum, product, max and min rule [20] at the confidence level.

Sum Rule: Let x_i be the feature vector presented to the ith classifier. The integration process is given using the evidence provided by n_r classifiers. The sum rule assigns the input data to class c such that

$$c = \underset{j}{\operatorname{argmax}} \sum_{i=1}^{n_r} s(c_j|x_i). \tag{7.4}$$

The sum rule assumes that the posteriori probabilities computed by the individual classifiers do not deviate much from the prior probabilities. This rule is applicable when there is a high level of noise leading to ambiguity in the classification problem.

Product Rule: The product rule assigns the input data to class c such that

$$c = \underset{j}{\operatorname{argmax}} \prod_{i=1}^{n_r} s(c_j|x_i). \tag{7.5}$$

This rule is based on the assumption of statistical independence of the multiple representations.

Max Rule: The max rule assigns the input data to class c such that

$$c = \underset{j}{\operatorname{argmax}} \max_{i} s(c_j|x_i). \tag{7.6}$$

The max rule approximates the mean of the posteriori probabilities by the maximum value.

Min Rule: The min rule assigns the input data to class c such that

$$c = \underset{j}{\operatorname{argmax}} \min_{i} s(c_j|x_i). \tag{7.7}$$

The min rule is derived by bounding the product of posteriori probabilities.

7.4 Experiments

7.4.1 Evaluation Methods

The number of images used in this chapter is 26,222 (14,214 male and 12,008 female images). Table 5.2 in Chapter 5 shows the amount of data in each age-group class. Age-groups are based on actual age, and not perceived age. The image size used in this chapter is a facial region of 32x32 pixels, as shown in Figure 5.1.

In our experiments, age-groups are divided into 11 classes as shown in Table 5.2, which are based on 5-year range classification. Our goal is to classify 11-class age-groups with a high degree of accuracy.

We evaluate our proposed fusion-based two-dimensional method (method 1+2+3+4) on WIT-DB and compare with Yang's 2DPCA (method 1) and Ye's 2DLDA. Here, method 1, 2, 3 and 4 are the R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA, respectively. For all the experiments, the Gaussian models are applied for classification and 2-fold cross validation is used for estimating the classification accuracy. The dimensionality in the transformed space is determined by experiments for each method.

Due to the difficulty in age-group classification, the classification rate in the 10-year range, which includes the contiguous class with the higher likelihood, and in the 15-year range, which includes both contiguous classes, are observed.

In addition to classification rates, the average error distance defined in 6.2.3 is used to measure the degree of the error.

7.4.2 Experimental Results

Table 7.5: Classification rates achieved by different normalized and fusion methods using method 1+2+3+4. Method 1, 2, 3 and 4 are the R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA, respectively.

	age-group range	Sum Rule	Product Rule	Max Rule	Min Rule
Z-Score	5-year	47.3	47.1	45.0	46.3
	10-year	68.5	68.1	66.7	67.2
	15-year	76.8	76.8	74.5	76.3
Min-max	5-year	47.3	47.1	44.3	46.4
	10-year	68.5	68.5	64.4	67.3
	15-year	76.9	76.9	71.0	76.2

Table 7.6: Average error distances assessed by different normalized and fusion methods using method 1+2+3+4. Method 1, 2, 3 and 4 are the R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA, respectively.

	Sum Rule	Product Rule	Max Rule	Min Rule
Z-Score	0.96	0.96	1.06	0.98
Min-max	0.96	0.96	1.16	0.98

Figure 7.2 and 7.3 present the comparisons of our proposed fusion-based method based on multiple two-dimensional algorithms and classical PCA, LDA, 2DPCA and 2DLDA approach on classification accuracy. The number

of R-2DPCA, C-2DPCA, R-2DLDA and C-2DLDA dimensions were chosen as 32x16, 20x32, 32x10 and 10x32, respectively by experiments based on the accuracy rates. The number of PCA and LDA dimensions were chosen as 64 and 10 as well. In Ye's 2DLDA, the dimension was firstly reduced to 10x10, and then reduced to 10 by LDA. In this figure, min-max normalization and sum rule fusion techniques are used for the reason that they are better than other techniques in terms of accuracy rates on the whole. Figure 7.2 and 7.3 show that by combining a row-based method and column-based method (method 1+2 and method 3+4), classification accuracy is improved. Figure 7.2 and 7.3 also show that by combining multiple classifiers based on 2DPCA+PCA and 2DLDA+LDA is further improved. Our experiments on WIT-DB have shown that integration of appropriate information can improve the age-group classification rates. This suggests that 2DPCA+PCA features and 2DLDA+LDA features have provided different information that can compensate for each other. Figure 7.4 also shows that the margin of error can be reduced, which means some obvious errors can be eliminated. As a result, our fusion method (method 1+2+3+4) has proven to be superior to conventional PCA, LDA, Yang's 2DPCA [49] and Ye's 2DLDA [50].

Table 7.5 shows classification rates achieved by different normalization and fusion methods using method 1+2+3+4. Table 7.6 shows average error distances achieved by different normalization and fusion methods based on method 1+2+3+4. These tables show that the sum rule consistently provides the best performance regardless of normalization methods. The product rule has been found to have comparable performance to the sum rule, however it is sensitive to errors. As the sum rule is much less affected by estimation errors, this may provide a plausible explanation for its superior performance. Therefore, choosing the sum rule would be the preferable choice.

Our proposed method achieved the best performance by using four types

of projections methods (R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA) and sum rule fusion technique based on min-max or z-score normalization. Experimental results also show that 7.2% and 5.7% relative reduction in error over a performance baseline of Yang's 2DPCA and Ye's 2DLDA in the 5-year range classification, 10.6% and 12.9% in the 10-year range classification, and 12.1% and 17.6% in the 15-year range classification.

7.4.3 Discussions

In order to check whether our system's accuracy was comparable to human evaluation in terms of assessing a person's actual age, additional experiments have been performed using perceived age, which was given by 6 subjects. When the age-group range is 5 years, average accuracy rate based on human evaluation was 50.8% (maximum rate: 54.7%, minimum rate: 47.1%), whereas our system's accuracy rate was 47.3%. Our system differed only slightly and but fell short of the accuracy achieved by human evaluations. Human's ability in judging the age seems to depend on various experiences and previous knowledge during their whole lives. On the other hand, the appearance-based approaches only use the training data and then they cannot get a full picture of what that person has experienced. One of the reasons why multiple two-dimensional methods are used is that the variation of the feature extraction can be increased instead of the variation of faces to boost the accuracy. If there are wider variations, we will reduce obvious errors from a single classifier model that do not occur with humans.

When the baseline classification rates are considered, there is room for improvement in the 5-year range, while there is less room for improvement in the 15-year range. In view of the difficulties in classification, however, 5-year range classification is more challenging than 15-year range classification. For these reasons, the improvement rates in all 5-year, 10-year and 15-year ranges

Table 7.7: Confusion matrices for females' age-group classification. Horizontal: true class, vertical: classified class. (a) Yang's 2DPCA (b) Ye's 2DLDA (c) fusion-based two-dimensional algorithm

(a)

age-group	3-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-85
3 - 14	76.4	8.2	5.7	1.3	2.7	2.7	1.4	0.7	0.3	0.3	0.2
15 - 19	6.8	60.0	15.7	6.3	3.6	2.8	3.0	1.2	0.4	0.2	0.2
20 - 24	6.0	26.8	25.0	12.8	13.5	4.8	5.2	4.4	0.7	0.3	0.5
25 - 29	2.7	4.6	12.4	38.2	18.0	7.8	6.7	6.2	2.5	0.7	0.4
30 - 34	3.5	3.9	12.7	17.7	17.8	10.9	12.1	13.4	4.7	2.3	1.0
35 - 39	4.0	3.4	6.0	11.5	10.6	23.9	21.4	11.4	4.7	2.2	0.9
40 - 44	2.2	5.7	3.6	3.9	8.4	19.7	23.1	19.3	6.6	4.8	2.8
45 - 49	0.2	3.2	4.5	1.4	9.8	9.0	17.5	25.9	13.2	10.2	5.1
50 - 54	0.1	0.4	1.9	1.0	7.0	4.1	7.3	17.3	20.8	24.1	16.0
55 - 59	0.1	0.1	0.9	0.4	2.2	1.6	5.2	11.1	21.9	32.7	23.8
60 - 85	0.1	0.0	0.2	0.2	1.6	1.1	2.9	6.9	18.1	28.5	40.4

(b)

age-group	3-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-85
3 - 14	71.4	10.9	3.7	1.5	4.0	3.1	2.5	1.5	0.7	0.2	0.3
15 - 19	5.9	73.2	6.7	3.8	2.5	4.4	1.9	1.2	0.3	0.1	0.1
20 - 24	6.8	29.7	10.9	11.9	14.6	7.4	9.9	5.3	1.7	0.8	1.0
25 - 29	2.4	9.4	5.8	29.6	18.3	14.0	10.5	6.1	3.3	0.7	0.0
30 - 34	2.8	5.5	6.2	15.8	17.4	15.1	14.9	12.7	5.8	1.7	2.0
35 - 39	5.3	5.7	3.4	7.9	9.9	27.2	20.3	12.3	5.6	1.4	1.0
40 - 44	3.6	8.0	1.8	2.9	9.9	19.3	25.8	16.0	7.2	2.6	3.0
45 - 49	0.6	8.5	1.5	1.3	8.8	10.1	19.0	24.3	11.8	7.3	6.8
50 - 54	0.5	1.3	1.1	0.7	4.0	2.4	11.5	17.5	24.7	16.0	20.4
55 - 59	0.0	0.4	0.4	0.5	3.3	0.7	7.9	13.2	22.3	18.3	33.0
60 - 85	0.0	0.2	0.2	0.1	0.9	1.6	4.9	10.3	20.4	19.9	41.6

(c)

age-group	3-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-85
3 - 14	78.3	8.1	4.1	1.8	2.5	2.7	1.5	0.6	0.2	0.1	0.1
15 - 19	4.8	72.7	7.9	4.8	2.9	3.5	2.2	1.1	0.1	0.1	0.0
20 - 24	5.5	30.3	18.4	12.5	14.7	6.3	7.4	3.7	0.8	0.3	0.3
25 - 29	0.9	5.9	7.3	41.6	19.6	8.6	9.4	4.8	1.7	0.1	0.0
30 - 34	2.2	2.9	6.5	19.4	21.9	14.6	14.6	11.8	4.4	1.2	0.7
35 - 39	3.4	3.5	3.6	10.5	11.7	27.6	22.2	12.5	3.8	0.6	0.8
40 - 44	1.4	5.5	1.9	2.2	9.2	21.8	29.2	20.3	4.9	1.9	1.8
45 - 49	0.2	4.4	2.5	1.5	6.5	9.0	22.0	28.8	14.3	6.3	4.5
50 - 54	0.0	0.1	0.6	0.5	4.0	2.8	11.0	19.4	23.7	17.3	20.6
55 - 59	0.0	0.1	0.1	0.3	1.6	0.5	5.7	11.6	25.3	21.1	33.6
60 - 85	0.0	0.0	0.0	0.0	1.1	1.1	4.2	7.2	18.1	17.9	50.4

Table 7.8: Confusion matrices for males' age-group classification. Horizontal: true class, vertical: classified class. (a) Yang's 2DPCA (b) Ye's 2DLDA (c) fusion-based two-dimensional algorithm

(a)

age-group	3-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-85
3 - 14	85.0	8.5	2.4	0.6	0.9	0.3	1.3	0.6	0.2	0.1	0.2
15 - 19	6.2	64.5	19.4	4.2	1.8	1.0	1.8	0.9	0.1	0.1	0.1
20 - 24	5.5	39.9	29.8	11.5	4.9	3.0	3.4	1.4	0.2	0.1	0.4
25 - 29	1.5	4.4	8.1	45.2	20.6	13.5	3.9	2.0	0.5	0.0	0.4
30 - 34	0.5	4.4	2.8	32.6	25.8	19.3	7.5	3.3	2.7	0.0	1.1
35 - 39	1.1	1.6	3.8	22.0	23.0	24.6	13.0	6.4	2.6	1.1	0.9
40 - 44	2.3	2.1	4.3	5.5	7.0	16.5	21.2	18.0	14.6	6.3	2.4
45 - 49	0.6	2.8	1.8	3.1	4.7	10.9	17.4	19.9	21.3	13.2	4.4
50 - 54	0.3	1.2	1.5	1.0	1.7	4.3	13.9	16.5	25.1	22.8	11.6
55 - 59	0.2	0.3	0.4	0.1	0.4	2.3	5.3	12.6	21.0	27.7	29.6
60 - 85	0.3	0.4	0.0	0.1	0.1	1.1	3.3	5.4	13.9	30.9	44.3

(b)

age-group	3-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-85
3 - 14	83.0	9.8	1.0	0.4	0.5	0.8	2.4	0.8	0.5	0.5	0.2
15 - 19	4.1	78.8	10.2	2.5	0.2	1.2	1.5	0.8	0.7	0.1	0.1
20 - 24	4.1	54.0	18.5	12.3	2.5	4.6	2.2	1.4	0.3	0.2	0.1
25 - 29	0.9	6.7	2.9	45.8	16.1	19.0	4.3	2.5	1.2	0.3	0.4
30 - 34	0.8	6.0	2.0	32.1	17.4	27.9	4.6	5.6	2.5	0.5	0.5
35 - 39	0.7	3.7	3.2	21.7	18.6	31.1	9.2	8.5	1.7	0.9	0.8
40 - 44	1.8	4.5	2.1	6.6	4.5	15.9	22.7	17.1	10.8	8.9	5.2
45 - 49	0.7	2.6	0.8	2.1	2.7	12.7	17.8	25.0	14.9	14.3	6.6
50 - 54	0.1	3.1	1.2	1.1	0.7	5.1	15.9	22.9	17.1	16.6	16.2
55 - 59	0.0	0.0	0.2	0.4	0.8	2.0	8.6	15.8	14.3	25.5	32.4
60 - 85	0.2	0.9	0.0	0.0	0.1	1.1	4.1	9.6	11.7	30.1	42.2

(c)

age-group	3-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-85
3 - 14	87.3	8.5	1.0	0.5	0.1	0.4	1.7	0.6	0.1	0.0	0.0
15 - 19	3.3	79.2	12.4	2.4	0.2	0.4	0.9	1.0	0.3	0.0	0.0
20 - 24	3.6	51.6	23.0	12.7	1.9	3.1	2.6	1.1	0.2	0.1	0.1
25 - 29	0.5	5.7	5.1	52.5	14.9	15.5	3.4	1.7	0.5	0.1	0.2
30 - 34	0.0	5.4	2.7	34.8	16.0	28.5	6.3	4.7	1.3	0.1	0.3
35 - 39	0.8	1.9	2.2	21.1	17.5	34.0	11.8	7.6	2.2	0.7	0.4
40 - 44	1.6	2.4	2.5	5.2	2.5	19.2	25.3	22.5	10.3	5.5	3.0
45 - 49	0.4	1.2	1.0	1.3	2.9	14.3	22.0	24.9	15.6	13.1	3.3
50 - 54	0.0	0.8	0.3	0.5	0.6	5.5	18.5	22.5	19.6	19.0	12.8
55 - 59	0.0	0.0	0.0	0.1	0.4	1.4	6.2	14.9	17.3	23.6	36.1
60 - 85	0.1	0.2	0.0	0.0	0.0	0.7	3.5	7.8	10.2	28.8	48.7

cancel each other out and they are approximately equivalent at around 3-5%. These improvements seem to be slightly lower then expected. As we mentioned, these classification results come close to approximating performance of human ability, which is considered as marginal performance. As far as we are concerned, a certain level of improvement can be achieved. If the dramatically higher accuracy is required, it is necessary to use not only facial region but also other regions, for instance, hair, clothes or height etc.

In addition, in order to confirm the age-group classification ability of 2DPCA and 2DLDA, projected training data are plotted. Figure 7.5 shows examples of projected male training data. This figure shows 1st dimension (x-axis) and 2nd dimension (y-axis) of R-2DLDA+LDA and R-2DPCA+PCA. The data samples in the R-2DPCA+PCA space are spread out irregularly, whereas the ones in the R-2DLDA+LDA are in order of age-group. For this reason, 2DLDA+LDA methods have better performance than 2DPCA+PCA in a 10 or 15-year range because of less obvious errors, and they might highly contribute to the integration of classifiers.

We also analyze the difference between males and female. Figure 7.2 and 7.3 show that classification rates in males are approximately 7-10% higher than the ones in females in every method, and this result shows classifying age-groups using female images to be difficult. These figures, however, indicate the tendency of improvement is almost the same in every method.

Here we will focus on the age-group-specific accuracy rates. In the first place, two different aspects of the imbalance between the amount of samples in each category (age-group) are considered. The first experiment is designed using the same number of training samples for every class to reduce the influence of imbalance, which means the number of training samples is limited. This experiment shows that the accuracy rates are generally 1% lower than the ones that use all training samples. However, the tendency of superiority

or inferiority over the methods is almost equivalent. The next experiment is designed using the same number of test samples for every class. These classification rates in 5, 10 and 15 year ranges are approximately 8%, 6% and 4% lower respectively, compared with the original results. This could be because there are many younger samples included in WIT-DB and they may be easier to be classified. In this experiment, the differences of superiority or inferiority between the methods are not observed as well. In the second place, we also provide detailed results of the age-group-specific accuracy rates to check if there are some categories that are easier to be classified than others. Table 7.7 shows the confusion matrices for females' age-group classification rates and Table 7.8 is for males'. In terms of younger age-groups (under 19) and the oldest age-group (over 60), classification rates are higher in each gender, however as for age-groups between 20 and 59, classification rates decrease.

In addition, our classification method approximately quadruples the computational cost. However, on a Pentium IV 3.07 GHz Windows machine, it takes only 1.34 milliseconds for one classification on average and works effectively as a real-time application.

7.5 Conclusion

In this chapter, new types of age-group classification methods are proposed to develop a demographic analysis system for market research purposes. A large data set, which includes more than 26,000 image samples, is constructed and age-groups are subdivided into smaller ranges such as a 5-year range. The problem of achieving high accuracy within a 5-year range is difficult even by human evaluations. Using a single classifier to solve this problem, the more categories, the more misclassifications seem to be made. Some data samples were misclassified into a quite different category, but the differences between the true classes' scores and the falsely predicted classes' scores are subtle in

some instances. Hence, extracting as many different features as possible from a single source and finding out another classifier, which confidently claims that this category is correct, is one of the solutions to improve the accuracy in the 5-year range.

In order to reduce as many errors as possible, two-dimensional-based dimensional reduction methods are used. There are three main reasons why we use 2DPCA (2DLDA) not other traditional methods such as PCA (LDA). One of them is the fact that Yang et al. [49] and Ye et al. [50] carried out experiments to compare the performance between 2DPCA (2DLDA) and PCA (LDA), and found that two-dimensional algorithms had better performance in terms of recognition rates. We also proved that the classification rates across all trials were almost the same or higher using 2DPCA (2DLDA) than PCA (LDA). The second reason is that the extraction of image features is computationally more efficient using 2DPCA (2DLDA) than PCA (LDA). In contrast to the covariance matrix (the scatter matrix) of PCA (LDA), the size of the image covariance matrix is much smaller. As a result, it is easier to evaluate the covariance matrix accurately and less time is required to determine the corresponding eigenvectors. The third reason, which is the most important point in this chapter, is that two-dimensional algorithms can extract two different features from two types of directions such as a row direction and column direction. Thus, higher accuracy rates can be expected by combining a row direction based method and column direction based method. Based on these reasons, we focused on the antisymmetry of two-dimensional feature extraction algorithms and constructed multiple classifiers with different error tendencies by preparing the rows and columns of images. To be more precise, the images are projected into a row direction and column direction separately using 2DPCA or 2DLDA for dimension reduction, and PCA or LDA is used for further dimensional reduction; R-2DPCA+PCA,

C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA. After the normalization process, multiple scores are integrated by using sum, product, max and min rules. Our proposed method showed the best performance by using four types of projections methods (R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA), and also achieved approximately the same accuracy as human evaluations, proving that our system would be substitutable for existing marketing research system.

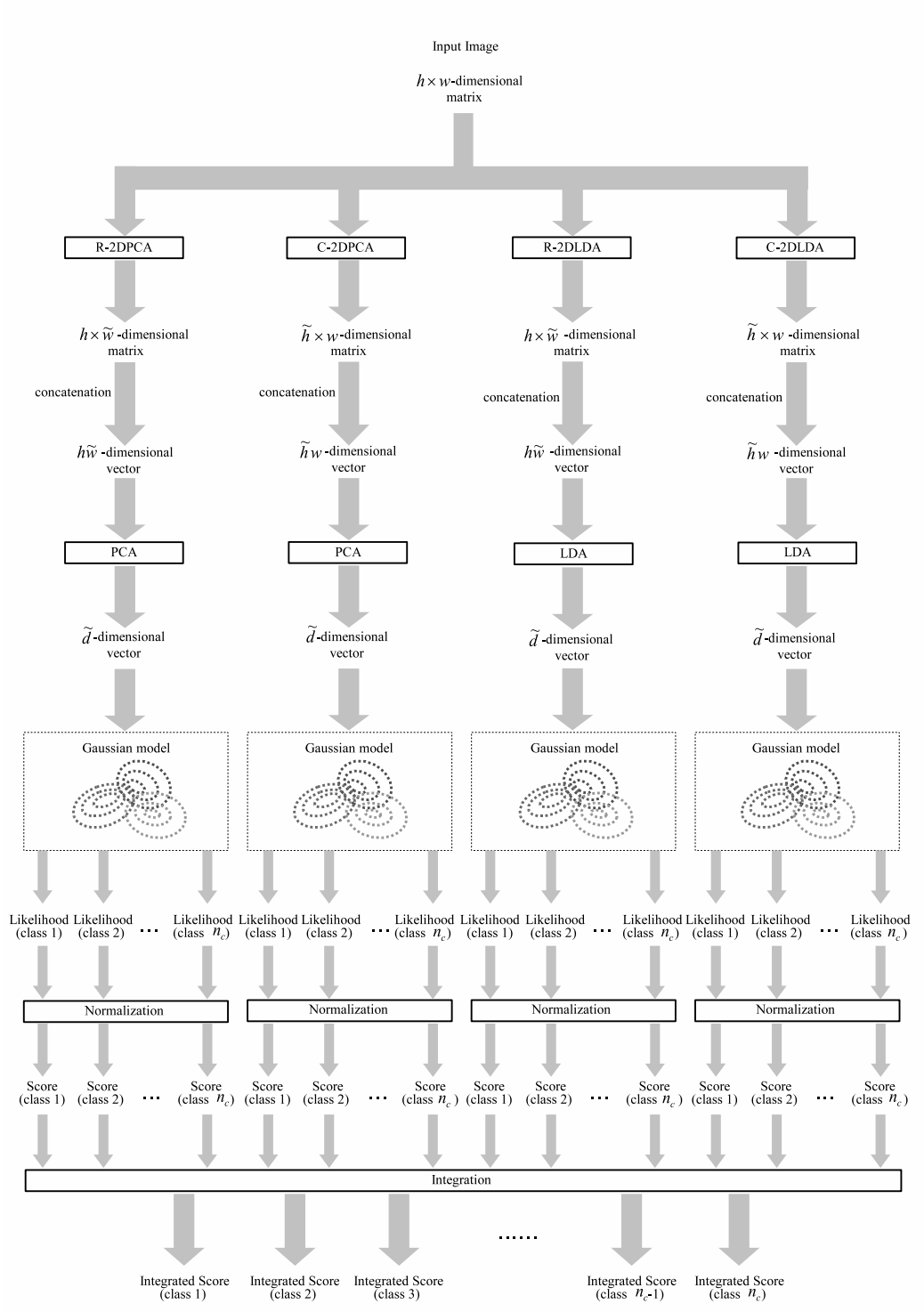


Figure 7.1: Our age-group classification scheme

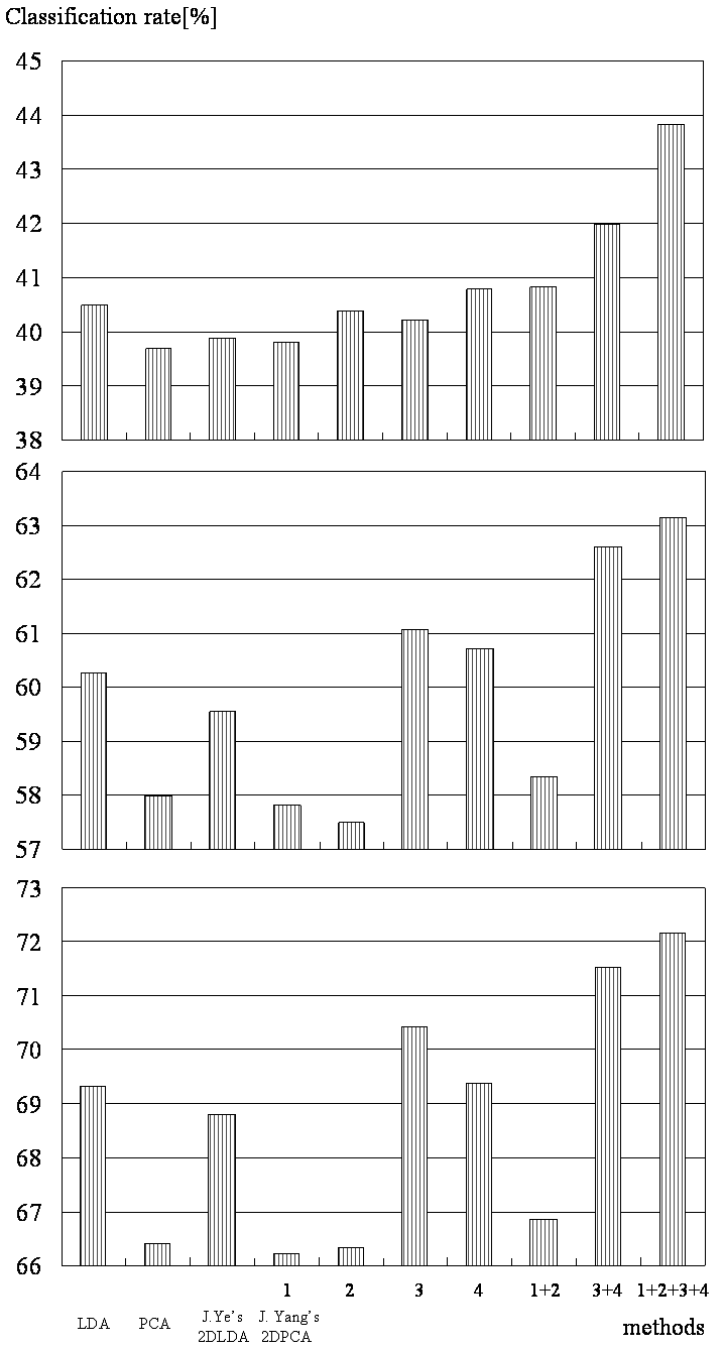


Figure 7.2: The classification accuracy of females' age-groups based on different approaches. Method 1, 2, 3, and 4 are the R-2DPCA, C-2DPCA, R-2DLDA, and C-2DLDA, respectively. Min-max normalization and sum rule fusion techniques are used. (Top: within the 5-year range; Middle: within the 10-year range; Bottom: within the 15-year range)

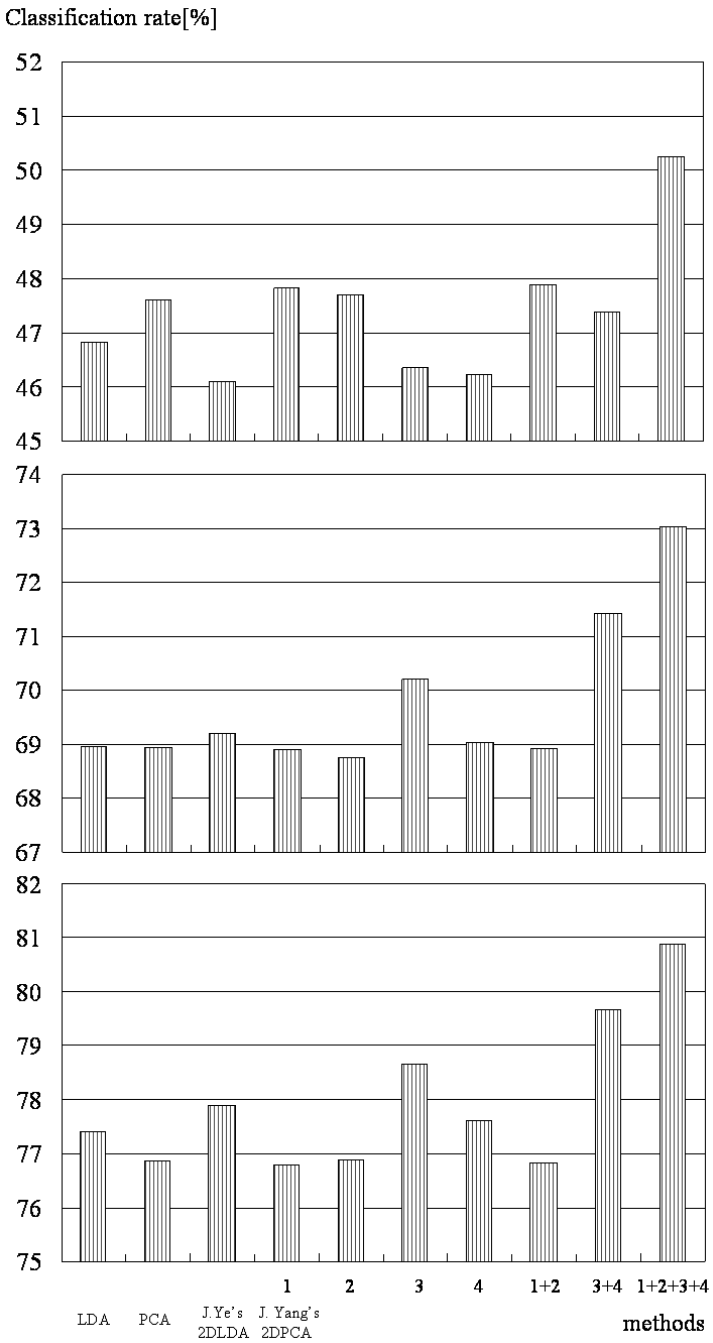


Figure 7.3: The classification accuracy of males' age-groups based on different approaches. Method 1, 2, 3, and 4 are the R-2DPCA, C-2DPCA, R-2DLDA, and C-2DLDA, respectively. Min-max normalization and sum rule fusion techniques are used. (Top: within the 5-year range; Middle: within the 10-year range; Bottom: within the 15-year range)

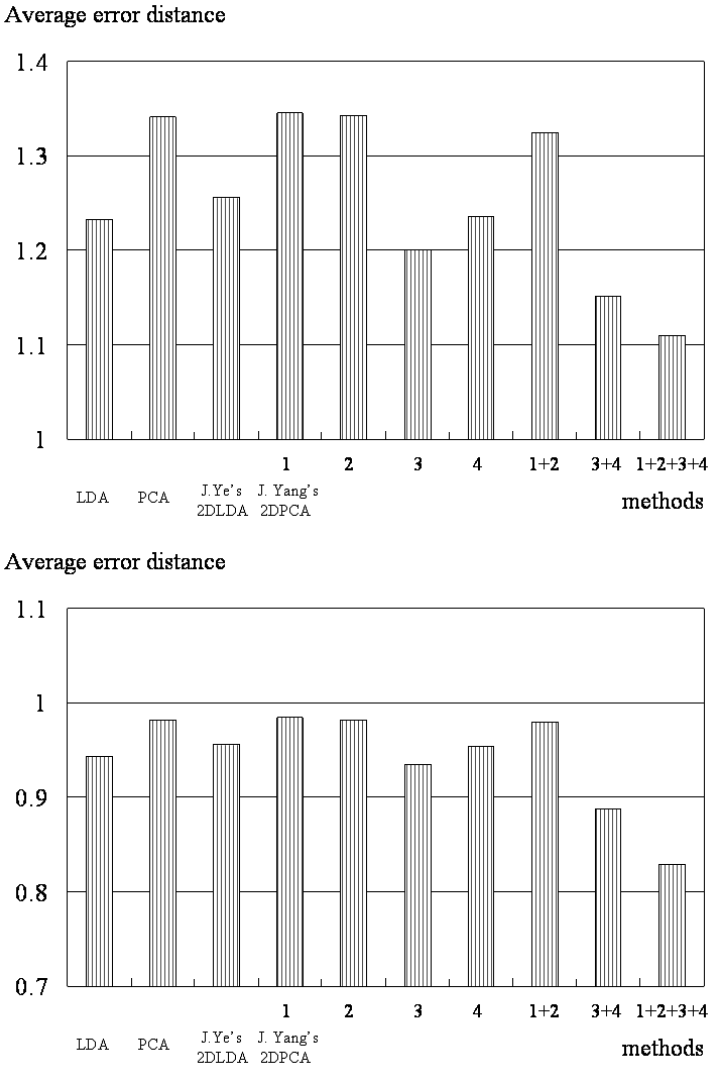


Figure 7.4: The average error distances based on different approaches. (Top: Females, Bottom: Males) Min-max normalization and sum rule fusion techniques are used. Method 1, 2, 3 and 4 are the R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA, respectively.

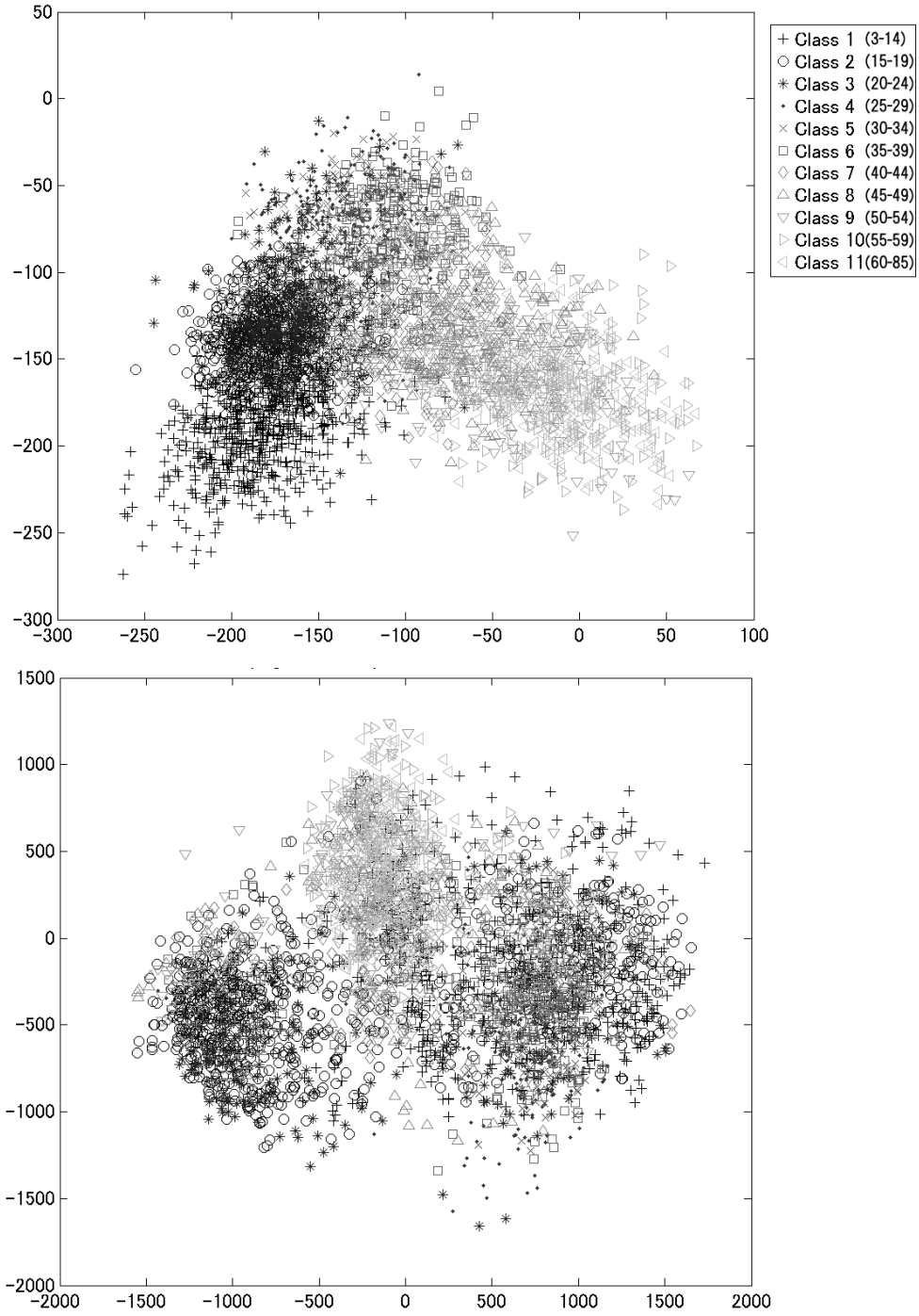


Figure 7.5: Examples of projected male training data using R-2DLDA+LDA (top) and R-2DPCA+PCA (bottom)

Chapter 8

Conclusion

8.1 Contributions

The main goal of this work was to improve the gender/age-group classification rates on a large data set. In this dissertation we have advanced the state of the art in gender and age-group classification in the following ways.

In Chapter 2, we have designed a large-scale database which includes more than 5,500 Japanese subjects (about 2,500 females and about 3,000 males), where there are more than 26,000 images. We have attempted to classify gender and age-groups using this database with a wide variety of age-groups and illumination changes. This set is much larger than those in other researchers' classification experiments and the classification performance turns out to be fairly good.

In Chapter 3 and 4, we have developed and implemented two new types of gender classification methods based on the integration techniques. In the first method, a person's hairstyle, tie and *décolletage* information is extracted and used as well as facial information. Bayes' rule is then applied to integrate these four types of information. In the second method, facial area and neck area have been separately analyzed and the final decision has been made based on integration techniques such as distance summation, GMM based

integration and SVM based integration. We have also proved that even if only one single source is used, feature extraction can produce variations and boost performance. We finally found that the best classification rate has been achieved with GMM based or SVM based integration.

In Chapter 5, we have proposed new algorithms called 2DLDA and 2DHLDA for age-group classification. Through new two-dimensional approaches, we have successfully achieved better performance than the conventional statistical learning methods such as PCA, LDA and 2DPCA.

In Chapter 6, we have carefully studied the differences between actual ages and perceived ages of the subjects. We have found that perceived ages can be used instead of actual ages, or can be even better in terms of data consistency. Using perceived age data, better class separability have been obtained.

In Chapter 7, finally, multiple two-dimensional feature extraction algorithms have been proposed for age-group classification. We have extended 2DPCA and 2DLDA such that they operate using two different orientations each (specifically row and column directions), named R-2DPCA, C-2DPCA, R-2DLDA and C-2DLDA. The best performance has been achieved using four types of projection methods (R-2DPCA+PCA, C-2DPCA+PCA, R-2DLDA+LDA and C-2DLDA+LDA).

We suggest that the results of our research can be used for developing real-world applications for gender and age-group classification.

8.2 Discussion of Future Work

The experimental results that have been presented in this dissertation are based mainly upon frontal facial images. Since the chances that we can obtain frontal view images are much lower in real-life environments, system performance would be less than optimal. In future work, more data based

on images from non-frontal views (perhaps at multiple angles) should be collected from more subjects. Thereafter, we would be able to make our system far more robust and useful.

At this point, when we perform the classification test, only one frame is utilized. In future work, extending the integration technique to multi-frame images (movies, for example) could be an important next step. With multiple frames, our gender/age-group classification system would most certainly become more reliable.

Age data has been categorized into 5, 10 or 15-year ranges, and classification algorithms have been applied. However, age data can be considered to be a continuous value, and this would be even more natural. Another idea is to treat age recognition problems not as clustering problems but as regression problems, for instance using Support Vector Regression (SVR) [54] or other regression methods.

Acknowledgments

I would like to thank all the individuals who have helped me during my Ph.D. study at Waseda University. First of all, I would like to express my gratitude to my advisor, Professor Tetsunori Kobayashi, for his guidance and support in academic research. His ideas, insights, suggestions, questions, and enthusiasm were great help in stimulating my thought and in bringing this research forward. I am grateful to my Ph.D. committee, Yasuo Matsuyama and Jiro Katto, for their valuable ideas, suggestions, and encouragement.

I would like to thank all the members of the Kobayashi lab in the Department of Computer Science and Engineering at Waseda University for their help.

I would also like to thank NEC Soft, Ltd. for their support, understanding and encouragement through my university degree.

I wish to thank all the NOVA instructors for their help with the English translation of this dissertation.

Finally, special thanks must go to my wife Akiko and the rest of my family for all the happiness they have shared with me and their unconditional love and support.

Bibliography

[1] R.A. Render and H.F.Walker, "Mixture densities, "Maximum likelihood and the EM algorithm," SIAM Review, vol.26, no.2, pp.195-239, 1984.

[2] G. W. Cottrell and J. Metcalfe, "EMPATH: Face, emotion, and gender recognition using holons," Advances in Neural Information Processing Systems, vol.3, pp.564-571, 1991.

[3] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, "SEXNET: A neural network identifies sex from human faces," Advances in Neural Information Processing Systems, vol.3, pp.572-577, 1991.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol.3, no.1, pp.71-86, 1991.

[5] R. Brunelli and T. Poggio, "HyberBF networks for gender classification," Proc. Image Understanding Workshop, pp.311-314, 1992.

[6] A. M. Burton, V. Bruce, and N. Dench, "What's the difference between men and women? Evidence from facial measurement," Perception, vol.22, pp.153-176, 1993.

[7] S. Yen, P. Sajda and L. Finkel, "Comparison of gender recognition by PDP and radial basis function networks," The Neurobiology of Computation, pp.433-438, 1994.

[8] H. Abdi, D. Valentin, B. Edelman, and A. J. O'Toole, "More about the difference between men and women: evidence from linear neural networks and the principal-component approach," Perception, vol.24, pp.539-562, 1995.

[9] D. M. Burt and D. I. Perrett, "Perception of age in adult Caucasian male faces: computer graphic manipulation of shape and colour information," Proc. of Royal Society, pp.137-143, 1995.

[10] C. Cortes and V. Vapnik, "Support vector networks," Machine Learning, vol.20, no.3, pp.273-297, 1995.

[11] S. Tamura, H. Kawai, and H. Mitsumoto, "Male/female identification from 8 x 6 very low resolution face images by neural network," Pattern Recognition, vol.29, no.2, pp.331-335, 1996.

[12] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.18, no.8, pp.831-836, 1996.

[13] L. Breiman, "Bagging predictors," Machine Learning, vol.24, no.2, pp.123-140, 1996.

[14] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.19, no.7, pp.711-720, 1997

[15] J. Fellous, "Gender discrimination and prediction on the basis of facial metric information," Vision Research, vol.37, no.14, pp.1961-1973, 1997.

[16] A. J. O'Toole, T. Vetter, H. Volz, and E. Salter, "Three-dimensional caricatures of human heads: distinctiveness and the perception of age," Perception, vol.26, pp.719-732, 1997.

[17] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," PhD thesis, John Hopkins University, Baltimore, 1997.

[18] S. Gutta, H. Wechsler, and P. J. Phillips, "Gender and ethnic classification of face images," Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition, pp.194-199, 1998.

[19] C. Burges, "Tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol.2, no.2, pp.121-167, 1998.

[20] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.20, no.3, pp.226-239, 1998.

[21] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol.10, no.5, pp.1299-1319, 1998.

[22] A. R. Martinez and R. Benavente, "The AR face database," Technical Report 24, Computer Vision Center (CVC) Technical Report, Barcelona, 1998.

[23] V. Vapnik, "Statistical learning theory," Wiley, 1998.

[24] A. J. O'Toole, T. Price, T. Vetter, J. C. Bartlett, and V. Blanz, "3D shape and 2D surface textures of human faces: The role of 'averages' in

attractiveness and age," Image and Vision Computing, vol.18, no.1, pp.9-20, 1999.

[25] C. Choi, "Age change for predicting future Faces," Proc. of IEEE International Fuzzy Systems Conference, pp.1603-1608, 1999.

[26] A. Lanitis, C. J. Taylor, and Timothy F. Cootes, "Modeling the process of aging in face images," Proceedings of IEEE ICCV99, pp.131-136, 1999.

[27] Y. H. Kwon and N. V. Lobo, "Age classification from facial images," Computer Vision and Image Understanding, vol.74, no.1 pp.1-21, 1999.

[28] L. Hong, A. Jain, and S. Pankanti, "Can multibiometrics improve performance?," Proc. AutoID'99, pp.59-64, 1999.

[29] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," Machine Learning, vol.37, no.3, pp.297-336, 1999.

[30] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," Proc. of IEEE Intl. Workshop on Neural Networks for Signal Processing IX, pp.41-48, 1999.

[31] A. Lanitis and C.J. Taylor, "Towards automatic face identification robust to ageing variation," Proc. of 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp.391-396 2000.

[32] A. Lanitis and C.J. Taylor, "Robust face recognition using automatic age normalization," Proc. of Electrotechnical Conference, vol.2, pp.478-481, 2000.

[33] S. Gutta, R. J. Huang, P. Jonathon, and H. Wechsler, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces," IEEE Trans. on Neural Networks, vol.11, no.4, pp.948-960, 2000.

[34] L. Chen, H. Liao, M. Ko, J. Lin and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," Pattern Recognition, vol.33, pp.1713-1726, 2000.

[35] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, no.10, 2000.

[36] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of face verification results on the XM2VTS database," Proceedings of the 15th International Conference on Pattern Recognition, Barcelona (Spain), vol.4, pp.858-863, September, 2000.

[37] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data - with application to face recognition," Pattern Recognition, vol.34, pp.2067-2070, 2001.

[38] W. B. Horng, C. P. Lee and C. W. Chen, "Classification of age groups based on facial features," Tamkang Journal of Science and Engineering, vol.4, no.3, pp.183-191, 2001.

[39] A. M. Martinez, and A. C. Kak, "PCA versus LDA," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23. no.2. 2001.

[40] B. Moghaddam and M.-H. Yang, "Learning gender with Support Faces," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.24, no.5, pp.707-711, 2002.

[41] Z. Sun, G. Bebis, X. Yuan, and S. J. Louis, "Genetic feature subset selection for gender classification: A comparison study," Proc. IEEE Workshop on Computer Vision, pp.165-170, 2002.

[42] A. Lanitis, C. J. Taylor, and Timothy F. Cootes, " Toward automatic simulation of aging effects on face images," IEEE Trans on PAMI, vol.24, no.4, pp.442-456. 2002.

[43] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," IEEE Trans. on Neural Networks, vol.13, no.6, pp.1450-1464, November 2002.

[44] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face Recognition: A Literature Survey," ACM Computing Surveys, pp.399-458, 2003.

[45] L. Walawalkar, M. Yeasin, A. M. Narasimhamurthy, and R. Sharma, "Support vector learning for gender classification using audio and visual cues," Intl. Journal of Pattern Recognition and Artificial Intelligence, vol.17, no.3, pp.417-439, 2003.

[46] A. Ross and A. Jain, "Information fusion in biometrics," Pattern Recognition Letters, vol.24, pp.2115-2125, 2003.

[47] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12), pp.1615-1618, 2003.

[48] B.-W. Hwang, H. Byun, M.-C. Roh, and S.-W. Lee, "Performance Evaluation of Face Recognition Algorithms on the Asian Face Database, KFDB", In Audio- and Video-Based Biometric Person Authentication (AVBPA), pp.557-565, 2003.

[49] J. Yang, D. Zhang, A. F. Frangi, J. Y. Yang, "Two-Dimensional PCA: a new approach to appearance-based face representation and recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.26, no.1, pp.131-137, 2004.

[50] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," Proc. Neural Information Processing Systems, pp.1569-1576, 2004.

[51] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face recognition across pose and illumination," Handbook of Face Recognition, Stan Z. Li and Anil K. Jain, ed., Springer-Verlag, June, 2004.

[52] G. Shakhnarovich and B. Moghaddam, "Face recognition in subspaces," Handbook of Face Recognition, Eds. Stan Z. Li and Anil K. Jain, Springer-Verlag, December, 2004.

[53] W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, "CAS-PEAL large-scale Chinese face database and evaluation protocols," Technical Report DJL-TR-04-FR-001, Joint Research & Development Laboratory, 2004.

[54] A. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing vol.14, no.3, pp.199-222, 2004.

[55] J. Ruiz-del-Solar and P. Navarrete, "Eigenspace-based face recognition: a comparative study of different approaches," IEEE Transactions on Systems, Man and Cybernetics, Part C, vol.35, no.3, pp.315-325. August, 2005.

[56] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," Pattern Recognition, vol.26, no.5, pp.527-532, 2005.

[57] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Two-dimensional FLD for face recognition," Pattern Recognition, vol.38, no.7, pp.1121-1124, 2005.

[58] J. Yang, D. Zhang, X. Yong, and J. Yang, "Two-dimensional discriminant transform for face recognition," Pattern Recognition, vol.38, no.7, pp.1125-1129, 2005.

[59] H. Kong, L. Wang, E. K. Teoh, J.-G. Wang, and R. Venkateswarlu, "A framework of 2D Fisher discriminant analysis: Application to face recognition with small number of training samples," Proc. of Intl. Conf. on Computer Vision and Pattern Recognition (CVPR2005), vol.2, pp.1083-1088, 2005.

[60] S. Chen, Y. Zhu, D. Zhang, and J.-Y. Yang. "Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA," Pattern Recognition, vol.26, no.8, pp.1157-1167, 2005.

[61] F. H. C. Tivive and A. Bouzerdoum, "A shunting inhibitory convolutional neural network for gender classification," Proc. of Intl. Conf. on Pattern Recognition, vol.4, pp.421-424, 2006.

[62] D. Kalamani and P. Balasubtamanie, "Age classification using fuzzy lattice neural network," Proc. of Sixth Intl. Conf. on Intelligent Systems Design and Application (ISDA'06), vol.3, pp.225-230, 2006.

[63] J. Wang, Y. Shang, G. Su, and X. Lin, "Age simulation for face recognition," Proc. of 18th Intl. Conf. on Pattern Recognition (ICPR2006), vol.3, pp.913-916, 2006.

[64] S. Baluja and H. Rowley, "Boosting sex identification performance," Intl. Journal of Computer Vision, vol.71, no.1, pp.111-119, 2007.

Published Work

Paters

[1] ○ Kazuya Ueki and Tetsunori Kobayashi, "Fusion-based age-group classification method using multiple two-dimensional feature extraction algorithms," IEICE Transactions on Information and Systems, 2007. (Accepted)

[2] ○ Kazuya Ueki, Teruhide Hayashida, and Tetsunori Kobayashi, "Two-dimensional heteroscedastic linear discriminant analysis for age-group classification," Proceedings of the 18th International Conference on Pattern Recognition (ICPR2006), vol.2, pp.585-588, August 2006.

[3] ○ Kazuya Ueki, Teruhide Hayashida, and Tetsunori Kobayashi, "Subspace-based age-group classification using facial images under various lighting conditions," Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FG2006), pp.43-48, April 2006.

[4] ○ Kazuya Ueki, Hiromitsu Komatsu, Satoshi Imaizumi, Kenichi Kaneko, Satoshi Imaizumi, Nobuhiro Sekine, Jiro Katto, and Tetsunori Kobayashi, "A method of gender classification by integrating facial, hairstyle, and clothing images," Proceedings of the 17th International Conference on Pattern Recognition (ICPR2004), vol.4, pp446-449, August 2004.

Lectures

[5] Kazuya Ueki and Tetsunori Kobayashi, "Age-group classification using facial images based on actual age and apparent age," Meeting on Image Recognition and Understanding 2006 (MIRU2006), pp.1308-1312, July 2006 (in Japanese).

[6] Kazuya Ueki, Teruhide Hayashida, and Tetsunori Kobayashi, "Age-group classification using facial images under various lighting conditions," IEICE Technical Report, vol.105, no.375, PRMU2005-95, pp.13-18, October 2005 (in Japanese).

[7] Teruhide Hayashida, Kazuya Ueki, and Tetsunori Kobayashi, "Integration of classifier for gender and age classification," IEICE Technical Report, vol.105, no.375, PRMU2005-96, pp.19-24, October 2005 (in Japanese).

[8] Rei Mochiki, Yuichi Uchiyama, Kazuya Ueki, Jiro Katto, and Tetsunori Kobayashi, "Robust human detection in a complicated background using multiple Gaussian mixture skin models," IEICE Technical Report, vol.105, no.375, PRMU2005-99, pp.37-42, October 2005 (in Japanese).

[9] Kazuya Ueki, "Facial feature recognition from video sequences," Institute of Image Electronics Engineers of Japan, the 14th Workshop for Visual Media Appliances, January 2005 (in Japanese).

[10] Hiromitsu Komatsu, Kazuya Ueki, Satoshi Imaizumi, Kenichi Kaneko, Nobuhiro Sekine, Jiro Katto, and Tetsunori Kobayashi, "The clothing discrimination using bosom images and its application to gender estimation," Meeting on Image Recognition and Understanding 2004 (MIRU2004), vol.1, pp.624-629, July 2004 (in Japanese).

[11] Satoshi Imaizumi, Kazuya Ueki, Kenichi Kaneko, Nobuhiro Sekine, Jiro Katto, and Tetsunori Kobayashi, "Method of gender and age-group estimation by integrating multiple information," IEICE Technical Report, vol.103, no.452, PRMU2003-142, pp.13-18, November 2003 (in Japanese).

[12] Kazuya Ueki, Kenichi Kaneko, Satoshi Imaizumi, Nobuhiro Sekine, Satoshi Imaizumi, Hiromitsu Komatsu, Jiro Katto, and Tetsunori Kobayashi, "A method of age and gender estimation using Gaussian mixture model," The 3rd Forum on Information Technology (FIT2003), vol.3, pp.125-126, September 2003 (in Japanese).

[13] Kenichi Kaneko, Kazuya Ueki, Satoshi Imaizumi, Nobuhiro Sekine, Hiromitsu Komatsu, Satoshi Imaizumi, Jiro Katto, and Tetsunori Kobayashi, "System of human feature extraction from a monitoring camera picture," The 3rd Forum on Information Technology (FIT2003), vol.3, pp.569-570, September 2003 (in Japanese).

[14] Jun Mizuno, Tasuya Watanabe, Kazuya Ueki, Kazuyuki Amano, Eiji Takimoto, and Akira Maruoka, "On-line estimation of hidden Markov model parameters," Proceedings of the 3rd International Conference on Discovery Science, DS'2000 - Lecture Notes in Artificial Intelligence, vol.1967, pp.155-169, December 2000.

[15] Jun Mizuno, Tasuya Watanabe, Kazuya Ueki, Kazuyuki Amano, Eiji Takimoto, and Akira Maruoka, "On-line estimation of hidden Markov model parameters," LA Symposium, no.21, February 2000 (in Japanese).