Available online at www.sciencedirect.com
SCIENCE DIRECT
Image and Vision Computing 23 (2005) 631–642
image AND vision COMPUTING
www.elsevier.com/locate/imavis

# Component-based LDA face description for image retrieval and MPEG-7 standardisation

Tae-Kyun Kim[a,*], Hyunwoo Kim[a], Wonjun Hwang[a], Josef Kittler[b]

[a]*Computing Lab., Samsung Advanced Institute of Technology, San 14-1, Nongseo-ri, Kiheung-eup, Yongin, Kyungki-do 449-712, South Korea*
[b]*Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 5XH, UK*

Received 29 January 2004; received in revised form 15 October 2004; accepted 2 February 2005

**Abstract**

We propose a method of face description for facial image retrieval from a large data base and for MPEG-7 (Moving Picture Experts Group) standardisation. The novel descriptor is obtained by decomposing a face image into several components and then combining the component features. The decomposition combined with LDA (Linear Discriminant Analysis) provides discriminative facial descriptions that are less sensitive to light and pose changes. Each facial component is represented in its Fisher space and another LDA is then applied to compactly combine the features of the components. To enhance retrieval accuracy further, a simple pose classification and transformation technique is performed, followed by recursive matching. Our algorithm has been developed to deal with the problem of face image retrieval from huge databases such as those found in Internet environments. Such retrieval requires a compact face representation which has robust recognition performance under lighting and pose variations. The partitioning of a face image into components offers a number of benefits that facilitate the development of an efficient and robust face retrieval algorithm. Variation in image statistics due to pose and/or illumination changes within each component region can be simplified and more easily captured by a linear encoding than that of the whole image. So an LDA encoding at the component level facilitates better classification. Furthermore, a facial component can be weighted according to its importance. The component with a large variation is weighted less in the matching stage to yield a more reliable decision. The experimental results obtained on the MPEG-7 data set show an impressive accuracy of our algorithm as compared with other methods including conventional PCA (Principal Component Analysis)/ICA (Independent Component Analysis)/LDA methods and the previous MPEG-7 proposals.


*Keywords:* Linear discriminant analysis (LDA); Face recognition; Image retrieval; Face descriptor; Facial component; Combining classifier; MPEG-7

## 1. Introduction

Recently, face descriptors as candidates for MPEG-7 (Moving Picture Experts Group) standardisation have been proposed for face image retrieval in video streams [1–6,8,14,22]. The various methods developed for multimedia description and retrieval involve many kinds of visual factors like shape, color and motion. Because of its importance in many applications, human face is a special topic in computer vision research, treated separately from those of generic shape information for visual discrimination and description. A face descriptor in MPEG-7 should meet the following requirements. It should be possible to extract it without any prior knowledge about the current image content (person identity) meaning that its statistical basis should be images of persons other than those contained in the test data base. Each image in the database is exploited as a query image in order to retrieve the other images of the same person from the data set. The descriptor should also be compact, even for a large data set. A challenging problem is to retrieve face images with large variations in lighting and pose when a single query image is given.

To compensate for image variation due to illumination changes, Wang and Tan [2] proposed the second-order Eigenface method and Kamei and Yamada [3] extended their work to use a confidence factor describing face symmetry and intensity variation due to illumination change. Kim et al. [5] developed the second-order PCA (Principal Component Analysis) Mixture Model (PMM)

* Corresponding author. Tel.: +82 31 280 1746; fax: +82 31 280 9257.
*E-mail address:* ktk22@hanmail.net (T.-K. Kim).

0262-8856/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.02.005

method. The second-order approaches attempt to remove effects due to changes in illumination by removing the component of the image lying in the sub-spaces spanned by the first few eigenvectors. However, the approaches seem to be weak under pose variation because they are describing a holistic pixel distribution, which is sensitive to pose change.

To compensate for image variation due to pose as well as illumination change, Nefian and Davies [6] used the DCT (Discrete Cosine Transform)-based embedded Hidden Markov Model (eHMM) for face description, while Kim et al. [4] proposed the eHMM method with the second-order Block-specific eigenvectors. The eHMM algorithm implicitly deals with pose variation using embedded states corresponding to facial regions and segmenting the observed image into overlapping blocks, but it may home on a local minimum if the initial solution is not close to the global minimum. Wiskott et al. [7] developed a Gabor wavelet-based algorithm called elastic bunch graph matching. However, these algorithms are computationally expensive. In face retrieval where the descriptor should be extracted from a single face image without any prior knowledge, eHMM-based methods [4,6,8] have been found to have poor performance.

In this paper, we propose a new approach to deal with pose and illumination variation resulting in a very efficient face description in terms of both accuracy and size. Preliminary works have been done in our studies [17–19,22]. We introduce a component-based LDA (Linear Discriminant Analysis) face representation. The closely related work is that of Heisele et al. [9]. They detect and align facial components to cope with pose changes. An SVM (Support Vector Machine) algorithm is applied to the geometrically aligned pixels and the extracted support vectors are used for binary classification. Although they show that their component-based algorithm gives a better accuracy than holistic image representations, SVM is very time consuming as far as huge and multi-class databases are concerned and the facial component detection is very difficult in natural environments.

Our algorithm combines the component-based representation with LDA. First, pose transformation is carried out based on a full face image analysis. The image is then partitioned into several facial components to simplify the modeling of image statistics. The components are encoded by LDA to compensate for the effect of illumination and expression variation. Another LDA is then applied to the collection of the component-based LDA representations yielding a compact description referred to as 'cascaded LDA'. The decomposition of the face image and its re-combination in the LDA space effectively solves the problem of face retrieval and person identification. Finally a recursive matching is proposed to further improve the face retrieval accuracy.

Section 2 briefly reviews LDA and Section 3 introduces the component-based representation and details the cascaded LDA. Sections 4 and 5 describe the pose compensation method and recursive matching, respectively. The experimental results are presented in Section 6 and conclusion is drawn in Section 7.

## 2. Review of LDA

When the training set data is labeled, supervised learning techniques offer a considerably more effective description of face images in terms of discriminative features than unsupervised learning. LDA effectively removes the effect of illumination and slight pose variation, provided such variations are reflected in the training set. LDA simply retains the identity information.

Given a set of $N$ images $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ with each image belonging to one of $c$ classes $\{X_1,X_2,\ldots,X_c\}$, LDA finds a linear transformation matrix $\mathbf{W}$ in such a way that the ratio of the between-class scatter and the within-class scatter is maximized. The between-class and within-class scatter matrices are defined by

$$\mathbf{S}_{\mathrm{B}} = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}} \tag{1}$$

and

$$\mathbf{S}_{\mathrm{W}} = \sum_{i=1}^{c} \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \mu_i)(\mathbf{x}_k - \mu_i)^{\mathrm{T}} \tag{2}$$

respectively, where $\mu_i$ denotes the mean image of class $X_i$, $\mu$ is a global mean, and $N_i$ denotes the number of images in class $X_i$. If the within-class scatter matrix $\mathbf{S}_{\mathrm{W}}$ is not singular, LDA finds an orthonormal matrix $\mathbf{W}_{\mathrm{opt}}$ maximizing the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix as

$$\mathbf{W}_{\mathrm{opt}} = \arg\max_{W} \frac{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{W}|}{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{W}|} = [\mathbf{w}_1\mathbf{w}_2\ldots\mathbf{w}_m]. \tag{3}$$

The set of bases of the solution $\{\mathbf{w}_i | i=1,2,\ldots m\}$ is constituted by generalized eigenvectors of $\mathbf{S}_{\mathrm{B}}$ and $\mathbf{S}_{\mathrm{W}}$ corresponding to the $m$ largest eigenvalues $\{\lambda_i | i=1,2,\ldots m\}$, i.e. $\mathbf{S}_{\mathrm{B}}\mathbf{w}_i = \lambda_i\mathbf{S}_{\mathrm{W}}\mathbf{w}_i$, $i=1,2,\ldots m$. Generally, to overcome the singularity of $\mathbf{S}_{\mathrm{W}}$, PCA first reduces the vector dimension before applying LDA [11,12]. Each image is represented by a vector projection, $\mathbf{y}_k = \mathbf{W}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{x}_k$, $k=1,2,\ldots,N$.

## 3. Component-based LDA representation

Although the holistic face image representation afforded by the LDA method described in Section 2 has many useful properties, it cannot cope with illumination and pose changes that have not been captured by the training set very well.

In order to enhance the generalisation capability of the LDA representation to changes underrepresented in the training data, we have developed a component-based LDA representation whereby the face is divided into a number of facial regions and a separate LDA is learnt for each region. The motivation for this is based on the following argument. By virtue of the 3D nature of human face, illumination changes will result in globally non-linear changes in the image intensity function. However, locally the effects can be approximated by a linear function which can be corrected for much more readily. A similar local versus global argument can be made for the effect of small variations in pose. In this paper, LDA applied to a whole face is called 'the holistic LDA' and LDA applied to image components is called 'the component-based LDA'. To overcome the problem that the component scheme lacks relational information between the components, combined LDA representations are considered which combines both sources of information.

### 3.1. Face representation by components

We partition a face image into facial components corresponding to forehead, eyes, nose and mouth regions. Statistical image variation due to illumination and/or pose change within each component patch may be smaller than that in the whole image space. Generally, holistic approaches based on PCA/ICA (Independent Component Analysis)/LDA encode the greyscale correlation among every pixel position statistically and different lighting and camera geometry result in a severe change of face representation. Since our component-based scheme encodes the facial components separately, variations in image statistics are limited to each component region. Fig. 1 illustrates an example of illumination effect on the statistics of the whole image and a component. The distribution parameters of the same identity group on the principal component of whole face regions largely change under the different lighting conditions, whereas the distribution parameters on the principal component of the eye region remain approximately consistent. Most of all, pre-processing within small patches is easier than that of the whole image region. The usual step of subtracting a mean vector or the best fitting plane from an image vector before any linear projection can provide a useful pre-processing in a small image patch.

The proposed component-based representation has other advantages. It exhibits a greater flexibility in similarity matching. Since each facial component can be considered as a separate classifier, the output can be weighted by its discriminability and prior knowledge. In this paper, weighting schemes are proposed for both 'component level' and 'feature level' fusion. We shall demonstrate that the proposed cascaded LDA which is a combining scheme at the feature level, provides a very compact face description while maintaining a good retrieval and identification accuracy. The details will be described in Section 3.3. In addition to the component weighting, lower

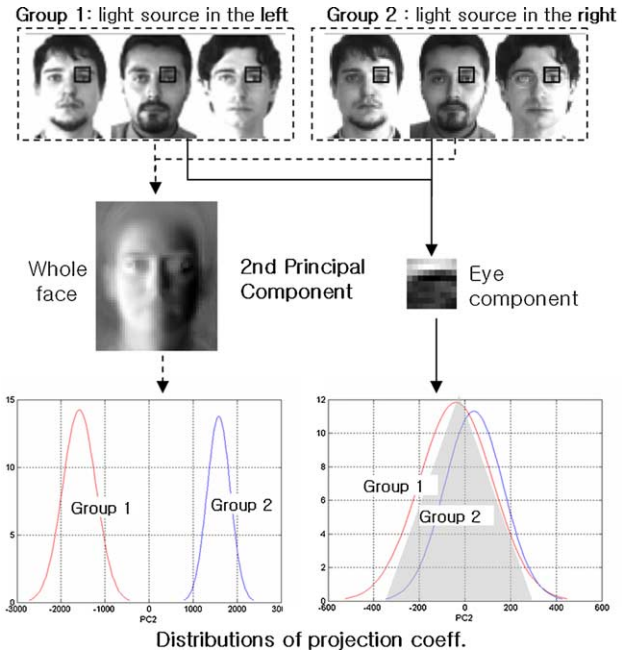



Fig. 1. Illumination effects on the statistics of a whole image in contrast with that of an eye component.

statistical complexity of a component region over that corresponding to the whole face image facilitates classification using a linear encoding via PCA/ICA/LDA. The experimental results presented in Section 6 show that the component encoding approach outperforms holistic encoding methods in face image retrieval experiments.

We considered the two schemes of face partitioning shown in Fig. 2. The size and position of the components are fixed relatively to the eye positions. One has 14 small or large components with a large overlap. The small components, which have their centres on physically meaningful points like eyes, nose or mouth, are expected to efficiently represent local statistics for face classification. The large components around head, cheeks and neck represent other important information about the face. The overlaps between neighboring components will promote the preservation of the component adjacency relationships. The other scheme is similar to the first but with the number of components reduced. This is achieved by merging two or three components of the original partitioning into one. In this way, the number of components is reduced to not more than five by searching for the best combination exhaustively, based on the retrieval performance on the training set. As a result, the retrieval accuracy as compared with the original partitioning is not compromised. Moreover, the descriptor size and the computational complexity are significantly reduced.

### 3.2. LDA encoding of components

Given a set of $N$ training images $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$, a set of LDA transformation matrices is extracted. First, all images

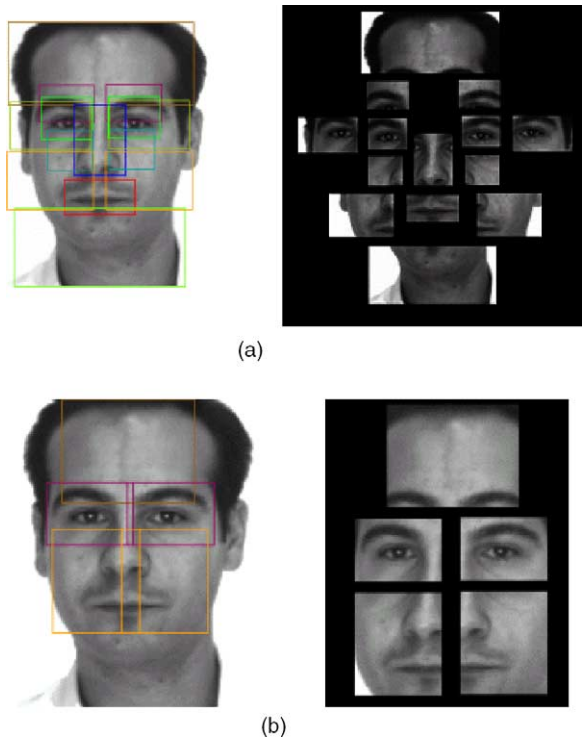



Fig. 2. Component definitions. (a) 14 components separation, (b) 5 components separation.

are partitioned into $L$ facial components. The image patches of each component are represented in a vector form with the $k$th component being denoted as $\{\mathbf{c}_1^k, \ldots, \mathbf{c}_N^k\}$. Then, for the $k$th facial component, the corresponding LDA transformation matrix $\mathbf{W}^k$ is computed.

During testing, the $L$ vectors $\{\mathbf{c}^1,\ldots,\mathbf{c}^L\}$ corresponding to facial component patches are extracted from a face image $\mathbf{x}$ of the test data set. A set of LDA feature vectors $\mathbf{y}=\{\mathbf{y}^1,\ldots,\mathbf{y}^L\}$ is extracted by transforming the component vectors by the corresponding LDA transformation matrices as

$$\mathbf{y}^k = (\mathbf{W}^k)^{\mathrm{T}}\mathbf{c}^k, \quad k = 1, 2, \ldots, L. \tag{4}$$

Thus for the component-based LDA method, a face image $\mathbf{x}$ is represented by a set of LDA feature vectors $\{\mathbf{y}^1,\ldots,\mathbf{y}^L\}$. This set is augmented by the LDA coefficients $\mathbf{y}^0$ of the holistic image to yield a combined representation $\{\mathbf{y}^0,\mathbf{y}^1,\ldots,\mathbf{y}^L\}$.

### 3.3. Combined LDA representations

The dimensionality of the simple collection of the holistic and component features derived in Section 3.2 is quite large. In order to reduce this dimensionality, and to determine the appropriate weighting for these features, the combined representation $\{\mathbf{y}^0,\mathbf{y}^1,\ldots,\mathbf{y}^L\}$ is transformed by another LDA. This is the weighting scheme at the feature level. The resulting two stage process is referred to as 'cascaded LDA'. It is illustrated in Fig. 3. A face image is represented as a merged column vector $\mathbf{f}=\{\mathbf{f}^0,\mathbf{f}^1,\ldots,\mathbf{f}^L\}$ obtained by concatenating the normalized LDA feature vectors by $\mathbf{f}^i=\mathbf{y}^i/\|\mathbf{y}^i\|$. An LDA transformation matrix $\mathbf{W}_{2\mathrm{nd}}$ is computed for the merged vectors of the training face images. This LDA allows us to control the dimension of the final descriptor $\mathbf{z}$. Note that large eigenvalues indicate the corresponding vectors to be more discriminative. Moreover, the elements of the LDA transformation matrix define the weights for the component features. The diagonal terms of the matrix reflect the importance of each feature and the off-diagonal terms assign weights for combinations of the features. The proposed final description $\mathbf{z}$ is given by

$$\mathbf{z} = (\mathbf{W}_{2\mathrm{nd}})^{\mathrm{T}}\mathbf{f} \tag{5}$$

CTU [14] applied Generalized Discriminant Analysis (GDA), a non-linear version of LDA, to our component-based LDA descriptor as a combiner. They achieved good retrieval performance by considering the facial space to be non-linear. However, the complexity of the feature extraction and matching in their method is too high to be applied to large data sets.

The other combining schemes which consider each component as a separate classifier have been considered. At the component level, weighting is considered in a similarity matching stage. To measure the similarity of two face images, each component quantifies the similarity of the two images first and then all the results are combined with appropriate weights. Given two face images $\mathbf{x}_1,\mathbf{x}_2$ represented by the component LDA feature vector set $\mathbf{y}_1,\mathbf{y}_2$, the similarity $d(\mathbf{y}_1,\mathbf{y}_2)$ is defined by the weighted sum of normalized-correlations

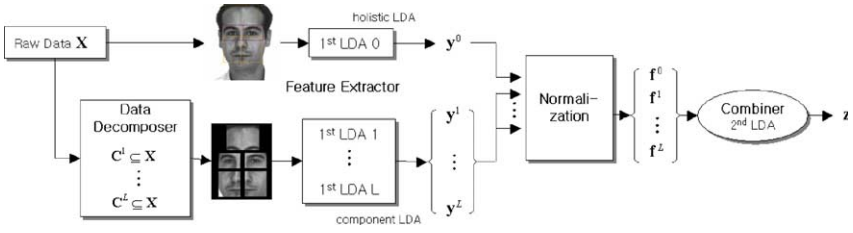



Fig. 3. Architecture of the cascaded LDA description.

$$\mathrm{Corr}_k(\mathbf{y}_1, \mathbf{y}_2) = \frac{\mathbf{y}_1^k \cdot \mathbf{y}_2^k}{\|\mathbf{y}_1^k\| \|\mathbf{y}_2^k\|}$$

between the corresponding components as

$$d(\mathbf{y}_1, \mathbf{y}_2) = \frac{1}{L} \sum_{k=1}^{L} w_k \mathrm{Corr}_k(\mathbf{y}_1, \mathbf{y}_2) \quad (6)$$

where $\mathbf{y}_1^k, \mathbf{y}_2^k$ are the LDA feature vector sets of the $k$th facial component of the two face images. The weights $w_k$ were heuristically chosen to reflect the retrieval performance of each component. This method requires some heuristics to select the weights and does not consider non-linear relationships of the components.

Any merit of non-linear modeling of the relationship of the components is investigated by using the non-linear Fisher Linear Discriminant (FLD). The input vector for the non-linear FLD has a $L+1$ dimensional space of $\{\mathrm{Corr}_1(\mathbf{y}_1, \mathbf{y}_2), \ldots, \mathrm{Corr}_L(\mathbf{y}_1, \mathbf{y}_2), \mathrm{label}\}$, which comes from the normalized-correlations of the corresponding $L$ components and the pose-pair label, which is an integer number, determined using the method described in Section 4. The input vector is then enlarged by its quadratic form. It is noted that the dimension $L+1$ is low. All the augmented input vectors of the training face image pairs of the same identities construct one group and the vectors of image pairs of different identities make the opposite group. A FLD transformation matrix is now computed to solve a binary classification problem. The FLD transformation matrix defines the weights for the respective components. The comparison will be given in Section 6 with the combining methods like a product, sum and weighted sum rule, consequently showing the benefits of the proposed cascaded LDA.

## 4. Pose classification and compensation

The component-based representation developed in Section 3 can cope with small deviations from the frontal face pose. When the component positions are well aligned, facial pose variation can be better compensated, yielding accuracy improvement [9]. More significant changes will require geometric correction. If deviations from the frontal pose change are not too severe, we can correct them by using affine transformation. Rather than estimating an exact transformation needed, which would require the detection of abundant facial features [21], we adopt a simple approach based on pose quantisation. The quantisation is achieved by classification of input face images into five face orientations based on a holistic PCA model. More specifically, during training, the eye positions are given for each face image. The face images are manually clustered according to predefined five pose sets. For each pose class, eigenfaces are extracted by PCA. During the pose classification stage a test image is projected into the five different eigen-sub-spaces corresponding to the first few eigenfaces of each pose class. The image is classified into the class with the smallest projection error [13].

The transformation from the each quantised pose class to the frontal pose class is determined by a linear matrix that establishes the correspondences of facial feature points between representative images of the two pose categories. All of the images are pose quantised and transformed to the frontal versions by using the corresponding affine transformation. Retrieval is then based on matching of the component-based representations of these transformed face images.

## 5. Recursive face matching

A novel face retrieval scheme based on recursive matching has been adopted. When a face image is encoded, the representation contains environmental variation as well as the face characteristic. When a query face is presented, the retrieved faces reflect certain encoding error of the query face image injected by the environmental variations. The faces obtained under similar imaging conditions to those of the query will be just retrieved successfully. Assuming that the most of the top ranked retrieved faces have been identified by the face characteristic, as the rank one face image will encapsulate slightly different environmental variations from that affecting the query face, this can be utilized as another query face which is helpful to find more faces of the same identity captured under slightly different environments from that of the new query. In this way, the gap in various conditioned faces and the original query face can be bridged as illustrated in Fig. 4(a). Should the first retrieved face be a wrong answer, the effect of the second retrieval by the new query is weakened by weighting the matching scores between the new query and the original query. The process can be conducted recursively.

As shown in Fig. 4(b), suppose that a query image $q$ (here, it is the final description vector $\mathbf{z}$ (5)) is given and we would like to retrieve $K$ images in ordered rank from an image database $\{I_i | i = 1, \ldots, N\}$ of size $N$. In the first step, from the query image, we obtain a sorted image set $\{q_i | i = 1, \ldots, M\}$ of size $M$ from the image database. It includes their corresponding score array $\{s(i) | i = 1, \ldots, M\}$, which is represented by $s(i) = D(q, q_i)$, where $D(\mathbf{z}_1, \mathbf{z}_2)$ denotes Euclidean distance of feature vectors $\mathbf{z}_1$ and $\mathbf{z}_2$. In the next step, the first-ranked retrieval image $q_1$ is selected as a new query $q'$ and the matching and sorting procedure is repeated only within the buffer of the size $M$, not the whole database. As a result, we have the re-sorted image set $\{q_i' | i = 1, \ldots, M\}$ and the corresponding score array $\{s'(i) | i = 1, \ldots, M\}$. To reflect the matching by the new query, the score array is updated by

$$s'(i) = s(i) + w_1 \cdot D(q', q_i') \quad (7)$$

where $w_1$ denotes a weighing constant. This procedure is performed recursively. In the $n$th step, we have a sorted data

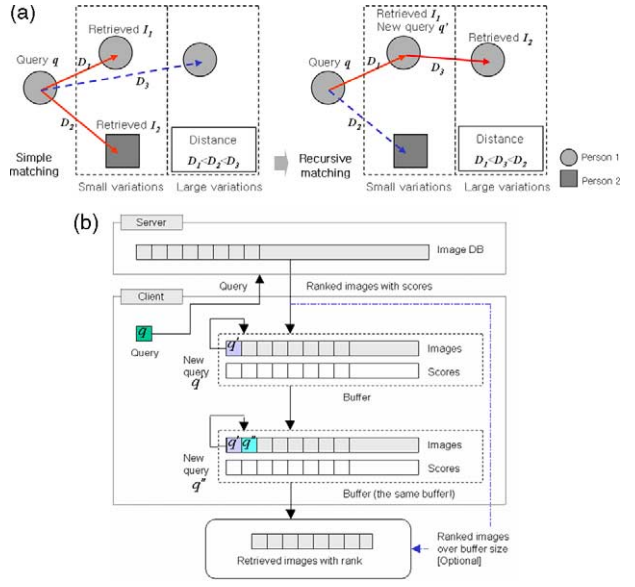



Fig. 4. Recursive face matching. (a) Conceptual drawing. (b) Block diagram.

set $\{q_i^{(n-1)} | i = 1, \ldots, M\}$ with corresponding score array

$$\{s^{(n-1)}(i) | i = 1, \ldots, M\}, \quad \text{where}$$
$$s^{(n-1)}(i) = s^{(n-2)}(i) + w_{n-1} \cdot D(q^{(n-1)}, q_i^{(n-1)}). \tag{8}$$

Note that the additional computational complexity of the recursive matching can be ignored in comparison with that of the simple matching which only utilizes an original query because the new matching and sorting procedure is repeated only within the buffer, not the whole database.

## 6. Experimental results and discussion

### 6.1. Database, protocol and measure

An experimental face database was obtained by MPEG-7 standardisation effort. It consists of five databases: the extended version 1 MPEG-7 face database (E1), Altkom database (A2), MPEG-7 test set in XM2VTS database (M3), FERET database (F4), and MPEG-7 test set in the Banca database (B5). The details are described in Table 1. The total number of images is 11,845. All the images in the database are normalized to $46 \times 56$ pixels by using the manually marked eye positions, giving fixed eye positions. For the training of the proposed pose classification and compensation algorithms, the face images are manually clustered into the five pose sets, which are quasi-frontal (F) and depth-rotated in four directions (R, L, U, D). Each rotated pose group has about $15 \sim 45^\circ$ rotation from the frontal one in each direction. The normalized sample images of the dataset with the labels of the pose class are given in Fig. 5.

The images for the experiments were strictly divided into training and test sets as shown in Table 2, which have different identities. All the parameters such as the basis vectors are extracted from the training set. All test images are utilized as a query image to retrieve the other images (ground truth) of the corresponding person in the test data set. Note that the test set has more than five images of one person. As a measure of retrieval performance, we use ANMRR (Average Normalized Modified Retrieval Rate) specified in [15]. ANMRR is 0 when all ground truth images are ranked on top, and it is 1 when all ground images are ranked out of the first $m$ images.

### 6.2. Component weighting scheme

For the weighted sum rule of normalized-correlations between the corresponding components (6), the weight of each component was determined proportionally to the square of a reciprocal of ANMRR on the training set. Table 3 gives the retrieval performance of each component. The components around forehead were dominant in recognition and this may be because the data set does not include large variations over time, relative to illumination and pose changes. The fixed hair-styles of the majority of the subjects provided a consistency of face images over time. The improvement in retrieval accuracy achieved by the weighting is shown in Table 4 for the five component encoding scheme.

### 6.3. Comparison of sub-space methods

Three kinds of sub-space methods applied to the whole face image, PCA, PCA–ICA, LDA were compared.

Table 1
Face dataset

| Ref. | | |
|---|---|---|
| E1 | The extended version 1 MPEG-7 face database | 635 persons (five images per person exhibiting illumination and view variations) |
| A2 | Altkom database | 80 persons (15 images per person: 5 views $\times$3 illuminations) |
| M3 | MPEG-7 testset in XM2VTS database | 295 persons (10 images per person: 5 views $\times$ 2 different times (sessions 1 and 4)) |
| F4 | FERET database | 4000 images of 875 persons selected for the 'background' at testing stage for both face image retrieval and personal identification |
| B5 | MPEG-7 testset in the Banca database | 52 persons (10 images per person: 4 office, 4 outdoor,2 ideal; each image taken at different time) |


Fig. 5. Sample face images of the dataset (E1∼B5). The manual pose class (F, R, L, U, D) is denoted with each picture.

Table 2
The summary on training and test dataset set for the face retrieval experiments

| Exp. no. | Train image | | | | Test image | | | |
|---|---|---|---|---|---|---|---|---|
| | DB | No. of persons | No. of images | No. of total images | DB | No. of persons | No. of images | No. of total images |
| 1-1 | E1 | 337 | 5 | 1685 | E1 | 298 | 5 | 1490 |
| 2-1 (train:test = 1:15) | E1 | 40 | 5 | 200 | E1 | 595 | 5 | 2975 |
| 2-2 (train:test = 1:3) | E1 | 160 | 5 | 800 | E1 | 475 | 5 | 2375 |
| 2-3 (train:test = 1:1) | E1 | 337 | 5 | 1685 | E1 | 298 | 5 | 1490 |
| 3-1 (train:test = 1:1) | A2 | 40 | 15 | 600 | A2 | 40 | 15 | 600 |
| | B5 | – | – | – | B5 | 52 | 10 | 520 |
| | E1 | 317 | 5 | 1585 | E1 | 318 | 5 | 1590 |
| | M3 | 147 | 10 | 1470 | M3 | 148 | 10 | 1480 |
| | F4 | – | – | – | F4 | – | – | 4000 |
| | Total | 504 | | 3655 | Total | 558 | | 8190 |
| 3-2 (train:test = 1:4) | A2 | 16 | 15 | 240 | A2 | 64 | 15 | 960 |
| | B5 | – | – | – | B5 | 52 | 10 | 520 |
| | E1 | 127 | 5 | 635 | E1 | 508 | 5 | 2540 |
| | M3 | 59 | 10 | 590 | M3 | 236 | 10 | 2360 |
| | F4 | – | – | – | F4 | – | – | 4000 |
| | Total | 202 | | 1465 | Total | 860 | | 10,380 |

Exp. no. 1-1 is identical to Exp. no. 2-3.

The Bartlett's PCA–ICA technique [10] was adopted without the first eight eigenfaces to remove illumination effects. From Table 4, we see that the PCA–ICA largely outperforms PCA and the proposed component and weight scheme also significantly enhance the performance of the PCA–ICA. The result of experiment 2-3 in Table 5, involving images obtained under the same conditions as those of experiment 1-1, shows that the supervised learning LDA outperforms both PCA and PCA–ICA. This is because the class-specific learning is much more profitable to eliminating various changes while keeping the identity information. It is noted that the proposed component scheme and the combined scheme significantly enhances the performance for both unsupervised and supervised feature extraction methods.

### 6.4. Generalization test: holistic vs. component LDA

Table 5 and the cumulative FIR (False Identification Rate) graphs in Fig. 6 present a comparison of the generalization performance of the holistic LDA, the component LDA and the combined LDA representation by the weighted sum of normalized-correlations. The retrieval performance measure, ANMRR and the person identification measure, FIR showed the same tendency for

Table 3
Retrieval performance of each component (Exp. no. 1-1)

| Five components | Forehead | Left eye | Right eye | Left cheek | Right cheek |
|---|---|---|---|---|---|
| ANMRR | 0.437 | 0.620 | 0.633 | 0.670 | 0.678 |

Table 4
Comparison of PCA, ICA, and component-based ICA (Exp. no. 1-1)

| Methods | ANMRR |
|---|---|
| PCA w/o first 8 | 0.499 |
| PCA–ICA w/o first 8 | 0.367 |
| Component-based PCA–ICA w/o first 8 (14 components) | 0.309 |
| Component-based PCA–ICA w/o first 8 (5 components) | 0.307 |
| Component-based PCA–ICA w/o first 8 (5 components) and weighting scheme | 0.252 |

Table 5
Generalization test: holistic vs. component LDA

| Unit: ANMRR | Exp. no. 2-1 | Exp. no. 2-2 | Exp. no. 2-3 |
|---|---|---|---|
| Holistic LDA | 0.524 | 0.198 | 0.100 |
| Component LDA | 0.159 | 0.104 | 0.107 |
| Combined LDA | N/A | N/A | 0.067 |

Exp. no. 2-3 is identical to Exp. no. 1-1 in Table 4.

superiority or inferiority. Note that the component LDA highly outperforms the holistic LDA in the case of small training data. The two have a similar performance in the experiment 2-3, which uses a half of the data set for training and the other half for the test. The holistic LDA can be over-trained giving a poor generalization. As shown in Fig. 7(a), most of the important facial information appears at a certain region of faces. Intensity variations of the holistic LDA basis images are around the forehead, and the eyes of faces. If the test faces had more discriminative information in other parts, the learned basis vectors would not differentiate faces effectively. Compared to the holistic approach, the component LDA learns evenly from the whole region of a face by separating the components. Fig. 6(d) shows that the combined method improves the performance in the first rank FIR dramatically. The first rank FIRs of the holistic and combined method are 0.0355 and 0.0208, respectively.

### 6.5. Pose compensation

The pose classification and affine transformation provided an additional enhancement in facial image retrieval as shown in Table 6. The pose classification algorithm yielded about 92% correct rate on the test set and the pose compensation combined with this automatic pose classification showing comparable retrieval accuracy with that of using manual pose label. The overall performance was enhanced more dramatically for the data sets which have large pose variation like Altkom(A2) and XM2VTS(M3) dataset. However, we see that the linear affine transformation could not solve the pose problem basically due to the

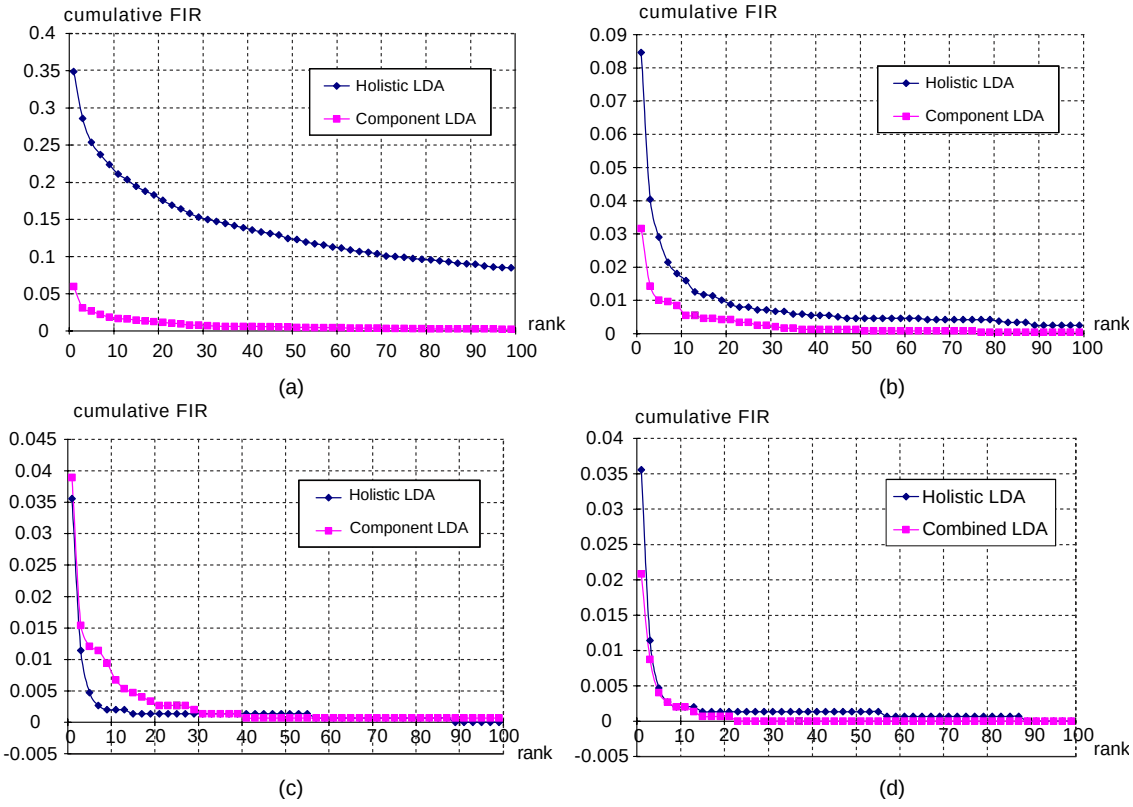
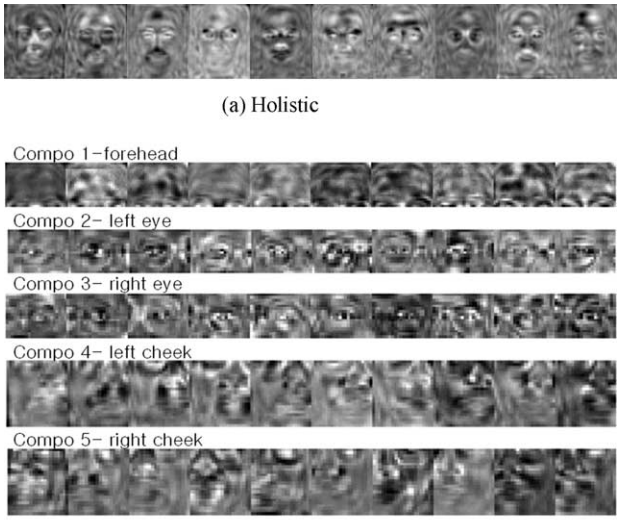



Fig. 6. Cumulative FIR plots. (a) Experiment 2-1. (b) Experiment 2-2. (c) and (d) Experiment 2-3.




(a) Holistic

(b) Facial components

(c) LDA transformation matrix at the second stage

Fig. 7. LDA basis images.

non-linearity of the face pose set. The results of the pose set, A2 and M3 are much worse than that of the quasi-frontal set E1. Any benefit of non-linear pose transformation should be further investigated taking the complexity into account.

### 6.6. Combined LDA

Clearly the weighted sum of normalized-correlations in (6) requires some annoying effort to choose weights heuristically and does not consider non-linear relationships of the components. The proposed two stages of LDA, cascaded LDA, yields an efficient representation in terms of both accuracy and descriptor size. In Table 7, compared with the weighted sum method, the cascaded LDA has much smaller descriptor size with a similar retrieval performance. The performance of the cascaded LDA was comparable to that of GDA [14] with approximately one tenth of GDA complexity of feature extraction and matching.

Fig. 8 shows the comparison for the combined LDA methods at the component level. The comparison was done in terms of false acceptance and false rejection. They were measured for (1) the non-linear FLD, (2) sum rule, (3) product rule and (4) the weighted sum rule of normalized-correlations. The best result of fusion at the component level was obtained by using the weighted sum of normalized-correlations and the non-linear FLD, which was not as good as the result of the cascaded LDA as mentioned above.

### 6.7. Recursive retrieval

The proposed recursive matching in conjunction with was the new face descriptor was then comparatively

Table 6
Face image retrieval results for database (Exp. no. 3-1)

| Unit: ANMRR | | | Total | A2 | B5 | E1 | M3 |
|---|---|---|---|---|---|---|---|
| w/o pose compensation | | Holistic LDA | 0.473 | 0.468 | 0.700 | 0.207 | 0.681 |
| | | Combined LDA | 0.438 | 0.421 | 0.556 | 0.157 | 0.705 |
| With pose compensation | Manual pose label | Holistic LDA | 0.440 | 0.435 | 0.655 | 0.181 | 0.645 |
| | | Combined LDA | 0.394 | 0.388 | 0.541 | 0.145 | 0.611 |
| | Automatic pose classification | Combined LDA | 0.405 | 0.392 | 0.538 | 0.167 | 0.619 |

Table 7
The cascade LDA/or GDA and recursive retrieval (Exp. no. 3-1)

| Common conditions | Combining method | Matching | Dimension | ANMRR |
|---|---|---|---|---|
| With pose compensation | Weighted sum | Simple matching | 240 | 0.394 |
| | Cascade GDA [14] | Simple matching | 50 | 0.388 |
| Combined LDA | Cascade LDA | Simple matching | 50 | 0.403 |
| | | | 150 | 0.387 |
| | | Recursive matching | 50 | 0.377 |
| | | | 150 | 0.359 |

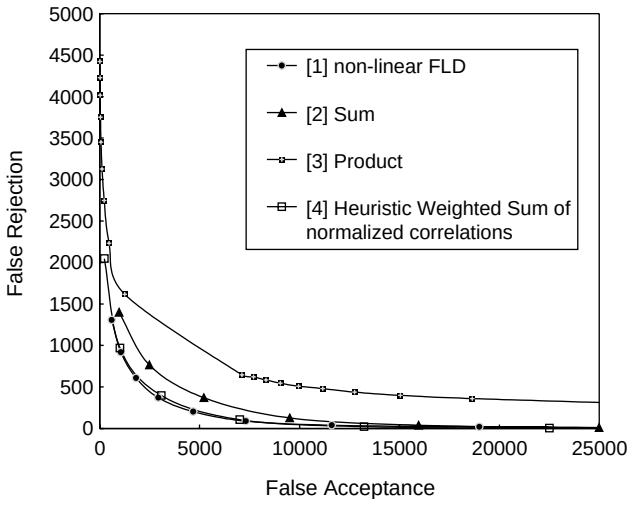



Fig. 8. Comparison of combining methods at the component level.

evaluated. As seen in Table 7, the performance of the advocated scheme was considerably better than the simple matching. The number of steps of the recursive matching scheme and the size of the buffers can be adjusted for a given database. In our experiment, the number of steps was set to 2 or 3 and the weight for the new queries was 0.8, as the probability of the first hit being correct hit is considerably high. It is noted that the recursive retrieval technique achieves the enhanced with negligible computational costs.

### 6.8. Computational complexity and size of descriptor

Table 8 compares our approaches in terms of computational complexity and size of descriptor. In feature extraction, the component LDA is simpler than the holistic LDA. Comparing the complexity and size of descriptors, the component LDA and the cascaded LDA require a more computations, compared with the holistic approach, but they show better generalization. The cascaded LDA reduces the matching complexity and the size of description for the combined LDA.

### 6.9. Comparisons with other MPEG-7 descriptors

Table 9 summarizes the comparitive results of the description techniques [2–4,8,22] proposed for the MPEG-7 meeting in May 2002. It is noted that the proposed component-based LDA method dramatically enhanced the retrieval performance of the previous methods. The extension work [16] of the proposed component and cascaded LDA techniques by using Fourier domains, which was recently proposed by Samsung AIT and NEC, yielded the best performance in the MPEG-7 working group, ISO/IEC JTC1/SC29/WG11. Its retrieval performance is 0.2728 and 0.3434 for the Experiment 3-1 and 3-2 respectively. It should be noted that the standardization

Table 8
Computational complexity and descriptor size

| | | Holistic LDA | Component LDA | Cascaded LDA |
|---|---|---|---|---|
| Feature extraction complexity | Additions | $N_0\times(N-1)=103{,}000$ | $5\times N_1\times(N_{avg}-1)=76{,}200$ | $\approx 179{,}200$ |
| | Multiplications | $N_0\times N=103{,}040$ | $5\times N_1\times N_{avg}=76{,}400$ | $\approx 179{,}440$ |
| Matching complexity | Additions | $3\times(N_0-1)=117$ | $5\times 3\times(N_1-1)+4=589$ | 147 |
| | Multiplications | $3\times N_0=120$ | $5\times(3\times N_1+1)=605$ | 150 |
| Size of descriptor in bits | | $40\times 4$ | $200\times 4$ | $50\times 4$ |

$N_0$, the number of elements of a holistic feature vector ($=40$); $N_1$, the number of elements of one component feature vector ($=40$); $N$, holistic input image size ($=46\times 56$); $N_{avg}$, average size of component input images ($=382$).

Table 9
Retrieval performance comparison with other MPEG-7 facial descriptors

| Methods (unit: ANMRR) | Exp. no. 2-3 | Exp. no. 2-4 |
|---|---|---|
| The second-order eigenface method [2] | 0.478 | N/A |
| The Fourier spectral PCA-based face description (with a confidence factor) [3] | 0.243 | N/A |
| The eHMM with the second-order eigenvectors [4] | 0.495 | N/A |
| The Pseudo2D-HMMs [8] | N/A | 0.554 |
| The component-based LDA [22] | 0.0678 | 0.1355 |

proposal submitted to the International Standardization body [20] in 2004 was successful.

## 7. Concluding remarks

In this paper, we proposed a face description method based on face image decomposition and the projection of each component by LDA. The component LDA augmented by the holistic LDA is then transformed by another LDA. The cascaded LDA method achieves impressive retrieval accuracy rates as compared with the conventional PCA/ ICA/LDA techniques and other description methods. The dimensionality of the descriptor and therefore its computational complexity are very low. The experimental results showed that the proposed description exhibits better generalization and that its application to large data sets is feasible.

## Acknowledgements

The authors would like to thank Seok-Cheol Kee in SAIT and Jiri Matas, Vojtech Franc in Center for Machine Perceptron for their helpful discussion, comment and efforts during MPEG-7 standardisation activity. We also thank Toshio Kamei in NEC and M.Z. Bober, the chair-person of the MPEG-7 working group and the many others working for the face descriptor in the MPEG-7 standard society. The database and protocol of this paper were obtained by their effort.

## References

[1] M. Abdel-Mottaleb, J.H. Connell, R.M. Bolle, R. Chellappa, Face descriptor syntex, Merging proposals P181, P551, and P650, ISO/MPEG m5207, Melbourne, 1999.
[2] L. Wang, T.K. Tan, Experimental results of face description based on the 2nd-order eigenface method, ISO/IEC JTC1/SC21/WG11/M6001, Geneva, May 2000.
[3] T. Kamei, A. Yamada, Report of core experiment on Fourier spectral PCA based face description, ISO/IEC JTC1/SC21/WG11 M8277, Fairfax, VA, May 2002. Also appeared in T. Kamei, Face retrieval by an adaptive Mahalanobis distance using a confidence factor, in: Proc. IEEE International Conference on Image Processing, Rochester, NY, 2002.
[4] M.-S. Kim, D. Kim, S. Lee, S. Jeong Kim, Experiment results of face descriptor using the embedded hmm with the 2nd-order block-specific eigenvectors, ISO/IEC JTC1/SC21/WG11 M8328, Fairfax, VA, May 2002.
[5] H.C. Kim, D. Kim, S.Y. Bang, Face retrieval using 1st- and 2nd-order PCA mixture model, International Conference on Image Processing, Rochester, NY, 2002.
[6] A. Nefian, B. Davies, Standard support for automatic face recognition, ISO/IEC JTC1/SC21/WG11/M7251, Sydney, July 2001. Also appeared in A. Nefian, M. Hayes, An embedded hmm-based approach for face detection and recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 6, 1999, pp. 3553–3556.
[7] L. Wiskott, J.-M. Fellous, N. Krüger, C. von der Malsburg, Face recognition by elastic bunch graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 775–779.
[8] C. Eckes, S. Eickeler, M. Larson, J. Löffler, K. Biatov, J. Köhler, Proposal of a face recognition descriptor based on Pseudo2D-HMMs, ISO/IEC JTC1/SC21/WG11 M8394, Fairfax, VA, May 2002.
[9] B. Heisele, P. Ho, T. Poggio, Face recognition with support vector machines: global versus component-based approach, in: Proceedings of the IEEE International Conference on Computer Vision, 2001.
[10] M.S. Bartlett, Face Image Analysis by Unsupervised Learning, Kluwer Academic Publishers, Dordrecht, 2001.
[11] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Recognition and Machine Intelligence 19 (7) (1997) 711–720.
[12] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Transactions on Pattern Recognition and Machine Intelligence 23 (3) (1997) 228–233.
[13] B. Moghaddam, A. Pentland, Face recognition using view-based and modular eigenspaces, SPIE Automatic Systems for the Identification and Inspection of Humans 2277 (1994).
[14] V. Franc, J. Matas, An extension of the component-based LDA descriptor by the Generalized Discriminant Analysis, ISO/IEC JTC1/SC21/WG11 M8727, Klagenfurt, AT, July 2002.
[15] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, New York, 2002.
[16] T. Kamei, A. Yamada, H. Kim, T.-K. Kim Wonjun Hwang, S. Cheol Kee, Advanced face descriptor using Fourier and inten-sity LDA features, ISO/IEC JTC1/SC29/WG11 M8998, Oct 2002.
[17] T.-K. Kim, H. Kim, W. Hwang, S. Cheol Kee, J. Ha Lee, Component-based LDA face descriptor for image retrieval, British Machine Vision Conference (BMVC), Cardiff, UK, Sept. 2–5, 2002.
[18] T.-K. Kim, H. Kim, W. Hwang, S. Cheol Kee, J. Kittler, Independent component analysis in a facial local residue space, IEEE International Conference on Computer Vision and Pattern Recognition, Madison, WI, 2003.
[19] T.-K. Kim, H. Kim, W. Hwang, S.-C. Kee, J. Kittler, Face description based on decomposition and combining of a facial space with LDA, IEEE International Conference on Image Processing, Spain, 2003.
[20] L. Cieplinski, A. Yamada, Text of ISO/IEC 15938-3/FPDAM1, ISO/IEC JTC1/SC29/WG11 N5695, July 2003.
[21] T. Vetter, T. Poggio, Linear object classes and image synthesis from a single example image, IEEE Transactions of PAMI 19 (7) (1997) 733–742.
[22] T.-K. Kim, H. Kim, W. Hwang, S. Cheol Kee, Component-based LDA face descriptor, ISO/IEC JTC1/SC29/WG11 M8243, Fairfax, VA, May 2002.

**Tae-Kyun Kim** is PhD student of Department of Engineering at the University of Cambridge. He received a B.Sc. and a M.Sc. degree in Dept. of EECS at Korea Advanced Institute of Science and Technology (KAIST) in 1998 and 2000 respectively. He worked as a research staff member of Samsung Advanced Institute of Technology. Korea during 2000-2004. This is also his period of obligatory military service. His research interests include computer vision, statistical pattern classification and machine learning. He reviews for the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) and is the korea delegate of MPEG-7. The joint face descriptor proposal of Samsung and NEC, for which he developed the main algorithms, has been adopted as the international standard of ISO/IEC JTC1/SC29/WG11.

**Hyunwoo Kim** received a BS degree in Electronic Communication Engineering from Hangyang University, Seoul, Korea, in 1994, and MS and PhD degrees in Electrical and Computer Engineering from POSTECH, Pohang, Korea, in 1996 and 2001, respectively. In 2000 he worked in the Institute for Robotics and Intelligent Systems at University of Southern California, Los Angeles, CA, as a visiting scientist. In 2001 he joined Samsung Advanced Institute of Technology, Korea, where he is currently a Research Scientist. His current research interests include computer vision, virtual reality, augmented reality and computer graphics.

**Wonjun Hwang** received both BS degree and MS degree from the Department of Electronics Engineering, Korea University, Seoul, Korea, in 1999, 2001, respectively. In 2002, he worked as an engineer for Samsung Electronics. He is now working as a researcher on face recognition for Samsung Advanced Institute of Technology. His research interests are in image processing, object detection, object recognition, and robot vision.

**Josef Kittler** is Professor of Machine Intelligence, and Director of the Centre for Vision, Speech and Signal Processing at the University of Surrey. He has worked on various theoretical aspects of Pattern Recognition and Image Analysis, and on many applications including personal identity authentication, automatic inspection, target detection, detection of microcalcifications in digital mammograms, video coding and retrieval, remote sensing, robot vision, speech recognition, and document processing. He has co-authored a book with the title 'Pattern Recognition: A statistical approach' published by Prentice-Hall and published more than 500 papers. He is a member of the Editorial Boards of Image and Vision Computing, Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications, and Machine Vision and Applications.