MMI-based Code Selection for Optimal Local Binary Patterns

Taewan Kim and Daijin Kim [1]

Department of Computer Science and Engineering
Pohang University of Science and Technology
San 31, Hyoja-Dong, Nam-Gu, Pohang, 790-784, Korea
{*taey16,dkim*}*@postech.ac.kr*

Abstract

Many variants of the local binary patterns (LBPs) are widely used for face analysis due to its simplicity and robustness. However, no one prove that they are optimal in the sense of minimization of the number of codes and minimization of the classification error. We propose an effective code selection method for the optimal LBP (OLBP), which is based on the maximization of mutual information (MMI) between features and class labels. We demonstrate the effectiveness of the proposed OLBP through many experiments of face recognition and facial expression recognition. Experimental results show that the OLBP outperforms other features such as LBP, ULBP, and MCT in terms of minimization of the number of codes and minimization of the classification error.

Key words: Local Binary Pattern, Feature Selection, Optimal LBP, Maximization of Mutual Information, Face Recognition, Facial Expression Recognition

1 Introduction

Face analysis including face recognition and facial expression recognition is a very active research area in computer vision, human computer interaction (HCI), and biometrics. In this field, the local binary patterns (LBPs) have been widely used as a powerful feature representation method. It has proven to be highly discriminative due to its invariance to the monotonic gray level changes [1] and to be highly efficient due to its fast computation. For these

[1] Corresponding author.
E-mail address: dkim@postech.ac.kr.

reasons, many research have been done using the LBPs [2]-[11]. Specifically, the LBP is one of the most successful representation for the face analysis.

However, the original LBP contains many less informative codes. It is observed that certain LBP codes, whose transitions from 1 to 0 or 0 to 1 in a circularly defined code are at most two, have been occurred frequently (more than 90%) in the natural images. Based on this observation, Ojala et al. [2] proposed the uniform LBP (ULBP) and applied it to face recognition. Lahdenoia et al. [4] proposed the symmetry ULBP which reduces the number of codes in the ULBP using the symmetry level of the code. However, it is not proven that they are optimal from the viewpoint of minimizing the number of codes and minimizing the classification error. Maximization of mutual information (MMI) is a key to guarantee the optimality.

Information theoretic feature selection, which utilizes the maximization of mutual information (MMI) between feature and class label, has been widely used for finding the optimal features because it guarantees the minimum classification error. Battiti [12] proposed the mutual information based feature selection (MIFS) and Kwak and Choi [13] proposed an improved version of MIFS (MIFS-U). Both MIFS and MIFS-U tried to find the optimal features one by one using the MMI. Principe [14] and Korkkola [15] used the steepest descent method to find the optimal projection basis vectors using the information theoretic error measures. Qiu and Fang [16] used the MMI to find the optimal projection basis vectors, and applied them to the face and car detection systems.

To find the optimal LBP (OLBP) without redundancy, we propose to use the information theoretic feature selection method based on the maximization of mutual information between the codes and the class labels. The proposed approach to code selection iteratively selects LBP codes which maximize their mutual information with respect to the class label, conditioned on any features previously selected.

This paper is organized as follows. Selection 2 describes the non-parametric local kernel based image representation such as LBP, ULBP, and MCT. Section 3 describes the MMI-based feature selection method. Section 4 describes a way of finding the optimal LBP (OLBP) codes using the MMI-based feature selection method. Section 5 explains the experimental results to demonstrate the effectiveness of the proposed OLBP in terms of minimizing the number of codes and minimizing the classification error. Finally, section 6 draws our conclusion.

2 Non-parametric local kernel based image representation

The original LBP [1] uses a 3 by 3 kernel that summarizes the local structure of an image. At a given pixel position (x_c, y_c), it takes the 3 by 3 neighborhood pixels surrounding of the given pixel and generates a binary 1 if the neighbor of the given pixel has a value greater than or equal to the given pixel or a binary 0 if the neighbor of the given pixel has a value smaller than the given pixel. The decimal form of the resulting 8-bit word (LBP code) can be expressed

$$LBP_{P,R}(x_c, y_c) = \sum_{n=0}^{7} \delta(i_n - i_c)2^n, \tag{1}$$

where i_c is a pixel value positioned at (x_c, y_c), i_n is one of the eight surrounding pixel values, and function $\delta(\cdot)$ is defined such that

$$\delta(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The subscripts P and R represent the number of neighboring pixels and the radius in multi-scale LBP, respectively [2]. For an example, $LBP_{8,2}$ denotes the LBP with 8 equally spaced pixels on a circle of radius 2.

Liao et al. [8] proposed the multi-scale block LBP (MB-LBP). It captures an n by n block-based local structure rather than pixel-based local structure so that it is less sensitive to noisy information.

Ojala et al. [2] observed that the natural images generally contain a small number of LBP codes, which are called the uniform LBP (denoted by superscript LBP^{u2}). ULBP contains two bitwise 0 to 1 or 1 to 0 transitions at most. These uniform patterns mainly represent the majority of micro structures such as lines, edges, and corners.

Lahdenoia et al. [4] proposed a method for reducing the number of the codes in the ULBP using the level of symmetry L_{sym} of ULBP, which is expressed as

$$L_{sym} = \min\Big[\sum_{i=1}^{P} B(i), \sum_{i=1}^{P} \overline{B}(i)\Big], \tag{3}$$

where the first and second term are the number of neighboring pixels with a binary value 1 and 0 in ULBP, respectively. They observed that the codes with high level of L_{sym} are more discriminative than those with low level of L_{sym} by qualitative visual inspection.

Zabih et al. [17] proposed the census transform (CT) that summarizes the local image structure as a bit string, where it is 0 if the intensity value at

a position in one image is less than the intensity value at the corresponding position in another image. This census transform has been extended to the modified census transform (MCT) [18] as

$$\Gamma_{MCT}(x_c, y_c) = \sum_{n=0}^{8} \delta(i_n - \bar{i}_c) 2^n, \tag{4}$$

where $\bar{i}_c$ denotes the mean of pixel values in a 3 by 3 local kernel positioned at (x_c, y_c), and i_n is one of the nine pixel values in the local kernel. The function $\delta(\cdot)$ is the same as Eq. (2). MCT can be referred to as an enlarged version of the original LBP, which means one pixel in the image is represented by 9 bit length. Hence, the number of codes is 512 in MCT while it is 256 in the LBP.

Fig. 1 illustrates that the LBP, ULBP, and MCT are invariant to the monotonic gray level changes.

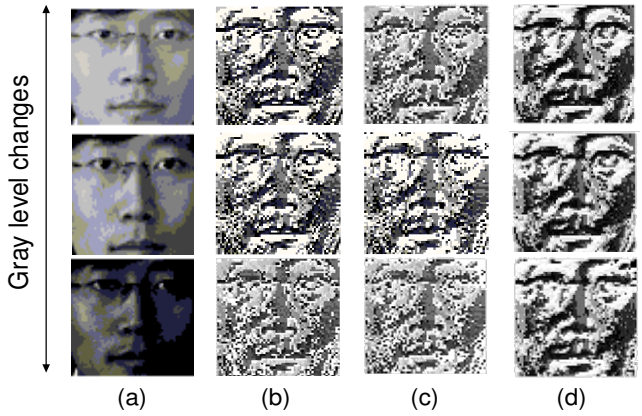

Fig. 1. Robustness to the monotonic gray level changes: (a) original image, (b) LBP, (c) ULBP, and (d) MCT.

3 MMI-based feature selection

Maximization of mutual information (MMI) is a powerful way of selecting the optimal feature that minimizes both the lower and upper bound of the Bayes error, simultaneously. Let $\mathbf{X} \in \mathbb{N}^{D \times N}$ be a data matrix, where D is the number of training images and N is the feature size of each training image. Let $F = \{f_1, \ldots, f_N\}$ and C be a discrete valued random variable for representing features and class labels, respectively. Fig. 2 illustrates a typical example of the data matrix $\mathbf{X}$, where $\mathbf{T}$ is a set of training samples as $\mathbf{T} = \{(f_1, C), \ldots, (f_N, C)\}$.

Let a function be $G(F) = \hat{C}$, where $\hat{C}$ is an estimate of C and the C has the class labels as $\nu = \{1, \ldots, N_c\}$, where N_c is the total number of classes. Then,

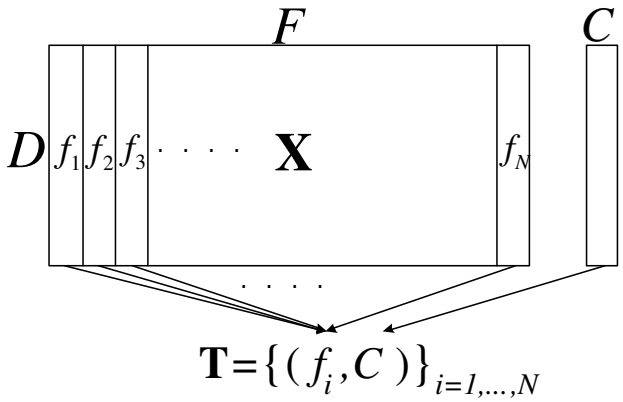

Fig. 2. A typical example of representing data matrix $\mathbf{X}$.

the lower and upper bounds of Bayes error probability $P_e \triangleq P(\hat{C} \neq C)$ are proven by Fano [19] and Hellman and Raviv [20] as

$$\frac{H(C) - I(C;F) - 1}{\log|\nu|} \leq P_e \leq \frac{1}{2}\Big(H(C) - I(C;F)\Big), \tag{5}$$

where $H(\cdot)$ denotes entropy, $I(\cdot;\cdot)$ denotes mutual information, and $|\nu|$ is the number of classes. The Bayes error probability P_e can be directly reduced by maximizing mutual information. From Eq. (5), we know that the maximization of mutual information (MMI) is equivalent to the minimization of the Bayes error probability. Also, we know that the maximization of mutual information (MMI) is equivalent to the minimization of the conditional entropy $H(C|F)$, since $I(C;F) = H(C) - H(C|F)$. Therefore, the optimal feature selection problem can be formulated as

$$\begin{cases} \arg\max_{f_i \in F} I(C; f_i) \\ \qquad \text{or} \qquad\qquad , \quad \forall i = \{1, \ldots, N\}. \\ \arg\min_{f_i \in F} H(C|f_i) \end{cases} \tag{6}$$

Since we need to find k features in a real application, Eq. (6) is modified as

$$\arg\max_{f_i \in F} I(C; f_1, f_2, \ldots, f_k), \quad k \ll N. \tag{7}$$

However, it is impossible to compute a joint mutual information Eq. (7) practically, because all possible combinations of feature sets is huge (the exact number of possible subsets of selected features is $\frac{N!}{(N-k)!k!}$). To overcome this problem, many researchers tried to find the approximated solutions. Battiti [12] proposed an iterative greedy feature selection strategy called 'Mutual Information Feature Selection (MIFS)'. He used the greedy feature selection criterion

as

$$\arg\max_{f_i \in F} \left[I(C; f_i) - \beta \sum_{f_s \in S} I(f_s; f_i) \right], \tag{8}$$

where f_i is a candidate feature in the feature set F, f_s is a previously selected feature, S is a set of the previously selected features, and β is a regularization parameter that adjusts the amount of redundance between the candidate features f_i and previously selected features f_s. Table 1 shows a typical algorithm of the MIFS method.

Table 1
Battiti's MIFS method.

1. Initialization

(1) $S \leftarrow \phi$.

2. Computation of the Mutual Information

(1) $\arg\max_{f_i \in F} I(C; f_i)$.

3. Selection of the first feature

(1) $F \leftarrow F \setminus \{f_i\}$, $S \leftarrow \{f_i\}$.

4. Greedy Selection

repeat until $|S| = k$

(1) For all pairs (f_i, f_s) with $f_i \in F$ and $f_s \in S$,

select the feature $f_i \in F$ using Eq. (8).

(2) Set $F \leftarrow F \setminus \{f_i\}$, and $S \leftarrow \{f_i\}$.

end

Kwak et al. [13] proposed the MIFS-U method that modifies the selection criterion as

$$\arg\max_{f_i \in F} \left[I(C; f_i) - \beta \sum_{f_s \in S} \frac{I(C; f_s)}{H(f_s)} I(f_s; f_i) \right]. \tag{9}$$

Peng et al. [21] proposed a modified MIFS method that uses the max-relevance and min-redundancy (mRMR) criterion as

$$\arg\max_{f_i \in F} \left[I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} I(f_s; f_i) \right], \tag{10}$$

where $|S|$ is the cardinality of the set S.

Estevez et al. [22] proposed another modified MIFS method that uses the

selection criterion as

$$\arg\max_{f_i \in F} \left[I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} NI(f_i; f_s) \right], \tag{11}$$

where NI is the normalized mutual information such that

$$NI(f_i; f_s) = \frac{I(f_i; f_s)}{\min\{H(f_i), H(f_s)\}}. \tag{12}$$

4 MMI-based OLBP code selection

We applied the MMI-based feature selection method to find the OLBP code. In this work, we took the Peng's modified MIFS method that used the max-relevance and min-redundancy (mRMR) as the selection criterion because it showed outstanding classification performance than the other iterative feature selection methods [23,24]. Fig. 3 shows a procedure of the proposed MMI-based OLBP code selection method, which consists of three consecutive stages. A detailed explanation about the MMI-based OLBP code selection is given below.

4.1 Stage I: MMI-based feature reduction

Suppose that we have a set of D training images with N_c classes, each of the images has a size of $N = h \times w$. All training images are transformed into LBP features. Then, we have a LBP feature matrix with a size of $D \times N$, $F_{LBP} = \{f_1, f_2, \ldots, f_N\}$, where f_i is a D dimensional LBP feature vector at the ith pixel position.

We compute the mutual information $I(C; f_i)$ between the class label C and the feature vector f_i for $i = 1, 2, \ldots, N$ and then obtain the selected feature index set $S_{LBP} = \{p_1, p_2, \ldots, p_M\}$ using the maximization of the mutual information given in Eq. (10), where M is the number of selected LBP feature vectors and the p_i denotes the index of the selected LBP feature vector at the ith iteration. Thus, we have a reduced LBP feature matrix with a size of $D \times M$, $F'_{LBP} = \{f_{p_1}, f_{p_2}, \ldots, f_{p_M}\}$, where f_{p_i} is a D dimensional LBP feature vector at the p_ith pixel position. Because $M \ll N$, we reduce the dimensionality of the original LBP feature matrix with removing less discriminative features greatly.

Fig. 4 shows two examples of the feature dimensionality reduction using the MMI-based feature reduction method on the POSTECH Face 2007 database

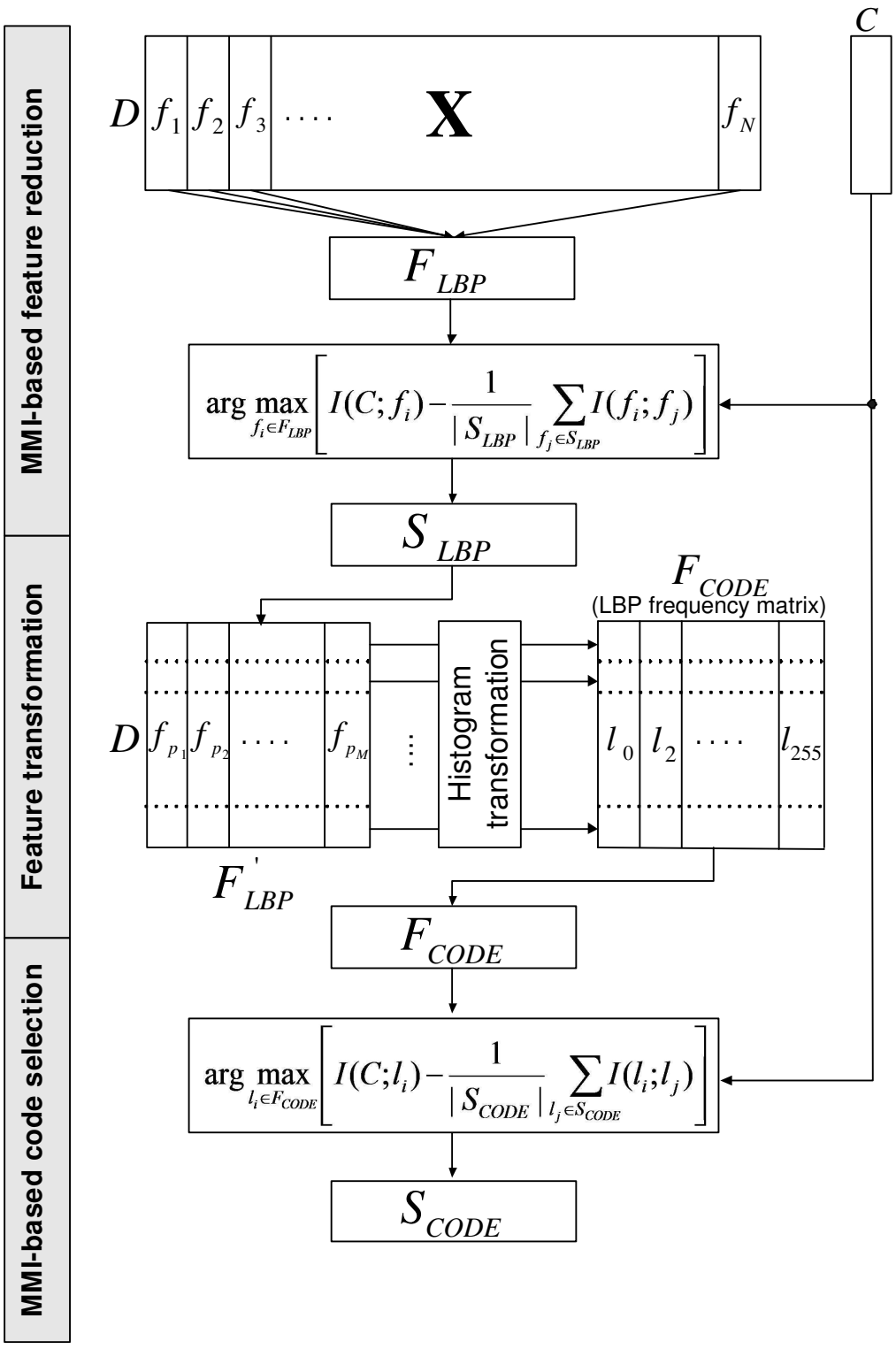

Fig. 3. Procedure of the proposed MMI-based OLBP code selection method.

(PF07) [25], where the horizontal axis denotes the number of selected features. Discriminative components of a face are eyes, eyebrows, and mouth because the selected features mainly contain information about them.

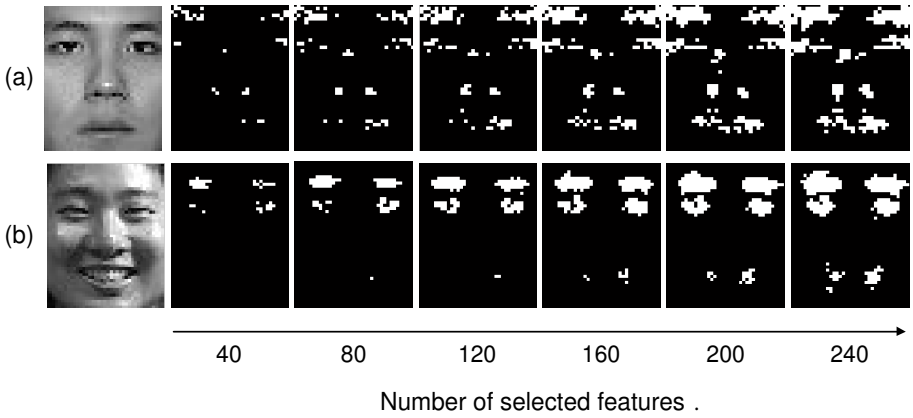

Fig. 4. Examples of feature dimensionality reduction using (a) face recognition data and (b) facial expression data on PF07.

4.2 Stage II: Feature transformation

Each row of the reduced LBP feature matrix F'_{LBP} is a LBP transformed training image with the reduced size of M. Each of the dimensionality reduced training image is transformed into a histogram of LBP (We call it histogram transformation). Then, we have a LBP code frequency matrix with a size of $D \times 256$, $F_{CODE} = \{l_0, l_1, \ldots, l_{255}\}$, where l_i is a D dimensional LBP code frequency vector at a specific LBP code i. Hence, the jth column of l_i has the number of pixels whose LBP code is i in the jth training image.

4.3 Stage III: MMI-based code selection

We again compute the mutual information $I(C; l_i)$ between the class label C and the LBP code frequency vector l_i for $i = 0, 1, \ldots, 255$ and then obtain the selected LBP code frequency set $S_{CODE} = \{c_1, c_2, \ldots, c_K\}$ using the maximization of the mutual information given in Eq. (10), where K is the number of the selected LBP code frequency vectors and the c_i is the index of the selected LBP frequency vector at the ith iteration. Therefore, S_{CODE} is OLBP codes we want to find.

Fig. 5 shows examples of four transformed images using the OLBP codes that are obtained from the MMI-based OLBP code selection method. The horizontal axis denotes the number of the OLBP codes. From the Fig. 5, we can represent many local structures in more detail as the number of OLBP codes

increases up to 40. However, we cannot observe a significant improvement of representing local structures in more detail after the number of OLBP codes exceeds 40. From this, we know that there exists an optimal number of LBP codes from the viewpoint of the number of OLBP codes and the classification error.

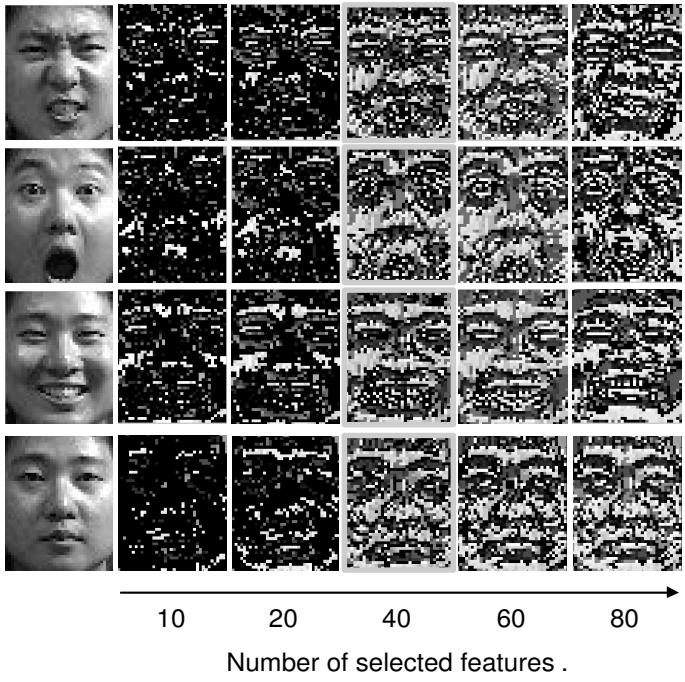

Fig. 5. Some examples of the OLBP transformed images.

5 Experimental results and discussion

To validate the effectiveness of the proposed MMI-based OLBP code selection method, we performed two types of experiments: face recognition and facial expression recognition. In these experiments, all input images were transformed using the OLBP that is obtained by the proposed MMI-based OLBP code selection method. We did not use any preprocessing methods or postprocessing methods to show that effectiveness of only the proposed method in terms of a number of codes and classification performance.

5.1 Database and experimental setting

We used the PF07 database [2] [25], which consists of 100 male and 100 female subjects with 320 images per subject. The 320 images has a size of 40×50

[2] The PF07 database is available on the web, `http://imlab.postech.ac.kr`

that includes consist 4 facial expressions, 16 illuminations, and 5 poses. Fig. 6 shows some example images of the PF07 database. All of these images are aligned according to the two eye locations, which are obtained manually. This database can be directly used for our experiment without any preprocessing or manual labelling.

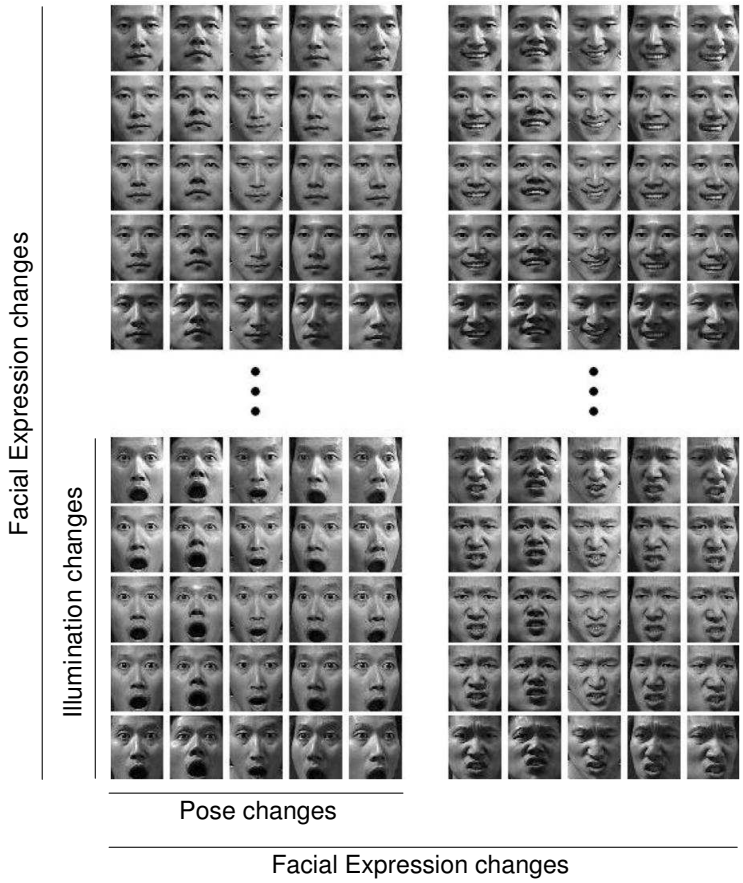

Fig. 6. Example images of PF07.

5.2 Face recognition

For the experiments of face recognition, we prepared a training set, three probe sets, and a gallery set as follows. The training set consisted of 32000 randomly selected images (= 100 persons × 16 different illuminations × 4 different facial expressions × 5 different poses). The three probe sets consisted of three differently conditioned image sets such as illumination changes, pose changes, and facial expression changes. First, illumination conditioned probe set consisted of 1500 images (=100 persons × 15 non-normal illuminations × the neutral expression × the frontal pose). Second, pose conditioned probe set consisted of 500 images (=100 persons × the normal illumination × the neutral expression × 4 non-frontal poses). Third, facial expression conditioned

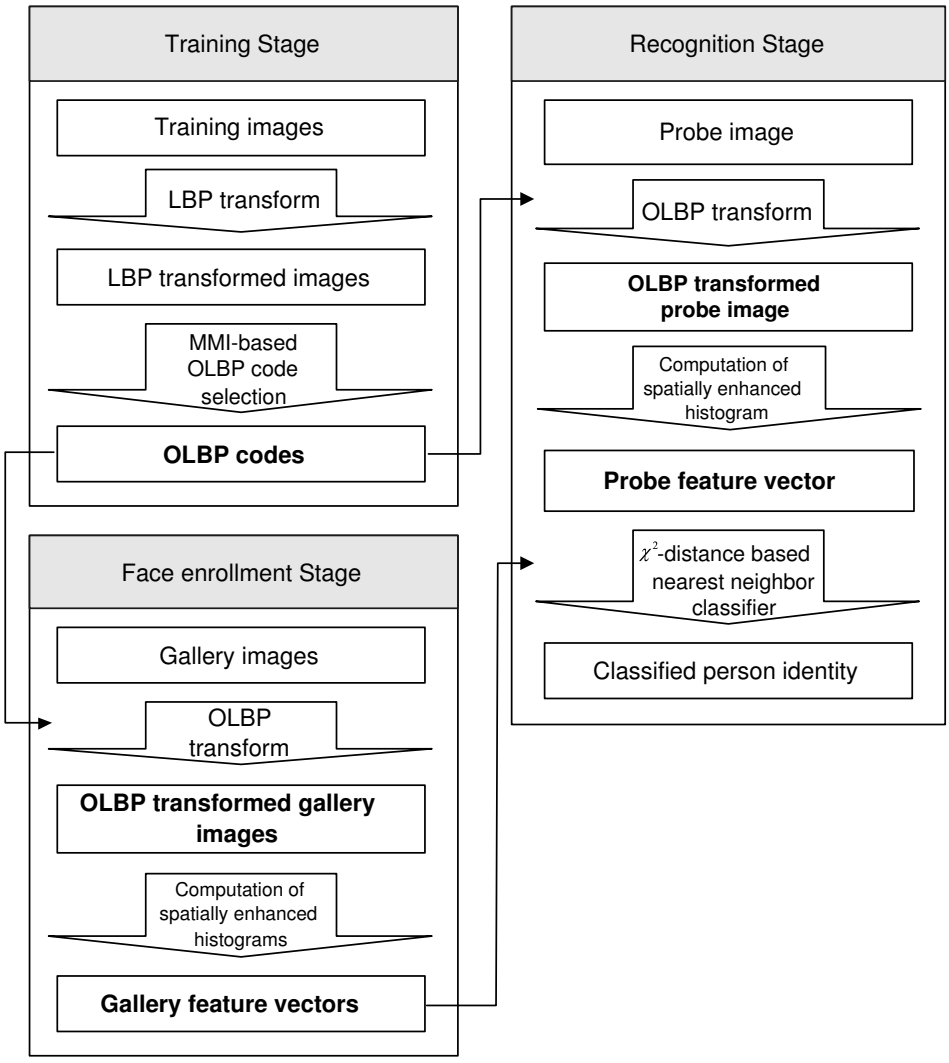

Fig. 7. Overall process of the face recognition.

probe set consisted of 300 images (=100 persons × the normal illumination × 3 non-neutral facial expressions × the frontal pose). A gallery set consisted of 100 images (=100 persons × the normal illumination × the neutral expression × the frontal pose). Hence, we performed three different experiments with a pair of both each probe set and the gallery set. To validate generality of the proposed method, individuals composing each training set and probe set were selected exclusively. However, the three probe sets and the gallery set contained same individuals.

Fig. 7 illustrates the overall process of the face recognition experiment, which consists of three stages.

1. Training stage
   We obtained the LBP transformed images by converting the training images with LBP and found the K OLBP codes using the MMI-based OLBP

code selection method in the section 4. In this experiment, we empirically reduced the dimensionality of the training images from 2,000 to 240 in the MMI-based feature reduction stage. The 240 features included some discriminative facial components such as eyes, eyebrows, mouth, and nose (See Fig. 4).

2. Enrollment stage
   We transformed the gallery images into the OLBP transformed gallery images by the OLBP codes which are obtained in the training stage. Then, we converted the OLBP transformed gallery images into a set of gallery feature vectors using the spatially enhanced histogram method [5], which is explained as follows. (1) Each OLBP transformed gallery image was divided into 30 local regions $R_i, i = 1, ..., 30$, (We empirically divided the image into 30 regions.) (2) the OLBP histograms of 30 local regions were computed independently in each OLBP transformed gallery image, and (3) the gallery feature vector was obtained by concatenating the OLBP histograms of 30 local regions sequentially.

3. Recognition stage
   We transformed the probe image into the OLBP transformed probe image by the OLBP codes. We converted the OLBP transformed probe image into a probe feature vector computed by the same method in the enrollment stage. Then, we measured the weighted χ^2 distance between the probe feature vector ($F_{P_{k,l}}$) and gallery feature vectors ($F^i_{G_{k,l}}, i = 1, \ldots, N_G, N_G = \{1500, 500, 300\}$) to find the best matched gallery image as

$$\arg\min_i \left[\chi^2_w \left(F_{P_{k,l}}, F^i_{G_{k,l}} \right) \right], \quad \forall i, \tag{13}$$

   where the weighted χ^2 distance is computed as

$$\chi^2_w \left(F_{P_{k,l}}, F^i_{G_{k,l}} \right) = \sum_{k,l} w_k \frac{\left(F_{P_{k,l}} - F^i_{G_{k,l}} \right)^2}{\left(F_{P_{k,l}} + F^i_{G_{k,l}} \right)}, \tag{14}$$

   where the indices l and k refer to lth bin in a histogram corresponding to the kth local region. w_k is the weight for region k. We empirically set the weights w_k.

The classification error for the face recognition could be computed such that

$$\text{Classification error } (\%) = \left(1 - \frac{N_{\text{CRI}}}{N_{\text{G}}} \right) \times 100, \tag{15}$$

where N_{CRI} and N_{G} is # of correctly recognized images and # of gallery im-

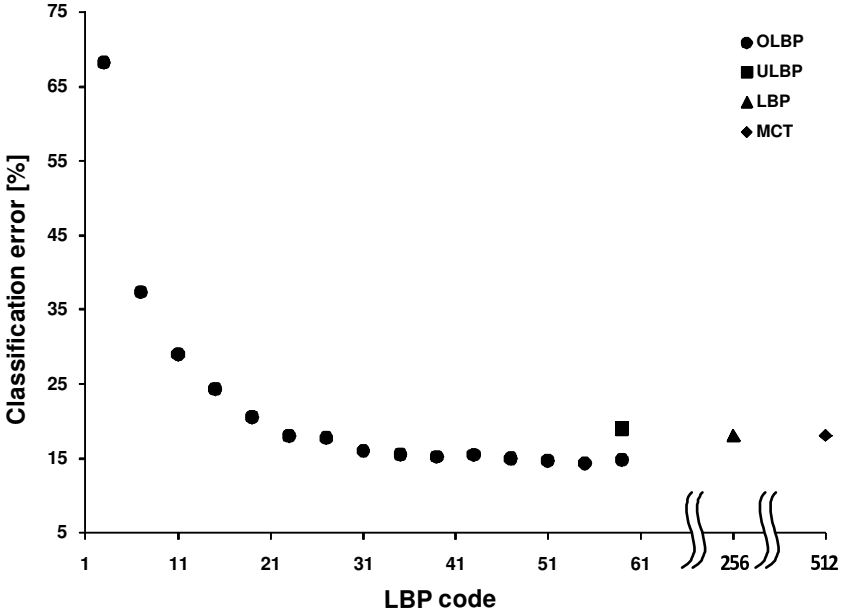

Fig. 8. Comparison of classification errors for face recognition among four different representation methods using the illumination conditioned probe set and the gallery set.

ages, respectively. Fig. 8 compares the classification errors of face recognition using the illumination conditioned probe set and the gallery set among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the horizontal axis denotes the number of OLBP codes. From the Fig. 8, we noticed that the classification error decreases drastically as the number of OLBP codes increases up to 23, and does not change much as the number of OLBP codes increases from 23 to 59. From the viewpoint of the number of codes and the classification error, the best number of OLBP code was 31. The classification error using 31 OLBP codes was 16.00%, which is the smallest than those of other representation methods such as ULBP (18.73%), LBP (18.07%), and MCT (18.07%). Fig. 9 compares the classification errors of face recognition using the pose conditioned probe set and the gallery set among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the horizontal axis denotes the number of OLBP codes. From the Fig. 9, we noticed that the classification error decreases drastically as the number of OLBP codes increases up to 19, and does not change much as the number of OLBP codes increases from 19 to 59. From the viewpoint of the number of codes and the classification error, the best number of OLBP code was 27. The classification error using 27 OLBP codes was 14.75%, which is the smallest than those of other representation methods such as ULBP (19.00%), LBP (20.25%), and MCT (20.25%). Fig. 10 compares the classification errors of face recognition using the facial expression conditioned probe set and the gallery set among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the hori-

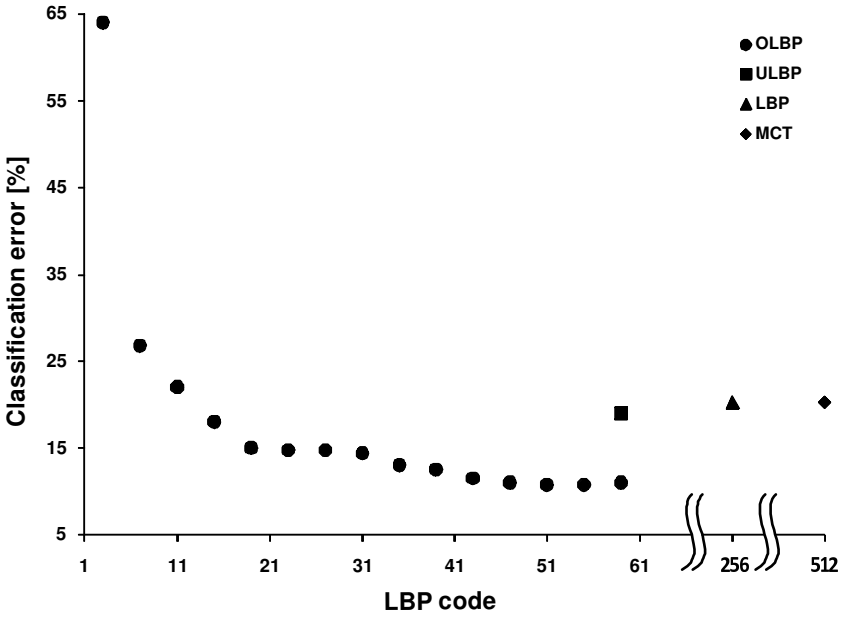

Fig. 9. Comparison of classification errors for face recognition among four different representation methods using the pose conditioned probe set and the gallery set.

zontal axis denotes the number of OLBP codes. From the Fig. 10, we noticed that the classification error decreases drastically as the number of OLBP codes increases up to 19, and does not change much as the number of OLBP codes increases from 19 to 59. From the viewpoint of the number of codes and the classification error, the best number of OLBP code was 35. The classification error using 35 OLBP codes was 18.16%, which is the smallest than those of other representation methods such as ULBP (21.50%), LBP (20.83%), and MCT (20.83%).

Optimal LBP codes gave reduced recognition time. In Ahonen's face recognition method [5], recognition time is totally dependent on the χ^2 distance based matching Eq. (16). Time complexity of Eq. (16) is represented as a linear function $\mathcal{O}(l)$, where $\mathcal{O}(\cdot)$ denotes Big-$\mathcal{O}$ function whose parameter l is the dimension of the feature vector. More specifically, $l = R \times B$, where R is the number of local regions, and B is the number of bins per region. Reducing the number of codes gives reducing the number of bins per region. Therefore, we can expect that if we reduce the original number of codes B to B', we reduce the operation as an amount of $R \times (B - B')$. Fig. 11 compares the computation times of face recognition using the illumination conditioned probe set and the gallery set among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the horizontal axis denotes the number of OLBP codes. From the Fig. 11, we noticed that the computation time increases almost linearly as the number of OLBP codes increases. The recognition time using 31 OLBP codes was 3.26 seconds, which is the smallest than those of other representation methods such as ULBP (4.95

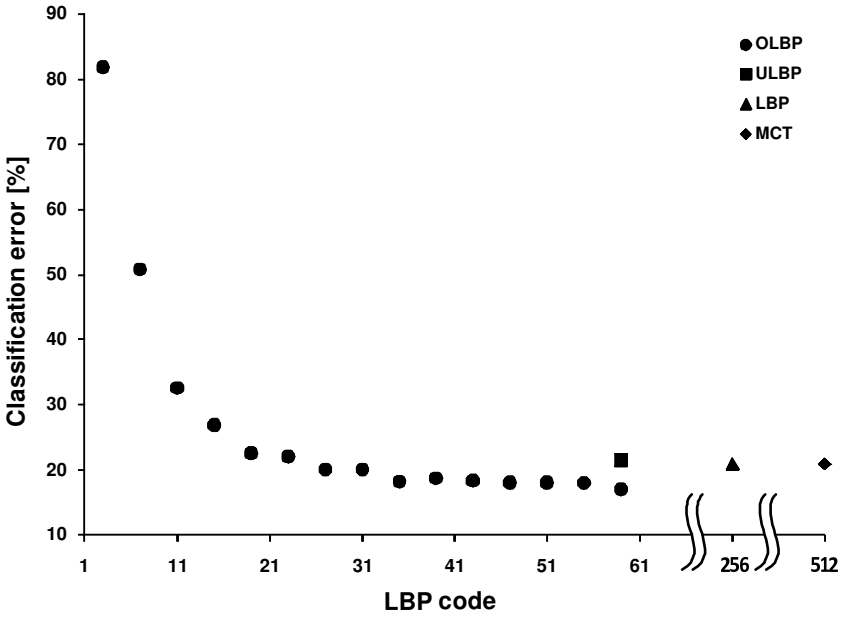

Fig. 10. Comparison of classification errors for face recognition among four different representation methods using the facial expression conditioned probe set and the gallery set.

seconds), LBP (15.65 seconds), and MCT (43.07 seconds). Hence, if we use 31 OLBP codes, we can reduce the computation time by about 1.5 times, 5 times, and 13 times when ULBP, LBP, and MCT are used respectively.

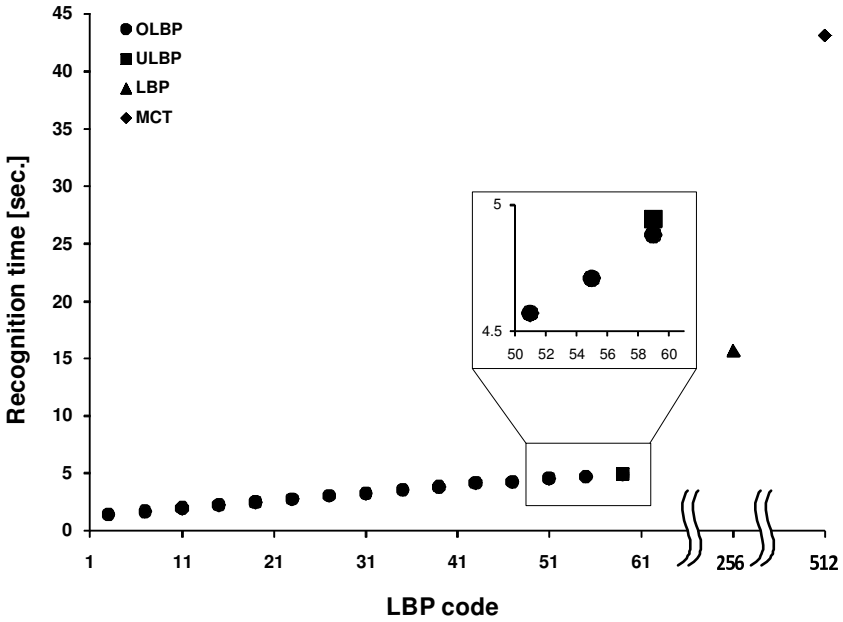

Fig. 11. Comparison of computation times for face recognition among four different representation methods using the illumination conditioned probe set and the gallery set.

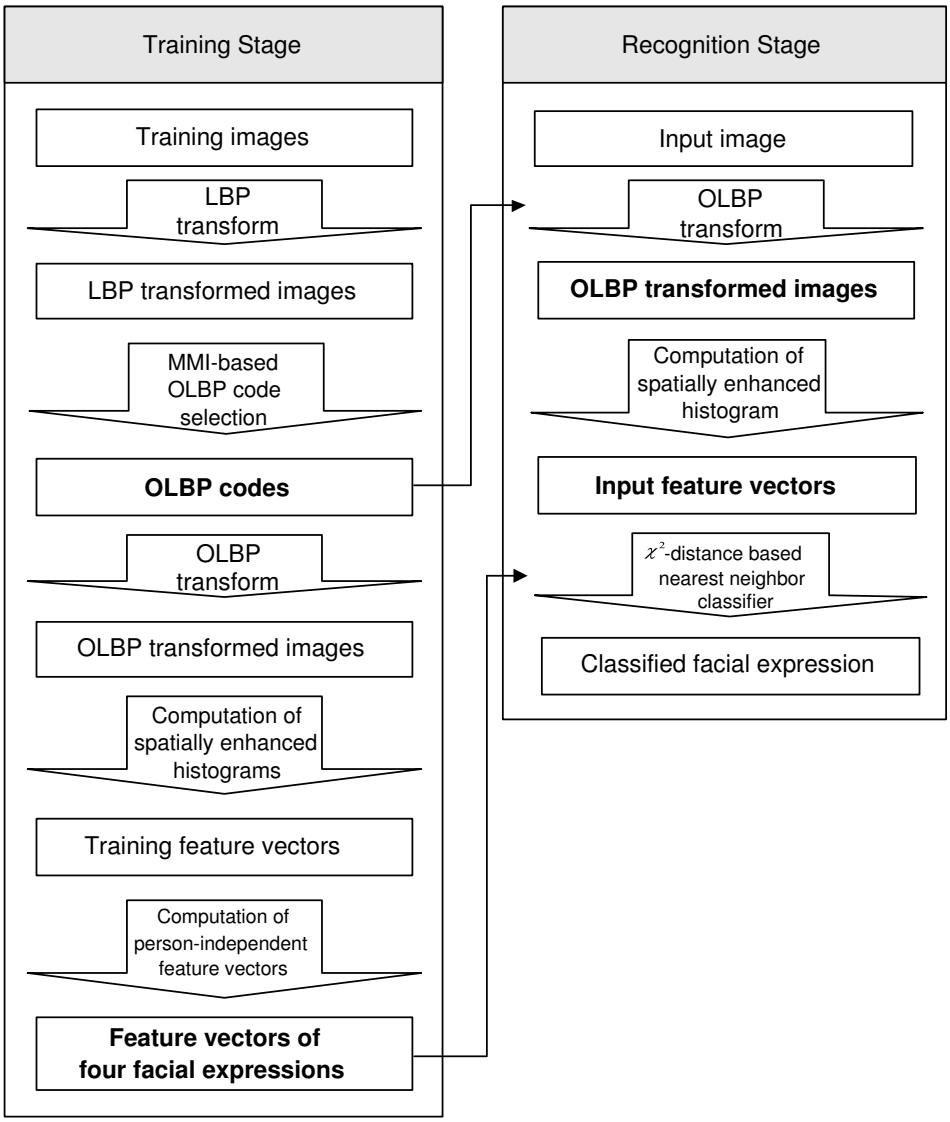

Fig. 12. Overall process of the facial expression recognition.

5.3 Facial expression recognition

For the experiments of facial expression recognition, we prepared the training and test image set as follows. First, we prepared 800 randomly selected images (= 200 persons × a normal illumination × 4 different facial expressions × a frontal pose). Among 800 images, we selected 700 images (= 175 persons × a normal illumination × 4 facial expressions × a frontal pose) for the training image set and selected the remaining 100 images (= the remaining 25 persons × a normal illumination × 4 facial expressions × a frontal pose) for the test image set. We performed 8-fold cross validation method to avoid the data tweak problem.

Fig. 11 illustrates the overall process of the facial expression recognition ex-

periment, which consists of two stages.

1. Training stage
   We obtained the LBP transformed images by converting the training images with LBP and found the K OLBP codes using the MMI-based OLBP code selection method in the section 4. In this experiment, we empirically reduced the dimensionality of the training images from 2,000 to 240 in the feature reduction stage. This 240 features included discriminative facial components such as eyes, eyebrows, and mouth (See Fig. 4). We converted the LBP transformed training images into the OLBP transformed training images by using the obtained K OLBP codes. Then, we converted the OLBP transformed training images into a set of training feature vectors using the spatially enhanced histogram method [5], which is the same as the method in section 5.2. Finally, we obtained four person-independent feature vectors, one per facial expression, by averaging all training feature vectors with a specific facial expression [10].

2 Recognition stage
   We transformed the input image into the OLBP transformed input images by the K OLBP codes. We converted the OLBP transformed training images into a input feature vector using the spatially enhanced histogram method [5], which is the same as the method in the section 5.2. Finally, we measured the χ^2 distance between the input feature vector (F_I) and the person-independent feature vectors of four facial expressions ($F_e,\ e \in \{$netural, happy, surprised, angry$\}$) to find the best matched facial expression such that

$$\arg\min_e \left[\chi^2\Big(F_I, F_e\Big) \right], \quad \forall e, \tag{16}$$

   where the χ^2 distance is computed as

$$\chi^2\Big(F_I, F_e\Big) = \sum_l \frac{\Big(F_I(l) - F_e(l)\Big)^2}{\Big(F_I(l) + F_e(l)\Big)}, \tag{17}$$

   where l is the component index of the feature vector F_I and F_e.

To get a reliable facial expression recognition performance, we performed the 8-fold cross validation method. In ith trial ($i = 1, \ldots, 8$), we had 100 test input images, 25 images for each facial expression and we counted the number of correctly recognized images in facial expression. Then, the classification error

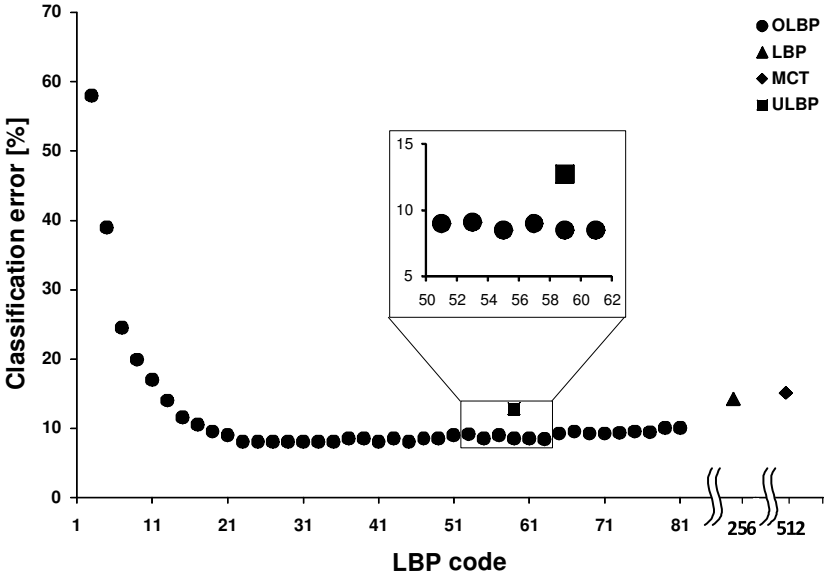

Fig. 13. Comparison of classification errors for facial expression recognition among four different representation methods.

for the facial expression recognition could be computed such that

$$\text{Classification error } (\%) = \left(1 - \frac{1}{8}\sum_{i=1}^{8}\frac{N_{\text{CRI}i}}{N_{\text{T}i}}\right) \times 100, \tag{18}$$

where $N_{\text{CRI}i}$ and $N_{\text{T}i}$ is # of correctly recognized images at the ith trial, and # of ith set of test images, respectively. Fig. 13 compares the classification errors of facial expression recognition among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the horizontal axis denotes the number of OLBP codes. From the Fig. 13, we noticed that the classification error decreases drastically as the number of OLBP codes increases up to 23, and does not change much as the number of OLBP codes increases from 23 to 80. From the viewpoint of the number of codes and the classification error, the best number of OLBP code was 23. The classification error using 23 OLBP codes was 8.00%, which is the smallest than those of other representation methods such as ULBP (12.73%), LBP (14.20%), and MCT (15.07%).

In this facial expression recognition, the time complexity is the same as the face recognition experiments because the χ^2 distance is used for finding correct matching. Fig. 14 compares the computation times of face recognition among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the horizontal axis denotes the number of OLBP codes. From the Fig. 14, we noticed that the computation time increases almost linearly as the number of OLBP codes increases. The recognition time using 23 OLBP codes was 0.5928 seconds, which is the smallest than those of other representation methods such as ULBP (0.8892

seconds), LBP (2.3088 seconds), and MCT (4.3056 seconds). Hence, if we use 23 OLBP codes, we can reduce the computation time by about 1.5 times, 4 times, and 7.2 times when ULBP, LBP, and MCT are used respectively.

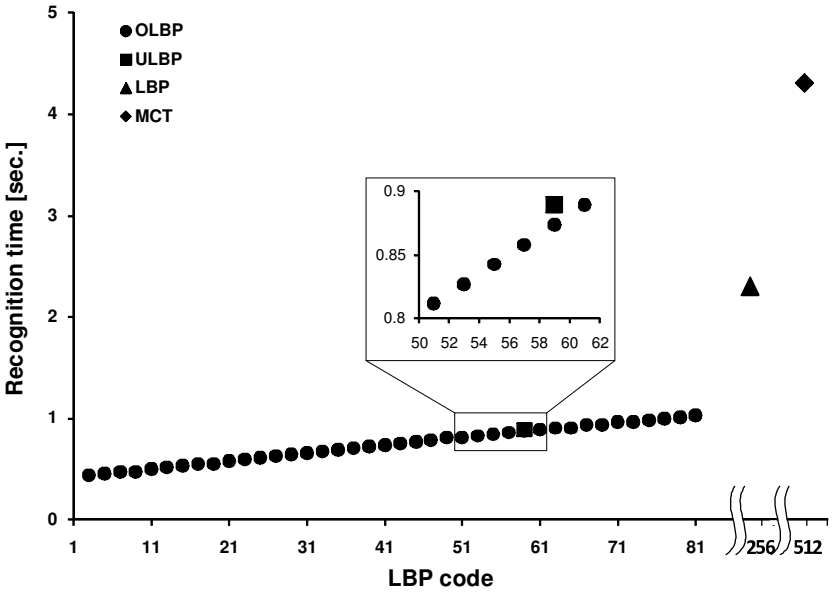

Fig. 14. Comparison of computation times for facial expression recognition among four different representation methods.

In previous facial expression recognition experiment, we used only randomly selected 10 person's images (=800 images). To validate the effectiveness of the proposed OLBP more reliably than the previous one, We performed another facial expression recognition experiment with large-scale data. We prepared 12800 facial expression images (12800 images = 4 different facial expressions × 200 persons × 16 different illuminations × a frontal pose). We divided the images into two sets. Each set contained 4 facial expression images with 200 person's face images × 8 different illuminations × a frontal pose. Between these two sets, one set was used to find OLBP codes (train), another set was used for test images. Overall procedure of this experiment was similar to the previous experiment (See Fig. 12), however distance measure slight differed from that of the previous experiment. In this experiment, we measured the weighted χ^2 distance between the input feature vector (F_I) and the person-independent feature vectors of four facial expressions ($F_e, e \in \{$netural, happy, surprised, angry$\}$) to find the best facial expression such that

$$\arg\min_{e} \left[\chi_w^2 \left(F_{I_{k,l}}, F_{e_{k,l}} \right) \right], \quad \forall e, \tag{19}$$

where the weighted χ^2 distance is computed as

$$\chi_w^2\left(F_{I_{k,l}}, F_{e_{k,l}}\right) = \sum_{k,l} w_k \frac{\left(F_{I_{k,l}} - F_{e_{k,l}}\right)^2}{\left(F_{I_{k,l}} + F_{e_{k,l}}\right)}, \tag{20}$$

where the indices l and k refer to lth bin in a histogram corresponding to the kth local region. w_k is the weight for region k. We empirically divided each face image into 30 blocks ($k = 1, \ldots, 30$), and empirically set the weights w_k. We used 2-fold cross validation method. Then, classification error was similarly computed as Eq. (18).

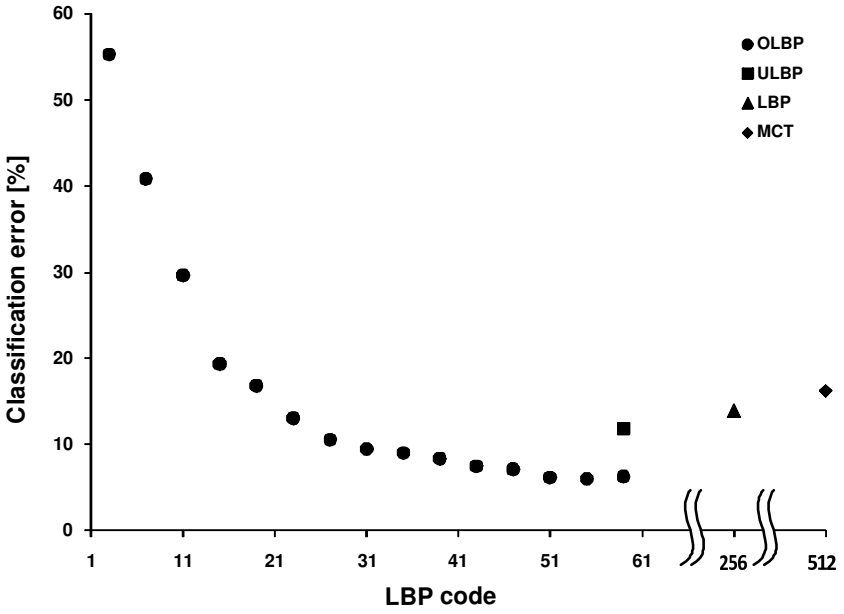

Fig. 15. Comparison of facial expression recognition for 100 person's facial expression among four different representation methods.

Fig. 15 compares the classification errors of 100 person's facial expression recognition among four different representation methods such as OLBP, ULBP, LBP, and MCT when the number of OLBP codes changes, where the horizontal axis denotes the number of OLBP codes. From the Fig. 15, we noticed that the classification error decreases drastically as the number of OLBP codes increases up to 27, and does not change much as the number of OLBP codes increases from 27 to 59. From the viewpoint of the number of codes and the classification error, the best number of OLBP code was 39. The classification error using 39 OLBP codes was 8.31%, which is the smallest than those of other representation methods such as ULBP (11.8%), LBP (13.86%), and MCT (16.16%). In this case, we needed more OLBP codes than the previous facial expression recognition (See Fig. 13), because database was much more complicated than that used in the 10 person's facial expression recognition experiment.

6 Conclusion

This paper proposed the MMI-based code section method for the optimal LBP, which provides the improved recognition performance and the reduced recognition time, simultaneously. Because the maximization of mutual information (MMI) between feature and class label assures the minimal classification error, we selected the codes of the optimal LBP iteratively, in the order of mutual information per code.

The proposed OLBP code selection method consisted of three stages: MMI-based feature reduction, feature transformation, and MMI-based code selection. In the first stage, we reduced the dimensionality of training images using mutual information. The selected LBP feature vectors were the most discriminative than any other remaining features. In the second stage, the dimensionality reduced training images were transformed into histograms of LBP by the histogram transformation. Then, we had a LBP code frequency matrix. Each column vector of the matrix represented a feature vector of corresponding LBP code. In the last stage, we selected several LBP code frequency vectors using MMI-based code selection, and then we finally could get the indices of the selected LBP frequency vectors.

To validate the effectiveness of the MMI-based OLBP code selection method, We applied it to two applications: face recognition and facial expression recognition. First, in face recognition experiments, we used three differently conditioned probe sets, the illumination conditioned probe set, the pose conditioned probe set, and the facial expression conditioned probe set. In the first experiment, the best number of OLBP codes was 31. The classification error using 31 OLBP codes was 16.00%, which is the smallest than those of other representation methods such as ULBP (18.73%), LBP (18.07%), and MCT (18.07%). We showed that if we use the 31 OLBP codes, we can reduced the computation time by about 1.5 times, 5 times, and 13 times when ULBP, LBP, and MCT were used respectively. In the second experiment, the best number of OLBP codes was 35. The classification error using 35 OLBP codes was 13.00%, which is the smallest than those of other representation methods such as ULBP (19.00%), LBP (20.25%), and MCT (20.25%). In the third experiment, the best number of OLBP codes was 35. The classification error using 35 OLBP codes was 22.60%, which is the smallest than those of other representation methods such as ULBP (30.00%), LBP (29.33%), and MCT (29.33%).

In the facial expression recognition experiments, the best number of OLBP codes was 23. The classification error using 23 OLBP codes was 8.00%, which is the smallest than those of other representation methods such as ULBP (12.73%), LBP (14.20%), and MCT (15.07%). Moreover, we showed that if

we used the 23 OLBP codes, we can reduced the computation time by about 1.5 times, 4 times, and 7.2 times when ULBP, LBP, and MCT were used respectively. From these experimental results, we concluded that the OLBP outperform other features such as LBP, ULBP, and MCT in terms of the number of codes, the classification error, and the computation time.

References

[1] T. Ojala and M. Pietikainen and D. Harwood, A comparative study of texture measures with classification based on feature distributions, *Pattern Recognition*, vol.29, pp.51-59, 1996.

[2] T. Ojala and M. Pietikainen and T. Maenpaa, Multiresolution grayscale and rotation invariant texture classification with Local Binary Pattenrs, *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp.971-977, 2002.

[3] T. Ahonen and A. Hadid and M. Pietikainen, Face Recognition with Local Binary Patterns, *The 8th European Conference on Computer Vision*, pp.469-481, 2004.

[4] O. Lahdenoia and M. Laiho and A. Paasio, Reducing the feature vector length in local binary pattern based face recognition, *IEEE International Conference on Image Processing*, pp.11-14, 2005.

[5] T. Ahonen and A. Hadid and M. Pietikainen, Face description with Local Binary Patterns: application to face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.12, pp.2037-2041, 2006.

[6] G. Hensch and Y. Rodriguez and S. Marcel, Local Binary Patterns as an Image Preprocessing for Face Authentication, *The 7th International Conference on Automatic Face and Gesture Recognition*, pp.9-14, 2006.

[7] Y. Rodriguez and S. Marcel, Face Authentication Using Adapted Local Binary Pattern Histograms, *The 10th European Conference on Computer Vision*, pp.321-332, 2006.

[8] S. Liao and X. Zhu and Z. Lei and L. Zhang and S. Z. Li, Learning Multi-scale Block Local Binary Patterns for Face Recognition, *The 2nd International Conference on Biometrics*, pp.828-837, 2007.

[9] L. Zhang and R. Chu and S. Xiang and S. Liao and S.Z. Li, Face Detection Based on Multi-Block LBP Representation, *The 2nd International Conference on Biometrics*, pp.11-18, 2007.

[10] C. Shan and S. Gong and P. W. McOwan, Facial expression recognition based on Local Binary Patterns: A comprehensive study, *Image and Vision Computing*, vol.27, pp.803-816, 2009.

[11] V. Kellokumpu and G. Zhao and S. Z. Li and M. Pietikainen, Dynamic Texture Based Gait Recognition, *Lecture Notes in Computer Science*, Springer, pp.1000-1009, 2009.

[12] R. Battiti, Using Mutual Information for Selecting Features in Supervised Neural Net Learning, *IEEE Transactions on Neural Networks*, vol.5, no.4, pp.537-550, 1994.

[13] N. Kwak and C. Choi, Improved mutual information feature selector for neural networks in supervised learning, *The 10th International Joint Conference on Neural Networks*, pp.1313-1318, 1999.

[14] J.C. Principe, Feature extraction using information theoretic learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.9, pp.1385-1392, 2002.

[15] K. Torkkola, Feature extraction by non parametric mutual information maximization, *Journal of Machine Learning Research*, vol.3, pp.1415-1438, 2003.

[16] G. Qiu and J. Fang, Car/non-car classification in an informative sample subspace, *The 16th International Conference on Pattern Recognition*, 2006.

[17] R. Zabih and J. Woodfill, A non-parametric approach to visual correspondence, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[18] B. Froba and A. Ernst, Face detection with the modified census transform, *IEEE the 6th International Conference on Face and Gesture Recognition*, pp.91-96, 2004.

[19] Fano, Transmission of information: A statistical theory of communications, *American Journal of Physics*, vol. 29, pp.793-794, 1999.

[20] M. E. Hellman and J. Raviv, Probability of Error, Equivocation, and the Chernoff Bound, *IEEE Transactions on Information Theory*, vol.16, no.4, pp.368-372, 1970.

[21] H. Peng and F. Long and C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, no.8, pp.1226-1238, 2005.

[22] P. A. Estevez and M. Tesmer and C. A. Perez and J. M. Zuraner, Normalized Mutual Information Feature Selection, *IEEE Transactions on Neural Networks*, vol.20, no.2, pp.189-201, 2009.

[23] C. Yun and J. Yang, Experimental comparison of Feature Subset Selection Methods, *IEEE International Conference on Data Mining*, pp.367-372, 2007.

[24] S. Li and Y. Zhu and J. Feng and P. Ai and X. Chen, Comparative Study of Three Feature Selection Methods for Regional Land Cover Classification using MODIS Data, *IEEE Congress on Image and Signal Processing*, pp.565-569, 2008.

[25] H. Lee and S. Park and B. Kang and J. Shin and J. Lee and H. Je and B. Jun and D. Kim, The POSTECH Face Database (PF07) and Performance Evaluation, *The 8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.