J Inf Process Syst, Vol.10, No.3, pp.443~458, September 2014
http://dx.doi.org/10.3745/JIPS.02.0004

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

Extreme Learning Machine Ensemble Using Bagging for Facial Expression Recognition

Deepak Ghimire* and Joonwhoan Lee*

Abstract—An extreme learning machine (ELM) is a recently proposed learning algorithm for a single-layer feed forward neural network. In this paper we studied the ensemble of ELM by using a bagging algorithm for facial expression recognition (FER). Facial expression analysis is widely used in the behavior interpretation of emotions, for cognitive science, and social interactions. This paper presents a method for FER based on the histogram of orientation gradient (HOG) features using an ELM ensemble. First, the HOG features were extracted from the face image by dividing it into a number of small cells. A bagging algorithm was then used to construct many different bags of training data and each of them was trained by using separate ELMs. To recognize the expression of the input face image, HOG features were fed to each trained ELM and the results were combined by using a majority voting scheme. The ELM ensemble using bagging improves the generalized capability of the network significantly. The two available datasets (JAFFE and CK+) of facial expressions were used to evaluate the performance of the proposed classification system. Even the performance of individual ELM was smaller and the ELM ensemble using a bagging algorithm improved the recognition performance significantly.

Keywords—Bagging, Ensemble Learning, Extreme Learning Machine, Facial Expression Recognition, Histogram of Orientation Gradient

1. INTRODUCTION

Automatic facial expression recognition (FER) and analysis has been an active topic in computer science for over two decades. Facial expression conveys non-verbal cues, which play an important role in interpersonal relations. Recent psychology research has shown that facial expressions are the most expressive way humans display emotions. The verbal part of a message contributes only 7% to the effect in the speaker's message as a whole, the vocal part contributes 38%, and facial expressions contribute 55% to the effect of the speaker's message [1]. The automatic recognition of facial expressions can be an important component for natural human-machine interfaces by responding to consumers' expressive conditions. It could also be used in behavioral science and in clinic practice. Although humans detect and interpret faces and facial expressions in a scene with little or no effort, accurate FER by machine is still a challenge. In general, there are three stages for detecting facial expressions, which are as follows: detection of image sequences as a face; extraction of the facial information as features; and classification of the facial features as a particular facial expression. In order to develop a reliable human-machine

※ This work was partially supported by a National Research Foundation of Korea Grant funded by the Korean government (2011-0022152) and BK21PLUS.
Manuscript received October 04, 2013; accepted March 03, 2014.
Corresponding Author: Joonwhoan Lee (chlee@jbnu.ac.kr)
* Department of Computer Engineering, Chonbuk National University, Jeonju, 561-756, Korea (deep@jbnu.ac.kr, chlee@jbnu.ac.kr)

interaction, the system should perform all of these operations accurately and in real time. In general, facial expressions are divided into the seven basic categories of neutral, happy, anger, surprise, fear, disgust, and sadness [2].

We expect there to be small training errors during the training phase in the neural network, but neural network training usually suffers from overtraining, which might degrade the generalization performance of the network. In such cases, the neural network can classify the training data without any errors but it can't guarantee that there will be good classification performance on the validation data set. The performance of the single neural network, as well as the generalization capability of the single neural network can be improved by using an ensemble of neural networks [3]. Two popular ensemble machine-learning methods are bagging [4] and boosting [5,6]. The ensemble method combines the output from several classifiers to improve the accuracy of the overall classification result. The performance of the bagging and boosting techniques are generally more accurate than the individual classifier performance if the base classifiers are unstable [4]. Huang et al. [7] proposed extreme learning machine (ELM), which is a fast learning algorithm for single-hidden-layer feedforward neural networks (SLFNs). An ensemble ELM based on a modified AdaBoost.RT algorithm for predicting the temperature of molten steel in ladle furnace is proposed in [8]. The ensemble of an online sequential ELM (OS-ELM) [9], at which several OS-ELMs are combined OS-ElMs output, improves the generalization capability of the network. Similarly, an ELM ensemble [10-12] has also been proposed. In this study we have selected bagging to generate the ensemble of the ELM for recognizing facial expressions. There are several reasons behind choosing ELM as a base classifier in our FER system. First, the ELM takes random weights between the input and hidden layer. We can then train the same dataset several times, which gives different classification accuracy with different output space. This makes ELM an unstable classifier. Second, the ELM is a much simpler learning algorithm for a feedforward neural network. Unlike the traditional neural networks, it does not need to calibrate the parameters, such as learning rate, learning epochs, etc. Another reason behind selecting ELM as a base classifier is that the learning speed of ELM is extremely fast.

Several researchers have given a lot of attention to the problem of developing methods for FER. The first survey in this field was published in 1992 [13]. Later, Lyons et al. [14] presented the coding of facial expressions with Gabor wavelets. Facial expression images are coded using a multi-orientation, multi-resolution set of Gabor filters and the similarity space derived from this representation is compared with one derived from semantic ratings of the images by human observers. Another detailed survey on the automatic analysis of facial expressions was recently presented by Pantic and Rothkrantz [15]. Support vector machine (SVM) based FER from real time video has been presented by Michel and Kaliouby [16]. An automatic facial feature tracker to perform face localization and feature extraction is employed. The facial feature displacements in the video stream are used as input to support the vector machine classifier. Several machine learning methods, like AdaBoost, SVM, and linear discriminant analysis, with different feature selection schemes along with Gabor wavelets for facial expression classification have been adapted in [17]. The best result was obtained by selecting a subset of Gabor filters using AdaBoost and then training SVM on the outputs of the filters selected by AdaBoost. Another SVM based facial expression classification, which uses geometric deformation features calculated from the images sequence is presented in [18]. Here, the geometric displacement of certain selected candid nodes, which are defined as the difference of the node coordinates

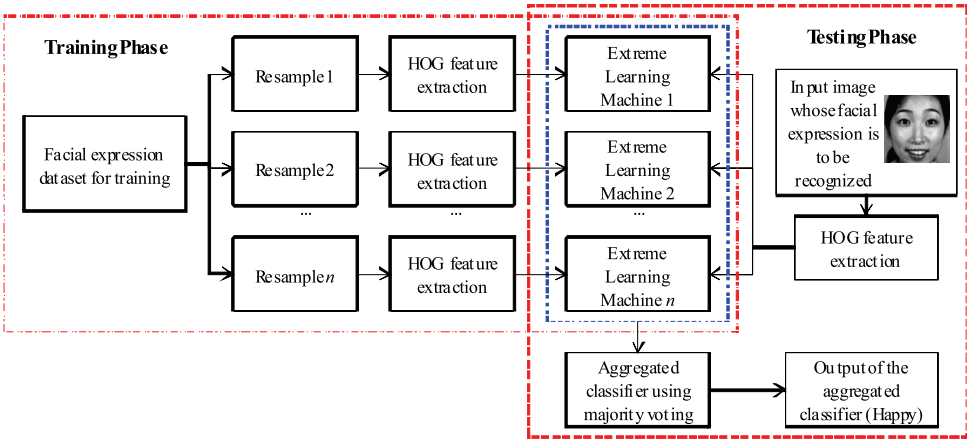


Fig. 1. The overall block diagram of the proposed facial expression recognition system. HOG= histogram of orientation gradient.

between the first and the greatest facial expression intensity frame, is then used as an input for a novel multiclass SVM classifier. Other more recent methods are presented in [19,20], which extract the geometric features by not only considering the single feature point but also a pair of facial feature points at the same time. The feature selection scheme is also presented in [19] using multi-class AdaBoost with a dynamic time warping similarity measure. A brief overview tutorial of FER that highlights the main system components and some research challenges are presented in [21]. Similarly, SVM based emotion classification for an emotion care service system has been developed by Lee et al. [22]. The feature displacements tracked by optical flow are used for input parameters to SVM. A similar method that applies differential flow for automatic FER has been presented in [23]. Recently, robust FER based on local directional patterns (LDPs) has been presented by Jabid et al. [24]. Another more recent method for robust FER, which is based on compressive sensing by employing three kind of facial features, such as the raw pixels, Gabor wavelets representation, and LBP, is presented in [25]. A meta-review of the FER and analysis challenge has recently been presented in [26]. The focus of this review is to clarify how far the field has come, to identify new goals, and to present the results of the baseline methods.

In this paper we present a system for FER using an ELM ensemble (Fig. 1). We have demonstrated that histogram of orientation gradient (HOG) features that have been trained using ELM base classifiers in a bagging algorithm produce better FER results as compared with other methods in the literature of FER. Here we applied the bagging algorithm to generate the number of training bags from the original training dataset. HOG features are calculated for each newly generated dataset and facial expression are trained by using individual ELMs. Finally, the result from individual ELMs is combined by using a majority voting scheme. Instead of using a single ELM for expression classification, we used the ELM ensemble using bagging algorithm. This is because it improves the generalization capability of the whole system, even if the classification accuracy of the individual ELMs is smaller.

The remainder of this paper is organized as follows. Section 2 describes the procedure of extracting features from the face image. Section 3 describes the ELM and bagging algorithm,

which are both used to make the ensemble classifier for FER. Our experimental setup and dataset description is given in Section 4. We report on our experimental results in Section 5 and present the conclusions of this study in Section 6.

2. HOG FEATURE EXTRACTION

The primary step in FER is to localize the face in the input image. Several methods are proposed in the literature for face detection in gray scale images [27], as well as in color images [28,29]. Face detection is also the primary step for face recognition [30,31]. In our system we used the popular face detection method for gray scale images, which was proposed by Viola and Johns [27]. The HOG features are extracted from the face images, which are used as an input to the ELM ensemble for recognizing particular facial expressions. The most important areas in the human faces for classifying expressions are eyes, the eyebrows, and mouth; and the remaining areas do not contribute much in this process. The HOG features are calculated from the whole face region and feed as an input for the FER system.

Different types of features are used by researchers for recognizing facial expressions from face image. These features are mainly divided into two categories: geometric features and appearance features. The HOG features presented by Dalal and Triggs [32], an appearance features provide excellent performance for object classification. The basic hypothesis is that local object appearance and shape can often be characterized relatively well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In addition, the HOG features are invariant to changes in illuminations or shadowing. Gradient calculation is the crucial stage in the HOG descriptor formation. Here we use the Sobel algorithm for gradient calculation. For each pixel $I(x, y)$, the gradient magnitude $G(x, y)$ and orientation $\theta(x, y)$ is computed. Now, the face image is divided into smaller cells and for each cell the one-dimensional orientation histogram of the gradient is formed from the gradient orientation of sample points within a cell. Each histogram divides the gradient angle range into a predefined number of bins (e.g., 9 bins). The gradient magnitudes vote on the orientation histogram. Cells are combined to create a block (e.g., 2×2, 3×3 cells). Each block is locally normalized using the *L1* and *L2* norms and the HOG feature extraction process is completed.

The original method of computing a histogram is not efficient. In this paper we used an integral image [33] to efficiently compute histograms over blocks. With the integral image, we can compute the sum of the elements within a rectangular region by using only 4 image access operations, as shown in Fig. 2.

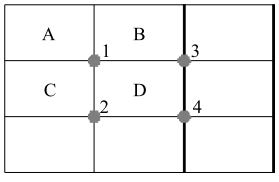


Fig. 2. The value of the integral image at location 1 is the sum of pixels in rectangle A. the value at location 2 is A+C, at location 3 is A+B, and at location 4 is A+B+C+D. The sum within D can be computed as 4+1-(3+2). The overall block diagram of the proposed facial expression recognition system.

Here we discretize each pixel's gradient orientation into 9 bins. Experimentally it has been shown that 9 bins are sufficient in HOG feature extraction. Next, we compute and store the integral image for each histogram bins. The HOG for any of the rectangular regions can then be computed by 9 × 9 = 36 image access operations, 4 image access operations for each of the 9 bins. Therefore, storing an integral image for each bin of a HOG can increase the efficiency in computing the HOG for any rectangular image region.

3. CLASSIFIER FOR FER

In this section we will describe the classifier that we have employed in our system for FER. In the following subsection we will give brief description of ELM and the bagging algorithm as well as propose of using ELM ensemble using bagging for FER.

3.1 Brief of ELM

As we know, traditionally all the parameters of feedforward networks need to be tuned and thus, there is a dependency between the different layers of parameters (weights and biases). The gradient descent-based methods have mainly been used in various learning algorithms of feedforward neural networks. But the disadvantage of these kinds of learning methods is that they are very slow and may easily converge to the local minima. They also require many iterative learning steps in order to obtain better learning performance.

ELM solves the problem of a gradient-based learning algorithm by analytically calculating the optimal weights of the SLFN. First, the weights between the input layer and the hidden layer are randomly selected and then the optimal values for the weights between the hidden layer and output layer are determined by calculating the linear matrix equations.

For N samples and $\tilde{N}$ hidden nodes, the SLFN neural network is defined as:

$$\sum_{i=1}^{\tilde{N}} \beta_i g_i\left(x_j\right)=\sum_{i=1}^{\tilde{N}} \beta_i g_i\left(w_i.x_j+b_i\right)=o_j, j=1,...,N \tag{1}$$

where $\boldsymbol{w_i}$ is the weight vector connecting the i-th hidden node and the input nodes, $\boldsymbol{\beta}_i$ is the weight vector connecting the i-th hidden node and the output nodes, and b_i is the threshold of the i-th hidden node, $g(.)$ is an activation function of the hidden node, and $\boldsymbol{o}_j$ is the output vector of the j-th input data. $\boldsymbol{w_i.x_j}$ denotes the inner product of $\boldsymbol{w_i}$ and $\boldsymbol{x_j}$.

The standard SLFNs with $\tilde{N}$ hidden nodes with the activation function $g(x)$ can approximate these N samples with zero error. This means that $\sum_{j=1}^{\tilde{N}}\left\|o_j-t_j\right\|=0$. For example, there exists $\boldsymbol{\beta_i}$, $\boldsymbol{w_i}$ and b_i such that:

$$\sum_{i=1}^{L} \beta_i g(w_i.x_j+b_i)=t_j, \quad j=1,..., N \tag{2}$$

where $\boldsymbol{t_j}$ is the target vector of the j-th input data.

Eq. (2) can be reformulated as a matrix equation to form Eq. (3) by using the output matrix of the hidden layer $\boldsymbol{H}$.

$$H\beta = T \tag{3}$$

where,

$$H = \begin{bmatrix} g(w_1.x_1 + b_1) & \cdots & g(w_L.x_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(w_1.x_N + b_1) & \cdots & g(w_L.x_N + b_L) \end{bmatrix}_{N \times \tilde{N}}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{\tilde{N} \times m} \quad \text{and} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{\tilde{N} \times m}$$

In Eq. (3), the target vector $\boldsymbol{T}$ and the output matrix of the hidden layer $\boldsymbol{H}$ can comprise a linear system. Thus, the learning procedure of the network becomes finding the optimal weight matrix $\boldsymbol{\beta}$ between the output layer and the hidden layer. Here $\boldsymbol{\beta}$ can be calculated by using the Moor-Penrose Generalized Inverse of $\boldsymbol{H}$, i.e.,

$$\boldsymbol{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{T} \tag{4}$$

With this process two effects can be established. The first is that we can take the minimum error condition, because the obtained $\boldsymbol{\beta}$ is a least-square solution. In addition, the optimal $\boldsymbol{\beta}$ is also the minimum norm among these solutions. Thus, ELM has a better generalization performance than the typical back-propagation network. In summary, the ELM algorithm can be written as follows:

ELM Algorithm: Given a training set $\{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1,...,N\}$, hidden node output function $g(w,b,x)$, and number of hidden nodes L,

1) Assign randomly hidden node parameters $(w_i, b_i), i = 1,...,L$.
2) Calculate the hidden layer output matrix $\boldsymbol{H}$.
3) Calculate the output weights $\boldsymbol{\beta}$: $\boldsymbol{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{T}$.

where $\boldsymbol{H}^{\dagger}$ is the Moore-Penrose generalized inverse of hidden layer output matrix $\boldsymbol{H}$.

3.2 The Bagging Algorithm

Bagging is a method for improving the results of machine learning classification algorithms. This method was formulated by Breiman [4] and its name was deduced from the phrase 'bootstrap aggregating.' The algorithm is shown in Fig. 3. The bagging algorithm works with training samples of a fixed size. The samples are resampled according to a given probability distribution. The algorithm constructs many different bags of samples. Each bag is a set of training samples and is collected by randomly and uniformly resampling the original training set. The algorithm then applies a base classifier to classify each bag. Finally, the decision is made by a majority vote of all the base classifier results. In our system, the base classifier is ELM.

Input: Training set D, base classifier L, integer M (number of bootstrap sample)
for $i = 1$ to M
{
D_i = bootstrap sample from D (i.i.d. sample with replacement)
$C_i = L(D_i)$
}
$C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} 1$ (the most often predicted label y)

Output: Compound classifier C^*

Fig. 3. The bagging algorithm.

3.3 ELM Ensemble for FER

It has been shown that if the base classifier is unstable in the bagging algorithm, then bagging may improve the classification accuracy significantly. The main reason here for selecting ELM as a base classifier is that the ELM is a highly unstable classifier. In ELM, even when we train the same training sample multiple times, it gives different performances. This is because each time there are different random weights between the input to hidden nodes and different random bias values for each hidden node. According to those random weights and bias values, the weight between the hidden and the output nodes is determined. Therefore, even we train the same training data, there will be a different output space ELM network each time. Again, if we use ELM as a base classifier in the bagging algorithm we have different bags of training data generated with randomly and uniformly resampling from the original training data set. Consequently, each base classifier is highly unstable and has different output spaces from each other. The main advantage of ELM with bagging is that even though the individual base classifier has a low performance, the performance of the compound classifier improves significantly. The size of the available dataset for facial expression is not big enough. Therefore, we also encourage the use of bagging algorithms. This is because, in general, if we have a small size dataset, bagging is an efficient algorithm for making a classifier.

4. Experimental Setup and Dataset Description

The performance of the proposed system is evaluated on two well-known facial expression datasets, which are the Japanese Female Facial Expression (JAFFE) dataset [14] and the Extended Cohn-Kanade (CK+) facial expression dataset [34].

The JAFFE dataset contains 213 grayscale images, which are each 256 × 256 pixels in size. This dataset includes facial expressions from 10 different female models, each assuming 7 distinct poses (i.e., 6 basic expressions and one neutral pose). For each expression there is an average of 3 different images from 1 model. That means there are 21 images per model and around 30 images per expression. At first, we divided the dataset into training and validation sets. For each expression type, 20 images were selected randomly for training and the remaining images are left for validation. Which resulted in the creation of 140 images in total for training and 73 images for validation.

The CK+ database contains 593 sequences from 123 subjects. The image sequence varies in duration (i.e., 10–60 frames) and incorporates the onset, which is also a neutral face, to the peak formation of the facial expressions. Image sequences from neutral to target displays were digitized into 640 × 480 or 640 × 490 pixel arrays. Only 327 of the 593 sequences were given the emotion label. From each image sequence we selected 2 or 3 of the middle and most expressive images, resulting in the creation of 1,037 expression images. For the case of neutral expressions, we selected the first image in the expression sequence. For each expression, one-fourth of the images were selected randomly for validation and rests of the image are used for training. These resulted in the creation of 781 images for training and 256 images for validation. All of the results presented in this paper for both the JAFFE and CK+ dataset belong to the validation set.

There is some background region in each image that appears in both datasets. Therefore, we only cropped the face area using the face detection method proposed in [27]. The upper row in Fig. 4 shows the sample images from the JAFFE dataset and the lower row in Fig. 4 shows the sample images from the CK+ database.

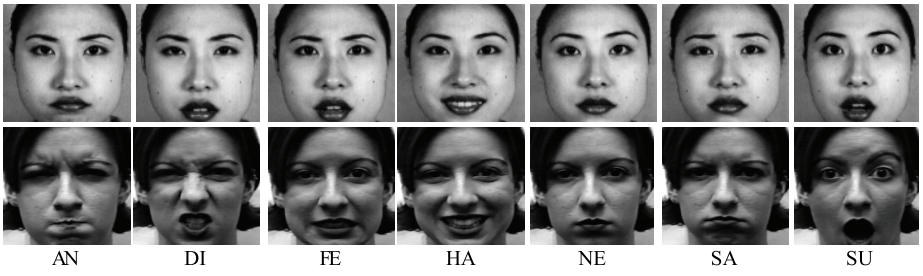


Fig. 4. Sample facial expression images of each prototypic expression from the JAFFE dataset (upper row) and from the CK+ dataset (lower row).

5. Results and Discussion

In this section we will explain several of the experimental results from both of the datasets what we used. First, we tried to investigate the suitable size for input HOG feature dimensions. HOG features are detected by dividing the face image into different number of cells. In our experiment, we took 2 × 2 cells to create a block and locally normalized it by using the L2-norm. The blocks were not overlapped, and we also tried taking overlapping blocks to find HOG features. However, there was not much difference in the performance of the classification system. Fig. 5 shows the performance of the ensemble classification system on the dataset with a different number of base ELM classifiers. We determined the HOG features by dividing the image into 8 × 8, 6 × 6, and 4 × 4 cells, with 9 bins per cell, which resulted in 576, 324, and 144 dimensional HOG features. The image size taken for this experiment was 100 × 100 pixels. From Fig. 5 we can see that classification performance using 324-D HOG features is better on both datasets.

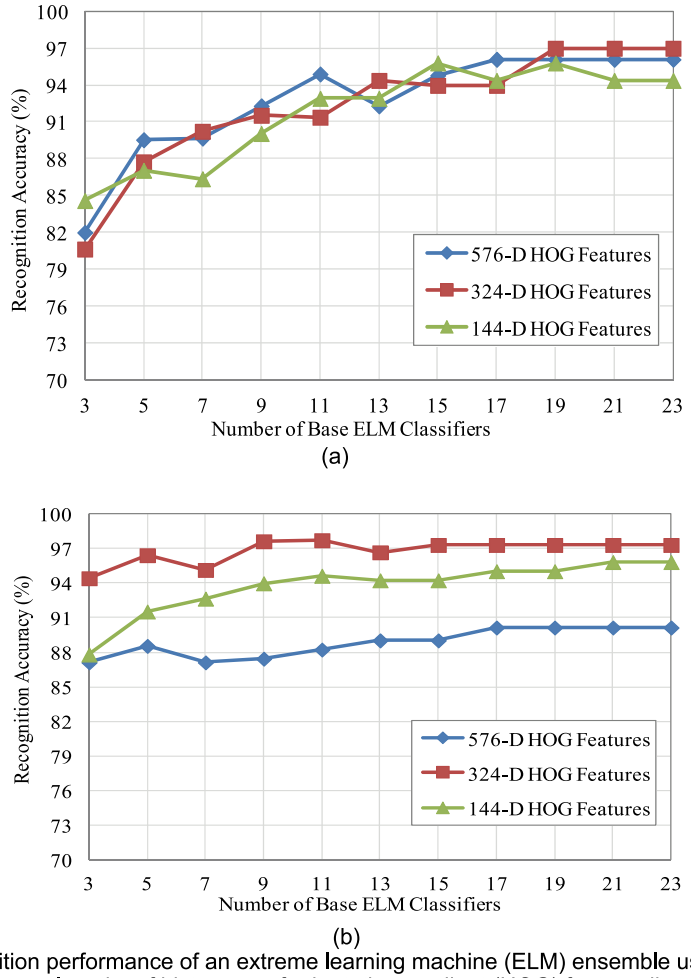


Fig. 5. Recognition performance of an extreme learning machine (ELM) ensemble using bagging with different lengths of histogram of orientation gradient (HOG) feature dimensions and a different number of base ELM classifiers in a JAFFE dataset (a) and a CK+ dataset (b).

Table 1. The average recognition accuracy (%) of 23 base ELM classifiers with different image resolutions

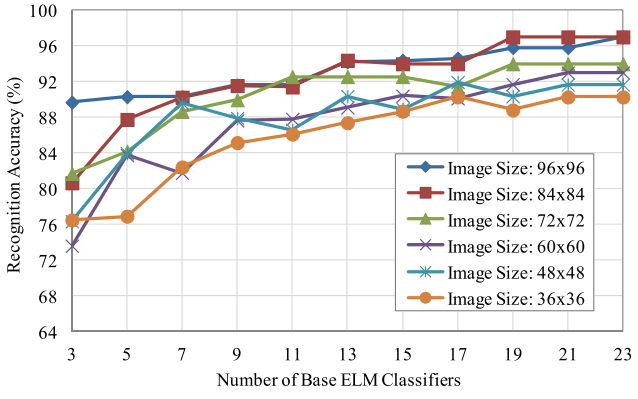
Database	Image size					
	96 × 96	84 × 84	72 × 72	60 × 60	48 × 48	36 × 36
JAFFE	74.51	70.46	70.29	65.45	63.14	58.02
CK+	87.90	87.72	82.67	81.22	79.50	75.54

ELM=extreme learning machine, JAFFE=Japanese Female Facial Expression, CK+=Extended Cohn-Kanade

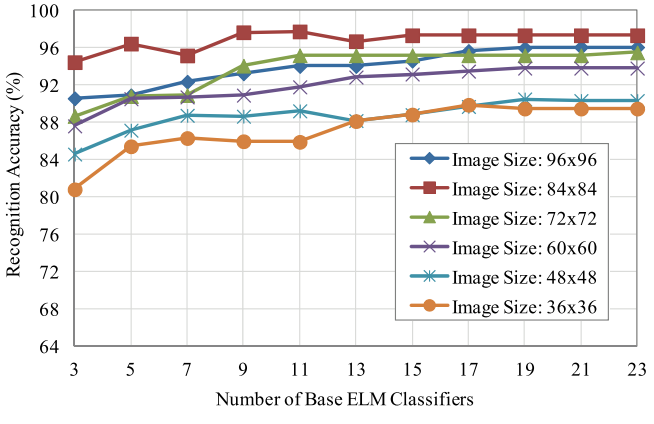
For the entire experiment we used the 324-D input feature by dividing the face image into 6x6 cells. The second experiment we performed was on expression classification with different resolutions of input face images. The resolution of the input image is very important. If we want to develop a real time facial expression classification system, then in general, the resolution of the detected input face image can vary from low resolution to high resolution. In our experiment

we found that by reducing the image resolution, the performance of the individual base ELMs also decreased. However, the bagging result was not significantly decreased, even if the classification performance of the individual base ELMs decreased. Table 1 shows the average classification accuracy of 23 base ELMs with a different resolution in both datasets. Regardless of the resolution of the image, we divided the face image into 6 × 6 cells and the 324-D HOG features were extracted.

From Table 1 it can be seen that the average classification accuracy was decreased from 74.51% to 58.02% by reducing the image resolution from 96 × 96 to 36 × 36 in the case of the JAFFE dataset and that the classification accuracy was reduced from 87.90% to 75.54% in the case of the CK+ dataset. Classifying facial expressions in the CK+ dataset is relatively easier than classifying expressions in the JAFFE dataset. There is big difference in the classification accuracy of a single ELM, as we reduced the image resolution. Fig. 6 shows the result of an ELM ensemble using bagging with different input face image resolutions. For low-resolution images, even though the base classifier accuracy was smaller, the bagging performance was



(a)



(b)

Fig. 6. Recognition performance of the extreme learning machine (ELM) ensemble using bagging for different image resolutions with a different number of base ELM classifiers in the JAFFE dataset (a) and the CK+ dataset (b).

better. With 23 base ELMs in bagging algorithms, the lowest classification accuracy was around 90% and the highest classification accuracy was around 97%, in the case of both the JAFFE and CK+ facial expression datasets. The classification results are relatively more stable in the case of the CK+ dataset.

As we know, the upper facial features play a more important role in face interpretation, as opposed to the lower facial features. The most important areas on human faces for classifying expressions are the eyes, eyebrows, and mouth. From among them, the eyes and eyebrows are the most important features.

Next, we will demonstrate the effect of removing features from a certain region of the face image in our system. By removing features from the eye and eyebrow region, the performance of the classification was decreased to 84% from 97% in the JAFFE dataset and it was decreased to 87% from 98% in the CK+ dataset. Therefore, if the eye and eyebrow region is covered, it becomes difficult to recognize the expression of the face image with high accuracy. We also removed the mouth and nose region features in order to see how the classification system performs. In this case, the performance did not decrease a lot. We still achieved a performance rate of around 94% for the ELM ensemble classifier in both datasets. Therefore, the performance of the proposed system is not lower even if we decrease the image resolution and remove the features from the regions of the eyes or mouth. Sometimes, there is a possibility of covering the mouth region by using a mask. Even in this case, the performance of the proposed system is around 94%. Fig. 7 shows the classification accuracy of the proposed system by removing features from the areas of the eyes and eyebrows or mouth and nose.

So far, we have discussed the average recognition accuracy for 7 distinct facial expression classes in different cases. To get a better picture of the recognition accuracy of individual expression types, the confusion matrices are given in table 2, 3, 4 and 5. The expressions are denoted by using the following notations: angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA), and surprise (SU). The confusion matrix is $n \times n$ matrix, at which each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The diagonal entries of the confusion matrix are the rates of facial expressions that are correctly classified, while the off-diagonal entries correspond to misclassification rates. In the several research [18, 35, 36], some authors also considered 6-class expressions by excluding neutral cases. However, in our experiment, we considered 7-class expressions by including a neutral case too, because in a real time system there are situations in which we also need to discriminate neutral cases from other expressions. Tables 2 and 3 show the confusion matrix of the proposed system by using 5 and 13 base ELMs in the JAFFE dataset, respectively. Similarly, Tables 4 and 5 show the confusion matrix by using the 5 and 15 base ELMs in the CK+ dataset, respectively. These tables show the individual recognition accuracy for each expression type. For the CK+ dataset the angry, fear, happy, and surprise expressions were detected with 100% accuracy, whereas, in the case of the JAFFE dataset the happy, neutral and sad expressions were detected with 100% accuracy. We used the ensemble scheme. Therefore, by varying the number of base ELMs, the performance on the confusion matrix also varies.

Table 2. Confusion matrix of the FER using bagging with 5 base ELM classifiers in the JAFFE dataset

%	AN	DI	FE	HA	NE	SA	SU
AN	**80**	10	0	0	10	0	0
DI	0	**88.89**	0	0	11.11	0	0
FE	0	0	**91.67**	0	0	8.33	0
HA	0	0	0	**90.91**	0	0	9.09
NE	0	0	0	0	**100**	0	0
SA	18.18	9.09	0	0	0	**72.73**	0
SU	0	0	0	0	10	0	**90**

FER=facial expression recognition, ELM=extreme learning machine, JAFFE=Japanese Female Facial Expression, AN=angry, DI=disgust, FE=fear, HA=happy, NE=neutral, SA=sad, SU=surprise.

Table 3. Confusion matrix of the FER using bagging with 13 base ELM classifiers in the JAFFE dataset

%	AN	DI	FE	HA	NE	SA	SU
AN	**90**	10	0	0	0	0	0
DI	0	**88.89**	0	0	11.11	0	0
FE	0	8.33	**91.67**	0	0	0	0
HA	0	0	0	**100**	0	0	0
NE	0	0	0	0	**100**	0	0
SA	0	0	0	0	0	**100**	0
SU	0	10	0	0	0	0	**90**

FER=facial expression recognition, ELM=extreme learning machine, JAFFE=Japanese Female Facial Expression, AN=angry, DI=disgust, FE=fear, HA=happy, NE=neutral, SA=sad, SU=surprise.

Table 4. Confusion matrix of the FER using bagging with 5 base ELM classifiers in the CK+ dataset

%	AN	DI	FE	HA	NE	SA	SU
AN	**96**	0	0	0	4	0	0
DI	0	**93.33**	0	0	3.33	3.33	0
FE	0	0	**100**	0	0	0	0
HA	0	0	0	**100**	0	0	0
NE	2.22	2.22	0	2.22	**91.11**	2.22	0
SA	5.56	0	0	0	0	**94.44**	0
SU	0	0	0	0	10	0	**100**

FER=facial expression recognition, ELM=extreme learning machine, JAFFE=Japanese Female Facial Expression, AN=angry, DI=disgust, FE=fear, HA=happy, NE=neutral, SA=sad, SU=surprise.

Table 5. Confusion matrix of the FER using bagging with 15 base ELM classifiers in the CK+ dataset

%	AN	DI	FE	HA	NE	SA	SU
AN	**100**	0	0	0	0	0	0
DI	0	**93.33**	0	0	3.33	3.33	0
FE	0	0	**100**	0	0	0	0
HA	0	0	0	**100**	0	0	0
NE	2.22	0	0	0	**93.33**	4.44	0
SA	0	0	0	0	5.56	**94.44**	0
SU	0	0	0	0	0	0	**100**

FER=facial expression recognition, ELM=extreme learning machine, JAFFE=Japanese Female Facial Expression, AN=angry, DI=disgust, FE=fear, HA=happy, NE=neutral, SA=sad, SU=surprise.

Table 6. Performance comparison of the proposed FER system with state-of-the-art methods

Reference	Method	No. of class	Classification accuracy (%)	
			JAFFE	CK+
Jabid et al. [24]	Local directional pattern features and SVM	7	86.78	93.61
Kotisa and Pitas [18]	Geometric deformation features and SVM	6	-	99.7
Zafeiriou and Pitas [35]	Discriminant expression-specific graphs	6	-	97.1
Lyons et al. [36]	PCA and LDA of labeled graph vectors	6	75-92	-
Thai et al. [37]	Canny, PCA and artificial neural network	7	87.7	-
Zhang et al. [25]	Sparse representation of sparse representation classifiers with raw pixels, Gabor Wavelet, and LBP	7	-	97.14
Proposed system	HOG features and ELM ensemble using bagging	7	94.37	97.3

FER=facial expression recognition, JAFFE=Japanese Female Facial Expression, CK+=Extended Cohn-Kanade, SVM=support vector machine, HOG= histogram of orientation gradient, ELM=extreme learning machine.

Even though the experimental setups are not the same, we also listed the overall recognition accuracy (%) of some methods from the literature [24, 18, 35, 36, 37, 25] and compared the FER results with our proposed system, which is shown in Table 6. Recently, authors in [24], by using LDP descriptors and SVM, achieved 86.78% and 93.61% classification accuracy in the JAFFE and CK+ datasets, respectively. The system in [18] has shown superior performance and has achieved a 99.7% recognition rate with 6-class expressions in the CK+ dataset. The drawback of the system in [18] is that it requires the Candide grid to be manually placed upon the facial area and moreover, it requires the manual detection of the neutral state in a video sequence. Another method proposed by Zafeiriou and Pitas [35] achieved a 97.1% recognition rate in the CK+ dataset. Elastic graph matching is used for facial feature point localization, at which the most discriminant facial landmark are selected for every facial expression. Lyons et al. [36] achieved up to 92% of their classification results by using PCA and LDA of labeled-graph vectors in the JAFFE dataset. Similarly Thai et al. [37] achieved an 85.7% recognition rate with 7-class expressions in the JAFFE dataset. Zhang et al. [25] achieved 97% recognition accuracy in a CK+ dataset by using a sparse representation classifier with LBP features. The recognition rates

of graph matching based methods are relatively better in comparison with other methods. But, in general, these methods need feature tracking in a sequence of images and it is difficult to determine facial expressions on a single image frame. However, our proposed method is very simple and can be easily implemented, but the complexity will increase with the increase in the number of base ELMs. We achieved 94.37% recognition results in the JAFFE dataset with only 13 base ELM classifiers. This is the best recognition accuracy to date, as compared with the methods in the literature of FER. We achieved a 97.30% recognition result in the CK+ dataset with 15 base ELM classifiers, which is also among the best results in the literature of FER.

6. CONCLUSION

This paper proposed a method for FER based on HOG features and an ELM ensemble using bagging. The ELM is an unstable classifier. Therefore, an ELM ensemble using bagging improved the classification accuracy significantly in comparison with the performance of the single ELM. HOG features calculated without overlapping the blocks are used as input for the classification system. We experimented with the proposed system on varying image resolutions, feature dimensions, and also removed features from the key regions of the face image. In each case, even if the recognition accuracy of the individual ELMs was smaller, the bagging result was considerably better. The recognition accuracy of the proposed method is better in the JAFFE dataset and is comparable with the best results in the literature of FER in the case of the CK+ dataset. Therefore, facial expression classification using HOG features and an ELM ensemble can achieve better results.

Further research will focus on specially combining the texture feature with geometric features, such as facial muscle movement, for recognizing facial expressions in pose invariant case.

REFERENCES

[1] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, no. 4, pp. 53-56, 1968.

[2] P. Ekman, "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique," *Psychological Bulletin*, vol. 115, no. 2, pp. 268-287, Mar. 1994.

[3] L. K. Hansen and P. Salamon, "Neural network ensemble," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, no. 10, pp. 993-1001, Oct. 1990.

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

[5] R. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197-227, Jun. 1990.

[6] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and Computation*, vol. 121, no. 2, pp. 256-285, Sep. 1995.

[7] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489-501, Dec. 2006.

[8] H. X. Tian and Z. Z. Mao, "An ensemble ELM based on modified AdaBoost.RT algorithm for predicting the temperature of molten steel in ladle furnace," *IEEE Transactions on Automation Science and Engineering*, vol. 7, no. 1, pp. 73-80, Jan. 2010.

[9] Y. Lan, Y. C. Soh, and G. B. Huang, "Ensemble of online sequential extreme learning machine," *Neurocomputing*, vol. 72, no. 13-15, pp. 3391-3395, Aug. 2009.

[10] Y. Liu, X. Xu, and C. Wang, "Simple ensemble of extreme learning machine," in *Proceedings of the 2nd International Congress on Image and Signal Processing*, Tianjin, China, October 17-19, 2009, pp. 1-5.

[11] N. Liu and H. Wang, "Ensemble based extreme learning machine," *IEEE Signal Processing Letters*, vol. 17, no. 8, pp. 754-757, Aug. 2010.

[12] M. van Heeswijk, Y. Miche, T. Lindh-Knuutila, P. J. Hilbers, T. Honkela, E. Oja, and A. Lendasse, "Adaptive ensemble models of extreme learning machines for time series prediction," in *Artificial Neural Networks–ICANN 2009, Lecture Notes in Computer Science Volume 5769*, C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, Eds., Heidelberg: Springer Berlin, 2009, pp. 305-314.

[13] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65-77, Jan. 1992.

[14] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba "Coding facial expressions with Gabor wavelets," in *Proceedings of the 3rd IEEE International Conference on Face and Gesture Recognition*, Nara, Japan, April, 14-16, 1998, pp. 200-205.

[15] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1434-1445, Dec. 2000.

[16] P. Michel and R. E. Kaliouby, "Real time facial expression recognition in video using support vector machines," *in Proceedings of the 5th International Conference on Multimodal Interfaces*, Vancouver, Canada, November 5-7, 2003, pp. 258-264.

[17] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from videos," *Image and Vision Computing*, vol. 24, no. 6, pp. 615-625, Jun. 2006.

[18] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172-187, Jan. 2007.

[19] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714-7734, Jun. 2013.

[20] D. Ghimire and J. Lee, "Automatic facial expression recognition based on features extracted from tracking of facial landmarks," *Proceedings of SPIE*, vol. 9069, pp. 90691O, Jan. 2014.

[21] C. C. Chibelushi and F. Bourel, "Facial expression recognition: a brief tutorial overview," [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/CHIBELUSHI1/CCC_FB_FacExprRecCVonline.pdf.

[22] B. Lee, J. Chun, and P. Park, "Classification of facial expressions using SVM for emotion care service system," in *Proceedings of the 9th ASIS International Conference on Software, Engineering, Artificial Intelligence*, Phuket, Thiland, August 6-8, 2008, pp. 8-12.

[23] A. Sánchez, J. V. Ruiz, A. B. Moreno, A. S. Montemayor, J. Hernández, and J. J. Pantrigo,, "Differential optical flow applied to automatic facial expression recognition," *Neurocomputing*, vol. 74, no. 8, pp. 1272-1282, Mar. 2011.

[24] T. Jabid, H. Kabir and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI Journal*, vol. 32, no. 5, pp. 784-794, Oct. 2010.

[25] S. Zhang, X. Zhao, and B. Lei, "Robust facial expression recognition via compressive sensing," *Sensors*, vol. 12, no. 3, pp. 3747-3761, Mar. 2012.

[26] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, K. Scherer, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 4, pp. 966-979, Aug. 2012.

[27] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, May 2004.

[28] D. Ghimire and J. Lee, "A robust face detection method based on skin color and edges," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 141-156, Mar. 2013.

[29] C. F. Juang and S. J. Shiu, "Using self-organizing fuzzy network with support vectors learning for face detection in color images," *Neurocomputing*, vol. 71, no. 16-18, pp. 3409-3420, Oct. 2008.

[30] S. J. Wang, C. G. Zhou, N. Zhang, X. J. Peng, Y. H. Chen, and X. Liu, "Face recognition using

second-order discriminant tensor subspace analysis," *Neurocomputing*, vol. 74, no. 12-13, pp. 2142-2156, Jun. 2011.

[31] W. Zong and G. B. Huang, "Face recognition based on extreme learning machine," *Neurocomputing*, vol. 74, no. 16, pp. 2541-2551, Sep. 2011.

[32] N. Dalal and B. Triggs, "Histogram of orientation gradients for human detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, June 25, 2005, pp. 886-893.

[33] P. Viola and M. Jones, "Rapid object detection using boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Kauai, HI, December 8-14, 2001, pp. 511-518.

[34] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specific expressions," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, CA, June 13-18, 2010, pp. 94-101.

[35] S. Zafeiriou and I. Pitas, "Discriminant graph structures for facial expression recognition," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1528-1540, Dec. 2008.

[36] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, Dec. 1999.

[37] L. H. Thai, N. D. T. Nguyen, and T. S. Hai, "A facial expression classification system integrating canny, principal component analysis and artificial neural network," *International Journal of Machine Learning and Computing*, vol. 1, no. 4, pp. 388-393, Oct. 2011.

Deepak Ghimire

He received his B.E. degree in Computer Engineering from Pokhara University, Nepal in 2007 and M.S. degree in Computer Science and Engineering from Chonbuk National University, Republic of Korea in 2011. Currently he is pursuing his Ph.D. degree in Computer Science and Engineering at Chonbuk National University, Republic of Korea since 2011. His main research interests include image processing, computer vision, pattern classification, facial emotion analysis etc.

Joonwhoan Lee

He received his B.S. degree in Electronic Engineering from the University of Hanyang, Republic of Korea in 1980. He received his M.S. degree in Electrical and Electronics Engineering from KAIST University, Republic of Korea in 1982 and Ph.D. degree in Electrical and Computer Engineering from University of Missouri, USA in 1990. He is currently a Professor in Department of Computer Engineering, Chonbuk National University, Republic of Korea. His research interests include image processing, computer vision, emotion engineering etc.