IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 22, 2019, accepted March 20, 2019, date of publication March 25, 2019, date of current version April 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907327

Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure

JI-HAE KIM[1], BYUNG-GYU KIM[1], (Senior Member, IEEE), PARTHA PRATIM ROY[2], (Member, IEEE), AND DA-MI JEONG[1]

[1]Department of IT Engineering, Sookmyung Women's University, Seoul, South Korea
[2]School of Computer Science and Engineering, IIT Roorkee, Roorkee, India

Corresponding author: Byung-Gyu Kim (bg.kim@sookmyung.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2016R1D1A1B04934750.

ABSTRACT With the continued development of artificial intelligence (AI) technology, research on interaction technology has become more popular. Facial expression recognition (FER) is an important type of visual information that can be used to understand a human's emotional situation. In particular, the importance of AI systems has recently increased due to advancements in research on AI systems applied to AI robots. In this paper, we propose a new scheme for FER system based on hierarchical deep learning. The feature extracted from the appearance feature-based network is fused with the geometric feature in a hierarchical structure. The appearance feature-based network extracts holistic features of the face using the preprocessed LBP image, whereas the geometric feature-based network learns the coordinate change of action units (AUs) landmark, which is a muscle that moves mainly when making facial expressions. The proposed method combines the result of the softmax function of two features by considering the error associated with the second highest emotion (Top-2) prediction result. In addition, we propose a technique to generate facial images with neutral emotion using the autoencoder technique. By this technique, we can extract the dynamic facial features between the neutral and emotional images without sequence data. We compare the proposed algorithm with the other recent algorithms for CK+ and JAFFE dataset, which are typically considered to be verified datasets in the facial expression recognition. The ten-fold cross validation results show 96.46% of accuracy in the CK+ dataset and 91.27% of accuracy in the JAFFE dataset. When comparing with other methods, the result of the proposed hierarchical deep network structure shows up to about 3% of the accuracy improvement and 1.3% of average improvement in CK+ dataset, respectively. In JAFFE datasets, up to about 7% of the accuracy is enhanced, and the average improvement is verified by about 1.5%.

INDEX TERMS Artificial intelligence (AI), facial expression recognition (FER), emotion recognition, deep learning, LBP feature, geometric feature, convolutional neural network (CNN).

I. INTRODUCTION

Technologies for communication have traditionally been developed based on the senses that play a major role in human interaction [1]. In particular, artificial intelligence voice recognition technology using the sense of hearing and AI speakers has been commercialized because of improvements in artificial intelligence (AI) technology [2]. Through the use of such technologies that recognize voice and language, there are artificial intelligence robots that can interact closely with real life, in such ways as managing the daily schedules of people and playing their favorite music. However, sensory acceptance is required for interactions more precisely mirroring those of humans. Therefore, the most necessary technology is a vision sensor, as vision is a large part of human perception in most interactions. In artificial intelligence robots using interactions between a human and a machine, human faces provide important information as a

The associate editor coordinating the review of this manuscript and approving it for publication was Peter Peer.

clue to understand the current state of the user. Therefore, the field of facial expression recognition has been studied extensively over the last ten years.

Recently, with the increment of relevant data and continued development of deep learning, a facial expression recognition system which accurately recognizes facial expressions in various environments has come to be actively studied. Facial expression recognition systems FERs) are fundamentally based on an ergonomic and evolutionary approach. Based on universality, similarity, physiological, and evolutionary properties, emotions in FER studies can be classified into six categories: happiness, sadness, fear, disgust, surprise, and anger. In addition, emotions can be classified into seven categories with the addition of a neutral emotion [1], [3].

A FER system for recognizing facial expressions requires four steps. First, we need a face detection step that localizes the human face. Representative algorithms include Adaboost [4], Haarcascade [5], and a histogram of oriented gradients (HOG) [6]. The second step involves a face registration with which to obtain the main feature points in order to recognize face rotations or muscle movement. The faces after detection step is inclined to be degraded in terms of recognition accuracy due to the potential for various illuminations and rotations. Therefore, it is necessary to improve the image by obtaining landmarks, which are the positions of the main muscle movements when one is making facial expression. The positions that define the contraction of the facial muscles are called action units (AUs), and the main positions include the eyebrows, eyes, nose, and mouth [7]. A typical algorithm is active appearance models (AAM) [8]. Third, features that can recognize facial expressions are extracted by acquiring the motion or position information of the feature points in the feature extraction step.

To this end, the approach can be divided into appearance feature-based and geometric feature-based methods. The appearance feature-based method is a feature extraction method for the entire facial image. It involves the method of dimension reduction through a fusion with binary feature extraction, which is widely applied in the field of facial studies. Principal component analysis (PCA) and linear discriminant analysis (LDA) are typical dimension reduction methods. The local binary pattern (LBP) and local directional pattern (LDP) techniques are binary feature extraction methods [9], [10] for presenting facial expression. The geometric feature-based method extracts the geometric position of the face or the value of the change in facial muscle movement. Finally, based on the obtained features, a classification step is needed to classify the defined emotions using a support vector machine (SVM) and the hidden markov model (HMM) [11], [12].

Despite the fact that many algorithms have been studied, some problems still remain, such as illumination changes, rotations, occlusions, and accessories [3]. These are not only classical problems involved in image processing, but also factors that cause hardship for capturing action units of facial recognition. Aside from environmental changes, there is a problem with the lack of appropriate datasets.

In this paper, we propose an efficient algorithm to improve the recognition accuracy by a hierarchial deep neural network structure which can re-classify the result (Top-2 error emotion). The feature extracted from the appearance feature-based network is fused with the geometric feature in a hierarchical structure. The proposed scheme combines two features to obtain more accurate result by considering the error associated with the second highest emotion (Top-2) prediction result.

The rest of this paper is organized as follows: In Section II, we discuss various existing algorithms for FER. Section III presents a new proposed FER algorithm using deep learning based on appearance feature and geometric feature. The experimental results are reported in Section IV. Finally, the concluding remarks are presented in Section V.

II. RELATED WORKS

We describe related works of facial expression recognition systems that have been studied to date. These algorithms can largely be divided into the classical feature extraction method and the deep learning-based method. The classical feature extraction methods can be roughly classified further into the appearance feature extraction method that extracts the features of the entire facial criterion and the geometric feature extraction method that extracts geometric elements of the facial structure and motion of the facial muscles. In the following sub-sections, we will describe some of the recent algorithms involving the appearance feature extraction method, the geometric feature extraction method, and deep learning-based facial expression recognition algorithms.

A. CLASSICAL FEATURE FER APPROACHES

Liu *et al.* [13] and Happy and Routray [14] reported a representative FER algorithm using LBP. In Liu *et al.* [13], the active patches were defined around 68 landmarks extracted by the active appearance model (AAM), and the features were extracted using LBP for the patches. In this way, they eliminated unnecessary parts of the face and reduced the effect on the environmental change. This improved the accuracy by using the more robust features obtained from the main facial muscles with patch centering. In Happy and Routray [14], rather than using other existing algorithms that extract facial landmarks, they detected the points of the eyebrows, eyes, nose, and mouth corners by applying the sobel edge, otsu algorithm, morphological operation, etc. By defining the active facial patches and extracting the LBP histogram feature of each patch, the features of the 19 patches have been extracted from the main facial muscles such as the forehead, nose, and mouth, which move when facial expressions are generated. The local directional ternary pattern (LDTP) feature extraction method based on the two largest directions after performing a matrix operation on the Robinson compass was used by Ryu *et al.* [15]. This algorithm formed a 17 $\times$ 17 block around 42 landmarks selected by the active pattern

through LDTP, and extracted an LDTP histogram from it. Through this process, robust features were extracted, and more accurate emotional recognition was achieved by extracting more information from the strong response regions.

When emotions are expressed, the formation of patches around the mainly changing facial muscles, the application of this information to the appearance-based feature extraction technique, have recently expanded to contribute to the development of various algorithms applying deep learning.

A typical algorithm that uses geometric features extracts the temporal or dynamic changes of the landmark of the face.

Kotsia and Pitas [16] used the candide wireframe model to predict facial emotions by extracting geometric features around face landmarks. In this algorithm, the grid was traced in the sequential dataset. The geometric and dynamic information regarding the emotion changes have been extracted by the difference between the first neutral face grid and the peak emotion grid of the last frame. Finally, when classifying the final emotion, the emotion was classified using support vector machine (SVM) by combining the values with the facial action units (FAUs), including the facial change according to the emotions.

A dynamic texture-based approach to classifying emotions using a free-form deformation (FFD) technique for tracking the motion direction of AUs in image sequences was proposed by Koelstra *et al.* [17]. The extracted representation based on motion history was used to derive the motion direction histogram descriptor in the spatial and temporal domains. The extracted features finally combined gentleboolst algorithm and HMM in order to classify the emotions.

These geometric feature extraction methods can reduce the effect of the degradation of accuracy due to illumination or external change by tracking the movement of geometric coordinates extracted from the main AUs. Therefore, those geometric feature-based algorithms have been studied by many researchers in order to improve accuracy by the fusion of the features extraction methods.

B. CNN-BASED FER APPROACHES

Recently, due to the development of big data and the improvement of hardware technology, many algorithms based on deep learning have been researched. Since the field of FER is being influenced by these advancements as well, more robust and efficient feature recognition has been achieved through the automatic learning of the extracted facial features. In this section, we introduce the CNN-based FER algorithms.

Lopes *et al.* [18] suggested a representative facial expression algorithm applying deep learning based on CNN. This takes the data argumentation process to resolve the scarcity of the FER dataset and to make robust facial emotions to changes such as rotation and transportation. In this algorithm, except for the parts with unnecessary elements around the face, the AUs are cropped into blocks at the center of the action unit, and the emotions are classified into six to seven emotions though CNN. In such algorithms, the lack of datasets required for deep learning algorithms are solved by argumentation methods, and FER research based on CNN is being actively studied.

The FER approaches using dual networks, which fuse both holistic features of the face and partial features focused on facial landmarks, have been studied in Jung *et al.* [19], Xie and Hu [20] and Yang *et al.* [21]. In these approaches, one CNN network extracted features using facial gray scale images while the other network extracted features using image patches or landmark changes. Finally, these features are fused by a weighted function or fully-connected learning.

A method of extracting temporal features and spatial features to combine softmax and predict emotions was used by Zhang *et al.* [22]. The temporal features were extracted so as to learn temporal landmarks using the part-based hierarchical bidirectional recurrent neural network (PHRNN), and the multi-signal convolutional neural network (MSCNN) involved extracting holistic features that extract the overall facial features. The PHRNN classifies the facial landmarks into each AU and hierarchically learns the movements of AUs, which change with time. The MSCNN learns the facial gray scale images in order to extract the entire appearance features. Using these two networks, more accurate facial expression recognition could be made possible by considering both temporal and spatial features.

Despite many studies, the recognition rate is still not high enough, due to the influence of various environmental changes such as lighting and accessories as well as the difference in the characteristics of individual people. Therefore, in this paper, we would like to propose an efficient FER scheme with robust features combining two types of features using deep learning.

III. PROPOSED ALGORITHM

The proposed algorithm is shown in Figure 1. In this study, we propose an efficient algorithm to improve the recognition accuracy by a hierarchial deep neural network structure which can re-classify the result (Top-2 error emotion), which is the most frequent error. The first network learns the convolutional neural network, which focuses on AUs using the LBP feature, which is a typical feature extraction technique in the field of facial studies [9]. The second network extracts the geometric changes of the landmarks of each AU and learns all pairs of six emotions. Based on two features, the proposed algorithm combines them using adaptive weighting function to give the final result.

A. OBSERVATION: ERROR RATIO OF TOP-2 SELECTION

The ratio of the correct answer was measured by using the 6-length softmax results in order to determine the cause of errors and ratio of the facial recognition errors when only using a single network. The experiment was performed using two datasets with 150 images per emotion using the appearance feature-based CNN. This experiment was conducted by using a 10-fold cross validation method.

The entire dataset was divided into 10 sets, 9 of which were employed for training while the other one was used for

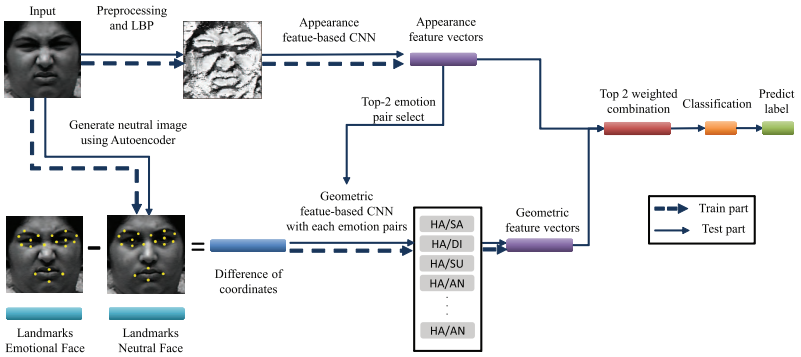


FIGURE 1. The overall procedure of the proposed FER algorithm.

TABLE 1. Top-2 error rate in CK + dataset.

	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6
Average of 10-fold (%)	95.3	4.2	0.3	0.1	0.0	0.1
Total of error (%)	4.7					
Ratio of Top-2 error (%)	90.5					

TABLE 2. Top-2 error rate in JAFFE dataset.

	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6
Average of 10-fold (%)	89.4	8.0	1.6	1.0	0.0	0.0
Total of error (%)	10.6					
Ratio of Top-2 error (%)	75.3					

verification. The softmax results for six emotions were sorted in descending order, and the rank of correspondence to the True answer was counted. The number of counts was divided by the total number of points of validation data in the fold, and the ratio was measured at each fold. The results are shown in Tables 1 and 2.

As shown in Tables 1 and 2, the case of including the second highest label in CK +, was observed by 4.2%. When Top-2 $\sim$ Top-6 were considered as errors and Top-1 is the correctly predicted result. The ratio of Top-2 resulted in 4.7%, which covers 90.5% of total error. In addition, the JAFFE dataset had the largest error rate of Top-2 at 8.0% which is occupied with 75.3% of 10.6% total error.

As a result, we can see that the error is biased in Top-2 error. The error occurs within the Top-2 range. In all datasets, Top-2 errors occurred at a rate of more than 82% of the total error. From this viewpoint, we can consider a structure to improve the recognition accuracy if we reduce Top-2 error rate by a refined classification.

To design an efficient scheme, robust features are extracted using hierarchical fusion of the two types of networks shown in Figure 1. The facial expression is predicted as one of six emotions: anger, disgust, fear, happy, sad, and surprise. First, the appearance feature-based network uses the LBP image in order to learn the holistic characteristics of the face in one frame. Secondly, the geometric feature-based network learns the change of eighteen x and y coordinates among the facial landmarks, which mainly move according to emotional changes. The predicted result of the appearance feature-based network, the two highest softmax values among the emotions are weighted with the results of the geometric feature-based network, respectively. Then, robust features are generated by fusing the different types of features. Finally, we can obtain a predicted final emotion.

From the next sub-section, each module will be explained in detail

B. PREPROCESSING

Before the main process of facial expression recognition, it is necessary to identify a face and recognize the facial area. Therefore, the face and non-face parts must be separated through the face detection process. Only preserving the important parts for emotion recognition prevents accuracy degradation due to changes in the environment surrounding the face. In this paper, we used the face detector model which was learned through P. F. Felzenszwalb *et al.* [23]: it is a detector that uses the HoG algorithm to determine the face boundary coordinates. We cropped the facial area using this detector. In this algorithm, the linear SVM was used to identify the facial region by training HoG features from positive (contain an object) and negative (not contain an object) samples composed of sliding window. The algorithm could be used as the get_frontal_face_detector defined in dlib [24] in order to identify and crop the facial area coordinates of left-top x, y, right-bottom x, y. The cropped facial images usually appear from the middle of the forehead to the chin, and from the leftmost face to the rightmost face.

After the facial region has been cropped, a blurring process is performed before creating the LBP image in order to remove noise, which is the input value of the appearance feature-based network. If the features are extracted from unfiltered facial images, it may lead to a

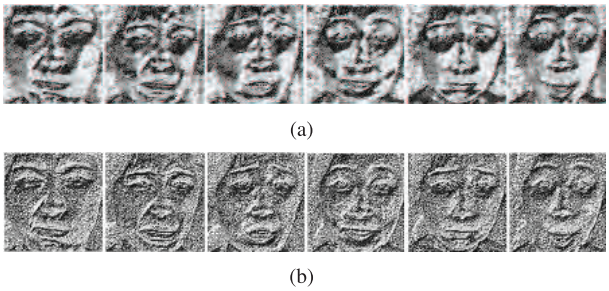


FIGURE 2. LBP images with applied preprocessing: (a)LBP images with bilateral blurring preprocessing, (b)LBP images without preprocessing.

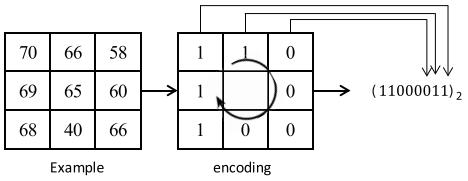


FIGURE 3. Example of encoding a LBP pattern.

degraded performance. In addition, since the importance of the AUs, which plays a major role in emotion change, is higher than that of other facial parts, they are converted to LBP images after preprocessing.

The result of using only the important parts extracted based on AUs is shown in Figure 2. We used a preprocessing technique involving bilateral blurring [25]. This filtering technique is one of the image analysis methods known as edge preserving smoothing. It performs Gaussian blurring while preserving the edge where the correlation with similar peripheral values is small.

C. THE APPEARANCE FEATURE-BASED NETWORK

The appearance feature-based CNN is a process for extracting the holistic features of the faces. In the proposed algorithm, we used LBP images which are robust in the FER system. This is a representative feature extraction technique used in the field of facial studies, because it extracts the texture of main facial AUs movement with a simple structure. It compares 3×3 neighbor pixels with 8 based on the center pixel, as shown in Figure 3.

Each pixel is encoded as a 1 if it is brighter than the center pixel, and as a 0 when it is darker. Then, these values are connected to each other in the clockwise direction from the upper left corner, and converted into decimal numbers to be used as feature values of a LBP image. These binary values are referred to as Local Binary Patterns or LBP codes [26]. The formula is written as the following:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P} s(i_p - i_c)2^p, \tag{1}$$

where $LBP_{P,R}$ means calculating the number of P neighboring pixels among the pixels in the radius of R from the center pixel. $LBP_{8,1}$ is used in the proposed method. (x_c, y_c) is the center coordinates of the block for making the LBP pattern. i_p and i_c denote the gray scale values of the neighboring pixels and the center pixel, respectively. $s(x)$, which is converted to a binary number based on the difference from the center pixel, is defined as the following:

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

The binary code extracted from the above equation is converted into a decimal number in order to form an LBP image. These images occupy less computational complexity and lower capacity than the original images, which enable learning and execution at higher speeds. These points are a great advantage, and they are used in facial recognition and facial emotion recognition because they can extract facial texture with good performance [26], [27]. For this reason, the LBP images are employed as input to the appearance feature-based network.

The convolutional neural network (CNN) has also been successfully used in computer vision applications. The CNN is a network that extracts feature maps by performing a convolutional operation with kernel on the original data. It is typically constructed of convolutional layers and pooling layers that extract feature maps expressing an image [28].

This CNN structure makes it possible to learn while preserving the shape characteristics of each component of the face image. In addition, maintaining the shape of the input and output data of each layer allows for the effective extraction of the facial expression features by considering the characteristics of the adjacent image. The formula used in the convolution operation is as:

$$x_{ij}^{l,d} = \sigma(b_{ij} + \sum_{m=0}^{K-1} \sum_{N=0}^{K-1} w_{mn} x_{i+m,j+n}^{l,d}), \tag{3}$$

where d is 2 in this study because of the 2-d convolution operation. The x denotes an output signal of i, j in the feature map of the l_{th} layer. The σ denotes a non-linear function, and b denotes a bias of i, j. w_{mn} is the weight value applied to the convolution operation with each kernel and the key to controlling the neuron.

As shown in Figure 4, the appearance feature-based network has a 128×128 size input, and passes through the convolution layers and the pooling layers a total of three times. The first convolutional layer performs a convolutional operation using 5×5 size kernel with the number of 4. Next, in the first pooling layer, the max-pooling, which involves selecting one pixel among the pixels in a 2×2 block, is processed. The kernel size was experimentally determined as 5×5 in order to fit 128×128 input sizes so as to effectively extract the image feature map while considering the typical application of a 3×3 kernel for a 64×64 size input. As a result, the 64 maps of 128×128 size obtained in the previous step are changed to 64 maps of 64×64 sizes, which represent a reduction in size by half.

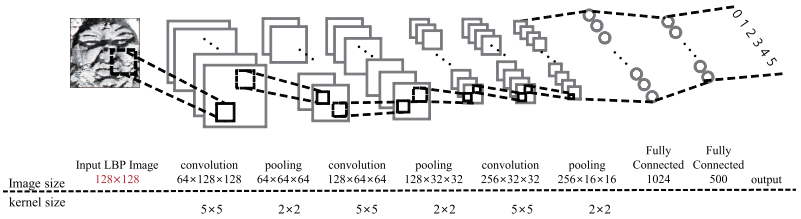


FIGURE 4. The proposed appearance feature-based CNN structure.

In a similar way, the of convolutional and pooling operations are repeated three times. Finally, 256 maps of 16×16 sizes as a result of the last pooling are derived. After the convolutional and pooling layers are finished, these values are flattened and passed through the fully connected layers, which are the hidden layers. The first fully connected layer has 1024 nodes, while the second has 500 nodes.

In the proposed network, we use the dropout operation between fully connected layers. When the network is learning, it turns off the neurons at random, thus disturbing the learning. This can prevent an over-fitting that is biased toward the learning data [29]. We use rectified linear unit (ReLU) as the activation function between the convolutional layer and fully connected layer. This activation function is a step for converting the quantitative value of the feature map through the convolution operation into a nonlinear value. At the end of the network, six emotions are extracted as continuous values using the softmax function. The formula of the softmax result for six emotions is as the following:

$$s_i = \frac{e^{a_i}}{\sum_{k=0}^{n-1} e^{a_k}}, \tag{4}$$

where n corresponds to the number of emotions requiring classification. s_i is the softmax function score of the i_{th} class. This value is the sum of the exponential values of the a_ks, which are the values of the entire category, and is then divided by the exponential value of the emotion score. The error is computed by the network through this process, and the error is reduced by using the cross entropy loss function. The calculation of the loss is considered as the following:

$$L = -\sum_{j=0}^{n-1} y_j log(s_j), \tag{5}$$

where y_j is the j_{th} element of the correct answer vector. Using the cross entropy function, it is possible to flexibly respond to various probability distributions of the model by obtaining the cross entropy through a negative log-likelihood. In addition, the process of finding a gradient is relatively simple as well [30].

To classify six categories, which is n, if the first element is correct, $y = [1, 0, 0, 0, 0, 0], y_1 = 1$ and the others are zero. s_j is the output value of the softmax function. In addition, we use a steepest gradient descent (SGD) as an optimizer along with the calculated cross entropy loss.

The results of two emotions with the highest value among the softmax results extracted using this learned model are later used for more accurate emotion prediction by fusing with the result of the geometric feature. Thus, the label information for these two high emotions are transmitted to the geometric feature-based network.

D. THE GEOMETRIC FEATURE-BASED NETWORK

We considered both types of the appearance feature-based feature and geometric feature in order to reduce recognition errors by using more robust features. In the case of using only one network, the recognition accuracy is inclined to be low because of various factors, such as rotations, illuminations, and peripheral accessories. Further, in the case of fine emotional change, it is difficult to recognize emotion only using the holistic features of the face.

In this paper, we additionally use a geometric feature-based CNN that captures the movements of the landmarks of emotion. The feature of the partial elements obtained by detecting the movement of the landmark is added to the overall features so that more robust features can be extracted. Furthermore, we detected and demonstrated that the facial expression recognition error most frequently occurred in the emotion of the second highest probability when only using the appearance network. Based on this assumption, the Top-2 values with the highest values among the results of 6-classes probabilities composed of the last softmax layer are selected. Finally, the emotion is classified through the max result of two network fusion by weighted sum calculation.

For extracting geometric features which contain the dynamic features of a face, a neutral facial image of the person depicted in the reference facial image is required. However, in the real system, there is not enough neutral facial images. There are also some FER datasets which do not have enough neutral image data. In this case, we need to obtain enough neutral image data to learn the dynamic feature of the facial expression. We suggest the autoencoder to generate neutral image data. This network can be used to learn the geometric feature-based network by obtaining the difference of coordinates between the generated neutral facial image and the emotional image, and to create dynamic features. The proposed autoencoder technique is presented in Figure 5.

This network is constructed from the VGG19 network structure [28]. The images of neutral faces can be generated using this structure. It can be divided into the encoding and

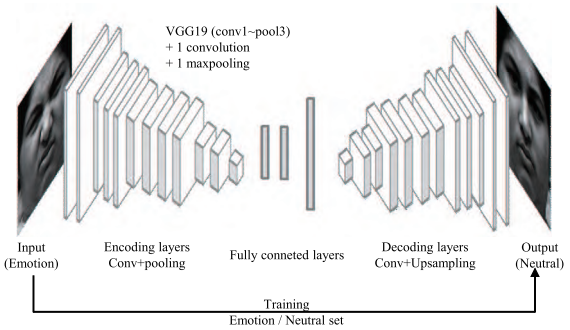


FIGURE 5. Autoencoder structure for generating neutral image.

decoding processes. First, in the encoding process, the input of the facial image with emotion is passed to the pool3 layer of VGG19. Next, the number of 1 convolution layer and 1 maxpooling layer are performed one more time, then the features are compressed through the twice fully connected layers of 4096 nodes.

In the decoding process, the fully connected layer of the 4096 extracted nodes is reshaped, and the upsampling and convolution processes are repeated as opposed to the encoding step. In this process, an output value equal to the input size is derived. The error function for reducing the loss is obtained by calculating the difference between the input facial image with emotion and the neutral facial image of the same person in the input image, unlike the previous case, in which the autoencoder narrows the difference between the input and output values. The error function can be written as:

$$error = MSE(G_n, X_n), \tag{6}$$

where G_n is the neutral facial image generated from the facial image with emotion, and X_n is the neutral facial image in the existing datasets. The ground truth X_n is used as the same person in the input image. The formula used for MSE to reduce this error is:

$$MSE = \frac{1}{m}\sum_{i=0}^{n-1}(g_i - x_i)^2, \tag{7}$$

where m is the number of pixels for an image and g_i and x_i are the i_{th} pixel of the generated neutral image, and the i_{th} pixel of the ground truth image, respectively. In this way, the autoencoder model is constructed so as to generate neutral facial data for input facial images by reducing loss.

The neutral facial images generated through this process are shown in Figure 6. The autoencoder had been trained for about 170 epochs for a week and used an Adam optimizer with a learning rate of 0.01. The CK+, JAFFE datasets with neutral image data were used as training data. Using the generated neutral image, dynamic features can be extracted using geometric feature-based CNN, even though we do not have enough such image data.

The geometric feature-based CNN is a process which involves capturing dynamic changes in facial expressions and extracting geometric features of landmarks. In the proposed

FIGURE 6. Generated neutral images with the autoencoder (Top: facial images with a emotion, Bottom: neutral facial images).

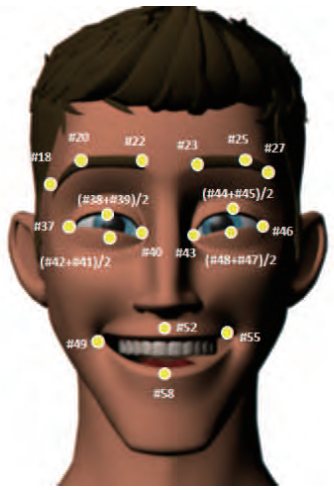


FIGURE 7. Landmarks using geometric feature-based CNN.

algorithm, 68 landmarks of a face defined by the iBUG 300-W face landmark dataset [31] were obtained using a real-time landmark extraction technique which used the ensemble of regression trees Kazemi [32]. Among the obtained landmarks, we calculate a difference of the main AUs's coordinates such as eyebrows, eyes, and nose between the neutral facial image and the emotion facial image. The coordinates are shown in Figure 7.

As shown in Figure 7, we selectively use the landmarks of the main AUs, such as the eyebrow edge, eye periphery, and mouth edge among the landmarks defined as 1 to 68. First, for eyebrows, we used 22 and 23, which are the points nearest to the middle of the forehead, as well as 18 and 27, which correspond to the ends of the eyebrows. In addition, we use 20 and 25 in the center of the landmark of each eyebrow area. Next, for eyes, we use 40 and 43, which are closest to the nose, and both ends of the eyes at 37 and 46.

In order to obtain each center of the eyes, two points, except for both ends of the eyes, are calculated as average values by dividing the top and bottom and used as each center coordinate of the eyes. At the edge of the lips, 49 and 55 are used at the ends of the lips, and the center of upper lip 52 and the center of lower lip 58 are used.

As a result, a total of 36-length x and y coordinates are constructed in order to extract the geometric features. The following is a vector used to calculate the difference between the coordinates of the neutral facial image and the facial

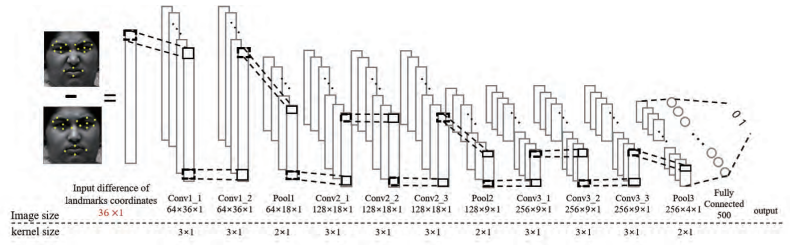


FIGURE 8. Geometric feature-based CNN structure.

TABLE 3. Fifteen pairs of six emotions for the geometric feature-based model (AN:Angry, DI:Disgust, FE:Fear, HA:Happy, SA:Sad, SU:Surprise).

	pair1	pair2	pair3	pair4	pair5
emotion1	AN	AN	AN	AN	AN
emotion2	DI	FE	HA	SA	SU
	pair6	pair7	pair8	pair9	pair10
emotion1	DI	DI	DI	DI	FE
emotion2	FE	HA	SA	SU	HA
	pair11	pair12	pair13	pair14	pair15
emotion1	FE	FE	HA	HA	SA
emotion2	SA	SU	SA	SU	SU

image with emotion.

$$V_e = [x_e^0 - x_n^0, y_e^0 - y_n^0, \cdots, x_e^{35} - x_n^{35}, y_e^{35} - y_n^{35}], \quad (8)$$

where V_e is the geometric vector of emotion e. Each x_e and y_e coordinate is listed in order from left to right, top to bottom, and the eyes edge and lips coordinate are constructed clockwise order, top, right, bottom, and left. The 36-length vector extracted from the above equation is learned through the CNN constructed by borrowing from the VGG16 Network [33].

The structure of the CNN is as shown in Figure 8.

Through the network composed as shown in Figure 8, one of the results among the fifteen pairs shown in Table 3 are weighted in the Top-2 result of the appearance feature-based CNN. The feature extraction is efficiently performed using part of the VGG16 structure verified as having a high performance in classification studies. In addition, geometric feature can be learned by maintaining the individual geometric order as well as the shape of components.

In the first convolution layer of Conv1_1, the feature maps are generated by performing convolutional operation using 64 kernels with 3×1 size. The convolution operation is also performed in the same way in Conv1_2, then the maxpooling is operated using the kernel of 2×1 size. Next, the Conv2_1 is operated using 128 kernels with 3×1 sizes, and this process is repeated in Conv2_2 and Conv2_3 as well. Next, the maxpooling process using the 2×1 kernel size is operated. Next, 256 convolution operations are repeated in Conv3_1, Conv3_2, and Conv3_3 using the same kernel size of Conv2. Following the Pool3 operation, the feature is flattened and connected with a fully-connected layer of 500 nodes of size, and we then obtaining softmax results of the two categories.

We employ the ReLU as the activation function in the same way as the appearance feature-based network, and the dropout is used to overcome the overfitting problem. The loss function for learning is calculated using the cross entropy loss function in the same way as the appearance feature-based network. The model for each pair is stored, and the model corresponding to the Top-2 pair from the result of the softmax of the previously obtained appearance model is selected. In this way, we consider a weighting function to determine the final emotion.

E. WEIGHTING FUNCTION FOR TOP-2 EMOTIONS

The emotion recognition error is the most frequent among the two highest results. The Top-2 emotion results of the appearance feature-based network are weighted by the geo- metric feature-based network result in order to more accurately predict emotion by taking into consideration the holistic feature and partial feature of a face. In the appearance feature-based CNN mentioned in Section 3.2, the two classes with the Top-2 softmax values among six classes are selected.

In the geometric feature-based CNN mentioned in Section 3.3, the 2-length softmax result by using the pair about the Top-2 classes from the appearance network is extracted. The extracted softmax values are fused by the following:

$$C_k = \alpha A_k + (1 - \alpha) G_k, \quad (9)$$

where α is a real-value between 0 and 1, corresponding to a weight for the combination of two features. In this thesis, the α value set a 0.8. The k is 0 or 1, as the class with the first highest softmax value is 0 and that with the second highest is 1. A_k is a softmax result of the appearance feature-based network corresponding to the k_{th} emotion class, while G_k is a softmax result of the geometric feature-based network. The softmax function for the operation is shown in Eq. (3.5). The Top-2 softmax results of the appearance feature-based CNN are normalized to the same ratio of geometric results. For example, they are operated so that the sum of the two classes is 1, before being combined with the geometric results as the following:

$$A_{Top_k} = \frac{a_{Top_k}}{\sum_{i=0}^{j} a_{Top_i}}, \quad (10)$$

where a is a softmax value for a class and Top_k is the class number with the k_{th} highest softmax value. The max value

of j is 1. The A_{Top_k} resulted by the formula is fused to the geometric result G_k at a rate of α, resulting in C_k of Eq. (9). The combined C_k using this value is re-scaled by the ratio occupied by Top-2 in the previous appearance feature-based network in order to finally determines the facial expression. The re-scaling process can be represented as:

$$R_k = (a_{Top_0} + a_{Top_1}) \times C_k. \tag{11}$$

The re-scaling value R_k, is obtained by multiplying the sum of a and C_k.

It is the combination of the class with the k_{th} highest softmax value and the ratio $(a_{Top_0} + a_{Top_1})$ occupied by Top-2 in the appearance feature-based network. Therefore, the softmax vector is obtained by fusing the two networks, and the predicted emotion are as follows:

$$F = [v_0, v_1, \cdots, v_5], \tag{12}$$

$$pred_e = \arg\max_i v_i. \tag{13}$$

The obtained R_k is used as the value v_i of the emotion class corresponding to 0 to 5 softmax values. Therefore, the result of the 6-length softmax function consists of the vector for the final prediction. The emotion i with the largest value of v like Eq. (12) and the final emotion is selected by Eq. (13). If the predicted i label value correspond to 0, then the final emotion is Angry. If i is from 1 to 5, then the final emotion will be Angry, Disgust, Fear, Happy, Sad, or Surprise.

IV. EXPERIMENT AND DISCUSSION

In this section, the experimental results and discussion are conducted in order to support the evidence of the suggested algorithm. We also demonstrate the superiority of the algorithm by comparing the accuracy with that of the state-of-the-art algorithm. In the following sub-sections, we describe the datasets for experimentation. The hardware and parameter settings in which the experiment were performed are described. Finally, we present the results of the experiments and analyze the performance.

A. DATASETS

1) EXTENDED COHN-KANADE (CK+)

The CK + is a dataset which has the labeled emotion number for the frame of sequences from neutral to peak states. A total of 123 subjects participated and 593 image sequences were included, with 327 of them being labeled with seven universal emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise) [34]. This dataset is employed by various algorithms related to facial expression recognition. Therefore, it is suitable for evaluation against the latest technology.

In this thesis, six emotions of anger, disgust, fear, happiness, sadness, and surprise were used as the data of the experiment. The last three frames in each sequence were used as peak emotion frames, and there were approximately 80-120 images for each emotion, so that a total number of 927 images were used. The verification method for accuracy uses a 10-fold cross validation method for the purpose of comparison with other algorithms. It divides the dataset into 10 sets, nine of which are used for learning, and the last one is used for validation. The accuracy of the facial expression recognition algorithm is averaged using these results. Figure 9 shows CK+ dataset examples of frontal facial images.

FIGURE 9. Extended Cohn-Kanade (CK+) dataset.

B. JAPANESE FEMALE FACIAL EXPRESSION (JAFFE)

JAFFE is a dataset consisting of gray scale frontal facial expression images of 10 Japanese women. It contains a total of 213 images including seven facial expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral) [35]. We use a total of 915 augmented points of data for learning and validation, including rotations, flip, and noise. The rotations were constructed 5 degrees clockwise and counterclockwise, respectively, and the noise was added to the original image with Gaussian noise with a variance of 0.01 and a zero mean.

In order to compare the proposed method with the latest algorithm, we verified the 10-fold cross validation method in the same way as the CK+ dataset, and measure the accuracy by averaging the results. Figure 10 shows examples of the JAFFE dataset.

C. EXPERIMENTAL ENVIRONMENT

They were based on Windows OS i7-8700 CPU with a clock speed of 3.20Ghz, RAM of 8GB, and GTX 1070 GPU. The proposed method is a deep learning-based algorithm. Therefore, we model CNN and Autoencoder using Tensorflow and Keras. It is also based on the python language, which is optimized for deep learning modeling and related libraries. The experimental results, including the accuracy, are verified for each dataset by learning and the 10-fold validation method.

In the training process for each network, 30 epochs were iterated for each data set and the learning rate was 0.01. The stochastic gradient descent (SGD) was used as the optimizer.

D. PERFORMANCE ANALYSIS

1) WEIGHT DETERMINATION FOR FUSION OF FEATURES

In order to improve the error in the Top-2 range, we designed an algorithm for fusing two networks. As shown in Eq. (3), the prediction result is calculated as the highest value by combining the results of the geometric feature-based network, which are obtained from the softmax results of the appearance

FIGURE 10. Japanese female facial expression (JAFFE) dataset.

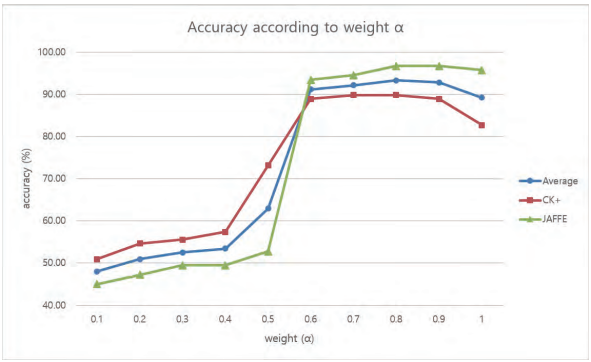


FIGURE 11. Accuracy according to weight value (α).

feature-based network. The value of α in Eq. (9) refers to the degree of contribution of the appearance feature-based network. It has a real value between 0 and 1.

We divide the experiment into 10 sets for each dataset. We used nine sets for training and the other set for testing. The result of the experiment is shown in Figure 11.

In this figure, The CK+ case is displayed with square points and the JAFFE case is represented with triangles. The circle points illustrate the average values. The experiments were carried out on the number of nine weights, ranging from 0.1 to 1.0 at intervals of 0.1. Each dataset is used in the same way as described in the previous sub-sections. A total of 2000 datasets were used. As a result, the CK+ dataset provided the highest accuracy of 89.81% when the weight α value was 0.7 or 0.8. The used of the JAFFE dataset resulted in the accuracy of 96.7%, where the weight value α was 0.8 or 0.9. The average results was 93.26% using an α value of 0.8. Therefore, the weight α of 0.8 was selected for combining two features.

2) QUANTITATIVE VERIFICATION

In this section, we will verify whether facial expressions are correctly recognized by the proposed algorithm for each dataset. The accuracy was measured using the 10-fold cross validation method for each dataset. First, we measured the confusion matrix using only the appearance feature-based network. In addition, we measured the confusion matrix of the proposed method, which is fused with the geometric feature-based network using the weight $\alpha = 0.8$; the confusion

TABLE 4. Confusion matrix of appearance feature-based CNN in the CK+ dataset. (AN:Angry, DI:Disgust, FE:Fear, HA:Happy, SA:Sad, SU:Surprise).

Actual values \ Predicted values	AN	DI	FE	HA	SA	SU
AN	**86.9**	2.2	0.5	0.8	9.6	0.0
DI	1.2	**93.2**	0.0	1.8	3.8	0.0
FE	0.0	0.0	**100.0**	0.0	0.0	0.0
HA	0.0	0.6	2.8	**96.6**	0.0	0.0
SA	3.1	0.0	0.0	0.0	**96.9**	0.0
SU	1.1	0.0	0.8	0.0	0.7	**97.4**

TABLE 5. Confusion matrix of the proposed method in the CK+ dataset. (AN:Angry, DI:Disgust, FE:Fear, HA:Happy, SA:Sad, SU:Surprise).

Actual values \ Predicted values	AN	DI	FE	HA	SA	SU
AN	**91.6**	0.0	1.4	0.8	6.2	0.0
DI	1.3	**95.4**	0.0	0.0	3.3	0.0
FE	0.0	0.0	**100.0**	0.0	0.0	0.0
HA	0.0	0.4	2.8	**96.8**	0.0	0.0
SA	3.1	0.0	0.0	0.0	**96.9**	0.0
SU	1.6	0.0	0.0	0.0	0.3	**98.1**

TABLE 6. Comparisons of our approach and the state-of-the-arts of FER approaches in CK+ dataset.

Method	Accuracy (%)
DCMA-CNNs [20]	93.46
Salient Facial Patches [14]	94.09
RBM [36]	95.66
Appearance feature-based network	**95.15**
The proposed method	**96.46**

matrix is the average of all results. In addition, we will compare the results of the proposed method using only the appearance feature-based network for each fold. We also describe the average values as the final results.

In the CK+ dataset, the accuracy was measured by dividing 927 points of data into 10 sets. The confusion matrix of the appearance feature-based network with LBP features is shown in Table 4. The horizontal axis presents a predicted class among six emotions, and the vertical axis is the actual class which is the correct answer. In the CK+ dataset, the accuracy for angry was 86.9% and the accuracy for fear was the highest. The angry emotion was often mistakenly predicted as sad emotion. This is because, in both of these emotions, the facial images appeared with the ends of the mouth pointing downward. Therefore, it seems that the probability of error is increased by similar the appearance-based features if only appearance feature-based network is used. The average for total emotions was 95.15% of the accuracy.

The results of the confusion matrix of the proposed method combining the geometric feature-based network extracting the changes of the AUs landmark on the face and the appearance feature-based network are shown in Table 5. The highest rate of recognition was for angry, which was the lowest accuracy of all. The ratio of increment was 4.7%, from 86.9% to 91.6%. In addition, the accuracy was maintained or increased across all emotions. The average for total emotions was 96.46%, showing an average increase of 1.3%.

TABLE 7. Confusion matrix of appearance feature-based CNN in the JAFFE dataset. (AN:Angry, DI:Disgust, FE:Fear, HA:Happy, SA:Sad, SU:Surprise).

Actual values \ Predicted values	AN	DI	FE	HA	SA	SU
AN	**95.7**	3.6	0.0	0.0	0.7	0.0
DI	7.1	**90.1**	0.0	0.0	0.6	0.0
FE	0.0	3.8	**86.9**	1.2	4.4	3.8
HA	0.0	0.5	0.8	**98.2**	0.0	0.5
SA	3.4	5.4	13.1	3.5	**74.6**	0.0
SU	0.0	0.0	6.4	3.1	0.0	**90.5**

TABLE 8. Confusion matrix of the proposed method in the JAFFE dataset. (AN:Angry, DI:Disgust, FE:Fear, HA:Happy, SA:Sad, SU:Surprise).

Actual values \ Predicted values	AN	DI	FE	HA	SA	SU
AN	**95.7**	3.6	0.0	0.0	0.7	0.0
DI	4.9	**92.3**	2.2	0.0	0.6	0.0
FE	0.0	4.4	**88.0**	0.6	4.4	2.6
HA	0.0	0.5	0.8	**98.2**	0.0	0.5
SA	3.2	1.2	14.2	2.2	**79.2**	0.0
SU	0.0	0.0	4.3	3.1	0.0	**92.6**

TABLE 9. Comparisons between our approach and the state-of-arts FER approaches in the JAFFE dataset.

Method	Accuracy
CNN [18]	84.48
Salient feature [37]	90.00
Multi-Level Haar Wavelet [38]	90.56
Appearance feature-based network	**89.33**
The proposed method with fusion	**91.27**

In Table 6, the proposed algorithm is compared with the state-of-arts algorithms. As mentioned previously, we used 10-fold cross validation using the CK+ dataset. Xie *et al.* [20] proposed a facial expression recognition algorithm using deep comprehensive multi-patches aggregation dual CNN (DCMA-CNNs) method. The similarity with our proposed method is that they both use a dual network. This method focused on static features with LBP image patches as opposed to dynamic features. As a result, the accuracy for the CK+ dataset was 93.46%. Happy and Routray [14] also focused on the appearance of each active patch in one frame rather than the dynamic feature. The method resulted in 94.09% accuracy in the CK+ dataset.

Another state-of-the-art is an RBM-based method [36] that combines temporal and spatial features. This method used facial-expressions and non-facial-expressions to learn emotional changes by using RBM-based models. The use of facial-expressions and non-facial-expressions is similar to our geometric feature-based network part, which used the neutral facial images and emotional face images. The accuracy of this method was measured at 95.66%.

Compared with these algorithms, the appearance feature-based network showed similar accuracy of 95.15%. However, the proposed fusion network with a geometric feature-based network showed better performance than other algorithms by a factor of 96.46% of accuracy.

In the JAFFE dataset, the accuracy of the 10-fold cross validation was measured as well, and the average value was calculated. This dataset contains black and white images with relatively high amounts of noise information. For this reason, the result was slightly lower accuracy than that of the CK+ dataset. In Table 7, the result of the appearance feature-based network was shown by the accuracy of 89.33%. The lowest rate was 74.6% in the sad and showed the most errors in fear and disgust. The highest value was 98.2% in happy.

The result of the confusion matrix of the proposed algorithm is shown in Table 8. The results were similar to those of the CK+ dataset. In particular, the error of the disgust emotion was the same as in the sad emotion which showed the lowest accuracy. As a result, the accuracy of the sad emotion was increased by about 4.6%. Finally, we achieved an average accuracy of 91.27%, which was increased by 1.7% on average.

In Table 9, the proposed algorithm was compared with the state-of-arts algorithms for JAFFE datasets. Lopes *et al.* [18] suggested a representative method that used a CNN in the FER. They learned data with shuffle training in order to change the order so that the method could be learned with less data. The accuracy of the basic CNN algorithm using only the static feature was 84.48%. Another recent algorithm is that presented by Liu *et al.* [37], which extracted features of salient areas using LBP and HoG features with gamma correction. This method resulted in 90% accuracy. Finally, Goyani and Patel [38] used feature vectors by constructing the Haar wavelet of multiple levels with the face, eyes, and mouth.

These methods also used spatial information in a frame. Therefore, the performance was similar to that of the appearance feature-based network. The proposed algorithm was verified with a rate of 91.27% and achieved higher accuracy than other methods.

As a result, we illustrated the improved accuracy and confusion matrix results of the proposed network. The hierarchical network which focuses on improving the most frequent Top-2 errors, has improved the accuracy of recognition in all datasets. We have also achieved better performance than those of the latest algorithms. The proposed method for all datasets achieved an average of at least 91.27%, up to 98.07% of recognition accuracy; an average of 94.67% was achieved.

To validate the performance in more wild dataset, we tested the proposed algorithm on the AffectNet dataset, which is a dataset of about 1M of face images collected by internet search engine as shown in a Figure 12 using about 1,250 emotions related with keywords. This dataset has 10 categories (0: Neutral, 1: Happiness, 2: Sadness, 3: Surprise, 4: Fear, 5: Disgust, 6: Anger, 7: Contempt, 8: None, 9: Uncertain, 10 : No-Face). In this experiment, the proposed hierarchical deep neural network scheme achieved about 88.3% of the accuracy.

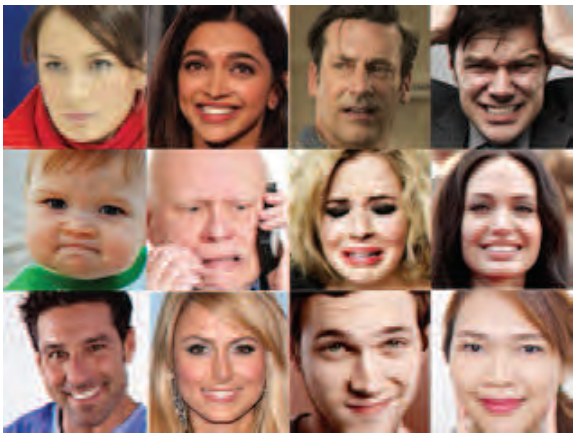


FIGURE 12. Examples of the affectNet dataset.

V. CONCLUSION

We have proposed an efficient facial expression recognition algorithm combining appearance feature and geometric feature based on deep neural networks for more accurate and efficient facial expression recognition. The appearance feature-based network extracts the holistic feature of the LBP feature containing the AUs information. The geometric feature-based network extracts the dynamic feature, which is the face landmark change centered on the coordinate movement between the neutral face and the peak emotion. As a result, we constructed more robust feature by combining static appearance feature from the appearance network and dynamic feature from the geometric feature-based network.

In the experiments, we have shown that the Top-2 error frequently occurred with average about 82% using only appearance feature-based network. As a result of improving this error with the proposed algorithm, we achieved about 96.5% accuracy with 1.3% improvement when comparing to the other algorithms in the CK+ dataset. In addition, the proposed algorithm yielded 91.3% of the accuracy which was improved by 1.5% when compared with other existing methods in the JAFFE dataset. From the experiments for all datasets, the accuracy of six emotions was at least maintained or enhanced significantly.

REFERENCES

[1] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[2] T. Yukitake, "Innovative solutions toward future society with AI, robotics, and IoT," in *Proc. VLSI*, Kyoto, Japan, Jun. 2017, pp. C16–C19.

[3] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2016.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Kauai, HI, USA, Dec. 2001, pp. I–I.

[5] P. I. Wilson and J. Fernandez, "Facial feature detection using Haar classifiers," *J. Comput. Sci. Colleges*, vol. 21, no. 4, pp. 127–133, Apr. 2006.

[6] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients.," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 1491–1498.

[7] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, to be published. doi: 10.1109/TAFFC.2017.2731763.2017.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[9] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, C, Appl. Rev.*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[10] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognit.*, vol. 45, no. 1, pp. 80–91, Jan. 2012.

[11] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007.

[12] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression recognition using hidden Markov models," *Signal Process., Image Commun.*, vol. 17, no. 9, pp. 675–688, 2002.

[13] Y. Liu *et al.*, "Facial expression recognition with PCA and LBP features extracting from active facial patches," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Angkor Wat, Cambodia, Jun. 2016, pp. 368–373.

[14] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Mar. 2015.

[15] B. Ryu, A. R. Rivera, J. Kim, and O. Chae, "Local directional ternary pattern for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 6006–6018, Dec. 2017.

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

[17] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.

[18] A. T. Lopes, E. de Aguiar, A. F. de Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[19] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 2983–2991.

[20] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211–220, Jan. 2019. doi: 10.1109/TMM.2018.2844085.2018.

[21] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.

[22] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[24] E. D. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.

[25] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.

[26] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, C, Appl. Rev.*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[27] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

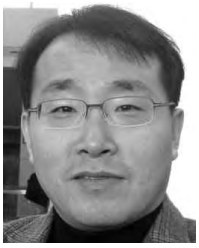
[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, Cambridge, MA, USA: MIT Press, 2016.

[31] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.
[32] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, Ohio, Jun. 2014, pp. 1867–1874.
[33] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[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-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.
[35] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. Third IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 200–205.
[36] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal RBM-based model for facial expression recognition," *Pattern Recognit.*, vol. 49, pp. 152–161, Jan. 2016.
[37] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors*, vol. 17, no. 4, 712, 2017.
[38] M. Goyani and N. Patel, "Multi-level haar wavelet based facial expression recognition using logistic regression," *Indian J. Sci. Technol.*, vol. 10, p. 9, Mar. 2017.

JI-HAE KIM was born in Seoul, South Korea, in 1994. She received the B.S. and M.S. degrees from the College of Engineering, Sookmyung Women's University, Seoul, South Korea, in 2017 and 2019, respectively, where she was a Researcher with the Intelligent Vision Processing Laboratory (IVPL), from 2017 to 2019. Her research interests include the development of computer vision processing and facial expression recognition techniques using deep learning and machine learning, fundamental study of image feature extraction, and image classification.



BYUNG-GYU KIM (M'04–SM'16) received the B.S. degree from Pusan National University, South Korea, in 1996, the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), in 1998, and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, KAIST, in 2004.

In 2004, he joined in the Real-Time Multimedia Research Team, Electronics and Telecommunications Research Institute (ETRI), South Korea, where he was a Senior Researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award, in 2007. From 2009 to 2016, he was an Associate Professor with the Division of Computer Science and Engineering, Sun Moon University, South Korea. In 2016, he joined the Department of Information Technology (IT) Engineering, Sookmyung Women's University, South Korea, where he is currently an Associate Professor. He has published over 200 international journal and conference papers, and holds patents in his field. His research interests include image and video signal processing for the content-based image coding, video coding techniques, 3-D video signal processing, deep/reinforcement learning algorithm, embedded multimedia systems, and intelligent information systems for image signal processing.

Dr. Kim is a Professional Member of the ACM and IEICE. He has received the Special Merit Award for Outstanding Paper from the IEEE Consumer Electronics Society, IEEE ICCE 2012, the Certification Appreciation Award from the SPIE Optical Engineering, in 2013, and the Best Academic Award from the CIS, in 2014. He also served or serves as an Organizing Committee Member of CSIP 2011, a Co-Organizer of CICCAT2016/2017, and a Program Committee Member of many international conferences. He is serving as a Professional Reviewer for many academic journals, including the IEEE, ACM, Elsevier, Springer, Oxford, SPIE, IET, MDPI, and so on. In 2007, he served as an Editorial Board Member for the *International Journal of Soft Computing*, *Recent Patents on Signal Processing*, the *Research Journal of Information Technology*, the *Journal of Convergence Information Technology*, and the *Journal of Engineering and Applied Sciences*. Since 2018, he has been the Editor-in-Chief of *The Journal of Multimedia Information System* and an Associate Editor of the IEEE Access Journal. He is serving as an Associate Editor for *Circuits, Systems and Signal Processing* (Springer), *The Journal of Supercomputing* (Springer), the *Journal of Real-Time Image Processing* (Springer), and the *International Journal of Image Processing and Visual Communication* (IJIPVC).

PARTHA PRATIM ROY (M'87) received the Ph.D. degree in computer science from the Universitat Autònoma de Barcelona, Spain, in 2010.

He was a Postdoctoral Research Fellow with the Computer Science Laboratory (LI, RFAI Group), France, and with the Synchromedia Laboratory, Canada. He was also a Visiting Scientist with the Indian Statistical Institute, Kolkata, India, for more than 6 times. He has gathered industrial experience while working as an Assistant System Engineer at TATA Consultancy Services, India, from 2003 to 2005, and as a Chief Engineer at Samsung, Noida, from 2013 to 2014. He is currently an Assistant Professor with the Department of Computer Science and Engineering, IIT Roorkee, Roorkee. He has participated in several national and international projects funded by the Spanish and French Government. He has published more than 160 research papers in various international journals and conference proceedings. His main research interest includes pattern recognition. In 2009, he received the Best Student Paper Award from the International Conference on Document Analysis and Recognition (ICDAR).

DA-MI JEONG was born in Bucheon, South Korea, in 1995. She received the B.S. degree from the College of Engineering, Sookmyung Women's University, Seoul, South Korea, in 2018, where she is currently pursuing the M.S. degree. Since 2018, she has been researching with the Intelligent Vision Processing Laboratory (IVPL), Sookmyung Women's University. Her research interests include the development of computer vision processing and real-time facial expression recognition techniques using deep learning, feature extraction, and image classification.