Applied Soft Computing Journal 84 (2019) 105724

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Recognizing learning emotion based on convolutional neural networks and transfer learning

Jason C. Hung [a,*], Kuan-Cheng Lin [b], Nian-Xiang Lai [b]

[a] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taiwan
[b] Department of Management Information Systems, National Chung Hsing University, Taichung, Taiwan

ARTICLE INFO

Article history:
Received 13 March 2019
Received in revised form 29 July 2019
Accepted 10 August 2019
Available online 21 August 2019

Keywords:
Learning emotion
Facial expression recognition
Convolutional neural network
Transfer learning
Model generalization

ABSTRACT

Learning effectiveness is normally analyzed by data collection through tests or questionnaires. However, instant feedback is usually not available. Learners' facial emotion and learning motivation has a positive relationship. Therefore, the system identifying learners' facial emotions can provide feedback that teachers can understand students' learning situation and provide help or improve teaching strategy. Studies have found that convolutional neural networks provide a good performance in basic facial emotion recognition. Convolutional neural networks do not require manual design features like traditional machine learning, they automatically learn the necessary features of the entire image. This article improves the FaceLiveNet network with low and high accuracy in basic emotion recognition, and proposes the framework of Dense_FaceLiveNet. We use Dense_FaceLiveNet for two-phases of transfer learning. First, from the relatively simple data JAFFE and KDEF basic emotion recognition model transferring to the FER2013 basic emotion dataset and obtained an accuracy of 70.02%. Secondly, using the FER2013 basic emotion recognition model transferring to learning emotion recognition model, the test accuracy rate is as high as 91.93%, which is 12.9% higher than the accuracy rate of 79.03% without using the transfer learning model, which proves that the use of transfer learning can effectively improve the recognition accuracy of learning emotion recognition model. In addition, in order to test the generalization ability of the Learning Emotion Recognition Model, videos recorded by students from a national university in Taiwan during class learning were used as test data. The original database of learning emotions did not consider that students would have exceptions such as over eyebrows, eyes closed and hand hold the chin etc. To improve this situation, after adding the learning emotion database to the images of the exceptions mentioned above, the model was rebuilt, and the recognition accuracy rate of the model was 92.42%. By comparing the output of maps, the rebuilt model does have the characteristics of success in learning images such as eyebrows, chins, and eyes closed. Furthermore, after combining all the students' image data with the original learning emotion database, the model was rebuilt and obtained the accuracy rate reached 84.59%. The result proves that the Learning Emotion Recognition Model can achieve high recognition accuracy by processing the unlearned image through transfer learning. The main contribution is to design two-phase transfer learning for establishing the learning emotion recognition model and overcome the problem for small amounts of learning emotion data. Our experiment results have shown the performance improvement of two-phase transfer learning.

1. Introduction

In the past researches [1,2] two kinds of emotions could be recognized by analyzing facial expression which include basic emotion and complex emotion. Basic emotion was proposed by Ekman [1] and divided into Anger, Disgust, Fear, Happiness, Sadness, and Surprise. The six basic emotions are instinctive response, no need to learn. Everyone has a similar reaction to the same situation. Basic emotions could be understood easily.

Social emotion is not nature emotion, and shall be obtained through specific learning process. This kind of emotion is more complicated than basic emotion. People, in other words, may have different reactions in the same situation, and different reactions may reflect different emotions. In this paper, we try to recognize this kind of complex emotion derived by cognitive affect. The Cognitive Affect was presented by Sidney D'Mello [2] in 2007

* Corresponding author.
E-mail addresses: jhung@nutc.edu.tw (J.C. Hung), kuanchenglin@gmail.com (K.-C. Lin), dryadf66449@gmail.com (N.-X. Lai).

https://doi.org/10.1016/j.asoc.2019.105724
1568-4946/

could be described as the learner's emotional reactions when they are in learning process. The learning processes of cognitive affect were defined as Learning Emotion. Complex emotions are not born with human beings. They must be learned through communication with others by socialization. Everyone has different complex emotions for the same event or thing. The learning emotion could be divided into six categories, including frustration, confusion, boredom, flow, delightful, and surprise. The detection of learning emotion is applied in the field of e-learning, which can be used for learning motivation assessment [3], learning state analysis [4] , concentration detection [5], etc. Complex emotions are usually caused by the differences in moral standards and values of each person. The learning emotion discussed in this paper is included in complex emotions.

General process for the implementation of facial emotion recognition can be summarized as follow. First, retrieving image(s) from recorded video stream. Second, to transfer original images into low-dimensional ones which could be kept the detail information of the original image by using PCA and feature selection [3], etc. Third, a specific model is expected by fusing the obtaining features into machine learning algorithms, such as Multilayer Perceptron (MLP) [6], Support Vector Machine (SVM) [7] and K Nearest Neighbor (KNN) [8] are examples to classify the images.

Traditional machine learning methods require specific features for learning but cannot learn via unknown data. On the other hand, the advantage of deep learning methods is getting higher accuracy through the characteristic of automatic learning. Zhuang Liu [9] proposed DenseNet based on deep CNN method to get high basic emotion recognition accuracy rate up to 77% in ImageNet. FaceLiveNet proposed by Zuheng [10] also get 68.60% accuracy rate in FER2013 dataset.

When you should use Convolution Neural Network or traditional Machine Learning method is a hard question to answer because it depends on the problem you are trying to solve. For every problem, a certain method is suitable and achieves good results while another method fails heavily. Traditional machine learning methods have some bottlenecks in emotion recognizing because it is hard to label feature points and high-quality image is required. Usually, Convolution Neural Networks are also more computationally expensive than traditional machine learning methods. State of the art deep Convolution Neural Network can take several weeks to train completely from scratch. Most traditional Machine Learning methods take much less time to train, ranging from a few minutes to a few hours or days. The amount of computational power required for a Neural Network depends heavily on the size of your data but also on how deep and complex your Network is. The problem of computational power was solved because of widely usage of Graphics Processing Unit (GPU). Convolutional Neural Network (CNN) architecture relies on a large amount of computing resources such as GPUs. The biggest characteristic of convolutional neural networks is that they can automatically learn the original data without feature engineering. The characteristics, however, through the interleaving of the convolutional layer and the pooled layer, the deeper network architecture can obtain more complicated features. The prediction model obtained by Convolutional Neural Network is better than Traditional machine learning in generalization ability [11].

CNN has excellent recognition performance for images because of the following two features:

1. Local receptive fields: Each neuron does not need to perceive the entire image but only need to perceive the local area.
2. Shared weights and biases: The convolution kernel is reusable. The weight of the shared convolution kernel ensures that the neurons used by the convolution of the network layer can learn the same features. These features still can be detected even if these features appear in different images.

The biggest problem of the traditional Deep Neural Network (DNN) is to ignore the relative relationship of original data when inputting data. The "shape" of the original data will be ignored because the traditional deep neural networks have a limitation that the input data must be one-dimensional.

So Convolutional Neural Networks also have better image recognition results than traditional machine learning methods [9, 12–14].

Past researches have demonstrated a basic emotion recognition model through open basic emotion datasets such as CK+ [15], KDEF [16], JAFFE [17], and FER2013 [18]. The Japanese Female Facial Expression (JAFFE) datasets contain 213 facial expression images of the neutral expression and six basic facial ones with same light source, angle, gender, and Action Unit(AU) labeling. The FER2013 dataset contains images of real life conditions such as different angles, light, gender, and ethnicity. FER2013 dataset consists of 35.887 grayscale, 48 × 48 sized face images with various seven emotions labeled as following.

0: -4593 images- *Angry*

1: -547 images- *Disgust*

2: -5121 images- *Fear*

3: -8989 images- *Happy*

4: -6077 images- *Sad*

5: -4002 images- *Surprise*

6: -6198 images- *Neutral*

To compare the recognizing JAFFE dataset [17] by using SVM [7] and Convolutional Neural Network(CNN) [19], the obtained test accuracy are 95.71% and 86.38%. In this case, the traditional machine learning algorithm has higher recognition accuracy than Deep Learning with clear features. Another example is to recognize FER2013 Data set by using Local Binary Patterns(LBP) combining SVM [7] and CNN [10], the obtained test accuracy are 44.33% and 68.60%. In this case, the traditional machine learning algorithm could not recognize the image with unclear features effectively, but in the same situation, CNN can recognize the image by the characteristic of automatic learning and obtained the higher accuracy.

CNN can recognize the image by the characteristic of automatic learning [19]. Therefore, we can build the model to deal with images with unclear features and still maintain the high recognition accuracy. Maxime Oquab [20] mentioned that if the number of datasets is insufficient, the overfitting problem might occur due to too many parameters in original CNN architecture. Therefore, using the original model through transfer learning could solve the overfitting problem more effectively.

FaceLiveNet CNN architecture proposed by Zuheng [10], has good recognition result in FER2013 dataset. Therefore, we try to adopt and improve FaceLiveNet to recognize learning emotion. We built the basic emotion recognition model first then build learning emotion recognition model by using the model though transfer learning [21].

Considering the above issues, this study first designs a general emotion recognition model to verify the suitability of all kinds of CNN architecture. Second, using the suitable model which is chosen in first step to build learning emotion recognition model according to transfer learning then verify the effectiveness of transfer learning. Third, we will describe the significant contribution of the proposed model by exporting the saliency maps [22] to verify the relationship between emotion recognition model and Action Unit. The performance of proposed model will be verified via the classroom teaching finally. The propose model is expected to provide learner's real-time learning emotion for instructors, and let instructors improve teaching strategy more effectively. This study will

Table 1
Example of facial action units.

AU1	AU4	AU5	AU7
			
Brow up	Brow gather	Eyelid elongation	Eyelid distance reduced
AU12	**AU15**	**AU20**	**AU27**
			
Mouth up	Mouth sinking	Mouth translation	
AU24	**AU25**	**AU26**	
			
Lips shrink	Lips slightly open	Chin down	Mouth stretch

1. Compare the applicability of various CNN architecture to basic emotion recognition.
2. Use transfer learning to build learning emotion recognition model and verify the recognition effective.
3. Analyze and describe the relationship between the saliency map which exported by CNN and AU.
4. Verify the generalization ability of the learning emotion recognition model through the data of classroom teaching.

The main contribution is to design two-phase transfer learning for establishing the learning emotion recognition model and overcome the problem for small amount of learning emotion data. After training the CK+ dataset with the lab basic emotions, the trained model can preserve the parameters to define the facial feature related to the lab basic emotions. The pre-trained model for CK+ will then be transferred to train the FER2013 dataset with the large amount of natural basic emotions and improve the recognition accuracy for FER2013 dataset. Finally, the pre-trained model for FER2013 will be transferred to train the learning emotion dataset contains small amount of data. Our experiment results verify the performance of two-phase transfer learning.

2. Related works

2.1. Definition of facial action unit

In 1978, Ekman & Friesen has developed a system named Facial Action Coding System. They based on the face of the muscle structure, divided into 46 independent the face action units. These face action units can be combined each other, and output six basic emotions. These action units divided two parts: upper face and lower face. Table 1 shows the example of facial action units. These facial action units can make up six basic expressions: happiness, anger, sadness, fear, surprise, and disgust [23]. These action units are combined into six basic emotions, identify the students positive or negative learning emotions, and learn the state of their learning.

2.2. DenseNet

DenseNet is the deep convolutional neural network architecture published by Liu et al. [9] in CVPR 2017, in order to solve the problem that the gradient disappears as the depth of the network deepens. It uses Dense Block to connect all the network layers to ensure the transmission of information in the overall network do not disappear. The structure is shown in Fig. 1.

One of characteristics of the DenseNet is that the overall network is deep, narrow and allows less quantity, thus become the key reason of the Dense Block design. Rather than the thousands of outputs of Inception structures, the Dense Block provides less quantity on feature map in each convolutional layer. This narrow and deep connection method makes the features and gradients transmission become more efficient, and it is conducive to the overall network training. Since each layer of the convolution layer in the Dense Block is small, only 1×1 and 3×3, each convolution kernel is responsible for learning the tiny features. The stacking of the layers is finally connected together to create a very large output in Feature Map. After each Dense Block, the Translation Layer reduces the action of the dimension. In other words, the Translation Layer as a 1×1 convolution pooling layer, uses Reduce parameter to drop to the original Feature Map Channel output into half. (The Reduce parameter is preset as 0.5)

DenseNet has the following advantages. First, it can save parameters. The DenseNet structure has the same performance accuracy as ResNet [24] in ImageNet, but it requires fewer than half the parameters of ResNet. The emergence of this small model can save storage expenditures for the industry. Secondly, it can reduce the amount of computation. DenseNet requires less amount of computation compared to ResNet, and it has excellent performance, even without using Depth Separable Convolution. It still can achieve better results than the existing methods.

2.3. FaceLiveNet

FaceLiveNet is a prediction model by Ming et al. [10] in early 2018 to identify facial emotions. The structure of the model is shown in Fig. 2 [10]. The entire model connecting by multiple InceptionBlocks via Residual Connect has two major features. First, the model uses only a small amount of parameters, 1.31M, to reduce model training time and computing resources. Second, it uses migration learning to improve emotion recognition. Through migration learning, this model first captured characteristics of the face and learn to recognize facial emotions. The advantage of this model is that it does not need to learn from facial contours but directly learns from more subtle features at the beginning to improve the prediction accuracy.

2.4. Swish

The self-gated activation function (Swish) [25] is a new activation function proposed by Google Brain in 2017. The Swish function formula is as shown in Eq. (1). Swish and ReLU [26] have

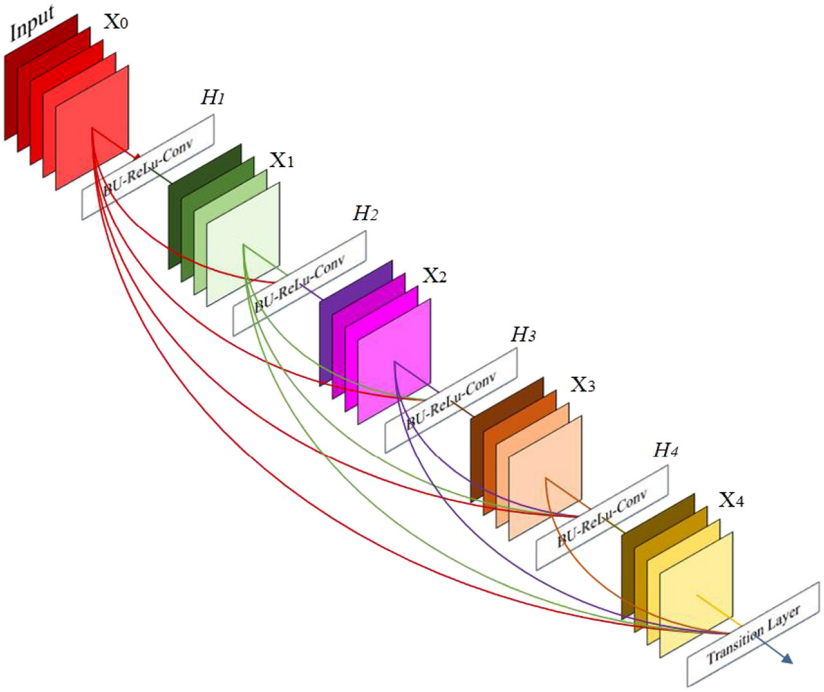


Fig. 1. Dense Block structure.

the same characteristics of no upper bound and lower bound. However, unlike ReLU, Swish has a smooth and non-monotonic function as shown in Fig. 3. Research has mentioned that Swish is more suitable for deep neural networks than ReLU in the case of using Batch Normalization [27].

The performance of using MobileNASNet-A on ImageNet is 0.9% higher than that of ReLU, and the performance on MobileNet is 2.2% higher than that of ReLU. When using Swish, just replace the formula of ReLU with Swish due to the same characteristics of Swish and ReLU.

$$f(x) = x \cdot sigmoid(x) \tag{1}$$

2.5. Transfer learning

In 2010, Pan et al. [21] proposed the concept of learning unknown knowledge through the existing knowledge called Transfer Learning. The core concept of this learning is to find the similarities between existing knowledge and unknown knowledge. Some knowledge domains are too abstract to learn, and resulting in high overall learning costs. Therefore, using existing knowledge to assist learning becomes important. For example, if people can write the JavaScript program and they can easily transfer the same learning model in Python.

The core concept of Transfer Learning is how to find the relevance between known and unknown and learn new knowledge. In Transfer Learning, the existing knowledge is usually called Source Domain and the unknown knowledge is called Target Domain. This learning mainly studies how to migrate knowledge from Source Domain to Target Domain. In the field of machine learning, Transfer Learning focuses on applying existing knowledge to unknown through the established model as shown in Fig. 4 [28]. For example, we can use the method of Transfer Learning to create a model of face recognition and learn about unknown facial emotions.

Transfer Learning can be divided into 4 categories (based on features, model-based, relationship, and sample) according to different learning methods.

- Feature-based transfer mainly involves mapping the Source Domain and Target Domain to the same space.
- Model-based transfer mainly combines the model with the sample to adjust model parameters.
- Relationship-based transfer mainly involves mapping the concept learning from Source Domain to the Target Domain, that is, the migration of knowledge.
- Sample-based transfer applies the weighted values of calibrated samples from Source Domain. In this article, we use model-based Transfer Learning in our study.

3. Model design for learning emotion

3.1. Research process

This section outlines two types of basic emotion datasets: (1) JAFFE and KDEF datasets that exhibit more exaggerated facial expressions and (2) a FER2013 dataset that exhibits more subtle facial expressions, with some expressions difficult for the human eye to distinguish.

Section 3.2 describes how JAFFE and KDEF datasets were used as training materials to compare the recognition models designed in this study with four other CNN architectures (i.e., VGG19, InceptionV3, Inception-ResNet V2, and DenseNet201) in terms of their applicability to basic emotion recognition. The recognition accuracies obtained after training were compared, and the correlation between the feature map of the emotion recognition model and a facial movement coding system was observed using the saliency map proposed by Simonyan et al. [29]. Subsequently, the FER2013 dataset was also used as training material, and a FER2013 basic emotion recognition model was established using the aforementioned five CNN architectures similar to the JAFFE and KDEF models through transfer learning.

Learning emotions are more complicated than basic emotions. Section 3.3 describes how the FER2013 basic emotion recognition model established in Section 3.2 was used to form a learning

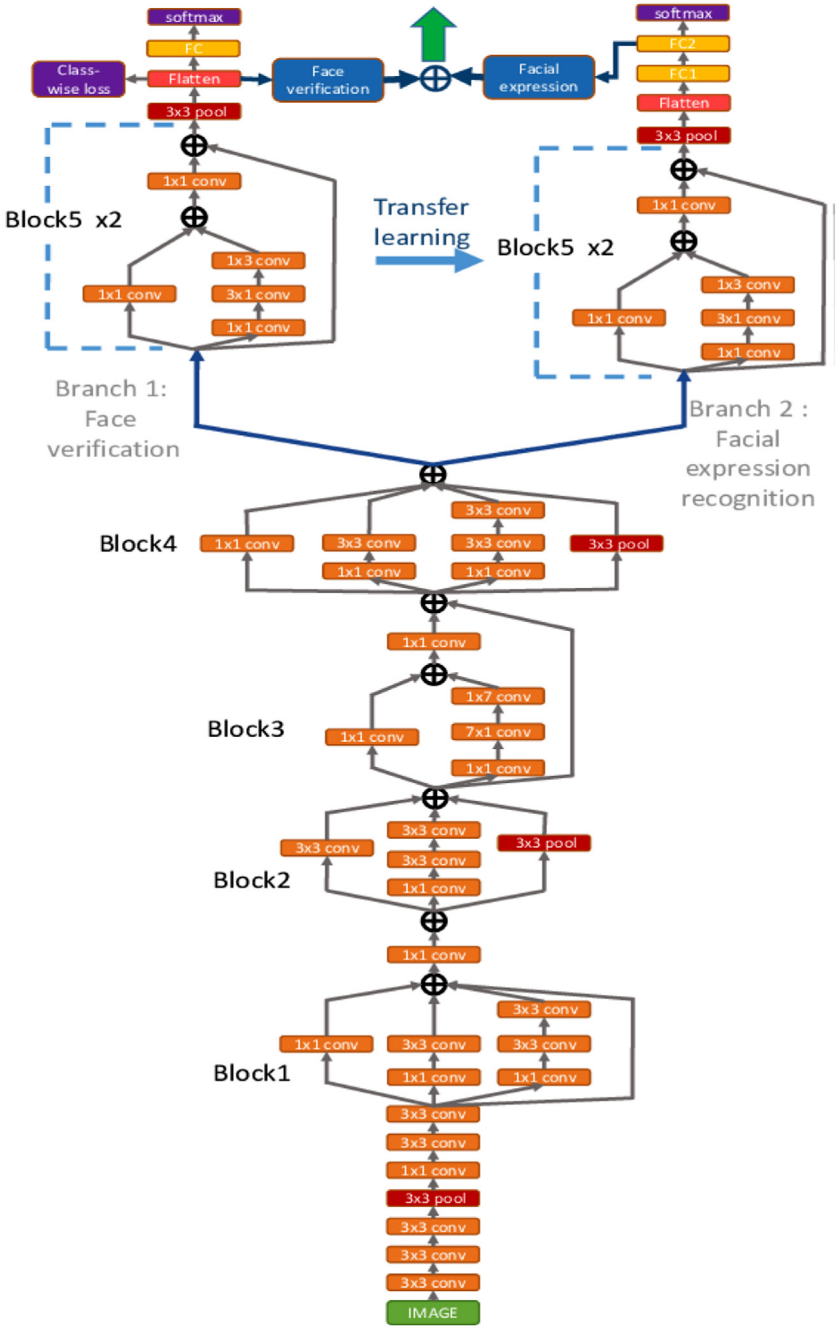


Fig. 2. FaceLiveNet model.

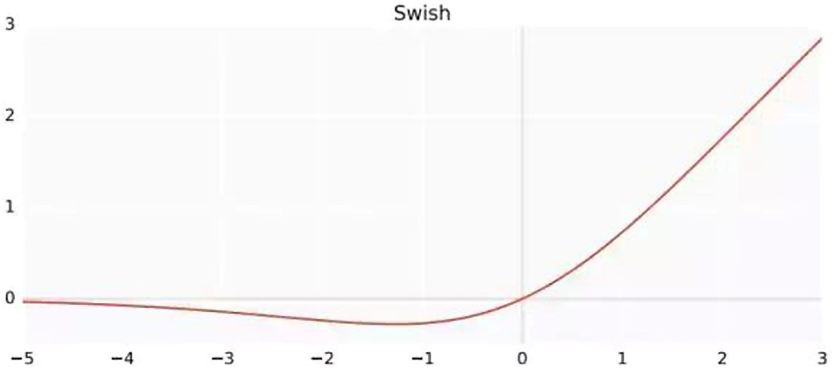


Fig. 3. Illustration of swish function.

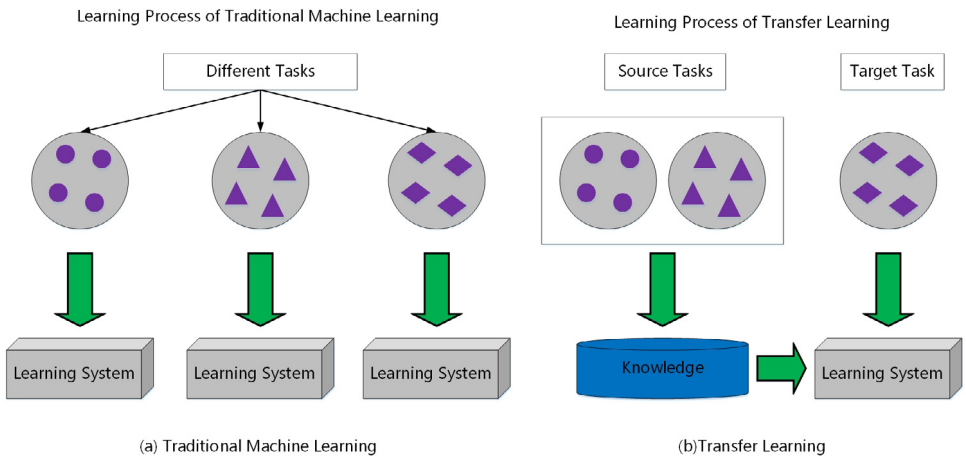


Fig. 4. The illustration of transfer learning.

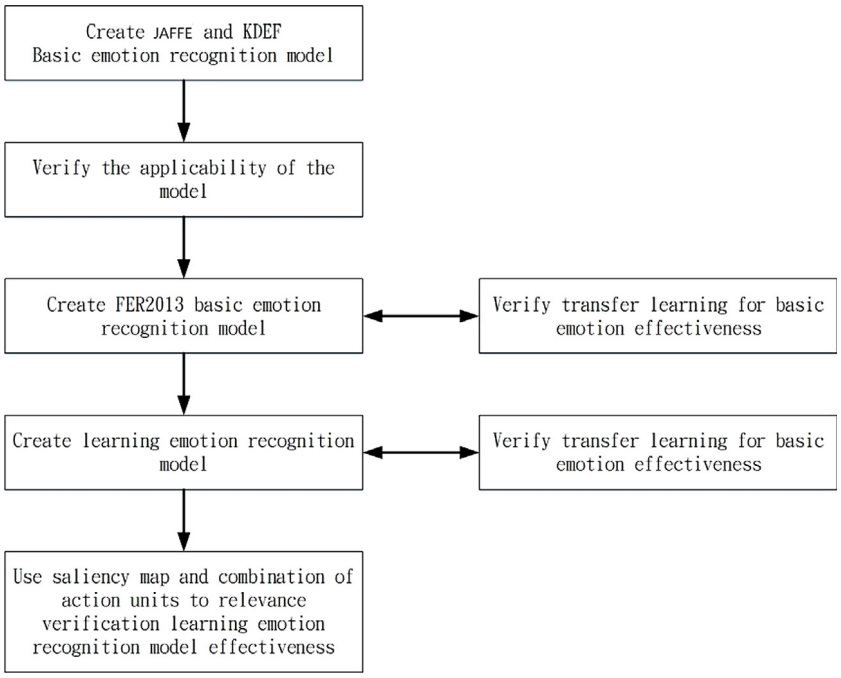


Fig. 5. Experimental process.

emotion recognition model through transfer learning. After establishing the model, indicators such as the saliency map, combinations of facial action units, and recognition accuracy were compared to verify the model's effectiveness. The research process is outlined in Fig. 5.

Column 1 — Establish basic emotion recognition models of JAFFE and KDEF; Verify models' applicability, Transfer learning; Establish basic emotion recognition model of FER2013, Transfer learning; Establish learning emotion recognition model; Verify the effectiveness of the learning emotion recognition model through the correlation between the saliency map and combinations of facial action units.

Column 2 — Verify the effectiveness of transfer learning on basic emotion; verify the effectiveness of transfer learning on learning emotion.

3.2. Basic emotion recognition model

This section outlines the steps for establishing the basic emotion recognition model (Fig. 6). First, image preprocessing details are provided for the basic emotion database and raw data, then followed by descriptions of the five CNN designs used in this study and adjustment of the hyper-parameters.

Furthermore, Calculations were performed on the saliency map output and accuracy of each model [22]. Finally, correlations between the facial action units and saliency map of each emotion are compared.

Block 1 — Basic emotion database; Image preprocessing; Architecture design of Dense_FaceLiveNet; Training of the Dense_FaceLiveNet model; Analysis results; Main process.

Block 2 — Visualization of the model's features; Accuracy of each classification; Mutual analysis and verification; Model analysis results; Subprocess.

Unlike conventional machine learning, using a CNN to build a model does not require manual design of features. Instead, CNNs use convolution kernels, the number of which is self-defined, as shared weights of different network layers. This enables the CNNs to automatically learn and classify the characteristics of different images during the training process. Therefore, the adjustment of hyper-parameters should only focus on the breadth, depth,

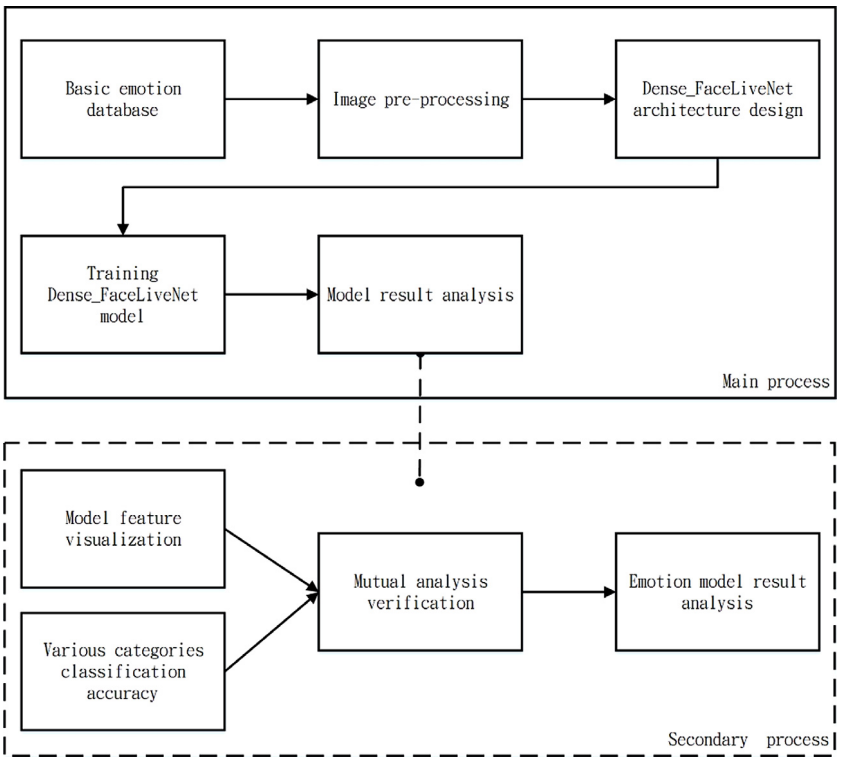


Fig. 6. Establishment of the basic emotion recognition model.

number of kernels, and size of the CNN when designing the model.

3.2.1. Data preprocessing

The original image was processed through face capture, image grayscale, and output size normalization (Fig. 7).

[17,30] found experimentally with two AdaBoost-based face detection systems that the detect rates may vary up to 10% by simply changing the parameters of the RGB to Gray conversion. We proposed to convert the original RGB image to gray scale to decrease the noises which caused by color in the first step. The second step is to detect the face through the face tracking and detection by using the gray scale image, and then to obtain the detected face image. We will normalize the face images by angle and size to benefit the model construction.

3.2.2. Design of the Dense_FaceLiveNet architecture

This study designed a CNN architecture named Dense_FaceLiveNet by modifying FaceLiveNet – proposed by [10] – and referring to the DenseNet design concept that serialized each Inception layer.

Dense_FaceLiveNet as proposed in this study is shown in Fig. 8.

The following three improvements have been made to the FaceLiveNet architecture:

(1) Replacement of the fully connected layer with global average pooling

The fully connected layer converts features learned by the stacking front convolution and pooling layers into a tag output, and the feature map output from the original convolution layer is converted into a vector. Subsequently, the vector is multiplied to reduce its dimensions before being output to the corresponding categories through softmax. Excessively large number of parameters is the major source of errors when using the fully connected layer. The parameters accounted for 85% of the entire network architecture, resulting in overfitting.

In [31], a global average pooling concept was proposed to replace many of the parameters in the entire network was proposed. The global average pooling sets the size of the average pooled stride to the input feature map size, and the output is set to the feature map channel support vectors, the number of which is similar to that of output categories. For example, if the basic emotion has seven categories of output, the number of feature map channels of the global average pooling will be seven, and the output will also be seven vectors.

To effectively reduce the parameters of the overall network and the overfitting problem, this study replaced the fully connected layer with global average pooling.

(2) Replacement of the residual block with a dense block

The main feature of the Inception network [14] is to use different convolution kernel sizes to learn features of different sizes. For example, larger convolution kernels are used to learn global features such as facial edges, whereas smaller convolution kernels are used to learn subtle local features such as movements of the eyes and mouth.

The concept of residual connections is similar to passing on a real-world message, where meanings may be misinterpreted as more people involved. A deeper network may cause problems related to gradient vanishing or gradient exploding. Through shortcuts, residual connections pass the message directly from the first person to the third person (i.e., the second person is skipped). Furthermore, $H(x)$ that originally must learn approximation is converted into function x through identity mapping [24] (2). This method can effectively reduce gradient vanishing or exploding. FaceLiveNet leverages residual connections in ResNet to incorporate the Inception structure, as shown in Fig. 9. Each layer of residual connections is connected with a 1×1 convolution kernel to reduce the dimensionality of the output multidimensional features and lead the reduced output feature map to the next network.

$$F(x) = H(x) - x(1) \tag{2}$$

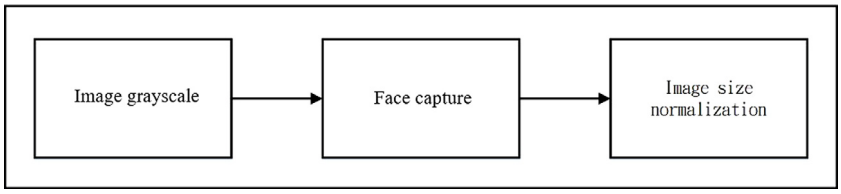


Fig. 7. Simple basic emotion recognition model.

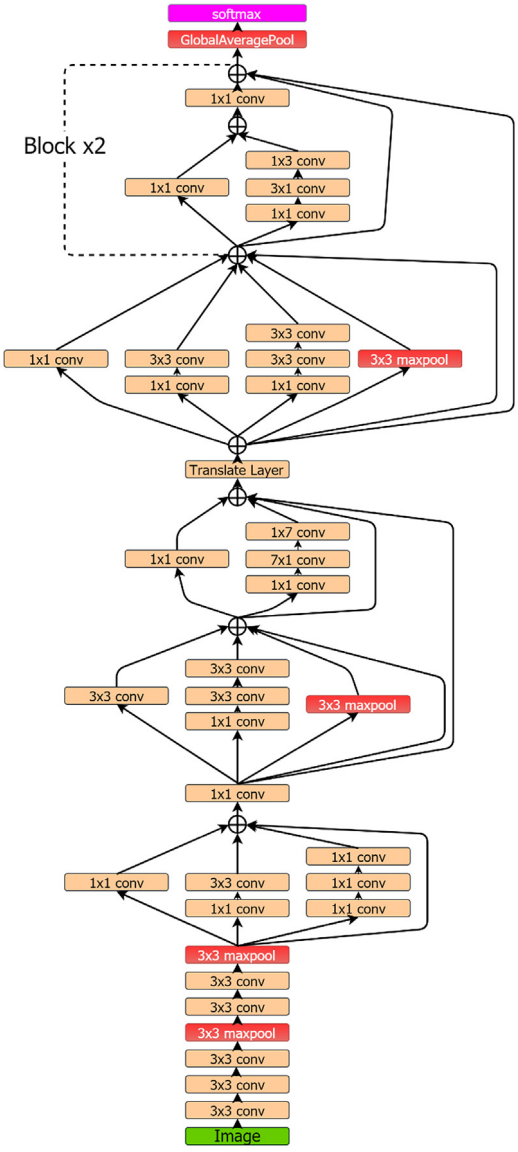


Fig. 8. Diagram of the Dense_FaceLiveNet architecture.

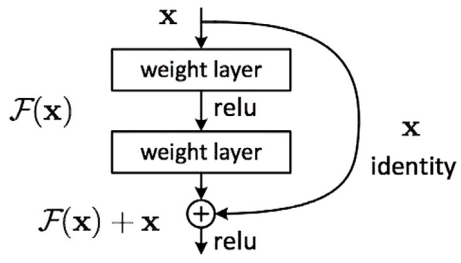


Fig. 9. Schematic of the residual network structure [32].

[9] noted that DenseNet possesses higher recognition accuracy for ImageNet than ResNet. Therefore, this study replaced the residual block with a dense block. The main concept of the dense block is that a neural network is not necessarily a progressive network. Features learned in one layer (*i*th layer) do not necessarily depend on the features transmitted by the upper layer ($i-1$ layer), but can backtrack to the features of the previous few layers. The dense block serializes each layer in the network to realize feature reusability. The formula of the dense block is as shown in (3) [9], where *XI* is the output and *HI* is the function of the *i*th layer. This study set each layer of Inception in the Dense_FaceLiveNet as *X* and formed *HI* by connecting each layer of Inception. The dense block passes the features learned by each layer of Inception to all subsequent Inceptions. In this study, all Inceptions were connected together to form an architecture, known as a dense Inception block. As shown in Fig. 10, a translation layer was added after *XI* to reduce the dimension of the learned features. Two layers of the dense block and a translation layer were used in Dense_FaceLiveNet.

$$x_I = H_I([x_0, x_1, \ldots, x_{\{I-1\}}]) \tag{3}$$

(3) Replacement of ReLU with Swish as the model activation function

The aforementioned improvement methods are to optimize the network architecture. This study sought to prove that Dense_FaceLiveNet can be improved using similar training methods without changing the network architecture. The direction of optimization efforts has shifted to using different activation functions to improve model quality.

A new activation function, Swish, was proposed by Google Brain [25] in 2017. Swish has the same features as ReLU, which is a smooth and nonmonotonic function. Therefore, Swish can be used to directly replace ReLU. Test accuracy with Swish was found to be 0.9% higher than ReLU in ImageNet, and Swish was therefore selected for use in Dense_FaceLiveNet.

3.2.3. Dense_FaceLiveNet training model

This research adjusts the super-parameter strategy as in the Inception-ResNet structure [33], and stated as follows:

(1) Use Adamax to train the model to converge.

A. Learning rate is set to 0.01
B. The other parameters according to [34]
C. If the accuracy rate does not decrease in each 10 epochs, multiply the learning rate by 0.2.

(2) Preserve the optimal model
(3) Use stochastic gradient descent strategy to train the model preserved in the second step to converge.

A. The initial learning rate is set to 0.001
B. Decay is set to 10-6
C. Momentum is set to 0.9
D. Using Nesterov momentum
E. Multiply the learning rate by 0.5 in each 2 epochs.

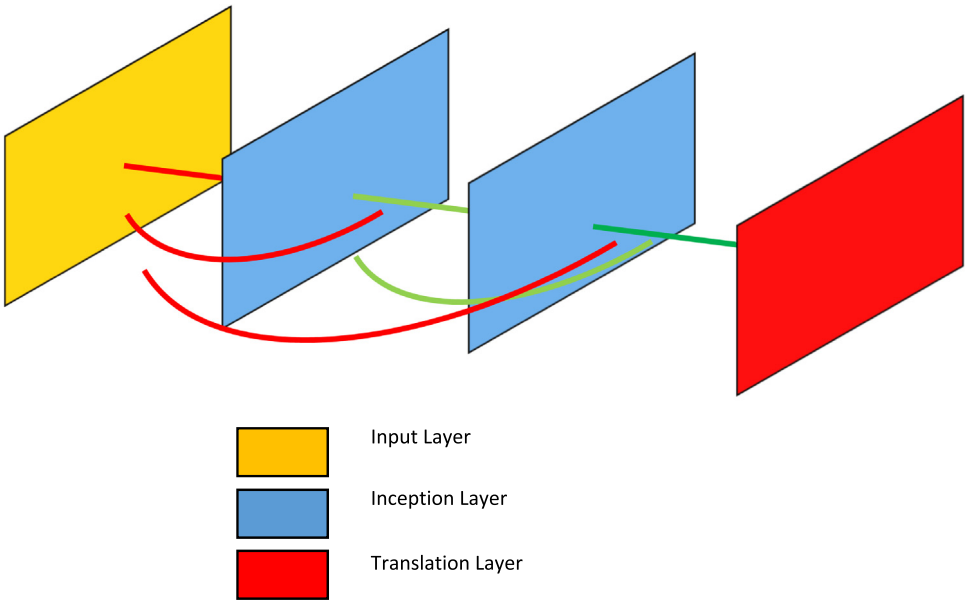


Fig. 10. Diagram of the dense Inception block architecture.

3.2.4. Transfer learning

The process of conventional transfer learning execution is illustrated in Fig. 11. First, the last layer of the model is removed, and the parameters of the remaining layers are set as untrainable. Second, an output layer is added, and the node is set as the number of output targets. The overall model is then trained to convergence. Subsequently, other layers are opened for the same training process to retrain the entire network to convergence. These processes are conducted to prevent large differences in the parameter values during transfer learning (i.e., the newly added output layer contains the initialization parameters, whereas the other network layers contained the training parameters). Differences can cause a large gradient update in the overall network, which causes the trained parameters to lose their effectiveness in the early stages of training. Dense_FaceLiveNet as proposed in this study does not use a fully connected layer in the last layer; thus, it can be directly trained during transfer learning by changing the node of the last output layer to a new number of datasets.

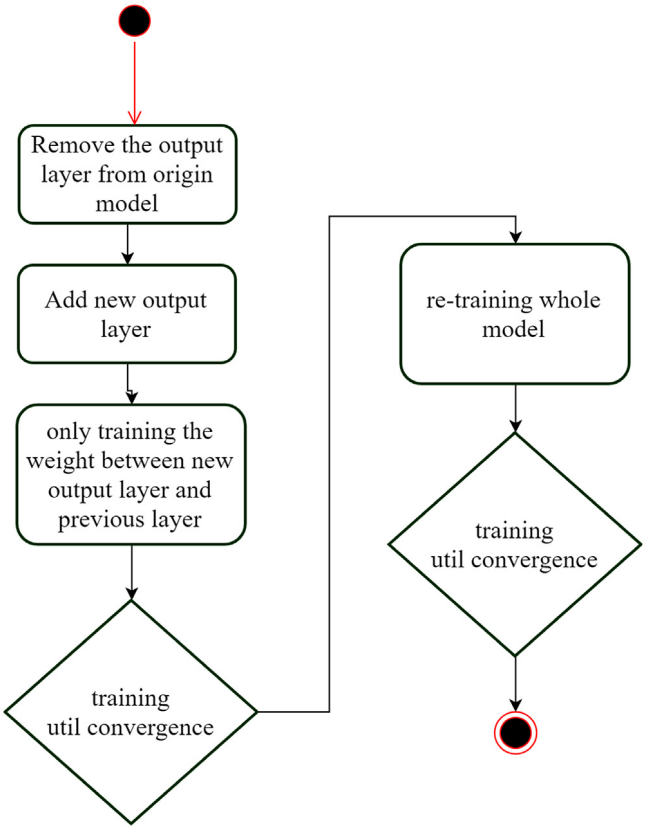


Fig. 11. Training flowchart of conventional transfer learning.

3.3. Learning emotion recognition model

This section explains the establishment process of the learning emotion recognition model, with Sections 3.3.1 and 3.3.2 outlining learning emotion database establishment and learning emotion recognition model establishment, respectively.

3.3.1. Learning emotion database

There is no learning emotion open dataset currently, and each person may have different responses for the same situation. These different reactions will lead to different learning emotions. Thus, it is very difficult to label learning emotions into classification. Four students of the department of information management at a university in central Taiwan were recruited to collect raw image data; they were asked to record videos on popular online platforms such as YouTube and VoiceTube. Ekman [7] noted three durations of emotional expression: short (1 s), normal (3 s), and long (4 s). However, short and long expressions were not employed in this study because their occurrences are rare. Each image was retrieved every 3 s from an original video, and the total image data is shown in Table 2. Learning emotions are difficult to identify. Therefore, the operational definition proposed by Jhong [35] was used to identify the learning emotions represented by each image, as described in Table 3 [35].

Row 1 — Learning emotion; Operational definitions
Row 2 — Frustrated; Facial expression, Frown and closed mouth movement; Description, It is difficult, difficult to understand.
Row 3 — Confused; Facial expression, Frown and mouth slightly open or closed; Description, Let me think about it.

Table 2
Sources of image data for the learning emotion database.

No.	Sample image	Sex	Data samples
1		Female	587
2		Female	549
3		Female	675
4		Female	189

Row 4 — Bored; Facial expression, Narrowed upper eyelids, listless, distracted or mouth in unnatural position; Description, 1. Can I quickly skip this concept? 2. Are there any other interesting topics?
Row 5 — Delighted; Facial expression, Smile or laugh; Description, 1. I understand! 2. It is interesting.
Row 6 — Surprised; Facial expression, wide-open eyes or open mouth and rising eyebrow; Description, Oh no! What is this?
Row 7 — Flow; Facial expression, Eyes focusing on the screen; Description, None

3.3.2. Model establishment

The network architecture used for learning emotion recognition establishment was Dense_FaceLiveNet, as proposed in Section 3.2.2. Data preprocessing, training methods, and data augmentation were consistent with the method in Section 3.2.3, and are thus not repeated in this section. The model was established using 5-fold cross-validation. A FER2013 basic emotion recognition model was employed to conduct two rounds of transfer learning in order to establish the final learning emotion recognition model.

4. Experimental results and discussions

4.1. Improvement of the FaceLiveNet architecture

As described in Section 3.2.2, this study improved the three main features of FaceLiveNet to enhance the accuracy of the basic emotion recognition model. As shown in Table 6, the FER2013 basic emotion dataset was employed to test the accuracy of the model. In [25], the accuracy of FaceLiveNet scored 68.60%. In this study, the fully connected layer was replaced with a global average pool. Although the recognition accuracy did not improve, but the overall number of parameters decreased sharply. Subsequently, the residual block was replaced with a dense block, which increased the recognition accuracy from 68.61% to 69.85%. Finally, ReLU was replaced with Swish to raise the recognition accuracy from 69.85% to 70.02%.

4.2. Correlation between CNN features and action units

Each basic emotion has a specific facial expression, which is composed of different face muscle movements. Based on the action units in a facial movement coding system, Ekman proposed that different emotions have different combinations of action units. In a CNN, the neural network enables the overall network to learn more effective features by continuously correcting the convolution kernel. Features learned by a CNN have a strong and positive correlation with the action units [32], that is, the features learned by the CNN are the action units. The present study verified whether each feature learned in Dense_FaceLiveNet can successfully correspond to an action unit. This section describes how the features were visualized using a saliency map and verifies the effectiveness of the model created by Dense_FaceLiveNet by comparing the correlations between the features and action units.

In this study, emotion images (except neutral emotion images) were randomly selected from the KDEF basic emotion dataset. To observe the positions of the action units corresponding to the features, outputs of the saliency map were overlapped with the original images. Because different convolution kernels tend to learn different action units [32], this study used all of the convolution kernels in the output layers of the basic emotion recognition model to output the saliency map.

The correspondence between the saliency map (output: fear, anger, disgust, happiness, sadness, surprise) and facial action units is as follows.

Fear

The experimental results are shown in Table 4. In the action unit (AU) coding system, fear is primarily composed of AU1, AU4, AU5, and AU25. AU1 represents raised inner brows, which correspond to the red boxes in the saliency map; AU4 represents a frown, corresponding to the green box in the saliency map; AU5 represents raised upper eyelids, corresponding to the blue boxes in the saliency map; and AU25 represents opened lips, corresponding to the purple box in the saliency map. This study verified that the basic emotion recognition model successfully learned the expression of fear after comparing the saliency map with the AUs units in Table 4.

Anger

Table 3
Operational definitions for learning emotion identification.

Learning emotion	Operational definition	
Frustration	Facial expression Description	Frown and mouth closed. Feeling unsure, discouraged, irritated, stressed and troubled in learning process.
Confused	Facial expression Description	Frown and mouth movements are slightly open or closed. Feeling cannot understand or confused in learning process.
Boredom	Facial expression Description	The upper eyelids are falling and appear to be listless, distracted or the mouth is unnatural. Feeling boring tired and distracted in learning process.
Delightful	Facial expression Description	Smile or laugh very happy. Feeling satisfied in learning process.
Surprised	Facial expression Description	Eyes wide open or mouth open and eyebrows rise. Feeling surprised or unexpected in learning process.
Flow	Facial expression Description	Eyes focused on watching the monitor. None.

Table 4
Correspondence of the basic emotion saliency map and AUs for fear.

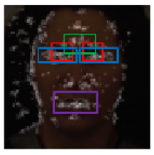
Saliency map	Facial expression	Description of AU	Corresponding feature marking area
		AU1 Raised inner brow	
		AU4 Frown	
		AU5 Raised upper eyelid	
		AU25 Slightly open lips	

Table 5
Correspondence of the basic emotion saliency map and AUs for anger.

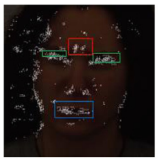
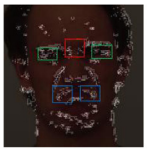
Saliency map	Facial expression	Description of AU	Corresponding feature marking area
		AU4 Frown	
		AU5 Raised upper eyelid	
		AU25 Slightly opened lips	

Table 6
Correspondence of the basic emotion saliency map and AUs for disgust.

Saliency map	Facial expression	Description of AU	Corresponding feature marking area
		AU4 Frown	
		AU7 Narrowed upper eyelid	
		AU15 Turned down mouth corners	

The experimental results are presented in Table 5. In the AU coding system, anger is mainly composed of AU4, AU5, and AU25. AU4 indicates a frown, corresponding to the red box in the saliency map; AU5 indicates raised upper eyelids, corresponding to the green boxes; and AU25 indicates slightly opened lips, corresponding to the blue boxes. This study verified that the basic emotion recognition model successfully learned the expression of anger after comparing the saliency map with the AUs in Table 5.

Disgust

The experimental results are listed in Table 6. In the AU coding system, disgust is primarily composed of AU4, AU7, and AU15. AU4 denotes a frown, corresponding to the red box in the saliency map; AU7 denotes narrowed upper eyelids, corresponding to the green boxes in the saliency map; and AU15 represents turned down mouth corners, corresponding to the blue boxes in the saliency map. This study verified that the basic emotion recognition model successfully learned the expression of disgust after comparing the saliency map with the AUs in Table 6.

Happiness

The experimental results are shown in Table 7. In the AU coding system, happy is mainly composed of AU7 and AU12. AU7 represents narrowed upper eyelids, corresponding to the red boxes in the saliency map and AU12 represents raised mouth corners, corresponding to the green boxes in the saliency map. This study verified that the basic emotion recognition model successfully learned the expression of happiness after comparing the saliency map with the AUs in Table 7.

Table 7
Correspondence of the basic emotion saliency map and AUs for happiness.

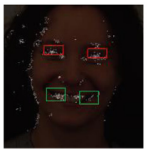
Saliency map	Facial expression	Description of AU	Corresponding feature marking area
		AU7 Narrowed upper eyelid	
		AU12 Raised mouth corners	

Table 8
Correspondence of the basic emotion saliency map and AUs for sadness.

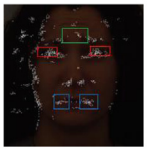
Saliency map	Facial expression	Description of AU	Corresponding feature marking area
		AU1 Raised inner brow	
		AU4 Frown	
		AU15 Turned down mouth corners	

Table 9
Correspondence of the basic emotion saliency map and AUs for surprise.

Saliency map	Facial expression	Description of AU	Corresponding feature marking area
		AU5 Raised Upper Eyelid	
		AU27 Mouth Opening	

Sadness

The experimental results are presented in Table 8. In the AU coding system, sadness is mainly composed of AU1, AU4, and AU15. AU1 signifies raised inner brows, corresponding to the red boxes in the saliency map; AU4 signifies a frown, corresponding to the green box in the saliency map; and AU15 signifies turned down mouth corners, corresponding to the blue boxes in the saliency map. This study verified that the basic emotion recognition model successfully learned the expression of sadness after comparing the saliency map with the AUs in Table 8.

Surprise

The experimental results are shown in Table 9. In the AU coding system, surprise is primarily composed of AU5 and AU27. AU5 represents raised upper eyelids, corresponding to the red boxes in the saliency map and AU27 represents an open mouth, corresponding to the green box in the saliency map. This study verified that the basic emotion recognition model successfully learned the expression of surprise after comparing the saliency map with the AUs in Table 9.

Table 10
Comparison of CNN architectures.

CNN architectures	Number of parameters	Number of layers
VGG16	138 M	23
InceptionV3	138 M	159
Inception-ResNet V2	2.5 M	572
DenseNet201	5.5 M	201
Dense_FaceLiveNet	15.3 M	129

Table 11
Accuracy comparison of the JAFFE and KDEF basic emotion models.

CNN Model	JAFFE	KDEF
VGG16	84.66%±5.70%	86.75%±2.87%
InceptionV3	82.59%±3.32%	90.25%±5.01%
Inception-ResNet V2	89.57%±6.33%	94.70%±1.19%
DenseNet201	90.23%±2.73%	92.52%±1.01%
Dense_FaceLiveNet	90.97%±3.95%	95.89%±0.76%

4.3. Analysis of the effectiveness of transfer learning for emotion recognition

4.3.1. Basic emotion model recognition results and analysis

This section is divided into two parts: the first part describes the training methods and results of the basic emotion models established using the JAFFE and KDEF basic emotion datasets, and the second part describes the training methods and results of the basic emotion model established using the FER2013 basic emotion dataset.

JAFFE and KDEF basic emotion recognition models

The JAFFE and KDEF datasets were employed as training materials to establish basic emotion recognition models. Dense_FaceLiveNet was verified to achieve satisfactory recognition accuracy for these small datasets compared with the other four CNN models. CNN architectures have proven credible choices for image recognition when tested with ImageNet's recognition challenge. Network architectures that have won the ImageNet image recognition competition are constantly changing. For example, InceptionV3 has a wide-but-shallow network architecture, DenseNet has a deep-but-narrow network architecture, and Inception-ResNet V2 has a deep-and-wide network architecture. This study verified the effectiveness of the proposed network architecture, Dense_FaceLiveNet, by comparing it with four other CNN architectures (Table 10).

Different methods were used according to the characteristics of the datasets when constructing the basic emotion recognition models. Because the JAFFE and KDEF data were small and simple, to avoid overfitting after training, k-folds cross-validation [36] was adopted as the training method to train the JAFFE and KDEF basic emotion recognition models.

The entire dataset was divided into five samples through 5-fold cross-validation. The training process was to select one sample each time as the data for model verification while as the remaining four samples used for training. This process was repeated five times before each training outcome was converged (Table 11).

According to Table 11, KDEF exhibited higher prediction accuracy than JAFFE in all network architectures. Test accuracy rates of the five CNN models that used JAFFE and KDEF datasets were tested with the Friedman test, and the results are shown in Table 12. From the verification results, the accuracy of the KDEF dataset was significantly different from that of the JAFFE dataset. The main difference between KDEF and JAFFE is data size; KDEF has 4900 images, whereas JAFFE has only 213 images. Therefore, this result verified that data size has a significant effect on the accuracy enhancement of a deep learning model.

FER2013 basic emotion recognition model

Table 12
Friedman test results of the JAFFE and KDEF basic emotion models.

Subject	Value
Number	5
Chi-square	5.000
DOF	1
P-value	.025

P-value $<$ 0.05: significant.

Table 13
Accuracy comparison of the FER2013 basic emotion recognition models.

Model	Accuracy
VGG [12]	65.80%
InceptionV3 [14]	68.86%
Mollahosseini [13]	66.40%
Inception_ResNetV2 [29]	69.72%
DenseNet201 [9]	68.52%
FaceLiveNet [10]	68.60%
Dense_FaceLiveNet	69.99%

The FER2013 basic emotion dataset was used as the training materials for a FER2013 basic emotion recognition model. The differences of the FER2013 dataset with the JAFFE and KDEF datasets are that the images in the FER2013 dataset are more complex. It contains people of different races, shot at different angles, and the same image appearing in different categories. With such complex data, the 28,432 images provided by the raw dataset were insufficient for the model to learn enough features. For the aforementioned JAFFE and KDEF basic emotion recognition models, increasing the data pool is helpful to improve the quality of the deep learning models. Therefore, a data augmentation method was adopted to increase the number of original training samples, thereby improving the model quality. Numerous methods can be used to augment data, such as random scaling, translation, vertical flipping, and horizontal flipping of the model. Through these methods, the amount of raw data can be greatly augmented. [37] employed translation, horizontal flipping, random scaling, and rotation methods, discovering that the accuracy of their model was increased by 10% compared with the initial accuracy. The present study referred to the parameters of data augmentation used with the data randomly scaled by 1–1.5 times, randomly rotated from $-5°$ to $5°$, and flipped horizontally. The augmented data are shown in Fig. 12.

The training datasets for the FER2013 basic emotion recognition model were divided into three sets: a training set, verification set, and test set. The training set was employed for learning, the verification set was used to verify each parameter value of the verification model, and finally the test set was used to obtain the model's accuracy. These data did not require splitting because they were arranged in the aforementioned three sets. Compared with other network architectures and the original architecture. FaceLiveNet and Dense_FaceLiveNet exhibited excellent recognition accuracy (Table 12). Although the accuracy difference was only 0.27% higher than Inception_ResNet V2. Dense_FaceLiveNet used only an eleventh of the parameters of Inception_ResNet V2 (Table 10). This proved the effectiveness of choosing Dense_FaceLiveNet for facial emotion recognition.

According to the results in Table 10, we can discover the accuracy of recognition on Dense_FaceLiveNet is higher than other models. Although the recognition accuracy only higher 0.27% than Inception-ResNet V2, but in Table 7, we can discover the amount of parameter in Dense_FaceLiveNet is lower 1/11 than Inception-ResNet V2.

Above contents, prove the effective on Dense_FaceLiveNet recognition facial emotion. (See Table 13.)

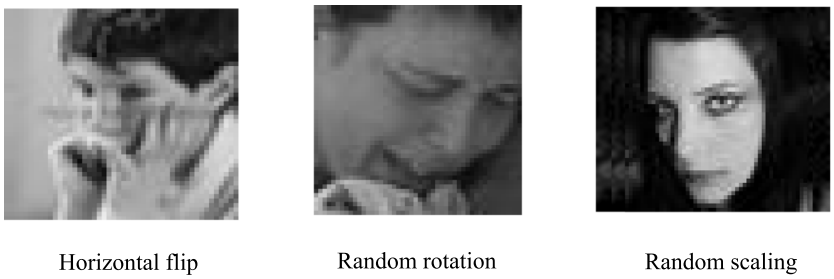


Fig. 12. Examples of data augmentation (images were selected from FER2013).

Table 14
Accuracy comparison of the basic emotion recognition models established through transfer learning.

Model	FER2013	JAFFE→FER2013	KDEF→FER2013
VGG16	65.80%	65.82%	66.20%
InceptionV3	68.86%	68.87%	69.12%
Inception-ResNet V2	69.11%	69.02%	69.15%
DenseNet201	68.82%	68.84%	68.91%
Dense_FaceLiveNet	69.99%	69.92%	70.02%

Table 15
Friedman test results of the JAFFE basic emotion recognition model undergoing transfer learning.

Subject	Value
Number	5
Chi-square	2.000
DOF	1
P-value	.655

P-value $<$ 0.05: significant.

Table 16
Friedman test results of the KDEF basic emotion recognition model undergoing transfer learning.

Subject	Value
Number	5
Chi-square	5.000
DOF	1
P-value	.025

P-value $<$ 0.05: significant.

4.3.2. Effectiveness analysis of basic emotion transfer learning

Influence of transfer learning on the quality of basic emotion models

The FER2013 basic emotion recognition model was established through transfer learning of the JAFFE and KDEF basic emotion recognition models constructed using the five network architectures (Table 14). Through Friedman tests, the JAFFE and KDEF basic emotion recognition models were tested for their respective effects on improving the quality of the FER2013 basic emotion recognition model. From Table 15, the FER2013 model established through transfer learning of the JAFFE model exhibited no significant difference in quality compared with direct use of the FER2013 dataset. However, use of the KDEF model significantly improved the FER2013 model (Table 16). Therefore, similar to the results in Section 4.3.1, the size of the dataset affects the quality of the model. The results of this section reveal that use of a larger and more complex model dataset can effectively improve the effects of transfer learning on model quality.

Dense_FaceLiveNet demonstrated excellent recognition accuracy compared with the other four architectures, with or without transfer learning (Table 14). The emotion recognition accuracy of each model established through weight initialization, transfer learning of JAFFE to FER2013, and transfer learning of KDEF to FER2013 was subsequently analyzed. The experimental results are shown in Fig. 13, revealing that the FER2013 basic emotion recognition model obtained from transfer learning using KDEF demonstrated higher ability than the other models to recognize anger and sadness, with the other emotion types being similarly recognized by all three models. Hence, use of a larger and more complex dataset could effectively improve the model's recognition ability.

Influence of transfer learning on training time of the basic emotion models

Training time and required resources are two of the main considerations for practical application of a model in industry. Therefore, in addition to accuracy, model establishment must consider overall training time. The training time from beginning to convergence for weight initialization using normalization, transfer learning of JAFFE to FER2013, and transfer learning of KDEF to FER2013 were compared (Fig. 14). The weight initialization model exhibited sizeable oscillations during the training process, and the total convergence time was more than 60 epochs. However, the JAFFE and KDEF models used only 17 and 15 epochs respectively. In addition, the final accuracy of the weight initialization model was lower than transfer learning of KDEF to FER2013 even after numerous iterations. Based on this result, this study inferred that transfer learning can improve the accuracy of a model and provide a noticeable improvement in the optimization of training convergence time.

4.3.3. Analysis of the effectiveness of learning emotion transfer learning

As shown in Section 4.3.2, transfer learning provided significant improvements to the accuracy of the basic emotion recognition model as well as training time optimization. Learning emotions is more complex because many expressions are subtle and difficult to recognize. Thus, the FER2013 basic emotion recognition model was employed for transfer learning with a learning emotion database, and the accuracy and difference between training times was compared for the learning emotion recognition models established with or without transfer learning. In this section, the effects of using transfer learning to establish learning emotion recognition models is examined with regard to two aspects: the influence of transfer learning on model quality improvement and the influence of transfer learning on model training time.

Influence of transfer learning on improving the quality of learning emotion models

Using the five network architectures introduced in Section 4.4.2, the KDEF basic emotion recognition model was employed to establish the FER2013 basic emotion recognition model through transfer learning. Subsequently, a learning emotion recognition model was established through transfer learning of the FER2013 basic emotion recognition model with a learning emotion database. Larger data size has a positive effect on the quality of the basic emotion recognition model after transfer

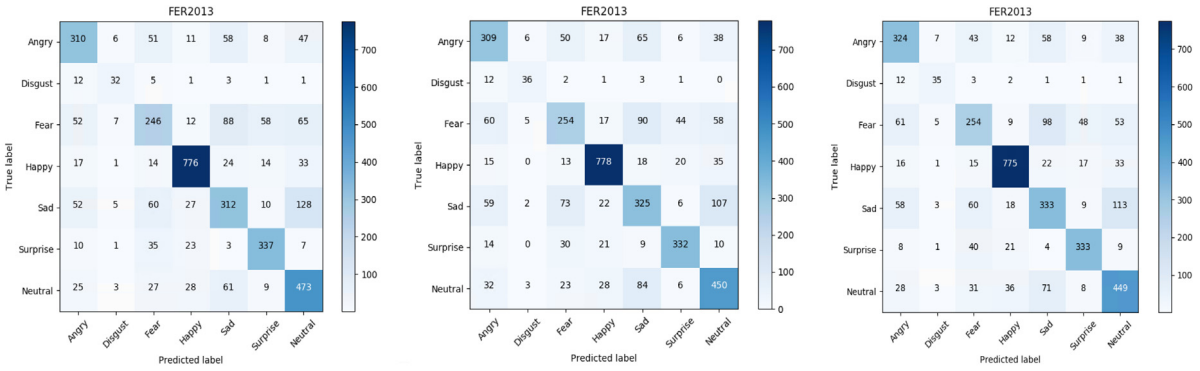


Fig. 13. Confusion matrices of basic emotion transfer learning for Dense_FaceLiveNet.

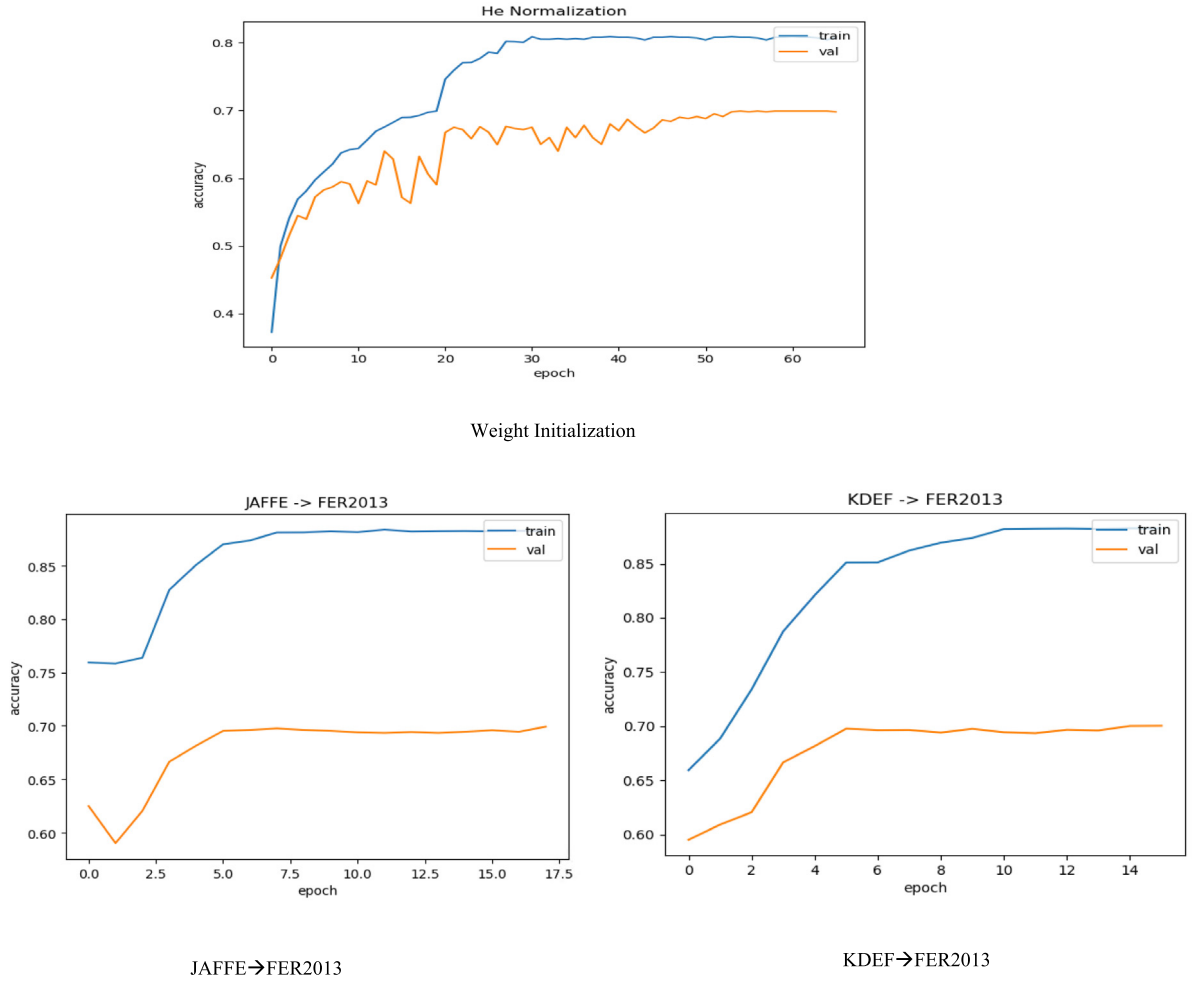
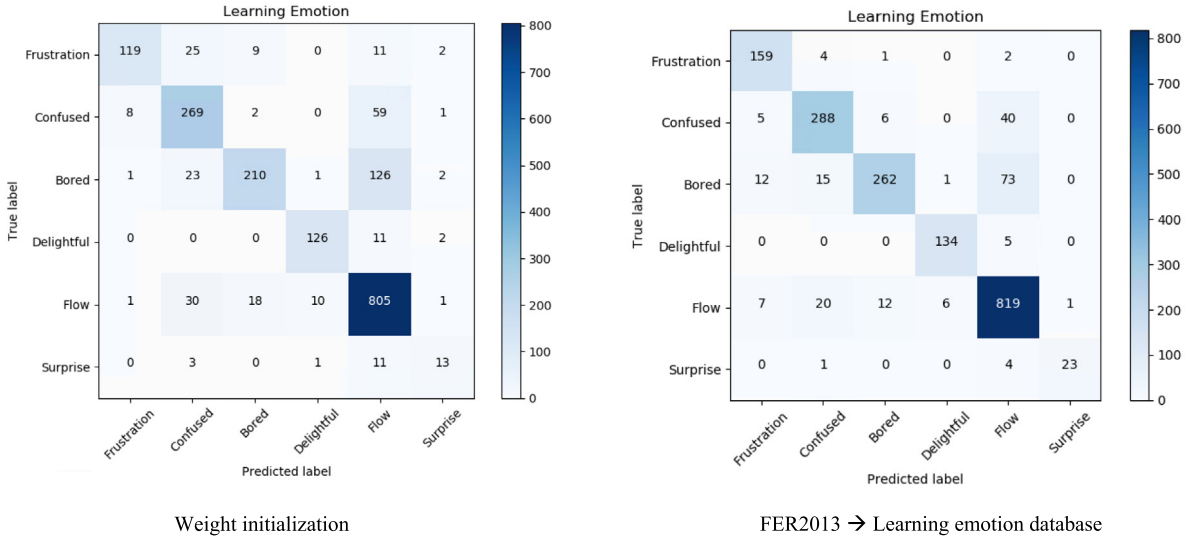


Fig. 14. Comparison of basic emotion transfer learning time for Dense_FaceLiveNet.

learning. This section examines whether transfer learning had a similar effect on the quality of the learning emotion recognition model.

The results of weight initialization using the normalization and the accuracy of transfer learning using the FER2013 basic emotion model are shown in Table 17. The accuracy of the five models with or without transfer learning was determined by performing Friedman test. The learning emotion recognition accuracy is presented to be significantly different among the five models in Table 18. Taking Dense_FaceLiveNet as an example, the accuracy of the weight initialization model was 79.03% and rose to 91.93% after transfer learning with the FER2013 basic emotion recognition model, showing a difference of 12.9%. The findings showed that the results for happiness and surprise were similar for both the learning and basic emotion models, thus there were fewer updates to the weights for happiness and surprise on the overall network when transfer learning was performed. Transfer learning focused only on the other expression types. From Fig. 15, transfer learning resulted in significant improvement in the prediction accuracy of each emotion type in the learning emotion recognition



Weight initialization

FER2013 → Learning emotion database

Fig. 15. Confusion matrices of learning emotion transfer learning for Dense_FaceLiveNet.

Table 17
Accuracy comparison of the learning emotion recognition models.

Models	Learning emotion database	FER2013 → Learning emotion database
VGG16	75.19%±3.09%	82.74%±2.24%
InceptionV3	77.84%±0.79%	86.64%±1.84%
Inception-ResNet V2	79.75%±3.99%	89.62%±1.66%
DenseNet201	77.54%±0.90%	86.32%±1.52%
Dense_FaceLiveNet	79.03%±1.86%	91.93%±1.33%

Table 18
Friedman test results of transfer learning for the learning emotion recognition models.

Subject	Value
Number	5
Chi-square	5.000
DOF	1
P-value	.025

P-value $<$ 0.05: significant.

model. Because the model already included facial edge features in the FER2013 basic emotion recognition model, more focus was placed on the learning of subtle features. Consequently, accuracy was greatly improved after transfer learning.

Influence of transfer learning on learning emotion model training time

As described in Section 4.4.1, model establishment must be efficient in addition to being highly accurate. Therefore, the training time of the models was compared from beginning to convergence for weight initialization using the normalization and for transfer learning with the FER2013 basic emotion recognition model. Training time is shown in Fig. 16. From the figure, the model with weight initialization using the normalization exhibited a rapid increase during the training process before reaching a bottleneck at 79%. For the FER2013 model with transfer learning, although a large oscillation occurred at the beginning, the increase of accuracy of the overall model in the later stages was relatively stable compared with the model using weight initialization, at only 18 epochs compared with 25 epochs in the weight initialization model. In addition, the FER2013 model had seven less iterations than the weight initialization model did. According to this result, using a basic emotion model for transfer learning in learning emotions can effectively improve accuracy and has considerable benefits in training convergence time optimization.

Table 19
Comparison of the learning emotion data in the video-based learning and interactive learning.

Label	Video-based learning	Interact learning
Frustration	7	21
Confused	82	103
Boredom	1477	574
Delightful	36	10
Flow	1570	469
Surprise	8	37
Total	3180	1214

4.4. Generalizability verification and correction of the learning emotion recognition model

To verify the feasibility of applying Dense_FaceLiveNet to a real environment, a generalizability test was performed on the established learning emotion recognition model.

Lin [38] and Liou [39] collected image databases for learning emotion transfer studies in the context of video-based learning and interactive learning. In the present study, four students from a first-year programming course at a national university in central Taiwan were randomly selected from these databases to participate in video-filming during the learning process, and images were taken from the videos every 3 s. First, the retrieved images were filtered to remove any blurred and severely cropped images, and manual calibration classification was performed based on the operational definitions to obtain learning emotion data from the video-based and interactive learning. The experimental data is presented in Table 19, revealing substantial differences in the total amount of data between the video-based and interactive learning. For the video-based learning, students only focused on watching the video without taking any other actions; whereas the interactive learning led to peer-to-peer communication and typing on the keyboard during the learning process, creating many unclear and indistinguishable facial images. Consequently, video-based learning obtained more effective images than interactive learning did.

This section presents the three phases of the experiments. Section 4.4.1 discusses prediction of the interactive learning emotion dataset through an emotion prediction model; Section 4.4.2 examines training performed on images with eyebrows covered, cheeks supported, and eyes closed; and Section 4.4.3 presents

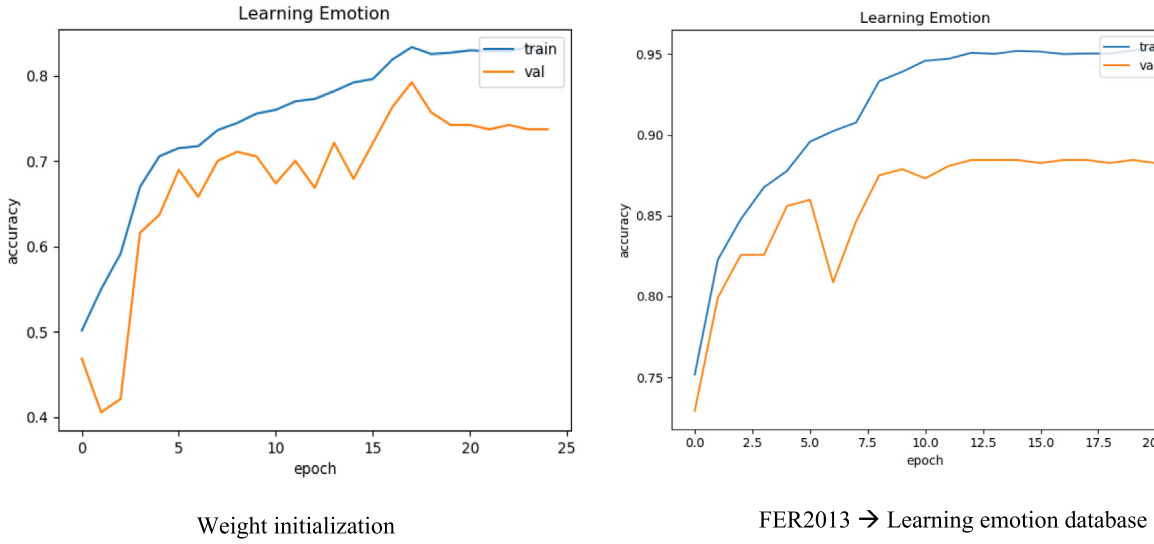


Fig. 16. Comparison of training time for learning emotion transfer learning for Dense_FaceLiveNet.

training using a combination of the original learning emotion database alongside the video-based and interactive learning emotion database.

4.4.1. Identifying interactive learning emotion datasets through the learning emotion recognition model

The learning emotion recognition model established with Dense_FaceLiveNet in Section 4.3.3 was used to predict the results of the video-based learning emotion dataset. The experimental results are displayed in Fig. 17. Many inputs were misclassified into bored, confused, and surprised, which resulted in an overall accuracy of only 43.73%. These results are because the video-based learning emotion dataset has many images in which the eyebrows are covered, cheeks are supported by the hand, and eyes are closed. These images were not found in the original emotion database. Initially, this study only focused on establishing facial expression data in a lab environment without considering exceptional situations that can occur in a real classroom environment. However, in response to these findings, a model was established for images with eyebrows covered, cheeks supported, and eyes closed.

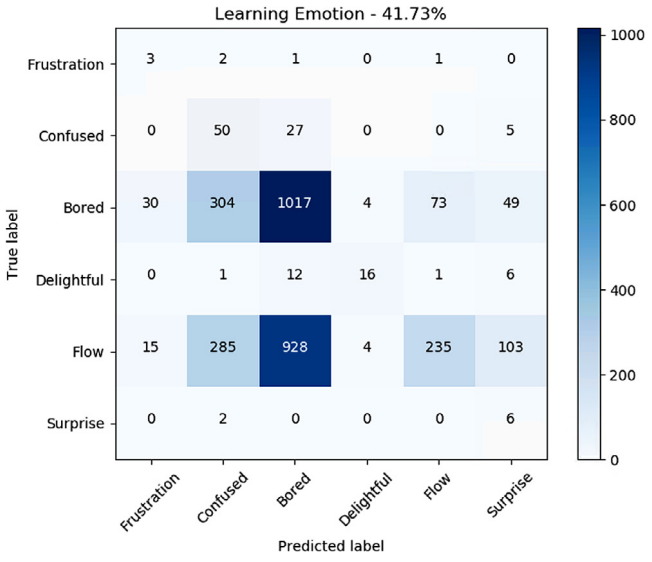


Fig. 17. Confusion matrix of video-based learning.

4.4.2. Transfer learning for a learning emotion recognition model for images with eyebrows covered, cheeks supported, and eyes closed

The results in Section 4.4.1 clearly show that the original learning emotion database did not account for images where the eyebrows covered, cheeks supported, and eyes closed, which resulted in poor model generalization. Therefore, these images were added to the original learning emotion database.

The learning emotion database was divided in a 70:10:20 ratio, representing the training, verification, and test sets, respectively. In addition, the transfer learning was performed using the FER2013 basic emotion recognition model. The experimental results are shown in Fig. 18, revealing a model accuracy of 92.42%. This shows that the learning emotion recognition model is still effective and can learn the corresponding features after the addition of images in which eyebrows are covered, cheeks are supported, and eyes are closed.

To verify whether the improved model could successfully capture the features of the exceptional images, this study randomly extracted eight images from the dataset for saliency map output. The output is shown in Fig. 19, revealing that the improved model has learned the corresponding features in situations where eyebrows are covered, cheeks are supported, and eyes are closed.

Because different students may use different hands to support their cheeks, four images were randomly selected for horizontal flipping and output to saliency maps. From the saliency map in Fig. 20, the horizontally flipped images still effectively captured the corresponding features which verified the improvement from using data augmentation on the training dataset.

The experimental results verified that a deep CNN can solve the problems inherent to conventional machine learning problems (i.e., inability to capture features other than facial expressions).

4.4.3. Enhancement of model generalizability by combining three sets of learning emotion datasets as training materials

From the results of Sections 4.4.1 and 4.4.2, the diversity and quantity of training data were correlated with the quality and generalizability of the prediction model. Therefore, increasing the amount and complexity of the data can effectively enhance the quality and generalizability of the model. The original learning

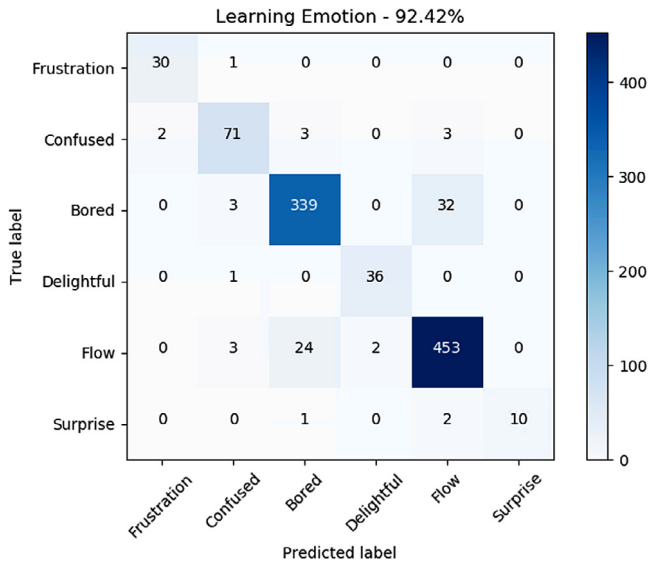


Fig. 18. Confusion matrix of the learning emotion database test set after adding the new images.

Fig. 19. Saliency map of covered eyebrows, supported cheeks, and closed eyes.

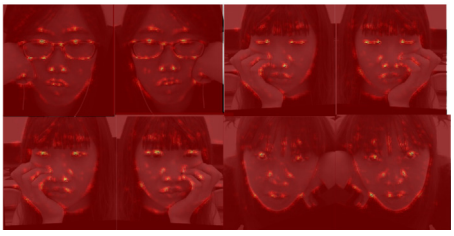


Fig. 20. Horizontally flipped image features were also effectively captured, verifying the results.

Table 20
Details of the combined learning emotion database.

Emotion	Data items
Frustrated	194
Confused	524
Bored	2414
Delighted	185
Flow	2904
Surprised	73
Total	6294

emotion database was combined with the video-based and interactive learning datasets to form a new learning emotion database. The experimental results are shown in Table 20.

The combined learning emotion database was divided into training, verification, and test sets in a 70:10:20 ratio, respectively, and transfer learning was performed using the FER2013 basic emotion database. The experimental results are presented

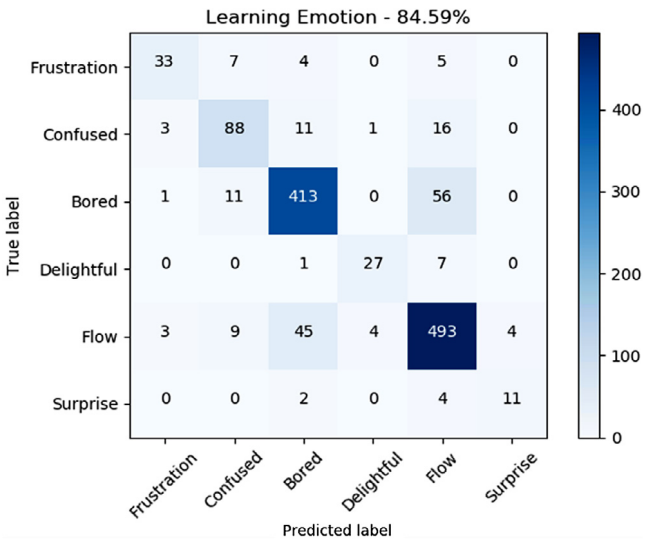


Fig. 21. Confusion matrix of combined learning emotion database test set results.

Table 21
Details of the small learning emotion database.

Emotion	Data items
Frustrated	6
Confused	15
Bored	95
Delighted	17
Flow	303
Surprised	1
Total	437

in Fig. 21. Based on the confusion matrix, "flow" and "bored" had the highest misclassification rates. The two main factors that most likely caused these errors were as follows. First, "bored" was determined by narrowing of the upper eyelids, and the proportion of participants with larger eyes was far greater than of participants with smaller eyes. Therefore, the model misjudged the expressions of the participants with smaller eyes as "bored". Second, the participants were only required to watch a computer screen during the learning process in the video-based learning environment. Therefore, the participants in the video-based learning were marked as bored whenever they looked downward. By contrast, the participants in the interactive learning environment had to type on the keyboard during the learning process. Consequently, these participants were marked as "flow" when they looked downward.

4.4.4. Small datasets testing

A 10-minute video produced by two male students in the laboratory was employed to perform image processing and categorization in order to obtain a learning emotion dataset (Table 21).

A 5-folds cross-validation was conducted on the small learning emotion database (Table 20), and transfer learning was performed using the FER2013 basic emotion database. The experimental results are presented in Table 22, showing that the learning emotion prediction model could effectively respond to small datasets, exhibiting recognition accuracy of 81.22% with only 437 pieces of data and a standard deviation of only 1.62%, which verified the stability of the model.

The experimental results are displayed in the previous description. There are four students randomly selected from the database for experiments to understand the applicability of the generalization ability of the learning emotion identification model.

Table 22
Test accuracy of the small learning emotion database.

Fold	The amount of dataset
1	82.02%
2	82.95%
3	81.60%
4	78.16%
5	81.39%
Average	81.22%±1.62%

The overall accuracy is only 43.73% because the video-based learning emotion dataset has many images in which the eyebrows are covered, cheeks are supported by the hand, and eyes are closed, which were not found in the original emotion database. This study only focused on establishing facial expression data in a lab environment without considering exceptional situations that can occur in a real classroom environment. Therefore, after adding the images which eyebrows are covered, cheeks are supported, and eyes are closed into original learning emotion database for re-training and re-building the learning emotion recognition model. The recognition accuracy rate can reach 92.42%. By the output of the saliency map, it can be found that the learning emotion recognition model has successfully learned the features of non-face emotions. This experimental results can verify that CNN can solve the problem that the original facial features which cannot be captured in traditional machine learning.

5. Conclusion

This study proposed a new model, namely Dense_FaceLiveNet, based on conventional CNN architecture to offer a solution for learning emotion recognition. First, dataset in JAFFE and KDEF was applied to develop fundamental emotion recognition model to identify the efficiency of Dense_FaceLiveNet. Results show a significantly high accuracy of 90.97% and 95.89% in JAFFE and KDEF. The results also indicate that there exists a positive correlation between number of training data and the quality of model based on Friedman test. Second, FER2013 model, with accuracy rate at 70.02%, was designed according to the results obtained from JAFFE and KDEF. Results indicate that the complexity and size of source domain may significantly impact the quality of target domain model based on Friedman test. Finally, the results also show that the learning emotion model powered by Dense_FaceLiveNet can retain critical action units from those basic learning emotions precisely. Over all, it demonstrates the effectiveness of proposed model in the recognition of learning emotion. Following the previous achievement [40], this study goes further to evaluate issue of compatibility through the 70/10/20 model. Our test result obtains an accuracy rate at 84.59% through the transfer learning of basic emotion model based on FER2013, and clearly demonstrates the importance of data complexity while deep learning model is designed.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.105724.

Acknowledgments

The work was supported by grant from the Ministry of Science and Technology of Taiwan (No. MOST-106-2221-E-025-015-). We would like to thank the Ministry of Science and Technology for funding this study.

References

[1] Paul Ekman, Richard J. Davidson, The Nature of Emotion: Fundamental Questions, 1994.
[2] Rafael A. Calvo, Sidney D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, IEEE Trans. Affect. Comput. (2010) 18–37.
[3] J. Wu, Facial Expression Analysis in E-Learning, 2014, Hualien.
[4] W. Chen, The Effects of Social Presence on Self-Regulatory Effcacy, Motivation and Achievement in E-Learning Environment using Senior Elementary Students as Example (Master's thesis), Information Management of National Yunlin University of Science and Technology, Yunlin, 2009.
[5] X. Su, The Image Processing of E-Learning Affective Detection on the Degress of Concentration (Thesis), Technology & Science Institute of Northern Taiwan, Taipei, 2007.
[6] M.W. Gardner, S.R. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ. (1992) 2627–2636.
[7] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Mach. Learn. (1995) 273–297.
[8] Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Statician (1992) 175–185.
[9] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, Densely Connected Convolutional Networks, 2016.
[10] Zuheng Ming, Joseph Chazalon, mluqma01, Muriel Visani, jcburie, FaceLiveNet: End-to-End Face Verification Networks Combining with Interactive Facial Expression-Based Liveness Detection, University of La Rochelle, Parise, France, 2018.
[11] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet classification with deep convolution neurl networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems. 2012, pp. 1097–1105.
[12] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014.
[13] Ali Mollahosseini, David Chan, Mohammad Mahoor, Going deeper in facial expression recognition using deep neural networks, in: IEEE Winter Conference on Applications of Computer Vision, 2016.
[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Rethinking the Inception Architecture for Computer Vision, 2015.
[15] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, Iain Matthews, The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 13–18.
[16] D. Lundqvist, A. Flykt, A. Öhman, The Karolinska Directed Emotional Faces - KDEF, 1998, http://www.emotionlab.se/resourses/kdef.
[17] M.J. Lyons, M. Kamachi, J. Gyoba, Japanese Female Facial Expressions (JAFFE), Database of Digital Images, 1997.
[18] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamnerand, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Challenges in Representation Learning: A Report on Three Machine Learning Contests, ICONIP 2013: Neural Information Processing, 2013, pp. 117–124.
[19] P.R. Dachapally, Facial Emotion Detection using Convolutional Neural Networks and Representational Autoencoder Units, 2017.
[20] Maxime Oquab, Leon Bottou, Ivan Laptev, Learning and transferring mid-level image representations using convolutional neural networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014.
[21] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. (2010) 1345–1359.
[22] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, CoRR 1312.6034 (2013).
[23] P. Ekman, W.V. Friesen, The Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, San Francisco, 1978.
[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, 2015.
[25] Prajit Ramachandran, Barret Zoph, Quoc V. Le, Searching for Activation Functions, 2017.
[26] Xavier Glorot, Antoine Bordes, Yoshua Bengio, Deep Sparse Rectifier Neural Networks, AISTATS, 2011.
[27] Sergey Ioffe, Christian Szegedy, Batch Normalization: Accelerating Deep Network Training By Reducing Internal Covariate Shift, Google, 2015.
[28] K.C. Lin, T.-C. Huang, J.C. Hung, N.Y. Yen, S.J. Chen, Facial emotion recognition towards affective computingbased learning, Libr. Hi Tech 31 (2013) 294–307.

[29] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 1600 Amphitheatre Pkwy, Google Inc, Mountain View, CA, 2016.
[30] Juwei Lu, Konstantinos N. Plataniotis, On conversion from color to gray-scale images for face detection, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009.
[31] Min Lin, Qiang Chen, Shuicheng Yan, Network in Network, 2013.
[32] R. Breuer, A Deep Learning Perspective on the Origin of Facial Expressions, 2017.
[33] Rupesh Kumar Srivastava, Klaus Greff, Jürgen. Schmidhuber, Highway Networks, 2015.
[34] Li Fei-Fei, Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Imagenet: A large-scale hierarchical image database, in: CVPR 2009, 2009, pp. 248–255.
[35] P.-R. Chung, Applying Facial Action Units and Feature Selection Methods to Develop the Learning Emotion Image Database and Recognition Model, National Chung Hsing University, Taichung, 2018.
[36] Kohavi Ron, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, 1995.
[37] David Eigen, Christian Puhrsch, Rob Fergus, Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, 2014.
[38] Y.-F. Lin, The Study on Developing a Learning Emotion Recognition and Emotion Transferring Model During Video-Based Programming Learning, National Chung Hsing University, Taichung, 2018.
[39] S.-T. Liu, The Study on Developing a Learning Emotion Recognition and Emotion Transferring Model During Interactive Programming Learning, National Chung Hsing University, Taichung, 2018.
[40] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. (2010) 433–459.

Jason C. Hung is an Associate Professor of Department of Computer Science and Information Engineering at National Taichung University of Science and Technology, Taiwan, ROC. His research interests include Multimedia System, e-Learning, Affective Computing, Artificial Intelligence and Social Computing. From 1999 to date, he was a part time faculty of the Computer Science and Information Engineering Department at Tamkang University. Dr. Hung received his BS and MS degrees in Computer Science and Information Engineering from Tamkang University, in 1996 and 1998, respectively. He also received his Ph.D. in Computer Science and Information Engineering from Tamkang University in 2001. Dr. Hung participated in many international academic activities, including the organization of many international conferences. He is the founder of International Conference on Frontier Computing. He served as Hon Treasurer of IET Taipei LN. In April of 2014, he was elected as Fellow of the Institution of Engineering and Technology (FIET). He was elected as vice chair of IET Taipei LN in Nov. 2014. From June 2015, He is Editor-in-Chief of International Journal of Cognitive Performance Support and serves as deputy editor of International Journal of Social and Humanistic Computing.

Kuan-Cheng Lin was born in Taiwan on September 13, 1964. He received a BS in chemistry from National Taiwan University in 1988 and a PhD in applied mathematics from the National Chung-Hsing University in 2000. From 2000 to 2006, he was an assistant professor with the department of information management at the Northern Taiwan Institute of Science and Technology, Taipei, Taiwan. From 2006 to 2008, he was an assistant professor with the department of management information systems at National Chung-Hsing University, Taichung, Taiwan. From 2008 to 2015, he was an associate professor with the department of management information systems at National Chung-Hsing University. Since 2015, he has been a professor with the department of management information systems at National Chung-Hsing University, Taichung, Taiwan. His current research interests include affective computing, intelligent tutoring system and data mining.

Nian-Xiang Lai received the M.S. degree in department of management information systems at National Chung-Hsing University, Taichung Taiwan in 2018. His research interests include e-Learning and Artificial Intelligence.