# Face Emotion Expression Recognition Using DLIB Model and Convolutional Neural Network Approach for Supporting Online Learning

Rita Wiryasaputra[1,2], Chin-Yin Huang[1], Jimmy Juliansyah[2], and Chao-Tung Yang[3,4(✉)]

[1] Department of Industrial Engineering and Enterprise Information, Tunghai University, Taichung 407224, Taiwan
huangcy@go.thu.edu.tw
[2] Department of Informatics, Krida Wacana Christian University, Jakarta 11470, Indonesia
rita.wiryasaputra@ukrida.ac.id
[3] Department of Computer Science, Tunghai University, Taichung 407224, Taiwan
ctyang@thu.edu.tw
[4] Research Center for Smart Sustainable Circular Economy, Tunghai University, Taichung 407224, Taiwan

**Abstract.** Many sectors have experienced the impact of the COVID-19 outbreak, without exception education. The method of process learning transformed from face-to-face meeting learning became online learning. Learners tried to adapt to this unexpected circumstance. In the online learning approach, the instructors only assumed the degree of learners' understanding with their face emotion expressions spontaneously. Advancement technology enables the machine to learn data fast and accurately. Mostly, the position of the learner's face in front of the camera when attending the online course, and the DLIB's shape detector model map the landmark of the captured face. Deep learning is a subset domain of machine learning. Convolutional Neural Network (CNN) model as a deep learning approach has characteristics in the high computation and ease of implementation. The work proposed a face-emotion expression recognition model for supporting online learning. The combination ratio images dataset was 80% data training and 20% data testing, and the condition expression was determined with a deep learning approach. The experimental results showed that the recognition accuracy of the proposed model achieved 97% for dataset image input.

**Keywords:** Convolutional Neural Network (CNN) · DLIB · Deep Learning · Face Emotion Expression Recognition · Machine Learning

## 1 Introduction

The COVID-19 outbreak changes many sectors of human life. The changes transform the learner and white-collar workers to have experience in the new


D.-J. Deng et al. (Eds.): SGIoT 2022, LNICST 497, pp. 144–149, 2023.
https://doi.org/10.1007/978-3-031-31275-5_15

unlimited world. The method of process learning and meeting transformed from physical face-to-face to online learning/meeting. Learners tried to adapt to this unexpected circumstance. In the online learning approach, the instructors only assumed the degree of learners' understanding with their face emotion expressions spontaneously. The instructor's engagement to be more active in the delivery of the course materials is important in synchronous learning. Early detection is needed to measure the involvement of learners' emotions and is the benchmark for the improvement of an online class. Human emotion reflects in their face. Basically, there are seven types of emotions, namely neutral, angry, happy, sadness, disgust, surprise, and fear [7]. The combination of emotions represents the person's understanding level [6]. In terms of face detection, the input mediums are video, images, and live video streaming [1]. Mostly, the position of learners' faces in front of the camera when attending the online course, and the DLIB's shape detector model map the landmark of the captured face into coordinates. The Dlib model provides 68 points of face form coordinates that landmark the vital of the face, eyes, nose, and mouth. DLIB enables identification of the faces from the front [2]. One of the classification models in the machine learning domain is the Convolutional Neural Network (CNN) which has advantages in speed and computation [1] [3]. Using segmented data, the CNN model can increase its accuracy value. The research contributes to building the model that recognizes a human facial emotion expression. The aim of the research is to investigate the performance of the Convolutional Neural Network model to support online learning. The paper is structured as follows: the first section reviews the research background, Sect. 2 describes the previous research relevant to this research, Sect. 3 presents the research methodology, and the experiment and conclusions are outlined in Sects. 4 and Sect. 5, respectively.

## 2 Related Works

This section describes the literature study stage where review and research article papers were retrieved from Google Scholar, Science Direct, and IEEEXplore repositories. Jadhav revealed the comparison between algorithms for the complex issue in automatic face detection. When comparing the four detection algorithms: Cascade classifier, DLIB CNN, DLIB Histogram of Oriented Gradients (HOG), and Multitask Convolutional Neural Network (MTCNN), the prediction of the Cascade Classifier algorithm was not accurate even though the algorithm supported the real-time detection process that ran on CPU. For easy implementation, DLIB CNN with different parameters such as from the front, with low light, in multiple and side faces robust with various face occlusions, however, the model ran slow on real-time images with CPU [1].

Kamal conducted the research with plant leaves images using background subtraction, segmentation, and Convolutional Neural Network. The foreground image was removed by the background subtraction function, and the segmentation was used to have the region of interest. Took several classes of datasets such as two, four and eight with the result where the fine-tuned DenseNet121

accuracy reached 98.7%, the Inception V3 accuracy 96.7%, and the DenseNet121 accuracy achieved 93.57% [3].

Mukhopadhyay state four continuous human emotions in concert with a learning session. The complexity of human emotions and the psychological states were not enough to reflect the learner through the basic emotion. During a period of time, the combination of two or more emotions enables one to capture the facial emotion. Using the CNN model, the accuracy of emotion classification 65% and 62% for identification of mind state [6].

In terms of the prevention of accidents and fatalities, Mohanty conducted the detection of drowsiness in drivers using the DLIB model. The input of the model was video and the ratios of eyes and mouth reflected as the drowsiness detection. DLIB library is used to localize the facial landmarks where the histogram form represented the frequencies of gradient direction. The maximum accuracy of recognition reached 96.71% [5].

## 3 Methodology

This section presents the overall research stages, from retrieving the image collection, and step face landmarking, to the evaluation of the model and the image classification. The images from dataset were carried out for the labeling process and the selection data were segmented with a DLIB face detector model which maps the coordinates on the face. The main purpose of data segmentation is easy classification. As a result of data segmentation, so the size of the image was reduced to $224 \times 224$ pixels. Then CNN model trained the data.

### 3.1 Face Emotion Expression

Detecting a person's emotions enable someone to understand other people's feeling. Facial expressions and body gestures reflect human emotions. There are seven kinds of human emotions, namely neutral, angry, sadness, happy, surprise, disgust, and fear [7].

### 3.2 DLIB

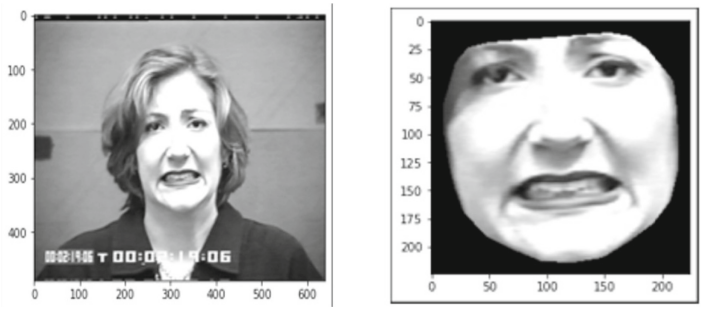
DLIB is an open-source library for shape detection. It maps in the 68 points coordinates of the facial landmarks. The DLIB's points are described as follows: point number 1 to 17 are facial shapes, point number 18 to 22 are used to show the left eyebrows, point 23 to 27 are right eyebrows pattern, point 28 to 22, and point 22 to 36 are the nose shape, point 37 to 42 are the left eye shape, while point 43 to the 48 are the right eye shape, and the mouth form use point 49 until point 68. In terms of face shape detection performance, the DLIB function used the Histogram of Oriented Gradients (HOG) [5].

### 3.3 Convolutional Neural Network

Convolutional Neural Network (CNN) is one of the deep learning models with its characteristics of ease of implementation and high computation [1]. The impact of the CNN model is experienced on Computer vision tasks, however the performance of the model decrease in the varied dataset images [3].

## 4 Experiment

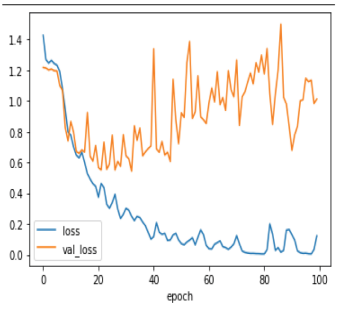
The experiment section explains the implementation of the proposed model. The experiment used a total of 901 images in CK+ dataset, which were divided into seven facial expressions: anger, happiness, sadness, fear, disgust, neutral, and surprise. All images in the dataset experienced with labeling process. The DLIB model was trained to identify 68 facial landmarks, then data segmentation eliminated the background images. The images as the result of segmentation were reduced according to the size of the model. Figure 1 presents the image results of before and after data segmentation. The result of data segmentation was adapted and learned by the model. In the architecture, ReLU activation function was used on 3 convolution layers, 2 pooling layers, and 1 fully connected layer. The SoftMax activation and the Adam optimizer were used to produce the output layer which had 7 nodes. The model was evaluated with confusion matrices in the simplest form of a table with two rows and two columns, and represents the four possibility classification outcome: True Positive (TP), False Positive, True Negative (TN) and False Negative (FN) [4]. The performance of the model did not only evaluated in the confusion matrix form, but also measure the model in accuracy, precision, recall, and the F1-score. The model gained 97% accuracy in 100 epochs. All the experiments used ratio of 20% data testing and 30% data testing consecutively. The first architecture model experiments used max-pooling $5 \times 5$, and dropout twice after dense. Generally, dense was used to decrease the unnecessary data so the model had an unfitting state. However, in these experiments, both models' results got overfitted and are shown in



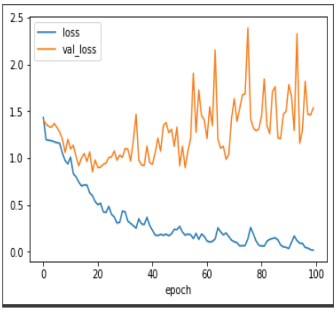

(a) Before (b) After

**Fig. 1.** Data Segmentation

Figure 2. The overall model architecture as follow: Input $224 \times 224 \times 1$; Convolution2D 64, $5 \times 5$; Max Pooling $5 \times 5$; Convolution2D 64, $3 \times 3$; Convolution2D 64, $3 \times 3$; Average Pooling $3 \times 3$; Convolution2D 128, $3 \times 3$; Convolution2D 128, $3 \times 3$; Average Pooling $3 \times 3$; Dense 1024; Dropout 0.2; Dense 1024; Dropout 0.2; Dense 7. The next experiments still used max-pooling $5 \times 5$ in 20% data testing or 30% data testing. Changed the position of dropout and only used it before dense. However, both models' results still experienced overfitting. The detail of model architecture was Input $224 \times 224$x1; Convolution2D 64, $5 \times 5$; Max Pooling $5 \times 5$; Convolution2D 64, $3 \times 3$; Convolution2D 64, $3 \times 3$; Average Pooling $3 \times 3$; Convolution2D 128, $3 \times 3$; Convolution2D 128, $3 \times 3$; Average Pooling $3 \times 3$; Dropout 0.2; Dense 1024; Dense 1024; Dense 7. Other experiments shifted the max-pooling with the average pooling $5 \times 5$ in 20% data testing or 30% data testing, only using dropout once and the position before dense. The pooling layer reduces the input images' spatial size and the number of computations in networks progressively. The purpose of simplifying the architecture was to get better results than previous experiments. The detail of model architecture was Input $224 \times 224$x1; Convolution2D 64, $5 \times 5$; Average Pooling $5 \times 5$; Convo-

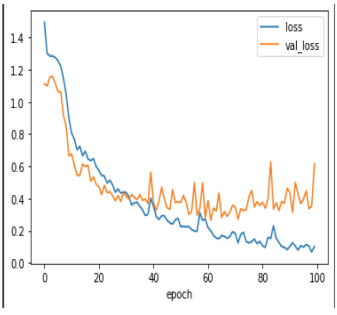



(a) Loss 20% data testing

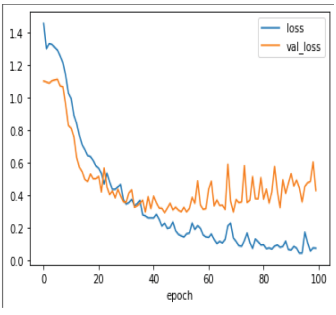



(b) Loss 30% data testing

**Fig. 2.** Comparison Accuracy of Architecture1




(a) Loss 20% data testing




(b) Loss 30% data testing

**Fig. 3.** Comparison Accuracy of Architecture3

lution2D 64, $3 \times 3$; Convolution2D 64, $3 \times 3$; Average Pooling $3 \times 3$; Dropout 0.2; Dense 1024; Dense 7. The graphs are better than those of the last experiments significantly, as illustrated in Fig. 3

## 5 Conclusion

Cutting-edge technology brings a new horizon to the education domain. For supporting the online learning process, the recognition of facial emotion expressions learners can be used as the evaluation for instructors when delivering the course material. With a distribution of 80% data training and 20% data testing, the accuracy of the proposed model to recognize achieved 97%. Although the achievement of the performance model is high, however, increasing the number of data images with high resolution (minimum: $640 \times 490$ pixels) in the dataset and equalizing the amount of data for each class will make the performance better.

**Acknowledgement.** This research was supported in part by the National Science and Technology Council (NSTC), Taiwan R.O.C. grants numbers 111-2622-E-029-003, 111-2811-E-029-001, 111-2621-M-029-004, and 110-2221-E-029-020-MY3.

## References

1. Survey on face detection algorithms. Int. J. Innov. Sci. Res. Technol. **6** (2021)
2. Bezerra, G.A., Gomes, R.B.: Recognition of occluded and lateral faces using MTCNN, DLIB and homographies, 11 (2018)
3. Kamal, K.C., et al.: Impacts of background removal on convolutional neural networks for plant disease classification in-situ (2021)
4. Krstinić, D., Braović, M., Šerić, L., Božić-Štulić, D.: Multi-label classifier performance evaluation with confusion matrix, pp. 01–14. Academy and Industry Research Collaboration Center (AIRCC) (2020)
5. Mohanty, S., Hegde, S.V., Prasad, S., Manikandan, J.: Design of real-time drowsiness detection system using DLIB (2019)
6. Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P.K.D., Dasgupta, N., Choudhury, P.: Facial emotion detection to assess learner's state of mind in an online learning system. In: ACM International Conference Proceeding Series, pp. 107–115 (2020)
7. Tarnowski, P., Kołodziej, M., Majkowski, A., Rak, R.J.: Emotion recognition using facial expressions. Procedia Comput. Sci. **108**, 1175–1184 (2017)