Using Dlib and CNN to Recognize the Facial Expression of People

[1] *Yung-Hsu,Chen*（陳詠絮） [2] *Chiou-Shann Fuh*（傅楸善）

[1]Department of Mechanical Engineering,
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan,
*E-mail: r11522502@ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

The project proposes a method to classify facial expressions by 68 facial landmarks. The facial landmarks will be detected by Dlib, and the facial expression will be classified by CNN model. We use Extended Cohn-Kanade dataset as training and testing dataset. In the dataset, there are seven emotions to be detected. The method advantage of our method is can quickly extract the facial features by pre-trained model in Dlib, which we don't need to spend lots of training time to extract the facial landmarks. To increase the accuracy of CNN model, we do data augmentation. After doing data augmentation, the testing accuracy is 69%. The research will show detail about CNN model and the method of data augmentation. We also discuss the performance of model by confusion matrix and PCA. By our research, we hope we can do some distribution about using facial landmarks to detect emotion in images.

Keywords: *Dlib, facial landmarks, CNN, emotion detection , CK+ dataset*

1. INTRODUCTION

In recent year, face detection has been a popular topic in academic research and daily life [1]. Face detection can be used in lots of applications, such as, mobile device, surveillance camera, and computer vision technique. Through face detection technique, the facial expression can be recognized automatically by a variety of neural network [2, 3].

Therefore, in the paper, we propose a method which detect the facial landmarks and use facial landmarks to tell the facial expression. When classifying the facial expression, accuracy and efficiency are important things to be consider. In nowadays, CNN is widely be used in detect human face [4], the advantage of it is can detect subtle details and get high accuracy in a variety of environments [5, 6]. However, the training time usually be long [7] and need expensive computational environment. Therefore, in the paper, when extracting facial feature, we replace Dlib with training CNN by ourself to reduce the training time. Dlib is an open-source library which has some pre-trained models to extract facial landmarks [8]. In recent year, there are some researches using Dlib to recognize people [9, 10]. With Dlib, eyebrows, eyes, nose, mouth and jaw can be presented by 68 landmarks quickly. There is also research using Dlib combined with CNN to do driver drowsiness detection [11].

Therefore, the objective of this paper is to develop a CNN model that can effectively learn and classify human emotions based on facial landmarks which detected by Dlib. We choose CK+ dataset as training and testing data [12]. In recent year, there are lots of research using CK+ dataset to train emotional classification neural network [13, 14].
There is also research about turning CK+ dataset to LBP image and training CNN to detect emotion [15]. In the research, we also modified CK+ dataset to train CNN model. Therefore, we use different method to modify CK+ dataset to make the dataset bigger. By training the model by our modified CK+ dataset, we aim to leverage the inherent hierarchical architecture of CNNs to automatically learn and extract relevant features from facial landmark positions. The testing accuracy in our research is 69%. Compare with the emotion recognition technique based on CNN in nowadays, the accuracy of our method is not high [16], but classifying emotions only depend on 68 facial landmarks without any pre-processing can shorten the training time of model. In addition, the training time of our model is short due to there are only several layers in it. In the end of our research we also use PCA to analysis the testing result [17], and give some ideals about improvement in the future.

The contributions of this paper are as follows: (1) an examination of Dlib's capabilities for facial feature detection, (2) the development of a CNN model that integrates Dlib's facial landmarks to classify human emotion, (3) experimental evaluation and validation of

the proposed method on CK+ dataset and modified CK+ dataset, and (4) Use confusion matrix and PCA to analysis the performance on our proposed method, and find the thing need to be improved in the future.

In nowadays, using images to detect emotion is a common thing in daily life. However, using whole images can cause some problems about personal privacy [18]. Using a part of picture as features can also reduce the influence of illumination [19]. Therefore, through this research, we aim to detect human emotion only depend on 68 facial landmarks by utilizing Dlib and CNN model in a safety way. By developing the integrated framework, we aspire to enhance the accuracy and efficiency of emotion recognition systems, opening new avenues for some applications, such as affective computing, human-robot interaction, and virtual reality.

2. METHOD

Our facial expression classifying algorithm is based on CNN in python language. Dlib, keras, numpy, pandas, matplotlib, PIL, scikit-learn, mpl_toolkits and csv modules are be used in the project. We download CK+ dataset and modified them as training and testing dataset. We collect the 68 facial landmarks by Dlib. During CNN training, the environment is AMD Ryzen 9 3950X 16-Core processor 3.70 GHz CPU. After training and testing, we use accuracy, confusion matrix and PCA to see and explain the performance of model. In the section, we will introduce the detail about several methods we use in the research.

2.1. CK+ dataset (Extended Cohn-Kanade dataset)

Extended Cohn-Kanade dataset (CK+) dataset is a widely used dataset in facial expression detection. There are 7 kinds of emotions in this it, including angry, happy, fear, sadness, disgust, surprise and contempt, as Figure 1 shows. To classify different expression, we label angry as 0, happy as 1, fear as 2, sadness as 3, disgust as 4, surprise as 5 and contempt as 6. Those labels will be used in training and testing our CNN model.

The pictures in the dataset are gray scale picture, and there are 981 annotation pictures in it. The size of picture is 48x48 pixels, and the person in the picture has different skin tone. By training the CNN model via the dataset, we try the detect the facial landmarks and classify the expression by neural network.

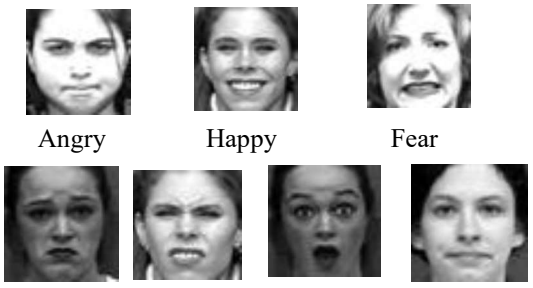

Fig. 1. Seven kinds of emotions in CK+ dataset.

The number of photos in each emotion are shown in the table below. Due to the number of each kind of emotion is a little imbalanced and the dataset is small, as Table 1 shows, we do data augmentation. We augment several kinds of photos, as Table 2 shown. We rotate the photo 5 to 10 degrees in both counterclockwise and clockwise direction to increase the diversity of dataset. By doing so, we hope we can enhance the robustness of our CNN model and avoid data imbalance problem.

Table 1. Original number of images in dataset.

Class of emotion	Number of images
Angry	135
Happy	207
Fear	75
Sadness	84
Disgust	177
Surprise	249
Contempt	54
Total	981

Table 2. Number of images after augmentation about rotating and horizontal flipping the photos.

Class of emotion	Number of photos
Angry	210
Happy	237
Fear	222
Sadness	208
Disgust	220
Surprise	249
Contempt	242
Total	1,588

After rotating the photo, we think shifting the photo in horizontal direction can also enhance the data diversity, because the human face is not bilateral symmetry. Furthermore, due to there a little difference of x coordinate in every photo in each emotion. Therefore, we decided to shift the photos in horizontal and vertical direction in 2 pixels. The number of images is shown in Table 3.

Table 3. Number of images after augmentation about rotating, shifting and horizontal flipping the photos.

Class of emotion	Number of photos
Angry	270
Happy	297
Fear	282
Sadness	262
Disgust	280
Surprise	309
Contempt	302
Total	2,008

2.2. Dlib

Dlib is a cross-platform C++ open-source library developed by Davis E. King for machine learning, which contains many common algorithms of machine learning and a large number of graphics model algorithms. There are many pre-trained models in Dlib, those models can used for face detection, facial landmark localization, object tracking. In the paper, we use a pre-trained model which is named "shape_predictor_68_face_landmarks.dat" to detect facial landmarks on the picture.
When detect human face, Dlib can output 68 facial landmarks in the picture. Figure 2 shows 68 landmarks in face.

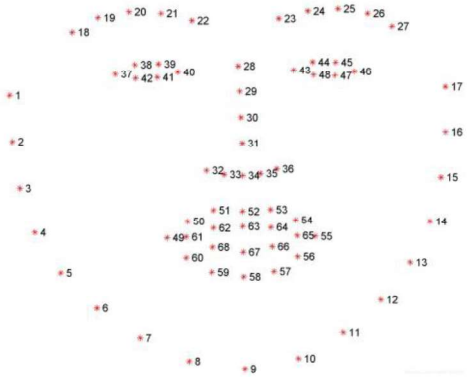

Fig. 2. 68 facial landmarks detected by Dlib.

2.3. CNN model

The reason of choosing CNN as model is because it is popular model in recent year. CNN can capture spatial locality features of input data. Therefore, we use CNN to classify the facial landmarks in each emotion.

After detecting the facial landmarks by Dlib, we save landmarks in the CSV file. We save facial landmarks as one-dimensional 136 vector. By doing so, we can use Con1D function in Keras module to do convolution. To classify several kinds of emotion, we train a CNN model. CNN is a popular and powerful class of supervised machine learning algorithms used for classification task. In the project, we design one CNN model, and compare the testing accuracy of different modified CK+ datasets.

Before doing data augmentation, we first split training data and testing data in 8:1 ratio. The epoch is set to 190, and the batch is set to 50. We use Adam optimizer, and the learning rate set to 0.0006. Because the size of our data is not large, we only use 2 convolutional layers, 2 dense layers and 2 max pooling layers to build the CNN model. In each layer, we use ReLU (Rectified Linear Unit) as activation function to avoid gradient descent and speed up the converge time. The loss function we used is categorical cross entropy, because there are seven kinds of emotion the model needs to classify. In the final layer, we set the unit of dense layer into 7, to classify 7 kinds of emotions in the dataset.

Figure 3 shows the layout of our CNN model. The input of our model is a one-dimensional vector which contain 68 landmarks in a face. Due to there are x coordinate and y coordinate in each landmark, there are 136 features in the one-dimensional input data. The size of kernels is set to be 5, and 9. And the strides of filter are set to be 3 to extract the feature in larger areas. With those kernels, we want to extract the useful features in face.
Between each convolutional layer, we use max pooling layers to decrease the size of features and only remain the signification features. By doing so, the max pooling layers can help CNN to extract important subtle features and avoid overfitting. After extracting features, we flatten the data and use dense layers to category the data. There are also dropout layers between flatten layer and dense layers to get rid of useless facial feature points. During training, dropout layer randomly sets some output values of neurons to zero to break the connections between them. This prevents individual neurons from relying too heavily on specific neurons, forcing the CNN network to learn more robust feature representations. In the dense layer, we use L2 regularization to reduce the probability of overfitting. In the final layer, the model will calculate the probability of each emotion in the photo and category the data into the predicted emotion. The parameters in each layer are shown in Table 4.

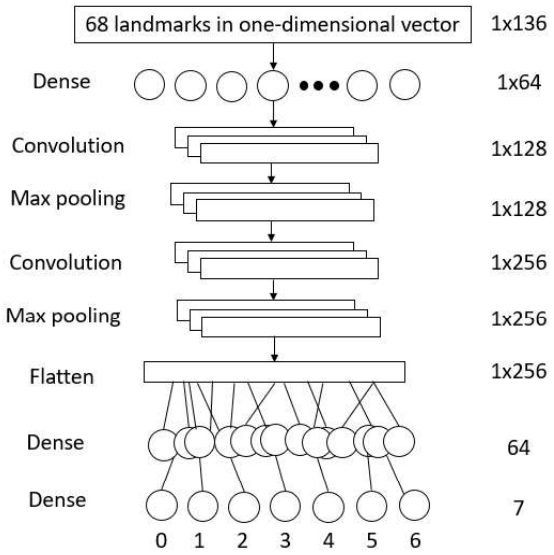

Fig. 3. CNN model.

Table 4. Details of each layer in the proposed network.

Layer Number	Layer Name	Parameters
1	Dense	64
2	Conv1D	Channel =128 Kernel = 5
3	MaxPooling1D	Size =2
4	Conv1D	Channel =256 Kernel =9
5	MaxPooling1D	Size =3
6	Flatten	
7	Dropout	Rate=0.5
8	Dense	Channel=64
9	Dropout	Rate=0.5
10	Dense	7

To see the performance of CNN model, we calculate the testing accuracy, confusion matrix and draw PCA figure. With the matrix, we can see which kind of emotion is easily to be detect or misjudge. By the confusion matrix, we can know which kind of emotion data need to be augmented more.

2.4. Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique commonly used in data analysis and machine learning. In our research, to see the distribution and relevant relationships in each emotion of testing result, we decrease the testing result into 3 dimensions and do PCA by scikit-learn module in python. And draw PCA stereogram by matplotlib module in python. By doing so, we try to understand and explain the model's ability about classifying different features in each emotion. Through PCA, we can also know which features of emotion are hard to tell, and can be improved in the future.

3. STEPS

Step 1. Download CK+ dataset.

Step 2. Do data augmentation to solve the data imbalanced problem in CK+ dataset.

Step 3. Use Dlib to detect facial landmarks in augmented CK+ dataset.

Step 4. Collect the detected facial landmarks and label the data by their emotions.

Step 5. Save the facial landmarks data and label data into CSV file.

Step 6. Spilt the data into training and testing dataset.

Step 7. Read training data from csv file and train a CNN network which can classify different emotion.
Step 8. Put testing dataset into CNN network and see the testing accuracy.

Step 9. Calculate the confusion matrix of the result to see the model performance.

Step 10. Draw PCA figure to see the distribution of features in each emotion.

4. RESULTS

After using Dlib, the face in the image can be detected automaticly, and be squared by green square as Figure 4 shown below. Due to there are too many pictures in the CK+ dataset, we choose one photo in angry emotion category as an example.

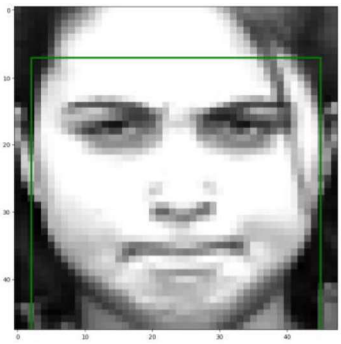

Fig. 4. The face in image detected by Dlib.

After detecting the face in image, the facial features can be labeled as 68 landmarks on picture, as Figure 5 shows. In Figure 5, the eyebrow, eyes, nose, mouth and jaw can be detected precisely. Each feature is mark by lots of dots and line in different color, to see the feature clearly. The position of those landmarks can be output precisely. By doing so, we can use those landmarks to do classify facial expressions.

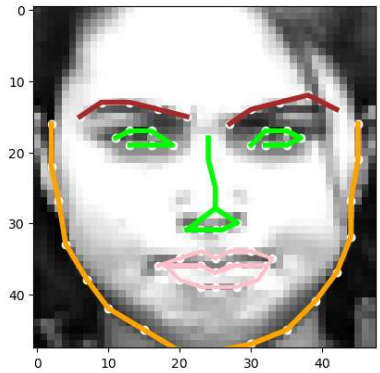

Fig. 5. The facial landmark extracted by Dlib.

To do data augmentation, we rotate 5 degrees in Figure 4 in counterclockwise direction, as Figure 5 shown. After rotating the picture, Dlib can still detect the landmarks precisely. By doing so, we want to augment the dataset and enhance the robustness of our CNN network.

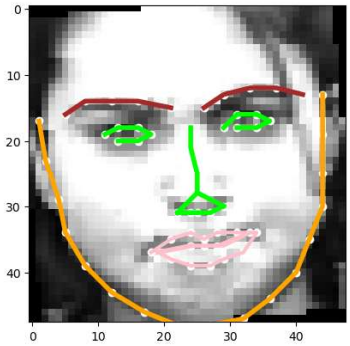

Fig. 6. The facial landmark in rotated image extracted by Dlib.

After collecting the facial landmarks, we train the CNN model and testing the performance of model. There are 196 images in testing dataset. The testing accuracy is 56%, the result is not satisfactory. The confusion matrix is shown as Figure 7. In this confusion matrix, the label in the left column represents the true label and the label in the bottom represent the predicted label. In Figure 7, there are each kind of emotion in testing dataset, however, due to the problem of data imbalanced, the result shows that lots of data are misclassified as label 4 (disgust). Label 3 (fear) and label 6 (contempt) almost cannot be detected due to there are few data of them. Among each kind of emotion, only the label 5 (surprise) is detected well, because the number of images in surprise category is larger than other a lot. From the confusion matrix, we know we need to solve the data imbalanced problem first.

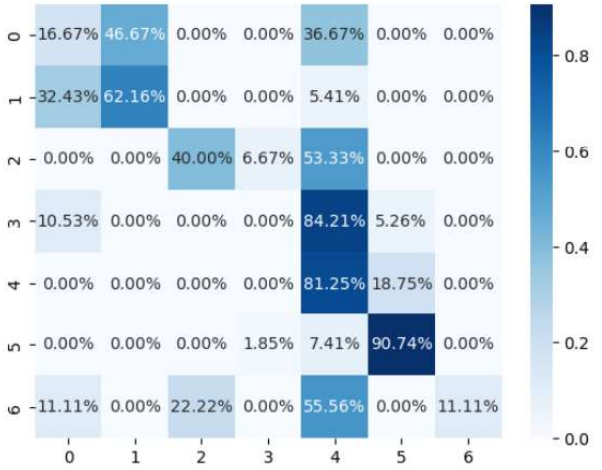

Fig. 7. Confusion matrix about detecting emotion before data augmentation.

Due to the unsatisfied result above, we try to do data augmentation. We increase the data by rotating images 5 to 10 degrees in counterclockwise and clockwise in CK+ dataset. We also flipping the images in horizontal direction. By doing so, the number of images in each emotion is similar than before. In this time, we also split training and testing data into 8:1 ratio, the number of images in testing dataset become to be 317. After increase the training dataset and train the CNN model again. The testing accuracy is 65%, which is better than before. In the confusion matrix, shown in Figure 8, we can see each emotion is detected better than before. With the augmented data, the accuracy of fear and contempt emotion increase a lot than before.

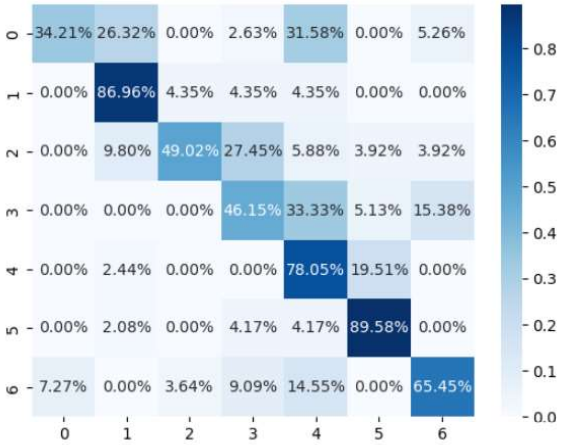

Fig. 8. Confusion matrix about detecting emotion after data augmentation.

There are several reasons about getting higher testing accuracy in CNN model than before. First, the number of training data is bigger than before, therefore the model can learn more information about each emotion, and predict better. Second, after doing data augmentation, the data imbalanced problem has been solved, which reduce the disparity of accuracy between

each emotion. Third, by rotating images in clockwise, counterclockwise and flipping in horizontal direction, model can learn more information, the diversity of information can improve the performance of model.

Furthermore, to enhance the robustness of the model, we keep doing data augmentation. We shift the photos in horizontal and vertical direction in 2 pixels. The number of images in testing dataset became 411. The testing accuracy is 69%, which increase 4% as before. The confusion matrix is shown in Figure 9. The result shows that shift the pixels can increase the testing accuracy sightly. From the confusion matrix, we can see the accuracy of some emotions have a significant disparity, such as angry emotion and surprise emotion, which shows that some features of emotion are hard to tell and there is also a room of improvement about classifying the emotion of our CNN model.

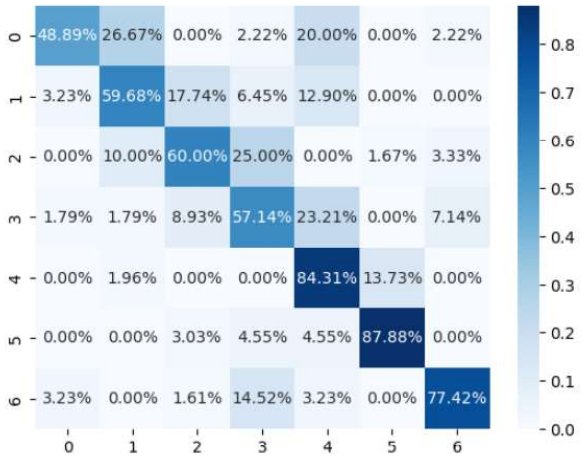

Fig. 9. Confusion matrix about detecting emotion after doing data augmentation again.

We compare the testing accuracy of different modified CK+ dataset in Figure10. In Figure 10, we can see the accuracy increases as the dataset is larger. The result shows than method of data augmentation can improve the accuracy of model.

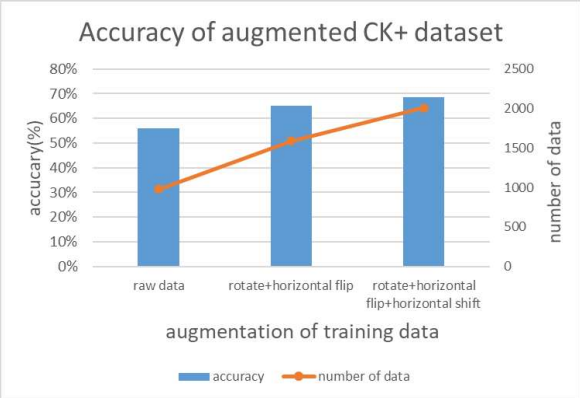

Fig. 10. Accuracy comparision of different augmented CK+ dataset.

However, compare with lots of emotion recognize techniques nowadays, the predict result of our method is not fine, there are several reasons we guess. First, facial expression is complicate, so there will be some loss information if it only uses 68 landmarks to present. Second, the x and y coordinates of pixels in each image are too similar, therefore, each kind of emotion is difficult to classify. If we do some data preprocessing, the result may be better. Third, we think our CNN model is not good enough to extract feature successfully. We should use some tricks to enhance the robustness our method.

Besides see the accuracy and confusion matrix of our testing result, we also want to see the distribution and relationships of different kinds of emotion. With PCA, we can explain and analyze the performance of testing result. Therefore, we reduce the testing result into 3 dimensions, and use module to make each kind of data into some points.

To observe the PCA result clearly, Figures 12 and 13 show the PCA result in two different angles of view, and each emotion is presented in different color. In the research, we only do PCA on the best result, which accuracy is 69%.

From Figure 12, we can see the points of surprise emotion cluster closely and separate far from other kinds of emotion, which indicate that the features of surprise emotion have large difference with other kinds of emotion. Therefore, the testing accuracy of it (87.88%) is also highest among other kinds of emotion. We think the accuracy of detecting surprise emotion is the highest is because of the mouth and eyes in surprise images are usually open big, as Figure 11 shows, which make lots of difference between other emotions. Therefore, the feature of surprise emotion can be identified easily in our neural network.

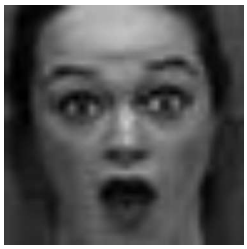

Fig. 11. Surprise emotion with big eyes and mouth in dataset.

Besides surprise emotion, from PCA figure and confusion matrix, we can see the features of contempt and sadness emotion have been detected well. Although the points of contempt and sadness emotion are close to other points of emotion, in Figures 12 and 13, the points of these two kinds of emotions cluster closely

individually. The testing accuracy are also better than other emotion.

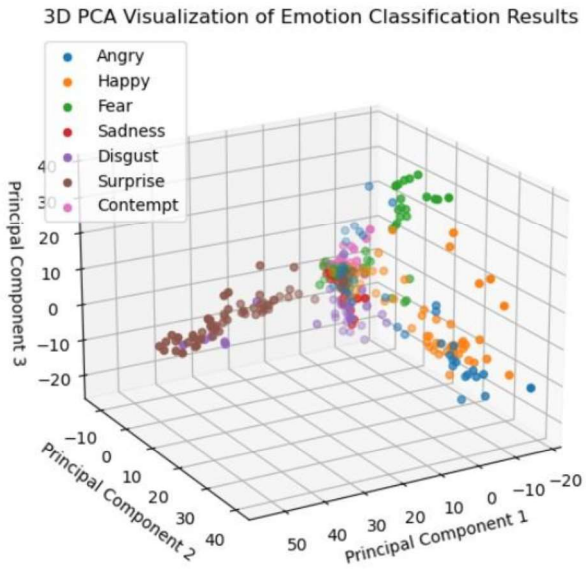

Fig. 12. PCA of testing result.

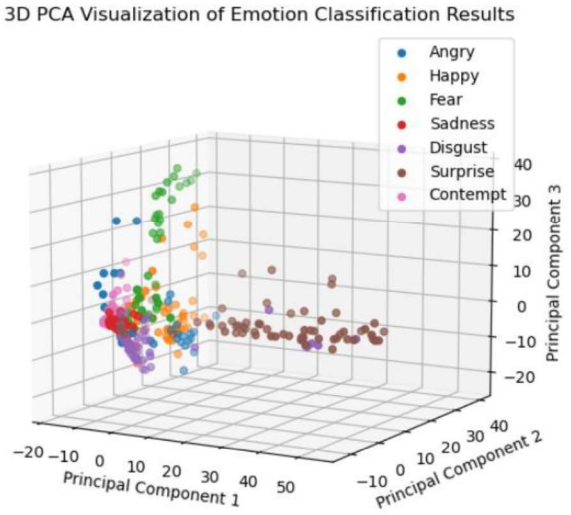

Fig. 13. PCA of testing result from another view point.

About Figures 12 and 13, we can see the points of angry are scatter in a large area, which indicate that the features of happy emotion are not consistent. Furthermore, from Figure 12, we can find out that there are overlapping area of the points of happy emotion and angry emotion. Therefore, the testing accuracy of happy emotion is also low. However, angry emotion and happy emotion are two opposite emotions. In the CK+ dataset, the mouth of these two kinds of emotions is close and open respectively in Figure 13.

In the future, we think we can do some data preprocessing or some tricks in CNN to extract the features better of angry emotion and happy emotion, especially in the landmarks of mouth. We think if we solve this problem, the testing accuracy can be enhanced a lot.

Angry Happy

Fig. 13. Angry and happy emotion in dataset which have closed and opened mouth respectively.

In addition, in confusion matrix in Figure 9, we found that some images of sadness emotion will be classified into disgust emotion. There is same result be found in PCA figure. In Figures 12 and 13, we can see several points of these two emotions are overlapping obviously. We think the reason is that the eyebrow and mouth are similar in these two emotions, as Figure 15 shown, so the CNN model cannot tell the difference between these two emotions well. How to improve the ability of telling the difference between these two kinds of emotions is also an important improvement goal in the future.

Sadness Disgust

Fig. 15. Sadness and disgust emotion in dataset which look similar.

Overall, use 68 facial landmarks as CNN input to detect emotion can reduce the parameters in model and make the calculation faster. Some kinds of emotion can also be detected well in our CNN model. In addition, using facial landmark can protect personal privacy, which only using a part of face instead of the whole face. By using a part of face, the impact of illumination of image can also be reduced, because the features are not too much. Therefore, we think using several facial landmarks to classify the emotion can be a potential and efficiency method in the future.

CONCLUSION

In this paper, we use CNN model to classify emotion by 68 facial landmarks extracted Dlib. The result shows that the data augmentation, such as rotating, shifting, flipping the images, in CK+ training dataset can rise the testing accuracy a lot. When using original data, the testing accuracy is 56% in our model. However, after doing data augmentation, such as rotation, picture shifting and picture flipping, the testing accuracy

become 69%. Among all kinds of emotions, the accuracy of recognize surprise and disgust emotions are the higher, which achieves 88% and 84% respectively. From PCA figure, we analysis the ability about classifying features of the model. We found that some features of emotion cluster successfully, however, the distribution of angry and happy emotion is overlapping, which lead to misclassify the emotion. The ability of classifying the features of these two emotions can be a room of improvement in the future. The advantage of our research is efficiency, which can only use 68 facial landmarks to recognize emotion instead of using the whole pixels in photo. The method of using facial landmark may be the trend in future. In the future, we think if the CNN model can be optimized, the model can classify the emotion more accurate, which make our method more useful.

REFERENCES

[1] Kumar, A., Kaur, A., and Kumar, M., "Face detection techniques: a review," *Artificial Intelligence Review*, 52, pp. 927-948, 2019.

[2] Ko BC., "A Brief Review of Facial Emotion Recognition Based on Visual Information," *Sensors*, Vol. 18, No. 2, pp. 401, 2018.

[3] S. Begaj, A. O. Topal and M. Ali, "Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN)," 2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications, Tirana, Albania, pp. 58-63, 2020.

[4] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua, "A convolutional neural network cascade for face detection," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 5325-5334, 2015.

[5] Minaee, Shervin & Luo, Ping & Lin, Zhe & Bowyer, Kevin. "Going Deeper Into Face Detection: A Survey," 2021.

[6] Fredj, Hana & Bouguezzi, Safa & Souani, Chokri, "Face recognition in unconstrained environment with CNN," *The Visual Computer*, 2021.

[7] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He Accurate, "Large Minibatch SGD: Training ImageNet in 1 Hour," 2018.

[8] King, Davis E., "Dlib-ml: A Machine Learning Toolkit," J. Mach. Learn. Res. 10: pp. 1755-1758, 2009.

[9] Xu, Min & Chen, Daijiang & Zhou, Guangheng, "Real-Time Face Recognition Based on Dlib, " pp.1451–1459, 2020.

[10] D. Zhang, J. Li, and Z. Shan, "Implementation of Dlib Deep Learning Face Recognition Technology," 2020 International Conference on Robots & Intelligent System (ICRIS), Sanya, China, pp. 88-91, 2020.

[11] N. Ali, I. Hasan, T. Özyer, and R. Alhajj, "Driver Drowsiness Detection by Employing CNN and Dlib," 2021 22nd International Arab Conference on Information Technology (ACIT), Muscat, Oman, pp. 1-5, 2021.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, pp. 94-101, 2010.

[13] S. A. -P. Raja Sekaran, C. Poo Lee, and K. M. Lim, "Facial Emotion Recognition Using Transfer Learning of AlexNet," 2021 9th International Conference on Information and Communication Technology, Yogyakarta, Indonesia, pp. 170-174, 2021.

[14] Haghpanah, Mohammad A., et al. "Real-time facial expression recognition using facial landmarks and neural networks," 2022 International Conference on Machine Vision and Image Processing. IEEE, 2022.

[15] M. Mukhopadhyay, A. Dey, R. N. Shaw, and A. Ghosh, "Facial emotion recognition based on Textural pattern and Convolutional Neural Network," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies, Kuala Lumpur, Malaysia, pp. 1-6, 2021.

[16] Akhand, M. A. H. & Roy, Shuvendu & Siddique, Nazmul & Kamal, M.A.S. & Shimamura, Tetsuya., "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," *Electronics*, 2021.

[17] Jolliffe, Ian & Cadima, Jorge, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* pp.374,2016.

[18] Chen, D., Chang, Y., Yan, R., & Yang, J, "Tools for protecting the privacy of specific individuals in video," *EURASIP Journal on Advances in Signal Processing*, 2007, pp.1-9, 2007.

[19] Lamsal, B., Kojima, N., & Matsumoto, N., "Impact of the stochastic resonance on dark and illumination variant images for face detection," *Journal of the Institute of Industrial Applications Engineers*, 3(4), pp.167-173. 2015.