International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 5
ISSN: 2088-5334

Emotion Recognition on Facial Expression and Voice: Analysis and Discussion

Kok-Why Ng [a,*], Yixen Lim [a], Su-Cheng Haw [a], Yih-Jian Yoong [a]

[a] *Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia*
*Corresponding author: *kwng@mmu.edu.my*

Abstract **- Emotion plays an important role in our daily lives. Emotional individuals can affect the performance of a company, the harmony of a family, the wellness or growth (physical, mental, and spiritual) of a child etc. It renders a wide range of impacts. The existing works on emotion detection from facial expressions differ from the voice. It is deduced that the facial expression is captured on the face externally, whereas the voice is captured from the air passes through the vocal folds internally. Both captured output models may very much deviate from each other. This paper studies and analyses a person's emotion through dual models -- facial expression and voice separately. The proposed algorithm uses a Convolutional Neural Network (CNN) with 2-dimensions convolutional layers for facial expression and 1-Dimension convolutional layers for voice. Feature extraction is done via face detection, and Mel-Spectrogram extraction is done via voice. The network layers are fine-tuned to achieve the higher performance of the CNN model. The trained CNN models can recognize emotions from the input videos, which may cover single or multiple emotions from the facial expression and voice perspective. The experimented videos are clean from the background music and environment noise and contain only a person's voice. The proposed algorithm achieved an accuracy of 62.9% through facial expression and 82.3% through voice.**

Keywords **- Emotion recognition; facial expression; voice; convolutional neural network; Mel-spectrogram.**

Manuscript received 21 Nov. 2022; revised 29 Apr. 2023; accepted 12 Jul. 2023. Date of publication 31 Oct. 2023.

I. INTRODUCTION

The escalating cost of living has created a myriad of unforeseen stressors for individuals, significantly impacting their psychological well-being [1]-[5]. In response, people often conceal their true emotions, leading to various problems, such as the alarming rise in suicide cases and the prevalence of mental health disorders. Recognizing the gravity of these challenges, emotion recognition applications have emerged as crucial tools in detecting and understanding human emotions. In the workplace context, where emotions play a pivotal role in employee well-being and organizational success, accurately assessing and addressing emotions has garnered substantial attention.

Historically, studies in emotion recognition have predominantly focused on facial expressions, employing advanced computer vision techniques to analyze images and detect emotions [6]-[10]. Promising results have been achieved, with accuracies surpassing 95% in numerous cases. For instance, a smiling expression, indicated by an upward curve of the mouth, is typically associated with happiness, while a downturned curve suggests sadness or unhappiness. However, relying solely on facial expressions to determine emotions can yield inaccurate results. A curved mouth or lips may, at times, indicate neutrality or moderate sadness, necessitating the inclusion of additional modalities for a more precise analysis. The following section discusses the historical context of facial and audio expression-based emotion recognition studies. The proposed approach is covered in Section 3, while the implementation results is presented in Section 4. Section 5 concludes the paper and suggests some future works.

II. MATERIAL AND METHOD

This section discusses some latest research works done in facial expression and voice separately, followed by the proposed method.

A. *Emotion Recognition by Facial Expression*

CNN is utilized for facial expression recognition. Among the well-known CNN architectures for image classification are VGGNet, AlexNet, and Inception [11]-[16]. There are numerous kernel sizes and numbers of filters in CNN, including 2, 4, 8, 16, 32, 64, 128, and 256. Different results and accuracy may result from different kernels of varying sizes and number of filters. It is very important to select a suitable kernel size because low kernel sizes (such as 2) may lead to a highly unstable network, but kernel sizes such as 32 and 64 will not achieve convergence for some merging of parameters in the model, while moderate and suitable kernel sizes (such as from 8 to 16) converge very well. One of the proposed models, which is not uniform and has different filter numbers across depth, can achieve a high accuracy of 65%, making it perform best for the Facial Expression Recognition 2013 (FER-2013) dataset.

By greedy training and stacking some layers of Restricted Boltzmann Machines (RBM), the Deep Belief Network (DBN) [17] is created. The ability to quickly find a suitable set of model parameters and perform well when computing latent variables in the deepest layer are two of DBN's advantages. However, the trained model may not interpret some ambiguous sensory input if the top-down influences on the inference process are ignored. Additionally, it only learns one layer of features at a time and does not adjust parameters at a lower level. This could lead to inefficiency and slowness.

An edge detector algorithm called Local Directional Patterns (LDP) is combined with the Local Binary Pattern (LBP) as the modification for the feature extraction originally designed for texture description [18]. LBP was later improved for face recognition due to its high accuracy. In order to use LBP, the image must first be converted to grayscale. After that, a label will be assigned to each pixel by thresholding each pixel's 3 x 3 region in line with the central pixel. The image is broken up into smaller regions for LDP, and the histograms serve as the basis for the LDP descriptors. This research uses four different classifiers: LDP alone (baseline), KNN, Voting Classifier, and AdaBoost. Suppose only LBP is applied (baseline). In that case, the lowest accuracy of the recognition is around 93% for KNN, and the highest is 97.5% for the AdaBoost classifier. In contrast, the lowest accuracy obtained for LDP with Extend Local Binary Pattern (ELBP) is around 94% for KNN and 99% for Voting Classifier. This research demonstrated that facial recognition accuracy with LBP can be improved using the LDP method in conjunction with the ELBP Feature Extractor and the Voting Classifier.

In this study [19], Haar Cascade classifier and Local Binary Pattern Histogram (LBPH) were also applied to low-resolution images from 76 x 76 pixels to 156 x 156 pixels. The histogram equalization and median filtering are used to pre-process the face images. The histogram is used to represent the appearance of each binary code in the image in the enhanced version of the LBP known as LBPH. The proposed method achieved a recognition rate of 92.67% and 99.67% for real-time images and local files, respectively. This study demonstrated that the face recognition rate could be improved with the use of a median filter and a Haar Cascade classifier because the recognition rate achieved was higher than that of previous studies.

The Viola-Jones algorithm combines Haar Cascade features used for detecting the face in [20]. The authors apply LBPH to convert the captured image into a binary vector. This is done to enhance the face detection of the Viola-Jones algorithm. A pre-trained model of CNN, VGG16 is used as the classifier with the combination of max pooling and SoftMax classifier. CNN was chosen as the classifier due to its ability to avoid a clear distinction between functions and indirect learning of training data. The accuracy of the model VGG-16 Face is 88% based on the validation provided by the Karolinska Directed Emotional Faces (KDEF) dataset. Because the Viola-Jones algorithm has an invariant detector that finds scales, it can scale features rather than the image itself. However, the Viola-Jones algorithm is extremely sensitive to lightning and cannot effectively detect faces that are tilted. Compared to other face detection algorithms, the processing time for Haar Cascade features is longer.

Principal component analysis (PCA) and LBP with the classifier SVM and KNN with Euclidean distance (L2) are used in [21]. One of the categories of feature extraction, which is also referred to as appearance-based techniques, includes both PCA and LBP. The entire face will be processed using statistical techniques and linear transformation in the search for face-illustrative feature vectors. AdaBoost is useful for selecting the most effective LBP features, and LBP has the ability to divide the face image into multiple regions. In contrast, PCA involves removing less important dimensions, organizing dimensions according to significance, and locating the data's orthogonal basis. PCA and LBP with SVM perform the best overall on the JAFEE and the MUFE databases, where the recognition rates are approximately 88% and 76%.

B. Emotion Recognition by Voice

Paper in [22] aimed to compare the performance of Support Vector Machines (SVM), Recurrent neural network (RNN) and Multivariate Linear Regression classification (MLR) on Speech Emotion Recognition (SER). They first extracted the speech signals' Mel-frequency cepstral coefficients (MFCC) and Modulation Spectral (MS) features. For some classifiers ([23][24]), they applied feature selection and speaker normalization to search for the relevant feature subset. Both regression and classification problems can be solved by using MLR. Before sending the data to MLR for classification, the authors altered the Linear Regression Classification (LRC). An optimal margin classifier in machine learning is SVM. SVM can perform better with limited data, so it can be used in experiments with only a small amount of training data. Last but not least, RNN is very good at learning time series data, but it has a problem with gradient vanishing that gets worse as the training sequences get longer. Consequently, feature extraction of MFCC+MS with RNN but without speaker normalization achieves 94% recognition accuracy.

Multilayer Perceptron (MLP), SVM, and Logistic Regression (LR) with MFCCs are also compared [25]. In the feature extraction process, three features will be extracted: Mel-Spectrogram, chroma, and MFCC. Mel-Spectrogram contains Mel-Scaled frequencies, and it takes emotion samples over time to act as the audio signal. Then, the signal is mapped from the time and frequency domain using the Fast Fourier Transform (FFT) and frequency and amplitude are shifted to construct the spectrogram. Chroma is used to extract the audio's vocal content and determine the pitch rotation's angle as the helix traverses while MFCC can capture the crucial phonetical characteristics of an audio file. As the result of the comparison, the accuracy of the MLP model is 84.62%, LR is 100% and SVM is 91.67%.

The researchers proposed a Deep Neural Network (DNN) model to recognize emotional states from a one-second frame of raw speech spectrogram because it contains acoustic and semantic features [26]. DNN is a feed-forward neural network with many hidden layers, and each layer consists of multiple neurons that hold a weight of the previous layer's output and an intercept term or bias. The result is passed to the next layer through a non-linear function such as the sigmoid function, SoftMax function, and Rectified Linear Unit (ReLU). Mini-batch Stochastic Gradient Descent (SGD) is used to update weights during training. This model also includes dropouts because DNN is prone to overfitting. The process of recognizing emotions from a spectrogram is made possible by the deep hierarchical architecture of DNN, the addition of data, and sensible regularization. The Surrey Audio-Visual Expressed Emotion (SAVEE) and eNTERFACE databases [27] are used to assess the proposed model's accuracy. For both databases, speech or voice emotion recognition accuracy is approximately 60%.

A deep Convolutional Neural Network (CNN) from spectrograms for SER was discussed in [28]. A spectrogram is a graphic representation of the intensity of a signal at various frequencies in a particular waveform over time. A time-frequency representation is created by computing the speech signal using FFT. To carry out SER, salient discriminative features are extracted from the spectrograms. The fact that these robust and discriminative features are automatically learned from spectrograms as the foundation for SER is one of the benefits of using them in this research. As CNN's inputs, the speech signals are generated to form spectrograms. CNN consists of three convolutional layers, several pooling layers, three fully connected layers, and the SoftMax activation layer, which classified the model's seven emotions. For anger, boredom, disgust, and sadness, the proposed model achieved accuracy greater than 50%, while for fear, happiness, and neutrality, it achieved an accuracy of less than 50%. Only anger, fear, and neutral emotions performed better than the proposed model when compared to the pre-trained AlexNet model, while the other emotions achieved lower accuracy. As a result, the proposed model outperforms the pre-trained model in terms of performance and complexity.

Multi-task learning (MTL) is used to implement SER with CNN [29]. Despite the fact that DNNs were reported to perform better than HMM and SVM, the authors stated that DNNs still suffer from a generalization error issue due to the lack of training data. As a result, they proposed MTL-based CNN (MTL-CNN), also known as transfer learning, which classifies emotion as the primary task and arousal level, valence level, and gender as minor tasks. They believe that the model's performance on the main task can be improved with the assistance of those auxiliary tasks. Emotional states, like neutral women's voices being confused with men's voices that are highly aroused, are gender specific. Therefore, gender-specific emotion classification may perform better than gender-neutral classification. An input layer, two convolutional layers, a fully connected layer, and an MTL output layer make up MTL-CNN. MTL-CNN's accuracy for emotion recognition with one main task and three auxiliary tasks is 89.59 %, which is higher than the model's accuracy of 86.56% for the main task alone.

C. Proposed Method

The proposed method to recognize facial expressions and voice in this paper is as follows.

1) Data Augmentation

This paper uses data augmentation to increase the size and range of the training data. With the balance and adequate training data, it helps improve the performance of the model and yields better precision. The process of gathering and labeling data is time-consuming and expensive. Thus, transforming the data into diverse aspects can significantly reduce the operational costs. The augmented data will be treated as new data after the augmentation process.

The ImageDataGenerator library and some methods such as horizontal flipping, rotation, width and height shifting, shearing and zooming are applied for image data. While for audio data, the augmentation techniques applied are noise-adding, pitch-tuning, shifting, dynamic-changing and speed-and-pitch-tuning.

2) Face Detection

ResNet SSD is a Single Shot-Multibox Detector (SSD) and uses ResNet-10 architecture as backbone [30]. ResNet-10 model has removed the fully connected layers and only uses the feature extraction layers. The SSD Head is embedded in its backbone. SSD predicts the bounding box and class at the same time in the same deep-learning model that analyses the image. This module supports three frameworks. They are Caffe, TensorFlow and PyTorch. In this paper, it applies the Caffe model. It is fundamentally cantered on the Visual Geometry Group (VGG) model.

The difference between the ResNet SSD in OpenCV to the other face detection methods like Dlib is that it separates the image into numerous grids instead of using the conventional sliding window algorithm. Each grid cell oversees recognizing items in a specific area of the image. "0" signifies that no object is located in the grid.

ResNet SSD is being utilized mainly due to it provides better accuracy than existing face detection methods like Haar Cascade Classifier. Besides, it is more efficient even on a CPU instantaneously. It can detect multiple face angles from different directions such as left, right, up and downside of the face. Contrasting to Dlib and Haar Cascade Classifier, it processes very fast even when there are a lot of obstacles.

3) Mel-Spectrogram

It is a signal defined as a change in a quantity over time and audio data is a variation of air pressure (amplitude) over time. Fourier transforms, or Fourier's theorem, allows us to decompose a signal into its frequencies and the amplitude of the frequencies. It lets us transform the signal from time domain to frequency domain and it is called a spectrum. The algorithm, Fast Fourier Transform (FFT) eases the calculation of the Fourier transform efficiently.

Some of the signals are non-periodic signals, which the signal's frequency content is different over time. Hence, there is an idea of calculating several spectrums from different segments of the audio data by using FFT and combine them together to form a piece of complete information called the Short-Time Fourier Transform (STFT) and the output is called the spectrogram by stacking all FFTs on top of each other. Due to human hearing can only concentrate in very small amplitude and frequency ranges, there is an adjustment needed, which is converting the frequency to a log scale and the amplitude to decibels it is known as the log scale of the amplitude.

Lower frequency differences are easier to notice in humans than higher frequency variations. It is also known as a non-linear transformation of the frequency scale. The formula of the Mel Scale is below:

$$M = 1127*\ln(1+f/700) \qquad (1)$$

which the *ln* is known as the natural logarithm (*log_e*).

4) Convolutional Neural Network (CNN)

CNN is a class of network comprising many layers of neurons that can study and adapt to their weights to categorize objects precisely. It takes less pre-processing as it can learn the characteristics or features by changing the weight of individuals neuron. CNN is mainly built for image classification process which the model examines in two-dimensional input and goes beyond feature learning and classification. CNN can also accomplish well on one-dimensional data like audio data. It can learn to extract important features from the given data at whatever of its dimensions or shape.

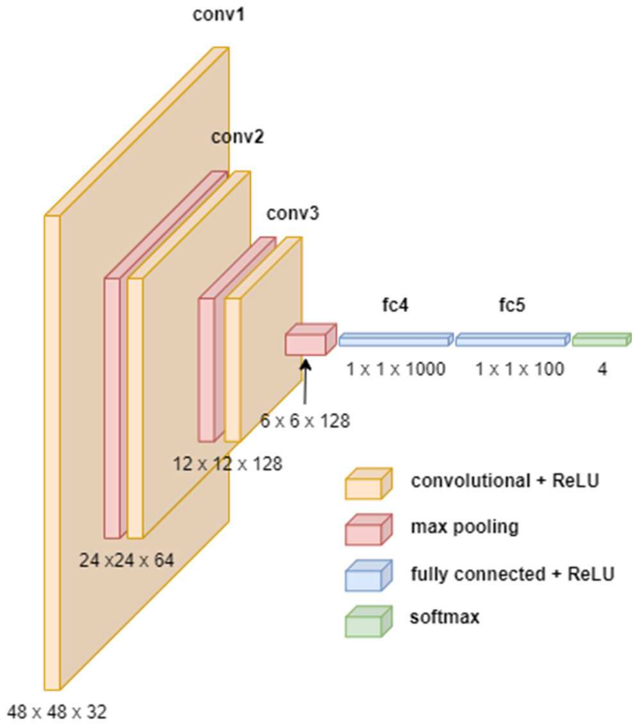


Fig. 1 CNN for image data

A 2-dimensions CNN was applied for image data as images are in 2D. ReLu function was selected as the activation functions for the CNN layers except the output layer. SoftMax was applied at the output layer as this is a classification of four classes of emotions. MAX pooling layers were also applied. The dataset used to train the CNN model is Facial Emotion Recognition (FER-2013). Initially, it consists of 28709 training data and 7178 testing data from seven emotions: angry, disgust, fear, happy, neutral, sad and surprise. All the images have the size of 48x48 pixels. Since only four emotions (angry, happy, neutral and sad) need to be categorized in this paper, after removing the unnecessary data, there are 21005 training data and 5212 testing data. Figure 1 shows the CNN built for emotion recognition by facial expression. It has three convolutional layers with MAX pooling layers and SAME padding. After the flattening process, there were 4608 neurons for the first layer of emotion classification. Then three FC layers were applied which have 1000, 100 and four neurons accordingly. The four neurons at the output layer represented the four target emotions and the activation function used is SoftMax function. Then the model was compiled with the Adam optimizer, categorical cross entropy for the loss function because this is a multiclass classification and accuracy metrics. Callbacks such as early stopping, learning rate reducing and model checkpoint are applied to ensure the model will not overfit.

While for audio data, a 1-dimension CNN was used, and the architecture is shown as Figure 2. The dataset used for the model training of emotion recognition by voice is the Ryerson Audio-Visual Database of Emotional Speech and Song dataset (RAVDESS), a validated multimodal database of emotional speech ang song. It consists of 24 professional actors whereas their genders are equivalent. Originally, the dataset contains 1440 total of speech and song of the seven emotions (angry, calm, disgust, fearful, happy, neutral, sad and surprise), but since only four emotions were needed in this paper, thus there are 672 final data left. Audio augmentation was applied as 672 data are insufficient for training and testing. Then, the audio data were loaded with Librosa library every 3 seconds and with a sample rate of 44100 Hz. By using the function to calculate Mel-Spectrogram provided by Librosa library, the feature extraction of Mel-Spectrogram was performed, and its amplitude was then converted to decibel to obtain the spectrogram. Finally, the spectrogram was averaged and combined with the emotions. There is a total of 259 features of Mel-Spectrogram obtained. Z-score normalization was applied, and the data were reshaped into three dimensions.

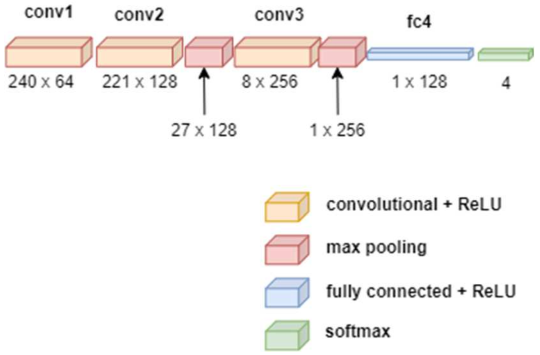
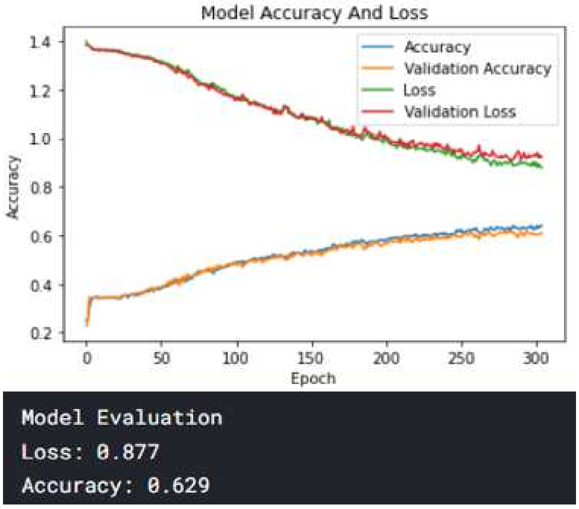


Fig. 2 CNN for audio data

III. Results and Discussion

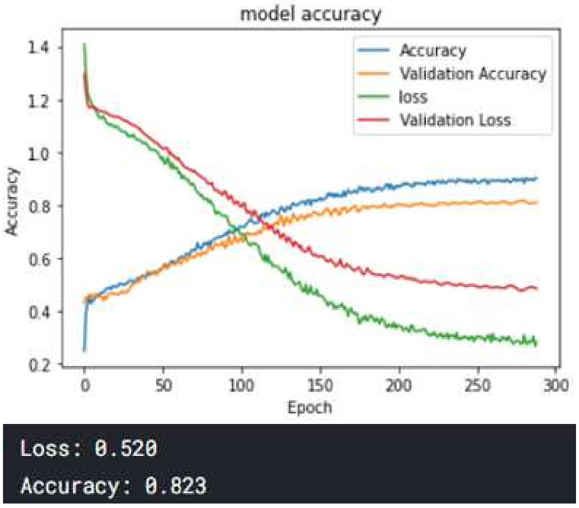
The CNN of facial expression is able to achieve the accuracy of 62.9% while the CNN of voice achieves 82.3% accuracy. The charts shown in Figure 3 and Figure 4 proved that the models did not overfit because the training loss is close to the validation loss until the training process is completed.

Emotional recognition by facial expression can work accurately most of the time under normal lighting conditions with frontal face position [31]. From the face images cropped by the face detection algorithm in Figure 5, it is observed that the employee is currently happy and the result generated was also correct. Happy emotion has the highest percentage which is about 56%.



```
Model Evaluation
Loss: 0.877
Accuracy: 0.629
```

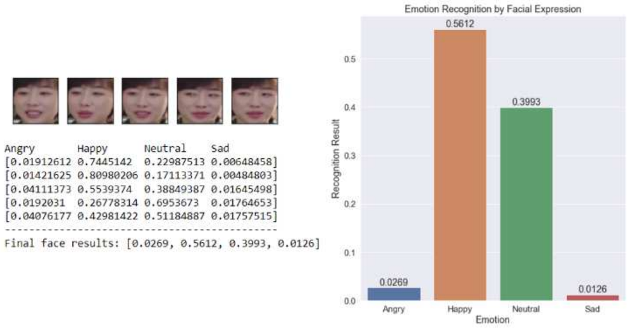
Fig. 3 Model evaluation of CNN of facial expression



```
Loss: 0.520
Accuracy: 0.823
```

Fig. 4 Model evaluation of CNN of voice

Other than that, the model can successfully detect hidden emotions by the employee. As shown in Figure 6, it is observed that the employee is pretending to be neutral on her face, however, her voice is a little sad and shaking. In the right chart, sad emotions are detected with a lower percentage as the employee is only a little sad but not crying. Furthermore, the proposed methodology is able to recognize combination of emotions. As shown in Figure 7, the predicted recognition result is that the employee is having a combination of sad and angry emotions, which matches with the scenario of the input video. Only sad and angry have higher detection percentage comparing to happy and neutral emotions. However, the algorithm has some limitations too. First, some emotions by facial expression might be wrongly classified as displayed in Figure 8. After some investigation, it is believed that the false prediction is caused by some wrongly classified images in the dataset FER-2013. There are some images of smiling or laughing categorized in the neutral emotion. This can cause the model to confuse between happy and neutral emotions and thus producing the incorrect result.



```
Angry       Happy      Neutral    Sad
[0.01912612 0.7445142  0.22987513 0.00648458]
[0.01421625 0.80980206 0.17113371 0.00484803]
[0.04111373 0.5539374  0.38849387 0.01645498]
[0.0192031  0.26778314 0.6953673  0.01764653]
[0.04076177 0.42981422 0.51184887 0.01757515]
---------------------------------------------
Final face results: [0.0269, 0.5612, 0.3993, 0.0126]
```

Fig. 5 Emotion recognition by facial expression of Happy.mp4.

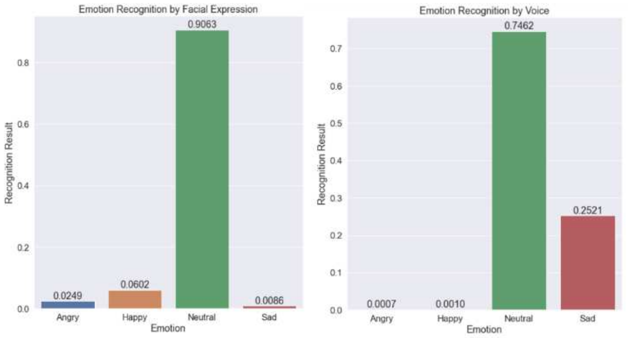
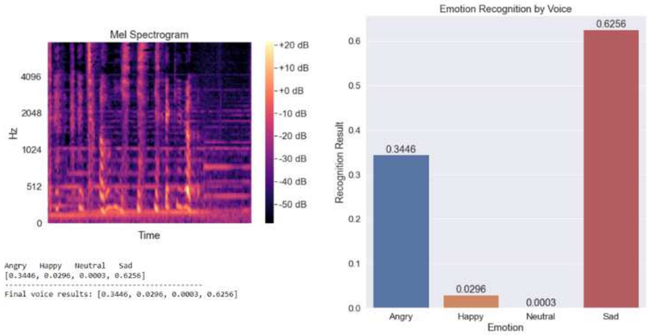
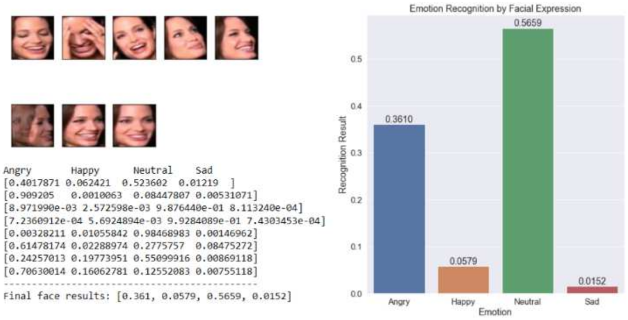


Fig. 6 Output of Neutral.mp4.



```
Angry   Happy   Neutral   Sad
[0.3446, 0.0296, 0.0003, 0.6256]
------------------------------------------
Final voice results: [0.3446, 0.0296, 0.0003, 0.6256]
```

Fig. 7 Result for combination of two emotions.

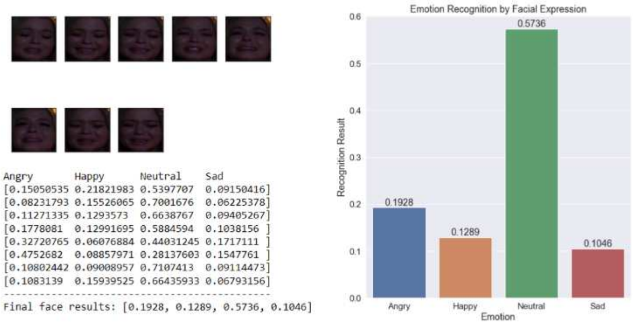


```
Angry      Happy      Neutral    Sad
[0.4017871 0.062421  0.523602  0.01219   ]
[0.909205   0.0010063  0.08447807 0.00531071]
[8.971990e-03 2.572598e-03 9.876440e-01 8.113240e-04]
[7.2360912e-04 5.6924894e-03 9.9284089e-01 7.4303453e-04]
[0.00328211 0.01055842 0.98468983 0.00146962]
[0.61478174 0.02288974 0.2775757  0.08475272]
[0.24257013 0.19773951 0.55099916 0.00869118]
[0.70630014 0.16062781 0.12552083 0.00755118]
---------------------------------------------
Final face results: [0.361, 0.0579, 0.5659, 0.0152]
```

Fig. 8 Happy face is classified as neutral of angry

Other than that, louder voice tends to be wrongly classified as angry emotion while lower voice can be recognized as sad emotion. Some people could laugh until they have a shaking or broken voice, especially when they laugh out loud. This is because angry people tend to have a louder voice and sad people tend to talk with a lower tone. Since the Mel-spectrogram only extracts features of the voice and not content, this could possibly make the model classify incorrectly.

Lastly, the emotion recognition by facial expression could not work well if the lighting condition is too dark or the face position is not frontal. As shown in Figure 9, the face images cropped out are very dark, thus losing useful information in the face region, causing the model unable to successfully recognize the correct sad emotion.



```
Angry      Happy      Neutral    Sad
[0.15050535 0.21821983 0.5397707  0.09150416]
[0.08231793 0.15526065 0.7001676  0.06225378]
[0.11271335 0.1293573  0.6638767  0.09405267]
[0.1778081  0.12991695 0.5884594  0.1038156 ]
[0.32720765 0.06076884 0.44031245 0.1717111 ]
[0.4752682  0.08857971 0.28137603 0.1547761 ]
[0.10802442 0.09008957 0.7107413  0.09114473]
[0.1083139  0.15939525 0.66435933 0.06793156]
---------------------------------------------
Final face results: [0.1928, 0.1289, 0.5736, 0.1046]
```

Fig. 9 Result produced under dark lighting condition

IV. Conclusion

The proposed emotion recognition algorithm which integrated with two CNN prediction models to predict the emotions from facial expression by a 2-Dimension CNN with 63% accuracy and voice by a 1-Dimension CNN with 82% accuracy are achieved. The experimented videos are all clean from the background music, environment noise and only contained of one's voice. The OpenCV ResNet SSD was selected due to its speed, accuracy, low rate of false positive detection and other aspects that perform the best among four face detection algorithms. Data augmentation, other feature selection of voice and transfer learning were also applied to find out the best combination to improve the algorithm's performance. Analysis was performed on the outputs based on the videos.

Some limitations in this work have been discovered. Part of the cropped face regions might be darker or in shadow, resulting in the reduce in accuracy of the 2-Dimension CNN model. To improve this limitation, some visual information processing techniques such as adjusting the brightness of the cropped face regions if is below the threshold could be carried out. The algorithm's performance was reduced as there are possibilities to identify the incorrect emotions when the voice extracted was too high or too low. This could be improved by training a more accurate categorized dataset or combination of datasets.

References

[1] Lim, Y., Ng, K. W., Naveen, P., & Haw, S. C., "Emotion Recognition by Facial Expression and Voice: Review and Analysis," *Journal of Informatics and Web Engineering (JIWE)*, 1(2), pp. 45-54, 2022. doi:10.33093/jiwe.2022.1.2.4.

[2] Naga, P., Marri, S. D., & Borreo, R., "Facial emotion recognition methods, datasets and technologies: A literature survey," *Materials Today: Proceedings*, 80(1), pp. 2824-2828, 2023. doi:10.1016/j.matpr.2021.07.046.

[3] Park, C. L., Kubzansky, L. D., Chafouleas, S. M., Davidson, R. J., Keltner, D., Parsafar, P., ... & Wang, K. H., "Emotional well-being: What it is and why it matters," *Affective Science*, 4(1), pp. 10-20, 2023. doi: 10.1007/s42761-022-00163-0.

[4] Anaam, E. A., Haw, S. C., Ng, K. W., Naveen, P., & Thabit, R., "Utilizing Fuzzy Algorithm for Understanding Emotional Intelligence on Individual Feedback," *Journal of Informatics and Web Engineering (JIWE)*, 2(2), pp. 273-283, 2023. doi: 10.33093/jiwe.2023.2.2.19.

[5] Lahat, L., & Ofek, D., "Emotional well-being among public employees: A comparative perspective," *Review of Public Personnel Administration*, 42(1), pp. 31-59, 2022. doi: 10.1177/0734371X20939642.

[6] Leong, S. C., Tang, Y. M., Lai, C. H., & Lee, C. K. M., "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing," *Computer Science Review*, 48, 100545, 2023. doi: 10.1016/j.cosrev.2023.100545.

[7] Cai, Y., Li, X., & Li, J., "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review," *Sensors*, 23(5), 2455, 2023. doi: 10.3390/s23052455.

[8] Lam, X. H., Ng, K. W., Yoong, Y. J., & Ng, S. B., "WBC-based segmentation and classification on microscopic images: a minor improvement," *F1000Research*, 10, 1168, 2021. doi: 10.12688/f1000research.73315.1

[9] Ang, J. S., Ng, K., & Chua, F. F., "Stock market prediction using deep learning approach," *Journal of Engineering Science and Technology*., Vol. 17, No. 5 pp. 3174-3186, 2022.

[10] Sarvakar, K., Senkamalavalli, R., Raghavendra, S., Kumar, J. S., Manjunath, R., & Jaiswal, S., "Facial emotion recognition using convolutional neural networks," *Materials Today: Proceedings*, 80, pp. 3560-3564, 2023. doi: 10.1016/j.matpr.2021.07.297.

[11] Agrawal, A., & Mittal, ·. N., "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *The Visual Computer* 36 (2), pp. 405-412, 2020. doi:10.1007/s00371-019-01630-9.

[12] Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., & Singh, S., "Facial expression recognition in videos using hybrid CNN & ConvLSTM," *International Journal of Information Technology*, 15(4), pp. 1819-1830, 2023. doi: 10.1007/s41870-023-01183-0.

[13] Chand, V., Chrisanthus, A., Thampi, A., Dayal, S., & Dhanup, S., "A Review on Various CNN-based Approaches for Facial Expression Recognition," *In 2023 International Conference on Inventive Computation Technologies (ICICT)*, pp. 465-471, 2023. IEEE. doi: 10.1109/ICICT57646.2023.10133947.

[14] De Ocampo, A. L. P., "Haar-CNN Cascade for Facial Expression Recognition," *In 2023 International Electrical Engineering Congress (iEECON)*, pp. 89-92, 2023. IEEE. doi: 10.1109/iEECON56657.2023.10126902.

[15] Nugraha, G. S., Darmawan, M. I., & Dwiyansaputra, R., "Comparison of CNN's Architecture GoogleNet, AlexNet, VGG-16, Lenet-5, Resnet-50 in Arabic Handwriting Pattern Recognition," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2023. doi: 10.22219/kinetik.v8i2.1667.

[16] A. Nainwal, G. Sharma, V. Kansal, S. Bhatla and B. Pant, "Comparative Study of VGG-13, AlexNet, MobileNet and Modified-DarkCovidNet for Chest X-Ray Classification," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 413-417.

[17] Agarwalla, N., Panda, D., & Modi, M. K., "Deep Learning using Restricted Boltzmann Machines," *International Journal of Computer Science & Information Security*, 7(3), pp. 1552-1556, 2016.

[18] Chengeta, K., & Viriri, S., "A Survey on Facial Recognition based on local directional and local binary patterns," *Conference on Information Communications Technology and Society (ICTAS)*, 2018. doi: 10.1109/ICTAS.2018.8368757.

[19] Isnanto, R. R., A. F., Eridani, D., & Cahyono, G. D., "Multi-Object Face Recognition Using Local Binary Pattern Histogram and Haar Cascade Classifier on Low-Resolution Images," *International Journal of Engineering and Technology Innovation*, vol. 11, no. 1, 2021, pp. 45-58, 2021. doi: 10.46604/ijeti.2021.6174.

[20] Hussain, S. A., & Balushi, A. S., "A real time face emotion classification and recognition using deep learning model," *Journal of Physics: Conference Series*, Vol.1432, No.1, pp. 012087, 2020. doi: 10.1088/1742-6596/1432/1/012087.

[21] Abdulrahman, M., & Eleyan, A., "Facial Expression Recognition Using Support Vector Machines," The 23nd *Signal Processing and Communications Applications Conference (SIU)*, pp. 276-279, 2015. doi: 10.1109/SIU.2015.7129813.

[22] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C., “Automatic Speech Emotion Recognition Using Machine Learning,” *Social media and machine learning*. IntechOpen, 2019. doi: 10.5772/intechopen.84856.

[23] Mannar Mannan J, Srinivasan L, Maithili K, & Ramya C., "Human Emotion Recognize Using Convolutional Neural Network (CNN) and Mel Frequency Cepstral Coefficient (MFCC)," *Seybold Report Journal*, 18(4), pp. 49-61, 2023.

[24] Mishra, S. P., Warule, P., & Deb, S., "Deep Learning Based Emotion Classification Using Mel Frequency Magnitude Coefficient," In 2023 *1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)* (pp. 93-98). IEEE, 2023. doi: 10.1109/IHCSP56702.2023.10127148.

[25] Rumagit, R. Y., Alexander, G., & Saputra, I. F., "Model Comparison in Speech Emotion Recognition for Indonesian Language," *Procedia Computer Science*, 179, pp. 789-797, 2021. doi: 10.1016/j.procs.2021.01.09.

[26] Fayek, H., Lech, M., & Cavedon, L., "Towards real-time speech emotion recognition using deep neural networks," The 9th *international conference on signal processing and communication systems (ICSPCS)*, pp. 1-5, 2015. doi: 10.1109/ICSPCS.2015.7391796.

[27] Sharafi, M., Yazdchi, M., & Rasti, J., "Audio-Visual Emotion Recognition Using K-Means Clustering and Spatio-Temporal CNN," In 2023 *6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pp. 1-6, 2023. IEEE. doi; 10.1109/IPRIA59240.2023.10147192.

[28] Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W., "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," *International conference on platform technology and service (PlatCon)*, pp. 1-5, 2017. doi: 10.1109/PlatCon.2017.7883728.

[29] Kim, N. K., Lee, J., Ha, H. K., Lee, G. W., Lee, J. H., & Kim, H. K., "Speech emotion recognition based on multi-task learning using a convolutional neural network," *The Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 704-707, 2017. doi: 10.1109/APSIPA.2017.8282123.

[30] Lu, X., Kang, X., Nishide, S., & Ren, F., "Object detection based on SSD-ResNet," In 6th *International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 89-92, 2019. IEEE. doi: 10.1109/CCIS48116.2019.9073753.

[31] S. A. Mithbavkar and M. S. Shah, “Recognition of Emotion in Indian Classical Dance Using EMG Signal,” International Journal on Advanced Science, Engineering and Information Technology, vol. 11, no. 4, p. 1336, Aug. 2021, doi: 10.18517/ijaseit.11.4.14034.