Development of a music recommendation application by using facial emotion recognition

Shengke Xie, Raenu Kolandaisamy, Ghassan Saleh
Institute of Computer Science and Digital Innovation, UCSI University, 56000 Kuala Lumpur, Malaysia

Heshalini Rajagopal
Department of Electrical and Electronics Engineering, MILA University, 71800 Nilai, Negeri Sembilan, Malaysia
E-mail: raenu@ucsiuniversity.edu.my

Abstract

Music is an important part of human life and culture, and it can affect people' s emotions and moods. However, choosing music from a large library can be a challenging and time-consuming task. In this paper, we propose a facial expression recognition-based music recommendation system that can recommend suitable music which matches the user' s current mood. The system uses a camera to capture the user's face and a convolutional neural network model which trained facial emotion recognition database to recognize seven basic emotions: anger, disgust, fear, happiness, sadness, surprise and neutral. The paper contributes to the research of facial emotion recognition and music recommendation and provides a convenient way for people to enjoy music.

Keywords: Facial Emotion Recognition, Music Recommend Systems, Convolutional Neural Network

1. Introduction

With the development of the Internet of Things (IoT) technology, various smart homes are entering the public eye, and AI voice assistants used to unify the management of various smart home appliances have also emerged. People control the use of various electrical appliances and furniture in the same network by giving commands to the AI voice assistants with their words, while many AI voice assistants have added many other functions in addition to the basic home control functions in order to increase the diversity of functions, such as music recommendations. When people are looking for music to listen to, they are often confused about the choice of music and are looking for some suggestions to help them choose, since it is often difficult to choose the most suitable music from a large library of thousands of songs, and music recommendations are one such feature that can reduce the difficulties users have in choosing music [1].

However, the music recommendation function sometimes does not meet the user's needs, and there are often situations where the recommended music conflicts with the user's current mood, such as recommending sad songs when the user is happy, or recommending songs with a strong rhythm when the user needs to calm down [2]. Therefore, how to make the computer quickly and accurately recommend the right song to the user is a problem that needs to be solved. It should be noted that facial expressions account for two-thirds of human communication and are one of the most important means of expressing human emotions. In the absence of facial expression recognition, AI is sometimes unable to give correct and rapid feedback through speech alone. For example, the same phrase what's up may have different meanings depending on the emotion, it may be a greeting between acquaintances, or it may be a concern to ask what difficulties you are experiencing, so people hope that computers will be able to understand human emotions and give correct feedback, bringing people a better experience with more intelligent human-computer interaction [3].

For this problem, we will use a camera to capture the user's image, use python deep learning based facial expression recognition technology to identify the user's current mood, and select the music that matches the user's current mood from the music categories in the database that apply to different moods to recommend music to the user, thus trying to reduce the user's anxiety in choosing music and improve the user's experience

2. Literature Review

Communication is a bridge to build interpersonal relationships, people send or obtain information through communication, and facial expressions are an important means and factor to help people understand the information conveyed by others in communication, according to the survey [4] the non-verbal component of communication reaches about 55% of interpersonal communication, and as an important category of non-verbal component, the study of human facial expressions is not only in medicine and psychology, its

wide application prospects have also led to its widespread interest in computer science [5].

As early as 1971, two American psychologists, Ekman and Friesen, had systematically studied facial expressions and in 1978 developed and defined the Facial Action Coding System (FACS), a system for recognizing various human emotions by representing different facial expressions in terms of different facial muscle changes. The system encodes specific facial muscle changes, called Action Units (AUs). Fig. 1 shows the main types of Aus.

Action Unit	Description	Facial Muscle
AU0	Neutral Face	
AU1	Inner Brow Raiser	frontalis (pars medialis)
AU2	Outer Brow Raiser	frontalis (pars lateralis)
AU4	Brow Lowerer	depressor glabellae, depressor supercilii, corrugator supercilii
AU5	Upper Lid Raiser	levator palpebrae superioris, superior tarsal muscle
AU6	Cheek Raiser	orbicularis oculi (pars orbitalis)
AU7	Lid Tightener	orbicularis oculi (pars palpebralis)
AU8	Lips Toward Each Other	orbicularis oris
AU9	Nose Wrinkler	levator labii superioris alaeque nasi
AU10	Upper Lip Raiser	levator labii superioris, caput infraorbitalis
AU11	Nasolabial Deepener	zygomaticus minor
AU12	Lip Corner Puller	zygomaticus major
AU13	Sharp Lip Puller	levator anguli oris (also known as caninus)
AU14	Dimpler	buccinator
AU15	Lip Corner Depressor	depressor anguli oris (also known as triangularis)
AU16	Lower Lip Depressor	depressor labii inferioris
AU17	Chin Raiser	mentalis
AU18	Lip Pucker	incisivii labii superioris and incisivii labii inferioris
AU19	Tongue Show	
AU20	Lip Stretcher	risorius
AU22	Lip Funneler	orbicularis oris
AU23	Lip Tightener	orbicularis oris
AU24	Lip Pressor	orbicularis oris
AU25	Lips Part	depressor labii inferioris, or relaxation of mentalis or orbicularis oris
AU26	Jaw Drop	masseter; relaxed temporalis and internal pterygoid
AU27	Mouth Stretch	pterygoids, digastric
AU28	Lip Suck	orbicularis oris
AU41	Lid Droop	relaxation of levator palpebrae superioris
AU42	Slit	orbicularis oculi
AU43	Eyes Closed	relaxation of levator palpebrae superioris
AU44	Squint	orbicularis oculi, pars palpebralis
AU45	Blink	relaxation of levator palpebrae and contraction of orbicularis oculi, pars palpebralis.
AU46	Wink	levator palpebrae superioris; orbicularis oculi, pars palpebralis

Fig. 1. The main types of Aus

It is well-known that human beings have six basic emotions (BEs), namely happiness, surprise, anger, sadness, fear and disgust, on the basis of which Du et al [6] have proposed 22 compound emotions (CEs) made up of combinations of basic emotions with each other, and different combinations of AU classify these facial emotions to facilitate identification, e.g. AU combinations 1, 4, 20, 25 can be identified as fear emotions. Fig. 2 shows the main types of human facial emotions represented by the various AUs combinations.

Category	AUs	Category	AUs
Happy	12, 25	Sadly disgusted	4, 10
Sad	4, 15	Fearfully angry	4, 20, 25
Fearful	1, 4, 20, 25	Fearfully surprised	1, 2, 5, 20, 25
Angry	4, 7, 24	Fearfully disgusted	1, 4, 10, 20, 25
Surprised	1, 2 , 25, 26	Angrily surprised	4, 25, 26
Disgusted	9, 10, 17	Disgusted surprised	1, 2, 5, 10
Happily sad	4, 6, 12, 25	Happily fearful	1, 2, 12, 25, 26
Happily surprised	1, 2, 12, 25	Angrily disgusted	4, 10, 17
Happily disgusted	10, 12, 25	Awed	1, 2, 5, 25
Sadly fearful	1, 4, 15, 25	Appalled	4, 9, 10
Sadly angry	4, 7, 15	Hatred	4, 7, 10
Sadly surprised	1, 4, 25, 26	-	-

Fig.2. The AUs combinations that can be observed in different human facial emotions [5]

3. Method

This system implements a music recommendation system based on facial expression recognition on the PC. Fig. 3 shows the steps for facial recognition.

Fig.3. Steps for facial emotion recognition

The aim of this system is to use the computer's powerful computing resources to quickly and accurately understand human emotions and recommend more appropriate songs to the user, bringing a smarter experience to people.

The system was developed on a PC with Windows 10, 64-bit operating system, Intel-i7-10510U CPU, 16G Byte RAM, PyCharm 2022.2.2, OpenCV-3.4.4.19 compilation environment and python as the main programming language. Tests of the system were also carried out on this machine.

The process of using this system can be simply summarized as capturing the user's image using the local camera, using deep learning based facial expression recognition technology to identify the user's current mood, and recommending music that matches the user's current mood from the downloaded music that has been categorized into different mood folders, thus trying to reduce the user's anxiety in choosing music and improve the user's experience. Fig.4 shows the workflow of this system.

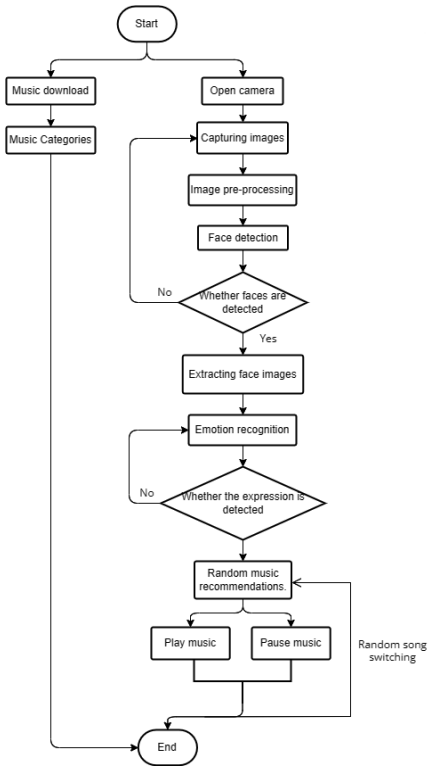

Fig. 4. System workflow diagram

According to the need, the design of this system as shown in Fig. 5 is mainly divided into seven modules such as Image pre-processing, Face detection, Emotion recognition, Music download, Music classification, Music recommendation and Music playing:

1) Image pre-processing module: the image captured by the local camera is pre-processed, the pre-processing process is mainly to convert the colour image into a grey scale image, which is convenient for face detection.

2) Face detection module: face detection is performed on the grayscale image, after the face is detected, a rectangular box is drawn on the original colour image to frame out the face, in preparation for expression recognition.

3) Emotion recognition module: The detected face image is fed into the expression recognition model for expression recognition. The recognition model is a convolutional neural network model with parameters trained to recognize the face expression and return the result.

4) Music download module: The crawler crawls the MP3 file of the searched music from the music playing website and downloads it to the local area, while crawling the corresponding lyrics and saving them in txt format, so as to prepare for music classification.

5) Music classification module: count the emotion value of each emotion word in the text according to the emotion dictionary, and classify the music files corresponding to the lyrics according to the emotion value.

6) Music recommendation module: according to the expression recognition result, a song is randomly loaded from the local music folder of the corresponding emotion

7) Music playing module: play or pause the loaded song.

Among them, the core function of the system is Emotion recognition, the recognition result will be directly related to the quality of the system, the appropriateness of the recommended music.

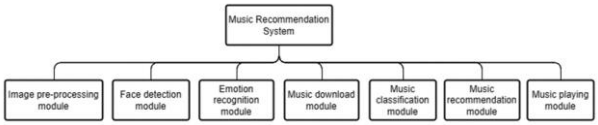

Fig.5. Structure diagram of the system modules

3.1. User Interface Design

A clean and aesthetically pleasing user interface (UI) provides a good user experience and helps users to interact easily with the software system. Here the UI is designed by using PyQt5, a GUI framework for the Python programming language, which makes it easy to implement powerful user interfaces using a variety of UI controls. The main controls used in this system are Label, Button, Line Edit and Timer.

Fig. 6 show the design blueprint of the UI and screenshots of the actual UI interface respectively, with the functions of each UI control as follows:

1) close_btn: Button control, used to close the whole system.

2) title_label: Label control for scrolling the name of the currently playing music file.

3) result_label: Label control, used to display the result of face expression recognition.

4) camera_btn: Button control for turning the local camera on and off.

5) play_btn: Button control to control the playing and pausing of the currently loaded music.

6) next_btn: Button control for randomly loading a piece of music from the folder matching the expression recognition result.

7) detect_btn: Button control that detects the user's current expression, returns the result and displays it in the result_label.

8) searchMusic_LEdit: Line Edit control for entering music names in preparation for music downloads.

9) download_btn: Button control for crawling and downloading mp3 files and lyric content of the music entered by the user.

10) camera_label: Label control to display the image captured by the camera.

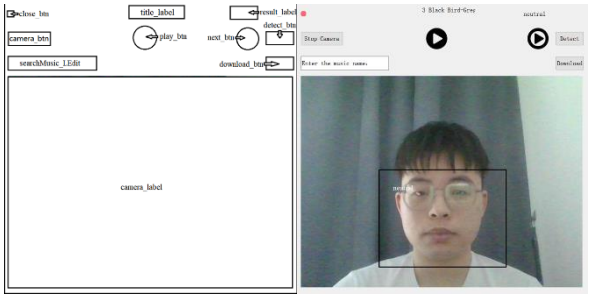

Fig.6. Design of the UI

3.2. *Design of Convolutional Neural Networks*

The design of the network structure affects the model performance and the accuracy of the classification. A good network structure can extract features more efficiently and reduce the training time. This system uses the convolutional neural network structure designed by Zhou [7] as the basis for classification. It achieved good recognition results on both the FER-2013 test set and the dataset [8]. The structure of this convolutional neural network consists of a 48×48 input layer and a 1×1 convolutional layer to form the first group of layer sets, and two consecutive convolutional layers with filter sizes of 3×3 and 5×5 followed by a 2×2 maximum pooling layer to form the second and third group of layer sets. Each convolutional layer uses ReLU as the activation function. The second group of layers has one more convolutional layer than the first group, which means that further refinement of features can be learned based on the first group of layers. The detailed structure of the network parameters is shown in Fig. 7.

Layer	Number of filters	Filter size	Stride	Padding	Feature image size
input	0	0	None	None	(48,48,1)
conv1-1	32	1×1	1	0	(48,48,32)
conv2-1	64	3×3	1	1	(48,48,64)
conv2-2	64	5×5	1	2	(48,48,64)
pool2	0	2×2	2	0	(24,24,64)
conv3-1	64	3×3	1	1	(24,24,64)
conv3-2	64	5×5	1	2	(24,24,64)
pool3	0	2×2	2	0	(12,12,64)
fc1	None	None	None	0	(1,1,2048)
fc2	None	None	None	0	(1,1,1024)
output	None	None	None	0	(1,1,8)

Fig.7. Network structure parameters [7]

3.3. Design and implementation of the main functions of the system

The music recommendation system based on facial expression recognition is divided into seven modules, and the main functions of each module are briefly introduced above.

Image pre-processing module

In order to ensure consistency in face size and position as well as face image quality in face images and to improve the efficiency and accuracy of face recognition, the images acquired by the camera need to be pre-processed prior to face detection, which can significantly reduce the computational effort of the device. The main task of pre-processing is image greyscale, which means that the image is unified into a grey-scale map of a specified size. In actual use will meet the image acquisition effect is not good, such as lighting and other factors cause image results are not good, and ultimately affect the recognition results, and the grayscale process can effectively remove these noise influence, after the grayscale, the influence of the noise is reduced to a minimum.

The method first normalises the input raw face image, scaling the pixel value range from 0-255 to between 0-1, then scales the raw face image to a specified image size of 48×48 and generates a set of face images of different sizes and orientations, providing more information for subsequent tasks such as face recognition and expression recognition, thus improving the accuracy and robustness of the algorithm.

The greyscale of the images is achieved using the COLOR_BGR2GRAY method in OpenCV.

Face detection module

The prerequisite for expression recognition is that the recognized image contains a face component, so face detection is a powerful guarantee for expression recognition. This module uses the face_detection_front.tflite and face_detection_back.tflite models from [8] for face detection.

The method first initialises the face detector, uses the detectFaces() method which from detector to detect faces in the input image, and returns a detection result containing the bounding box, key points and scores for each detected face. Finally, the returned detection results are used to obtain and return the absolute coordinates of the face bounding boxes.

Emotion recognition module

Emotion recognition is the core function of this system, image pre-processing, face detection modules are all paving the way for the implementation of this module.

In this method we first load a pre-trained facial emotion recognition model, read the video frames from the local camera using the capture.read() method in the OpenCV computer vision library, resize the video frames and greyscale them. Next, the faces returned by the face detection module are recognized for their emotions using the predict() method, and the predicted results are annotated in the original image and displayed in the control camera_label.

Music download module

The music download module mainly uses python's powerful crawler function to crawl music from music websites. The module consists of three main parts: music search, music crawling and music saving. Firstly, the searchSong() method is used to send an HTTP GET request to the url with the search content, and return a list of songs in JSON format to be prepared for the music download.

Then the MusicDownload() method uses the returned JSON list of songs to extract the song information for the first song in the list, including the song name, artist and song ID. Then the downloadSong() method is called and passed the extracted song information.

Finally, in the downloadSong() method, a request is sent to the url consisting of the song ID and a response is received. The music content is found in the JSON format response, read and saved as an MP3 file in the local download folder, thus completing the download of the music and also preparing the music for classification.

Music classification module

In the music classification module, we will classify the downloaded music according to the corresponding emotion and transfer the files to the corresponding emotion folder. For this function we rely on a sentiment dictionary from [9], which assigns sentiment values to each English word and the corresponding Chinese word for eight basic emotions (anger, fear, expectation, trust, surprise, sadness, joy and disgust) and two emotions (negative and positive), and saves them in a csv file for easy searching. We used this sentiment dictionary to analyse the lyrics of the song to find the emotion with the largest sentiment value among all the emotions in the text of the lyrics as the final classified emotion, thus determining the sentiment classification of the song.

The code first read the sentiment dictionary and map the words to their corresponding sentiment values. Then the classify_sentiment() method below counts the number of each sentiment word in the input text and the total sentiment value based on the sentiment value of the word in the dictionary, thus returning the sentiment with the highest sentiment value as the result.

Music recommendation module

The music recommendation module is mainly based on the result of face emotion recognition, the music in the emotion folder corresponding to the recognition result will be played randomly using the method in the random library in python. The next_song() method in this module is also bound to the control next_btn, which is also a method for randomly switching music.

The code first opens the path to the folder where the emotional music of the category is stored based on the results of the face emotion recognition, searches the folder for files ending in '.mp3', '.wav', '.ogg ' or '.m4a' extension, saves the path to one of the files at random, and finally loads and plays the music using the method in the pygame library.

Music playing module

The Music playing module is used to control the playing and pausing of loaded music, and is implemented by the methods provided in the pygame library.

This method uses the Boolean variable ‘playing’ to determine if the music is playing and to control the playing and pausing of the music. When the playing variable is false, the music is not playing and can be played via the button, when the playing variable is true, the music is already playing and can be paused via the button.

4. Result and discussion

4.1. *Under different light conditions*

The tests were first carried out under different lighting conditions and the results are shown in Figs. 8

and 9. As can be seen from the above diagram, under different lighting conditions, Fig. 8 shows the condition of insufficient lighting and Fig. 9 shows the condition of sufficient lighting, it can be seen from the figures that the function of the system expression recognition is not affected by the lighting, but the accuracy of expression recognition may be affected and reduced as a result.

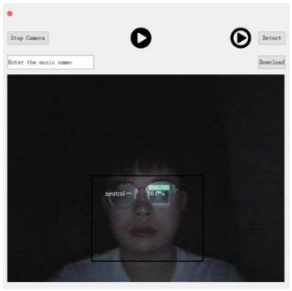

Fig.8. Insufficient light

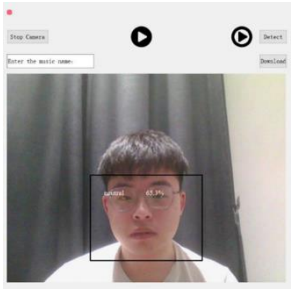

Fig.9. Sufficient light

4.2. *Testing at different distances*

The next test was conducted for different distances, with the face being closer to the camera and further away. The test results are shown in Figs.10 and 11.

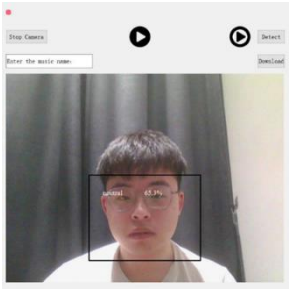

Fig.10. Distance from camera 0.3m

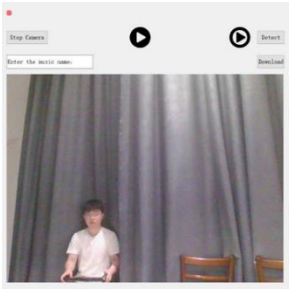

Fig.11. Distance from camera 2m

Fig. 10 shows the results of the test at a distance of 0.3m from the camera and Fig. 11 shows the results of the test at a distance of 2m from the camera. After testing, it was found that the recognition of faces within two metres from the camera would not be affected; however, once the face is more than two metres away from the camera, the recognition will be affected and the face may not be detected, so when using this system, the user needs to be as close to the camera as possible so that the face can be better captured for recognition.

4.3. *Recognition accuracy test*

In this study, the 7 categories of "neutral", "surprise", "happy", "angry", "sad", "fear" and "disgust" were made in turn as shown in Fig. 12. The first five expressions matched the actual expressions, while the last two did not match the actual expressions. The highest recognition accuracy was achieved for the "happy" and "surprise" expressions, with over 90%, and for the "neutral", "sad" and "angry" expressions, with over 60%. The intended "fear" expression was identified as "sad" and the intended "disgust" expression was identified as "fear". The reason for this may be, on the one hand, the low recognition rate of these two expressions by the model, and on the other hand, the ambiguity of these two expressions, which could also be identified as "sad" and "fear" if judged by the human eye.

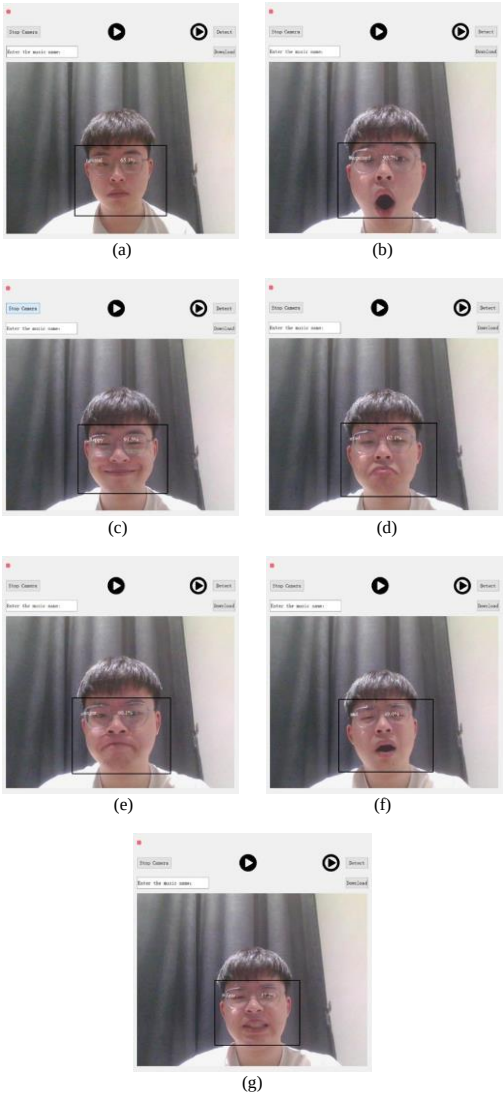

Fig. 12 (a) Neutral, (b) Surprised, (c) Happy, (d) Sad, (e) Anger, (f) Fear, (g) Disgusted expressions

5. Conclusion

Facial expression recognition has application needs in many scenarios, and today more and more fields have introduced expression recognition technology, such as the education field and the medical field. In order to meet the requirements of a facial expression recognition based music recommendation system, a convolutional neural network model based on the FER-2013, JAFFE, CK+ expression datasets was used in this project. A music recommendation system based on facial emotion recognition was successfully designed and implemented in the PyCharm development environment using the python programming language. It mainly includes seven modules: image pre-processing, face detection, emotion recognition, music download, music classification, music recommendation and music playing. After the design was completed, the system was also tested in actual use. The results showed that the system has a high accuracy rate of expression recognition and is highly practical. However, the expression recognition function in this system is still inadequate for the occlusion, side face and distance cases. At a later stage, it is considered to add the expression images for the occlusion and side face cases to the existing dataset to improve the face detection rate and thus the expression recognition rate. In addition, the music classification function of the system designed in this project is limited to classifying music with lyrical content. In the future, the authors will consider improving the music classification algorithm and adding the classification of pure music without lyrics to the music classification module of this system.

References

1. S. Metilda Florence and M. Uma, 'Emotional Detection and Music Recommendation System based on User Facial Expression', IOP Conf. Ser.: Mater. Sci. Eng., vol. 912, no. 6, p. 062007, Aug. 2020, doi: 10.1088/1757-899X/912/6/062007.
2. Song, Y., Dixon, S., & Pearce, M. (2012, June). A survey of music recommendation systems and future perspectives. In 9th international symposium on computer music modeling and retrieval (Vol. 4, pp. 395-410).
3. D. O. Melinte and L. Vladareanu, 'Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer', Sensors, vol. 20, no. 8, Art. no. 8, Jan. 2020, doi: 10.3390/s20082393.
4. K. Kaulard, D. W. Cunningham, H. H. Bülthoff, and C. Wallraven, 'The MPI Facial Expression Database — A Validated Database of Emotional and Conversational Facial Expressions', PLOS ONE, vol. 7, no. 3, p. e32321, Mar. 2012, doi: 10.1371/journal.pone.0032321.
5. B. C. Ko, 'A Brief Review of Facial Emotion Recognition Based on Visual Information', Sensors, vol. 18, no. 2, Art. no. 2, Feb. 2018, doi: 10.3390/s18020401.
6. S. Du, Y. Tao, and A. M. Martinez, 'Compound facial expressions of emotion', Proceedings of the National Academy of Sciences, vol. 111, no. 15, pp. E1454–E1462, Apr. 2014, doi: 10.1073/pnas.1322355111.
7. Kolandaisamy, R., Subaramaniam, K., & Jalil, A. B. (2021, March). A Study on Comprehensive Risk Level Analysis of IoT Attacks. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 1391-1396). IEEE.
8. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, 'Coding facial expressions with Gabor wavelets', in Proceedings Third IEEE International Conference on

Automatic Face and Gesture Recognition, Apr. 1998, pp. 200–205. doi: 10.1109/AFGR.1998.670949.
9. https://github.com/luanshiyinyang/FacialExpressionRecognition

Authors Introduction

Shengke Xie

He has completed his Bachelor of Computer Science (Hons) Mobile Computing and Networking, UCSI University, Malaysia.

Dr. Raenu Kolandaisamy

He received his PhD from the Faculty of Computer Science & Information Technology, University Malaya in 2020. He is currently an Assistant Professor in UCSI University, Malaysia. His research interest areas are Wireless Networking, Security, VANET and IoT.

Dr. Ghassan Saleh Hussein Al-Dharhani

He received his Ph.D. degree in Computer Science from Universiti Kebangsaan Malaysia (UKM). He is currently an Assistant Professor with the Institute of Computer Science and Digital Innovation, UCSI University, Malaysia. His research interests include Artificial Intelligence, Data Mining and Knowledge Discovery, etc.

Dr. Heshalini Rajagopal

She received her PhD and Master's degree from the Department of Electrical Engineering, University of Malaya, Malaysia in 2021 and 2016, respectively. She received the B.E (Electrical) in 2013 Currently, she is an Assistant Professor in UCSI University, Kuala Lumpur, Malaysia. Her research interest includes image processing, artificial intelligence and machine learning.