TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 5, October 2020, pp. 2463~2471
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i5.16717

Development of video-based emotion recognition using deep learning with Google Colab

Teddy Surya Gunawan[1], Arselan Ashraf[2], Bob Subhan Riza[3], Edy Victor Haryanto[4], Rika Rosnelly[5], Mira Kartiwi[6], Zuriati Janin[7]
[1,2]Department of Electrical and Computer Engineering, International Islamic University Malaysia, Malaysia
[1,3,4,5]Faculty of Engineering and Computer Science, Universitas Potensi Utama, Indonesia
[6]Departement of Information Systems, International Islamic University Malaysia, Malaysia
[7]Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia

Article Info

Article history:

Received Jan 17, 2020
Revised Apr 29, 2020
Accepted May 11, 2020

Keywords:

Convolutional neural networks
Deep learning
Emotion recognition
Google Colab
Machine learning

ABSTRACT

Emotion recognition using images, videos, or speech as input is considered as a hot topic in the field of research over some years. With the introduction of deep learning techniques, e.g., convolutional neural networks (CNN), applied in emotion recognition, has produced promising results. Human facial expressions are considered as critical components in understanding one's emotions. This paper sheds light on recognizing the emotions using deep learning techniques from the videos. The methodology of the recognition process, along with its description, is provided in this paper. Some of the video-based datasets used in many scholarly works are also examined. Results obtained from different emotion recognition models are presented along with their performance parameters. An experiment was carried out on the fer2013 dataset in Google Colab for depression detection, which came out to be 97% accurate on the training set and 57.4% accurate on the testing set.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Teddy Surya Gunawan,
Department of Electrical and Computer Engineering,
International Islamic University Malaysia, Malaysia.
Email: tsgunawan@iium.edu.my

1. INTRODUCTION

In the past few years, emotion recognition has become one of the leading topics in the field of machine learning and artificial intelligence. The tremendous increase in the development of sophisticated human-computer interaction technologies has further boosted the pace of progress in this field. Facial actions convey the emotions, which, in turn, convey a person's personality, mood, and intentions. Emotions usually depend upon the facial features of an individual along with the voice. Nevertheless, there are some other features as well, namely physiological features, social features, physical features of the body, and many more. More and more work has been done to recognize emotions with more accuracy and precision. The target of emotion recognition can be achieved broadly using visual-based techniques or sound-based techniques. Artificial intelligence has revolutionized the field of human-computer interaction and provides many machine learning techniques to reach our aim. There are many machine learning techniques to recognize the emotion, but this paper will mostly focus on video-based emotion recognition using deep learning. Video-based emotion recognition is multidisciplinary and includes fields like psychology, affective computing, and human-computer

interaction. The fundamental piece of the message is the facial expression, which establishes 55% of the general impression [1].

To make a well-fitted model for video-based emotion recognition, there must be proper feature frames of the facial expression within the scope. Instead of using conventional techniques, deep learning provides a variety in terms of accuracy, learning rate, and prediction. convolutional neural networks (CNN) is one of the deep learning techniques which have provided support and platform for analyzing visual imagery. Convolution is the fundamental use of a filter to an input that outcome in an actuation. Rehashed use of a similar filter to an input brings about a map of enactments called a feature map, showing the areas and quality of a recognized element in input, for example, a picture. The development of convolution neural systems is the capacity to consequently gain proficiency with an enormous number of filters in equal explicit to a training dataset under the requirements of a particular prescient displaying issue, for example, picture characterization. The outcome is profoundly explicit highlights that can be distinguished anywhere on input pictures. Deep learning has achieved great success in recognizing emotions, and CNN is the well-known deep learning method that has achieved remarkable performance in image processing.

There has been a great deal of work in visual pattern recognition for facial emotional expression recognition, just as in signal processing for sound-based recognition of feelings. Numerous multimodal approaches are joining these prompts [2]. Over the past decades, there has been extensive research in computer vision on facial expression analysis [3]. The objective of this paper is to develop video-based emotion recognition using deep learning with Google collab.

2. VIDEO-BASED EMOTION RECOGNITION USING DEEP LEARNING ALGORITHMS

In this section, details of the general architecture for building a video-based emotion recognition model using a deep learning algorithm is described. Moreover, the architectural diagram, along with various pre- and post-processing processes, are briefly described. The overview of the system using CNN is shown in Figure 1. Before CNN comes into action, the input video has to go through several processes.

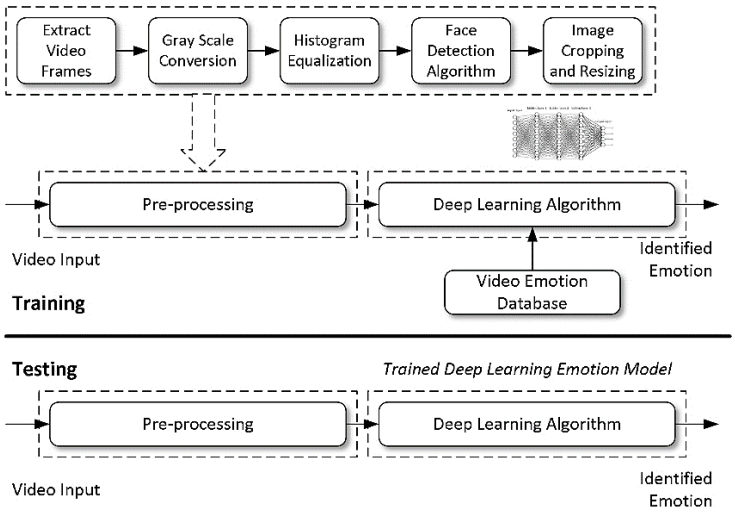


Figure 1. Video-based emotion recognition using deep learning algorithms

2.1. Pre-processing

This is the first process that is applied to the input video sample. Emotions are usually categorized as happy, sad, anger, pride, fear, surprise, etc. Hence, frames are to be extracted from the input video [4]. The number of frames varies for different researchers based on complexity and computational time. The frames are then converted to the grayscale. The frame obtained after gray scaling is somewhat black and white or gray monochrome. The contrast with low-intensity results in grey and that with strong intensity results in white [5]. This step is followed by the histogram equalization of the frames. Histogram equalization is a computer picture handling strategy used to improve contrast in pictures. It achieves this by viably spreading out the most successive intensity esteems, for example, loosening up the intensity scope of the picture. A histogram is a graphical portrayal of the intensity dissemination of a picture. In straightforward terms, it represents the number of pixels for every intensity value considered [6].

2.2. Face detection

Emotions are featured mainly from the face. Therefore, it is crucial to detect the face to obtain facial features for further processing and recognition. Many face detection algorithms are used by many researchers like OpenCV, DLIB, Eigenfaces, local binary patterns histograms (LBPH), and Viola-Jones (VJ) [7]. Conventional algorithms included face acknowledgment work by distinguishing facial highlights by extricating highlights, or milestones, from the picture of the face. For instance, to extricate facial highlights, a calculation may examine the shape and size of the eyes, the size of the nose, and its relative situation with the eyes. It might likewise dissect the cheekbones and jaw. These extracted highlights would then be utilized for looking through different pictures that have matching features. Throughout the years, the industry has moved towards deep learning. CNN has been utilized recently to improve the exactness of face acknowledgment calculations. These calculations accept a picture as information and concentrate a profoundly intricate arrangement of features out of the picture. These incorporate features like the width of the face, the stature of face, the width of the nose, lips, eyes, proportion of widths, skin shading tone, and surface. Essentially, a convolutional neural network separates an enormous number of highlights from a picture. These highlights are then coordinated with the ones put away in the database.

2.3. Image cropping and resizing

In this phase, the face detected by the face detection algorithm is cropped to obtain a broader and clearer look of the facial image. Cropping is the expulsion of undesirable external regions from a photographic or illustrated picture. The procedure, as a rule, comprises of the expulsion of a portion of the fringe regions of a picture to expel incidental rubbish from the image, to improve its surrounding, to change the perspective proportion, or to highlight or disengage the topic from its background. After performing cropping operation on the frames, the size of the images varies. Therefore, to attain uniformity, these cropped images are subjected to resizing, say for our example 80×80 pixels. A digital image is just information numbers showing varieties of red, green, and blue at a specific area on a framework of pixels. More often than not, we see these pixels as smaller than normal square shapes sandwiched together on a PC screen. With a little inventive reasoning and some lower-level control of pixels with code, in any case, we can show that data in a horde of ways. The size of the frame determines its processing time. Hence, resizing is very important to shorten the processing time. Moreover, better resizing techniques should be used to preserve image attributes after resizing [8]. The accuracy of the classification depends on whether the features are well representing the expression or not. Therefore, the optimization of the selected features will automatically improve classification accuracy [9].

2.4. CNN structure with ConvNet

A CNN is a deep learning algorithm that can take in an info picture, allocate significance (learnable loads and bias) to different viewpoints in the picture and have the option to separate one from the other. The pre-preparing required in a ConvNet is a lot of lower when contrasted with other algorithms. While in crude techniques, filters are hand-designed, with enough preparation, ConvNets can get familiar with these qualities. The engineering of a ConvNet is pretty much similar to that of neurons in the human brain and was enlivened by the association of the Visual Cortex. Singular neurons react to improvements just in a confined locale of the visual field known as the receptive field. An assortment of such fields overlaps to cover the whole visual zone [10].

ConvNet is an arrangement of layers, and each layer of a ConvNet changes one volume of initiations to another through a differentiable function. There are three primary kinds of layers to construct ConvNet models: convolutional layer, pooling layer, and fully-connected layer as shown in Figure 2. The general architecture of the ConvNet consists of the following [11]:

- Input [80×80×2] will hold the raw pixel estimations of the picture, right now a picture of width 80, height 80.
- The convolutional layer will evaluate the yield of neurons that are associated with nearby locales in the info, each processing a dot product between their loads and a little area they are associated with the input volume. This may bring about volume, for example, [80×80×12] if we chose to utilize 12 filters.
- RELU layer will be applied for an element-wise actuation work, for example, the max (0, x) thresholding at zero. This leaves the size of the volume unaltered [80×80×12].
- POOL layer will play out a downsampling activity along with the spatial measurements, bringing about volume, for example, [40×40×12].
- Fully connected layer (FC) will process the class scores. The input to this layer is all the outputs from the previous layer to all the individual neurons.

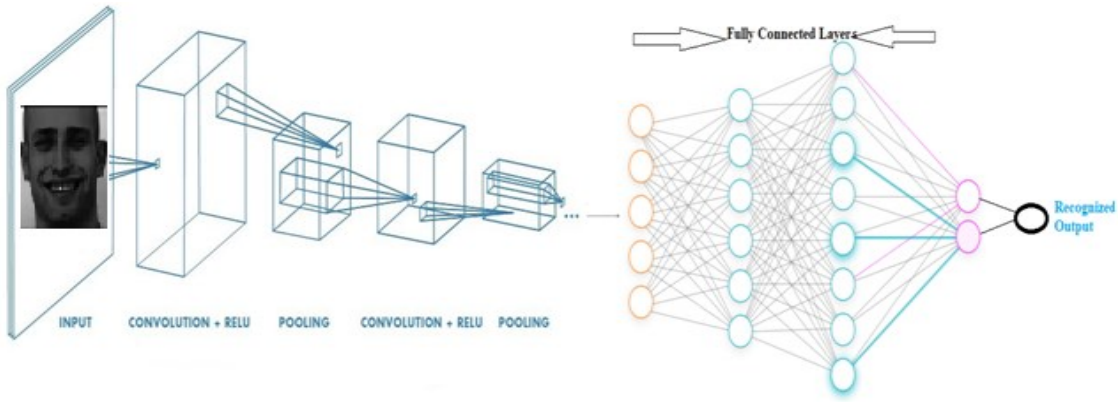


Figure 2. General Architectural structure of convolutional neural networks

3. VIDEO EMOTION DATABASES

This is one of the critical steps in building a model. The data involved in the form of videos should be effective in terms of detailing and usability. There are many datasets available that can be used to frame the video-based emotion recognition models. For example, real-world affective faces database (RAFDB) [12], which gathers real-world pictures from a vast number of people, has been discharged to energize more true research on facial emotion recognition. RAF-DB contains around 12271 training samples and 3068 test data downloaded from the Internet, giving unconstrained articulations under distinctive natural conditions.

Table 1. Video datasets available for emotion recognition

Database	Facial expression	Number of Subjects	Number of images/videos
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	Calm, happy, sad, angry, fearful, surprise, disgust, and neutral.	24	7356 video and audio files
F-M FACS 3.0 (EDU, PRO & XYZ versions)	Neutral, sadness, surprise, happiness, fear, anger, contempt and disgust	10	4877 videos and images sequences
Japanese Female Facial Expressions (JAFFE)	neutral, sadness, surprise, happiness, fear, anger, and disgust	10	213 static images
MMI Database	-	43	1280 videos and over 250 images
DISFA	-	27	4,845 video frames
Multimedia Understanding Group (MUG)	Neutral, sadness, surprise, happiness, fear, anger, and disgust	86	1462 sequences
Indian Spontaneous Expression Database (ISED)	Sadness, surprise, happiness, and disgust	50	428 videos
Real-World Affective Faces Database (RAFDB)	Surprise, fearful, disgusted, happy, sad, angry, fearfully surprised, sadly angry, sadly fearful, angrily disgusted, angrily surprised, sadly disgusted, fearfully disgusted, disgustedly surprised, happily surprised, sadly surprised, fearfully angry, happily disgusted.	-	29672 real-world pictures
Acted Facial Expressions in the Wild (AFEW)	-	-	1809 video samples
ADFES-BIV	Anger, disgust, fear, sadness, surprise, happiness, pride, contempt, embarrassment, and neutral.	-	370 short video samples

The other database, acted facial expressions in the Wild (AFEW) [13], is built up for the emotion recognition in the wild challenge (EmotiW). It comprises of training samples (773), validation samples (383) and test (653) video cuts gathered from TV shows or motion pictures. ADFES-BIV is an augmentation of the ADFES dataset, which was first presented by Van der Schalk et al. [14]. ADFES is acted by 12 North European subjects (five females, seven guys) and 10 Mediterranean entertainers (five females, five guys) communicating the six fundamental feelings in addition to the three complex feelings of contempt, pride, and shame, what is more to impartial.

Wingenbach et al. [15] made the ADFES-BIV dataset by altering the 120 recordings played by the 12 North European entertainers to include three degrees of force. They made three new recordings, showing a similar feeling at three unique degrees of force - low, medium, and high-, for a sum of 360 recordings. Each tape of ADFES-BIV begins with a neutral articulation and closures with the most elevated expressive casing. Another data set is WSEFFEP, which contains 210 high-quality pictures from 30 people [16]. Some of the well-known video databases for emotion recognition are listed in Table 1.

4. RELATED RESEARCH WORKS

Introducing deep learning techniques in the field of emotion recognition with videos, images, voice, or handwritten words as input has achieved a promising result. More and more researchers are developing their interests in their contribution to this field. Figure 3 shows the graphical representation of the rise in the scholarly works in emotion recognition using deep learning [17]. Many researchers have implemented different training architectures of CNN to improve their recognition accuracies. In [18, 19], CNN was trained with many CNN architectures obtained from the pre trained model (ImageNet) as shown in Table 2. The results become more promising with the advent of more and more researches. The video-based emotion recognition models developed to be more and more accurate. The percentage of accuracies of Wingenbach et al. [15] and Sonmez [16] model for different emotions are listed in Table 3.

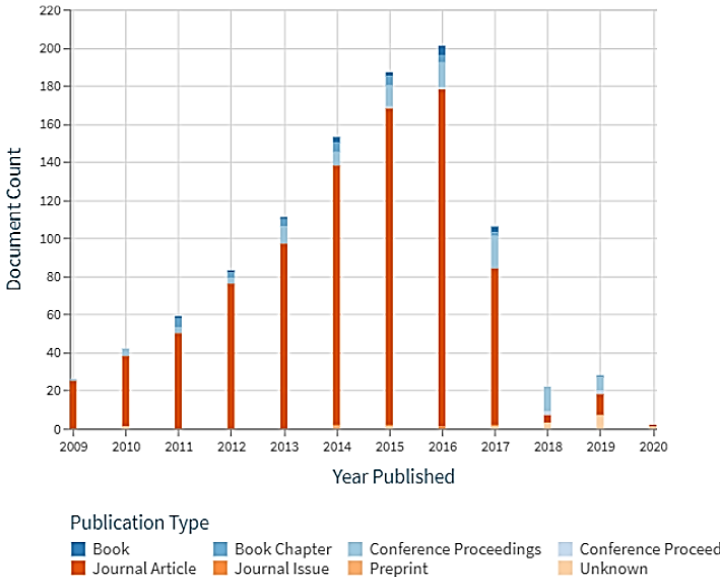


Figure 3. Yearly scholarly works on emotion recognition using deep learning from 2009 to 2020

Table 2. Testing accuracies for CNN architectures

Architecture Used	Pre-Trained	Accuracy
GoogLeNet	With ImageNet	62.96
CaffeNet	With ImageNet	68.05
VGC16	With ImageNet	68.24
Residual Network	With ImageNet	69.65

Table 3. Comparison between Wingenbach et al. [15] and Sonmez [16] accuracies

Emotion	Wingenbach et al. (2016) Accuracy (%)	Sonmez (2018) Accuracy (%)
Happy	84.6	86
Sad	79.3	67
Angry	74.6	94.6
Surprise	92.3	83.3
Disgust	65	97.3
Fear	61.6	61
Neutral	89	36
Pride	42.3	91.6
Contempt	35	50
Embarrassment	64.6	89

5. GOOGLE COLAB IMPLEMENTATION AND RESULTS

5.1. Experimental setup

The experiment was carried on the fer2013 dataset. The data consists of 48×48-pixel grayscale images of faces. This dataset was prepared by Pierre-Luc Carrier and Aaron Courville as a part of the Kaggle Challenge named as Challenges in Representation Learning: Facial Expression Recognition Challenge in 2013. It consists of 28709 train samples and 3589 test samples. It included seven emotions, namely Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral.

The integrated development environment (IDE) used for the process was Google Colaboratory or Google Colab in short. Google Colab is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. With Google Colab, it is possible to write and execute code, save and share our analyses, and access powerful computing resources, all for free from the browser.

The fer2013 dataset was mounted to the Google Colab using Google Drive in the form of a CSV file. After Loading the dataset, the batch size was set to 256, and the training epochs set to 25. The software used was Python 3 with machine learning libraries, including Keras 2.1.6 and Tensorflow 1.7.0. After initializing the training and the testing instances, the data was given to the convolutional neural network (CNN), which consisted of 3 convolutional layers and one fully connected neural network. The model trained for about 4 hours.

5.2. Experimental results

This experiment used 25 epochs for training the data samples. With each epoch, the training accuracy increased while reducing the loss. Performance evaluation is shown in Table 4, while the confusion matrix is shown in Figure 4. While some emotion recognition samples are shown in Figure 5. The testing results were made more accurate by including the Haar cascade face detection process. It is a machine learning object detection algorithm used to identify objects in an image or video. It detected the face from the image to reduce the additional noise. It worked as illustrated in Figure 6.

There was a tremendous increase in efficiency after using Haar cascade on a random test image, which is reflected in Figure 7. It can be found that before using Haar cascade, as shown in Figure 7 (a), the fear emotion is more dominant than happy. While after using Haar cascade, as shown in Figure 7 (b), the only emotion that can be recognized is happy. Therefore, it can be concluded that the Haar cascade improves emotion recognition accuracy.

The scope for future improvements is very appealing in this field. Different multimodel deep learning techniques can be used along with different architectures to improve the performance parameters [20-27]. Apart from recognizing the emotions only, there can be further addition of intensity scale. This might help to predict the intensity of the recognized emotion. Also, multi modals can be used in future works; for example, video and speech can both be used to design a model along with the use of multi-datasets.

Table 4. Performance evaluation based on training and testing accuracy and loss

Performance Parameters	Performance Metrics (%)
Training Loss	0.0948
Training Accuracy	97.07
Testing Loss	2.657
Testing Accuracy	57.509

```
array([[466,   1,   0,   0,   0,   0,   0],
       [ 56,   0,   0,   0,   0,   0,   0],
       [496,   0,   0,   0,   0,   0,   0],
       [895,   0,   0,   0,   0,   0,   0],
       [653,   0,   0,   0,   0,   0,   0],
       [415,   0,   0,   0,   0,   0,   0],
       [607,   0,   0,   0,   0,   0,   0]])
```

Figure 4. Confusion matrix

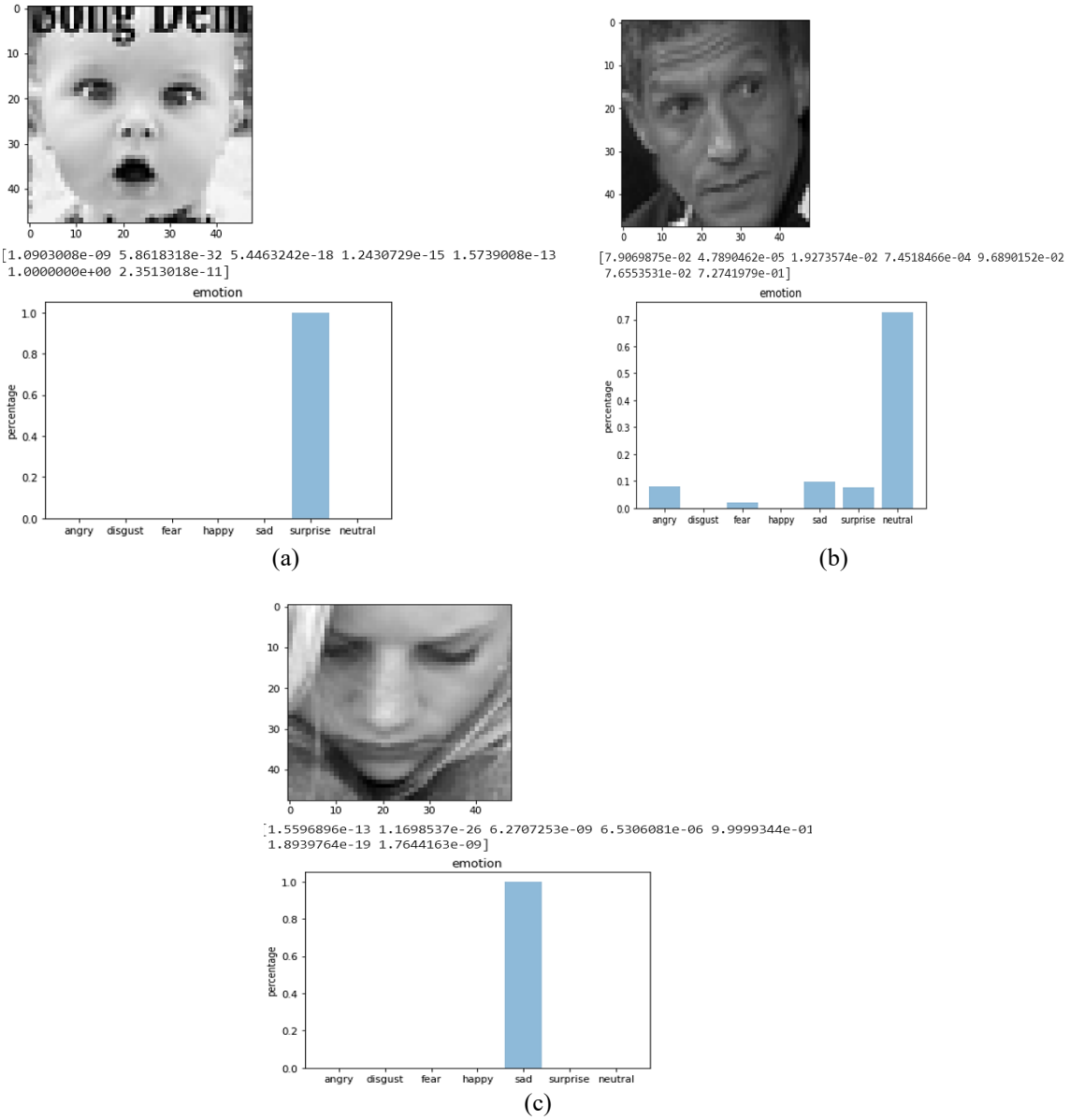


Figure 5. Sample of recognized emotion accuracy percentage graph; (a) surprise emotion, (b) neutral emotion and (c) sad emotion

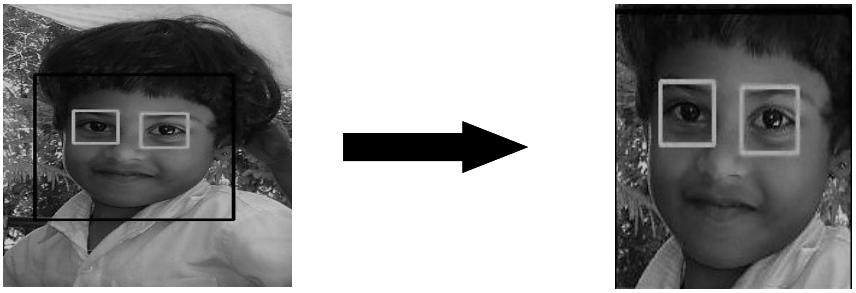
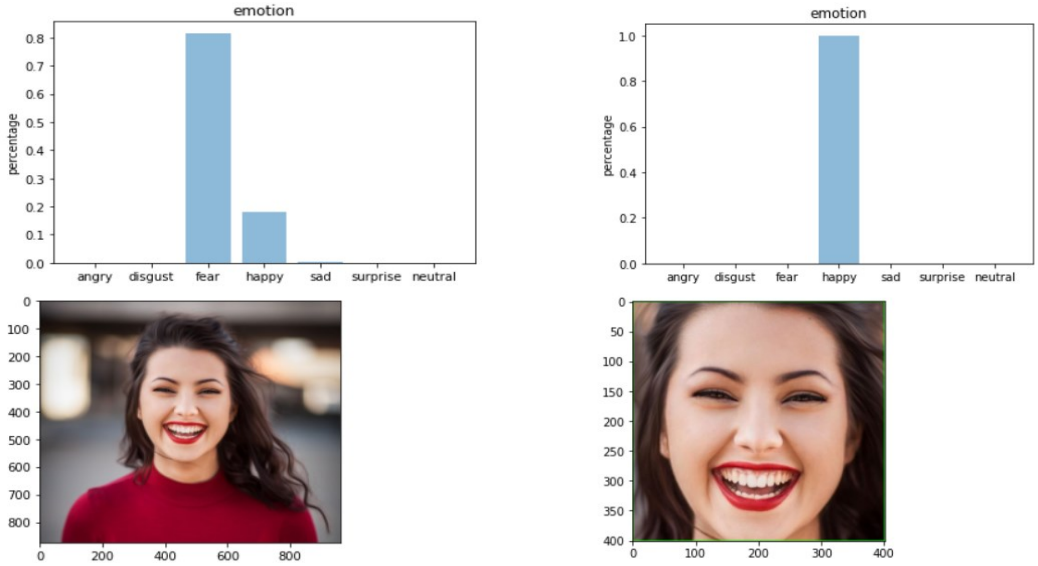


Figure 6. Haar cascade face detection process



(a) (b)

Figure 7. Accuracy improvement using Haar cascade; (a) before Haar cascade and (b) after Haar cascade

6. CONCLUSIONS

This paper presented the development of video-based emotion recognition using deep learning with Google Colab. The success of this approach in recognizing emotions has been tremendously improving over time. Introducing deep learning techniques like CNN, DNN, or other multimodal methods has also boosted the pace of recognition accuracy. Our work demonstrated the general architectural model for building a recognition system using deep learning (more precisely CNN). The aim was to analyze pre and post processes involved in the methodology of the model. There is extensive work done on image, speech, or video as input to recognize the emotion. This paper also covered the datasets available for the researchers to contribute to this field. Different performance parameters were benchmarked on different researches to show the progress in this sphere. The experimental observation was carried out on the fer2013 dataset involving seven emotions, namely angry, disgust, fear, happy, sad, surprise, neutral, which yielded in the be 97% accuracy on the training set and 57.4% accuracy on the testing set when Haar cascade technique is applied.

ACKNOWLEDGMENTS

The author would like to express their gratitude to the Malaysian Ministry of Education (MOE), which has provided research funding through the Fundamental Research Grant, FRGS19-076-0684. The authors would also like to thank International Islamic University Malaysia (IIUM), Universiti Teknologi MARA (UiTM) and Universitas Potensi Utama for providing facilities to support the research work.

REFERENCES

[1] C. H. Wu, J. C. Lin and W. L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.
[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2008.
[3] P. C. Vasanth and K. R. Nataraj, "Facial Expression Recognition using SVM Classifier," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 3, no. 1, pp. 16-20, 2015.
[4] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 14-19, 2019.
[5] S. Johnson, "Stephen Johnson on digital photography," *O'Reilly*, 2006.
[6] Y. C. Hum, K. W. Lai, and M. I. M. Salim, "Multiobjectives bihistogram equalization for image contrast enhancement," *Complexity*, vol. 20, no. 2, pp. 22-36, 2014.

[7] S. Al-Sumaidaee, S. S. Dlay, W. L. Woo, and J. A. Chambers, "Facial expression recognition using local Gabor gradient code-horizontal diagonal descriptor," *Proceedings of the 2nd IET International Conference on Intelligent Signal Processing 2015 (ISP),* 2015.
[8] D. Zhou, X. Shen, W. Dong, "Image zooming using directional cubic convolution interpolation," *IET Image Processing*, vol. 6, no. 6, pp. 627-634, 2012.
[9] H. Boubenna, D. Lee, "Image-based emotion recognition using evolutionary algorithms," *Biologically Inspired Cognitive Architectures*, vol. 24, pp. 70-76, 2018.
[10] I. Wallach, M. Dzamba, and A. Heifets, "AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery," *arXiv*, 2015.
[11] CS231n, "Convolutional Neural Networks for Visual Recognition," 2016. [Online]. Available: http://cs231n.github.io/convolutional-networks/.
[12] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2584-2593, 2017.
[13] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 3, pp. 34-41, 2012.
[14] J. Van Der Schalk, S. T. Hawk, A.H. Fischer, and B. Dooseje, "Moving faces, looking places: Validation of the Amsterdam dynamic facial expression set (ADFES)," *Emotion*, vol. 11, no. 4, pp. 907-920, 2011.
[15] T. S. Wingenbach, C. Ashwin, and M. Brosnan, "Validation of the Amsterdam dynamic facial expression set–bath intensity variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions," *PloS One,* vol. 11, no. 1, pp. 1-1, 2016.
[16] M. Olszanowski, G. Pochwatko, K. Kuklinski, M. Scibor-Rylski, P. Lewinski, and R.K. Ohme, "Warsaw set of emotional facial expression pictures: a validation study of facial display photographs," *Frontiers in Psychology*, vol. 5, 2015.
[17] "Search Articles the Lens - Free & Open Patent and Scholarly Search," The Lens - Free & Open Patent and Scholarly Search. [Online]. Available: https://www.lens.org [Accessed: 15-Feb-2020]
[18] Y. Cai, W. Zheng, T. Zhang, Q. Li, Z. Cui, and J. Ye, "Video Based Emotion Recognition Using CNN and BRNN," *Chinese Conference on Pattern Recognition*, pp. 679-691, 2016.
[19] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445-450, 2016.
[20] T. S. Gunawan, M. F. Alghifari, M. A. Morshidi, and M. Kartiwi, "A review on emotion recognition algorithms using speech analysis," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 6, no. 1, pp. 12-20, 2018.
[21] M.F. Alghifari, T.S. Gunawan, and M. Kartiwi, "Speech emotion recognition using deep feedforward neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, pp. 554-561, 2018.
[22] S. A. A. Qadri, T. S. Gunawan, M. F. Alghifari, H. Mansor, M. Kartiwi, and Z. Janin, "A critical insight into multi-languages speech emotion databases," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1312-1323, 2019.
[23] M. F. Alghifari, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and Z. Janin, "On the use of voice activity detection in speech emotion recognition," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1324-1332, 2019.
[24] T. S. Gunawan, M. Kartiwi, N. A. Malik, and N. Ismail, "Food Intake Calorie Prediction using Generalized Regression Neural Network," *2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA),* pp. 1-4, 2018.
[25] T. S. Gunawan, and M. Kartiwi, "On the Use of Edge Features and Exponential Decaying Number of Nodes in the Hidden Layers for Handwritten Signature Recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 722-728, 2018.
[26] E. Ihsanto, K. Ramli, D. Sudiana, and T.S. Gunawan, "An Efficient Algorithm for Cardiac Arrhythmia Classification Using Ensemble of Depthwise Separable Convolutional Neural Networks," *MDPI Applied Sciences*, vol. 10, no. 2, pp. 1-16, 2020.
[27] E. Ihsanto, K. Ramli, D. Sudiana, and T. S. Gunawan, "Fast and Accurate Algorithm for ECG Authentication using Residual Depthwise Separable Convolutional Neural Networks," *MDPI Applied Science*, vol. 10, no. 9, pp. 1-15, 2020.