Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 3, June 2020, pp. 1208~1219
ISSN: 2302-9285, DOI: 10.11591/eei.v9i3.2030

Deep learning based facial expressions recognition system for assisting visually impaired persons

Hendra Kusuma, Muhammad Attamimi, Hasby Fahrudin
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

Article Info

Article history:

Received Aug 6, 2019
Revised Nov 3, 2019
Accepted Dec 4, 2020

Keywords:

Deep learning
Facial expressions recognition
Visual recognition
Visually impaired persons
Wearable devices

ABSTRACT

In general, a good interaction including communication can be achieved when verbal and non-verbal information such as body movements, gestures, facial expressions, can be processed in two directions between the speaker and listener. Especially the facial expression is one of the indicators of the inner state of the speaker and/or the listener during the communication. Therefore, recognizing the facial expressions is necessary and becomes the important ability in communication. Such ability will be a challenge for the visually impaired persons. This fact motivated us to develop a facial recognition system. Our system is based on deep learning algorithm. We implemented the proposed system on a wearable device which enables the visually impaired persons to recognize facial expressions during the communication. We have conducted several experiments involving the visually impaired persons to validate our proposed system and the promising results were achieved.

Corresponding Author:

Muhammad Attamimi,
Department of Electrical Engineering,
Institut Teknologi Sepuluh Nopember,
Raya ITS, Sukolilo 60111, Surabaya, Indonesia.
Email: attamimi@ee.its.ac.id

1. INTRODUCTION

In everyday life, the visually impaired persons face many challenges such as finding a path in unknown environments as well as visually recognizing their surroundings including the objects, human body or face. Many studies have been done to enable the computer and/or the machine to see the world as human does through developing the algorithm in visual recognition systems including object detection, object recognition, and so forth [1-15]. One of the ability in visual recognition is to recognize the facial expressions of a person during the interaction. Generally, non-verbal information such as body movements, gestures, facial expressions; takes a great role in a face-to-face interaction including communication [16-18]. Moreover, the facial expression is strongly related to human emotion that indicates the inner state of the speaker and/or the listener during communication. Such information is important in engaging a good communication as well as interaction with the others. The visually impaired persons that in general have difficulties to obtain such information having the limitation in social interaction [19]. Especially, the ones who lost vision early in their life will have more difficulties in social interaction [20]. To the best of our knowledge, there are only few assistive devices available for supporting visually impaired persons in capturing non-verbal information during social interaction.

Without the ability to see, human can process the information which usually obtained through sensing using the other senses such as tactile and/or auditory [21-23]. The textual information such as text can be delivered in the auditory form or touching a medium designed to convey such information. One knows

that the braille is commonly used to deliver textual information for the visually impaired persons. This fact inspires us to develop a wearable device that can capture non-verbal information for assisting the visually impaired persons. In this study, we focus on facial expressions recognition as one of the important non-verbal information during the social interaction. We propose a facial expression recognition system which is based on deep learning. The proposed system is implemented on a wearable device shown in Figure 1. The device is equipped to the visually impaired persons. Thanks to proposed system, the input image captured from the device can be classified into facial expressions which consist of anger, disgust, fear, neutral, happy, surprised, and sadness. To deliver the results to the users which are visually impaired persons, it is natural to consider a touch or an auditory information as a choice. We have investigated and chosen the auditory information as the final output of our wearable device.

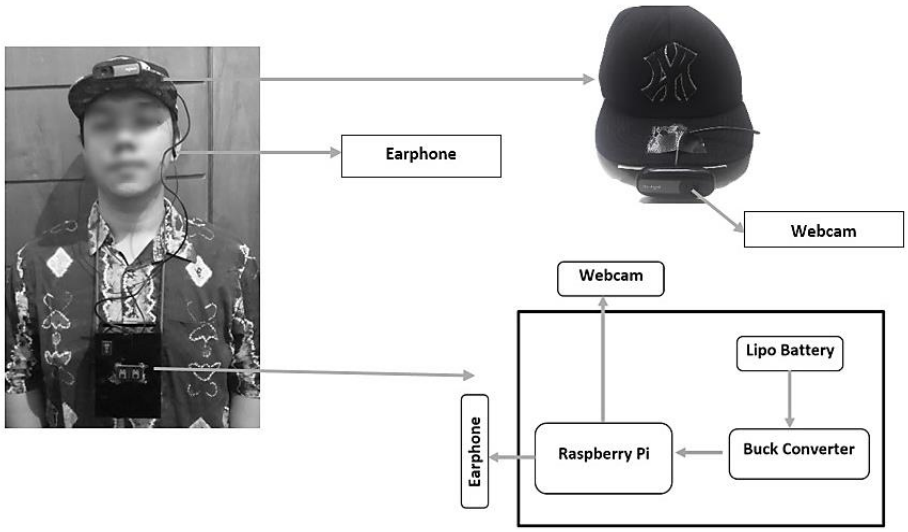


Figure 1. A wearable device developed in this study for assisting visually impaired persons

2. RESEARCH METHOD

This study focuses on proposing a facial expression recognition system and implementing the proposed system in a wearable device to assist the visually impaired persons. There are several related studies. First, the studies related with the assistive tools for visually impaired persons. For examples, a vibrotactile has been used in [21-22] and sound has been used in [23] for conveying information to the users. Particularly the study in [21] developed a wearable device to convey the facial expressions to the haptic information via a vibrotactile belt. In [21], the device consisted of a Microsoft Surface Pro 4 tablet (6th Gen 2.2-GHz Intel Core i7-6650U processor with Intel Iris graphics 540, Windows 10 operating system) and a webcam mounted on a cap to capture the image. FaceReader 6™ (Vicar Vision, Amsterdam, The Netherlands) was used for face and facial expression recognition. The integrated system was excellent, but still difficult to realize a cheap device. Therefore, we try to develop a cheaper wearable device that can support the visually impaired persons.

Second, the studies related with the datasets for facial expressions recognition [24-26]. The Cohn-Kanade (CK) dataset aims to promote research on the detection of individual facial expressions automatically. CK dataset has become one of the datasets commonly used for the development of algorithms and their evaluations. However, according to [24], CK dataset still has some limitations and the Extented Cohn-Kanade (CK+) provided the solutions to overcome those limitations. In [24], the number of sequences increased by 22% and the number of subjects increased by 27%, and the target expressions for each sequenced coded FACS and emotional labels have been validated and revised. Another dataset is the AffectNet [25]. AffecNet is a database created by querying different search engines (Google, Bing, and Yahoo) using 1250 emotional-related tags in six different areas (English, Spanish, Portuguese, German, Arabic, and Farsi). AffectNet has more than one million images with faces. Twelve experts annotated each 450,000 images in terms of categories and dimensions. To calculate the similarity of perception, 36,000 images were annotated by two people. AffectNet has so far been the largest dataset of facial expressions on static images that include both category and dimension models.

In this study, we combine the mentioned datasets with the Japanese Female Facial Expressions (JAFFE) dataset [26], to realize the adaptable facial expression recognition system. JAFFE dataset consists of 213 images of seven facial expressions (six basic facial expressions and neutral) with 10 Japanese women as the model. Each image is rated based on six emotions from 60 subjects. JAFFE was designed and created by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. Photo taken at the psychology department of Kyushu University. Each expressor was shot while looking at the semi-reflective plastic sheet towards the camera. Hair was tied away from the face to reveal all regions of expression on the face. The tungsten lamp was positioned to add illumination to the face. A box covered the area between the camera and plastic to reduce back reflection. The JAFFE dataset has been used as research data to prove cultural differences in the interpretation of facial expressions. In [27], there was a finding that there were differences in the interpretation of facial expressions for each person who has a different cultural identity.

Third, the studies that related to image recognition, particularly, the methods for image classification. There are two paradigms in image classification that are feature based ones such as in [12-15], [28, 29] and end-to-end ones which is popular with neural network [30]. The latter paradigm become powerful since deep learning had been found. One popular method in deep learning is Convolutional Neural Network (CNN) and had been proven to outperform others method and become the state-of-the-art in visual recognition [31]. However, this method needs high computational power. Fortunately, there are some frameworks that help us to implement algorithm in a low computational power device by reduce it performance to certain degree [32]. Those problems could be anticipated by training deep learning model in cloud service.

3. PROPOSED METHOD

In this study, we propose a facial expressions recognition system which can classify seven human expressions that are sadness, anger, happy, disgusted, fear, surprise, and neutral expression. Our proposed method is based on deep learning framework through transfer learning using publicly available model based on Convolutional Neural Networks (CNN). Figure 2 shows the illustration of our proposed method. The mechanism of transfer learning is similar to our previous work in [33] which is by collecting datasets consist of seven types of facial expressions and training the publicly available model with the new dataset. Since model building is time consuming, the training phase is conducted offline. Once the model is trained, we can use it for facial expressions recognition.

One can see from Figure 2 that the proposed method is divided into two blocks which are preprocessing for input data and the prediction phase. To predict the preprocessed input image, a pretrained networks is used. The proposed method is then implemented on a wearable device shown in Figure 1. The main processor of our device is Raspberry Pi 3, which is able to control the camera mounted on the cap to capture the images and process them to output the recognition results. Thanks to our proposed device the facial expressions results can be conveyed into auditory information which is allow the visually impaired persons to understand the expressions during the social interaction. The details of proposed method is explained as follows.

3.1. Implementation of a wearable device

In this paper, we also implement a facial expression recognition system to a wearable device as shown in Figure 2. We use auditory information instead of vibration pattern information. Visually impaired persons obtain the information of facial expressions of their partner by remembering the tone of each expression that have been assigned beforehand. We consider that auditory information can be conveyed facial expressions simpler than the one using a vibration. We used earphone as a tool to convey auditory information because it has relatively small size and available in cheap price. Raspberry Pi 3 used as the main processing unit because it already compatible with the latest deep learning frameworks, so we can implement our trained network without any compatibility issues. A webcam is responsible on providing a raw input data for trained network frame by frame. The webcam will be placed on a cap to be more precisely parallel to the eye sight. There are push buttons that serve as pause, play, and shutdown command. We decide to add this feature to help visually impaired persons to use the device independently. The user can pause the recognition system when taking a break instead of turning off the power which is not energy efficient. Raspberry Pi 3, push buttons, and the power supply will be placed in a small box. With this design, we can predict at least once in 1.2 seconds which is not very fast but enough to maintain a good conversation.

3.2. Data Preprocessing

In Figure 2, data preprocessing block have several steps. Data preprocessing block is used to maintain consistency and quality of the raw input data. As one can see, in the first step we convert the RGB

image into grayscale image. There are two reasons we used grayscale image, i.e., 1) the computational cost is low since the data is reduced in 1/3 (using one channel instead of three channels), and 2) to identify facial expression we only need shape, texture, and depth of human face that can be extracted even using the grayscale image.

We used boosted-cascade to localize human face in captured image [34]. Boosted-cascade has a good accuracy with less computational cost. Capturing picture of an object in the wild usually did not have good lighting condition which can cause decrease of trained network accuracy. Therefore, by applying histogram equalization method can adjust contrast of the image, so the main face feature will not be damaged or overlooked.

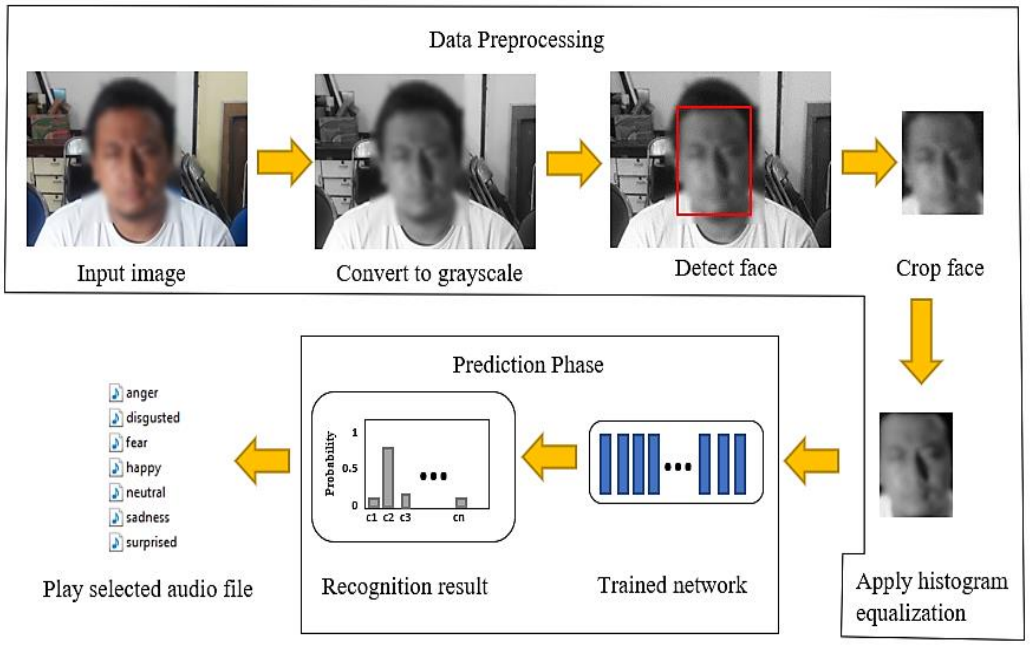


Figure 2. Overview of proposed facial expressions recognition system

3.3. Deep Learning for Facial Expressions Classification

We used transfer learning method to develop our own facial expressions recognition model. In this study, one of our goals is to reduce cost production, so using a cheap microcomputer is one of the solutions. Cheaper devices generally have low computational power. Therefore, we need to use a lightweight model. In this paper, we use MobileNetV2 as a pretrained network due to its excellent performance while provides very efficient mobile-oriented model [35]. The intuition is that the bottleneck encodes intermediate its input and output while inner layer designed to allow model transform from the pixel information to general descriptors such as image category. MobileNetV2 use residual connection in their model architecture to increase performance while enabling shortcuts to increase model speed.

In this study, we combined CK+ dataset [24], AffectNet dataset [25], and JAFFE dataset [26] for training and test sets. We divide the dataset based on six basic human emotion which are anger, sadness, disgust, happy, surprise and fear with addition of neutral expression. In Figure 2, we processed our image in grayscale format to maintain input quality. To maintain the consistency among the datasets, we processed all the data in grayscale format. The data needs to be checked manually by human before split into training data and validation data to assure data integrity. Manual check guidance will be based on [24-26]. On the training phase, we focused on learning rate, epoch, and dropout layer. We believe the parameter tuning can greatly improve networks performance. Those parameters manually tuned according to experiment results.

3.4. Outputs

Prediction results are presented in the form of audio files which each of facial expression is assigned to different audio files that are made by simple pattern of three different tones called audio_1, audio_2, and audio_3. Instead of using seven kind of tone to address each expression, using simple pattern with fewer kind of tone will be easier to remember. Table 1 shows the audio representation of each facial expression.

Table 1. Auditory information

Facial expression	Audio representation
Anger	Playing audio_1 three times
Disgust	Playing audio_1 two times
Fear	Playing audio_1
Neutral	Playing audio_2
Happy	Playing audio_3
Surprise	Playing audio_3 two times
Sadness	Playing audio_3 three times

4. EXPERIMENTS AND DISCUSSION

To evaluate our proposed facial expressions recognition system, we have conducted several experiments. First, the experiment is to compare and report the result of transfer learning using the MobileNetV2 model [35], Xception model [36], and VGG16 model [37]. Here, we compared the computational cost and performance of those models. Next, we have chosen the best model and performed parameter tuning to get a better model. Finally, we have implemented the proposed facial expressions recognition system to a wearable device and tested the performance of a whole system. The details of experimental settings and the results are described as the following.

4.1. Experimental setup

One of major factors of the quality in the deep learning based algorithm is the dataset. The more representative the dataset used, the better the model's performance. In this study, we used the combination of three publicly available dataset, i.e., CK+ dataset [24], AffectNet dataset [25], and JAFFE dataset [26]. Combining the datasets was needed for training the deep learning model implemented on the proposed system. The images on the JAFFE dataset and AffectNet dataset have been separated according to the type of emotions as many as seven categories, i.e., sadness, happy, fear, disgust, anger, surprise, and neutral by storing images in the folder whose title matches the facial expressions. However, we needed to extract the image in the CK+ dataset because it was provided in the form of a video. Therefore the image needed to be selected manually and sorted according to emotion category. The total image on the combined dataset was approximately 40,000 images. The images were then divided into 32,000 images for a training set and 8,000 images for a test set.

The next step is to separate the face image with the background. This aims to ensure the learning process the model will focus on studying features on the face. To ensure the quality of the dataset, we have checked the image one by one. Images with inappropriate facial expressions will be labeled accordingly. The image that contained ambiguous facial expression was deleted. At this stage, human perception was needed in classifying facial expressions. Figure 3 shows some examples of images in the dataset with the corresponding labels. Once the dataset preparation was done we performed the training task. To realize a better implementation of facial expressions recognition system, we need to choose a best model according to the computational cost and the performance. Section 4.2. discusses the model comparison used in this study. After the best model was found, the model need to be tuned to obtain a better performance. The results of parameter tuning is discussed on Section 4.3.

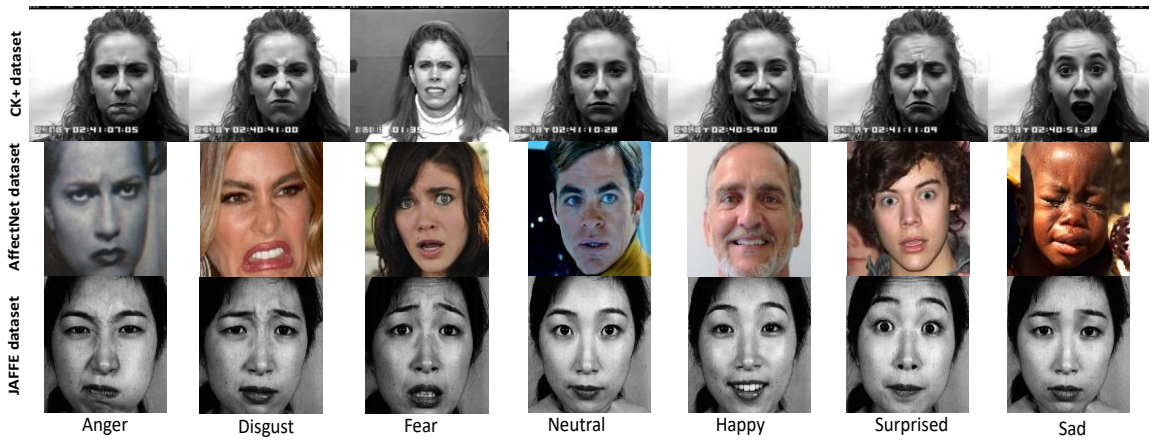


Figure 3. The datasets used in this paper

Finally, since the ultimate goal of this study is to assist the visually impaired persons recognizing the facial expressions of their speaking partners in daily activities, we conducted the experiment to evaluate our wearable device in indoor and outdoor environments. We tested our proposed system on visually impaired users. Figure 4 shows the experimental scenes in indoor and outdoor environments participated by visually impaired subjects. From the Figure 4, the person using proposed device guessed the facial expressions of his/her partner in some communication scenarios. The distance between the speaker and listener was one meter. It should be note that in ice breaking phase, we trained the user for about 10 minutes to obtain the information from auditory information outputed from proposed device. We then compared the correct information collected from the facial expressor with one come from the user equipped with the wearable device. This experiment participated by three facial expressors. Each expressor was performed seven types of facial expressions for five times. Due to the limitations, we have only invited three visually impaired persons to use our wearable device and participated in our experiment. We reported the result on Section 4.4.

Figure 4. Experimental scenes in, (a) Indoor environment, and (b) Outdoor environment

4.2. Model comparison

In this study, we compared three publicly available deep learning models that are MobileNetV2 model [35], Xception model [36], and VGG16 model [37] in two aspects, i.e., computational speed and performance. Here, using the training set mentioned in Section 4.1, we trained the models via transfer learning and compared the accuracies of test set. In general, to train a deep learning model, a personal computer with powerful Graphical Processing Unit (GPU) is needed. We used a 3.0-GHz Intel Core i5-7400 processor with GeForce GTX 1060 6 GB and Linux Mint 19 operating system to create a model to be used in our proposed wearable device that is based on Rasberri Pi 3.

Table 2 shows the comparison results in a personal computer whereas Table 3 shows the results in Rasberri Pi 3. From Table 2, the difference in accuracy and speed was not too significant. If the model speed is less than 0.1 seconds, then the model is enough to predict the images continuously in a social interaction scenario. There is no significant difference in accuracy of the three architectures. On the other hands, one can see from Table 3 that there was a significant difference in speed. In the aspect of accuracy, the three models have relatively the same results. Therefore, MobilenetV2 was suitable for implementation on proposed wearable device because it was the fastest model in prediction the input images.

Table 2. The results on personal computer

Model architecture	Speed [seconds/predictions]	Accuracy [%]
VGG16	0.05123	58
Xception	0.03244	62
Mobilenet V2	0.01982	60

Table 3. The results on Rasberri Pi 3

Model architecture	Speed [seconds/predictions]	Accuracy [%]
VGG16	9.132	53
Xception	11.983	56
Mobilenet V2	1.2783	55

4.3. Parameter tuning

Tuning the parameters of deep learning models is important process to obtain a robust model. In this study, the parameters to be tuned were learning rate and momentum. Using the training set, we trained several times by changing the momentum parameters to find the optimal value. It should be note that each model was also tested using the test set. We also conducted parameter tuning for learning rate to obtain the optimal value. First, we tuned the momentum value. The momentum can help to know the direction of change of the parameters that refer to the results of the previous step and can also prevent oscillations on gradient descent. Figure 5 shows the results of model accuracy and model loss when the momentum values of 0.3, 0.5, and 0.7 were set to the model. Based on the figure, it can be concluded that changing momentum values can significantly influence the learning process. At values of 0.9 and 0.7, the accuracy of the training set increased to reach 0.9 on 15th epoch and changes in the test set were less significant. Loss from training set at this value (0.9 and 0.5) increased with the learning session progressing. This causes no increase in accuracy in the test set. From this tendency it can be concluded that the values of 0.9 and 0.7 were not suitable and fallen in overfitting. At 0.5 momentum value the accuracy of training set increased gradually but in the same epoch the training session accuracy of the momentum value of 0.5 was not as high as the other values (0.9 and 0.7). However, the accuracy in test data was tended to increase. Loss in the test set decreased with the learning session progressing. From these trends, it can be concluded that the momentum value suitable for this application was approximately 0.5.

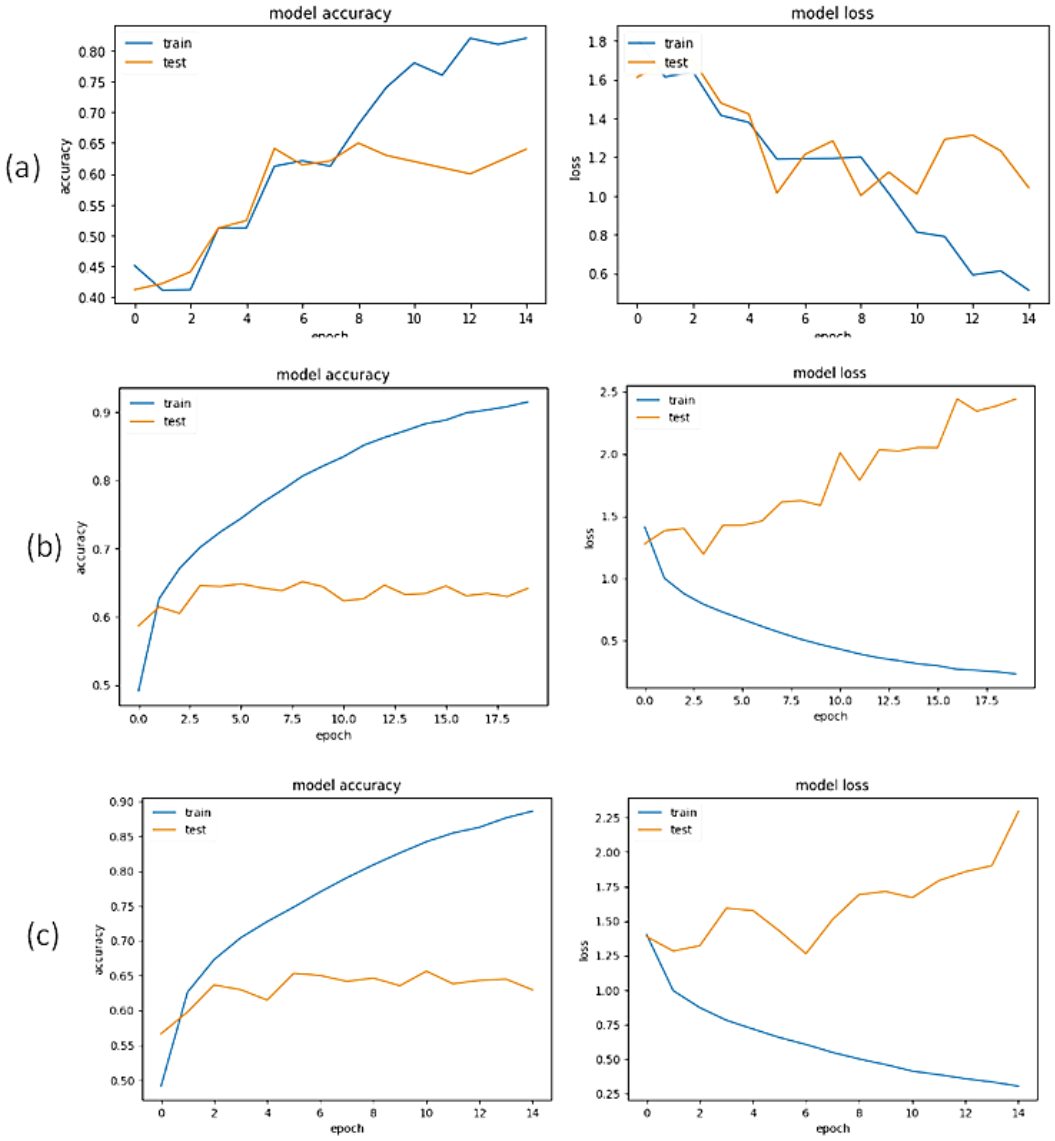


Figure 5. Model accuracy and model loss using the momentum value of: (a) 0.5, (b) 0.7, and (c) 0.9

Next, after selecting the momentum value, the parameter tuning was made on learning rate. The learning rate determines how quickly the model updates the model parameters. By finding the appropriate learning rate value, the model can learn input features well and quickly. Figure 6 depicts the results of model accuracy and model loss when the learning rate values of 0.001, 0.0001, and 0.00001 were set to the model. Based on the figure, changes in the learning rate value can affect the learning session. At a value of 0.001, the model has difficulty recognizing the test set. It can be seen in the training history, the value of 0.001 has a very high loss even though it has high training accuracy and low training loss. From these characteristics, it can be concluded that the model experiences overfitting when using a learning rate of 0.001. The values of 0.0001 and 0.00001 have the same tendency, that is the accuracy of training set and test set shows an increase. As the learning session progressed, the loss of test set and training set have decreased. However, the value of 0.00001 required a longer learning session to match the ability of the model at a value of 0.0001. When observed in more detail, the accuracy of training set shows stagnation in the last few epochs. So, if we want to match the performance of the model with a learning rate value of 0.0001 it will take a very long time. Therefore, the suitable value of the learning rate was approximately 0.0001.

Finally, after finding the optimal parameters that are 0.55 for momentum and 0.0004 for the learning rate, we trained our models with those parameters. Figure 7 shows the model accuracy. One can see that the accuracy of the training set almost reached 99%. It should be noted that the accuracy of the test set has exceeded the minimum limit of close to 85%. Therefore, the model of the learning session can be used as a final model that will be implemented on a wearable device.

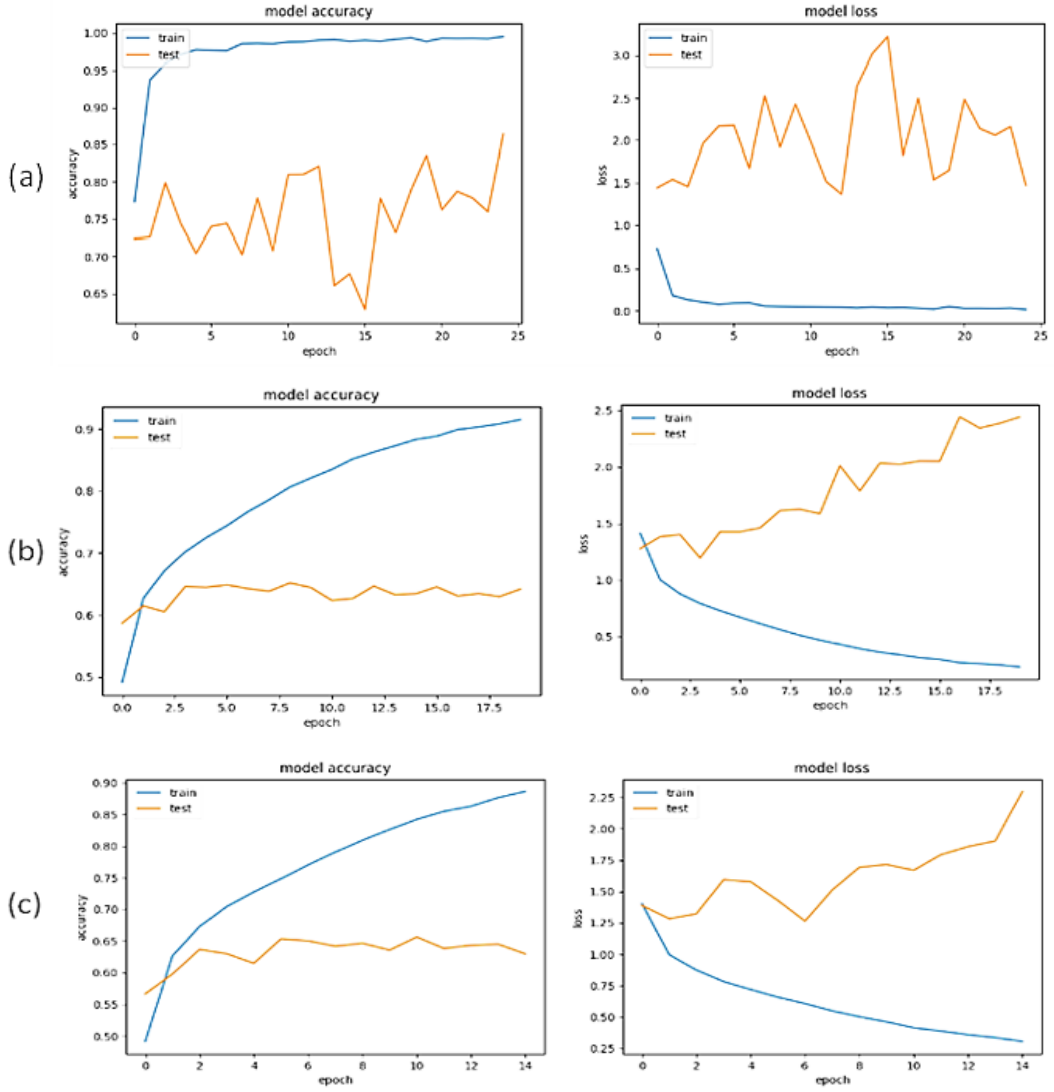


Figure 6. Model accuracy and model loss using the learning rate value of, (a) 0.001, (b) 0.0001, and (c) 0.00001

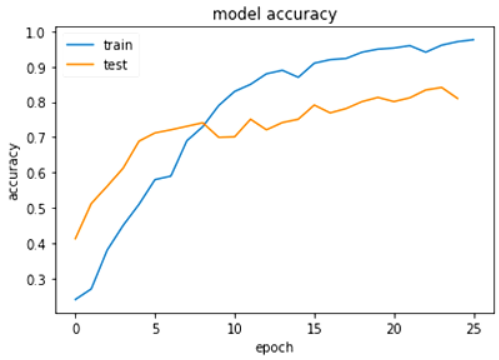


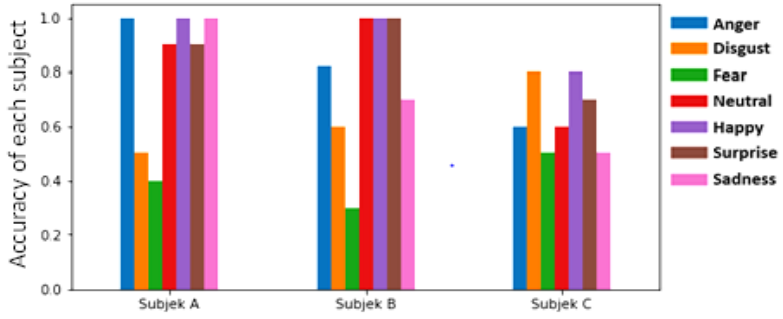
Figure 7. Model accuray using the optimal momentum and learning rate

4.4. Evaluation results of the whole system

In this paper, to validate the whole system we tested on visually impaired persons. The result of indoor experiment and outdoor experiment were respectively shown in Figure 8 and Figure 9. One can see from Figure 8(a), subjects A and B have success rates relatively the same. However, there was a significant difference with the subject C. To investigate the results, we have also interviewed the subject C. The subject was explained that the subject did not really understand the explanation during the ice breaking. However, subject C felt more comfortable after taking data on the second expressor. The confussion matrix of the result was shown in Figure 8(b). From the figure, one can see that the subjects can guess the facial expressions correctly by accuracy of 80% with natural room conditions without any additional lighting. Fear expression was the biggest facial expressions that have errors. On expression disgusted, the error of the answer was always in an angry expression. This can be considered that the expression of anger and disgust having similar facial features or the user was not precise in digesting auditory information because the angry expression has a sound pattern that was almost the same as the expression of disgust (see Table 1). Other expressions have errors that can be tolerated. In general, the device can be said to work properly.

One can see from Figure 9(a), subjects A, B, and C have almost the same trend on the accuracy of predictions. The fear expressions have the lowest success rates. Neutral expressions and smiles are always ranked first on the percentage of success predictions. From Figure 9(b), the user can guess the facial expressions correctly by 78% in an outdoor environment with no additional lighting.

(a)



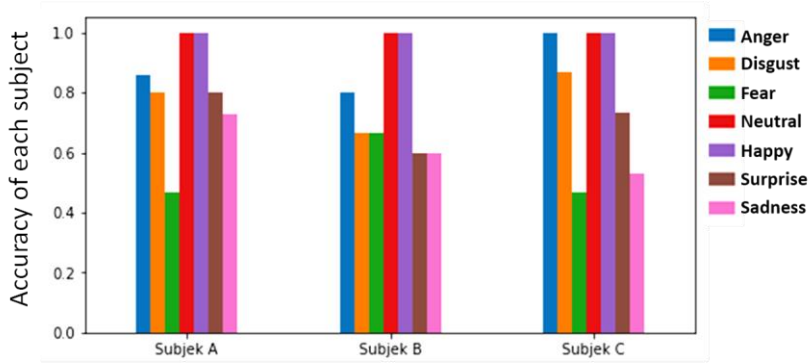
(b)

Prediction results

Correct expressions	Anger	Disgust	Fear	Neutral	Happy	Surprised	Sadness
Anger	1	-	-	-	-	-	-
Disgust	0.44	0.56	-	-	-	2	-
Fear	-	-	0.32	0.16	0.32	0.2	-
Neutral	-	-	-	0.96	0.04	-	-
Happy	-	-	-	-	1	-	-
Surprised	-	0.04	-	-	0.04	0.92	-
Sadness	-	-	-	0.16	-	-	0.84

Figure 8. The results of indoor experiment, (a) Success rates of each subjects, (b) Confussion matrix of facial expressions recognition

(a)



(b)

Prediction results

Correct expressions	Anger	Disgust	Fear	Neutral	Happy	Surprised	Sadness
Anger	0.889	0.112	-	-	-	-	-
Disgust	0.112	0.778	-	-	0.044	-	0.066
Fear	0.177	0.112	0.533	-	0.112	-	-
Neutral	-	-	-	1	-	-	-
Happy	-	-	-	-	1	-	-
Surprised	-	0.066	-	0.066	0.1334	0.711	-
Sadness	-	0.377	-	-	-	-	0.622

Figure 9. The results of outdoor experiment, (a) Success rates of each subjects, and (b) Confussion matrix of facial expressions recognition

5. CONCLUSION

In this paper, we have proposed the facial expressions recognition system. Our proposed system was based on deep learning framework utilizing the transfer learning of available models. To boost the accuracy of proposed method, we have combined several publicly available datasets and performed comprehensive parameter tuning. We have also implemented the proposed system on a wearable device. The device was designed to assist the visually impaired persons in social interaction. We have used the auditory information to convey the results of facial expressions predicted during the social interaction. To validate our proposed system, we have invited visually impaired persons to have conversations with some facial expressors in indoor and outdoor environments. The results were 80% for indoor environment and 78% for outdoor environment. These results were good enough for our proposed wearable device and can be considered assisting the visually impaired persons in social interaction. In the future, we are planning to improve this device in accuracy, ergonomics aspects, and power management. Therefore, to improve accuracy we plan to collect more data and possibly make our own dataset. We also plan to a study about characteristic of basic human facial expressions which can help us building a better facial expression dataset.

REFERENCES

[1] T-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020.
[2] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," *IEEE Xplore*, pp. 9308-9316, 2018.
[3] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4203-4212, 2018.
[4] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *ArXiv: 1804.02767Comment: Tech Report*, pp. 1-6, 2018.
[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517-6525, 2017.
[6] J. Dai et al., "Deformable convolutional networks," *IEEE Xplore*, pp. 764-773, 2017.
[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.
[8] H. Bilen et al., "Weakly supervised deep detection networks," *Conference on Computer Vision and Pattern Recognition*, pp. 2846-2854, 2016.
[9] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "SoftNMS improving object detection with one line of code," *International Conference on Computer Vision*, pp. 5562-5570, 2017.

[10] Z. Cai and N. Vasconcelos, "Cascade RCNN: Delving into high quality object detection," *Conference on Computer Vision and Pattern Recognition*, pp. 6155-6162, 2018.
[11] K. Chen et al., "Hybrid task cascade for instance segmentation," *Conference on Computer Vision and Pattern Recognition*, pp. 4974-4983, 2019.
[12] M. Attamimi, T. Araki, T. Nakamura, and T. Nagai, "Visual recognition system for cleaning tasks by humanoid robots," *International Journal of Advanced Robotic Systems*, vol. 10, no. 11, pp. 1-14, 2013.
[13] M. Attamimi, T. Nagai, and D. Purwanto, "Particle filter with integrated multiple features for object detection and tracking," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 16, no.6, pp. 3008-3015, 2018.
[14] M. Attamimi et al., "Real-Time 3D visual sensor for robust object recognition," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4560-4565, 2010.
[15] M. Attamimi, T. Nakamura, and T. Nagai., "Hierarchical multilevel object recognition using markov model," *21st International Conference on Pattern Recognition*, pp. 2963-2966, 2012.
[16] M. Attamimi, Y. Katakami, K. Abe, T. Nakamura, and T. Nagai, "Modeling of honest signals for human robot interaction," *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 415-416, 2016.
[17] K. Abe, C. Hieida, M. Attmimi, and T. Nagai, "Toward playmate robots that can play with children considering personality," *2nd International Conference on Human-Agent Interaction*, pp. 165-168, 2014.
[18] M. Attamimi, M. Miyata, T. Yamada, T. Omori, and R. Hida, "Attention estimation for child-robot interaction," *Proceedings of the Fourth International Conference on Human Agent Interaction*, pp. 267-271, 2016.
[19] M. G. Frank., "Facial expressions," *International Encyclopedia of the Social & Behavioral Sciences*, pp. 5230-5234, 2001.
[20] M. D. Naraine, D. Mala, and P. H. Lindsay, "Social inclusion of employees who are blind or low vision," *Disabil Soc.*, vol. 26, no. 4, pp. 389-403, 2011.
[21] H. P. Buimer et al., "Conveying facial expressions to blind and visually impaired persons through a wearable vibrotactile device," *PLoS One*, vol. 13, no. 3, pp. 1-16, 2018.
[22] Y. Jin, J. Kim, B. Kim, and R. Mallipeddi, "Smart cane: Face recognition system for blind," *Proceedings of the 3rd International Conference on Human-Agent Interaction*, pp. 145-148, 2015.
[23] Y. Zhao, S. Wu, L. Reynolds, and S. Azenkot, "A face recognition application for people with visual impairments: Understanding use beyond the lab," *Conference on Human Factors in Computing Systems*, pp. 1-4, 2018.
[24] P. Lucey et al., "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition -Workshops*, pp. 96-101, 2010.
[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor., "AffectNet: A database for facial expression, valence, and arousal computing in the wild*," IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18-31, 2017.
[26] M. J. Lyons, M. G. Kamachi, and J. Gyoba, "Japanese female facial expressions (JAFFE)," *Database of Digital Images*, pp. 200-205, 1997.
[27] M. N Dailey, C. Joyce, M. J. Lyons, and M. G. Kamachi "Evidence and a computational explanation of cultural differences in facial expression recognition," *Emotion*, vol. 10, no. 6, pp. 874-893, 2010.
[28] R. Osada, T. A. Funkhouser, B. M. Chazelle, and D. P. Dobkin, "Shape distributions," *ACM Transactions on Graphics*, vol. 21, no. 4, pp. 807-832, 2002.
[29] D. G. Lowe, "Distinctive image features from scale-invariants keypoints," *J. of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[30] C. M. Bishop, "Pattern Recognition and Machine Learning," *Springer*, 2006.
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097-1105, 2012.
[32] C. Sakr and N. Shanbhag, "Per-tensor fixed-point quantization of the back-propagation algorithm," *International Conference on Learning Representations*, pp. 1-26, 2019.
[33] M. Attamimi, R. Mardiyanto, and A. N. Irfansyah, "Inclined image recognition for aerial mapping using deep learning and tree based models," *TELKOMNIKA*, vol. 16, no. 6, pp. 3034-3044, 2018.
[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511-518, 2001.
[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4510-4520, 2018.
[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1800-1807, 2017.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, pp. 1-14, 2015.

BIOGRAPHIES OF AUTHORS

Hendra Kusuma received the BS and PhD degree from Institut Teknologi Sepuluh Nopember Surabaya both in electrical engineering, in 1988 and 2016, respectively. He also received the MS degree from Curtin University of Technology in renewable energy engineering. From 1989 until now, he is a lecturer in the Department of electrical engineering, Institut Teknologi Sepuluh Nopember Surabaya. His research interests are in artificial intelligence, machine learning, pattern recognition, IoT as well as in applied electronic.

Muhammad Attamimi received his BE, ME, and DE degrees from the University of Electro-Communications in 2010, 2012, and 2015, respectively. He received scholarship from Ministry of Education, Culture, Sports, Science and Technology Japan (MEXT) for his BE and ME courses. From 2012 to 2015, he was also a Research Fellow (DC1) of Japan Society for the Promotion of Science (JSPS). He was with a postdoctoral researcher at the department of Mechanical Engineering and Intelligent Systems, The University of Electro-communications (UEC) from April 2015 to December 2015. Since January 2016, he was with Tamagawa University Brain Science Institute as a commission researcher for six months. Currently, he is a lecturer at Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His research interests are computer vision, visual recognition, machine learning, multimodal categorization, artificial intelligent, probability robotics, intelligent systems, intelligent robotics.

Hasby Fahrudin has completed elementary education at SDIT Ghilmani Surabaya in 2009. After that, the author continued his education at SMPN 6 Surabaya and SMAN 2 Surabaya. After graduating from high school, writer proceed to tertiary education at Institut Teknologi Sepuluh Nopember. While pursuing bachelor degree, he spent a lot of time in developing electrical workshop. He was also the assistant of basic electronics laboratory. The author became the participant of ASEAN SCIENS GKS held in South Korea in 2018.