Sign language detection using convolutional neural network for teaching and learning application

Wan Mohd Yaakob Wan Bejuri[1,3], Nur'Ain Najiha Zakaria[1], Mohd Murtadha Mohamad[2],
Warusia Mohamed Yassin[1], Sharifah Sakinah Syed Ahmad[1], Ngo Hea Choon[1]
[1]Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]School of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[3]Institute of IR 4.0, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Article Info

ABSTRACT

Teaching lower school mathematic could be easy for everyone. For teaching in the situation that cannot speak, using sign language is the answer especially someone that have infected with vocal cord infection or critical spasmodic dysphonia or maybe disable people. However, the situation could be difficult, when the sign language is not understandable by the audience. Thus, the purpose of this research is to design a sign language detection scheme for teaching and learning activity. In this research, the image of hand gestures from teacher or presenter will be taken by using a web camera for the system to anticipate and display the image's name. This proposed scheme will detects hand movements and convert it be meaningful information. As a result, it show the model can be the most consistent in term of accuracy and loss compared to others method. Furthermore, the proposed algorithm is expected to contribute the body of knowledge and the society.

Corresponding Author:

Wan Mohd Yaakob Wan Bejuri
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia
Email: yaakob@utem.edu.my

1. INTRODUCTION

As a result of the disabilities, the persons with hearing disabled and mute find it hard to blend with the world [1]-[8]. They are unwillingly viewed unfairly by society [9], [10]. In certain fields of interaction, they cannot do well. For instance, the educational system for individuals with disabilities is not like others as they do not have any special commodity to buy tools, they have trouble seeking jobs, and much more [11]-[15]. It establishes a difference between those with disabilities and those without them. This gap is constantly growing day by day. Therefore, to close the gap between them, sign language is a very excellent way of communicating for people who are deaf and mute [16]-[20]. Sign language is a gesture-based language that enables communication between hearing and speech impaired people. It is a non-verbal language that deaf and mute people typically utilize to communicate more efficiently with normal people. Sign language has its own set of rules and grammar for efficient communication. In this paper, we aim to design a new scheme of sign language detection. In this shcme, tt consists of two major steps which are the detection and recognition of signs. It detects a particular shape that separates the object from the remaining shapes. The technical aspects of sign language communication include its social path and purpose, which are simple to use, technical, and technological. In this paper, we structure this paper as shown in: section 2 presents the literature review. Section 3 presents the proposed design. Section 4 presents the experiment setup. Section 5 presents the result of proposed method. Section 6 concludes with a discussion and recommendations for further research.

2.   PROPOSED METHOD

The traditional methods of sign language have several issues, particularly creating some problems. In this section, we list into three (3). First is, understanding the exact context of the symbolic expressions of the deaf and mute is difficult work in real life unless it is correctly specified. Second is, disable people have no confidence when they face normal people as they are not always treated as ordinary people. Third is deaf and mute people have found it difficult to engage with normal society these days due to a lack of appropriate platforms. In order to counteract this problem, we aim to establish a scheme of sign language detection for deaf and mute people by applying a convolutional neural network (CNN) to identify sign language using hand gestures. This is due to the CNN is a type of multilayer perceptron that is used to accelerate the processing of information [21]-[26].

This section includes the image classification system's design flow, as well as the functions and components that will be included. Generally, there are five main stages involved in this project such as data acquisition, pre-processing, feature extraction, segmentation, and recognition as shown in Figure 1. The first step is image acquisition which the image is obtained from Kaggle and some of it is self-created. The hue, saturation, value (HSV) color algorithm is used by this system to identify the hand gesture and change the background to black. Then, the images were processed using several computer vision techniques, including grayscale conversion, dilatation, and masking. The hand segmentation is performed in the second stage, in which the hand is extracted from the captured image and then used to determine the return on investment (ROI) in the frame. Lastly, the feature extraction procedure is employed, where are the sign language will be predicted. The output of this sign language detection scheme is consisted in 26 alphabetical and the numerical between 0 to 10 and some other simple gestures such as like, 'love', 'name', and 'you'. As detail, the explanation will be described in two part which are image processing and feature extraction.
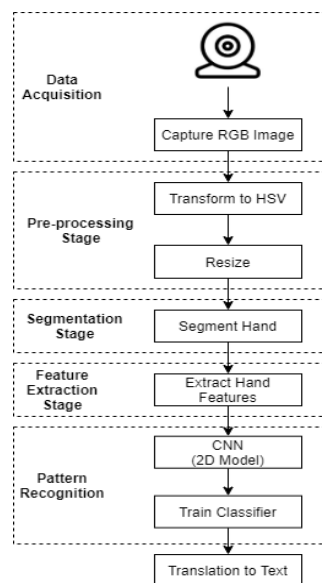

Figure 1. The overview of entire system

2.1. Image processing

The image will be start transformed into HSV colorspaces since it is acquired in red, green, and blue (RGB) colorspaces. This is due to the difficulty of segmenting hand gestures based on skin color only. The approach divides an image color into three components which are hue, saturation, and value. HSV is a helpful method for improving image stability by separating brightness from chromaticity. Moreover, the Hue element does not affect any type of lighting, shadings, or shadows and can thus be considered for removal of the background. A track bar with H values range from 0 to 179, S values range from 0-255, and V values range from 0 to 255 which can be seen in Figure 2 and Figure 3. It is used to detect the hand gesture and set the background to black. Following that, segmentation of the hand gesture is accomplished by removing all unrelated components from the image and then retaining only the most related component. The frame has been reduced to 64 by 64 pixels in size. After the segmentation process is accomplished, binary images of size 64 by 64 pixels are formed, with the white region representing the hand gesture and the black region representing the remaining of the image.
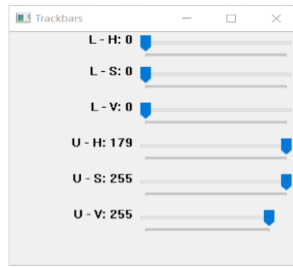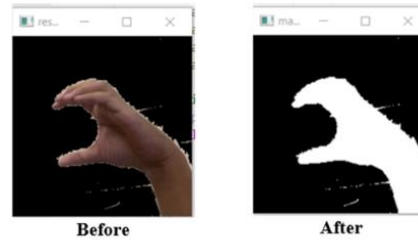
Figure 2. The overview of entire system   Figure 3. HSV Process

2.2. Feature extraction

The convolution layers scan the images with a 3 by 3 filter (see Figure 4). The dot product of the frame pixel and the filter weights is calculated. Thus, this step captures important features from the input image for further processing. After each convolution layer, the pooling layers are applied. One pooling layer decrements the activation map of the previous layer. Following that, it incorporates all the activation map features identified in the previous layers. Additionally, this assists in the elimination of overfitting and the implementation of the network's features. The CNN input layer in this case has 32 feature maps of size 3 by 3 and the activation function is a rectified linear unit (ReLU). The padding and maximum pool layers are 2 by 2 filters in size. This combination of convolution and max pool layers is maintained throughout the architecture. Additionally, the dropout value for this project is set at 0.5, and the layer is flattened. The last layer of the network is a fully connected output layer with 36 units and a SoftMax activation function. Following that, the model is constructed using category cross-entropy as the loss function and SGD as the optimizer with a learning rate of 0.01.
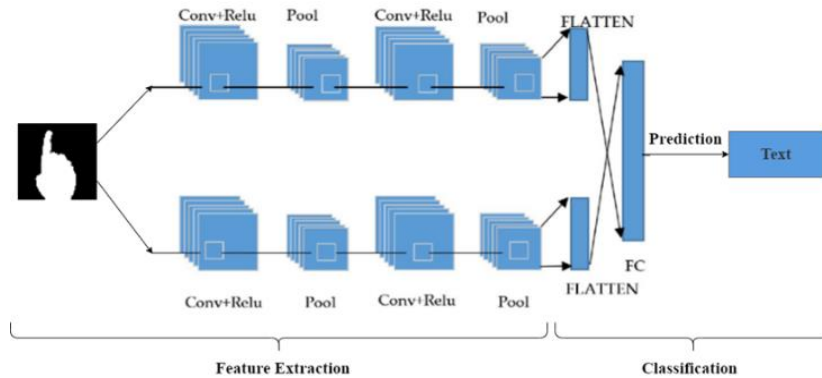

Figure 4. Process of image classification using CNN

3.   RESULT AND DISCUSSION

This chapter will describe the strategy for evaluating the performance of the models, as well as the implementation of the testing and discussion of the results. In this step, we will evaluate our system with models other than 2-D CNN. These models include VGG-16, a 16-layer convolutional neural network, and VGG-19, a 19-layer convolutional neural network. Apart from that, we will evaluate this system using the same dataset for each model. All images in the dataset are captured in portable network graphics (PNG) format with a resolution of 64 by 64 pixels. The test data for this project is divided into two folders which are labeled as train and test. Apart from that, it will be identified in the application using the PyQt5 module, which is retrieved from the model structure stored in a JavaScript object notation (JSON) file and the weights recorded in hdf5 format. Furthermore, a different CNN model with a different layer of parameters will be utilized to evaluate sign language detection using CNN. These include the layer's size, the image's input size in pixels, the convolutional layer's size, the filter's size, the ReLU layer's size, the max-pooling layer's size, the optimizer's type (SGD or Adam), the dropout layer's size, and the SoftMax layer's size. However, during training, all models are trained with the same number of epochs, which is 50, with 200 steps per epoch, and 5000 steps for validation.

3.1. Parameter

This section discusses in detail the parameters used to test the data. Which were specified differently for each CNN model to determine whether CNN models can obtain a good result when classifying the data. As a result, we can ascertain which CNN model is more effective in detecting sign language through hand gestures (see Table 1).

Table 1. Parameters in different CNN models

Layer	Type of Model in CNN		
	2-D	VGG-16	VGG-19
Size of Layer	32	32	32
Image Input Size (Pixels)	64×64	64×64	64×64
Convolutional Layer	4	13	16
Filter Size	32, 64, & 256	32, 64, & 128	64, 128, 256, & 512
ReLU	3×3	3×3	3×3
Max Pooling	2×2	2×2	2×2
Optimizer	0.01 (SGD)	0.001 (Adam)	0.001 (Adam)
Drop Out	0.5	0.2	0.2
SoftMax	1	1	1

3.2. Graphical test

In the graphicl test, each CNN model is required to be trained in 50 epochs with 200 steps each epoch (see Figure 5). After that, the model need to be validated in 5000 steps. As, a result, the 2-D CNN model considered as the most consistent model. It can achieves the second-highest test accuracy, 81.59%, and the second-lowest validation loss, 0.7214, among all the models on the final epoch.
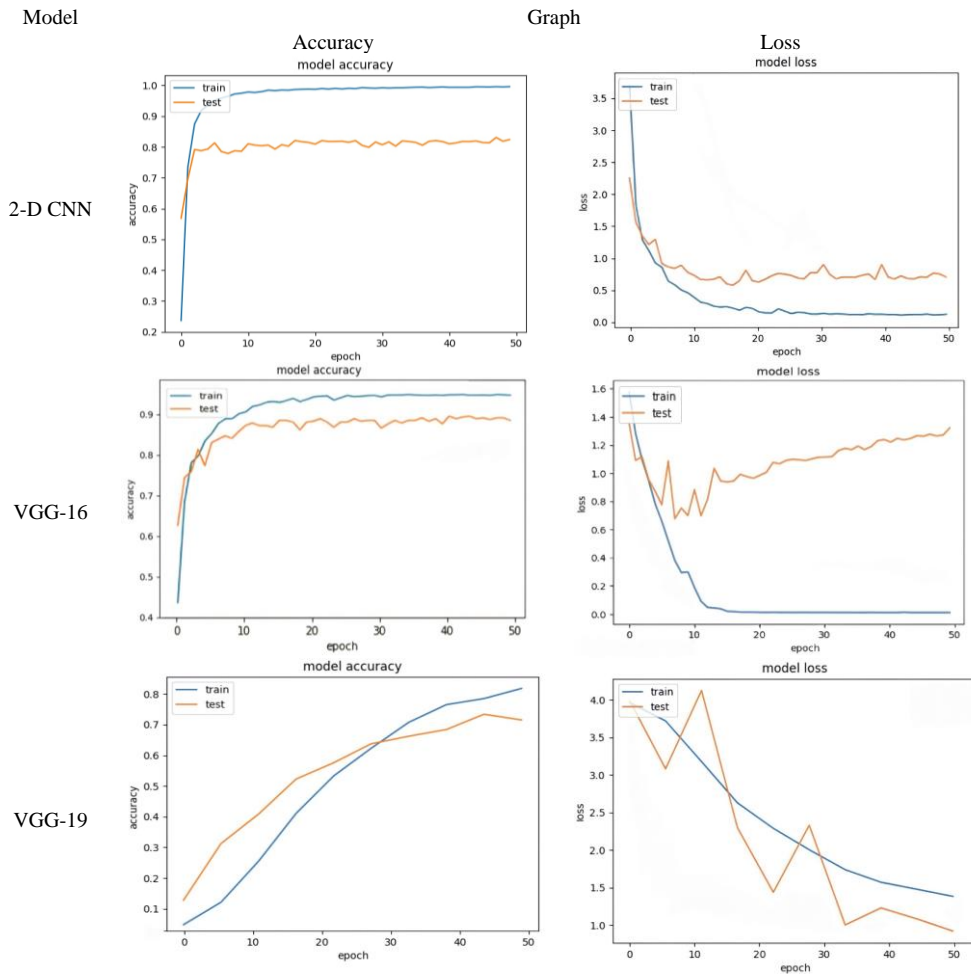

Figure 5. Graphical test result

Meanwhile, the VGG-16 model achieves the highest test accuracy of 87.83%, it also has the highest validation loss of 1.3034 on the 50th epoch, eventhough the model appears to overfit at the 7th. Lastly, the VGG-19 model has the lowest test accuracy, 67.78%, and the lowest validation loss, 0.5417 at the 50th epoch. The reason behind is, the VGG-16 and VGG-19 model appears to be underfitting (test accuracy > train accuracy), because of the implementation of extra dropout layers after the connected layer.

4. CONCLUSION

A person have a disability to speak has difficulty teaching using sign language, especially when the sign language is not understandable by the audience. In this paper, we developed a real-time Sign Language Detection scheme that uses hand gestures and techniques acquired in computer vision and machine learning to assist the deaf and mute in interacting with others. As a result, it show the sign language detection scheme can be the most consistent in term of accuracy and loss compared to others method. Although the proposed classification technique is suitable, it requires improvement in terms of recognition speed and detection efficiency, which are currently lacking. To be more precise, this research is sometimes unable to identify the hand gesture used in sign language since some of them have a similar identical shape to the one being predicted. In addition, the training period of the hand gesture's image for this project is depending on computational power, which indicates that low-end hardware could have a longer training period. Apart than that, we also planning to optimize the performance of the VGG-16 and VGG-19 models, by avoiding training using Adam optimizer and change it to the SGD optimizer.

REFERENCES

[1] R. Bhadra and S. Kar, "Sign language detection from hand gesture images using deep multi-layered convolution neural network," in *2021 IEEE 2nd International Conference on Control, Measurement and Instrumentation, CMI 2021 - Proceedings*, Jan. 2021, pp. 196–200, doi: 10.1109/CMI50323.2021.9362897.

[2] N. Hasfi and T. Rahardjo, "The Disabled people virtual communities in social media from the perspective of public sphere theory," *Jurnal Komunikasi Ikatan Sarjana Komunikasi Indonesia*, vol. 4, no. 2, pp. 65–76, Dec. 2019, doi: 10.25008/jkiski.v4i2.327.

[3] M. F. A. Malik, "Barrier free park: designing a therapeutic environment inclusive of special needs community at Central Park One Utama," PhD Thesis, Universiti Teknologi MARA, Selangor, Puncak Alam Campus, 2018.

[4] A. Pandey and P. Gautam, "Key informant methods: an innovative social mobilization strategy to enable communitybased diagnosis, treatment and rehabilitation for people with disability," *Journal of Nepal Health Research Council*, vol. 18, no. 1, pp. 147–149, Apr. 2020, doi: 10.33314/jnhrc.v18i1.1826.

[5] A. Dewani, S. Bhatti, M. A. Memon, W. A. Arif, Q. Arain, and S. B. Zehra, "Sign language e-learning system for hearing-impaired community of Pakistan," *International Journal of Information Technology (Singapore)*, vol. 10, no. 2, pp. 225–232, Jun. 2018, doi: 10.1007/s41870-018-0105-4.

[6] S. Amir, M. M. Osman, N. Samsudin, S. Bachok, and N. S. Rabe, "Level of satisfaction and preferences on location and design elements of houses among the disabled community in Kuala Lumpur and Selangor," *Advanced Science Letters*, vol. 24, no. 10, pp. 7030–7033, Oct. 2018, doi: 10.1166/asl.2018.12401.

[7] K. Sohelrana, S. F. Ahmed, S. Sameer, and O. Ashok, "A review on smart gloves to convert sign to speech for mute community," in *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, Jun. 2020, pp. 1262–1264, doi: 10.1109/ICRITO48877.2020.9197947.

[8] E. Baldry, D. B. Briggs, B. Goldson, and S. Russell, "'Cruel and unusual punishment': an inter-jurisdictional study of the criminalisation of young people with complex support needs," *Journal of Youth Studies*, vol. 21, no. 5, pp. 636–652, May 2018, doi: 10.1080/13676261.2017.1406072.

[9] K. Renz, N. C. Stache, S. Albanie, and G. Varol, "Sign Language Segmentation with Temporal Convolutional Networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Jun. 2021, vol. 2021-June, pp. 2135–2139, doi: 10.1109/ICASSP39728.2021.9413817.

[10] N. Darapaneni *et al.*, "American sign language detection using instance-based segmentation," in *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*, Apr. 2021, pp. 1–6, doi: 10.1109/IEMTRONICS52119.2021.9422601.

[11] P. Abiud and O. Akach, "Application of South African sign language (SASL) in a bilingual-bicultural approach in education of the deaf," *International Journal of Special Education*, vol. 2, no. 3, pp. 54–59, 2010.

[12] A. Chernaya and Y. Obukhova, "Psychological support for professional development of people with disabilities and special needs: An overview," *SHS Web of Conferences*, vol. 70, p. 10003, 2019, doi: 10.1051/shsconf/20197010003.

[13] P. J. Robertson, "Building capabilities in disabled job seekers: A qualitative study of the Remploy Work Choices programme in Scotland," *Social Work and Society*, vol. 16, no. 1, 2018.

[14] E. Nadai, A. Gonon, and F. Rotzetter, "Costs, risks and responsibility. Negotiating the value of disabled workers between disability insurance and employers," *Swiss Journal of Sociology*, vol. 44, no. 3, pp. 405–422, 2018, doi: 10.1515/sjs-2018-0018.

[15] P. J. Belknap, K. A. Korwin, and N. M. Long, "Job coaching: A means to reduce unemployment and underemployment in the deaf community," *Journal of the American Deafness & Rehabilitation Association*, vol. 28, no. 4, pp. 21–38, 1995.

[16]   O. Robinson and J. Henner, "Authentic voices, authentic encounters: cripping the university through American Sign language," *Disability Studies Quarterly*, vol. 38, no. 4, Dec. 2018, doi: 10.18061/dsq.v38i4.6111.

[17]   P. K. Athira, C. J. Sruthi, and A. Lijiya, "A signer independent sign language recognition with co-articulation elimination from live videos: An Indian scenario," J*ournal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 771–781, Mar. 2022, doi: 10.1016/j.jksuci.2019.05.002.

[18]   K. Parvez *et al.*, "Measuring effectiveness of mobile application in learning basic mathematical concepts using sign language," *Sustainability (Switzerland)*, vol. 11, no. 11, p. 3064, May 2019, doi: 10.3390/su11113064.

[19]   M. De Meulder and H. Haualand, "Sign language interpreting services," *Translation and Interpreting Studies*, vol. 16, no. 1, pp. 19–40, May 2021, doi: 10.1075/tis.18008.dem.

[20]   A. Young, "Deaf children and their families: Sustainability, sign language, and equality," *American Annals of the Deaf*, vol. 163, no. 1, pp. 61–69, 2018, doi: 10.1353/aad.2018.0011.

[21]   J. Feng *et al.*, "CNN-based multilayer spatial-spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 4, pp. 1299–1313, Apr. 2019, doi: 10.1109/JSTARS.2019.2900705.

[22]   M. R. Kapadia and C. N. Paunwala, "Improved CBIR system using multilayer CNN," in *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, Jul. 2018, pp. 840–845, doi: 10.1109/ICIRCA.2018.8597199.

[23]   M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, and M. M. Bin Lakulu, "A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017*," Sensors (Switzerland)*, vol. 18, no. 7, p. 2208, Jul. 2018, doi: 10.3390/s18072208.

[24]   R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou, "Sign language production: A review," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2021, pp. 3446–3456, doi: 10.1109/CVPRW53098.2021.00384.

[25]   A. Wadhawan and P. Kumar, "Sign language recognition systems: a decade systematic literature review," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 785–813, May 2021, doi: 10.1007/s11831-019-09384-2.

[26]   S. N. Chua, K. Y. Chin, S. F. Lim, and P. Jain, "Hand gesture control for human vehicle interaction," *International Journal of Modern Trends in Engineering & Research*, vol. 5, no. 3, pp. 103–106, Mar. 2018, doi: 10.21884/IJMTER.2018.5071.CW0RT.

BIOGRAPHIES OF AUTHORS

Wan Mohd Yaakob Wan Bejuri ⓘ 🄶 sc Ⓟ in Computer Science from Universiti Teknologi Malaysia in 2019, 2012 and 2009. Previously, he was received Diploma in Electronic Engineering in 2005 from Politeknik Kuching Sarawak. He is currently a lecturer at Universiti Teknikal Malaysia Melaka and associate member of Institute IR 4.0, Universiti Kebangsaan Malaysia. He can be contacted at email: yaakob@utem.edu.my.

Nur' Ain Najiha Zakaria ⓘ 🄶 sc Ⓟ currently persuing her degree in artificial intelligence at Universiti Teknikal Malaysia Melaka. She expecting completing her degree within this year (2022). She can be contacted at email: b031910466@student.utem.edu.my.

Mohd Murtadha Mohamad ⓘ 🄶 sc Ⓟ received his Ph.D. in Electrical and Master Degree in Embedded System Engineering from Heriot-Watt University. He received his Bachelor Degree in Computer from Universiti Teknologi Malaysia. He is currently an, associate professor at Universiti Teknologi Malaysia and also a member of IEEE. He can be contacted at email: murtadha@utm.my.

Warusia Mohamed Yassin (iD) (g) SC P is a senior lecturer in Department of Computer Systems and Communication at the Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka (UTeM). He is a member of information security, digital forensic and computer networking (INSFORNET) research group. He completes his Bachelor Degree in Computer Science (2008), Master of Science (2011) and PhD (2015) at Universiti Putra Malaysia (UPM). His research interests include Security in Computing, Machine Learning and Cloud Computing. He can be contacted at email: s.m.warusia@utem.edu.my.

Sharifah Sakinah Syed Ahmad (iD) (g) SC P received his master degree from Universiti Sains Malaysia and PhD degree from University of Alberta, Canada. She is currently as an associate professor in the faculty. She can be contacted at email: sakinah@utem.edu.my.

Ngo Hea Choon (iD) (g) SC P received his master degree from UNSW, Australia and PhD degree from Universiti Sains Malaysia. He is currently as a senior lecturer in the faculty. He can be contacted at email: heachoon@utem.edu.my.