InstantSL: A Sign Language Model to Support Two-Way Communication between Aurally Impaired Communities with Others

Deshinta Arrova Dewi, Harprith Kaur Rajinder Singh, Mohammad Abbas Husaini

Abstract: ***Research works in Sign Language (SL) mostly give prominence to the SL detection and solve the complexity of the hand gesture alphabets. The other works focus on real-time SL recognition and solve various technical problems such as hand self-occlusion, limited resolution due to the sensor quality or others. However, lack of attention is given to the essence of two-way communication i.e. communication channel or platform that accommodate the need of aurally impaired people to speak with others in daily life. This paper introduces InstantSL, an alternative model of Sign Language (SL) application to overcome issues of two-way interaction among users who are familiar and unfamiliar with SL. In addition, the demand of more flexible SL becomes significant along with the usage of a mobile application that has been growing rapidly and makes current SL no longer reliable to run only on local devices like desktop and laptop. Hence, InstantSL proposes a cohesive approach that not only covers on technologies of SL recognition like machine learning for detection accuracy but also voice translation for user-friendliness and mobile operating system Android for flexibility. Promoting interactive communication in delivering messages from SL users and be understood by others are always challenging. Therefore, InstantSL is designed to run as an instant messaging (video call) for remote and face-to-face communication. The aim is to intensify the interaction and conversation between SL users and non-SL users in anytime-anywhere when necessary.***

Index Terms: Sign Language, Machine Learning, Android, Instant Messaging.

I. INTRODUCTION

Sign Language (SL) is well known as a complex language that employs signs made by moving the hands combined with facial expressions and postures of the body. For the aurally impaired communities, SL is substantial for communication because it emphasizes on a vision to communicate and receive information; the moving hands correspond to the spoken words.

SL is increasingly widespread that everyone needs to learn or understand. This could be an arduous task for most people. In order to support fast and better communication between aurally impaired (SL user) and unimpaired people (non SL user), dedicated software is incorporated during communication. This software requires enhanced and relevant features to convert SL into a clear format that is familiar for non-SL users; for example, a software feature that provides conversion of SL into a text format that can be read by both users. Subsequently, the non-SL users can reply in the text to the SL users. This model is simple and workable although slightly time-consuming to perform text-typing.

In order to expedite the communication, voice recognition is proposed in InstantSL to replace text-typing job. This feature able to keep the conversation alive simultaneously. Voice recognition is well-known as a computer program to decode human voice to operate a device, perform a command or write/text without having to press any button on keyboards. Nowadays, voice recognition is referred to as ASR technology (Automatic Speech Recognition). The common ASR has been exploited for voice dialing, call routing, search or simple data entry. In InstantSL, ASR is more on speech-to-text processing that enables non SL users to respond to messages of SL users by a text translation.

Apart from that, the communication channel is an important part of InstantSL. A chat application is very convenient for most people and building a real-time chat application has been simplified for easy use, instant messaging is considered relevant as a platform for InstantSL. The instant messaging or chat application can be generated and replicated following current instant messaging or chat application like WhatsApp, wechat, telegram, etc.

The motivation of using chat application as a platform for InstantSL is because a very limited chat application that supports SL as the main feature. The existing chat application is more towards non-SL user's perspective that comprises of text, voice, and video messaging that cannot be exploited fully by the SL users. This shows a lack of attention and feature-friendly for them. Hence, InstantSL suggests SL implementation in the form of video call and camera for direct communication. This idea is transformed into two ways of application. First, SL without internet access for direct communication or face-to-face. Second is SL with an internet connection. The first application is used to translate the SL to text using camera between two adjacent users. The second application is used for remote communication through a video call with a direct translation of SL to text and voice to text features. With both features, InstantSL offers a new approach to make SL more interactive and friendly.

Revised Manuscript Received on June 8, 2019

Deshinta Arrova Dewi, Faculty of Information Technology (FIT) INTI International University, Nilai, Malaysia.

Harprith Kaur Rajinder Singh, Faculty of Information Technology (FIT) INTI International University, Nilai, Malaysia

Muhammad Abbas Husaini, Faculty of Information Technology (FIT) INTI International University, Nilai, Malaysia.

The main objectives of this paper are to introduce an alternative model of SL application that aims to:

- Minimize the communication gap between aurally impaired (SL user) and unimpaired people (non SL user).
- Increase interactive engagement in the conversation.
- Promote flexibility through the Android operating system and instant messaging services.

Altogether, those objectives need an advance model and supporting technology. Minimizing the communication gap can be done with a medium that supports two-way communication of two parties. This such medium should increase an interactive engagement during SL conversation, hence the alternative model suggests voice-to-text translation to ease the non-SL user to respond. With the Android operating system, the SL app offers a variety of user experience in communication and interaction.

This paper is organized as follows: section II describes the related works relevant to our research. The discussion of techniques that currently employed into modules and how it is used or exploited in other studies. Section III presents the instantSL, the proposed model designed to ease the aurally impaired communities to do two-way communication with others. Section IV concludes the research that has been carried out, follows with acknowledgment and references.

II. RELATED WORKS

Many previous works attempted ways and techniques to create a better SL in helping the aurally impaired people to communicate. Most of the works were categorized as SL recognition using image processing that highlight problems in capturing hands gesture and provide an accurate translation. In this section, American Sign Language (ASL) recognition is emphasized. For example, a researcher focused on solving the problem in a gesture-to-speech interface for American Sign Language (ASL) [1]. The technique is able to reach 88.26% of recognition rate within 0.5 seconds in various illumination. Another researcher uses a multilayer perceptron model in Beale and Edwards' posture recognizer. The idea is to classify sensed data into five postures in ASL [3]. Newby worked on the recognition of the letters and numbers of the ASL manual alphabet based upon statistical similarity [3].

A more simplified method, using approximate spline, was proposed by Watson [4]. The method treats gestures as a representation of a sequence of critical points (local minima and maxima) to illustrate the motion of the hand and wrist [5][6][7]. This approach is more flexible in matching a gesture both spatially and temporally and thus reduces the computational complexity [8][9][10]. In our research, a machine learning algorithm is exploited using similarity computation to produce an accurate gesture model for translation. More explanation is in section III.

In term of speech to text translation, some papers have shared and introduced techniques in their research. For example, the following paper focuses on speech synthesis model that comprises of the acoustic model and an audio synthesis module [11]. This paper highlighted a deep study of text to voice and voice to text translation. Another paper includes a realistic multi-speaker dataset using unsupervised term discovery (UTD) [12]. This paper focused on translating speech to text in low-resource scenarios whereby none of the current technique explores the matter, neither automatic speech recognition (ASR) nor machine translation (MT). On the other hand, a paper introduces combined techniques of convolutional neural network and recurrent neural network in doing the speech to text translation and to train larger models [13]. The motivation behind the research is the awareness of speech-to-text translation that has many potential applications but the typical approach often to be impossible. The paper introduces a technique that enables the listener to adjust their phonetic categories to cope with speech signal variations [14]. This study is important to ensure the speech is clearly transmitted and accepted by the receiver.

Besides speech to translation, the alternative model has suggested an instant massaging platform as a medium of communication. The suggestion was inspired by the literature that has highlighted the effectiveness of using instant messaging in collaborative works. For example, the work of that compare various tools for collaboration and resulted in the instant managing promise more social interaction than others [15]. Another work tests the effectiveness of using instant messaging for online learning. The research proves that instant messaging makes learning more effective [16]. The work used instant messaging for interaction between teacher and students outside the classroom. The instant messaging found to be supportive of social bounding communication. Interesting research in Hong Kong conducted focusing on WhatsApp application to support teaching and learning [15][16]. The result shows positive perception and acceptance towards WhatsApp although the usability issues are found to be valid. The alternative model that we propose in this paper mostly was influenced by the work of [15] and [16].

Another important module that constitutes instantSL is the Android operation system. Previous research stated that the emergence of smartphones has changed the phenomena from just a communication tool, but also an essential part of the people's communication and daily life [17]. In this context, InstantSL implementation as the mobile application is considered relevant.

III. INSTANTSL: THE PROPOSED MODEL

This section focuses on the explanation of the instantSL as an alternative model that attempts to make current SL application become more interactive and useful for the aurally impaired community in speaking with others including the non SL users. The model has two modules that consist of translation SL to text module and voice to text module that is supported by an instant messaging platform. Two environments are included in the application i.e. face-to-face and remote interaction. The instant messaging service is activated during remote interaction. It needs an internet connection to support the video call. In a face-to-face interaction, the system just utilizes a built-in camera in Android smartphone to do translation from SL to text

and voice to text; hence no internet access required. The overview of InstantSL is depicted in figure 1.

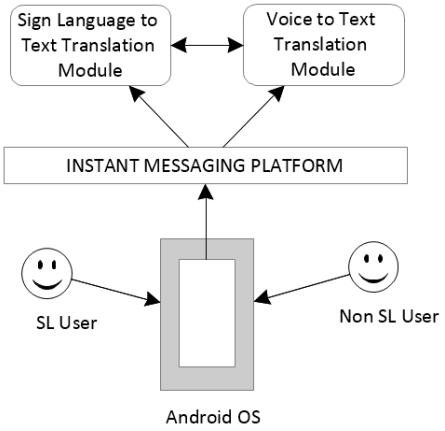

Figure 1. The overall model of InstantSL

As we see in figure 1, two important modules constitute InstantSL and supported by Instant Messaging platform to run on the Android operating system. In the SL-to-text translation module, the main task is to recognize the SL alphabets and numbers which mainly depends on the human's hand and fingers. The process is spread into few stages: preprocessing the input images, computation of the properties region of the preprocessed image, and transliteration from the images to text. This module can be enhanced with features to recognize gestures of SL. Expanding the training set of the main database allows human gestures to be included in the recognition process.

In the voice-to-text translation module, an inbuilt feature is already provided by the Android operating system. Supported by Android studio software, the module is developed using the Android environment. This module is responsible for voice-to-text translation which can be done easily. The speech or voice input is streamed into a voice server whereby the voice server will convert the voice to text and this text will be sent back to the InstantSL. In here, a two-way of communication is established and promote friendlier social interaction between the SL and non-SL users. The sequence diagram of both module's interaction is illustrated in the following figure 2.

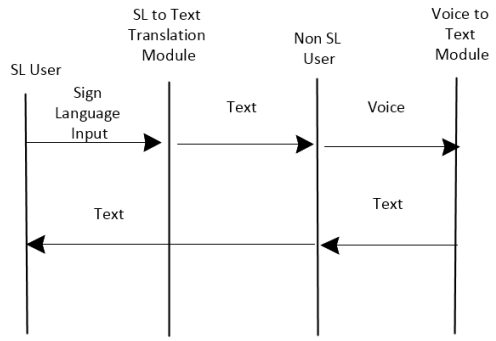

Figure 1. The interaction model of InstantSL

Machine learning technology using similarity computation against training set data is employed for InstantSL. The success ratio is improved when the number of the training dataset is increased. This module cover SL translation of alphabets, numbers, and gestures. With four main processes involved (pre-processing, feature extraction, training images, and machine learning), the InstantSL works in real-time manner and results are sent to non SL user to activate a voice-to-text module for a quick response. The framework of SL to text module is described in the following figure.

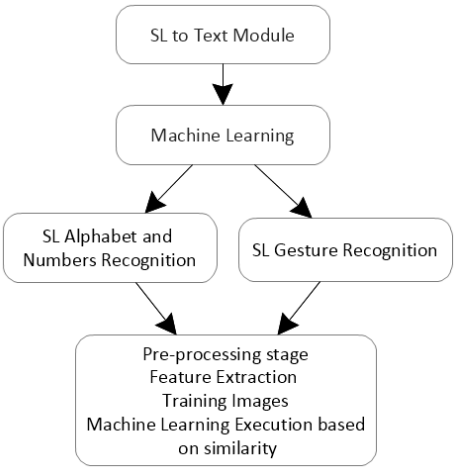

Figure 2. SL to Text module method

The similarity computation is employed in the machine learning module to predict the gesture input and find relevant text. The execution is based on similarity computation when comparing the gesture input from the camera and dataset training in the database. The most common way of doing this is using Deep Learning model as the feature extractor. The other way is using similarity measurement of the image vector. The input image can be converted to vectors and distance measurement computation is performed accordingly. There are three distance measurements available i.e. Euclidean distance, Mahalanobis distance, and Chord distance. The overall process is depicted in the following figure.

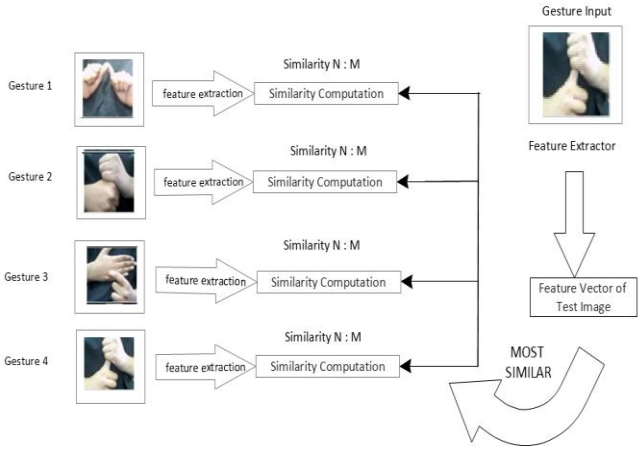

Figure 3. Similarity Computation on Machine Learning

Subsequently, dataset needs to be gathered to support development. The role of the trained dataset is significant in machine learning because the output is based on the

similarity computation against the trained dataset. The trained dataset is kept as an image database is it keep running during a matching process is performed. The following figure explains how machine learning access a trained dataset in the database of InstantSL.

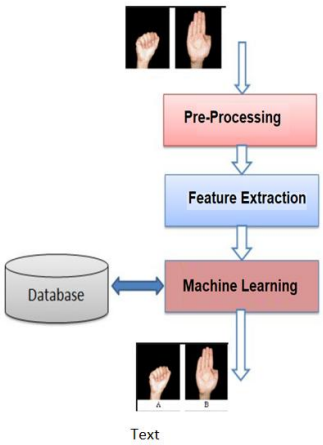

Figure 4. Database of the trained dataset in the translation

After the pre-processing stage, feature extraction requires a central mass of region of the preprocessed image. The system needs to find the area and perimeter (the distance around the boundary of the region) to capture the gesture of the hand movement.

In order to do this, the following equation is adopted.

$$RND = (4 * \pi * A) / P^2 \quad (1)$$

Whereas *A*=area and *P*=perimeter.

Afterward, peak offset point computation from the central mass of the region and compute a number of peaks that are identified. Lastly, to recognize the gestures using the value of the roundness and the number of peaks obtained. So in InstantSL to build machine learning component need several steps to do. First is training dataset. Dataset can be pre-built or using public datasets that are accessible online. For example, the dataset from the University of Surrey that provides two types of ASL and hand gesture dataset: color images and depth images (number of bits used to indicate the color of a single pixel). Another ASL dataset is provided by the Massey University that contains 2542 close-up of color images, cropped in a way the hands touch all four edges of a frame. Both can be used to enrich the system like InstantSL.

Every image of the dataset has to be converted into a feature vector (*x*) and every image will have a label corresponding to the sign language alphabet that it denotes (*y*). The SL recognition is performed by comparing the roundness values and number of peaks between the input image and dataset. If they are matched the gesture is recognized and the system displays the specified alphabet.

In order to do this, the dataset is organized user-wise and the directory structure of the dataset can be arranged in a form of indentation. The image names indicate the alphabet represented by the image. It is separated into training and validation. The directory can accommodate many trained datasets that can be compared during the translation process. The example of dataset arrangement is depicted as follows.

```
Dataset
  |----user_1
        |---A0.jpg
        |---A1.jpg
        |---A2.jpg
        |---...
        |---Y9.jpg
  |----user_2
        |---A0.jpg
        |---A1.jpg
        |---A2.jpg
        |---...
        |---Y9.jpg
  |----                                                    ...
  |-----...
```

Next, is to build classifier as part of machine learning. The key is to use an appropriate strategy to vectorize the image and extract meaningful information to feed to the classifier. InstantSL uses a Histogram of Oriented Gradients (HOG) approach as it has been proven to yield good results. The relevant feature extractors that can be used include Local Binary Patterns and Haar Filters. In order to train multi-class classifier, two functions are built: *functions-crop()* and *convertToGrayToHog()***.** They are used to get the required hog vector and append it to the list of vectors. The following fragment of code written in Panda shows how to load data for multiclass classification.

```
# returns a hog vector of a particular image vector
def convertToGrayToHOG(imgVector):
    rgbImage = rgb2gray(imgVector)
    return hog(rgbImage)
# returns cropped image
def crop(img, x1, x2, y1, y2, scale):
    crp=img[y1:y2,x1:x2]
    crp=resize(crp,((scale, scale)))
    return crp
#loads data for multiclass classification
def get_data(user_list, img_dict, data_directory):
  X = []
  Y = []
  for user in user_list:
    user_images = glob.glob(data_directory+user+'/*.jpg')
    boundingbox_df = pd.read_csv(data_directory + user + '/'
 + user + '_loc.csv')

    for rows in boundingbox_df.iterrows():
      cropped_img = crop( img_dict[rows[1]['image']],
                          rows[1]['top_left_x'],
                          rows[1]['bottom_right_x'],
                          rows[1]['top_left_y'],
                          rows[1]['bottom_right_y'],
                          128
                        )
       hogvector = convertToGrayToHOG(cropped_img)
       X.append(hogvector.tolist())
       Y.append(rows[1]['image'].split('/')[1][0])
    return X, Y
```

As mentioned earlier, managing communication between two parties over voice or video calling on the web is significant for InstantSL. Currently, people often use voice and video calls via various platforms of social media like Skype, Messenger, Facebook, WhatsApp, Line, etc. Generally, both voice and video call need media to stream message between two parties that are connected to each other. In this view, WebRTC (Web Real-Time Communication) is selected to support real-time communication of InstantSL. WebRTC allows voice and video call to be

performed within web browsers on the client side. WebRTC also requires other technology to support real-time communication. First, is *signaling* to initiate a connection between two parties. Second is *STUN* (Session Traversal Utilities for NAT – Network Address Translation) server to establish a peer-to-peer connection by obtaining the IP address. Third, is TURN (Traversal Using Relay NAT) server in case the peer-to-peer connection is failed to establish. The main architecture of InstantSL for a video call is depicted in the following figure.

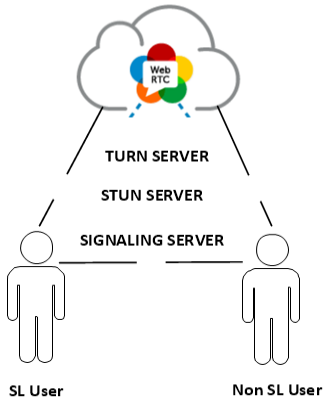

Figure 5. The communication model of InstantSL

In term of speech recognition, InstantSL uses an inbuilt feature that is available in the Android operating system, however, the overview of speech recognition is explained below. Speech recognition works with speech extraction from microphone to capture voice signal. The speech processing module (as stated in Figure 6) is performed to recognize the words and it produces recognition results in the form of text after speech processing and matching process with a prebuilt database. In InstantSL, this feature is important as a platform for the non-SL user to respond to the SL user. Besides, it reduces type-texting job and makes communication faster.

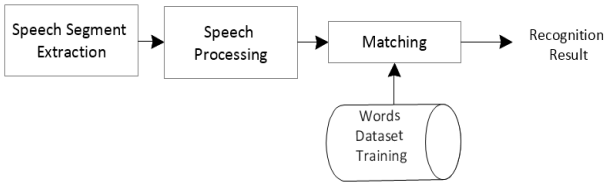

Figure 6. Automatic Speech Recognition System

In the stage of testing or system evaluation, the concerns are some challenges such as: environmental (e.g. lighting sensitivity, static/dynamic background, and camera position), occlusion (e.g. all or some fingers, or an entire hand can be out of the field of frame view), sign boundary detection (when a sign ends and the next begins) and co-articulation (when a sign is affected by the succeeding or preceding sign). All these concern are addressed during the pre-processing stage when receiving gesture input from the camera. The next step is to encode the output labels (the y-values) to numerical values. Once this is done, the model can be trained using any multi-class classification algorithm and can be adopted in many platforms including Android.

To measure the system accuracy, InstantSL has gone through two evaluation metrics i.e. metrics of recognition and real-time user model testing. Metrics of recognition is to identify the accurate detection and real-time user model testing is to indicate the speed of InstantSL in translating the SL. The following equation hence is employed.

$$R = \Sigma no\_of\_gesture\_recognized / \Sigma no\_of\_gesture\_tested \quad (2)$$

whereby R= recognition rate.

Real-time user testing is performed by employing real users (volunteer) to test InstantSL. Both main features (SL recognition and voice translation) are verified in a real-time manner from the user's perspective. This phase can be done repeatedly to ensure the functionality of InstantSL is met and convenience. By performing such testing, a real user is given a chance to evaluate and validate the functionality of InstantSL, without looking at the internal code structure, knowledge of internal paths of the software and implementation details, which inclusive of machine learning classifier and matching process, whether or not the training dataset is able to tackle a variety of input images.

IV. CONCLUSION

InstantSL's technology is a strong alternative model for SL users and non-users to communicate in more convenient way. InstantSL is capable of handling the different input sign images of alphabets and gestures; translates them into text and produce voice translation as a quick response over the instant messaging platform. It has lower execution period so that the recognition and respond can be achieved easily at very less time. Future enhancement is to expand the dataset, train more images to capture another type of SL to achieve higher accuracy.

ACKNOWLEDGMENT

The authors would like to thank the Dean, Head of Programs and colleagues of Faculty of Information Technology (FIT) in INTI International University, Nilai campus, Malaysia for their support and encouragement.

REFERENCES

1. Jayshree R. Pansare, Maya Ingle, "Vision-Based Approach for American Sign Language Recognition Using Edge Orientation Histogram", International Conference on Image, Vision, and Computing, pp.86-90, 2016
2. A. Kuznetsova, L. L. Taixe, and B. Rosenhahn. Real-time sign language recognition using a consumer depth camera. Computer Vision Workshops, pp. 83-90, 2013.
3. B. Leo. Random Forests. Machine Learning, Vol. 45, pp. 5-32, 2011.
4. C. Nölker, and H. Ritter. Detection of Fingertips in Human Hand Movement Sequences. Gesture and Sign Language in Human-Computer Interaction, Vol.1371, pp. 209-218, 1998.
5. C. Nölker, and H. Ritter. GREFIT: Visual Recognition of Hand Postures. Gesture and Sign Language in Human-Computer Interaction, Vol.1739, pp. 61-72, 1999.
6. C. Oz, and M. C. Leu. Recognition of finger spelling of American sign language with the artificial neural network using position/orientation sensors and a data glove. Advances in Neural Networks, pp. 157-164,2005.
7. C. Oz, and M. C. Leu. Linguistic Properties Based on American Sign Language Recognition with Artificial Neural Networks Using a Sensory Glove and Motion Tracker. Computational Intelligence and Bioinspired Systems, pp.1197-1205, 2005.

8. C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun. Real-Time Hand Pose Estimation using Depth Sensors. 2011 Computer Vision Workshops, pp. 1228-1234, 2011.
9. C. R. Mihalache, B. Apstol. Hand pose estimation using HOG features from RGB-D data. System Theory, Control, and Computing (ICSTCC), pp. 356-361, 2013.
10. D. Comaniciu, and P. Meer. Mean shift: A robust approach toward feature space analysis. IEEE Trans. PAMI, pp. 603-619, 2002.
11. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le QV. Tacotron: A fully end-to-end text-to-speech synthesis model. arXiv preprint arXiv:1703.10135. 2017 Mar.
12. Bansal S, Kamper H, Lopez A, Goldwater S. Towards speech-to-text translation without speech recognition. arXiv preprint arXiv:1702.03856. 2017 Feb 13.
13. Bansal S, Kamper H, Livescu K, Lopez A, Goldwater S. Low-resource speech-to-text translation. arXiv preprint arXiv:1803.09164. 2018 Mar 24.
14. Keetels M, Schakel L, Bonte M, Vroomen J. Phonetic recalibration of speech by text. Attention, Perception, & Psychophysics. 2016 Apr 1;78(3):938-45.
15. Sun, Z., Lin, C. H., Wu, M., Zhou, J., & Luo, L. (2018). A tale of two communication tools: Discussion- forum and mobile instant- messaging apps in collaborative learning. *British Journal of Educational Technology*, *49*(2), 248-261.
16. Choi, Jongmyung, and Chae-Woo Yoo. "Connect with things through instant messaging." *The Internet of Things*. Springer, Berlin, Heidelberg, 2008. 276-288.
17. Chen, C. C., & Chu, H. T. (2005, July). Similarity measurement between images. In *29th Annual International Computer Software and Applications Conference (COMPSAC'05)* (Vol. 2, pp. 41-42). IEEE.

AUTHORS PROFILE

Deshinta Arrova Dewi received her Ph.D. in computer science from University Kebangsaan Malaysia (UKM). She is currently is a senior lecturer in INTI International University, Nilai, Malaysia where she has actively teaching and conducting research in the Faculty of Information Technology (FIT). She has authored 30 papers in international conferences, journals, newspaper, book chapter and posters. Her research interest is software engineering, green technology and teaching and learning.

Harprith Kaur Rajinder Singh receives her master degree in computer science from University of Putra Malaysia (UPM). She is currently a senior lecturer in INTI International University Nilai. She involves actively in InstantSL research and development. She successfully leads the project to develop a prototype of InstantSL with her final year project student. Her research interest is software development, leadership and teaching, and learning.

Mohammad Abbas Husaini is a final year student under Bachelor of Computer Science at INTI International University, a dual award program with Coventry University in the UK. He will be completing his degree in a two-semester time. His interest is more on programming and software development. He successfully develops a prototype of InstantSL in around six months.