American Sign Language Recognition Using Deep Learning

PADA Seneviratne[1] and N Wedasinghe[1]

[1]*General Sir John Kotelawala Defence University, Sri Lanka*

36-cs-0006@kdu.ac.lk

Abstract*: American Sign Language (ASL) is a visual gestural language used by the deaf community for communication. There exists a communication gap between hearing-impairedhearing and the normal people because most normal people do not understand the sign language. Conversations with the hearing-impaired people becomes more difficult as most of us do not know the sign language. Hand movements are one of the most powerful nonverbal communication methods which uses both non-manual and manualcorrespondence. ASL-to-text ASL to text interpreting technology using hand gesture recognition could fill up this communication gap. Recently, the hand gesture recognition systems received a great attention and many researchers have been doing studies on the methods for hand gesture recognition for many different purposes. Sign Language recognition is one main purpose among those purposes. Among these the Finger Spelling method is a very interesting research problem in computer vision which has being addressed for years with different kinds of applications in various domains. Inthis paper a survey of existing hand gesture recognition systems and sign language recognition systems are presented for the recognition of Static Finger Spelling method in the American Sign Language. This sign language recognition can be achieved by using sensor- based or vision-based approaches. In this paper, both these approaches are reviewed along with the background of the problem and the pros and cons are also discussed algorithms.*

Keywords: *Sign Language Recognition, Hand Gesture Recognition, American Sign Language*

1. Intorduction

Each individual utilizes a language to communicate with others but there exists a communication barrier between the hearing impaired and speech-impairedspeeche and normal people. When it comes to the conversations with hearing-impairedhearing mostly, they use the sign language to express their thoughts and to understand what the other person says. Hand movements are one of the most powerful nonverbal communication methods which uses both non-manual and manual correspondence. Conversations with the hearing-impaired people become more difficult as most of us do not know the sign language. Sign language is a visual language and it mainly consist of 3 components. They are Fingerspelling, Sign vocabulary and non-manual features. Among these the Finger Spelling method is a very interesting research problem in computer vision that has been addressed for years with different kinds of applications in various domains. In this paper, a critical review has been done to identify existing systems that have been developed using many different technologies on the purpose of hand gesture recognition and sign language recognition.

Furthermore, a systematic review is presented on computer vision techniques as well as other techniques which are used to the utilize the Finger-spelling method and the American Sign Language (ASL) to translate the sign language into text in . This proposedg system aims to develop algorithms and methods to correctly identify a sequence of demonstrated signs and then translate them into its meaning. Several features of sign language must be obtained in order to recognize a sign. Manual markers such as handshape, hand orientation, location and movement expressing lexical meaning are some of those necessary features. This system will not only be used as a communication tool between hearing impaired people and normal people, but also as a system for self-learning assessment. Deaf children will be able to use this system as a self-assessment tool in learning the ASL alphabet.

The rest of the paper is organized as follows. Section 2 presents the background information on the deaf community, sign languages and ASL. Section 3 gives a literature review to study some of the existing systems which uses computer vision-based techniques as well as other techniques on hand gesture recognition and sign language recognition. Section 4 gives the proposed solution to the problem and the following sections contain the methodology, results and the conclusion together with future developments.

2. Background

A. Deaf Community

(Anon., 2021a) Any person who is having problems in hearing and talking like a normal person can be categorized into the deaf community. This community shares a common language to communicate with others. That is called as the Sign Language (SL).

B. Sign Language

(Perera and Jayalal, n.d.) Sign Language is a visual language that uses hand gestures, facial expressions and other body parts like mouth and eyes to convey a message. As mentioned above, there are three main components in sign language. In Finger-spelling method, a word is spelled out using hand signs character by character whereas the Sign vocabulary contains an entire gesture for one word. Non-manumanualtures use tongue and facial expressions to convey the message to the other person. Among these, the finger spelling method uses an alphabet which represents letters by hand signs. All around the world there exist more than 100 different sign language alphabets including Sinhala Sign Language (SSL), American Sign Language (ASL), (Hosoe, Sako and Kwolek, 2017) Japanese Sign Language (JSL) and (Association for Computing Machinery and International Conference on Intelligent Computing and Applications (8th: 2019: Melbourne, n.d.) Thai Sign Language (TSL). American Sign Language is one of the most famous sign languages among these.

C. American Sign Language (ASL)

ASL is a complete, natural language which is the primary language of Northern American deaf community. (Anon., 2021b)ASL has the same linguistic properties as any spoken language, but grammar may differ from normal spoken English. The ASL alphabet is shown as Figure 1.

Figure 1: ASL Sign Language Alphabet

3. Literature Review

(Chen et al., 2014) Sign language recognition is one of the widely used applications that come under the hand gesture recognition technology. This hand gesture recognition has gained a wide research interest for years now. There have been many publications related to this topic. When it comes to the sign-language recognition systems, there are some prominent works done by researchers which helps this study in many ways. (Shukor et al., 2015) Hand gesture recognition technology can be divided into main two categories: wearable data glove method and computer vision-based method.

A. Wearable Data-glove Method

Using a wearable data glove is one of the most used methods in the early days. (Oudah, Al-Naji and Chahl, 2020) Sensors are used to capture the position and motion of the hand. Using this glove technique, the exact coordinates of finger locations and palm, exact orientation and configuration can be obtained easily. These each data glove is outlined with 10 flex sensors, two on every finger and those sensors recognize the bending of eacheach finger and transmit the signals to the microcontroller. (Shukor et al., 2015) In this data glove technique, the flex sensors function as variable resistance sensors, it means that when we bend our fingers the change of the resistance is indicated by the flex sensors. Because of that this framework requires less computational force to recognize the wanted hand gesture.

Figure 2: Wearable data-glove

(Oudah, Al-Naji and Chahl, 2020) This data glove is highly suitable in recognizing both sign motions and fingerspelling which include both dynamic and static signs. However, these gloves cost more money as they were made using high priced sensors. Also, the user had to wear the glove and needs to be physically connected to the computer to keep the interaction between the user and the computer. (Aryanie and Heryadi, 2015) It is possible to make less expensive gloves using less sensors, but they are more vulnerable to noise and will give less accurate results in the sign language interpretation process.

In (Wang and Popovi´c, n.d.), a glove with different color markers is used to track the movement of the hand and capture the gestures, shown in Figure 3. This method has been called “Color-based recognition using a glove marker”. This glove consists of 20 patches colored randomly with a set of distinct 10 colors. The different color patches on the glove enable the camera sensors to track and detect the location of the fingers and palm. This technique is sufficiently distinctive that the system could reliably recognize the movement or the sign of the hand from a single frame. (Wang and Popovi´c, n.d.) Compared to the

data glove method this method is way cheaper but wearing a glove limits the degree of natural interaction between the user and the computer.

Figure 3: Glove used for the color-based recognition using glove marker

B. Computer Vision-based method

1. *Finger Segmentation method*

In (Chen et al., 2014), the authors detect the hand gesture inputs using a method called the "finger segmentation method". In this finger segmentation method, the images are captured with a normal camera and then hand is detected from the background using background subtraction method. All the images are taken under the same conditions and the background is identical. The hand detection process outputs a binary image where the white pixels indicate the hand region, while blacblackels indicate the background. Palm point is found by the distance transform method and palm mask is drawn with the help of the palm point and the wrist points. Next, the fingers are discovered and segmented using the palm mask. (Tay et al., n.d.) Using a labeling algorithm finger regions are marked, and the center points of fingers are detected. Based on the segmentation results, a simple rule classifier is then used to get the recognition done. Using this classifier, the hand sign shown to the camera is recognized according to the content and the number of fingers detected by the segmentation process. (Chen et al., 2014) As the experimental results show, this approach performs well in the real-time applications but the performance of it depends on the result of the hand detection.

2. *Multi Feature Fusion*

Mainly hand gesture features fall into 2 categories: the apparent-based model and the 3-D hand gesture model. In (Liu, Zhang and Zhang, 2012), a hand gesture recognition technology is introduced using the apparent-based model approach. In apparent-based model methods the images are directly used for the hand gesture identification. Using the multi-feature fusion method, the recognition results are improved by extracting the angle count, non-skin color angle, skin color angle in combination with 'Hu invariant moment features' of the large regions of the hand for the target image recognition and for the training of the sample. First, the extracted images from the camera are transformed from RGB space into HSV space for skin color detection. By using the HSV space the hand shaped region can be located effectively. (Liu, Zhang and Zhang, 2012) Next the extraction of the contour of the hand gesture region is obtained comparing to ensure the integrity of the gesture. In the obtained contour area, the center point, radius and the angle of the connection points in the edges are computed and then feature vector is selected. Multiple feature fusion makes the feature analysis data of the extracted images more comprehensive and more differential between the feature values of different hand signs. After extracting extracting features of each test imagimage, value of each feature is matched by Euclidean distance with multi-feature fusion method. Then, the system matches the angle count of hand sign images and selects the possible images through the threshold filter. Skin color angle and non-skin color angle values are matched next through the threshold selection in the same way and further narrow the selection belongs to the hand sign. Finally, the system obtains the result through one of the above 3 categories by matching the Hu variant moments features and then detedetermining classification of the hand sign image.

3. *Scale Invariant Feature Transform*

In (Perera and Jayalal, n.d.), a model is presented which is developed combining CNN (Convolutional Neural Network) and SIFT (Scale Invariant Feature Transform). This model is capable of achieving higher accuracy in sign language recognition using less training data. This proposed system consconsists main 4 stages: data acquisition, image preprocessing, feature extraction, classification and displaying text. As a low-cost implementation, a simple web camera is used for capturing the image. When the images are captured using the web camera, they are preprocessed to enhance the features. Captured RGB images are converted into HSV color space in order to enhance the features. Then a mask is applied to separate the hand region from the background. In the feature extraction stage key points on the preprocessed images are localized and SIFT feature descriptors are generated for each key point, as shown in the figure 4.

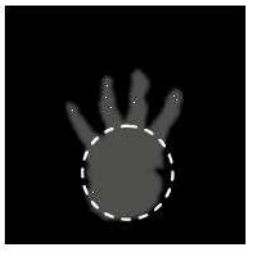

Figure 4: SIFT key point mapped on binary image

(Mahmud et al., n.d.) The sizes of the descriptors differ from each other, because of that a uniform size vector is generated using K-means clustering. To scale the variations

of the hand, this result is combined with a feature map from CNN to improve the robustness. Then the hand sign images are classified into relevant classes. The classification model consists of a channel from CNN and another channel from the SIFT and the final fully connected layer will concatenate both vectors from the CNN and SIFT layers to generate the output of the gesture recognition model. Then the classifier gives a gesture ID to the image and the process of predicting gesture is done by mapping the gesture ID based on the predefined gesture database. When the mapping is successful the relevant text for the sign input is displayed.

4. Solution

There are few main identified problems of this research. They are the communication gap between the hearing-impaired people and the common society, and the lack of opportunities for the disable children to learn the American Sign Language. So, the proposed system of this research will give solutions for these problems.

The final deployment of this proposed system is a mobile application which has the capability of translating the captured sign language gestures into text. This mobile application will bridge the communication gap between the disabled persons and the common society. These days the mobile phone have become an essential part of the daily life of every person. So, the most suitable solution for a communication problem like this will be a mobile application. The disabled person only has to perform the gestures in front of the camera of the mobile phone. The app will capture the images from the video feed and preprocess them using the computer vision techniques. Then these preprocessed imaged will become the inputs for the CNN model. The trained model will classify the images according to the alphabetical letters and map with the corresponding letter after the recognition. Then the translated letters will be displayed on the mobile screen. The users can create words as well as sentences using this application. Using this application, disabled people will also be able to communicate with others freely at any place if they have the mobile phone with them. If someone's intention is to learn the American Sign Language, they can use the front camera of the mobile phone and perform the gestures in front of it. Then the mobile app will translate the gesture and display the corresponding letter for the sign. In this way, the proposed system can provide the solutions for the core problems of the deaf and speech impaired people.

5. Methodology

This proposed system consists of main 4 stages: data acquisition, image preprocessing, feature extraction, classification and displaying text. Data acquisition is done by getting the inputs, or the signs into the system using the camera of the device. Then the image preprocessing, feature extraction and the classification of the signs are done. Finally, as the output of the system the translated text of the signs is displayed.

A. Dataset

A dataset which consists of the letters in the ASL alphabet is created. 2000 images were captured for each individual letter in the ASL alphabet and they are divided into training and test datasets.

Figure 5: Dataset images

B. Input

The main input of this system is the images of the hand signs which are captured by the mobile camera. As this system is based on the American Sign Language there are basically more than 26 sign inputs. In the initial stage the proposed system will be implemented on these signs and later on with the modifications and add-ons, the number of inputs will be increased. Numbers will be represented by using some other signs and another sign will be used to switch between numbers and characters.

Figure 6: Input image

C. Process

When the mobile application is opened, the user is asked to choose from the options given in the application. Then the hand gesture recognition will be done by the mobile application using the developed CNN model. Then the classified signs will be mapped to the corresponding letters. Finally, the words which are spelled using the ASL finger spelling method will be translated into ordinary English words and the output will be given as sentences in the text box.

D. Output

The output of the system is the English alphabet character which is translated using the input sign image. As we use the ASL Finger spelling method in this system the output will be given letter by letter. Adding those letters sentences will be generated and it will be displayed by the system. When this system is used as a learning tool learning sable people, they can choose the learning option and select the letter or the word that they wish to learn using the system and practice it with the help of this system.

6. Result and Discussion

Early days the sign language recognition was done by using the data glove technique. That technique was very costly because those gloves were made using expensive sensors.

So that, researchers tend to search cheaper ways of fulfilling this need. Computer vision-based techniques were the best option found. Using techniques such as Finger Segmentation, SIFT, HMM, CNN and Multi-feature Fusion the above task could be done using just a web camera as hardware. Not only that, when the glove techniques are used, the user had to stay physically attached to the system. Using the computer vision-based techniques it was not necessary stay attached to any hardware devices. As we can see, the drawbacks and the limitations that existed in the wearable glove techniques could have been overcome by the computer vision-based techniques with the development of the new technologies and research. When it comes to the computer vision-based systems, above we have mentioned several systems which uses feature extraction, segmentation, and CNN techniques. When we carefully analyze all these techniques, advantages and the theitation can be listed down as follows.

Technique	Advantages	Disadvantages
Hidden Markov Model (HMM)	Can use for both static and dynamic recognizing, faster, efficient	Need large amount of training data, higher number of parameters used
Convolutional Neural Network (CNN)	Better performance and higher accuracy in achieved, a sub-network improves the classification accuracy, data overfitting is avoided	On higher level dynamic features the classifiers need to improve more.
Artificial Neural Network (ANN)	For the performance the illuminations and complex background affects more	Need to work in hybrid systems (HMM with ANN or CNN with ANN)
Adaptive Probabilistic Model	Can achieve better results with limited power and processing time for the embedded system applications	Due to the multi camera usage the accuracy needs to be improved

Table 1: Advantages and limitations of gesture recognition techniques

As for now the developed model for the above-mentioned purpose works really well and translates the ASL alphabet letters accurately. Figure 8 shows one of the translated letters from the developed system and in Figure 7 the accuracy graph of the developed model is shown.

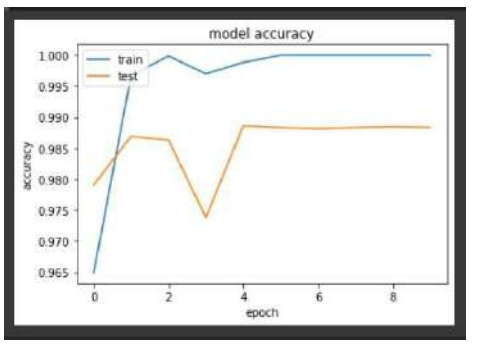

Figure 7: Accuracy graph of the developed model

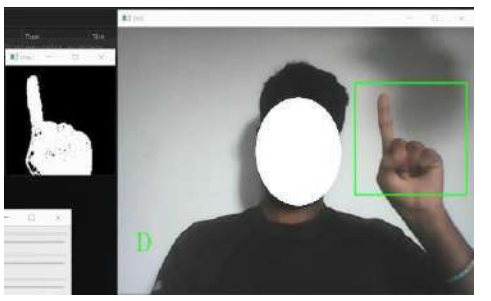

Figure 8: Translation from the developed system

7. Conclusion and Future Works

In this paper, we did a critical review on the existing sign language and hand gesture recognition systems, and we compare and contrast the topics relevant to the sign language recognition systems, different hand gesture recognition models, pros and cons of each model and various techniques used in this relevant research area. After analyzing all these different kinds of methods and techniques closely we can state that the SIFT-CNN technique performs better than others. SIFT-CNN sign language recognition system has less drawbacks and more advantages when compared with other systems. When it comes to the system, we are developing using the above-discussed aboveods and techniques, an android device that we use in our daily life will be more than enough. This system will be able to address the above- mentioned problems as it shows more pros than cons as well as good results. Current datasets are developed only using 3 persons hand gesture images. In the future we will be using more images from many different persons and train the model more accurately. Also, this system will support more than one language and the users will be able to translate ASL into any language they wish.

References

Anon. 2021a. *Deaf culture and community: Why it is important*. [online] Available at: <https://www.healthyhearing.com/report/52285-The-importance-of-deaf-culture> [Accessed 19 October 2021].

Anon. 2021b. *What Is American Sign Language (ASL)? | NIDCD*. [online] Available at: <https://www.nidcd.nih.gov/health/american-sign-language> [Accessed 19 October 2021].

Aryanie, D. and Heryadi, Y., 2015. American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier. In: *2015 3rd International Conference on Information and Communication Technology, ICoICT 2015*. Institute of Electrical and Electronics Engineers Inc.pp.533–536. https://doi.org/10.1109/ICoICT.2015.7231481.

Association for Computing Machinery and International Conference on Intelligent Computing and Applications (8th : 2019 : Melbourne, Vic.), n.d. *Proceedings of the 11th International Conference on Computer Modeling and Simulation, ICCMS 2019 ; Workshop, the 8th International Conference on Intelligent Computing and Applications, ICICA 2019 : January 16-19, 2019, Melbourne, Australia.*

Chen, Z.H., Kim, J.T., Liang, J., Zhang, J. and Yuan, Y.B., 2014. Real-time hand gesture recognition using finger segmentation. *Scientific World Journal*, 2014. https://doi.org/10.1155/2014/267872.

Hosoe, H., Sako, S. and Kwolek, B., 2017. *Recognition of JSL Finger Spelling Using Convolutional Neural Networks*. [online] Available at: <http://home.agh.edu.pl/~bkw/research/data/mva/jsl.zip.>.

Liu, Y., Zhang, L. and Zhang, S., 2012. A hand gesture recognition method based on multi-feature fusion and template matching. In: *Procedia Engineering*. pp.1678–1684. https://doi.org/10.1016/j.proeng.2012.01.194.

Mahmud, H., Hasan, K., al Tariq, A., Hasan, A.-A.-T. and Mottalib, M.A., n.d. *Hand Gesture Recognition Using SIFT Features on Depth Image Brain-Computer Interface View project Human Emotion Recognition View project Hand Gesture Recognition Using SIFT Features on Depth Image*. [online] Available at: <https://www.researchgate.net/publication/303519047>.

Oudah, M., Al-Naji, A. and Chahl, J., 2020. *Hand Gesture Recognition Based on Computer Vision: A Review of Techniques*. *Journal of Imaging*, https://doi.org/10.3390/JIMAGING6080073.

Perera, L.L.D.K. and Jayalal, S.G.V.S., n.d. *Sri Lankan Sign Language to Sinhala Text using Convolutional Neural Network Combined with Scale Invariant Feature Transform (SIFT)*.

Shukor, A.Z., Miskon, M.F., Jamaluddin, M.H., Ali Ibrahim, F. bin, Asyraf, M.F. and Bahar, M.B. bin, 2015. A New Data Glove Approach for Malaysian Sign Language Detection. In: *Procedia Computer Science*. ElsevierB.V.pp.60–67. https://doi.org/10.1016/j.procs.2015.12.276.

Tay, Y.H., Tunku, U., Rahman, A. and Phu, J.J., n.d. *Computer Vision Based Hand Gesture Recognition Using Artificial Neural Network Deep Learning for Forestry View project Deep Learning for Security and Surveillance View project Computer Vision Based Hand Gesture Recognition Using Artificial Neural Network*. [online] Available at: <http://www.sign-lang.uni-hamburg.de/fa/>.

Wang, R.Y. and Popovi´c, J.P., n.d. *Real-Time Hand-Tracking with a Color Glove*.

Author Biography

PADA Seneviratne is a final year undergraduate of General Sir John Kotelawala Defence University. Following the BSc (Hons) Computer Science Degree Programme. Studied at Richmond College, Galle.

Dr. Nirosha Wedasinghe was graduated in Computer and Information systems from The London Metropoliton University in UK in 2005 and has completed her MSc from Charles Sturt University in Austrailia in 2008, As well as her PhD in Information Systems from General Sir John Kothalawala Defense University Rathmalana. She has worked in international universities such as Charles Sturt University, Austrailia and Wolverhampton University UK as a program manager and in University of Greenwich and University of Northampton as a visiting lecturer. Dr.Nirosha Wedasinghe held the positions of the Research coordinator of the faculty, the faculty student counsellor, the incharge of MBA in E-Governance program, the program cordinator for Information Systems degree program and the chairperson of the curriculum development committee of Faculty of Computing at General Sir John Kothalawala Defense University. Currenty she is working as a senior lecturer of the Faculty of Computing in General Sir John Kothalawala Defense University.