A Portable Hong Kong Sign Language Translation Platform with Deep Learning and Jetson Nano

Zhenxing Zhou
Yisiang Neo
King-Shan Lui
Vincent W.L. Tam
Edmund Y. Lam
Ngai Wong
zxchow@connect.hku.hk
neoys97@connect.hku.hk
kslui@eee.hku.hk
vtam@eee.hku.hk
elam@eee.hku.hk
nwong@eee.hku.hk
Department of Electrical and Electronic Engineering, Faculty of Engineering, The University of Hong Kong
Hong Kong, China

ABSTRACT

As hearing loss is arousing more and more public concern, different researches have been conducted on translating the sign language into spoken language. However, most of these researches remain in a theoretical level and few of them investigate how to realize a real system. In this paper, we introduce an effective and portable Hong Kong sign language recognition platform which can translate the Hong Kong sign language within a few seconds. In this platform, there are mainly two parts: a mobile application and a Jetson Nano. The mobile application accounts for preprocessing the sign video and transferring the videos to Jetson Nano. Then, Jetson Nano will translate sign videos into spoken language with the pretrained deep learning model and return the results to the mobile application. With this platform, non-disabled people can easily translate and understand the sign performed by deaf people through mobile phones quickly. We believe that this platform can significantly facilitate the daily communication between deaf people and the others in Hong Kong.

CCS CONCEPTS

• **Applied computing** → *Language translation.*

KEYWORDS

sign language recognition, deep learning, jetson nano, video recognition

ASSETS '20, October 26–28, 2020, Athens, Greece

ACM ISBN 978-1-4503-7103-2/20/10.
https://doi.org/10.1145/3373625.3418046

ACM Reference Format:
Zhenxing Zhou, Yisiang Neo, King-Shan Lui, Vincent W.L. Tam, Edmund Y. Lam, and Ngai Wong. 2020. A Portable Hong Kong Sign Language Translation Platform with Deep Learning and Jetson Nano. In *ASSETS '20: International ACM SIGACCESS Conference on Computers and Accessibility, October 26–28, 2020, Athens, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3373625.3418046

1 INTRODUCTION

As hearing loss is affecting more and more people, sign language recognition has become one of the topics with highest potential impact all around the world [1]. In Hong Kong along, more than 155 thousand people are deaf or hard of hearing (DHH) and use Hong Kong Sign Language (HKSL) for daily communication [2]. However, Hong Kong Sign Language is actually quite different from Chinese/Cantonese in its linguistic rules. Thus, it is difficult for a non-disabled Hong Kong resident to understand Hong Kong Sign Language without professional training. Worse still, there are less than 100 registered Hong Kong Sign Language interpreters which builds a strong communication barrier between the deaf people and the others [3]. A prior study has indicated that technologies for real-time communication are needed [4].

This paper creatively introduces and demonstrates a portable Hong Kong Sign Language translation platform with deep learning and Jetson Nano. Basically, there are two parts in this platform, a front-end mobile application and a back-end Jetson Nano. The mobile application provides a user-friendly interface and preprocesses the sign videos. The back-end Jetson Nano recognizes the sign videos with the help of a pre-trained deep learning model. With this platform, Hong Kong residents without any knowledge in sign language can understand DHH people. Our work is unique that it is the first mobile platform on Hong Kong sign language translation. It also realizes the research advancement in image/video recognition in a practical system.

The rest of this paper is organized as follows: some related work in the field of sign language will be presented in Section II; And

then, the overall structure of the proposed platform together with its components will be introduced in Section III. Lastly, some conclusion remarks and future directions will be provided in Section IV.

2 RELATE WORK: SIGN LANGUAGE RECOGNITION

As one of most important research directions for helping deaf people, many researches have been conducted in the area of sign language recognition since the 21^{th} century and most of them focused on enhancing the recognition accuracy through different methods. At first, some scholars tried to adopt some traditional machine learning approaches such as support vector machine [5] to make the classification based on some basic features extracted from the videos and frames. After that, most researchers started to apply the 2D convolutional neural network (CNN), which has reached a great success in image recognition, for sign language recognition to extract more 2D features from frames and then feed those features into a sequence model such as recurrent neural network for classification [6]. Afterwards, as different 3D models were proved to reach an outstanding performance in video recognition, some scholars also attempted to adopt 3D CNN for sign language recognition [7]. Nevertheless, although many aforementioned researches have reached a decent accuracy in sign language recognition, most of them are not sufficient in supporting Hong Kong deaf people in two aspects:

First of all, most of them focused on the recognition of the American sign language [8] and there is rarely any prior research study focusing on Hong Kong sign language. In fact, the linguistic rules of Hong Kong sign language are quite different from American sign language. Therefore, the recognition methods adopted for American sign language may not be suitable for Hong Kong sign language. To deal with this issue, we introduce a new Hong Kong sign language dataset and pre-train a deep learning model for the proposed platform based on this dataset [9].

Secondly, nearly all the existing works in sign language recognition emphasize the theoretical aspect of the problem but not how to build a practical system. Some existing platforms [10, 11] can only support image-level sign language recognition but not video-level. To the best of our knowledge, none of them explores how to build a practical real-time and portable sign language translation platform. Thus, this paper presents our real-time Hong Kong sign language translation system.

3 PORTABLE HONG KONG SIGN LANGUAGE TRANSLATION PLATFORM

Our goal is to develop a real-time and portable system. That is, the system is able to translate a sign video very soon after it is captured. A fast translation time facilitates more efficient and smooth communication. On the other hand, the system should be portable (small enough in size) so that a deaf person can carry it anywhere. As mobile phones are widely used by the public, the built-in connectivity and camera make mobile phones the ideal input devices for the platform. However, most mobile phone operating systems do not support 3D convolution (which is a crucial part of the deep learning model), making an extra container is needed for the deep learning model. Existing phones in the market are then either not powerful enough or not fast enough. Due to the availability of the cloud, it is also possible to send a captured video from the phone to the cloud for translation. In this way, the system will be still portable. Unfortunately, the network delay may be huge and there may be data cost. Besides, if we want to develop a personalized learning model, some privacy may need to be sacrificed if we rely on the cloud. Therefore, we adopt the edge learning approach that the learning model is kept in a portable device that is "closer" to the users.

Figure 1: Overall Structure of the Proposed Platform

It is worth noting that to develop a translation system that is both fast and small in inherently challenging. Fast translation requires a powerful machine while a small embedded system is limited in both computational and memory capacities. To balance the tradeoff, we identify Jetson Nano [12] which is equipped with Quad-Core ARM®-based CPU and NVDIA Maxwell™ architecture GPU to be our portable server. In our system, each deaf person could have his/her own device so that the training model can be personalized to further increase the accuracy and time performance. The basic structure of the proposed HKSL translation platform is illustrated in Figure 1. The sign videos captured by our application will be firstly preprocessed in the mobile phone and then further transferred to the Jetson Nano through local network. After that, the Jetson Nano will recognize and translate the sign video through the pretrained deep learning model and return the results back to the mobile phone. At the end, the application in the mobile phone will display the results to our users directly. Under this platform, when users want to understand the sign performed by deaf people, he/she only needs to turn on the mobile phone and capture the sign through the camera. Then, the translation results will be automatically showed on our application quickly. Listed below are several scenarios where this platform would be helpful:

- In daily life setting where other means of communication is not easily accessible, non-disabled person can use the platform to understand the sign language from the deaf people during occasions such as shopping and dining in a restaurant.
- In a classroom setting where a deaf student is trying to present an idea to the whole class, the teacher and all students together could utilize the platform to recognize the sign language.

3.1 Hong Kong Sign Language Dataset

To train the deep learning model, a Hong Kong sign language dataset was created by us. In this dataset, 45 most common used sign words were selected. For each sign word, we recorded at least

30 videos from different signers. Thus, there are totally 1500 sign videos in the proposed dataset. Each sign video contains the full sign of one sign word and lasts for 6 to 10 seconds with a resolution of 480 × 640 and a sample rate of 30 fps. Also, we are still collecting more sign videos for different new sign words to enlarge our dataset. More detailed information regarding to this Hong Kong sign language dataset can be found in [9].

3.2 Mobile Phone and Frontend Application

The mobile phone and client application are used as the user interface of our platform which is illustrated in Figure 2. In the current version, the client application is developed and tested on iOS platform. To use this platform, the mobile phone should be connected to our Jetson Nano through the "cloud" button in the top-left corner of the interface. Users will then start recording video of sign language and the recorded video will be preprocessed. Preprocessed frames are transmitted to Jetson Nano after clicking the "send" button in the bottom. At the end, the translation results will be displayed on the screen.

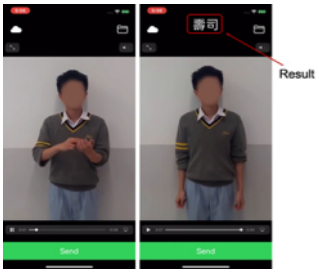

Figure 2: Interface of the Application

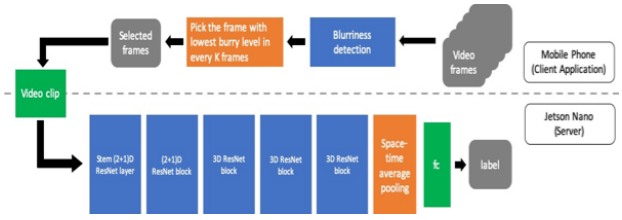

Figure 3: Overall Structure of the Proposed Recognition Method

The preprocessed method is exhibited in Figure 3 and can be explained as follows: Given a sign video with a length of T seconds, we select the clearest frame in every K frames. After the frame selection, totally N frames are picked up in this video. Video clips will then be constructed by every L selected frame with an overlap of P frames. The number of the constructed video clips is decided by K, L and P, which are all hyperparameters. In this platform, after several trials, K, L and P are set to be 5, 16 and 8, respectively.

3.3 Jetson Nano and Backend Deep Learning Model

Jetson Nano and the backend deep learning model are used for recognizing the sign videos and returning the results to the mobile phone. The deep learning model adopted in this platform is (3+2+1)D ResNet model [9] which reached an impressive accuracy of 94.6% in our HKSL dataset. There are one stem (2+1)D ResNet model and one (2+1)D ResNet block [13] followed by three 3D ResNet blocks in the (3+2+1)D ResNet model as illustrated in Figure 3. More information regarding to the detail structure and the Hong Kong Sign Language dataset used for pre-training the deep learning model can be found in [9]. As the Jetson Nano is employed, the portability of the proposed platform is highly increased as the Jetson Nano is actually a product designed for Internet of Thing whose size is similar to a mobile phone. This platform can be deployed at any place as long as there is a network around it. In addition, it is also possible to specifically pre-train a deep learning model for this signer and install it on his/her Jetson Nano so that the recognition accuracy can be increased.

Number of Translated Sign Videos	30
Accuracy in the Experiments	93.3%
Average Preprocess time in phone	0.22s
Average Round-trip Transmission Time	0.54s
Average Recognition Time in Jetson Nano	5.06s
Average Total Time	5.82s

3.4 Experimental Results of the Proposed Platform

To prove the effectiveness of the proposed Hong Kong Sign Language Translation Platform, an experiment was conducted to test the performance in terms of both accuracy and time. In this experiment, 30 sign videos were randomly selected from the testing set of our Hong Kong Sign Language dataset as the input sign videos. Each sign video was firstly preprocessed by the mobile phone and then transferred to Jetson Nano through a local WiFi network. In addition, the inference process of the sign videos was all conducted in the GPU of Jetson Nano. The experimental setting and all the experimental results are listed in Table I.

The total time is the time it takes for the user to get the translation result after hitting the send button. It is the sum of preprocess time in phone, transmission time in the network, and recognition time in Jetson Nano. According to Table I, the major delay in the proposed platform lies in the recognition process of the sign video in the Jetson Nano through deep learning model. This is actually a trade-off between recognition accuracy and speed. As accuracy is more important than the speed in most of the real-world application scenarios, the (3+2+1)D ResNet model is adopted in this platform to guarantee the recognition accuracy, which results in the delay because of the 3D operations in the deep learning model. Fortunately, there are many ways to further reduce the processing delay in a fully developed system. First, a more powerful device can be used instead of Jetson Nano. Second, in our experiments, we are developing a generic recognizer that can recognize the gestures performed by anyone. We believe that the processing time can be further reduced if only the gestures of a single person have to be recognized.

4 CONCLUSION REMARKS AND FUTURE DIRECTION

To facilitate the communication between the deaf residents and the others in Hong Kong, this paper creatively proposed a Portable Hong Kong Sign Language Translation Platform. In this platform, there are one mobile phone with our frontend application, which is used as the user interface, and one Jetson Nano with our pre-trained backend deep learning model for recognizing the sign video. The sign video captured and preprocessed by the mobile will be transferred to Jetson Nano for recognizing and translating.

Most importantly, this work not only supports inclusive education but also opens numerous directions for future exploration. Firstly, it is important to further increase application scenarios of the proposed platform by enlarging the vocabulary size and supporting streamed content. Secondly, the platform is remaining in a word-level translation and can be further upgraded to translating sentence-level sign videos. Thirdly, this work proves the possibility of building up a platform with deep learning and Jetson Nano to help deaf people and there is no doubt that we can build up more similar platforms for helping people in other types of disability. Last but not least, the proposed platform can be easily combined with other existing learning analytics frameworks, such as [14], so that deaf students can blend into normal school.

REFERENCES

[1] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudrealt, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler and M. Ringel Morris. "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective". In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2019.

[2] Census and Statistics Department of Hong Kong Special Administrative Region, "Special Topics Report No.62" p.194, 2017.

[3] The Hong Kong Council of Social Service, "The List of Sign Language Interpreter in Hong Kong" [Online]. Available: https://www.hkcss.org.hk/. [Accessed: 26-Feb-2020]

[4] L. Elliot, M. Stinson, J. Mallory, D. Easton, M. Huenerfauth. "Deaf and Hard of Hearing Individuals' Perceptions of Communication with Hearing Colleagues in Small Groups", In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2016.

[5] S. Nagarajan and T. Subashini. "Static hand gesture recognition for sign language alphabets using edge-oriented histogram and multi class svm." International Journal of Computer Applications, 82(4), 2013

[6] P. Kishore, G. A. Rao, E. K. Kumar, M. T. K. Kumar, and D. A. Kumar. "Selfie sign language recognition with convolutional neural networks". International Journal of Intelligent Systems and Applications, 2018

[7] Y.Q. Liao, P.W. Xiong, W.Q. Min, W.D. Min and J.H. Lu. "Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks". IEEE Access, 2019

[8] M. Huenerfauth, K. Patel, L. Berke. "Design and Psychometric Evaluation of an American Sign Language Translation of the System Usability Scale." In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2017.

[9] Z.X. Zhou, V. Tam, K.S. Lui and E.Y. Lam, "Applying (3+2+1)D Residual Neural Network with Frame Selection for Hong Kong Sign Language Recognition", submitted for publication.

[10] B. Taylor, A. Dey, D. Siewiorek, A. Smailagic. "Real-Time Depth-Camera Based Hand Tracking for ASL Recognition". In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2019.

[11] M.J. Cheok, Z. Omar, M. H. Jaward. "A Mobile Application of American Sign Language Translation via Image Processing Algorithms". IEEE Region 10 Symposium (TENSYMP), 2016.

[12] Nvidia, Inc. "Jetson Nano Developer Kit." [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit. [Accessed: 29-Jun-2020]

[13] T. Du, H. Wang, T. Lorenzo, J. Ray, Y. LeCun, P. Manohar. "A Closer Look at Spatiotemporal Convolutions for Action Recognition". Proceedings of the IEEE conference on computer vision and pattern recognition, 2018

[14] Z.X. Zhou, V. Tam, K.S. Lui, E.Y. Lam, A. Yuen, X. Hu and N. Law, "Applying Deep Learning and Wearable Devices for Educational Data Analytics", Proceedings of the 2019 IEEE 31th International Conference on Tools with Artificial Intelligence, p. 871-878, 2019