2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Applying (3+2+1)D Residual Neural Network with Frame Selection for Hong Kong Sign Language Recognition

Zhenxing Zhou, King-Shan Lui, Vincent W.L. Tam and Edmund Y. Lam
Department of Electrical and Electronic Engineering
The University of Hong Kong
Hong Kong, China
email:zxchow@connect.hku.hk

***Abstract*—As reported by Hong Kong Government in 2017, there are more than 1.5 million residents suffering from hearing impairment in Hong Kong. Most of them rely on Hong Kong Sign Language for daily communication while there are only 63 registered sign language interpreters in Hong Kong. To address this specific social issue and also facilitate the effective communication between the hearing impaired and other people, this paper introduces a word-level Hong Kong Sign Language(HKSL) dataset which currently includes 45 isolated words and at least 30 sign videos per word performed by different signers(more than 1500 videos in total now and still enlarging). Based on this dataset, this paper systemically compares the performances of various deep learning approaches, including (1) 2D histogram of oriented gradients(HOG) feature/pose estimation/feature extraction with long-short term memory(LSTM) layer; (2) 3D Residual Neural Network(ResNet) (3) (2+1)D Residual Neural Network, in HKSL recognition. Meanwhile, to further improve the accuracy of sign language recognition, this paper proposes a novel method called (3+2+1)D ResNet Model with Frame Selection which adopts blurriness detection with Laplacian kernel to construct high-quality video clips and also combines both (2+1)D and 3D ResNet for recognizing the sign language. At the end, the experimental results show that the proposed method outperforms other deep learning approaches and attains an impressive accuracy of 94.6% in our dataset.**

***Index Terms*—Sign Language Recognition, Residual Neural Network, Video Recognition**

I. Introduction

According to the statistics from Hong Kong Government, the number of residents with hearing impairment increases dramatically from 70 thousand in 2000 to more than 155 thousand in the past twenty years [1]. However, Hong Kong Sign Language, as the most important communication method in the deaf community, does not receive enough attention in Hong Kong. Among those people suffering from hearing impairment in Hong Kong, it was reported that only 2.9% of them could communicate with others through sign language. Yet there are only 63 official sign language interpreters in Hong Kong which prevents deaf people from communicating with others significantly [2].

This phenomenon is even worse in children. Due to the prejudice for sign language in the 1900's, most deaf students paid their attention to learning oral language, i.e. speak like a normal person. However, the absence of listening ability makes it very difficult for deaf students to develop a normal speaking skill, which significantly impedes them from communicating with others and hurts their confidence in learning. In addition, according to a research from a local university [3], it was reported that 42.9% of students with hearing impairment in normal school failed in developing their communication skills normally and some of them could only pronounce limited words even though they were already six years old. Different from oral language, it is reported that sign language is beneficial for deaf students in communicating with their teachers and most deaf students prefer using sign language in school according to research from Hong Kong scholars [4]. Although the demand of providing equal education opportunity and environment for deaf students is arousing more and more public attention, there is only one school for deaf students in Hong Kong which is far away from enough. Therefore, to address this social issue and break down the communication barrier between the hearing impaired residents and the whole society, the objectives of this paper is to introduce a new video dataset for Hong Kong Sign Language and propose a new deep learning method for recognizing sign language efficiently.

In order to achieve this target, a new Hong Kong Sign Language dataset is firstly created. In this dataset, there are more than 1500 sign videos which include the most frequently used sign words in Hong Kong. And then, various deep learning methods were adopted to perform sign language recognition, including (1) 2D HOG feature/pose estimation/feature extraction with Long-Short term memory(LSTM) layer; (2) 3D ResNet and (3) (2+1)D ResNet for sign video recognition. After that, to further improve the recognition accuracy in Hong Kong Sign Language, this work innovatively develops a new deep learning method called (3+2+1)D ResNet Model with Frame Selection. In this method, blurriness detection with Laplacian kernel is adopted to select the clearest frame in every K frames to construct high quality video clip so that the video clip can contain clear and enough information for approximately one word which will be elaborated in Section IV. After that, this method creatively combines 3D and (2+1)D ResNet, which is called (3+2+1)D ResNet in this paper, to

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412075

process the video clips. To find out the best model for the recognition of HKSL, this work systematically compares the performance of different combinations of the 3D and (2+1)D ResNet. The detailed information and structure of different combinations of the 3D and (2+1)D model together with their performance will be presented in Section III.

The rest of this paper is organized as follows: Some related works in the field of sign language recognition will be discussed firstly in Section II. After that, in Section III, the proposed Hong Kong Sign Language dataset and the different deep learning methods used for sign language recognition, including six baseline methods for comparison and the proposed (3+2+1)D ResNet model, will be presented. In Section IV, the proposed frame selection mechanism will be explained in detail. In Section V, The experimental results will be provided and the effectiveness of the proposed (3+2+1)D ResNet Model with Frame Selection will be proved. Lastly, some concluding remarks and future directions will be shared in Section VI.

II. Related Work: Sign Language Recognition

As one of the hottest directions in area of video recognition, the studies in sign language recognition/translation have been started since early 21^{th} century. The recognition methods developed gradually from traditional machine learning approaches to 2D deep learning and then further expanded to 3D deep learning model recently. In general, these sign language approaches mainly include two steps: feature extraction and classification. At first, researchers began with extracting hand-crafted features, including the histogram of oriented (HOG) [5], [6] and features in the frequency domain [7] followed by applying different machine learning approaches for classification, such as support vector machine [8]. After that, accompanied by the success of Convolution Neural Network (CNN) in image classification, most scholars paid their attentions to using different pre-trained CNN models, such as VGG16 and AlexNet, to extract 2D features from image frames and then feeding those features into a Recurrent Neural Network (RNN) for considering the information in time domain, such as [9], [10]. In addition to 2D CNN features, following with the fast development in pose estimation, some researchers also attempted to extract the coordinates of the key points in body skeleton through some famous pose estimation models, such as open-pose [11], and reshaped them into a feature vector before passing it through a RNN network [12]. Afterwards, as more and more powerful 3D models, such as I3D [13] and C3D [14] were promoted for video classifications, some scholars also adopted 3D deep learning in sign language recognition [15]. For example, in [9], the authors systematically proved the superiority of 3D model by comparing the performance of different methods in sign language recognition task

However, although many researches mentioned above in sign language recognition have already reached an accuracy between 80% to 90%, they suffered from three important limitations:

First of all, in the preprocessing stage of sign language recognition, most studies simply extracted several consecutive frames by applying sliding window approach into the original videos to construct video clips before passing them through 3D model. However, on the one hand, different from other video classification tasks, video clips for sign language recognition should contain enough signs for approximately one word since the separated signs have no valuable meaning. Therefore, the recognition accuracy will be decreased if the video clip is too short. On the contrary, the time and memory space required for recognizing and training will be enlarged if the video clip is too long and the model will become difficult to train.

On the other hand, in sign language video, the movement of signers' hands are usually fast and large which may probably result in some motion blurs in the video. In this case, it is difficult to recognize the sign with blurry frames. Thereby, to construct high-quality video clips, it is important to measure the blurry level of the frames and select the one with less motion blur. In this research, we proposed to adopt Laplacian operator into each frame and measure the blurry level based on the variation of the flatten Laplacian frames. According to their blurry level, one frame will be selected in every K frames to perform a K time down-sampling to the original videos and construct the high-quality video clips. After the selection and down-sampling, video clip can be constructed by clear but non-consecutive frames. In this case, the total number of video clips extracted from each original video will be reduced but the information in each video clip will be increased.

Secondly, although many research studies have tried to apply I3D and C3D for sign language recognition, there is rarely any research in investigating the possibility of constructing a new deep learning model with traditional 3D ResNet model and the newly proposed (2+1)D ResNet model [16] in sign language recognition. These two ResNet architectures have been proved to attain a state-of-art level in most video recognition task thus probably can be applied into the area of sign language recognition. To fill in this research gap, this paper creatively proposed to adopt (3+2+1)D ResNet Model, which combines both 3D ResNet and (2+1)D ResNet layers for sign language recognition. In this model, there are four ResNet blocks. The top layer is a (2+1)D ResNet block while the last three layers are tradition 3D ResNet blocks. By comparing its performance with other existing or similar models, we systematically prove the superiority of this proposed model. The structure of this model will be presented in more detail in section IV.

Thirdly, most of the researches in sign language recognition focused on recognizing American Sign Language (ASL), followed by Indian Sign Language. But there is seldom any prior studies in recognizing Hong Kong Sign Language (HKSL). As a matter of fact, the linguistic rules of Hong Kong Sign Language is quite different from ASL. Thus, the models proposed for ASL may not achieve a similar performance in HKSL. To solve this problem, we collected a new Hong Kong Sign Language video dataset and explored the appropriate method for recognizing sign video with high accuracy in this

research.

III. The Proposed Dataset and Different Methods for Sign Language Recognition

In this section, the details of the proposed Hong Kong Sign Language video dataset will be presented. Meanwhile, six baseline deep learning methods and the proposed (3+2+1)D ResNet model for sign language recognition will also be explained in detail. In general, these methods can be mainly divided into two categories: 2D approaches and 3D approaches. For 2D approaches, the 2D features of each selected frame, such as HOG, 2D pose coordinate and feature vector from 2D ResNet, will be extracted and fed into a RNN layer thus all of them have the almost identical structure as showed in Figure 3. For 3D approaches, the full video will be firstly cut into small video clips and then all the video clips will be directly used as the input for the 3D model to output the label.

A. The Proposed Hong Kong Sign Language Video Dataset

In this dataset, there are 45 isolated sign words and at least 30 videos for each isolated sign word currently. In total, there are more than 1500 sign videos in this dataset and we are still enlarging it by collecting more sign videos for different sign words. The English meaning of the vocabulary list of this dataset can be found in Table I. All the sign videos were produced by high school students, who have all gone through a professional training in Hong Kong Sign Language, from the Christian and Missionary Alliance Sun Kei Secondary School in Hong Kong. Also, some technical details about the dataset can be listed as follows: $SampleRate : 30fps$; $Resolution : 480 \times 640$; $Duration : 6 - 10s$ and Figure 1 shows one video example in the proposed dataset. This dataset will be kept updated and available to public for research purpose.

B. 2D Approaches for Hong Kong Sign Language Recognition

1) 2D HOG Feature with LSTM: In this method, each frame will firstly be resized to 256×256 and center cropped into 224×224. And then, the histogram of oriented gradients(HOG) of each selected frame was computed in each 16×16 pixels cell and each block with 1×1 cell to construct feature vector with a dimension of 1568. After that, Long short term memory was adopted for considering the temporal information between different frames. Figure 3 displays the structure of this method and it can be explained as follow: After cutting video into frames and computing HOG on each frame, the HOG features of different frames will be integrated and fed into a LSTM layer to consider their temporal relation. After that, a fully-connected layer followed by a softmax layer will be applied to output the label of this video.

2) 2D Pose Estimation with LSTM: Pose estimation targets at extracting the 2D coordinates of the body key points, such as eye and mouth, from a image or a video. In the paper, openpose [11] is adopted to perform the pose estimation and Figure 2 demonstrates the 52 key points that we can get from openpose on one frame in the proposed dataset. Figure 3 displays the structure of this method and it is quite similar to the one of HOG feature except the HOG features were replaced by the 2D coordinates of the body key points.

3) 2D Feature Extraction with LSTM: As also described in Figure 3, this method will firstly adopt a 2D ResNet layer, rather than a pose estimation or HOG feature extractor, to perform a feature extraction. After that, the extracted features from the selected frames will be again combined together and fed into a LSTM to consider temporal dependence between them. Lastly, similar to pose estimation, a fully-connected layer followed by a softmax will be employed to predict the label for the whole video. In this paper, the ResNet-50 model which was pre-trained on ImageNet is adopted to perform the feature extraction. And the feature vectors are extracted before the last layer of the ResNet-50 with a dimension of 2048.

4) Integrated Features with LSTM: As its name implies, this method integrates the 2D HOG feature, 2D pose estimation and 2D feature extraction together. As exhibited in Figure 3, this structure can be explained as follow: the HOG features, the coordinates of key points and the features extracted from ResNet model will be all concatenated together to construct a larger feature vector before being fed into the LSTM and fully-connected layer for final classification. This method considers all the features that extracts by different methods that we can get from the videos to improve the recognition accuracy.

C. 3D Approaches for Hong Kong Sign Language Recognition

1) 3D ResNet for Sign Language Recognition: Compared with traditional 2D convolution, 3D convolutional networks are able to consider not only the spatial information of each frame but also the temporal relationship between the consecutive frames. Thus, this work also considers to use 3D ResNet for sign language recognition which can be taken as baseline methods in 3D approaches. The networking structure of 3D ResNet is illustrated in Figure 4.

After the stem 3D residual convolutional block with a kernel size $3 \times 7 \times 7$, there are four residual blocks and each block contains four 3D convolution layers with a kernel size of $3 \times 3 \times 3$ followed by a 3D batchnorm layer and ReLu activation layer. After that, the results from all residual block are accumulated and processed by the top pooling and fully connected layer to output the label of the video clips.

2) (2+1)D ResNet for Sign Language Recognition: The theory of (2+1)D model was firstly brought up in 2018 [16]. The key idea of this theory is that the operation of 3D convolution actually can be decomposed into a 2D convolution followed by a 1D convolution operation in which the 2D convolution is responsible for processing the spatial information while 1D convolution can preserve the temporal dependence from the input. The advantages for this decomposition include increasing the nonlinearities of the model and simplifying the optimization process by separating the processes for spatial and temporal dimensions. The (2+1)D ResNet model adopted in this paper have the same structure with Figure 4 except the 3D convolutional kernel will be decomposed into a 2D

TABLE I: The English Meaning of the Vocabulary List in the Proposed Hong Kong Sign Language Dataset

Sandwich	Afternoon Tea	Not Tasty	Ginseng	Mushroom
Coca	Having Meals	Monosodium Glutamate	Taste	Coffee
Sushi	How Are You	Dinner	Pear	Coconut
Soup	Fry	Milk	Corn	Noodles
Noodles Roll	Tea	Vegetable	Salad	Continue
Good Evening	Good Morning	Pasta	Pay Bill	See You Next Time
Student	Milk Powder	Senior Citizen	Spoon	Beer
Cold	Curry	Lemonade	Hello	Free of Charge
Menu	Rice	Good Afternoon	Waiter	SoftDrink

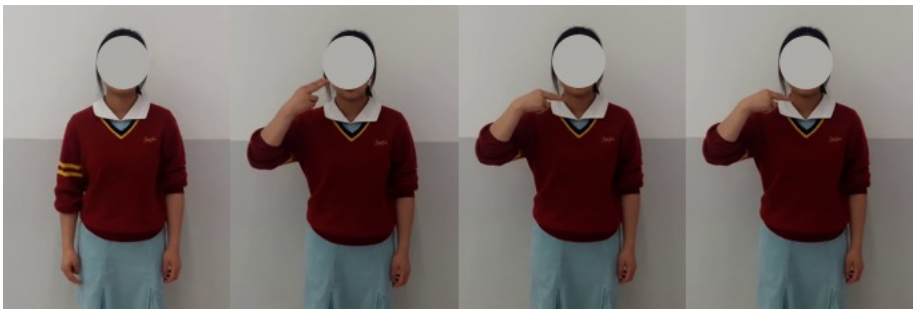

Fig. 1: Example of the Proposed Hong Kong Sign Language Dataset of Word: Having Meals

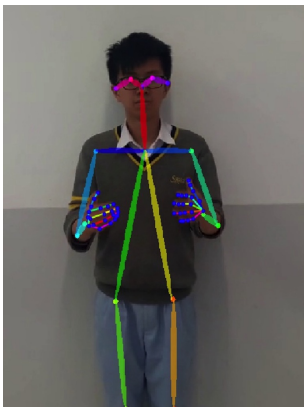

Fig. 2: A Demonstration For Pose Estimation

convolutional kernel and a 1D convolutional kernel which is demonstrated on Figure 5.

3) Proposed (3+2+1)D ResNet for Sign Language Recognition: Although (2+1)D ResNet has its own advantage, it inevitably has a weaker ability in considering the relation between the temporal and spatial information compared with traditional 3D convolution, as these two dimensions are processed by two separated layers at different time. Therefore, to incorporate the advantages of both 3D and (2+1)D ResNet, a new (3+2+1)D ResNet Model is proposed in this work. Figure 6 elaborates the structure of this model. In this structure, after the first stem (2+1)D Residual layer, there are one (2+1)D ResNet block and three 3D ResNet blocks. In the (2+1)D ResNet block, there are four (2+1)D ResNet layers followed by the 3D batch-norm layer and ReLu activation layer while each 3D ResNet block consists of four 3D ResNet layers followed by the 3D batch-norm layer and ReLu activation layer. In this structure, the output of the i^{th} residual layer can be formulated as follow:

$$Y_i = Y_{i-1} + F(Y_{i-1}; \theta_{i-1}) \tag{1}$$

In addition, in the (2+1)D block, the N_i 3D convolutional kernel, where N_i represents the number of filters that we used in the i_{th} block, with shape $N_{i-1} \times t \times d \times d$ will be replaced by N_i group of a 2D convolutional kernel with shape $N_{i-1} \times 1 \times d \times d$ and a 1D convolutional kernel with shape $M_i \times t \times 1 \times 1$, in which the M_{i-1} is formulated as

$$M_i = int(\frac{td^2 N_{i-1} N_i}{d^2 N_{i-1} + tN_i}) \tag{2}$$

so that the number of parameters in this block is similar to the one in 3D convolutional block as suggested by [16]. Moreover, different from 2D convolution methods, the inputs for both (2+1)D and 3D ResNet model are required to be a 4D vector with shape $C \times T \times H \times W$ in which C denotes the number of color channel, T represents the length of video clips while H and W mean the height and width of each frame, respectively.

The hypothesis of constructing the (3+2+1)D ResNet model in this way can be explained as follow: at the beginning of the model, the major task of the top deep learning layer is to extract some basic features from both temporal dimension

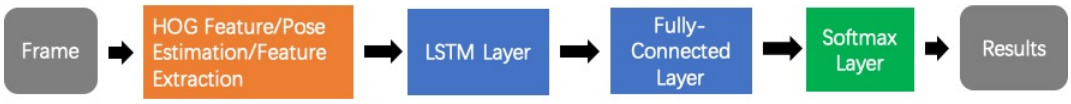

Fig. 3: The General Structure for 2D HOG Feature/Pose Estimation/Feature Extraction/Integrated methods

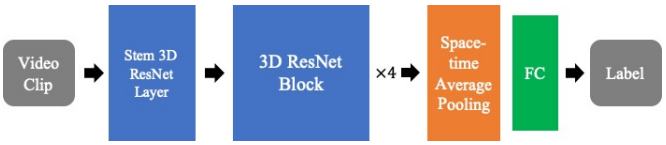

Fig. 4: The Structure of 3D ResNet model

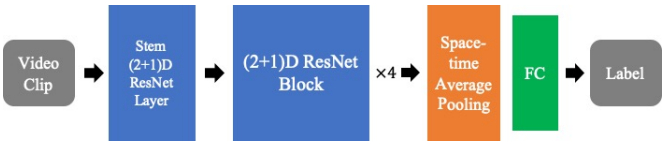

Fig. 5: The Structure of (2+1)D ResNet model

and spatial dimension, such as edge detection. In this state, the weak ability of (2+1)D ResNet in considering the relation between these two dimensions will be probably overweighted by its advantages in extracting features from different dimensions separately and increasing the complexity. Thus, we will benefit from the (2+1)D ResNet layer in this situation. However, when it comes to the deeper part of the model, the relation between temporal dimension and spatial dimension becomes more important than before. In this case, it would be better for us to adopt 3D ResNet layer as it can process both spatial and temporal information at the same time and consider the relation between them.

To prove the validity of this hypothesis, we constructed different types of (3+2+1)D ResNet model by adjusting the position of (2+1)D and 3D ResNet blocks for comparison. These architectures include one (2+1)D ResNet block followed by three 3D ResNet blocks (named as $hybrid\_1\_3$ in the following section), two (2+1)D ResNet block followed by two 3D ResNet blocks ($hybrid\_2\_2$), three (2+1)D ResNet blocks followed by one 3D ResNet block ($hybrid\_3\_1$) and their reversal versions, including one 3D ResNet block followed by three (2+1)D ResNet blocks ($r\_hybrid\_1\_3$), two 3D ResNet block followed by two (2+1)D ResNet blocks ($r\_hybrid\_2\_2$) and three 3D ResNet block followed by one (2+1)D ResNet blocks ($r\_hybrid\_3\_1$). The experimental results of all these structures clearly demonstrate the effectiveness of the proposed model.

IV. Constructing the Video Clips with Frame Selection

As mentioned in Section II, in the preprocessing stage of video recognition task, each video will firstly be cut into separated video clips for training the model. However, in this preprocessing stage, most studies simply applied sliding window to randomly extract several consecutive frames and then merged them together to be a small video clip before passing them through the model. The most common length of the video clip includes 8, 16 or 32 frames in traditional video recognition tasks. However, this may not work well in sign language recognition task.

Firstly, in traditional video recognition tasks, the label can be predicted at the first several frames sometimes. For example, if you find a basketball and basketball rim from several frames at the beginning of the video, you can already recognize this video as playing basketball actually. On the contrary, it is very difficult for you to recognize the meaning of the signer from a small video clip because the separated signs have not valuable meanings before grouping together into a word. To explain and provide a more intuitive understanding about this issue, two frames are extracted from our datasets in Figure 4. These two frames are actually the first and 32^{th} frame of our video dataset. As shown in Figure 4, although the length of video clip has been expended to 32 frames, it is still nearly impossible for you to recognize the sign language in this clip as it is actually the very beginning of the whole sign video.

Secondly, in sign language video, the movements of signers are usually fast which probably results in motion blurs. An example of motion blur is also exhibited in Figure 4. Therefore, to construct video clips with high quality, it is important to remove the blurry frames and select high-quality frames from the video stream.

To address these issues, in this research, the variation of Laplacian operation is adopted to determine the blurry level of a frame. The Laplacian operator is defined as $\Delta f = (\partial^2 f)/(\partial x^2) + (\partial^2 f)/(\partial y^2)$ and according to this definition, we can obtain the Laplacian kernel as $K_f = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$. After that, the blurry level of the image can be formulated as follow:

$$B = -Variation(conv2d(frame, K_f)) \tag{3}$$

The variation in formulation is calculated on the flatten Laplacian frame. The processing procedure is elaborated in Figure 8 and it can be explained as follow: firstly, the blurry level of each frame in the video was calculated. Then, based on the blurry level, pick the frame with lowest blurry level in every K consecutive frames for downsampling and denoising, where K is set to be five during the experiment in this paper. Finally, all the selected frames will be used to construct video clips and the length of each video clip is set to be 16 frames. In this case, the actual frame rate of our videos will be decreased from 30fps into 6fps. Although we can get less video clips in this method, the quality of the constructed video clips will be

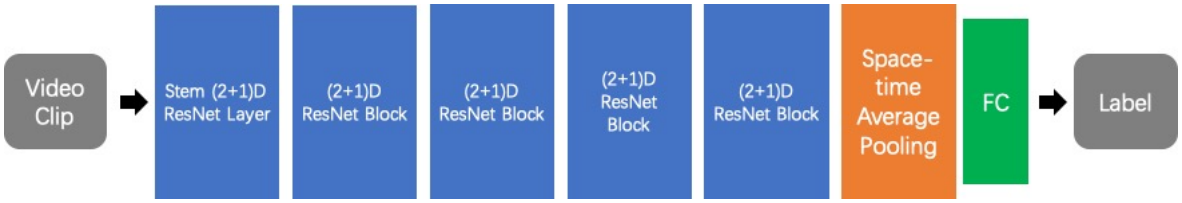

Fig. 6: Proposed (3+2+1)D ResNet Model

Fig. 7: Two Example Frames Extracted from Dataset

increased so that it can contain the clear and enough signs of approximate one word.

V. Implementation Details and Experimental Results of Different Methods

A. Experimental Results of the proposed methods in our Hong Kong Sign Language dataset

In this research, 75% of the sign videos are used as training set while the rest 25% are reserved for testing. To guarantee the effectiveness of the result, the video clips constructed from the testing video will not be used for training. Also, the frame extracted from the video will be resized and center cropped into a shape of 112 × 112 before the normalization. In addition, to further improve the accuracy, all of the 3D, (2+1)D ResNet and the proposed (3+2+1)D ResNet Model are firstly pretrained on the Kinetics-400 datasets. [17]

Table II summaries the experimental results of different methods and structures in the proposed dataset. In this table, the structure of one (2+1)D ResNet block followed by three 3D ResNet blocks is named as $hybrid\_1\_3$, the structure of two (2+1)D ResNet blocks followed by two 3D ResNet block is named as $hybrid\_2\_2$ and the rest are in the same naming rules. For their reversal versions, the structure of one 3D ResNet block followed by three (2+1)D ResNet blocks is named as $r\_hybrid\_1\_3$ and the rest are also in the same naming rule. For all the hybrid structures and 2D methods, frame selection was adopted.

In general, 3D approaches, including both (2+1)D and 3D ResNet, outperform all the 2D approaches significantly by around 20%. Among the 3D approaches, by adopting the proposed frame selection preprocessing methods, the accuracy of (2+1)D ResNet model is 3.8% higher than the one without it which successfully verify the significance of frame selection. Meanwhile, compared with other models, the proposed (3+2+1)D ResNet Model achieves the best performance and reaches the highest accuracy of 94.6% with frame selection. These experimental results strongly prove the effectiveness of both the frame selection method and the proposed (3+2+1)D ResNet model.

B. Experimental Results of the proposed methods in public Chinese Sign Language dataset

To prove the effectiveness of the proposed (3+2+1)D ResNet Model with Frame Selection, we also evaluated this method in one public Chinese Sign Language(CSL) Dataset I [20]. This Dataset I contains 100 sign words performed by one signers with 5 repetitions thus 500 videos in total. Table III summaries the results of the proposed method and other methods that listed in the website of the CSL dataset [21]. Among all these methods, our proposed (3+2+1)D ResNet Model with Frame Selection reaches a highest accuracy of 96.0%. This experimental result strongly proves the effectiveness of the proposed method.

VI. Concluding Remarks and Future Direction

To address the social issue of the low popularization level of sign language in Hong Kong and help hearing impaired residents to blend into the normal community, this paper introduces a new Hong Kong Sign Language(HKSL) dataset which includes 45 most common used sign words in Hong Kong and at least 30 videos for each sign word. Based on this dataset, we systematically compare the performance of different deep learning approaches in sign language video recognition and propose a new methods called (3+2+1)D ResNet Model with Frame Selection. The experimental results strongly verify the effectiveness of the proposed method which reaches the highest accuracy of 94.6% in the proposed dataset.

Most importantly, this work shed lights on numerous directions for future investigation. Firstly, it is meritorious to examine the performance of other 2D or 3D models in sign language recognition. Secondly, the dataset introduced in this paper is a word-level dataset and it would be more attractive and interesting if we could collect a sentence-level dataset. Thirdly, to expand the application scenarios of the proposed deep learning methods, the possibility of developing a mobile version of it is worth exploring in the future [22]. Last but not least, the proposed HKSL recognition approaches can be easily integrated into other existing learning analytics platforms, such

Fig. 8: The Procedure of Frame Selection

TABLE II: Performance of Different Deep Learning Methods on the Proposed HKSL Dataset

Methods	Accuracy	Methods(with Frame Selection)	Accuracy
2D HOG feature with LSTM	62.7%	(2+1)D ResNet with Frames Selection	93.5%
2D Pose Estimation with LSTM	66.7%	Reversal Hybrid ResNet model r_hybrid_1_3	90.7%
2D Feature Extraction with LSTM	71.1%	Reversal Hybrid ResNet model r_hybrid_2_2	90.3%
Integrated Feature with LSTM	75.8%	Reversal Hybrid ResNet model r_hybrid_3_1	91.1%
3D ResNet without Frames Selection	89.1%	Hybrid ResNet model hybrid_3_1	93.2%
(2+1)D ResNet without Frames Selection	89.7%	Hybrid ResNet model hybrid_2_2	93.8%
3D ResNet with Frames Selection	92.2%	**Proposed Hybrid ResNet Model(hybrid_1_3)**	**94.6%**

TABLE III: Performance of Proposed Method on CSL Dataset

Methods	Accuracy
fc7-3DCNN+fc-LeNet [18]	85.8%
LSTM_fc2 [19]	86.2%
eSC+HOG [20]	92.0%
Proposed (3+2+1)D ResNet Model with Frame Selection	**96.0%**

as [23], so that deaf students can also get equal access to education.

REFERENCES

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

[2] 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]

[3] Sign Bilingualism and Co-Enrolment Education program, "The Statue of Education for Deaf People in Hong Kong". [Online]. Available: http://www.cslds.org/slco/tc/intro1.php. [Accessed: 26-Feb-2020]

[4] W.P. Shi, J.Y. Lu, "The relationship between early deaf education and the development of sign language in Hong Kong", *Journal of Education*, p139-156, 2011.

[5] S. Liwicki and M. Everingham. "Automatic recognition of fingerspelled words in british sign language". *Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, p.50–57. 2009

[6] H. Cooper, E.J. Ong, N. Pugeault, and R. Bowden. "Sign language recognition using sub-units". *Journal of Machine Learning Research*, 2012

[7] P. C. Badhe and V. Kulkarni. "Indian sign language translator using gesture recognition algorithm". *Proceeding of 2015 IEEE International Conference on Computer Graphics, Vision and Information Security*, p.195–200, 2015

[8] 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

[9] D.X. Li, R. Cristian, X. Yu and H.D. Li. "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison". *Proceedings of the IEEE Winter Conference on Applications of Computer Vision(WACV)*, 2020

[10] 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

[11] Z. Cao, T. Simon, S.E. Wei and Y. Sheikh, "Realtime Multi-person 2D pose Estimation using Part Affinity Fields". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017

[12] S.K. Ko, C. J. Kim, H. Jung, and C. Cho. "Neural sign language translation based on human keypoint estimation". *Applied Sciences*, 2019

[13] T. Du, B. Lubomir, F. Rob, T. Lorenzo, P. Manohar. "Learning Spatiotemporal Features with 3D Convolutional Networks". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015

[14] C. Joao and Z. Andrew. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017

[15] 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

[16] 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(CVPR)*, 2018

[17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev. "The kinetics human action video dataset". 2017

[18] T. Liu, W. Zhou, and H. Li, "Sign Language Recognition with Long Short Term Memory," *IEEE International Conference on Image Processing (ICIP)*, 2016.

[19] J. Pu, W. Zhou, and H. Li, "Sign Language Recognition with Multimodal Features," *Pacific-Rim Conference on Multimedia (PCM)*, 2016.

[20] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese Sign Language Recognition with Adaptive HMM," *IEEE International Conference on Multimedia and Expo (ICME)*, 2016.

[21] Chinese SLR Dataset, "Chinese Sign Language Recognition". [Online]. Available: http://home.ustc.edu.cn/ pjh/openresources/cslr-dataset-2015/index.html [Accessed: 14-Jul-2020]

[22] Z.X. Zhou, Y.S. Neo, K.-S. Lui, V.W.L. Tam, E.Y. Lam, N. Wong. "A Portable Hong Kong Sign Language Translation Platform with Deep Learning and Jetson Nano" *In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2020.

[23] Z.X. Zhou, V.W.L. 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* 31^{th} *International Conference on Tools with Artificial Intelligence*, p. 871-878, 2019