Sign Language Recognition Using Modified Convolutional Neural Network Model

Suharjito
Computer Science Department. BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
E-mail: suharjito@binus.edu

Herman Gunawan
Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
E-mail: herman.gunawan@binus.ac.id

Narada Thiracitta
Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
E-mail: narada.thiracitta@binus.ac.id

Ariadi Nugroho
Computer Science Department. BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
E-mail: ariadi.nugroho@binus.ac.id

***Abstract*—Sign Language is an interesting topic and similar to Action Recognition. Especially along with the great development of Deep Learning. Video-based Sign Language Recognition is our concern because we want to recognize a sign not only by the shape but also by the action the signer does. The problem is sign language is very complex and vary. The variation of sign language is making the system harder to recognize all the words accurately. Many researchers have been researching Sign Language Recognition for a long time. So many methods had been used to find out which one is the best method. Because of similarity between Sign Language Recognition and Action Recognition, we are trying to implement one of the top-tier models in Action Recognition which is i3d inception this model is also a new Action Recognition model with very high accuracy. So we can know is it possible to adopt Action Recognition behavior into Sign Language Recognition. The goal of this paper is to implement the i3d inception model to Sign Language Recognition with transfer learning method. From the test we've been done, we got 100% accuracy on training with 10 words and 10 signers with 100 classes but the validation accuracy is pretty low. This model is too overfit.**

Keywords: deep learning;convolutional neural network; recognition; comparation; sign language;

I. INTRODUCTION

Deaf is an adjective word with meaning unable to hear [1]. Sometimes deaf happened to some peoples and make them restricted to do communication with each other. Gladly, there is a method called Sign Language which is a very good way to make them to be able communicate with others. But, the problem is not all peoples are understood or willing to learn about Sign Language. With this research, we hope we can build a better communication with deaf peoples.

Because of the complexity and diversity of Sign Language. It makes this topic even more harder. For example, every country has its own Sign Language and standard. Like America and Argentina Sign Language is completely different thing. And also sometimes Sign Language has some similar sign with different meaning. Which is, it made Deep Learning harder to recognize them. And then we also should consider of differences between people who do the Sign Language. Now, you can imagine how hard Deep Learning should learn many things about this topic to get a good accuracy and decent to be implemented as a device to interpret the Sign Language. We found that some researchers had been doing well in this topic. Like, J. Huang tried to make Sign Language Recognition System using his proposed 3D Convolutional Neural Network model [5] And L.Pigou tried his proposed Convolutional Neural Network with video input and some preprocessing [8]

Sign Language is very complex and divers. Even using different dataset of the same method could make a significant difference of result. That's why we want to try to implement the i3d inception model into Sign Language and analyze the result. And improve this topic to be better.

II. RELATED WORK

Researchers are trying their best to solve the problem of this topic. There are so many research had been done. But still need to be improved to be a decent to solve this topic's problem. We are here to discuss about other's research. And we picked [5], [6], [8] for the main research to be discuss.

In 2014 Pigou did research about Sign Language Recognition. He used his proposed model. First one, he used CLAP14 for his dataset [3]. CLAP14 dataset consists of 20 variations of Italian sign language and performed by 27 peoples with different surroundings, clothes, lighting

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

intensities and movement. The dataset was made by Microsoft Kinect. He used 6600 videos in the development set of CLAP14, 4600 for the training set and 2000 for the validation set. He did some preprocessing also. At first, he cropped the highest hand and the upper body using the give joint information. The preprocessing will make 4 results which are hand, upper-body and both gray-scale and depth of resolution 64x64x32 (HxWxFrame). He also did thresholding to reduce the noise in depth maps, background removal using the user index, and median filtering.

His proposed Convolutional Neural Network model consist of 6 layers including input and output. His input layer need 2 inputs. So, the shape would be 2x2x64x64x32 (Gray-scale and Depth x Hand and upper-body x Height x Width x Frames). His proposed model is not 3D Convolutional Neural Network, so all the kernels are 2D. All the neurons are using Retrificted Linea Units (ReLU). The proposed model consists of 3 layers of Convolutional Neural Network and continue by classical Artificial Neural Network or fully connected layer. With his proposed model he got 91.70% accuracy and 8.30% error rate.

In 2015 J Huang did research with the same topic. He used his proposed model but this one is 3D Convolutional Neural Network. He made his own dataset using Microsoft Kinect. He didn't mention which Sign Language's country or standard. He got 25 Vocabularies (Classes) that are commonly used in daily life. Each word performed by 9 signers and every signer performed 3 times. In total it would be 25x9x3 = 675 videos with 27 videos each word. He selected 18 videos each word randomly to be his training set. The data is recorded by Kinect, capturing color image, depth map and body joints locations simultaneously.

He didn't do any preprocessing. Then let's continue to his proposed model. His proposed model consists of 8 layers including input and output layer. The input shape of his model is 5x64x48x9 (channels x height x width x frames). He selected 9 frames centered according to the videos. And the 5 channels are color-R, color-B, color-G, depth and body skeleton. The model has 4 layers of 3D Convolutional Network which is all the kernels are 3D. He got 94.2% average accuracy for this method.

In 2017 until 2018 J.Carriera bring his innovation into Action Recognition topic. Which is has a good impact for that topic. His innovation is to make Inception V1 [11] to be a 3D Convolutional Neural Network model. And also he used pre-training or known as transfer learning method to his research. He used ImageNet[2] and Kinetic Dataset [7] as his pre-trained dataset. And they trained the pre-trained model using 2 Dataset which are UCF-101 [12] and HMDB-51 [13]. His research has many combinations of models between RGB model, flow model, pre-trained Kinetic and pre-trained ImageNet. The best accuracy on UCF-101 dataset is 98.0% and on HMDB-51 is 80.9%.

In our opinion [6] did really well and decent to be implemented into Sign Language. As we mentioned before, that Action Recognition has a similar characteristic with Sign Language. So, we can analyze if it's suitable for Sign Language, why is it suitable? And why not? We hope we can solve that kind of question after this research.

III. METHOD

A. Data Collection

Data collection is important part of this research. Because dataset can affect to the result significantly. Because of it, we used a public dataset LSA64 [9] that we show you in **Figure 1**. We used 10 vocabularies each word is performed by 10 signers and each signer performed 5 times. So, it would be 500 videos in total.

Fig. 1. LSA64 Dataset example.

B. Data Processing

First, we turn the video into sequence of images and produced varies number of frames. After that we padded the number of frames into averages frame of all videos. The average frame is 126. We add blank screen which is RGB (0,0,0) for every frames that has less than 126 frames and cut every videos that has more than 126 frames.

After the convert process, we resized the frames into 224x224 height and width with bilinear interpolation. Then, pixel values were being normalized into a scale of -1 to 1. So, the numpy shape would be (1, 126, 224, 224,3). The numpy were being store into an HDF5 file using h5py library.

C. 3D Convolutional Neural Network Architecture

For the 3D CNN architecture we used [6], it's called i3d inception or inflated inception. Because this model is based on the inception v1 model [11]. Inception v1 was being modified into 3D CNN. We used this architecture because it can improve the result from the previous researcher who used ResNet-50 models [4], C3D Ensamble [14] and Two-Stream Fusion + IDT [15].

I3d inception consist of 67 convolutional layers including input and output. There are 9 inception modules. The detail of

inception module is shown in **Fig 2**. For the training, we distribute the dataset into 6:2:2 ratio. So, there would be 300 videos for training set, 100 for validation set and 100 for testing set. For training set we distribute 3 videos for each signer. Then it will be 3 x 10 x 10 (words) = 300 videos. For the validation is 1 x 1 x 10 (words) = 100 videos. And for the testing set, it is same with validation set, but with different videos not duplicated.

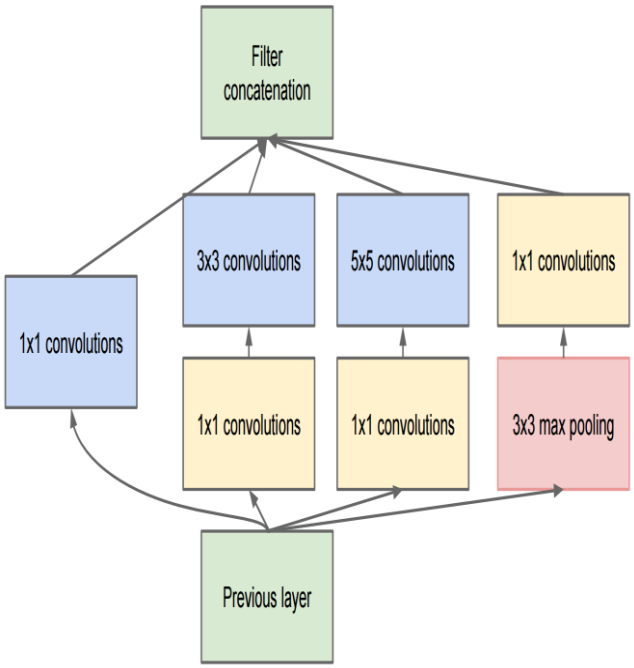

Fig. 2. Details of inception module.

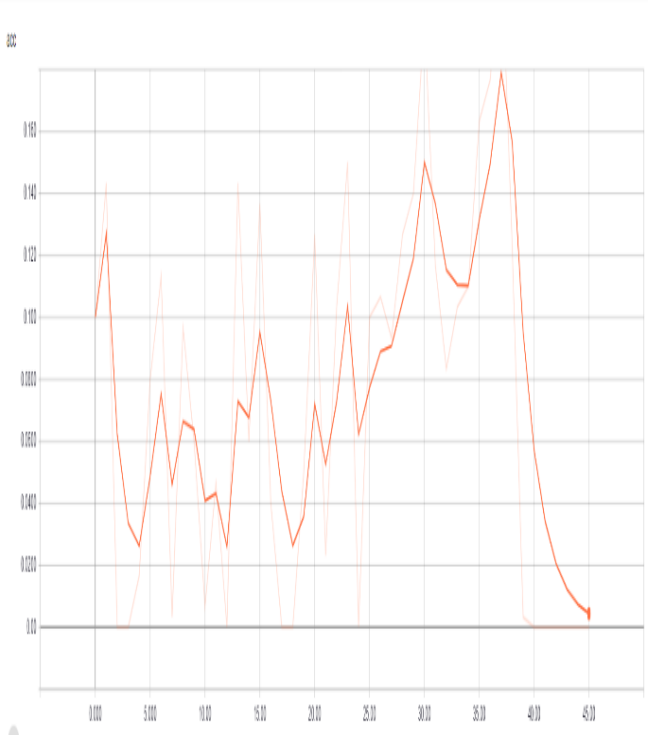

Fig. 3. 10 Signer with 10 classes (500 videos) result.

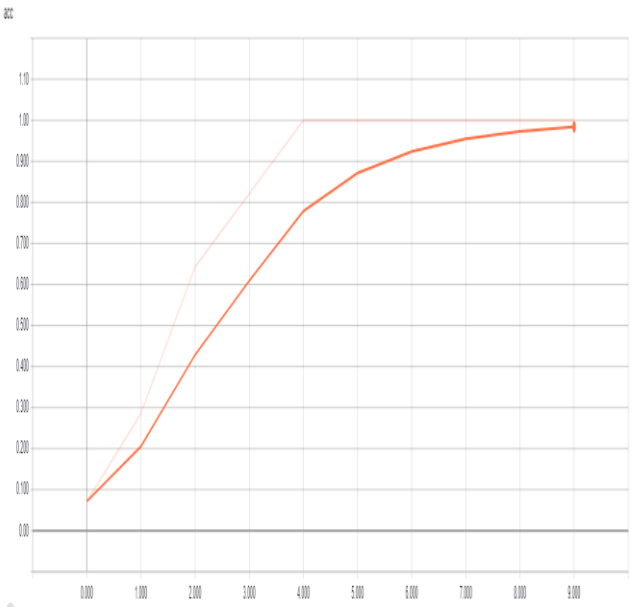

Fig. 4. 1 Signer with 10 classes (50 videos) result.

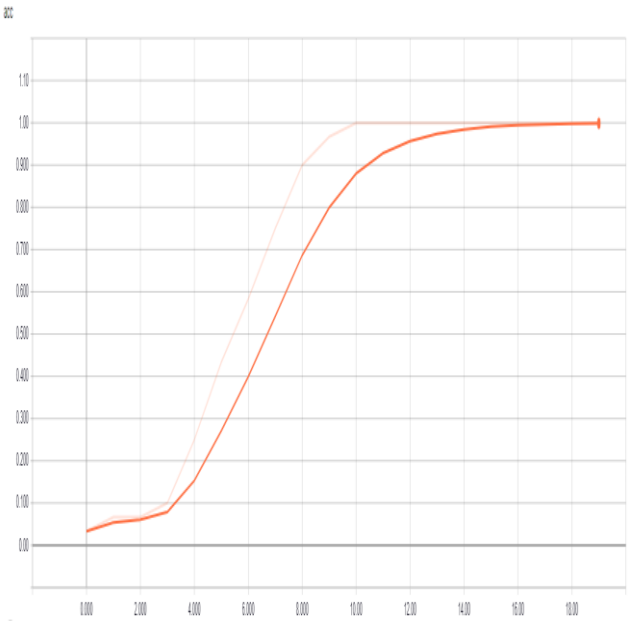

Fig. 5. 2 Signer with 10 classes (100 videos) result.

IV. Result and Discussion

The result of the first training is terrible, the training accuracy was around 0.00% - 20.00% with 45 epochs. We stopped that because we thought something went wrong. Then we tried we a single signer with 10 class. So, it would be 10 (words) x 5 (performed) x 10 (signer) = 500 videos used. And got 100.00% accuracy shown as **Fig 3, 4**.

Because of those result, we thought our model is overfitting. It learnt too detail until it could classify even the signer. After that, we want to make sure our hypothesis. So,

we trained again with different dataset structure. Now we used 2 signers for 10 classes with total 100 videos, 2 signers for 20 classes with total 100 videos and 4 signers for 40 classes with total 200 videos. For the 2 signers 10 classes the training accuracy around 50.00%-80.00%, for 2 signers 20 classes we got 100.00% training accuracy and 4 signers around training 20.00% and still climbing shown as **Fig 5, 6, 7**.

TABLE I. THE COMPARISON OF TESTING RESULT FOR EACH TRAINING

Signer	**Classes**	**Testing Accuracy**
1 Signer	*10 Classes*	100%
2 Signer	*20 classes*	75%
2 Signers	*10 classes*	25%
4 Signers	*40 classes*	0%

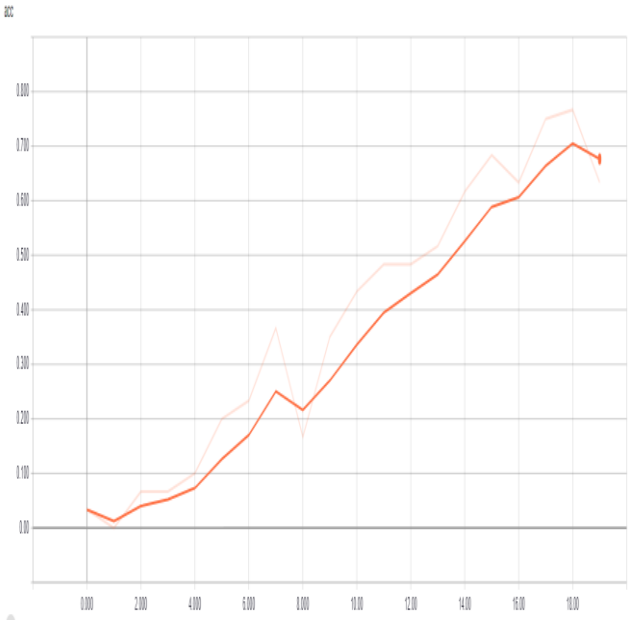

Fig. 6. 2 Signer with 20 classes (100 videos) result.

Fig.7. 4 Signer with 40 classes (200 videos) result

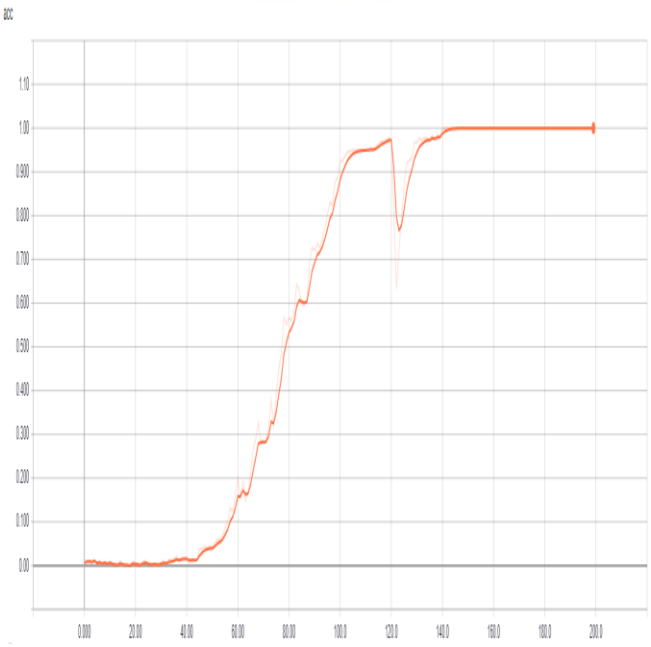

Fig. 8. 10 Signer with 100 classes (500 videos) training accuracy.

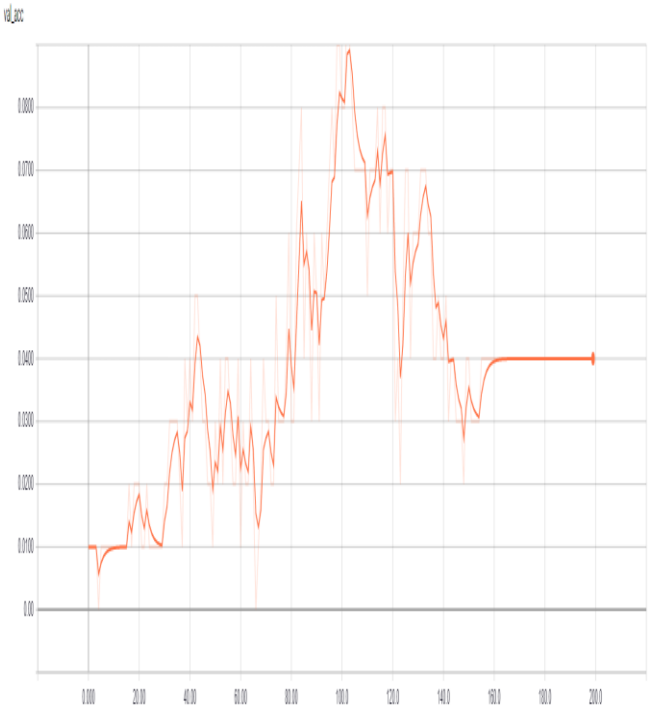

Fig. 9. 10 Signer with 100 classes (500 videos) validation accuracy.

V. CONCLUSION AND FUTURE WORK

After several of trainings, in our opinion i3d inception without modification is too overfit because of the results, we trained the model with 10 signers and 100 classes (500 videos) with 200 epochs, which are has a good training accuracy, but very low validation accuracy. We can do a lot more things with this model, like freeze the layers, remove some inception module, remove the transfer learning, and change the fully connected layer into another deep learning model. In our

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

opinion the fully connected layer is not much the matter. We think we should concern about the convolutional neural network layer more, since it's the detector.

The other reason of these results, it might be from the dataset, because of the dataset. The differences of background lightning of LSA64 could be the cause of the overfit. Because the differences of background lightning could be a feature to the machine learning. So, the machine learning would learn from the wrong feature.

ACKNOWLEDGMENT

This research was supported by Research Grant No. 036/CR.RTT/IV/2018from Ministry of Research Technology and Higher Education of Republic of Indonesia.

REFERENCES

[1] Deaf Cambridge Dictionary. (2018). Retrieved from Cambridge Dictionary: https://dictionary.cambridge.org/dictionary/english/deaf

[2] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE. Miami, FL, USA .

[3] Escalera, S., Baró, X., Gonzàlez, J., Bautista, M., Madadi, M., Reyes, M., . . . Guyon, I. (2014). ChaLearn Looking at People Challenge 2014: Dataset and Results. Workshop at the European Conference on Computer Vision (pp. 459-473). Springer, . Cham.

[4] Feichtenhofer, C., Pinz, A., & Wildes, R. P. (2016). Spatiotemporal Residual Networks for Video Action Recognition. Advances in neural information processing systems, (pp. 3468-3476)

[5] Huang, J., Zhou, W., & Li, H. (2015). Sign Language Recognition using 3D convolutional neural networks. IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). Turin: IEEE.

[6] Jaoa Carriera, A. Z. (2018). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 4724-4733). IEEE. Honolulu.

[7] Kay, W., Carriera, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., . . . Zisserman, A. (2017). The Kinetics Human Action Video Dataset. Computer Vision and Pattern Recognition, arXiv:1705.06950v1, 1-22.

[8] Pigou, L., Dieleman, S., Kindermans, P.-J., & Schrauwen, B. (2014). Sign Language Recognition Using Convolutional Neural Networks. Workshop at the European Conference on Computer Vision (pp. 572-578). Springer, Cham. (pp. 572-578). Springer Link.

[9] Ronchetti, F. |., Estrebou, C. A., Lanzarini, L. C., & Rosete, A. (2016). LSA64: An Argentinian Sign Language Dataset. XIII Workshop Bases de datos y Minería de Datos (WBDMD), (pp. 794-803).

[10] Suharjito, Anderson, R., Wiryana, F., Ariesta, M. C., & Kusuma, G. P. (2017). Sign Language Recognition Application Systems for Deaf-Mute People: A Review Based on Input-Process-Output. 2nd International Conference on Computer Science and Computational Intelligence . Bali.

[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9). (pp. 1-9). IEEE.

[12] Soomro, K., Zamir , A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Computer Vision and Pattern Recognition, arXiv:1212.0402v1, 1-7.

[13] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 2556-2563). IEEE

[14] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. International Conference on Computer Vision (pp. 4489-4497). Las Condes: IEEE.

[15] Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. Compoter Vision and Pattern Recognition (pp. 1933-1941). Las Vegas Valley: IEEE.