American Sign Language Interpreter using an RGB-D camera and Machine Learning

A Thesis

presented to

Mapua University Makati

In Partial Fulfillment of the Requirements in

the course of Thesis

By:

Fernandez, Kurt Christian G.

Paredes, Quintin Brian A.

Perfecto, Janiño I.

2021

MAPÚA UNIVERSITY

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

According to the World Health Organization (WHO) the total percentage of people suffering from hearing impairment is five percent (5%) [1]. These are a collection of people that were born deaf and people that lose their hearing throughout their lives. That approximates to 430 million people worldwide having hearing impairment. It is projected that this number can rise to 630 million by 2030 and 900 million by 2050 [2]. That means five percent of the world's population communicate through sign language and gestures.

One of the most powerful means of communication among humans is through gestures [3]. Sign Language is the language that uses the visual manual modality to explain the meaning of a certain gesture [4]. To date there are a total of one thirty-eight (138) different types of sign language that are known to be used and most of them have the same linguistics but differ in gestures [5]. In addition, the researchers chose the most prominently used sign language in the world which is the *American Sign Language* (ASL). This paper will focus on developing a system that can detect the mentioned sign languages to bridge the gap between the people who are unable to comprehend and those sign users.

1.1 Background of the Study

The emergence of technology for the impaired is increasing — examples are: hearing aids, cochlear implants, and visual alert systems that can help various individuals to have dependable technology that can assist their everyday living [6].

One such technology is gesture recognition, there are two ways for a computer to recognize a certain hand gesture; these are through computer vision and sensors [7]. Sensors use a data glove that measures local hand motions. Within this data glove there are three sensors; the flex sensor, which senses the bending of each finger. Contact sensor, two metal plates that send information when they touch each other. And lastly, an accelerometer that determines the x, y, and z axis [8]. In contrast to this is computer vision that uses an RGB or D-RGB to recognize the hand gesture. There are two types; static and dynamic. Static uses still images[9][10], while the dynamic uses videos for the dataset[11][12]. The researchers proposed a solution that might be able to bridge the gap between the impaired and able-bodied individuals.

The researchers will base their research on an existing study that presented a concept on hand gestures recognition used for interacting with the computer. An analysis and experimentation regarding the application of hand gestures was developed using a depth camera and neural networks — the study was conducted by Tran et al. and according to their findings, they proposed a practical system capable of detecting interactive gestures that can command a computer to do certain tasks[12]. Apart from this, the model was able to achieve a 92.6% accuracy that is higher than the other models. Furthermore, the mentioned work would be useful due to its relevance to our study of recognizing words of ASL using a RGB-D camera.

1.2 Gap or Opportunity

According to Tran et al. in the discussion part of their paper, they intended to expand their system to handle more gestures by applying their methods towards more practical uses[12]. The capabilities of the previous model only focuses on HCI - mainly for computer control. From that proposal, the researchers saw an opportunity to utilize the methods from the

mentioned study in developing a dynamic gesture detection model. In addition to this, a new dataset will be fed and tested to the said model. This dataset is a collection of *American Sign Language* hand gesture words.

1.3 Statement of the Problem

The study is focused on utilizing the previous model towards the ASL hand gesture dataset. Moreover, the study aims to answer the following:

- What will be the performance of the proposed model in terms of accuracy when introduced to a new gesture dataset?
- How will the Long Short-Term Memory Model affect the prediction of dynamic hand gestures?

1.4 Objectives:

The following statements will be used as guidelines in implementing the *ASL*.

- To determine the model of (Tran et al, 2020) in distinguishing the ASL gesture dataset.
- To introduce and utilize the LSTM model when predicting the sequence of hand gestures.

1.5 Significance of the Problem

The aim of the study is to identify and classify hand gestures of ASL by utilizing the proposed model of Tran et al. [12]. Through the usage of 3D neural networks, features will be extracted from each hand gesture video. This research can help in future projects that aim to create a solution for bridging the communication gap between signers and non signers. With the use of the model, the results will be beneficial towards the following:

1. **Future Researchers** - this study can be used to further expand the knowledge of researchers regarding the improvement of communication using sign languages.
2. **Non- Signers** - People who are part of this group will eventually improve their reaction time and boost their communication skills as well as create more opportunities and use this new learning when needed.
3. **Signers** - Socializing with other individuals will drastically improve since people outside of their community are now learning to understand their way of communication.

1.6 Scope and Limitations

This study is focused on adding the capability of the model to detect and recognize the *American Sign Language* (ASL) gesture words using the depth camera developed by Microsoft called *Kinect Sensor*. Furthermore, the researchers decided to use version 2 since it offers more features compared to its predecessor – some of the noticeable advancements are enhanced depth sensing, improved resolution and better skeletal joint tracking [13][14].

The limitations of the study is it will only recognize the most used *ASL* dynamic gestures consisting of 11 classes (most common gestures in ASL)[15][16]. – the model will only focus on detecting ASL. Moreover, these gestures can only be detected when the hand is facing the camera.

CHAPTER 2

REVIEW OF RELATED LITERATURE

2.1 Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network

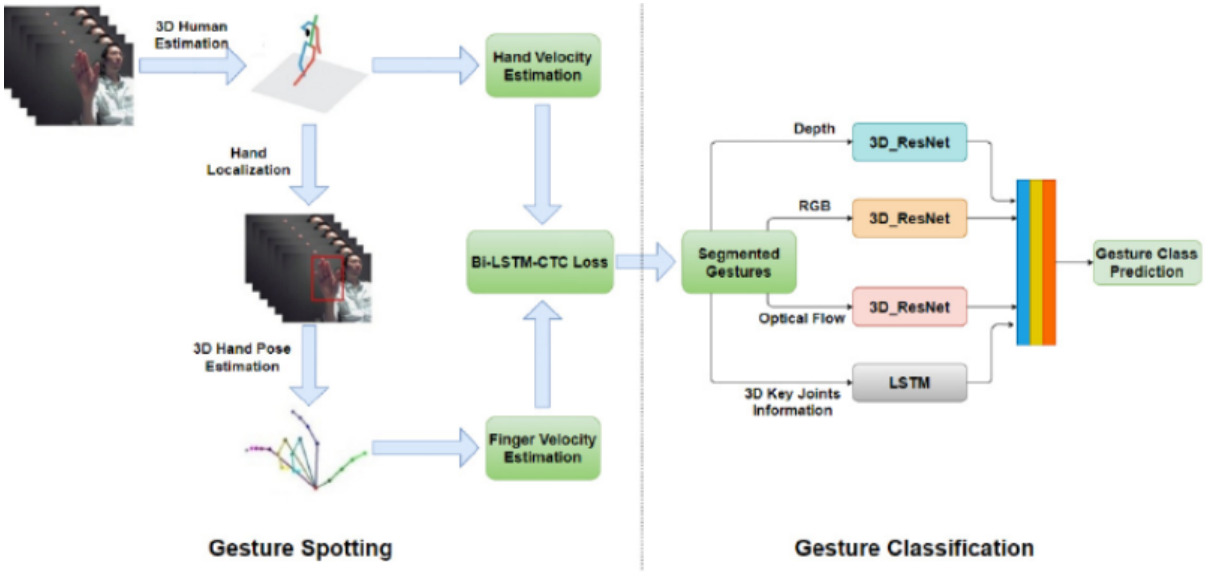
According to Tran et al, the traditional means of interacting between humans and computers are lacking flexibility and to address this concern various methods are implemented such as speech recognition and body-based language interaction. The latter is more suitable for most cases since this is a familiar method of humans during communication.

The researchers proposed a system that can recognize the hand gestures using a Microsoft Kinect Sensor v2 then applying a number of algorithms to extract and detect the fingertip location based on the contoured coordinates of the detected hands[12].

2.2 3D Skeletal Joints-Based Hand Gesture Spotting and Classification

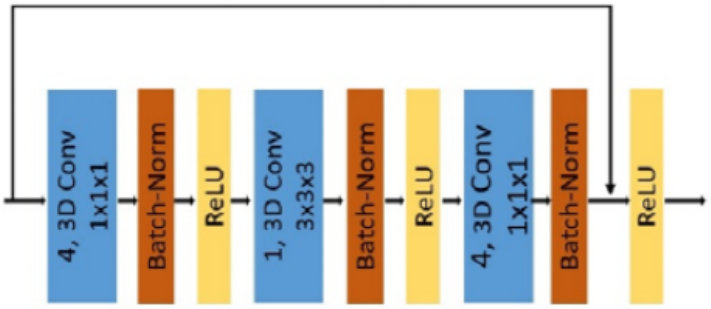
In the paper of Ngoc-Hoang Nguye et al, They proposed a system that uses a spotting-classification *algorithm* for continuous dynamic hand gestures. They focused on their

two main modules; which are gesture spotting and gesture classification [17].



2.2.1 Ngoc-Hoang Nguye et al. Gesture Spotting-Classification Module

For their gesture spotting module. For the research to extract the 3D human pose they utilized all the frames of the continuous gesture sequences. Through RGB hand ROI localized from the 3D hand palm position, when the hand palm stands still and over the spine base joint. They extracted the 3D position of the finger joints using the 3D hand pose estimation algorithm. They also estimated the 3D hand pose by using the real-time 3D hand joints tracking network of OccludedHands. Lastly, with the use of the 3D_ResNet architecture, the researchers can classify the gestures[17].



2.14.2 3D_ResNet architecture

2.3 Hand Gesture Recognition for Sign Language Using 3DCNN

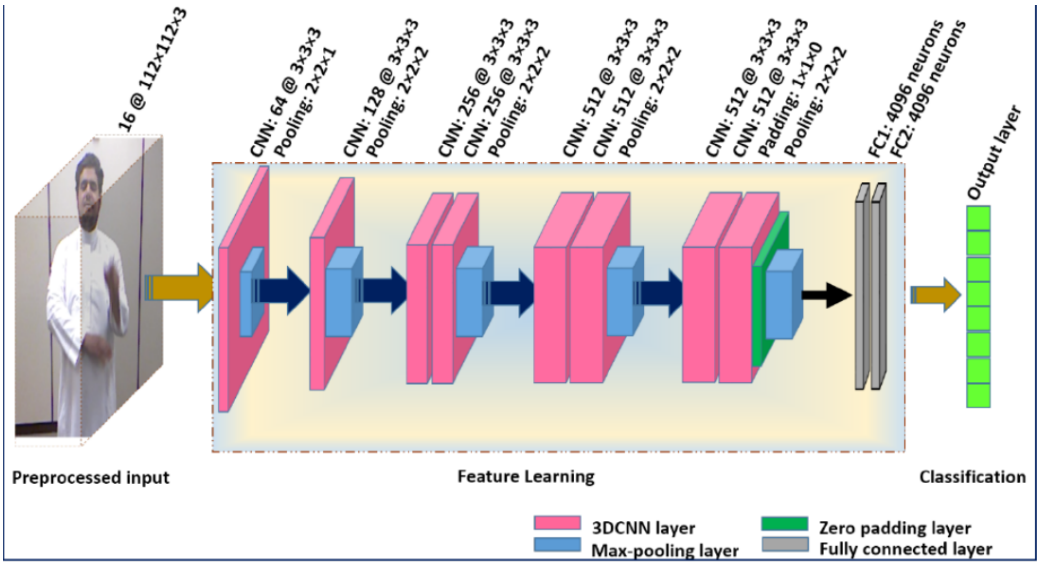
A study by Muneer Al-Hammadi, et al. claimed that there are two main purposes of gesture recognition, these two purposes are: To help with the growing community of deaf and hard-of-hearing population. And it has an extended use of vision-based and touchless applications. The proposed system utilizes the 3DCNN architecture for spatiotemporal feature learning using two approaches. In the first approach, features from the entire video sample were extracted with the use of 3DCNN. In their second approach, with the use of the 3DCNN they aimed to enhance the temporal dependency of the video frames[18].

Their first approach consists of three main phases: video preprocessing, feature learning, and classification. In the video preprocessing phases, the input video was converted into RGB frames sequence. Linear sampling was applied to preserve the order of the selected frames and fix the length of 16 frames; the corresponding indices of these frames are calculated in fig 2.15.1 where len(input) is the length of the input sequence. Spatial dimension normalization was used to overcome variations in the heights and distances of the signers from the camera. The final step was to resize the cropped square frames into 112x112 pixels[18].

$$\text{index}_i = \text{round}\left(\frac{\text{len(input)}}{16} \times i\right), i \in \{1,16\} \quad (1)$$

2.3.1 index of each frame calculation

In feature learning, to extract the local spatiotemporal features of gesture sequences 3DCNN was used. Transfer learning was also employed to beat the scarcity of a large labeled database of gestures. Their single 3DCNN-based structure is shown in fig 2.15.2 [18].



2.3.2 3DCNN Structure proposed by Muneer Al-Hammadi

To classify, the features extracted in the previous phases are fed into the SoftMax layer. Which in return would output the probability of each class.

$$SoftMax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}}$$

2.3.3 SoftMax calculation

`

2.5 Human Action Recognition using CNN and LSTM-RNN with Attention Model

According to Kuppusamy, The paper proposed the integration of a convolutional neural network and long short term memory recurrent neural network for processing the video. The process for the convolutional is the given output produces the informative spatial features.

The features that are extracted are directed to the long-short term memory module to generate temporal features. The feature maps of the long short-term memory component fed to the proposed attention element. This process captures the highly valuable informative features in the frame video. [39]

2.6 3D-CNN+LSTM: Deep Neural Networks for No-Reference

According to Varga et al, The purpose of no-reference video quality assessment algorithms that are based on the long-short term memory network is a pretrained convolutional neural network that is introduced. The study proposed the result of the experiments on KoNVID-1k demonstrate the proposed method outperforms the state-of-the-art algorithm. The results of the study are confirmed using the test on the LIVE Video Quality Assessment Database, which consist of artificially distorted videos. \

The study uses a reliable method of assessing the quality of digital videos through subjective evaluation. The paper also introduced an architecture for NR-VQA that utilizes deep features extracted from a pretrained CNN and LSTM network for sequence-to-one

regression The main purpose of the objective VQA is to design mathematical models that are able to predict the quality of the video assessed by humans. [40]

2.7 Google Colab - Build Large Deep Learning Models on your Machine!

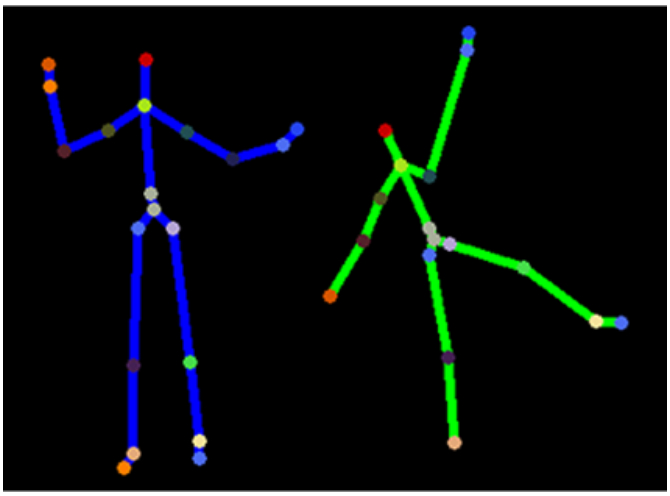
Google Colab is an online free cloud-based Jupyter notebook environment that allows users to train machine learning and deep learning models. In addition, various problems on local machines during executions are nowhere to be found since most of the dependencies are already installed and training time is also faster due to better hardware provided by Google[20].

THEORETICAL FRAMEWORK

According to the study of Tran et al, they mentioned that the hand gesture recognition is one of the other methods that humans use for interaction, in other words hand gesture recognition is another way on how humans can interact with another human with the help of computers. [12] In this study the researchers will use Hand Region Extraction, Fingertip Detection, and Hand Gesture Recognition:

Skeletal Tracking

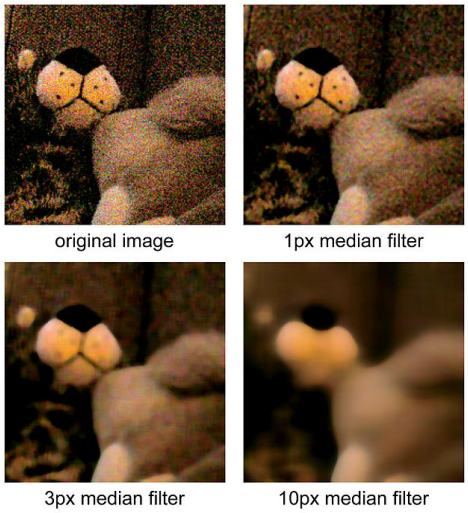
According to Intel, Skeletal Tracking uses sensors that are widely used in depth cameras to track the motion or the action of a human body [21]; this is similar to the motion capture that is being used in Game developing to Produce CGI characters. The researchers will use the Kinect v2 that has an improved Skeletal Tracking. This will help the researchers to detect the motion in Sign Language Gesture.



2.4.1 Skeletal Tracking System Sample Image [28]

Median Filter

According to ScienceDirect Media Filtering is a common nonlinear method for noise suppression that is unique. [22] Median filter is most useful in reducing random noise and especially when there is a dense noise amplitude. Median filtering is accomplished by sliding a window over the image. The researchers will use media filtering because this is very useful in the image processing for preserving the edges during the removal of the noise aptitudes in the image.



2.4.2 Sample Median Filter [31]

Morphological Processing

According to Nickson, Morphological processing is a collection of non-linear operations related to shape or morphology features in an image. Morphological processing is a technique that gives an opening to the image and erodes it then dilates the eroded image, using the same structuring element for the both operations.[23] The researchers will use this to remove small objects on the image while preserving the shape and size of the larger objects in the images.

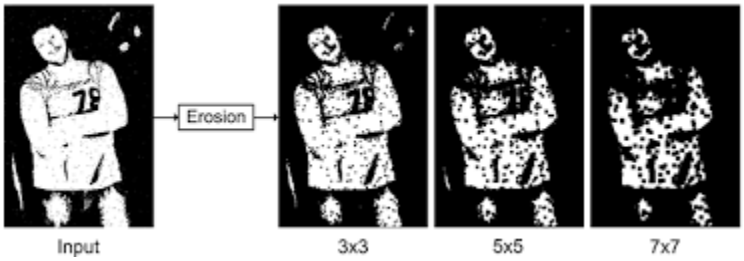


Fig 2.4.3 Erosion with different sized structuring elements [32]

Border-Tracing algorithm

According to a study of the school of Technical University of Chj-Napoca, [24] Border-Tracing algorithm is used to extract the contour of the objects or regions from an image.

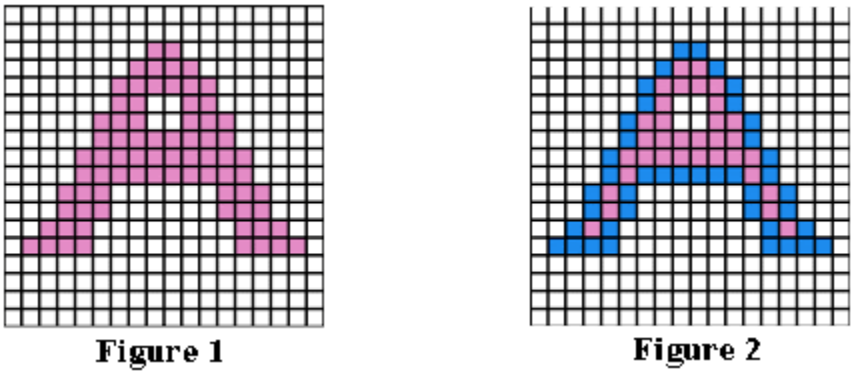


Fig 2.4.4 Sample Border Tracing Algorithm [33]

Moore–Neighbor algorithm

According to Image processing place, the main idea of Moore–Neighbor algorithm is every time you hit a black pixel, P, backtrack i.e. go back to the white pixel you were previously standing on then, go around pixel P in a clockwise direction, visiting each pixel in the Moore neighbor algorithm, until a black pixel value is found. [34]

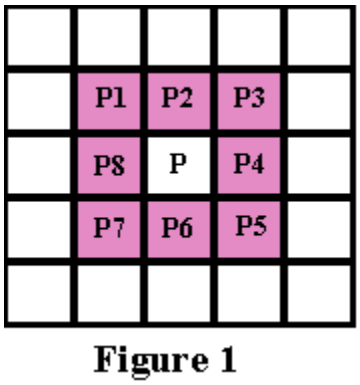


Fig.2.4.5 Moore Neighborhood [34]

K-cosine Corner Detection

The K-cosine algorithm is a major requirement in this study, because K-cosine will have a vital role in image processing and computer vision. Corners will be the representative for the contour-based corner detection algorithm. [34]

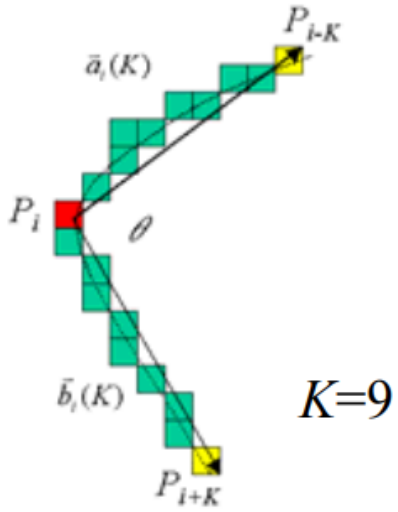


Fig.2.4.6 Curvature Measurement with K-cosine [35]

3D-CNN

According to Shen et al, Convolutional Neural Networks are the types of deep learning models that can directly extract data on raw inputs. In 3DCNN, the model extracts the features from the spatial and temporal dimension by performing a 3D convolutions to capture the motion information and to be encoded in every multiple adjacent frames. 3DCNN is a deep learning model that mostly analyzes the gesture and motions of the object. [36]

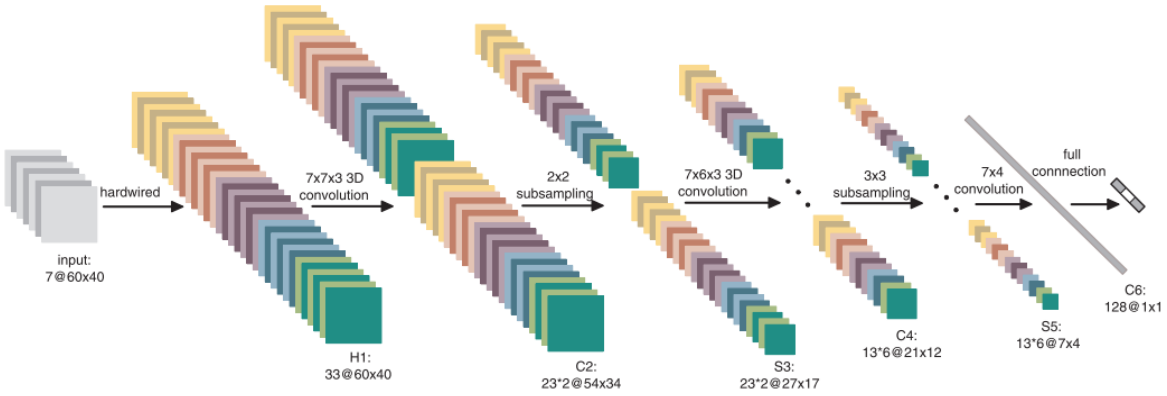


Fig 2.4.7 3DCNN Structure [36]

Long Short-Term Memory (LSTM)

According to the study of Brownlee. Long Short-Term Memory is a neural network that is capable of learning, that depends on the sequence of predictions. This occurs when there is a set of complex data like speech recognition, Machine translation and more. [27].

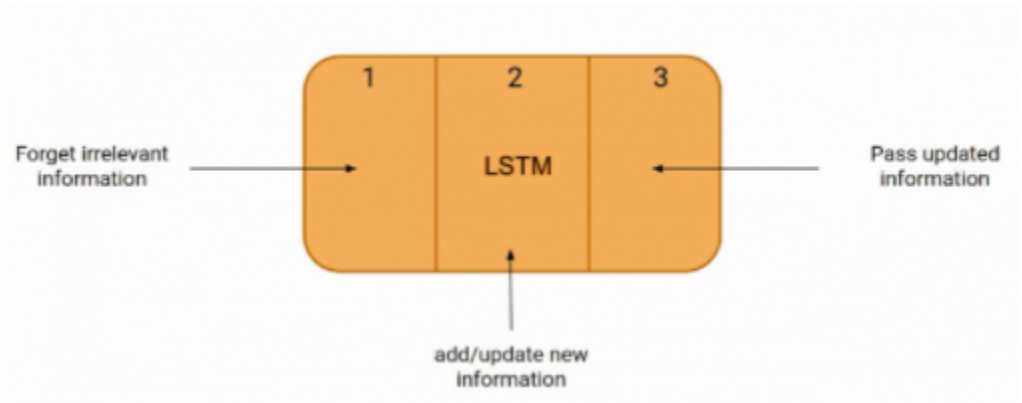


Fig 2.4.8 Long Short-Term Memory Network [37]

CHAPTER 3

METHODOLOGY

This section covers the implementation of the proposed *ASL* recognition model with the LSTM networks and the research methodology mentioned from the study of (*Tran et al., 2020*) . The proposed model will developed using Google Colaboratory :

Processor	Intel(R) Xeon(R) @2.3GHz
Memory	12GB
Graphics Card	Nvidia K80 12GB

3.1 Conceptual Framework

This is an overview of the proposed system based on the methods of Tran et al. [12] as presented in figure 3.1.1. The researchers would collect their input data using the Kinect Sensor v2 as their camera. After the data collection, the hand region of interest and the center of the palm would be extracted from the given data with the use of the Kinect Skeletal tracker. To remove the noise of the extracted data, the researchers would utilize the image processing technique of both morphological processing and median filtering – the hand contours are extracted and described using a border-tracing algorithm. After this, the hand contours are computed using the Moore-Neighbor algorithm. With the extracted hand contours, the researchers would use the K-cosine corner algorithm to detect the fingertip position based on the hand-contour coordinates model. The result will be transformed into the gesture initialization for the hand gesture spotting. Gesture Spotting will determine the beginning and end point of the hand point trajectory. Finally, the 3DCNN model will be used to extract unique features to the video dataset – the output will be used in the LSTM model to create

sequential features that will be utilized to correctly predict the hand gesture during actual detection using the Kinect Sensor.

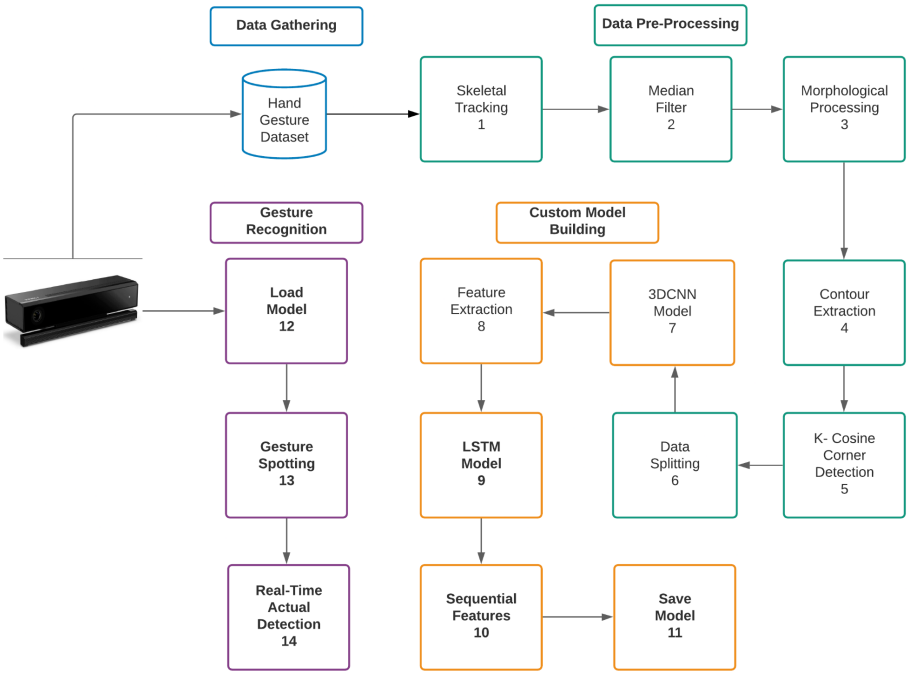


Fig 3.1.1 Proposed Conceptual Framework

3.2 Data Gathering

For this section, the researchers would create and provide their own dataset containing one handed ASL gestures using only the right hand for simplicity, these words are: *Police, Abstain, Commute, Reason, Badger, Airplane, Daily, Accident, Dart, Glad, Glance.* The inputs are taken using the Kinect Sensor v2 where the participants will be standing in front of the camera during the recording of the dataset – the participants will be recording themselves in full body at their homes while having a plain background with no obstructions or objects behind them. The participants will be composed of relatives living together with the researchers due to the restrictions caused by the pandemic. In addition, these participants are both genders and their

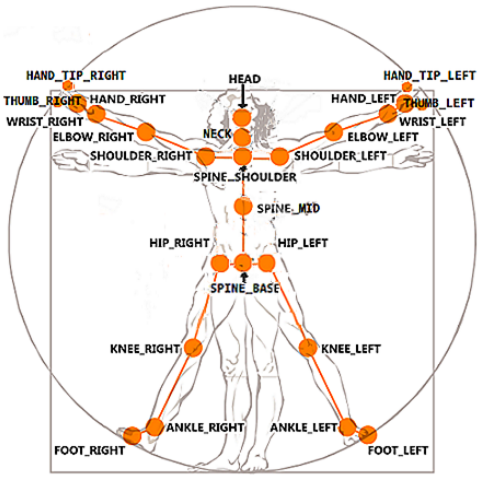
age group will fall under the ages of 17 to 50 to be standard as bones stop growing at the said age and variations happen as they reach the latter one[41]. The first five (5) videos of each hand gesture will be shot in normal lighting while the other five (5) videos will be shot with faint lighting. Moreover, each gesture is performed ten (10) times by the fifty (50) participants resulting in five-thousand five-hundred (5500) videos in total. The researchers will commission three (3) domains or experts to validate and annotate the dataset.

Fig 3.2.1 Sample Dataset of ASL words

3.3 Data Pre-processing

3.3.1 Skeletal Tracking (Kinect Sensor v2)

The depth images that were captured using the camera will be processed using the skeletal tracker and this method will map the coordinates of the subject's body parts in accordance to the known joints (i.e. *head, spine, hip & shoulder*) of the sensor. This function has greatly improved on the version 2 as it can track more joints (25) with higher precision compared to its predecessor[14].



3.3.1.a The Identified Joints of the Kinect Sensor v2 [38]

3.3.2 Image Processing (Median Filter & Morphological Processing)

Median Filter is used to reduce noise in an image. Additionally, the main goal of this process is to check the surrounding pixels for any difference in values – if found, that pixel value will be replaced by the median of neighboring values[26]. On the other hand, Morphological processing is more inclined on the ordering of pixel values which means this process is related to the processing of binary images[30].

3.2.3 Contour Extraction

Moore-Neighbor tracing is a common algorithm that is applied to extract the contours of a pattern or object from a digital image. This process works by finding the group of black pixels from a background of white pixels and then continuing until the start pixel has been found again[24].

3.2.4 K-Cosine Detection

The *K-Cosine Algorithm*[22][19] will be used to extract the fingertips based on detected hand contours. The method is used to determine the angle between the vectors of a finger. It computes the fingertip points using the coordinates of the detected hand contour. In the end, this process would give the researchers a set of pixels belonging to the hand and fingertip locations.

3.3 Custom Model Building

3.3.1 Training the Model

The normal CNN uses a 2D matrix on its filters and inputs while the 3DCNN uses the same operations but differs with 3D layers and filters[42].The training process will begin by installing the required dependencies or libraries from Keras and Scikit-Learn in Google Colab. The dataset will then be splitted into train, test and validation. In addition, the train_test_split function will be used to split the data into the said subsets – 70% of the sample data will be used to train the custom model[42]. The 3D Convolutional Neural Network will be used to extract features on the data input as it captures spatio-temporal information on videos[12]. Additionally, the features extracted from the previous process will be fed to the LSTM network in order to obtain sequential features from each video[43] which will be used to predict the hand gestures.

3.3.2 Validating the Model

For the Validation, 20% of the input data will be processed, two metrics from Scikit-Learn will be used to assess the performance of the model. The first is *accuracy*,

which will provide the correct predictions per input and the partial performance of the model. Second is the *confusion matrix* from the same library, to oversee the classification score for each gesture.

3.3.3 Testing the Model

After the whole process of training and validation, the proposed model will be tested using the remaining 10% of unseen data to avoid overfitting. To finalize, a real-time actual detection will be implemented to test the final model. This step will ensure that the proposal model is well-made and prediction shall be done using the dataset to assess the whole model's accuracy.

REFERENCES:

[1]World Health Organization. (2021, April 1). *Deafness and hearing loss*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2]World Health Organization. (2021, March 1). *WHO: 1 in 4 people projected to have hearing problems by 2050*. World Health Organization. https://www.who.int/news/item/02-03-2021-who-1-in-4-people-projected-to-have-hearing-problems-by-2050.

[3] Oudah, M., Al-Naji, A., & Chahl, J. (2020, July 23). *Hand Gesture Recognition Based on Computer Vision: A Review of Techniques*. MDPI. https://www.mdpi.com/2313-433X/6/8/73.

[4] Quer, J., & Steinbach, M. (2019, February 19). *Handling Sign Language Data: The Impact of Modality*. Frontiers. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00483/full.

[5] Media, A. I. (2020, October 27). *Sign Language Alphabets From Around The World - Ai-Media*. AiMedia. https://www.ai-media.tv/sign-language-alphabets-from-around-the-world/#:~:text=There%20are%20somewhere%20between%20138,sign%20language%20as%20each%20other.

[6] Shrestha, U. (2019, May 9). The rise of technology and impact on skills. Retrieved March 28, 2021, from https://www.tandfonline.com/doi/full/10.1080/14480220.2019.1629727

[7]Hussain, Z. (2019, June 12). *Different Approaches for Human Activity Recognition: A Survey*. ResearchGate. https://www.researchgate.net/publication/333745638_Different_Approaches_for_Human_Activity_Recognition_A_Survey.

[8] Shubham Jadhav, Pratik Shah, Avinash Bagul, Parag Hoshing, Ashutosh Wadhvekar. (2016, November 10) Review on Hand Gesture Recognition using Sensor Glove. https://www.ijarcce.com/upload/2016/november-16/IJARCCE%20120.pdf

[9] Wangchuk, K., Riyamongkol, P., & Waranusast, R. (2020, September 03). Real-time Bhutanese sign Language Digits recognition system using convolutional neural network. Retrieved March 28, 2021, from https://www.sciencedirect.com/science/article/pii/S2405959520301685

[10] Alok, K., Mehra, A., Kaushik, A., & Verma, A. (2020, January). Hand Sign Recognition using Convolutional Neural Network. Retrieved April 17, 2021, from https://www.irjet.net/archives/V7/i1/IRJET-V7I1294.pdf

[11]Hakim, Noorkholis L.; Shih, Timothy K.; Kasthuri Arachchi, Sandeli P.; Aditya, Wisnu; Chen, Yi-Cheng; Lin, Chih-Yang. 2019. "Dynamic Hand Gesture Recognition Using 3DCNN and LSTM with FSM Context-Aware Model" *Sensors* 19, no. 24: 5429. https://doi.org/10.3390/s19245429

[12]Tran, D., Ho, N., & Yang, H. (2020, January 20). *https://www.researchgate.net/publication/338701325_Real-Time_Hand_Gesture_Spotting_and_Recognition_Using_RGB-D_Camera_and_3D_Convolutional_Neural_Network*. ResearchGate. https://www.researchgate.net/publication/338701325_Real-Time_Hand_Gesture_Spotting_and_Recognition_Using_RGB-D_Camera_and_3D_Convolutional_Neural_Network.

[13]Szymczyk, M. (2017, December 19). *How Does The Kinect 2 Compare To The Kinect 1?* Zugara. http://zugara.com/how-does-the-kinect-2-compare-to-the-kinect-1#:~:text=%E2%80%9CThe%20Kinect%20v2%20face%20recognition,and%20motion%20of%20certain%20objects.&text=Although%20the%20first%20Kinect%20used,has%20greatly%20improved%20upon%20it.

[14]Ghost, S. (2018, October 16). *The difference between Kinect v2 and v1*. The Ghost Howls. https://skarredghost.com/2016/12/02/the-difference-between-kinect-v2-and-v1/#:~:text=Well%2C%20first%20of%20all%20if,to%20carry%20and%20to%20install.

[15]Rizvia, M. S. (2020, October 19). *CNN Image Classification: Image Classification Using CNN*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/.

[16]Ghost, S. (2018, October 16). *The difference between Kinect v2 and v1*. The Ghost Howls. https://skarredghost.com/2016/12/02/the-difference-between-kinect-v2-and-v1/#:~:text=Well%2C%20first%20of%20all%20if,to%20carry%20and%20to%20install.

[17]Blog, C. (2020, May 22). *Most Popular American Sign Language (ASL) Phrases - You Need to Know*. Cudoo Blog. https://cudoo.com/blog/most-popular-sign-language-phrases-you-need-to-know/.

[18]Al-Hammadi, M., Muhammad, G., & Abdul, W. (2020, April 9). *Hand Gesture Recognition for Sign Language Using 3DCNN*. ResearchGate. https://www.researchgate.net/publication/340972266_Hand_Gesture_Recognition_for_Sign_Language_Using_3DCNN.

[19] Tran, D.-S., Ho, N.-H., Yang, H.-J., Kim, S.-H., & Lee, G. S. (2020, November 23). *Real-time virtual mouse system using RGB-D images and fingertip detection*. Multimedia Tools and Applications. https://link.springer.com/article/10.1007/s11042-020-10156-5.

[20] Sharma, A. (2020, December 29). *Use Google Colab for Deep Learning and Machine Learning Models*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/google-colab-machine-learning-deep-learning/#1

[21] *Skeletal Tracking Overview*. Intel® RealSense™ Depth and Tracking Cameras. (2020, March 11). https://www.intelrealsense.com/skeletal-tracking/.

[22] *Median Filter*. Statistics.com: Data Science, Analytics & Statistics Courses. (2021, May 5). https://www.statistics.com/glossary/median-filter/.

[23]Joram, N. (2020, January 1). *Morphological Operations in Image Processing*. Medium. https://himnickson.medium.com/morphological-operations-in-image-processing-cb8045b98fcc.

[25]Admin. (2019, April 18). *Top ASL Signs for Medical Emergencies*. Learn Sign Language. http://www.learnsignlanguage.com/2019/04/18/top-asl-signs-for-medical-emergencies/.

[26]*Median Filter*. Spatial Filters - Median Filter. (2017.). https://homepages.inf.ed.ac.uk/rbf/HIPR2/median.htm.

[27]Brownlee, J. (2021, July 6). *A Gentle Introduction to Long Short-Term Memory Networks by the Experts*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/.

[28]*Kinect for Windows SDK Beta*. Microsoft Research. (2020, March 14). https://www.microsoft.com/en-us/research/project/kinect-for-windows-sdk-beta/.

[24] Seo, J., Chae, S., Shim, J., Kim, D., Cheong, C., & Han, T.-D. (2016, March 9). *Fast Contour-Tracing Algorithm Based on a Pixel-Following Method for Image Sensors*. Sensors (Basel, Switzerland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4813928/.

[30] Sundararajan D. (2017) Morphological Image Processing. In: Digital Image Processing. Springer, Singapore. https://doi.org/10.1007/978-981-10-6113-4_8

[34]Ghuneim, A. (2016.). Contour Tracing. http://www.imageprocessingplace.com/downloads_V3/root_downloads/tutorials/contour_tracing_Abeer_George_Ghuneim/moore.html.

[35]]T. Sun, C. Lo, P. Yu and F. Tien, "Boundary-based corner detection using K-cosine," (2016) https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.6940&rep=rep1&type=pdf

[34] *corner detection algorithm*. corner detection algorithm: Topics by Science.gov. (2016.). https://www.science.gov/topicpages/c/corner+detection+algorithm.html.

[31] Shipra SaxenaShipra is a Data Science enthusiast, 2021 https://www.wikiwand.com/en/Median_filter

[32] *Morphology (Introduction to Video and Image Processing) Part 2*. whatwhenhow RSS. (2016.). http://what-when-how.com/introduction-to-video-and-image-processing/morphology-introduction-to-video-and-image-processing-part-2/.

[33] Ghuneim, A. (2016.). *Contour Tracing Algorithms*. Contour Tracing. http://www.imageprocessingplace.com/downloads_V3/root_downloads/tutorials/contour_tracing_Abeer_George_Ghuneim/alg.html.

[36] Three-dimensional convolutional neural network (3D-CNN) for Satellite behavior discovery (April,12,2021) https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11755/117550C/Three-dimensional-convolutional-neural-network-3D-CNN-for-satellite-behavior/10.1117/12.2589044.short?SSO=1&tab=ArticleLinkCited https://sci-hub.se/10.1109/TPAMI.2012.59

[37] Shipra SaxenaShipra is a Data Science enthusiast. (2021, March 18). LSTM: Introduction to LSTM: Long Short Term Memory. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/.

[38]Ghuneim, A. (2016.). *Contour Tracing Algorithms*. Contour Tracing. http://www.imageprocessingplace.com/downloads_V3/root_downloads/tutorials/contour_tracing_Abeer_George_Ghuneim/alg.html.

[39] Pothanaicker, Kuppusamy. (2019). Human Action Recognition using CNN and LSTM-RNN with Attention Model. 8. 1639-1643.

[40] Varga, Domonkos & Szirányi, Tamás. (2019). No-reference video quality assessment via pretrained CNN and LSTM networks. Signal, Image and Video Processing. 13. 10.1007/s11760-019-01510-8.

[41]*Causes: Age and bone strength*. Royal Osteoporosis Society - Osteoporosis Charity UK. (2016, July 21). https://theros.org.uk/information-and-support/osteoporosis/causes/age-and-bone-strength/#:~:text=Bones%20stop%20growing%20in%20length,slowly%2C%20until%20your%20late%20twenties.

[42] Chan, C. H. M. (2021, July 26). *Step by Step Implementation: 3D convolutional neural network in Keras*. Medium. https://towardsdatascience.com/step-by-step-implementation-3d-convolutional-neural-network-in-keras-12efbdd7b130.

[43]Ouyang, X. (2019, March 16). *A 3D-CNN and Lstm Based Multi-Task learning architecture for action recognition*. IEEE Xplore. https://ieeexplore.ieee.org/document/8677269.