ECE 445

SENIOR DESIGN LABORATORY

PROJECT PROPOSAL

An Intelligent Assistant Using Sign Language

Team #27

HANWEN LIU (hanwenl4@illinois.edu)
YIKE ZHOU (yikez3@illinois.edu)
HAINA LOU (hainal2@illinois.edu)
QIANZHONG CHEN (qc19@illinois.edu)

TA: Xiaoyue Li

March 8, 2023

Contents

1 Introduction

1.1 Problem

An intelligent assistant (IA) is software that can provide services and interact with the user, typically by performing automated tasks and assisting with daily activities. With the advent of computer vision and natural language processing technologies and the emergence of smart home accessories, intelligent assistants have revolutionized how people interact with technology.

Most intelligent assistants use voice user interface (VUI) as a primary means of communication. Some of the most prevalent examples include Siri from Apple, Cortana from Microsoft, and Alexa from Amazon. VUI provides several advantages, such as hands-free operation, faster input, and greater convenience. However, VUI is not always suitable for people with hearing or speech problems, hindering their ability to use these intelligent assistants.

People with hearing or speech impairments often face significant challenges in accessing information, participating in social interactions, and performing daily activities. Therefore, developing technologies that meet their unique needs and facilitate their communication and engagement can significantly improve their quality of life

1.2 Solution, Visual Aid

fix grammar and tense

We propose to develop an intelligent assistant that uses sign language as its primary communication standard. Sign language will enable people with hearing or speech impairments to interact with intelligent assistants effectively. By leveraging the latest advancements in computer vision and natural language processing, our intelligent assistant will recognize sign language and respond in real-time, making it a powerful and accessible tool for a broader range of users.

In recent years, intelligent assistants are becoming more personalized. Some assistants use data analysis to learn about individual users and their preferences and provide more accurate, relevant, and reasonable services and recommendations. Our intelligent assistant using sign language consists of four subsystems: Input and output subsystem, Gesture recognition subsystem, System and control subsystem, and Bionic hand subsystem.

The input and output subsystem includes a camera that receives input from the user through sign language and a displayer that reveals the interaction between the user and the intelligent assistant to those who do not understand sign language.

The gesture recognition subsystem receives the visual signal from the input and output subsystems. Gesture recognition subsystem applies *You Only Look Once* (YOLO)[1], an object detection algorithm, to detect the 3-dimensional position of the user's hand. Then it utilizes *Mediapipe*[2], developed by Google, to collect the relevant position for the wrist in real-time.

The system and control subsystem receives the decision of the gesture, which the bionic hand needs to do from the gesture recognition subsystem. It translates it to Pulse-Width Modulation (PWM) signals to control the movement of servo motors. Our control system uses Microcontroller Unit (MCU) to output signals and an advanced computing unit to deploy the machine learning model.

The bionic hand subsystem is responsible for communication with the user.

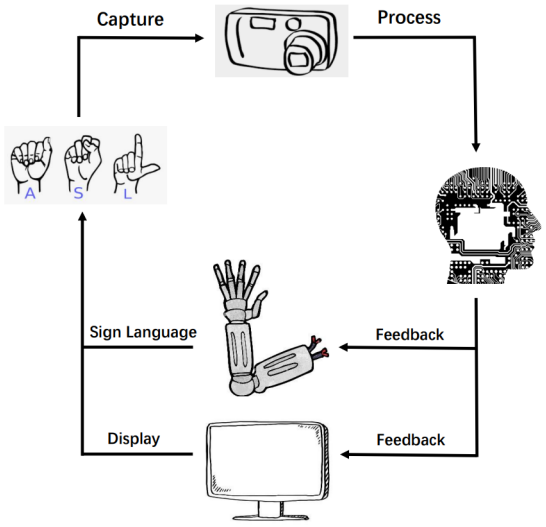

Figure 1: Visual Aid

1.3 High-level Requirements

Build an end-to-end model using the Mediapipe framework combined with different machine learning models, including Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Gate Recurrent Unit (GRU).

To implement a good interaction experience, the time used by the user doing sign language to display the dialogue and the response of the bionic hand should be at most 30 seconds.

The bionic hand can move free and fluently as designed, all of the 12 degrees of freedom fulfilled; the movement of a single joint of the finger does not interrupt or be interrupted by other movements; the bionic hand could work for one hour in a roll and two years in total.

2 Design

2.1 Block Diagram

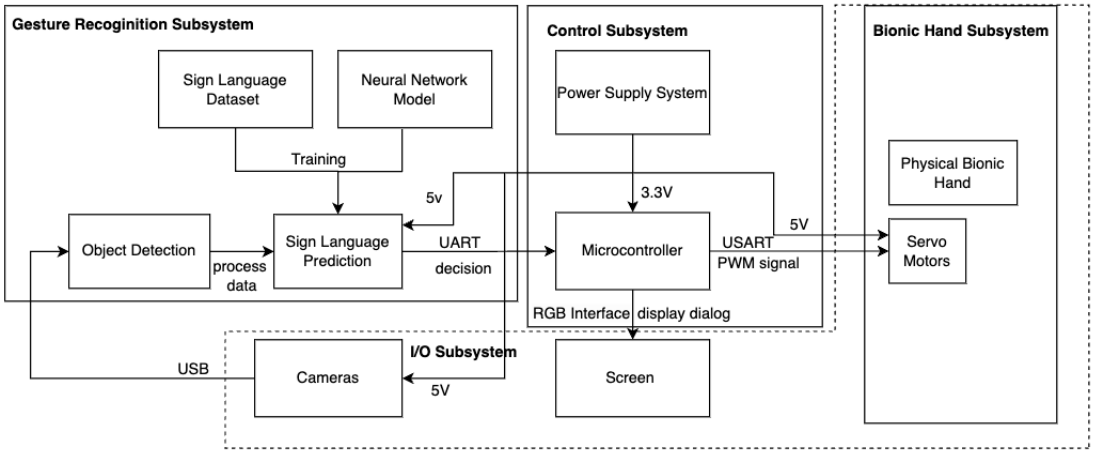

Figure 2: High Level System Overview

2.2 Pilot Mechanical Design

update mechanical design

In this part, we demonstrate the system overview in Fig. 3 and Fig.4. Our system includes two bionic hands, one camera for gesture recognition, one STM32 micro-controller and development board, and one Liquid Crystal Display (LCD) screen to visualize the information.

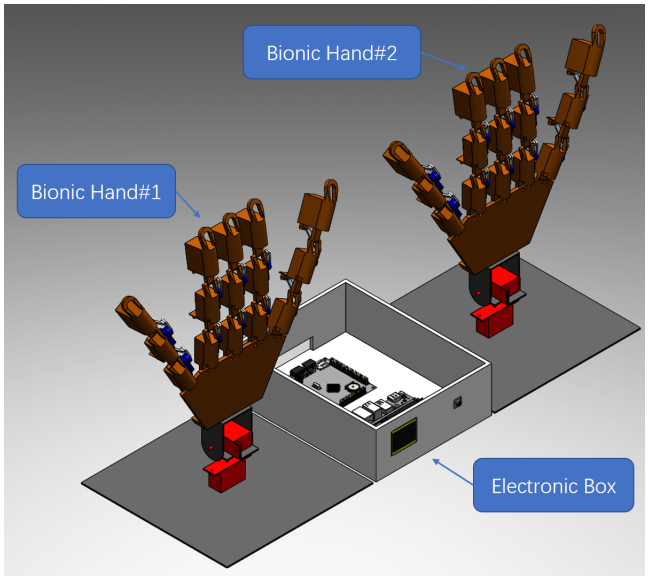

Figure 3: Overview of the System Physical Design

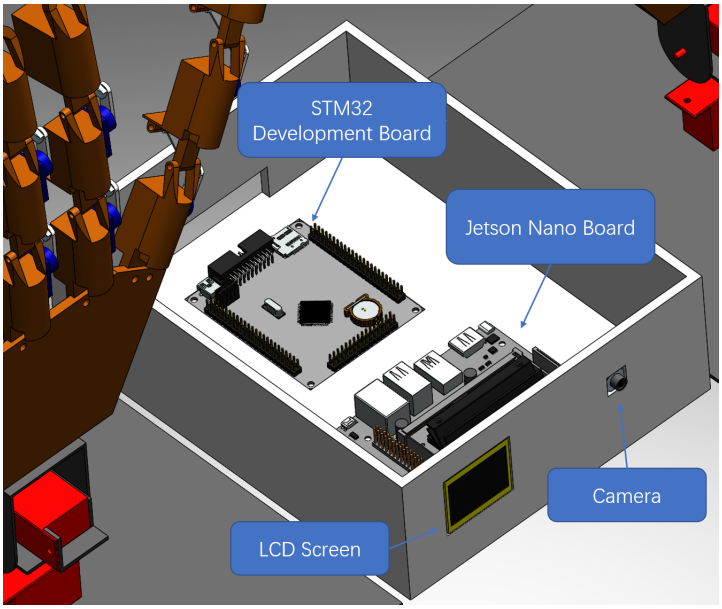

Figure 4: Detailed Layout of Electronic Box

2.3 Subsystem Overview

2.3.1 Gesture Recognition Subsystem

The Gesture Recognition Subsystem which consists of object detection and sign language prediction is mainly used to extract features from the images passed by the camera and then feed them into a pre-trained model for further prediction.

Connection with Other Subsystems:

- Connection to the input and output subsystem: receive the image delivered by the camera.
- Connection to The System & Control Subsystem: pass the prediction results for further control used.

2.3.2 System & Control Subsystem

The System & Control Subsystem is aimed at translating from sign language predictions to PWM signals and delivering the signals to servo motors, which makes it possible for communication and control of the whole systems. It contains one development board with a MCU and a computing unit which supports real-time gesture recoginition.

Connection with Other Subsystems:

- Connection to the power source: Our MCU is driven by 3.3V.
- Connection to the gesture recognition subsystem: Receives predictions from machine learning model.

- Connection to the bionic hand subsystem: Output PWM signals to each servo motor to control the motion of the bionic hand.
- Connection to the input and output Subsystem: The camera will be connected to our computing unit to capture the gesture, and LCD will be connected to MCU via I2C.

2.3.3 Bionic Hand Subsystem

The bionic hand subsystem is consist with two identical bionic hands, forming a system that has 24 degree of freedom (DOF) in total. The bionic hand subsystem is responsible of delivering the motion planned by control system and interact with system's goal customers directly. It will interpret the information generated by AI and execute the decision with sign language.

Connection with other subsystems:

- Connection to the power source: All of the servo motors are driven by 4.8V - 7.2V DC power provided by micro-controller development board in Control Subsystem.
- Connection to the Control Subsystem: Motion of each servo motor is controlled with PWM signal generated by micro-controller in control system.

2.3.4 Input & Output Subsystem

The input and output subsystem includes one camera and one LCD. Camera module is used to capture user's hand gesture as input data. LCD is used to display sign language dialogs to other people as text.

Connection with other subsystems:

- Connection to the power source: The camera and LCD are connected to a 5V power source. Jetson Nano board and MCU supply the camera and LCD power, respectively.
- Connection to the control subsystem: Jetson Nano board and MCU supply power to the camera and LCD, respectively. The camera inputs hand gesture data through a USB wire, and the LCD displays sign language dialogues.

2.4 Subsystem Components

2.4.1 Gesture Recognition Subsystem

add detailed description

Object Detection:

Considering the cost-effectiveness and convenience, we proposed adopting computer-vision-based techniques to detect objects rather than sensor-based ones. However, most

computer vision-based methods consisting of gesture segmentation and hand shape estimation have high demands on high computing power, which indicates that a platform with robust processors is needed. Given that we do not have the equipment to support those methods, we select an open-sourced framework developed by Google, called MediaPipe, to detect the users' body movements. By using the machine learning pipeline under MediaPipe's Hand Tracking solution and Pose solution, we can accurately get high-fidelity 3D coordinates (i.e., x, y, z-axis) key points quickly, which is lightweight enough to run in real-time devices.

Sign Language Prediction:

We treated recognizing static signs image as a baseline, and the final goal is completing a dynamic recognition subsystem. As shown in Figure 5, we need a pre-trained model to predict the meaning of American Sign Language input images. Before introducing the machine learning algorithms and models we selected, we must have enough high-quality datasets for training and testing our models. We plan to take four hundred photos of each sign language that must be classified for the static recognition dataset and combine open-sourced datasets on Kaggle if more samples are needed. For dynamic recognition, each sign movement sample will be recorded in a video with the same length. Each video should contain 30 frames, an appropriate length for us to conduct a complete sign language. Since video recording is time-consuming, we plan to record 40 videos for each type of sign movement. In case of lacking training data, we may record more on demand.

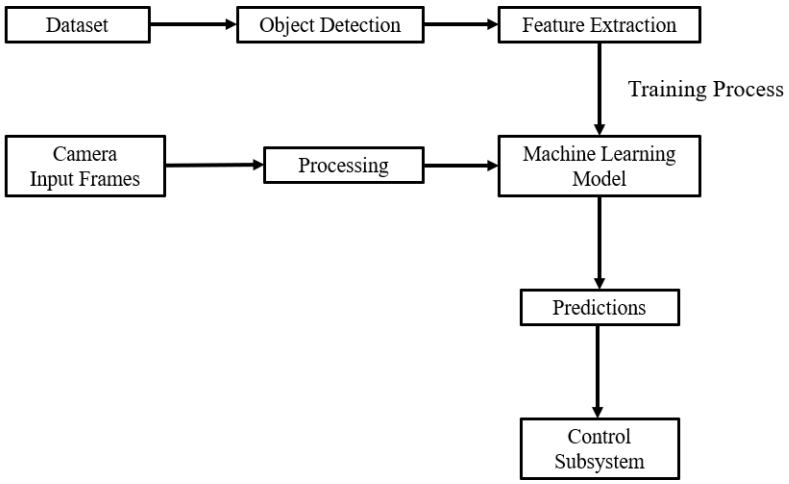

Figure 5: Gesture Recognition Subsystem Workflow

After obtaining sufficient datasets, we could utilize MediaPipe to extract the critical landmarks on hands and poses. We learned that the returned landmarks change as the relative position of the hand in the image changes. To solve the problem, we will select one fixed point as a reference and then adjust all other landmarks accordingly relative to the selected point. In this case, we could successfully eliminate the influence of location effects

on the outcome. After pre-processing, we could use the updated coordinates information as the data features and feed them to models for training.

We will not deploy complicated neural networks in static recognition tasks because many traditional machine learning algorithms can perform well. Considering the number of features is much lower than the number of samples when dealing with static classification tasks, we adopted Support Vector Machine (SVM), which performs better in high-dimensional space than other traditional models[3]. We will test the model in the different kernels to find the one that best fits our dataset. In the case of dynamic recognition, we proposed adopting recurrent neural networks (RNNs), which contain memory storing information from previous states' computations. Thus, they can deal with time series and sequential data. While traditional RNN may occur problems of gradient vanishing, we planned to adopt LSTM, a variant of RNN, to prevent the potential risk. Due to the possible lack of data, we may encounter the problem that our data set is insufficient to train LSTM with a large number of parameters well. Hence, GRU is our alternative solution for its less parameter, which means we could perform better with limited data and more efficiently[4]. In this case, we proposed to adopt GRU and LSTM as our machine learning.

2.4.2 System & Control Subsystem

MCU:

We choose STM32 as our MCU. It is responsible for controlling the motion of 24 servo motors on the bionic hand by sending PWM signals. The specific motion signal should be generated based on the decision from computing unit through uart.

Computing Unit:

We will use Jetson Nano as our computing unit in this project. Jetson Nano is a small, powerful computer that can run multiple neural networks in parallel for applications like image classification, object detection and speech processing [5]. We use this platform to deploy our sign language recognition machine learning model. It will intake data from the camera unit, and perform computing onboard. After getting the result from the Gesture Recognition Subsystem. The decision will be sent to MCU via uart. It should be powered by development board.

2.4.3 Bionic Hand Subsystem

The bionic hand subsystem consists of two identical bionic hands, forming a system with 24 degrees of freedom (DOF). The bionic hand subsystem is responsible for delivering the motion planned by the control subsystem and interacting with the user directly. From bottom to top, each hand has a moveable platform, 10 SG-90 servo motors, and a plastic hand. The moveable platform, driven by two RDS3115 digital servo motors, holds the plastic bionic hand and provides two extra DOF. The SG-90 servo motor is directly fixed inside the fingers, and its output shaft will connect the finger's moveable part with a linkage and drive the finger to rotate around the joint. Therefore, the finger part, motor,

and linkage will form a basic 4-bars link system and move smoothly. The combination of fingers' movements will form different gestures. The bionic plastic hand comprises 26 parts, as shown in Fig. 6. We plan to manufacture them with 3D printing using PLA material.

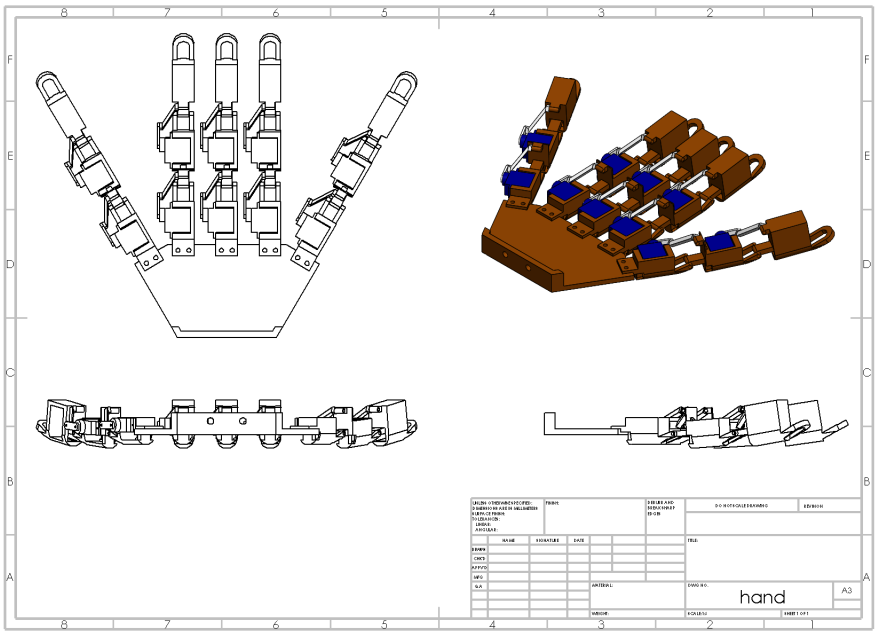

Figure 6: Engineering Drawing of the Bionic Hand

2.4.4 I/O Subsystem

Camera:

A webcam is used for capturing real-time hand gesture of the user. The input resolution should be 1080p. It will be connected to Jetson Nano board by USB wire.

LCD:

When user using sign language to communicate with our robot, not only our bionic hands will give users feedback but also the dialog between users and robots will be displayed on LCD as text.

2.5 Subsystem Requirement

add more quantitative and detailed requirements

2.5.1 Gesture Recognition Subsystem

1. The Gesture Recognition Subsystem should recognize different American Sign Language correctly and react with corresponding gestures without obvious delay.

2.5.2 System & Control Subsystem

1. The microcontroller needs to output 20 stable PWM signals with 50 Hz frequency. Stable means the error of motor deflection angle should not exceed 15 degrees.
2. The edge computing platform we choose should have high performance when running the dynamic gesture recognition model.
3. The delay from microcontroller receives decision message from Jetson Nano board to output corresponding PWM signal should be less than 5 second.
4. The degree of servo motors should not exceed the maximum degree. Otherwise, the mechanical parts will be damaged.

2.5.3 Bionic Hand Subsystem

1. For each hand, 12 DOF can be fulfilled as designed.
2. Each finger's two move-able links can move freely, rapidly, and accurately under the drive of servo motor.
3. Under the PWM signals generated by control subsystem, the movement combinations of two hands' fingers and platform form a series of gestures.
4. The bionic hand subsystem is stable and durable, without the risk of stuck or even falling apart.

2.5.4 I/O Subsystem

1. Camera should have enough resolution to capture the characteristics of hand.
2. User standing 0.5m - 1.5 m in front of our camera can be captured by Jetson nano camera easily.
3. LCD driven voltage needs to be 3.0V-5.0V.
4. Character shown on the LCD screen should display at a fast speed. Whole dialog should be shown on the screen in 5 seconds.

2.6 Tolerance Analysis add more detailed mathematical tolerance analysis

Given that our machine learning models' performance depends to a large extent on the dataset's quality. Especially in dynamic sign language recognition, our dataset should be established by recording as clips of videos, and it is inevitable for us to generate a lot of low-quality data due to various reasons, including the camera's resolution, lighting conditions, and the correctness of the recorders' sign language movement.

We will evaluate the results using performance matrices, including accuracy, precision, recall, and F1 score, to evaluate our outcome quantitatively. Accuracy represents correctly predicted labels from the whole dataset, as shown in Equation 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision measures the number of actual positives in all the positives predicted by the

model, which is a good measurement when the cost of False Positive is high. Equation 2 gives the mathematical formulation. Recall calculates the number of actual positives predicted correctly by our model, which is a good measurement when the cost of a False Negative is high. Equation 3 gives the mathematical formulation. The F1 score combines Precision and Recall and represents both properties. Equation 4 gives the mathematical formulation.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Recall = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

In the above equations, TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively. Confusion Matrices are also proposed to understand the models' performance better.

Concerns related to the bionic hand subsystem mainly circled with fingers' motion smoothness and material strength. The core system can be abstracted to a four-bar linkage for motion smoothness. Therefore, detailed kinematics and dynamics simulation is necessary. Initially, we referenced the simulation tool used in UIUC TAM 212 course**four-bar-linkages**. Due to the space limitation, we had the presupposed ground link length (g) and input link length (a). By changing the output link length (b) and floating link length (f), we have combinations of four-bar linkages that can deliver satisfied trajectories on the output node, which links to the finger part driven by the motor. We conduct dynamics simulation for those candidate four-bar linkages using the MATLAB program developed in UIUC ME 370 lab and project. By calculating the position, velocity, acceleration, and torque on the output node for the whole cycle, we chose the best four-bar linkage that suffers little impact and vibration. The kinematic and dynamic simulation results are shown below in Fig.7 and Fig. 8. The formulas we used are listed below:

$$\text{Grashof index:} \quad \boldsymbol{G = s + l - p - q \geq 0} \tag{5}$$

$$\text{Validity index:} \quad \boldsymbol{V = l - s - p - q \geq 0} \tag{6}$$

where $\boldsymbol{l}$ is the longest link, $\boldsymbol{s}$ is the shortest link, and $\boldsymbol{p}$, $\boldsymbol{q}$ are the rest two links.

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ -R_{12y} & R_{12x} & -R_{32y} & R_{32x} & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & R_{23y} & -R_{23x} & -R_{43y} & R_{43x} & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & u & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} F_{12x} \\ F_{12y} \\ F_{32x} \\ F_{32y} \\ F_{43x} \\ F_{43y} \\ F_{14y} \\ T_{12} \end{bmatrix} = \begin{bmatrix} m_2 A_{CG2x} \\ m_2 A_{CG2y} \\ I_{CG2}\alpha_2 \\ m_3 A_{CG3x} - F_{Px} \\ m_2 A_{CG2y} - F_{Py} \\ I_{CG3}\alpha_3 - R_{Px}F_{Py} + R_{Py}F_{Px} \\ m_4 A_{CG4x} \\ 0 \end{bmatrix} \tag{7}$$

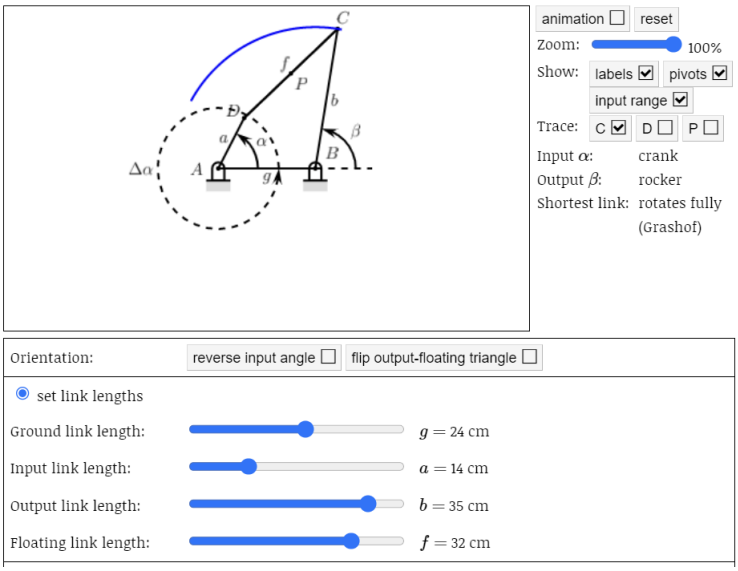

Figure 7: Kinematics Simulation Results of Proposed Four-bar Linkages

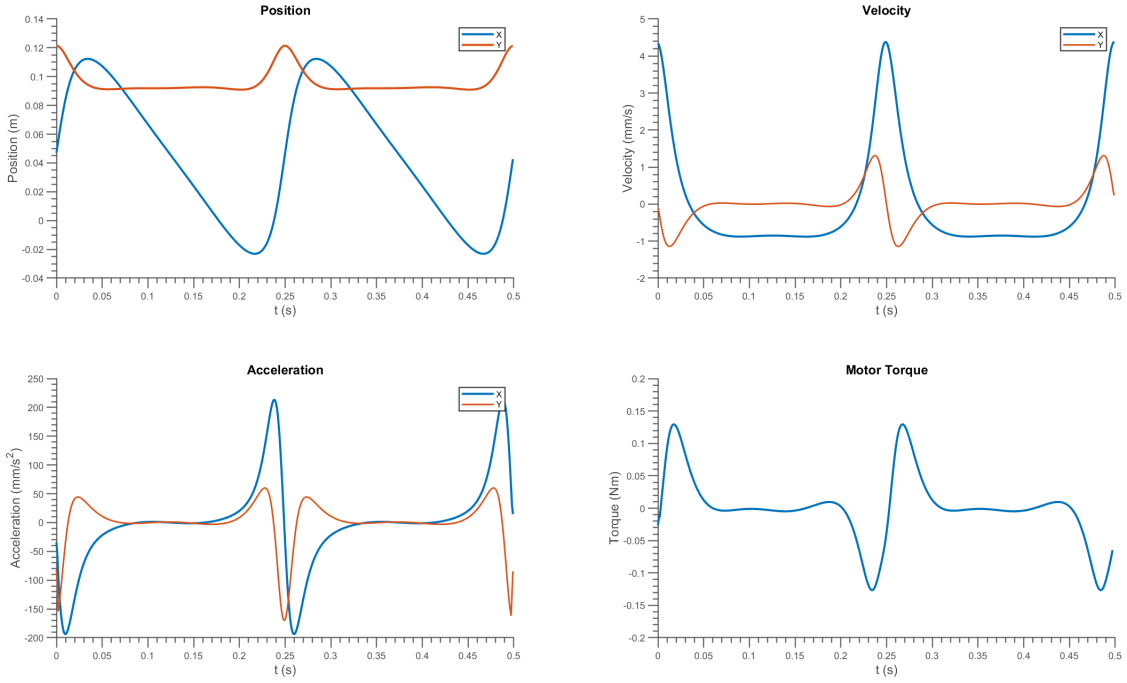

Figure 8: Dynamics Simulation Results of Proposed Four-bar Linkages

For critical parts, we conduct statics simulation and failure analysis with computer aids engineering (CAE) software to verify and improve our design. For instance, as demonstrated in Fig .9, the palm we initially designed have right-angle sides, which may cause stress concentration and even yielding under impact. Afterwards we strengthen the part and fillet the right-angle sides. The optimized design performs much better under the same load conditions as shown in Fig. 10.

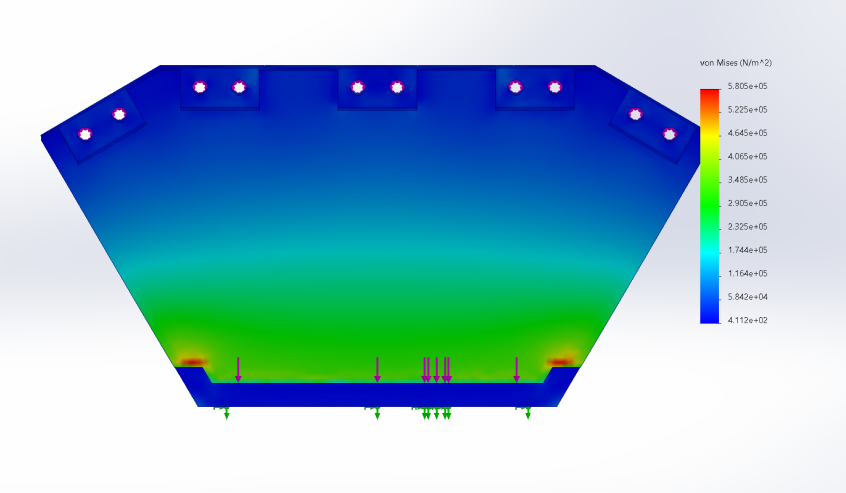

Figure 9: Statics Simulation for Initial Palm

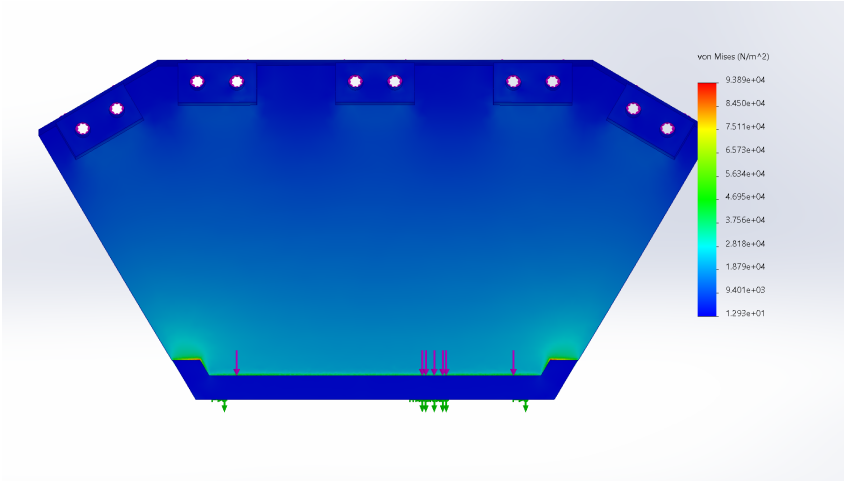

Figure 10: Statics Simulation for Optimized Palm

The tolerance analysis lies on control subsystem concentrates on the stability of PWM signals. In order to make sure the signals would not interfere with each other, and servo motors would be move correctly, some other control modules might also be added to our development board.

3 Ethics & Safety

When designing a product, safety is our top priority. To ensure "the safety, health, and welfare of the public"[6], it is vital to notify users of potential hazards and minimize the possibility of systematic danger caused by misuse of our work. This means guaranteeing that users are aware of any potential risks associated with using our intelligent assistant and are given clear instructions, warnings, and possible solutions when danger happens. In addition, it is essential to implement safety features and safeguards that can help prevent accidents and minimize the consequences and risks of any incidents that could happen.

It's our responsibility, as engineers, to address unforeseen risks to ensure the safety of users. First, the movement of the fingers is controlled by pulling on strings. However, if the strings are pulled too hard or in the wrong direction by mistake, it may cause the hand to jerk or twist unexpectedly, which could potentially cause harm to the user or others nearby. Moreover, if the strings get tangled or wrapped around the user's neck or other body parts, they could cause discomfort or even choking. Meanwhile, our product includes sharp or protruding parts that may accidentally puncture or scratch the user's skin, which could lead to bleeding, infection, or other medical injuries. It's essential to identify which parts of the intelligent assistant pose a risk and mitigate them, such as adding protective covers, smoothing edges, or even a whole redesign the pinpoints. Additionally, if the system malfunctions or short circuits, it could cause a fire or explosion, further endangering the user and the surrounding area.

The soul of design is to help people to life easier and work better. As a society, striving for respect, inclusivity, fairness, and equilibrium is vital, ensuring everyone has access to the tools and resources they need to live fulfilling lives without "discrimination based on characteristics such as race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression"[6]. Our core is to help people with hearing and speech problems could also interact with intelligent assistants.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection." (Jun. 2016).

[2] C. Lugaresi, J. Tang, H. Nash, *et al.* "Mediapipe: A framework for perceiving and processing reality." (2019), [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

[3] A. Halder and A. Tayade, "Real-time vernacular sign language recognition using mediapipe and machine learning," *Journal homepage: www. ijrpr. com ISSN*, vol. 2582, p. 7421, 2021.

[4] G. H. Samaan, A. R. Wadie, A. K. Attia, *et al.*, "Mediapipe's landmarks with rnn for dynamic sign language recognition," *Electronics*, vol. 11, no. 19, p. 3228, 2022.

[5] N. Developer. "Jetson Nano Developer Kit." (2021), [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit (visited on 03/08/2023).

[6] IEEE. "IEEE Code of Ethics." (2020), [Online]. Available: https://www.ieee.org/about/corporate/governance/p7-8.html (visited on 03/08/2023).