UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Plataforma de Aprendizaje de Lengua de Señas y Accesibilidad de Textos para Personas con Discapacidad Auditiva en Guatemala

Trabajo de graduación presentado por Evelyn Andrea Amaya Malin para optar al grado académico de Licenciada en Ingeniería en Ciencias de la Computación y Tecnología de la Información

Guatemala,

2023

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Plataforma de Aprendizaje de Lengua de Señas y Accesibilidad de Textos para Personas con Discapacidad Auditiva en Guatemala

Trabajo de graduación presentado por Evelyn Andrea Amaya Malin para optar al grado académico de Licenciada en Ingeniería en Ciencias de la Computación y Tecnología de la Información

Guatemala,

2023

Vo.Bo.:

(f) _____

M.Ed. Carmen Lucía del Pilar Guerrero Abascal de Prado

Tribunal Examinador:

(f) _____

M.Ed. Carmen Lucía del Pilar Guerrero Abascal de Prado

(f) _____

MBIA Ing. Carlos Jorge Valdéz Bautista

(f) _____

MSc. Douglas Leonel Barrios Gonzalez

Fecha de aprobación: Guatemala, 04 de diciembre de 2023.

Prefacio

La elaboración del presente trabajo de graduación surgió del interés personal de disminuir la barrera de comunicación entre personas oyentes y personas sordas. En Guatemala, la accesibilidad para las personas con cualquier tipo de discapacidad es baja, sino es que nula. El principal reto en este trabajo consiste en crear un modelo de predicción de lengua de señas a texto que sea posteriormente integrado en una plataforma de acceso libre y gratuito.

El objetivo es crear un modelo sustentado por la comunidad, el cual inicialmente cuenta con un conjunto de datos de 100 palabras, el cual se busca que crezca por sí solo haciendo uso de la aplicación Deaflens Studio.

Índice

Lista de figuras

Lista de cuadros

Resumen

Este trabajo se enfoca en un modelo que pueda ser integrado con una plataforma funcional y con interfaz amigable al usuario para que las personas sordas se puedan comunicar de forma eficaz con personas oyentes. El trabajo abarca tres temas principales: discapacidad auditiva, aprendizaje del lenguaje en personas sordas y el modelo propuesto para predecir lengua de señas a texto.

En el aspecto de la discapacidad auditiva, se examinan los tipos de sordera y cómo afecta la misma en aspectos pisocológicos desde niños. El aprendizaje del lenguaje en personas sordas se centra en el aprendizaje viso gestual para aprender a darle sentido a las palabras, además de la ayuda de la lectura fácil para facilitar el aprendizaje y se habla de un segundo método que es el deletreo manual. La dactilología es una parte importante en la comunicación de la comunidad sorda al hacer uso del deletreo en lengua de señas. Por último, se describen los modelos propuestos para la predicción de lengua de señas a texto. El cual se construyó haciendo uso de redes neuronales convolucionales, árboles de decisión y máquinas de vectores de soporte. El modelo recibe un video como entrada y devuelve como máximo tres palabras con mayor probabilidad de ser la correcta. En el caso del deletreo, se recibe la(s) imagen(es) en secuencia y se devuelve la palabra o letra predicha.

En conclusión, este trabajo ofrece un flujo completo para mejorar la comunicación entre personas sordas y oyentes tras proveer un modelo que será integrado en Deaflens Studio, una aplicación disponible para Android la cual permitirá visualizar las palabras con su intérprete, subir nuevas palabras para ayudar a la comunidad y un sistema eficiente para predecir lengua de señas a texto.

Abstract

Según la OMS, se estima que 430 millones de personas presentan pérdida auditiva de gravedad moderada y a su vez esto impacta en su capacidad para comunicarse y participar activamente dentro de la sociedad. La tecnología sigue avanzando todos los días y se han conseguido avances con el reconocimiento de lengua de señas, pero aún no se presenta una solución final debido a la complejidad de todos los factores que influyen: variaciones de lengua de señas, predicción de palabras no entrenadas con anterioridad, patrones complejos. La mayoría de trabajos existentes buscan el reconocimiento de lengua de señas a través del procesamiento de imágenes estáticas como lo es el abecedario, limitando las posibilidades de predicción del modelo. Este trabajo propone reconocer lengua de señas teniendo como entrada un video y procesando las coordenadas de ambas manos. El resultado es un modelo de predicción con dos conjuntos de datos disponibles: ASL (American Sign Language) y LENSEGUA (Lengua de Señas Guatemalteco).

Palabras clave: Lengua de señas, Movimiento de manos, Redes neuronales convolucionales, Árboles de decisión, Máquinas de vectores de soporte

CAPÍTULO 1

Introducción

Las Naciones Unidas estima que existen aproximadamente 70 millones de personas sordas en todo el mundo, de las cuales un 80 % de ellas viven en países en desarrollo[1]. En Guatemala, alrededor de 250 mil personas viven con algún grado de discapacidad auditiva[2] además de ser un país en desarrollo que no provee las herramientas necesarias para que una persona con sordera interactúe sin ningún problema con el resto de la comunidad oyente. El ejemplo más claro de esto es que hasta en 2020, el Congreso de Guatemala aprobó el Decreto 3-2020, el cual reconoce LENSEGUA como la lengua de señas oficial en Guatemala y hasta 2023 este sigue sin tener un reglamento oficial.

La tecnología juega un papel importante para superar la barrera de accesibilidad que tienen las personas con discapacidad auditiva. La solución planteada consiste en crear tres modelos que trabajen en conjunto y que sean fácilmente integrados con cualquier plataforma para traducir lengua de señas (video o secuencia de imágenes) a texto.

La importancia de crear un modelo que reciba como entrada un video en lengua de señas se debe a que las personas con sordera encuentran dificultades en el aprendizaje de la lectoescritura debido a su déficit auditivo y lingüístico. Comienzan a aprender lengua escrita a través de leer imágenes o por medio de un intérprete y así asociar palabra-significado. Y, las palabras aisladas del contexto no tendrán significado alguno si no se acompaña con experiencias concretas como ver, oler, probar o sentir.

Por otra parte, los modelos de predicción proveen una comunicación más sencilla con alguien que haga uso de la lengua de señas como su forma de comunicación principal. Donde además de dar la opción de grabar video, se puede hacer uso del deletreo en dado caso el modelo no haya sido entrenado previamente con la palabra buscada o bien, subir la palabra en la aplicación del estudio para futuras búsquedas. Se busca disminuir la brecha de accesibilidad para las personas con discapacidad auditiva y, a su vez, proveer una herramienta de apoyo de aprendizaje de lengua de señas para las personas oyentes.

CAPÍTULO 2

Antecedentes

Actualmente se están haciendo varios trabajos de investigación con relación a proveer un sistema de comunicación para las personas sordas. Bioingenieros de UCLA han diseñado un guante para traducir Lengua de Señas Americano a inglés en tiempo real haciendo uso de una aplicación móvil[3]. Así mismo, la compañía BrightSign cuenta con un guante para traducir más de 300 lenguas de señas, la persona se lo coloca y cuenta con un dispositivo de salida de ayuda para que el oyente entienda la seña[4]. Algunos trabajos con realidad aumentada están siendo desarrollados como Content4All, la cual busca automatizar un intérprete virtual para mostrar lengua oral a lengua de señas y viceversa[5]. Existen otro tipo de productos como SignAll el cual cuenta con dos pantallas y tres cámaras para reconocer los movimientos del cuerpo, expresiones faciales y manos, además de una cámara que reconoce color, donde se espera que el usuario utilice guantes con color para que sean detectado[6]. También existen diferentes aplicaciones móviles y de escritorio para transcribir voz a texto. Por último, existen varios modelos de machine learning para reconocer el abecedario estático en lengua de señas[7].

CAPÍTULO 3

Justificación

El artículo 26 de los Derechos Humanos [8] establece que toda persona tiene derecho a la educación. Es decir, la alfabetización es un derecho humano y según la UNESCO [9] se define alfabetización como "la capacidad que tiene una persona para leer y escribir". Además, la Convención Internacional sobre los Derechos de las Personas con Discapacidad dicta en el artículo 9 que las personas con cualquier discapacidad tienen derecho a los mismos accesos que el resto de la sociedad, incluyendo el entorno físico, el transporte, la información y las comunicaciones, y otras instalaciones y servicios públicos [10].

¿Por qué la plataforma propuesta para proveer mejor accesibilidad a las personas con sordera es viable y aportará solución al problema?

1. La tecnología es una gran herramienta para el mejoramiento de la educación y la inclusión social para las personas con discapacidad auditiva.
2. Se proveerá un modelo de predicción de lengua de señas a texto el cual será de acceso gratuito a todas las personas a través de una aplicación: Deaflens Studio.
3. Se proveerá una herramienta gratuita la cual busca disminuir la barrera de comunicación entre personas oyentes y personas sordas.

CAPÍTULO 4

Objetivos

4.1. Objetivo general

- Desarrollar un modelo de predicción de lengua de señas para el apoyo y aprendizaje para personas sordas.

4.2. Objetivos específicos

- Desarrollar un modelo con precisión de al menos 0.75 el cual se pueda utilizar dentro de una plataforma intuitiva para facilitar la comunicación entre personas oyentes y sordas.
- Crear un modelo con un conjunto de datos amplio y actualizado de manera periódica para proporcionar variedad de palabras y expresiones.
- Desarrollar un sistema preciso para reconocer el abecedario en lengua de señas y predecirlo a texto.

CAPÍTULO 5

Alcance

El alcance del proyecto es la predicción de 100 palabras en total. Se tienen 75 palabras en lengua de señas guatemalteco (LENSEGUA) el cual está compuesto por 3 conjuntos de palabras: 25 palabras de uso frecuente, 25 palabras relacionadas a supermercados y 25 palabras relacionados a médicos/farmacias. De esta misma lengua se incluye el abecedario para predicción a través del deletreo. Por último, se incluyó un listado 25 palabras de uso frecuente en lengua de señas americano (ASL).

Algunas limitaciones del proyecto son:

- No se puede obtener la predicción correcta de una palabra que no haya sido entrenada con anterioridad.
- La predicción correcta de la palabra es directamente afectada por la cantidad de puntos detectados por el modelo Mediapipe Hand Landmarks de Google.
- Actualmente, no es posible para el modelo predecir la lengua de señas utilizada en el video, por lo que son dos modelos distintos LENSEGUA y ASL.

CAPÍTULO 6

Marco teórico

6.1. Discapacidad auditiva

La Organización Mundial de la Salud (OMS) define sordo como toda persona cuya agudeza auditiva le impide aprender su propia lengua, aprovechar de enseñanzas básicas y participar en las actividades normales de su edad.

6.1.1. Clasificación de la sordera

Desde el punto de vista educativo, existen dos clasificaciones[11]:

- Niños hipoacúsicos: Tienen dificultades con la estructuración del lenguaje, pero su vía auditiva no les impide adquirir el lenguaje oral. Necesitan la ayuda de prótesis auditivas.
- Sordos profundos: Se les dificulta la adquisición del lenguaje oral a través de la vía auditiva incluso haciendo uso de prótesis o amplificadores.

6.1.2. Psicología del niño sordo

Las repercusiones psicológicas serán distintas dependiendo de la edad a la que al niño presente indicios de sordera. Solo los niños que adquieran la sordera luego de los 4 o 5 años siguen utilizando normalmente el lenguaje debido a que su cerebro fue asociando los sonidos lingüísticos.

Es inevitable que la sordera aísle a la persona, ya que la falta de audición inhibe el desarrollo emocional, por lo que frecuentemente incrementa el sentimiento de soledad y aumenta el deseo de comunicarse socialmente[11]. Un ejemplo, es que interrumpan las conversaciones para que se les indique de qué se habla.

6.1.3. Aspectos del perfil psicológico del niño sordo

Además Sabina Pabón también menciona 7 aspectos psicológicos del niño sordo [11], de los cuales se seleccionaron los 4 con mayor impacto al trabajo actual:

- Problemas de atención: Una persona sorda interrumpe sus actividades con frecuencia para controlar en forma visual el ambiente y está pendiente de cualquier estímulo a su alrededor.
- Acentuada afectividad: Una persona sorda no puede inferir emociones a través del tono de las expresiones. Por lo que se les es difícil inferir condiciones de proximidad e identificar sentimientos.
- Mayor dependencia: El interlocutor deberá hablar más lento, repetir y vocalizar bien, lo cual dependerá de la voluntad y paciencia de él. A veces en necesario la intervención de un tercero.
- Sentimiento de inferioridad y frustración: Debido a sus limitaciones, puede llegar a demostrar conductas de irritabilidad, alejamiento, agresividad e inferioridad.

6.2. Aprendizaje del lenguaje en niños sordos

La lectura y comprensión son inseparables, un niño oyente cuando comienza a leer y visualiza las letras, les atribuye un significado a través del sentido auditivo, construyendo mentalmente una imagen acústica.

En la década de los 70 [12], Dodd, demostró que es posible que un niño sordo reconozca palabras (conciencia fonológica), ya que no sólo proviene de origen acústico, sino también de la visual y kinésico. Es por esto, que la lectura labiofacial facilita la compresión del habla, pero resulta insuficiente para la construcción de representaciones fonológicas precisas y exactas, sobre todo para fonemas similares.

Según Ester Ruiz [13], la palabra complementada es un sistema de apoyo para la lectura labiofacial que suprime las ambigüedades del habla al haber ausencia acústica, permitiendo que se visualicen los fonemas, y así permitir que el niño sordo reconozca palabras.

Aún así, este sistema no es aplicado siempre y se emplea el método de la memorización de palabras, lo cual es un trabajo difícil y laborioso, ya que el niño carece del sentido de la palabra. Debido a que no tiene sentido la mayor parte de las palabras memorizadas, el niño no conseguirá ser un lector eficaz y autónomo; podrá leer textos y necesitará un gran esfuerzo para darle contexto a cada palabra, y al leer una palabra desconocida, no podrá atribuirle ningún significado.

El aprendizaje de la lectoescritura es un proceso largo y complejo, que requiere de mucho trabajo y esfuerzo por parte de quien quiere desarrollarlo.

6.3. Lectura fácil

La lectura fácil es un método de redacción que permite a las personas con dificultades lectoras comprender textos [14]. El texto escrito es el medio con mayor potencial para que las personas sordas accedan al conocimiento del mundo. Aún así, el 80 % de la población sorda es analfabeta, por lo que es necesario intervenir y brindarle acceso a la lectoescritura tanto desde el ámbito escolar como familiar.

Los textos en lectura fácil deben de variar dependiendo de lo que se quiere transmitir a las personas con discapacidad auditiva. Si la finalidad del texto es transmitir conocimiento del entorno y no potenciar la adquisición de habilidades lingüísticas, siguiendo la guía "*Accesibilidad auditiva. Pautas básicas para aplicar en los entornos*", de Antonio Espínola [15], se recomienda:

- Evitar oraciones subordinadas siempre que sea posible.
- Evitar el estilo poético.
- Aunque resulte repetitivo, el sujeto debe de estar en forma nominal. Los prenombres personales ("lo(s)", "la(s)", "le(s)", "me", etc.) son una barrera mayor para las personas sordas.
- Hacer uso de los tiempos verbales más comunes.
- Hacer uso de sinónimos sencillos.
- No hacer uso de expresiones en doble sentido, metáforas, sarcasmo, etc.
- Hacer uso de elementos visuales como apoyo para conceptos complejos.
- De ser posible, hacer uso de un formato cómic como medio de expresión.

6.4. Dactilología

El comunicarse con una persona sorda deletreando una palabra en lengua de señas hace referencia a la dactilología. El cual se basa en el alfabeto latino y cada letra del alfabeto es representada manualmente por un movimiento de la mano único y discreto. Es decir, se entrega información lingüística de carácter secuencial [16].

La dactilología es parte del sistema de comunicación de la comunidad sorda. Es un puente altamente utilizado entre la lengua oral y la lengua de señas, ya que contiene información viso-gestual.

En 1987 Hirsh-Pasek [17] descubrió que la dactilología puede ser usada como una estrategia para identificar palabras, pero es una herramienta limitada para aprender sintaxis. Por lo que propuso la idea de bilexicalismo, la cual se refiere a aprender conceptos a partir de la seña de la lengua de señas y la dactilografía de la palabra. Y así, permitir a las personas sordas complementar el vocabulario y aumentar la habilidad para identificar palabras.

6.5. Sistema propuesto

En la lengua de señas las manos se utilizan para el vocabulario, mientras la cara y el cuerpo son utilizados para expresar y dar énfasis en las palabras y frases. Por lo que se extraen 84 coordenadas en total por ambas manos haciendo uso del modelo Hand Landmarker [18] de Google para predecir lengua de señas a palabra. Y, se hará uso de la dactilología para construir palabras que no están dentro del vocabulario del modelo. Es decir, se entrenará el alfabeto de la lengua de señas guatemalteca, para que se pueda hacer uso del deletreo en señas y predecir la palabra completa.

6.5.1. Método propuesto

Se hará uso de tres algoritmos de aprendizaje supervisado para el procesamiento de videos: Red neuronal convolucional, Árbol de decisión y Máquina de vectores de soporte. Y, únicamente del algoritmo Árbol de decisión para el procesamiento de imágenes. Estos algoritmos nos permitirán categorizar videos e imágenes al ser algoritmos que aprenden iterativamente de los datos de problemas conocidos, es decir, datos etiquetados. Al necesitar datos correctamente etiquetados para identificar futuras palabras/letras, se ignora en el preprocesamiento los videos e imágenes que no tengan detectadas las manos necesarias. Un ejemplo de esto es la letra "ñ" que hace uso de ambas manos, la cual podría confundir al algoritmo con una "n" si solo hay una mano detectada, por lo cual se descarta del conjunto de datos de entrenamiento.

Cada video será de un máximo de 5 segundos, el cual será preprocesado para tener 30fps y se ralentizará o acelerará para tener 1 segundo de longitud. Es decir, cada video al ser procesado por el modelo Hand Landmarker tendrá una salida de las 84 coordenadas con 30 filas, donde cada fila representa un frame del video. Si dentro del frame no se detectan la(s) mano(s) el modelo se encarga de rellenar a cero la fila y así todas las salidas sean un vector con tamaño (1, 30, 84).

A diferencia de la Red neuronal convolucional la entrada del Árbol de decisión y de la Máquina de vectores de soporte es únicamente de dos dimensiones, por lo que se realiza un aplanamiento de los conjuntos de datos donde cada fila representa un video/palabra. En el caso de un video, la entrada pasa a ser de tamaño (1, 2520) y en el caso de una imagen, la entrada es de tamaño (1, 84). Es decir, las 30 filas correspondientes de un video se esparcen de forma horizontal donde se tendrán las 84 coordenadas de cada fila representadas por columnas, mientras que, en una imagen solo se tiene 1 frame, es decir, una única fila. Por último, para abordar el desequilibrio de clases en el conjunto de datos se hizo uso de la librería *class_ weights* de Scikit-Learn para asignarle distintos pesos a la palabra/letra y así prestarle atención a las clases minoritarias. Es decir, cada palabra y cada letra del abecedario tiene la misma importancia dentro del modelo.

6.5.2. Red neuronal convolucional

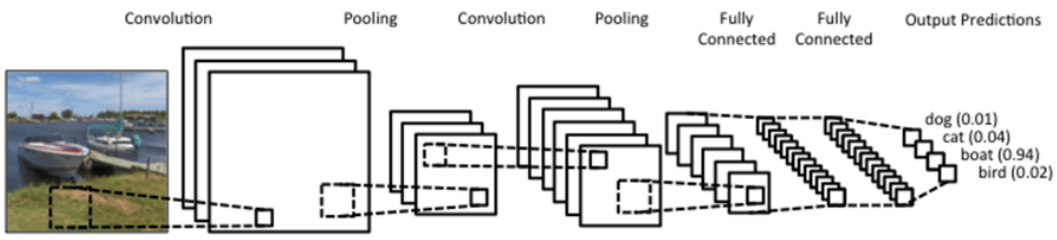

Figura 1: La red convolucional toma una imagen como entrada y produce una salida numérica que representa el contenido de la imagen.

Una red neuronal es un método de inteligencia artificial que tiene como objetivo simular el funcionamiento del cerebro humano. Relacionan los datos de entrada y salida que no son lineales y son complejos aprendiendo las relaciones que hay entre ellos. Se utilizan para reconocer patrones numéricos contenidos en vectores, por ejemplo, imágenes.

Una red convolucional es una clase de red neuronal que se especializa en procesar imágenes al extraer características relevantes que ayudan en su posterior clasificación. Cada capa oculta extrae y procesa diferentes características de la imagen, como los bordes, el color y la profundidad. La capa *Pooling* sirve para reducir las matrices generadas por la capa convolucional. Al final, se conectan todas las capas y la predicción será una lista de probabilidades por cada posible resultado. Es decir, si tenemos 25 palabras dentro de nuestro modelo, el resultado será una lista de 25 probabilidades, de la cual se toma la probabilidad mayor como nuestra predicción resultante.

Arquitectura del modelo

Una arquitectura simple para redes convoluciones cuenta con tres capas interconectadas las cuales son: capa de entrada, capa oculta y capa de salida. La red convolucional considerada para mi estudio Figura 2 está compuesta por tres redes convolucionales. La capa *Flatten* para transformar los datos multidimensionales en un vector unidimensional, la capa *Dense* para conectar todas las neuronas y que aprendan patrones y, la capa *Dropout* para evitar el sobre ajuste y mejorar la generalización del modelo. Además, se hace uso de la función de activación ReLU para remover números negativos y colocarlos en 0 y Softmax para la capa de salida debido a que se tiene un problema de clasificación multiclase por lo que se obtiene un vector con la probabilidad normalizada por cada posible output.

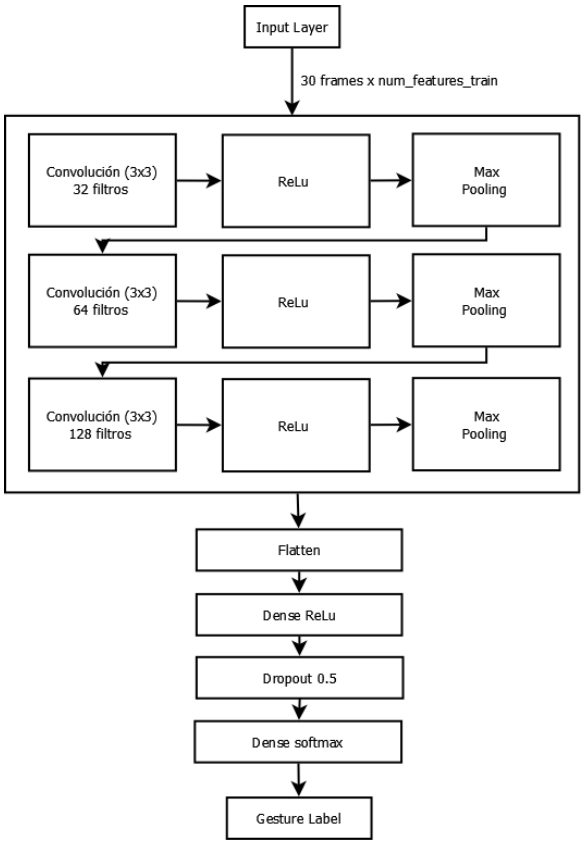

Figura 2: Arquitectura del modelo

6.5.3. Árboles de decisión

Un Árbol de decisión es un algoritmo de Machine Learning que se utiliza para tareas de clasificación o regresión. Consta de nodo raíz, ramas, nodos de decisión y nodos hoja o terminal. Se emplea una estrategia de divide y vencerás mediante la búsqueda de puntos óptimos dentro del árbol. Es un proceso recursivo de arriba hacia abajo hasta que todos o la mayoría de los registros se hayan clasificado.

Arquitectura del modelo

Se construyó un Árbol de decisión con profundidad máxima cantidad de palabras – 1, para ajustar la cantidad de divisiones dentro del árbol según la complejidad (cantidad de palabras) del modelo. Entre los hiperparámetros del árbol se definió un valor pequeño para la cantidad mínima de muestras requeridas para dividir un nodo interno del árbol en dos nodos hijos y la cantidad mínima de muestras requeridas en un nodo hoja. Esto nos permite que el modelo capture detalles de los datos de entrenamiento al ser un modelo complejo que depende estrictamente de las coordenadas de las manos.

Cabe mencionar que los hiperparémtros son escogidos por la técnica de validación cruzada *GridSearchCV* para ejecutar diferentes parámetros y extraer la mejor combinación.

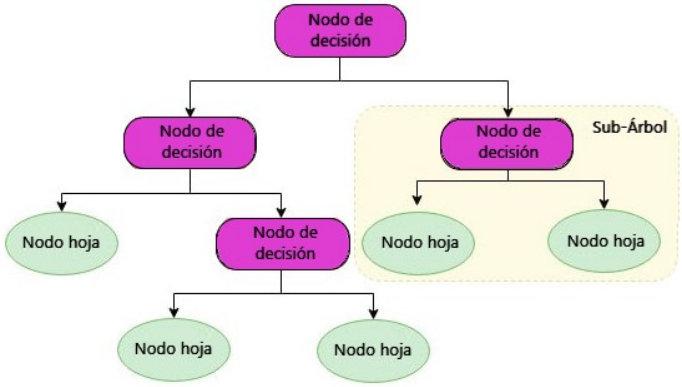

Figura 3: Árbol de decisión con sus terminologías

Además, implementa *AdaBoost* el cual es un algoritmo de aprendizaje automático para mejorar el rendimiento del Árbol de decisión y así dar más peso a las muestras que se clasificaron incorrectamente en las iteraciones anteriores. Se agregó la implementación de Scikit-Learn para mejorar la precisión del modelo al ser un conjunto de entrada compuesto únicamente de números flotantes.

6.5.4. Máquinas de vectores de soporte

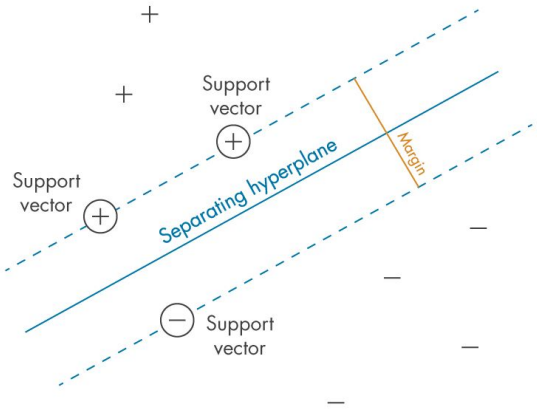

Figura 4: Hiperplano para separar dos clases con un margen establecido

El objetivo del algoritmo Máquinas de vectores de soporte (Support Vector Machine) es encontrar un hiperplano que separe de la mejor forma posible dos o más clases de puntos de datos. Este algoritmo pertenece a los métodos kernel, donde se puede hacer uso de una función de kernel para transformar las características y así asignar a un espacio dimensional diferente cada clase. Este algoritmo es útil para resolver problemas de clasificación de patrones difíciles que, en nuestro caso, serán las 2520 coordenadas por cada entrada.

Arquitectura del modelo

Se hace uso de un clasificador con un kernel lineal y un margen de 0.001 entre clases al contar con un problema complejo. Dependiendo de las características de los datos de entrenamiento se podrá separar entre grupos y así devolver la palabra predicha por el algoritmo.

6.6. Inteligencia artificial y ética

La inteligencia artificial es la capacidad de una máquina para utilizar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano [19]. En este proyecto, se ve claramente reflejado en la toma del conjunto de coordenadas de las manos de los usuarios para su posterior predicción de la palabra o su uso para el entrenamiento de los tres modelos utilizados.

La responsabilidad al hacer uso de los datos y desarrollar una inteligencia artificial es esencial. AI4People propone que se puede aumentar las capacidades de la sociedad sin reducir el control humano [20]. Es decir, la inteligencia artificial ofrece posibilidades para aumentar la inteligencia de los humanos además de proveer soluciones para problemas nuevos y viejos pero siempre las decisiones de estos sistemas autónomos deben de permanecer bajo la supervisión de humanos.

Según la Federación Mundial de Sordos, en el mundo se utilizan más de 300 lenguas de signos. Por lo que, si en el futuro se logra predecir de video a texto cada una de sus variantes será gracias a la inteligencia artificial. Como lo es en este caso, la predicción de un video en lengua de señas a texto no sería posible sin el uso de algoritmos de predicción y el modelo Hand Landmarker de MediaPipe Google el cual es es un modelo de aprendizaje automático que puede detectar y rastrear 21 puntos clave en las manos.

Es un problema realmente complejo extraer 84 coordenadas entre ambas manos y los ejes x y y para dar como resultado una palabra. Lo cual es posible porque la solución está basada en algoritmos que hacen uso intensivo de una gran cantidad de datos para proveer la solución correcta al problema.

El enfoque ético de la inteligencia artificial se basa en garantizar resultados a problemas de la sociedad y a su vez mitigar los daños potenciales al hacer un mal uso de esta tecnología. Se debe de siempre cumplir con la ley, además de proveer confianza pública y gestionar los riesgos posibles. Aún así, la ética es un arma de doble filo, un claro ejemplo de esto es proveer una solución necesaria y socialmente aceptable que a su vez cuenta con riesgos potenciales pero prevenibles o minimizables que no están protegidos bajo la ley [20].

Es importante tener en cuenta que la solución planteada en este proyecto es una herramienta para predecir lengua de señas a texto. Para el conjunto de datos de entrenamiento y validación se obtuvo el consentimiento de las personas involucradas además de explicar de forma transparente el funcionamiento del modelo o se hizo uso de videos de uso público. Y, la forma en la que se utilice este modelo depende de los desarrolladores que creen aplicaciones con él. Es importante que los desarrolladores sean conscientes de las implicaciones éticas de su trabajo y que tomen medidas para mitigar los riesgos potenciales.

CAPÍTULO 7

Metodología

7.1. Recolección de datos

7.1.1. Generación de nuevas entradas

Se hizo uso del modelo de Hand Landmarker de Google para detectar 21 coordenadas de la mano izquierda y 21 coordenadas de la mano derecha, en ambos ejes x y y. Se creó un programa para extraer las 84 coordenadas de un video o lista de videos seleccionados. Siendo así, que la predicción de palabras se hace por medio de 84 coordenadas con n cantidad de filas correspondiente a cada frame del video y en caso de una imagen solo se cuenta con un único frame.

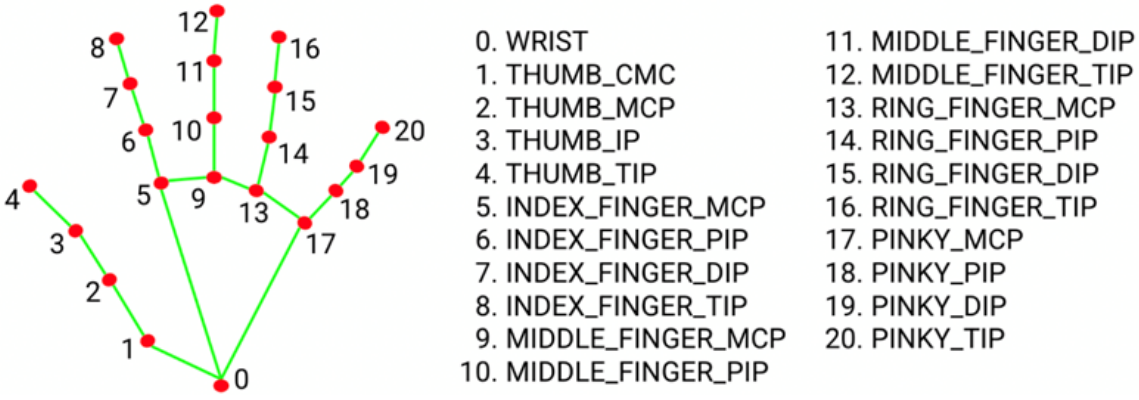

Figura 5: Coordenadas detectadas en la palma de la mano por Hand Landmarker
Fuente: Google (2023)

Cada video para entrenamiento o predicción se pasará por el mismo preprocesamiento, el cual cuenta de los siguientes pasos:

- Se cambia la duración del video a un segundo. Se acelera el video en dado caso sea mayor que un segundo o realentiza en caso de durar menos tiempo.
- Se revisa los fotogramas por segundo (fps) para solo aceptar una velocidad de 30 fps. Si un video contiene más fotográmas, estos se recortan a 30 fps.
- En caso de un video, se toma una captura por cada frame para procesarlo como un set de imágenes y que así no afecte la rapidez del video.
- Se normalizan las coordenadas de cada frame entre 0 y 1 para que las coordenadas sean similares cuando la mano esté muy cerca o muy alejada de la cámara. Esto ayuda con la predicción del modelo al tener la mano sobre el mismo eje además de centrarla.

7.1.2. ASL: Lengua de Señas Americano

Para iniciar la base de datos, se descargó el video *25 ASL Signs You Need to Know | ASL Basics | American Sign Language for Beginners* del canal *Learn How to Sign* en YouTube. Este video se recortó por cada vez que la enseñante repetía la seña de alguna de las 25 palabras, dando un total de 108 videos con una longitud entre 0.5 a 2 segundos.

Se extrajeron las coordenadas de los 108 videos haciendo uso del Hand Landmarker separando por cada palabra un video para su posterior predicción. Teniendo así, 4-6 videos en el conjunto de prueba para validar la precisión del modelo por palabra. Mientras que para el conjunto de entrenamiento procedí a grabarme repitiendo las 25 palabras de 25-40 veces. Y, al ser mutuamente excluyentes los conjuntos de validación y entrenamiento se consigue que el modelo no produzca sobre ajuste.

Figura 6: Detección de las coordenadas para la palabra: no

7.1.3. LENSEGUA: Lengua de señas guatemalteco

Para iniciar la base de datos de texto-intérprete se les pidió ayuda a estudiantes de la Universidad del Valle de Guatemala con conocimiento en lengua de señas. Para que interpreten cerca de 75 palabras y usar esto como base de datos de entrenamiento para el modelo. Al igual que con el conjunto de lengua de señas americano, se dejó mutuamente excluyente el conjunto de datos de validación y entrenamiento para evitar el sobre ajuste del modelo.

Al no contar con la LENSEGUA oficializada, se contó con la ayuda de docentes de la Universidad del Valle de Guatemala que están trabajando directamente en oficializar la lengua de señas por lo que se tiene un acercamiento más preciso a la lengua de señas de las palabras.

Se trabajará por 3 fases, cada fase con 25 palabras. La primera fase son palabras comunes como saludos, verbos fáciles, entre otros. La segunda fase son palabras que pueden ser útiles dentro de un supermercado. Y, la tercera fase son palabras que pueden ser útiles dentro de una farmacia o consulta médica.

Figura 7: Detección de las coordenadas para la palabra: feliz

CAPÍTULO 8

Resultados

El método propuesto fue probado con una base de datos inicial con 25 palabras en lengua de señas guatemalteco LENSEGUA. Por cada palabra, se cuenta con más de 20 vídeos para entrenamiento además de 2 vídeos de prueba y un video extra para validación del modelo exportado. Ninguna de las personas utilizadas para validar la precisión del modelo son parte de los datos de entrenamiento para evitar sesgos y validar que el modelo realmente puede predecir palabras haciendo uso de personas terceras.

En el Cuadro 1 se puede observar que el mejor resultado individual fue con el modelo Árbol de decisión con una precisión de 0.81, pero al hacer uso de los tres modelos en conjunto la precisión aumenta hacia un 0.88. Cabe mencionar que, debido a que es un conjunto de datos pequeño, el pipeline de los tres modelos falla únicamente 5 palabras (Tabla 2), donde solo la palabra *escuela* no es predicha en ninguno de los dos videos de prueba. La idea detrás de hacer uso de los tres modelos se debe a la complejidad del modelo en el cual al usuario se le mostrarán como máximo 3 palabras distintas predichas, donde él es el encargado de elegir la palabra correcta.

Modelo	Precisión
CNN	0.61
Árbol de decisión	0.81
SVC	0.67
Pipeline tres modelos	0.88

Cuadro 1: Precisión de los modelos contra el conjunto de datos de prueba

CNN	Árbol de decisión	SVC	Valor real
Escuela	Tiempo	Médico	Familia
Hola	Hola	Hola	Adios
Casa	Casa	Adios	Hola
Ayuda	Salud	Ayuda	Escuela
Ayuda	Salud	Ayuda	Escuela

Cuadro 2: Resultados contra el conjunto de datos de prueba haciendo uso de 25 palabras

La precisión de los tres modelos se ve directamente afectada por la proporción de en cuantos frames donde sí aparezca(n) la(s) mano(s) sean detectados correctamente por el modelo Hand Landmarker de Google. Un ejemplo de esto, es la palabra *hola* la cual en los primeros segundos la mano se encuentra posicionada sobre la frente del usuario, y luego de eso se mueve el codo en un ángulo aproximado de 30° hacia afuera, mientras la palabra *gracias* comienza exactamente igual pero se mueve el codo en un ángulo aproximado de 90° hacia afuera. Dicho esto, la mano está estática en la misma posición los primeros segundos en ambas palabras, pero al hacer el movimiento rápido hacia afuera el modelo de Google suele perder la mano y no detectar las coordenadas en esos frames, afectado directamente la precisión del modelo.

Para comprobar esto, se grabaron 7 palabras extra por una persona distinta a las del grupo de entrenamiento y prueba. Los resultados obtenidos se observan en el Cuadro 3, con los cuales la precisión del pipeline de los tres modelos fue de un 100 %.

CNN	Árbol de decisión	SVC	Valor real
Amor	Feliz	Feliz	Amor
Familia	Escuela	Familia	Familia
Casa	Casa	Casa	Casa
Gracias	Hola	Por favor	Hola
Triste	Bebida	Bebida	Bebida
Tiempo	Tiempo	Tiempo	Tiempo
Amigo	Amigo	Amigo	Amigo

Cuadro 3: Resultados contra el conjunto de datos de validación

Se usó la misma metodología y los tres modelos descritos para aumentar el tamaño del conjunto de datos inicial a 50 palabras en lengua de señas guatemalteco LENSEGUA. Las palabras agregadas corresponden a la temática supermecado. En el Cuadro 4 se puede observar que la precisión de los tres modelos se redujo, pero al hacer uso del pipeline se consiguió una precisión de 0.81. Aún así, el modelo Árbol de decisión sigue siendo el mejor modelo por sí solo, pero otro punto interesante es que al hacer uso de 50 palabras los tres modelos se ayudan más entre sí. Haciendo uso de 25 palabras la precisión del Árbol de decisión contra el pipeline tenía una diferencia de 0.07, mientras que haciendo uso de las 50 palabras, la precisión aumenta 0.11 haciendo uso del pipeline. Por lo que se le da validez a hacer uso de los tres modelos como una primera iteración del proyecto.

Modelo	Precisión
CNN	0.55
Árbol de decisión	0.70
SVC	0.57
Pipeline tres modelos	0.81

Cuadro 4: Precisión de los modelos contra el conjunto de datos de prueba haciendo uso de 50 palabras

En el Cuadro 5 se pueden observar las palabras incorrectas no predichas por ninguno de los tres modelos. Como mencionado anteriormente, por cada palabra hay 2 videos de validación dentro del conjunto de datos. De las cuales las siguientes palabras no fueron predichas correctamente ninguno de los dos videos: Escuela, Azúcar, Mantequilla, es decir, se fallaron un 100 % únicamente 3 palabras del conjunto de datos.

CNN	Árbol de decisión	SVC	Valor real
Médico	Médico	Queso	Familia
Hola	Hola	Enfado	Adiós
Adiós	Casa	Casa	Hola
Pañuelos	Shampoo	Salud	Escuela
Feliz	Ayuda	Triste	Escuela
Frutas	Sopa	Sopa	Azúcar
Sal	Tiempo	Sal	Aceite
Casa	Sopa	Escuela	Mantequilla
Enfado	Querer	Querer	Jabón
Casa	Casa	Escuela	Shampoo
Por favor	Shampoo	Escuela	Mantequilla
Dinero	Dinero	Dinero	Galletas
Sopa	Sopa	Sopa	Azúcar
Feliz	Feliz	Feliz	Frutas

Cuadro 5: Resultados contra el conjunto de datos de validación

Para la siguiente iteración, se aumentó el tamaño del conjunto de datos inicial a 75 palabras en lengua de señas guatemalteco LENSEGUA. Las palabras agregadas corresponden a la temática farmacia o médico. En el Cuadro 6 se puede observar que la precisión de los tres modelos se redujo, pero al hacer uso del pipeline se consiguió una precisión de 0.79. Y, al igual que en las iteraciones pasadas el Árbol de decisión sigue siendo el modelo por sí solo, pero ahora el pipeline ayuda a mejorar el modelo en una precisión de 0.15.

Modelo	Precisión
CNN	0.57
Árbol de decisión	0.64
SVC	0.47
Pipeline tres modelos	0.79

Cuadro 6: Precisión de los modelos contra el conjunto de datos de prueba haciendo uso de 75 palabras

De las palabras incorrectas obtenidas en esta iteración, cabe mencionar que ninguna de las palabras nuevas agregadas fue predicha incorrectamente un 100 %, recordando que solo se cuentan con dos videos en el conjunto de datos de validación. De las palabras incorrectas de esta iteración se encuentran: crema humectante, efectos secundarios, termómetro digital, jarabe, pomada, vendaje, dosificación, loción.

Para implementar la dactilología, se hizo uso del modelo Árbol de decisión para el método propuesto de deletreo. Se entrenaron 27 letras del abecedario en lengua de señas guatemalteco LENSEGUA incluyendo las letras F, J y S las cuales son con movimiento. Por cada letra se tienen alrededor de 25 imágenes para entrenamiento, además de tener dos imágenes para prueba y una palabra deletreada para validación. Al igual que con los videos, las personas de prueba y validación no son parte del conjunto de datos de entrenamiento para comprobar que el modelo funciona como se espera.

Se obtuvo una precisión de 0.93, donde las letras incorrectamente predichas una vez fueron la letra *e* contra la letra *o* y la letra *z* contra la letra *q*. Al igual que con los videos, la precisión de la predicción depende directamente de las coordenadas que detecte el modelo Hand Landmarker de Google. Y, se espera obtener una precisión de 100 % al aumentar el conjunto de datos de entrenamiento. Además, se comprobó el modelo tras darle como entrada cuatro imágenes deletreando *h, o, l, a* con las cuales se obtuvo una precisión de 100 % y como resultado la palabra *hola*.

Figura 8: Ráfaga de imágenes para la palabra hola

Haciendo uso de exactamente los mismos modelos y procedimiento se procedió a entrenar el conjunto de datos de lengua de señas americano (ASL), el cual como se menciona en la metodología cuenta de 25 palabras. En el Cuadro 7 se puede observar los resultados individuales y del pipeline al correr las palabras en ASL. Al igual que en los resultados anteriores, el mejor modelo por sí solo es el Árbol de decisión con una precisión de 0.73, pero al correr el pipeline se logra aumentar su precisión a un 0.84, por lo que es útil hacer uso de los tres modelos en conjunto.

Modelo	Precisión
CNN	0.67
Árbol de decisión	0.73
SVC	0.56
Pipeline tres modelos	0.84

Cuadro 7: Precisión de los modelos contra el conjunto de datos de prueba haciendo uso de 25 palabras en ASL

En el Cuadro 8 se pueden observar las palabras incorrectas no predichas por ninguno de los tres modelos. A diferencia de en LENSEGUA, en ASL los resultados cuando falla el modelo son idénticos. En ASL se tienen más videos en el conjunto de validación, pero la palabras con más fallos es *repeat* con 75 % videos predichos incorrectamente, seguida por *sign* y *what* con 50 % videos predichos incorrectamente y *more* y *see you later* con 40 % videos predichos incorrectamente.

CNN	Árbol de decisión	SVC	Valor real
dog	dog	yes	see you later
sign	learn	learn	sign
finish	learn	finish	more
want	sign	sign	what

Cuadro 8: Resultados contra el conjunto de datos de prueba en ASL

Al igual que con el conjunto de datos pasado, se realizó una segunda validación con una tercera persona del género opuesto para corroborar la predicción del modelo. Con el cual se grabaron 14 de las 25 palabras. El resultado fue de 0.72 donde se predijo incorrectamente 4 palabras de las 14 grabadas (Cuadro 9). Cabe mencionar que las palabras *sign* y *see you later* fueron grabadas en este conjunto de datos y sí fueron predichas correctamente.

CNN	Árbol de decisión	SVC	Valor real
go to	learn	go to	sign
fine	fine	fine	father
want	repeat	learn	help
dog	dog	sign	finish

Cuadro 9: Resultados contra el conjunto de datos de validación en ASL

CAPÍTULO 9

Conclusiones

El poder de la inteligencia artificial para reconocer la(s) mano(s) de los usuarios para su predicción a texto fue utilizada en este trabajo. El modelo propuesto provee una nueva forma de predecir lengua de señas al hacer uso de las coordenadas de las manos ignorando la postura del usuario.

Este trabajo demuestra que los algoritmos Red neuronal convolucional, Árbol de decisión y Máquinas de vectores de soporte pueden ser usados en conjunto para reconocer de forma precisa 75 palabras en lengua de señas LENSEGUA con una precisión de 0.79 y 25 palabras en lengua de señas ASL con una precisión de 0.83. Y, que el Árbol de decisión puede ser utilizado para predecir correctamente las imágenes del abecedario en lengua de señas LENSEGUA con una precisión de 0.93.

La precisión de este modelo fue evaluada con tres conjuntos de datos creados con ayuda del club de lengua de señas de la Universidad del Valle de Guatemala: LENSEGUA con 75 palabras, LESENGUA con 24 letras del abecedario y ASL con 25 palabras.

Este trabajo puede contribuir al campo de investigación sobre el reconocimiento automático de la lengua de señas al demostrar que es posible recibir un video y dar como resultado una palabra.

CAPÍTULO 10

Recomendaciones

Se recomienda que los videos utilizados para entrenamiento sean revisados por un humano para detectar que al menos un 80 % de los *frames* que sí incluyen la(s) mano(s) de los usuarios sean correctamente detectados y así aumentar la capacidad de predicción al tener un conjunto de datos completamente limpio antes de entrenar el modelo.

Además, es importante investigar sobre diferentes algoritmos que puedan mejorar el modelo y así seguir cumpliendo con una precisión por arriba de 0.75 sin importar la cantidad de palabras dentro del conjunto de datos. Así mismo validar si es posible la implementación de un solo modelo para diferentes tipos de lengua, es decir, que además de la palabra se prediga la variante de la lengua de señas.

Por último, se incentiva a compañeros y colegas interesados en el área de tecnología a seguir investigando sobre propuestas para cerrar la brecha de comunicación entre personas sordas y oyentes.

CAPÍTULO 11

Bibliografía

[1] "Día internacional de las lenguas de señas, 23 de septiembre," Naciones Unidas. (2017), dirección: https://www.un.org/es/observances/sign-languages-day#:~:text=Seg%C3%BAn%20la%20Federaci%C3%B3n%20Mundial%20de,300%20diferentes%20lenguas%20de%20se%C3%B1as.

[2] "LENSEGUA, Ley que fomenta la inclusión social," Congreso de la República. (2022), dirección: https://www.congreso.gob.gt/noticias_congreso/9131/2022/4#gsc.tab=0.

[3] M. Chin. "Wearable-tech glove translates sign language into speech in real time." (2020), dirección: https://newsroom.ucla.edu/releases/glove-translates-sign-language-to-speech.

[4] "The BrightSign Glove." Recuperado de https://www.brightsignglove.com/. (2019), dirección: https://www.brightsignglove.com/.

[5] N. C. Camgoz, B. Saunders, G. Rochette et al., *Open Research Sign Language Translation Datasets*, 2021. arXiv: 2105.02351 [cs.CV].

[6] M. Nochel, "SignAll Invents Revolutionary Tool to Translate Sign Language," 2022.

[7] I. Adeyanju, O. Bello y M. Adegboye, "Machine learning methods for sign language recognition: A critical review and analysis," *Intelligent Systems with Applications*, vol. 12, pág. 200 056, 2021. DOI: 10.1016/j.iswa.2021.200056. dirección: https://doi.org/10.1016/j.iswa.2021.200056.

[8] "Declaración Universal de los Derechos Humanos," Naciones Unidas. (1948), dirección: https://www.ohchr.org/sites/default/files/UDHR/Documents/UDHR_Translations/spn.pdf.

[9] "Alfabetización para el desarrollo." (2021), dirección: https://www.unesco.org/es/articles/alfabetizacion-para-el-desarrollo.

[10] "Convención sobre los derechos de las personas con discapacidad," Naciones Unidas. (2008), dirección: https://www.un.org/esa/socdev/enable/documents/tccconvs.pdf.

[11] S. Pabón, "La discapacidad auditiva. ¿Cómo es el niño sordo?" *Revista CSIF*, n.º 16, págs. 81-91, 2009. dirección: https://archivos.csif.es/archivos/andalucia/ensenanza/revistas/csicsif/revista/pdf/Numero_16/SABINA_PABON_2.pdf.

[12] A. B. Domínguez y P. Alonso, *La educación de los alumnos sordos hoy. Perspectivas y respuestas educativas.* Málaga: Ediciones Aljibe, 2004.

[13] E. Ruiz, "El aprendizaje de la lectoescritura en los niños y niñas sordos," *Caleidoscopio, revista de contenidos educativos del CEP de Jaén*, n.º 2, págs. 129-144, 2009. dirección: https://dialnet.unirioja.es/servlet/articulo?codigo=3095863.

[14] "¿Qué es la lectura fácil (LF)?" Lectura Fácil. (), dirección: https://www.lecturafacil.net/es/info/1-que-es-la-lectura-facil-lf/.

[15] A. Espínola, "Accesibilidad auditiva. Pautas básicas para aplicar en los entornos.," *Colección democratizando la accesibilidad*, vol. 7, 2015. dirección: https://sid.usal.es/idocs/F8/%20FDO27110/Accesibilidad%20auditiva.pdf.

[16] V. Herrera, A. M. Puente, J. Alvarado y A. Ardila, "Códigos de lectura en sordos: la dactilología y otras estrategias visuales y kinestésicas," *Revista Latinoamericana de Psicología*, vol. 39, n.º 2, págs. 269-286, 2007. dirección: http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S0120-05342007000200005&lng=pt&tlng=es.

[17] K. Hirsh-Pasek, "The metalinguistic of fingerspelling: An alternate way to increase reading vocabulary in congenitally deaf readers," *Reading research Quarterly*, vol. XXII, n.º 4, págs. 455-473, 1987.

[18] "MediaPipe Hand Landmark," Google Developers. (2023), dirección: https://developers.google.com/mediapipe/solutions/vision/hand_landmarker.

[19] L. Rouhiainen, *Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro.* Alienta, 2018.

[20] L. Floridi, J. Cowls, M. Beltrametti et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines*, vol. 28, n.º 4, págs. 689-707, 2018.

CAPÍTULO 12

Anexos

Github: https://github.com/andreamalin/megaproyecto_model

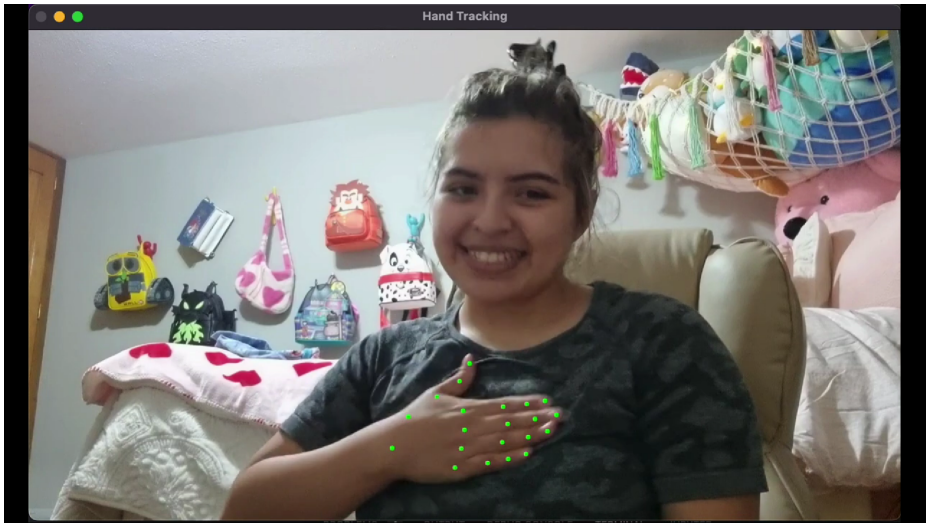

Figura 9: Detección de las coordenadas para la palabra: *please*

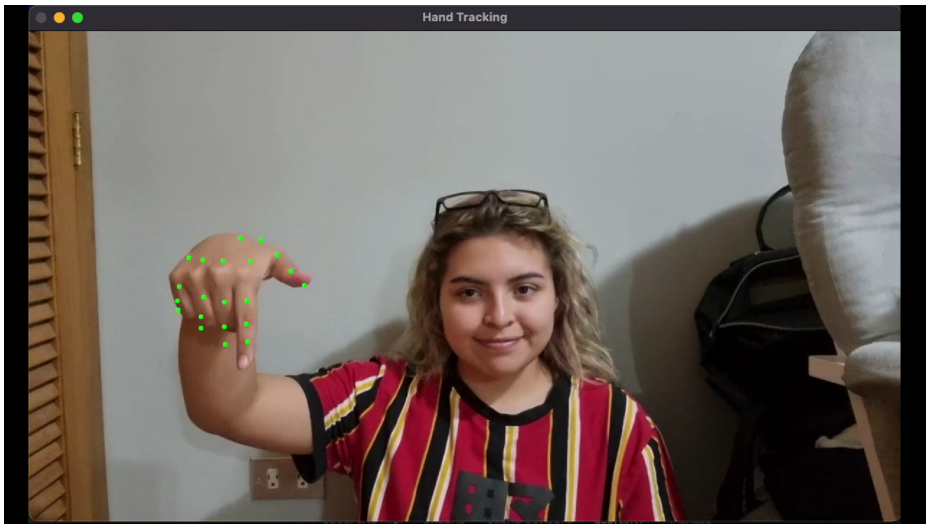

Figura 10: Detección de las coordenadas para la palabra: *see you later*

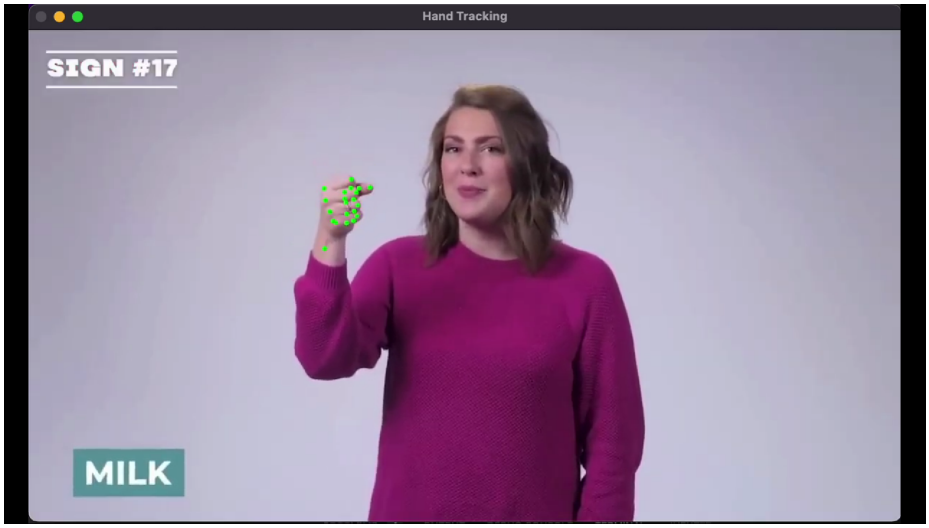

Figura 11: Detección de las coordenadas para la palabra: *milk*

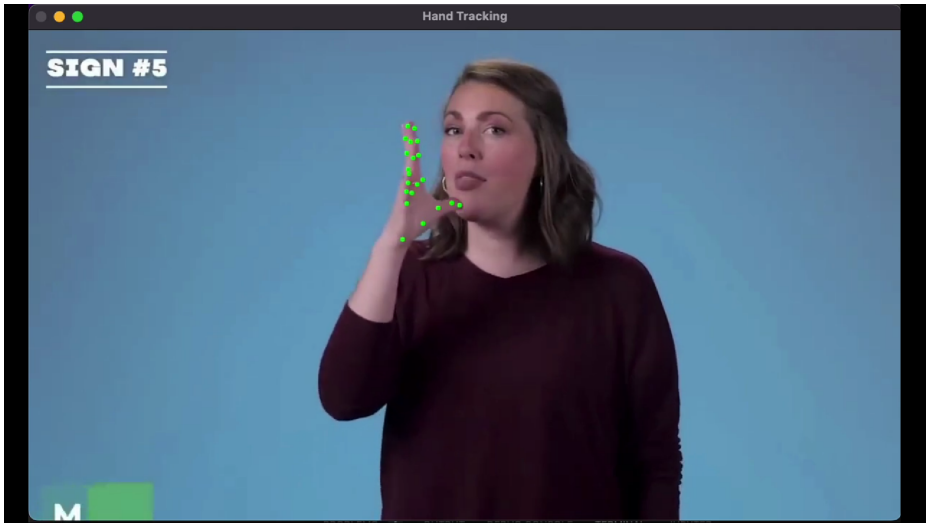

Figura 12: Detección de las coordenadas para la palabra: *father*

Figura 13: Detección de las coordenadas para la palabra: termómetro digital

Figura 14: Detección de las coordenadas para la palabra: querer

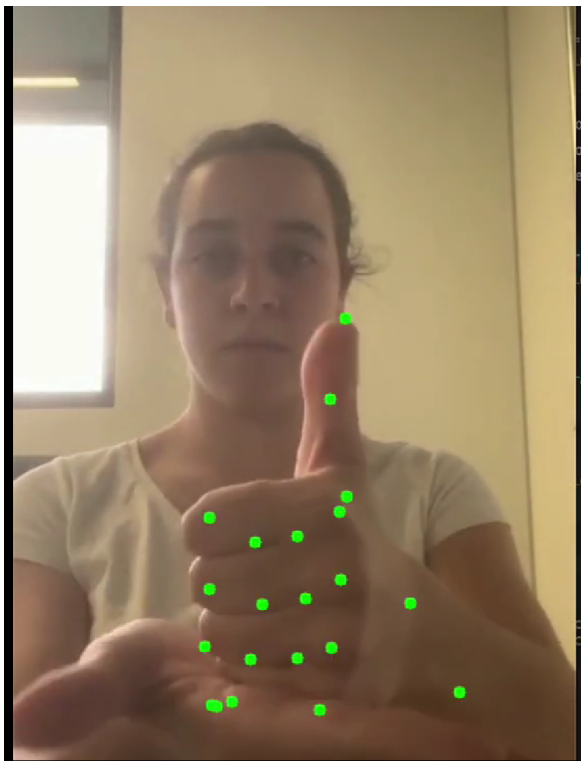

Figura 15: Detección de las coordenadas para la palabra: ayuda

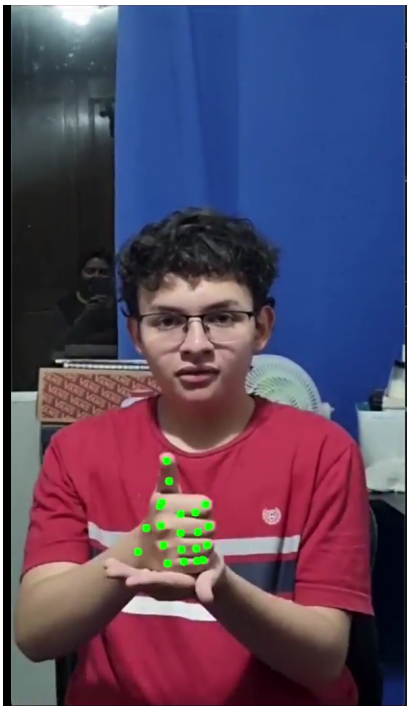

Figura 16: Detección de las coordenadas para la palabra: ayuda

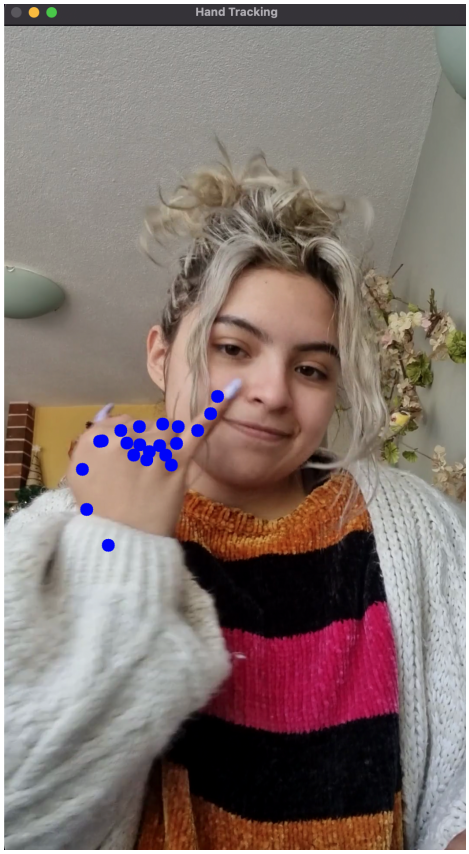

Figura 17: Detección de las coordenadas para la letra con movimiento: j