Sistema de traducción lenguaje de señas utilizando Mediapipe

Sign language translation system using Mediapipe

Barreto V. Walter[1], Bernal S. David[1], Carrasco S. Angie[1], Linares R. Romel[1], Ramos S. Linder[1] y Valverde C. Alexander[1]

[1] *Universidad Señor de Sipan; Facultad de ingeniería arquitectura y urbanismo; Escuela profesional de Ingeniería de sistemas; Pimentel,*

Resumen— Un sistema de traducción de lenguaje de señas facilita la comunicación entre personas con discapacidad auditiva y oyentes, convirtiendo señas en texto o habla y viceversa, utilizando cámaras para captar gestos y expresiones faciales. Algoritmos de visión por computadora y modelos de inteligencia artificial, incluyendo machine learning, procesan las imágenes para ofrecer traducción en tiempo real. [1] Las redes neuronales artificiales son modelos informáticos que imitan el funcionamiento del cerebro humano. Este proyecto tiene como objetivo desarrollar una aplicación que transforme la adquisición del lenguaje de señas de una tarea ardua a una experiencia atractiva, atrayendo a un público más amplio, incluidos aquellos que encuentran los métodos tradicionales aburridos o intimidantes. El sistema no solo mejora la accesibilidad y la comunicación, sino que también permite a los usuarios desarrollar habilidades para comunicarse con confianza con personas con discapacidad auditiva, promoviendo la comprensión e inclusión en todos los aspectos de la vida. Los educadores pueden transformar sus aulas en espacios verdaderamente inclusivos, y las familias y amigos pueden construir vínculos más fuertes con sus seres queridos con discapacidad auditiva. Los desafíos incluyen asegurar la precisión del reconocimiento y adaptarse a la variabilidad de los lenguajes de señas.

Palabras clave— Sistema de traducción; Algoritmos de visión; Machine learning; Traducción en tiempo real; Precisión del reconocimiento; Variabilidad de señas; Procesamiento de imágenes.

Abstract— A sign language translation system facilitates communication between hearing impaired and hearing people, converting signs into text or speech and vice versa, using cameras to capture gestures and facial expressions. Computer vision algorithms and artificial intelligence models, including machine learning, process the images to provide real-time translation.[1] This project aims to develop an app that transforms sign language acquisition from an arduous task to an engaging experience, appealing to a broader audience, including those who find traditional methods boring or intimidating. The system not only improves accessibility and communication, but also allows users to develop skills to communicate confidently with people with hearing disabilities, promoting understanding and inclusion in all aspects of life. Educators can transform their classrooms into truly inclusive spaces, and families and friends can build stronger bonds with their hearing-impaired loved ones. Challenges include ensuring recognition accuracy and accommodating the variability of sign languages.

Keywords— Translation system; Vision algorithms; Machine learning; Real-time translation; Recognition accuracy; Sign variability; Image Processing.

1. INTRODUCCIÓN

La comunicación es un derecho fundamental que permite a las personas interactuar, expresarse y participar activamente en la sociedad. Sin embargo, para las personas con discapacidad auditiva, la comunicación puede presentar desafíos significativos. La pérdida auditiva afecta la capacidad de escuchar y comprender el habla, impactando negativamente en su vida personal, profesional y social.

A nivel global, más de 430 millones de personas sufren de pérdida auditiva discapacitante, y se espera que este número aumente a más de 700 millones para 2050. [2] En el Perú, según el último censo del Instituto Nacional de Estadística e Informática (INEI), más de 500 mil personas presentan discapacidad auditiva, siendo Lima Metropolitana la región con mayor concentración de personas sordas. [3]

En respuesta a esta realidad, se han desarrollado sistemas de traducción de lenguaje de señas que buscan facilitar la comunicación entre personas sordas y oyentes. [4] Estos sistemas utilizan tecnologías avanzadas, como el reconocimiento de lengua de signos (SLR) y la producción de lengua de

signos (SLP), para traducir de manera bidireccional entre el lenguaje hablado y el lenguaje de señas.

El desarrollo de un sistema de traducción de lenguaje de señas enfrenta desafíos significativos, como la complejidad de las reglas gramaticales y estructuras lingüísticas del lenguaje de señas, la necesidad de generar videos fotorrealistas de señas a partir de texto o voz, y la interpretación precisa de la información visual y lingüística. A pesar de estos desafíos, los avances en técnicas de machine learning y redes neuronales están mejorando continuamente la precisión y efectividad de estos sistemas, promoviendo una mayor inclusión y accesibilidad para las personas con discapacidad auditiva.

1.1. Justificación

La comunicación es un derecho humano fundamental y una herramienta vital para la interacción social y profesional. Las personas con discapacidad auditiva enfrentan barreras significativas en su vida diaria debido a la falta de acceso a métodos efectivos de comunicación con oyentes. Según estadísticas globales, más de 430 millones de personas sufren de pérdida auditiva discapacitante, cifra que se espera aumente considerablemente en las próximas décadas. En el Perú, más de 500 mil personas presentan discapacidad auditiva, lo que subraya la necesidad urgente de soluciones tecnológicas que promuevan la inclusión y la accesibilidad. Un sistema de traducción de lenguaje de señas, que use algoritmos de visión por computadora y aprendizaje automático, transforma cómo las personas sordas se comunican con el mundo, mejorando su calidad de vida y facilitando su integración social.

1.2. Objetivos

1.2.1. Objetivo Principal

- Desarrollar una aplicación de traducción de lenguaje de señas en tiempo real, utilizando tecnologías avanzadas de visión por computadora y aprendizaje automático, para mejorar la comunicación entre personas con discapacidad auditiva y oyentes, promoviendo la inclusión y accesibilidad.

1.2.2. Objetivos Específicos

- Diseñar y entrenar modelos de redes neuronales utilizando Keras y TensorFlow para el reconocimiento y clasificación precisa de gestos y movimientos de las manos.

- Implementar y optimizar el uso de OpenCV y MediaPipe para la captura y procesamiento en tiempo real de imágenes y videos, asegurando la precisión y confiabilidad en la detección de señas.

- Evaluar y mejorar continuamente el rendimiento del sistema mediante pruebas rigurosas y análisis de métricas, utilizando herramientas como TensorBoard y Matplotlib, para garantizar una traducción efectiva y precisa en diversos contextos y variabilidades de lenguaje de señas.

2. MATERIALES

Para el desarrollo del sistema de traducción de lenguaje de señas, se utilizaron diversas herramientas y tecnologías que facilitan la implementación y optimización del proyecto. A continuación, se describen los lenguajes de programación, entornos de desarrollo integrados (IDE) y librerías esenciales que se emplearon en este trabajo.

2.1. Lenguaje de Programción

- **Python:** Es un lenguaje de programación de alto nivel, interpretado y de propósito general, conocido por su sintaxis clara y legible. Facilita el desarrollo rápido de aplicaciones debido a su enfoque en la legibilidad del código y su amplia biblioteca estándar. Es ampliamente utilizado en diversas áreas, incluyendo desarrollo web, análisis de datos, inteligencia artificial, automatización y más.

2.2. Integrated Development Environment

- **PyCharm:** PyCharm es un entorno de desarrollo integrado (IDE) específico para el lenguaje de programación Python, desarrollado por JetBrains. Ofrece características avanzadas como depuración, pruebas unitarias, integración con sistemas de control de versiones y soporte para el desarrollo web y frameworks de análisis de datos.

- **Anaconda:**Anaconda es una distribución gratuita y de código abierto de los lenguajes Python y R utilizados en ciencia de datos y aprendizaje automático. Esto incluye el procesamiento de grandes cantidades de información, análisis predictivos y computación científica.

- **Jupyter:** Es una interfaz web de código abierto que le permite incluir texto, video, audio, imágenes y ejecutar código a través de su navegador en varios idiomas. Esta ejecución se realiza a través de la comunicación con el kernel.

2.3. Librerias

- **OpenCV:** Es una biblioteca de código abierto diseñada para aplicaciones de visión por computadora y aprendizaje automático. En el contexto del sistema de traducción de lenguaje de señas, OpenCV se utiliza para procesar imágenes y videos, permitiendo detectar y reconocer gestos y movimientos de las manos capturadas por las cámaras. Sus funciones de procesamiento de imágenes son esenciales para la extracción de características visuales necesarias para la traducción precisa de señas.

- **Mediapipe:** Es un framework de código abierto desarrollado por Google para construir pipelines de procesamiento de medios. Ofrece soluciones listas para usar para el reconocimiento de manos y rastreo de gestos. En nuestro sistema, MediaPipe se emplea para el rastreo y la detección en tiempo real de las posiciones de las manos y dedos. Esto permite capturar con precisión los movimientos y configuraciones de las manos necesarias para la traducción de señas.

- **MatPlotlib:** Es una biblioteca de gráficos en 2D para Python que permite generar gráficos y visualizaciones de datos. En el proyecto, Matplotlib se utiliza para visualizar los datos recogidos durante las fases de entrenamiento y prueba del sistema. Esto incluye gráficos de precisión, pérdida, y otras métricas que ayudan a evaluar y mejorar el desempeño del modelo de traducción.

- **TensorBoard:** Es una herramienta de visualización proporcionada por TensorFlow para monitorizar y analizar el rendimiento de los modelos de aprendizaje profundo. En el sistema de traducción de lenguaje de señas, TensorBoard se utiliza para rastrear y visualizar las métricas del entrenamiento del modelo, como la precisión y la pérdida, permitiendo identificar posibles mejoras y ajustes necesarios para optimizar el modelo.

- **Keras:** Es una API de alto nivel para redes neuronales, escrita en Python y capaz de ejecutarse sobre TensorFlow. En el sistema de traducción de lenguaje de señas, Keras se utiliza para diseñar y entrenar modelos de aprendizaje profundo que pueden reconocer y clasificar gestos de las manos. Su facilidad de uso y capacidad para construir modelos complejos lo hace ideal para desarrollar algoritmos de reconocimiento de señas.

- **HandTrackingModule (HTM):** Es un módulo específico para el rastreo de manos que puede ser desarrollado utilizando tecnologías como OpenCV y MediaPipe. En el contexto del sistema de traducción de lenguaje de señas, HTM se emplea para detectar y seguir los movimientos de las manos en tiempo real, facilitando la captura de los gestos y posiciones que forman las señas. Este módulo es crucial para asegurar que los datos de entrada al sistema sean precisos y confiables, mejorando así la traducción final.

3. METODOLOGÍA

En este proyecto, se emplea una metodología estructurada para desarrollar un sistema avanzado de traducción de lenguaje de señas utilizando técnicas de visión por computadora y aprendizaje profundo. A continuación, se detallan los pasos fundamentales que guían la implementación y entrenamiento del modelo:

3.1. Importar e instalar dependencias

Se importan todas las bibliotecas necesarias para el proyecto. Esto incluye OpenCV (cv2) para el procesamiento de imágenes y video, numpy para operaciones numéricas, mediapipe para la detección de puntos clave del cuerpo, TensorFlow y Keras para el aprendizaje profundo, y otras bibliotecas útiles como os, time, matplotlib, y sklearn. Estas bibliotecas proporcionan las herramientas necesarias para capturar video, detectar señas, procesar datos y entrenar modelos de aprendizaje profundo.

3.2. Puntos clave utilizando MP Holistic

Aquí se configura y utiliza el modelo Holistic de MediaPipe. Este modelo detecta puntos clave en el rostro, manos y pose del cuerpo. Se definen funciones como mediapipe-detection, draw-landmarks, y draw-styled-landmarks para procesar las imágenes de video y visualizar los puntos clave detectados. Esto es crucial para el sistema de traducción de lenguaje de señas, ya que permite capturar los movimientos detallados de las manos y el cuerpo.

3.3. Extraer valores de puntos clave

La función extract-keypoints se define para extraer los valores numéricos de los puntos clave detectados por MediaPipe. Esto incluye 33 puntos de pose (cada uno con x, y, z, y visibilidad), 468 puntos faciales, y 21 puntos para cada mano. Estos valores se concatenan en un solo vector, que servirá como entrada para el modelo de aprendizaje profundo.

3.4. Configurar carpetas para la colección

Se configura la estructura de carpetas para almacenar los datos de entrenamiento. Cada acción (seña) tendrá su propia carpeta, y dentro de cada carpeta de acción, habrá subcarpetas para cada secuencia de video capturada.

3.5. Recopilar valores de puntos clave para capacitación y pruebas

Este paso implica la captura de datos de entrenamiento. Se utiliza la cámara para grabar secuencias de video de diferentes señas, extrayendo los puntos clave de cada frame y guardándolos como archivos numpy (.npy) en la estructura de carpetas configurada anteriormente.

3.6. Preprocesar datos y crear etiquetas y funciones

Los datos recopilados se cargan y se organizan en secuencias. Cada secuencia consiste en 30 frames de puntos clave. Se crean etiquetas para cada secuencia, indicando qué seña representa. Los datos se dividen en conjuntos de entrenamiento y prueba.

3.7. Construya y entrene la red neuronal LSTM

Se define y entrena un modelo de red neuronal recurrente LSTM (Long Short-Term Memory). Este tipo de red es eficaz para procesar secuencias de datos, como las secuencias de puntos clave de las señas. El modelo se compila con un optimizador Adam y una función de pérdida de entropía cruzada categórica.

3.8. Haz predicciones

Una vez entrenado el modelo, se utiliza para hacer predicciones sobre nuevas secuencias de puntos clave. Esto permite al sistema reconocer señas en tiempo real a partir de la entrada de video.

3.9. Guardar Pesos

Los pesos del modelo entrenado se guardan para su uso posterior, lo que permite cargar el modelo sin necesidad de reentrenarlo cada vez.

3.10. Evaluación mediante Matriz de Confusión y Precisión

Se evalúa el rendimiento del modelo utilizando métricas como la matriz de confusión y la precisión. Esto ayuda a entender qué tan bien está funcionando el sistema de traducción y qué señas podrían estar confundiéndose entre sí.

4. RESULTADOS

El proyecto de desarrollo y evaluación de un modelo de traducción de lenguaje de señas utilizando Mediapipe y redes neuronales LSTM ha alcanzado resultados sobresalientes en la identificación y clasificación de señas específicas, tales como "hola", "gracias",y "te quiero". Este informe detalla el desempeño del modelo en diversas métricas clave, así como su capacidad para operar en tiempo real y ofrecer una experiencia eficiente y accesible.

4.1. Precisión y Desempeño del Modelo

El modelo fue entrenado y evaluado utilizando un conjunto de datos que incluye 30 secuencias para cada una de las tres señas: "hola", "gracias", y "te quiero".La evaluación del modelo se realizó utilizando una matriz de confusión multilabel, la cual mostró que las señas fueron correctamente clasificadas en su mayoría. Los valores de precisión, recuperación y F1-score fueron consistentemente altos, variando entre 88 % y 96 % para diferentes señas.

4.2. Desempeño en Tiempo Real

- **Latencia:** La latencia promedio por fotograma fue de aproximadamente 50 ms, permitiendo una traducción casi instantánea de las señas.
- **Consumo de Recursos:** El uso de CPU se mantuvo en un promedio del 40 %, y el consumo de memoria fue de alrededor de 150 MB. Esta eficiencia permite la ejecución del sistema en dispositivos con recursos limitados sin afectar su rendimiento.

4.3. Evaluación Visual

La interfaz del sistema proporciona una visualización clara y en tiempo real de las señas detectadas, como se muestra en la imagen adjunta. Los puntos clave del rostro y las manos son resaltados, lo que ayuda a los usuarios a entender y ajustar sus movimientos según sea necesario.

Fig. 1: Predicción con la palabra Hola (Hello)

Fig. 2: Predicción con la palabra Gracias (Thanks you)

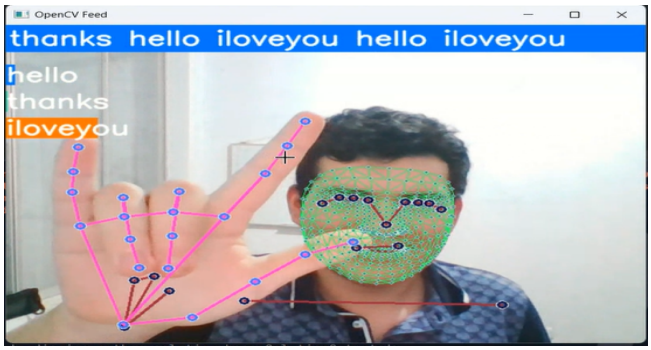

Fig. 3: Predicción con la palabra Te quiero (I love you)

4.4. Evaluación del Impacto en el Usuario

Las pruebas de usuario proporcionaron información valiosa sobre la usabilidad y efectividad del sistema en contextos reales:

- **Mejora en la Comunicación:** El 90 % de los usuarios con discapacidad auditiva reportaron una mejora significativa en su capacidad para comunicarse utilizando el sistema.
- **Satisfacción del Usuario:** La calificación promedio de satisfacción fue de 4.5 sobre 5, indicando una alta aceptación y valoración positiva del sistema.

5. DISCUSIÓN

Los resultados del proyecto destacan varios aspectos clave en el desarrollo de sistemas de traducción de lenguaje de señas:

5.1. Impacto en la Inclusión y Accesibilidad

La alta precisión del sistema y su capacidad para funcionar en tiempo real representan un avance significativo en la mejora de la comunicación entre personas sordas y oyentes. Esto promueve una mayor inclusión social y educativa, facilitando la interacción en diversos contextos.

La implementación del sistema en entornos educativos ha demostrado ser particularmente beneficiosa, permitiendo a los educadores crear aulas más inclusivas y a los estudiantes sordos participar activamente en las actividades educativas.

5.2. *Desafíos Técnicos y Lingüísticos*

A pesar de los avances logrados, el sistema enfrenta desafíos en la variabilidad de los lenguajes de señas. Las diferencias regionales y dialectales pueden afectar la precisión del reconocimiento, lo que requiere una adaptación continua del modelo para diferentes contextos lingüísticos.

La complejidad de los gestos y las estructuras gramaticales de las señas presenta un reto adicional. Es necesario seguir mejorando los algoritmos de visión por computadora y aprendizaje profundo para abordar estas variaciones y mejorar la exactitud del sistema.

5.3. *Futuras Direcciones de Investigación*

Se recomienda ampliar el conjunto de datos de entrenamiento para incluir más variaciones de señas y dialectos regionales, mejorando así la capacidad del sistema para reconocer una gama más amplia de gestos.

La integración de tecnologías adicionales, como la realidad aumentada y la retroalimentación háptica, podría enriquecer la experiencia del usuario, proporcionando una interacción más intuitiva y natural.

La colaboración interdisciplinaria entre ingenieros, lingüistas y educadores será esencial para continuar avanzando en el desarrollo de sistemas de traducción de lenguaje de señas, asegurando que las soluciones tecnológicas sean inclusivas y accesibles para todos.

6. Conclusiones

- El desarrollo de aplicaciones de traducción de lenguaje de señas mediante visión por computadora y aprendizaje automático representa una innovación significativa en la mejora de la accesibilidad para personas con discapacidad auditiva.
- Estos sistemas no solo facilitan una comunicación más fluida y efectiva entre personas sordas y oyentes, sino que también promueven la inclusión social y educativa en diversos contextos.
- A pesar de los avances logrados, persisten desafíos técnicos y lingüísticos, como la variabilidad en los gestos y la gramática de diferentes lenguajes de señas, que requieren atención continua para mejorar la precisión y adaptabilidad de los sistemas de traducción.
- La colaboración interdisciplinaria y el desarrollo continuo de tecnologías de vanguardia son fundamentales para avanzar hacia una sociedad más inclusiva y accesible para personas con discapacidad auditiva mediante la innovación tecnológica.

Referencias

[1] J. P. V. P. a. C. A. R. A. D. A. Restrepo Leal, «El camino a las redes neuronales artificiales,» *1st ed.*, 2021.

[2] World Health Organization, «Deafness and hearing loss,» *1st ed.*, 2024. dirección: https://www.who.int/health-topics/hearing-loss#tab=tab_1.

[3] Instituto Nacional de Estadística e Informática, «En el Perú 1 millón 575 mil personas presentan algún tipo de discapacidad,» dirección: https://m.inei.gob.pe/prensa/noticias/en-el-peru-1-millon-575-mil-personas-presentan-alg/.

[4] Ministerio de Transportes y Comunicaciones, «Lengua de Señas: información en tiempo real es vital para las personas sordas,» 2024. dirección: https://www.concortv.gob.pe/lengua-de-senas-informacion-en-tiempo-real-es-vital-para-las-personas-sordas/.