Estado de Arte – Lengua de Señas Mexicano (LSM)

Christopher Flores González – A01795419

Luis Felipe Nicanor Gutiérrez - A01795631

Luis Ángel Seda Marcos - A01795301

Instituto Tecnológico y de Estudios Superiores de Monterrey

Estado del Arte:

Reconocimiento y Traducción de Lenguaje de Señas Mediante Inteligencia Artificial

Introducción

La comunicación es un pilar fundamental de la interacción humana. Sin embargo, existen barreras significativas entre las comunidades de personas sordas, que utilizan lenguajes de señas como la Lengua de Señas Mexicana (LSM), y la población oyente. Para mitigar esta brecha, la investigación en visión por computadora e inteligencia artificial ha explorado, durante más de tres décadas, el desarrollo de sistemas de Reconocimiento de Lenguaje de Señas (SLR) y Traducción de Lenguaje de Señas (SLT) (Cooper, Holt, & Bowden, 2007; Ramírez & Esquivel). Lo que comenzó como un subcampo del reconocimiento de gestos (GR) ha evolucionado hacia un área de estudio compleja que integra lingüística, procesamiento de video y modelos avanzados de aprendizaje automático. Este documento presenta una revisión del estado del arte, destacando la evolución en la adquisición de datos, las técnicas de extracción de características y los modelos de clasificación y traducción, para finalmente identificar los vacíos de investigación y contextualizar la relevancia de un nuevo proyecto.

1. Evolución en la Adquisición de Datos y Detección de Manos

El primer paso crucial en cualquier sistema de SLR/SLT es la captura precisa de los gestos. Las metodologías han evolucionado desde dispositivos intrusivos hacia soluciones basadas en visión, volviéndose cada vez más accesibles y naturales.

- **Dispositivos Invasivos y Sensores Especializados:** Los primeros sistemas dependían de hardware especializado como guantes de datos (data gloves) y acelerómetros para medir directamente los ángulos de las articulaciones, la orientación y el movimiento de las manos (Cooper et al., 2007). Aunque precisos, estos métodos son costosos, limitan la

naturalidad del movimiento del señante y no son escalables para un uso generalizado (Oszust & Wysocki, 2013).

- **Sistemas Basados en Visión y Profundidad:** La transición hacia cámaras de visión artificial eliminó la necesidad de dispositivos portátiles. Inicialmente, se enfrentaron desafíos como la segmentación de la mano del fondo, lo que se solucionaba con fondos controlados o el uso de guantes de colores (Cooper et al., 2007). La llegada de sensores de profundidad como el Microsoft Kinect representó un avance significativo, ya que proporcionaba un "mapa de profundidad" y un "esqueleto" del cuerpo con 20 articulaciones clave, simplificando enormemente la segmentación y el seguimiento de las manos al ser insensible a las condiciones de iluminación y al color de fondo (Oszust & Wysocki, 2013; Prasad & Shibu, 2018).
- **Detección de Puntos Clave en Cámaras RGB (Estado Actual):** El avance más reciente y democratizador es el desarrollo de modelos de aprendizaje profundo capaces de detectar con alta precisión la estructura de la mano a partir de una sola cámara RGB estándar. La solución **MediaPipe Hands** de Google es un ejemplo paradigmático de esta tecnología. Implementa un pipeline de dos etapas: primero, un detector de palmas (BlazePalm) que localiza la mano en la imagen, y segundo, un modelo de puntos de referencia (landmarks) que predice con alta fidelidad la ubicación de **21 puntos clave tridimensionales** en la mano (Zhang et al., 2020). Esta tecnología elimina la necesidad de hardware especializado y permite el desarrollo de aplicaciones en tiempo real en dispositivos móviles, siendo la base tecnológica sobre la cual se proponen muchos proyectos actuales (Ramírez & Esquivel).

2. Metodologías para la Extracción de Características

Una vez detectada la mano y sus puntos clave, es necesario extraer un vector de características (feature vector) que describa de manera única y robusta la configuración del gesto.

- **Características Basadas en Apariencia:** Estos métodos analizan la imagen de la mano en su conjunto. Técnicas como el **Scale-Invariant Feature Transform (SIFT)** han sido utilizadas para extraer "puntos de interés" que describen la textura y forma del objeto, ofreciendo invarianza a la escala, rotación e iluminación (Prasad & Shibu, 2018). Sin embargo, pueden ser computacionalmente costosos y sensibles a la calidad de la segmentación.
- **Características Geométricas y Estructurales a partir de Puntos Clave:** El paradigma actual, impulsado por herramientas como MediaPipe y OpenPose, se centra en la caracterización geométrica a partir de los 21 puntos de referencia de la mano. Este enfoque es el núcleo del proyecto propuesto por Ramírez y Esquivel. La posición relativa de estos puntos permite calcular descriptores robustos como ángulos entre los dedos, distancias normalizadas, y curvaturas, que caracterizan la postura de la mano de forma independiente a su posición o tamaño en la imagen (Ansar et al., 2022).
- **Características Dinámicas para Gestos en Movimiento:** Para gestos dinámicos, que componen gran parte del lenguaje de señas, no solo importa la postura, sino también el movimiento. Se han propuesto técnicas como las **curvas de Bézier** para modelar la trayectoria de los puntos clave, el **diferenciamiento de fotogramas** (frame differencing) para capturar el cambio de posición y las **nubes de puntos 3D** para analizar la evolución de la forma de la mano en el tiempo (Ansar et al., 2022).

3. Modelos de Clasificación y Traducción

Con un vector de características definido, el siguiente paso es clasificar el gesto o traducir la secuencia de gestos.

- **Clasificadores Tradicionales:** Los primeros enfoques utilizaron algoritmos de aprendizaje automático clásicos. El **Support Vector Machine (SVM)** ha demostrado ser eficaz para clasificar gestos estáticos a partir de características SIFT (Prasad & Shibu,

2018). Para secuencias dinámicas, técnicas como **Dynamic Time Warping (DTW)** han sido empleadas para comparar series temporales de características de longitud variable, comúnmente en conjunto con un clasificador de vecino más cercano (Oszust & Wysocki, 2013).

- **Redes Neuronales Convolucionales (CNN):** Las CNNs se han convertido en una herramienta poderosa, tanto para la extracción de características directamente desde los píxeles de la imagen como para la clasificación de gestos. Se ha demostrado que arquitecturas de CNN profundas alcanzan altas tasas de reconocimiento, superando a los clasificadores tradicionales en tareas de reconocimiento de lenguaje de señas (Rao et al., 2018).
- **Modelos Secuencia a Secuencia (Seq2Seq) y el Reto de la Traducción Continua:** El lenguaje de señas no es una colección de gestos aislados, sino una secuencia continua con una estructura gramatical propia (Cooper et al., 2007). Esto hace que el problema sea más análogo a la traducción automática que a la simple clasificación. Los modelos **Seq2Seq**, basados en Redes Neuronales Recurrentes (RNN) como LSTM o GRU, se han adaptado para esta tarea. Estos modelos utilizan un codificador (encoder) para procesar la secuencia de fotogramas de video y un decodificador (decoder) para generar la secuencia de texto correspondiente (Ananthanarayana et al., 2021).
- **Mecanismos de Atención y Modelos Transformer (Estado del Arte Actual):** Una limitación de los modelos Seq2Seq básicos es su dificultad para manejar secuencias largas. La introducción de **mecanismos de atención** permitió al decodificador "enfocarse" en las partes más relevantes de la secuencia de entrada en cada paso de la traducción. El avance más disruptivo ha sido el **modelo Transformer**, que prescinde de la recurrencia y se basa completamente en mecanismos de auto-atención (self-attention), permitiendo procesar secuencias de manera paralela y capturar dependencias a larga

distancia de manera más efectiva. Estudios comparativos demuestran que los modelos Transformer, combinados con características de entrada robustas (como las de OpenPose o embeddings de CNNs), superan a otros modelos Seq2Seq en tareas de SLT, especialmente en datasets controlados (Ananthanarayana et al., 2021).

4. Desafíos Persistentes y la Importancia de los Datasets

A pesar de los avances, la comunidad científica enfrenta desafíos significativos que definen las fronteras actuales de la investigación.

- **Reconocimiento vs. Traducción:** Es crucial distinguir entre SLR (reconocer qué seña se hizo) y SLT (traducir el significado contextual). La traducción es un problema mucho más complejo que requiere no solo reconocer las señas manuales, sino también interpretar características no manuales (expresiones faciales, postura del cuerpo) y la gramática espacial del lenguaje (Cooper et al., 2007).
- **Manejo de Secuencias Largas:** Las frases largas en lenguaje de señas presentan problemas de dependencias a largo plazo y redundancia de información. Investigaciones recientes proponen soluciones innovadoras como el **Frame Stream Density Compression (FSDC)**, un algoritmo que detecta y elimina fotogramas redundantes o similares para acortar las secuencias sin perder información semántica, facilitando el trabajo de los modelos de traducción (Zheng, Zhao, Chen, et al., 2020).
- **La Necesidad de Datos Diversos y a Gran Escala:** El rendimiento de los modelos de aprendizaje profundo depende intrínsecamente de la cantidad y calidad de los datos de entrenamiento. Muchos datasets existentes son limitados en tamaño, número de señantes, condiciones de iluminación y fondos. Para abordar esta carencia, se ha desarrollado el dataset **HaGRID (Hand Gesture Recognition Image Dataset)**, que contiene más de 550,000 imágenes de 37,583 personas diferentes, ofreciendo una diversidad sin

precedentes que es fundamental para entrenar modelos robustos y generalizables (Kapitanov et al., 2023).

5. Identificación de Vacíos y Justificación del Proyecto Propuesto

El análisis del estado del arte revela que, si bien existen modelos de traducción de extremo a extremo (end-to-end) muy avanzados como los Transformers, su éxito depende de dos componentes fundamentales: (1) la disponibilidad de enormes datasets etiquetados y (2) una representación de características de entrada que sea rica y discriminativa.

En este contexto, se identifica un vacío de investigación clave: mientras gran parte del esfuerzo se ha centrado en la complejidad de la arquitectura del modelo de traducción, el estudio sistemático y la propuesta de **nuevas formas de caracterizar gestos estáticos a partir de los 21 puntos clave** ha recibido menos atención. La mayoría de los trabajos utilizan las coordenadas de los puntos directamente o con transformaciones simples.

El proyecto propuesto por Ramírez y Esquivel se posiciona estratégicamente para abordar esta necesidad fundamental. A diferencia de los enfoques que buscan una traducción completa, este trabajo se concentra en la etapa crucial de la **caracterización de gestos estáticos de la LSM**. Su objetivo es explorar y validar técnicas para transformar la información espacial de los 21 puntos en un vector de características optimizado que capture la esencia de cada seña. Este enfoque es relevante por varias razones:

1. **Fundacional:** Una clasificación precisa de gestos estáticos es un bloque de construcción indispensable para sistemas de reconocimiento de señas continuas más complejos.
2. **Eficiencia:** Desarrollar un conjunto de características robusto puede permitir el uso de clasificadores más ligeros y eficientes, ideal para aplicaciones en tiempo real.
3. **Especificidad Lingüística:** Al centrarse en la LSM, el proyecto puede desarrollar características que sean particularmente descriptivas para los gestos de esta lengua, aportando un valor localizado que los modelos genéricos podrían pasar por alto.

En resumen, mientras el estado del arte avanza hacia modelos de traducción más complejos, el proyecto propuesto refuerza una de las bases del problema: cómo representar de la mejor manera un gesto para que una máquina pueda entenderlo. Al hacerlo, contribuye directamente a mejorar la calidad de la entrada para los sofisticados modelos de traducción y abre la puerta a soluciones más eficientes y especializadas para la Lengua de Señas Mexicana.

Bibliografía

1. **Ansar, H., Ksibi, A., Jalal, A., Shorfuzzaman, M., Alsufyani, A., Alsuhibany, S. A., & Park, J. (2022).** Dynamic Hand Gesture Recognition for Smart Lifecare Routines via K-Ary Tree Hashing Classifier. *Applied Sciences*, *12*(13), 6481. https://doi.org/10.3390/app12136481
    - *[Citado en las secciones: 2 y 5 del presente documento.]*
2. **Ananthanarayana, T., Srivastava, P., Chintha, A., Santha, A., Landy, B., Panaro, J., ... & Nwogu, I. (2021).** Deep Learning Methods for Sign Language Translation. *ACM Transactions on Accessible Computing (TACCESS)*, *14*(4), 1-30. https://doi.org/10.1145/3477498
    - *[Citado en las secciones: 3 y 5 del presente documento.]*
3. **Cooper, H., Holt, B., & Bowden, R. (2007).** Sign Language Recognition. En *Guide to Visual Analysis of Humans* (pp. 537-562). Springer.
    - *[Citado en las secciones: Introducción, 1, 3 y 4 del presente documento.]*
4. **Kapitanov, A., Kvanchiani, K., Kraynov, R., Nagaev, A., & Makhliarchuk, A. (2023).** *HaGRID - HAnd Gesture Recognition Image Dataset*. (v2). arXiv. https://doi.org/10.48550/arXiv.2206.08219
    - *[Citado en la sección: 4 del presente documento.]*
5. **Oszust, M., & Wysocki, M. (2013).** Polish Sign Language words recognition with Kinect. En *2013 6th International Conference on Human System Interactions (HSI)* (pp. 219-226). IEEE.
    - *[Citado en las secciones: 1 y 3 del presente documento.]*
6. **Prasad, P. K., & Shibu, A. P. (2018).** Intelligent Human Sign Language Translation using Support Vector Machines Classifier. *International Journal of Research and Analytical Reviews (IJRAR)*, *5*(4), 461-466.

- *[Citado en las secciones: 1, 2 y 3 del presente documento.]*

7. **Ramírez, R. V., & Esquivel, O. A. (s.f.).** *Propuesta Proyecto Integrador LSM* [Presentación de proyecto]. Tecnológico de Monterrey, Maestría en Inteligencia Artificial Aplicada.
   - *[Citado en las secciones: Introducción, 1, 2 y 5 del presente documento.]*
8. **Rao, G. A., Syamala, K., Kishore, P. V. V., & Sastry, A. S. C. S. (2018).** Deep Convolutional Neural Networks for Sign Language Recognition. En *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 194-197). IEEE.
   - *[Citado en la sección: 3 del presente documento.]*
9. **Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. (2020).** *MediaPipe Hands: On-device Real-time Hand Tracking*. arXiv. https://doi.org/10.48550/arXiv.2006.10214
   - *[Citado en la sección: 1 del presente documento.]*
10. **Zheng, J., Zhao, Z., Chen, M., Chen, J., Wu, C., Chen, Y., Shi, X., & Tong, Y. (2020).** An Improved Sign Language Translation Model with Explainable Adaptations for Processing Long Sign Sentences. *Computational Intelligence and Neuroscience*, *2020*, 8816125. https://doi.org/10.1155/2020/8816125
   - *[Citado en la sección: 4 del presente documento.]*