



Universidad de Córdoba
Escuela Politécnica Superior de Córdoba
Grado de Ingeniería Informática, mención en Computación

Petición de tema de TFG:

Aprendiendo a ver el Lenguaje: Aplicaciones de Modelos de Visión y Lenguaje para Lenguas de Signos

Learning to See the Language: Vision and Language Models Applications for Sign Languages

Autor: Carlos David López Hinojosa

Director: Francisco José Madrid Cuevas

Tipo de TFG

Este trabajo es una obra de carácter original realizada por Carlos David López Hinojosa bajo la tutela de Francisco José Madrid Cuevas con tipología de investigación aplicada, o de desarrollo, sobre algún tema en concreto de la tecnología específica de cada titulación, Cualquier uso de fuentes externas está y será citado de la manera correspondiente.

Los apartados que formarán el documento de este trabajo serán los indicados en la sección 2.2 de la guía de realización para el TFG, para esta tipología en concreto son las partes listadas a continuación:

- Introducción
- Estado de la técnica
- Formulación del problema y objetivos
- Metodología de trabajo
- Desarrollo y Experimentación
- Resultados y discusión
- Conclusiones y Recomendaciones
- Bibliografía

Resumen

Este Trabajo de Fin de Grado tiene como objetivo investigar nuevas aplicaciones de la inteligencia artificial multimodal, combinando modelos de visión artificial basados en redes convolucionales tridimensionales con modelos de lenguaje, para abordar problemas de alta complejidad y relevancia social como el procesamiento de las lenguas de signos.

Como aplicación práctica, se propone el desarrollo de un sistema capaz de clasificar vídeos de lengua de signos según su categoría temática, explorando la capacidad de estos modelos para reconocer patrones visuales y semánticos en secuencias tridimensionales.

Además, se construirá un conjunto de datos propio a partir de material audiovisual disponible en internet, recopilado y etiquetado mediante técnicas de *web scraping*. Este proceso incluirá la selección, limpieza y normalización de los vídeos, así como su etiquetado automático por temática.

El trabajo pretende aportar una visión innovadora sobre el uso de la inteligencia artificial multimodal en contextos lingüísticos y de accesibilidad, demostrando cómo los modelos de visión y lenguaje pueden contribuir a la comprensión y análisis de sistemas de comunicación visual complejos.

Palabras clave: Visión artificial, Aprendizaje profundo, Lenguas de signos, Modelos multimodales, Clasificación temática.

Abstract

This Bachelor's Thesis aims to explore new applications of multimodal artificial intelligence by combining three-dimensional convolutional neural networks for visual processing with language models, addressing complex and socially relevant challenges such as sign language understanding.

As a practical implementation, the project focuses on developing a system capable of classifying sign language videos according to their thematic category, analyzing how these models can capture both visual and semantic patterns in three-dimensional sequences.

A complete and labeled dataset will be created using *web scraping* techniques to collect, clean, and normalize publicly available audiovisual material, followed by automatic thematic annotation.

This work seeks to provide an innovative perspective on the use of multimodal AI in linguistic and accessibility contexts, showing how vision and language models can contribute to the understanding and analysis of complex visual communication systems.

Keywords: Computer Vision, Deep Learning, Sign Languages, Multimodal Models, Thematic Classification.

Índice general

Índice de figuras	v
Índice de cuadros	vi
1. Introducción	1
2. Objetivos	2
3. Antecedentes	3
3.1. Investigaciones propias previas	4
3.2. Modelos utilizados como base	4
3.2.1. Modelo visual	4
3.2.2. Modelo de lenguaje	5
3.2.3. Modelo visual y de lenguaje	5
4. Fases del proyecto	7
4.1. Distribución temporal	9
5. Recursos y requerimientos	10
Bibliografía	11

Índice de figuras

3.1. Arquitectura propuesta para la extracción de características a partir de videos de lenguas de signos, procesamiento temporal usando transformers y clasificación por temática	6
--	---

Índice de cuadros

4.1. Cronograma estimado de trabajo del proyecto.	9
---	---

Lista de abreviaciones

AI	Artificial Intelligence — Inteligencia Artificial.
ASL	American Sign Language — Lengua de Signos Americana.
BSL	British Sign Language — Lengua de Signos Británica.
CNSE	Confederación Nacional de Personas Sordas de España.
CNN	Convolutional Neural Network — Red Neuronal Convolucional.
I3D	Inflated 3D ConvNet — Red Convolucional Inflada en 3D para video.
LSE	Lengua de Signos Española.
LSF	Langue des Signes Française — Lengua de Signos Francesa.
mBART	Multilingual Bidirectional and Auto-Regressive Transformer — Modelo Transformer multilingüe de codificación y decodificación.
ML	Machine Learning — Aprendizaje Automático.
NLP	Natural Language Processing — Procesamiento del Lenguaje Natural.
TFG	Trabajo de Fin de Grado.
XLM-R	Cross-Lingual Language Model - RoBERTa — Modelo Transformer multilingüe basado en RoBERTa.
GPU	Graphics Processing Unit — Unidad de Procesamiento Gráfico.
CPU	Central Processing Unit — Unidad Central de Procesamiento.
IDE	Integrated Development Environment — Entorno de Desarrollo Integrado.
VSCode	Visual Studio Code — Entorno de desarrollo multiplataforma.
API	Application Programming Interface — Interfaz de Programación de Aplicaciones.
PyTorch	Framework de aprendizaje profundo en Python.
HuggingFace	Plataforma de modelos y herramientas para procesamiento del lenguaje y visión.
BSLDict	British Sign Language Dictionary Dataset — Conjunto de datos de lengua de signos británica.

Capítulo 1

Introducción

La inteligencia artificial está transformando nuestra forma de comprender el mundo. Cada día surgen nuevas técnicas que no solo automatizan tareas, sino que abren posibilidades inéditas para analizar fenómenos complejos. Entre los campos donde estas tecnologías pueden marcar una diferencia significativa se encuentra el estudio de las lenguas de signos: sistemas de comunicación visuales, tridimensionales y de enorme riqueza expresiva.

Las lenguas de signos no son universales; cada país o comunidad posee la suya propia, como la Lengua de Signos Española (LSE), la *American Sign Language* (ASL) o la *Langue des Signes Française* (LSF), todas con gramáticas, léxicos y estructuras únicas. Esta diversidad, junto con la naturaleza visual y dinámica de la comunicación signada, plantea grandes retos para su análisis automatizado mediante técnicas de aprendizaje profundo.

El presente Trabajo de Fin de Grado tiene como objetivo desarrollar un sistema basado en modelos de inteligencia artificial multimodales, capaz de clasificar vídeos de lengua de signos según su categoría temática. Para ello, se combinarán modelos de visión artificial tridimensional, encargados de extraer las características visuales de los gestos, con modelos de lenguaje tipo *transformer*, que permitirán modelar la información temporal y semántica contenida en cada secuencia. Mediante el uso de mecanismos de atención, el sistema podrá identificar y concentrarse en las partes más relevantes de cada vídeo, ignorando información redundante o poco significativa.

Una parte fundamental del proyecto consistirá en la creación de un conjunto de datos propio, obtenido mediante técnicas de *web scraping* a partir de material audiovisual público disponible en distintas lenguas de signos. Este proceso incluirá la recopilación, limpieza y normalización de los vídeos, así como su etiquetado temático. Dicho conjunto servirá como base para entrenar y evaluar el modelo propuesto.

El trabajo se enmarca en el ámbito de la lingüística computacional multimodal, explorando cómo la inteligencia artificial puede aplicarse al análisis y comprensión de sistemas de comunicación visual complejos. Además de su relevancia técnica, el proyecto busca contribuir a la accesibilidad y al conocimiento de las lenguas de signos desde una perspectiva tecnológica e inclusiva.

Capítulo 2

Objetivos

El presente proyecto tiene como objetivo principal el desarrollo de un sistema multimodal basado en inteligencia artificial capaz de analizar y clasificar vídeos de lengua de signos según su categoría temática, combinando modelos de visión artificial tridimensionales y modelos de lenguaje preentrenados. Con este enfoque se busca explorar cómo las representaciones visuales y semánticas pueden integrarse para comprender mejor la comunicación gestual humana.

A partir de este objetivo general, se establecen los siguientes objetivos específicos:

1. Diseñar y construir un conjunto de datos propio de vídeos de lengua de signos, recopilado mediante técnicas de *web scraping*, e incluir etiquetas de categoría temática que permitan el entrenamiento supervisado del modelo.
2. Implementar un modelo de visión artificial tridimensional (I3D o equivalente) para la extracción de características visuales discriminativas a partir de los vídeos.
3. Integrar las representaciones visuales en un modelo *transformer* preentrenado con el fin de capturar dependencias temporales y semánticas en las secuencias de signos.
4. Entrenar y evaluar el sistema propuesto en la tarea de clasificación temática, utilizando métricas estándar de clasificación como la precisión y el *top-k accuracy*.
5. Analizar las regiones de atención y las representaciones aprendidas por el modelo para identificar qué fragmentos o patrones gestuales resultan más relevantes para cada categoría.

Con estos objetivos, el trabajo busca contribuir tanto al avance de las técnicas multimodales aplicadas a la visión y el lenguaje, como a la promoción del estudio computacional de las lenguas de signos, un ámbito de gran interés social y tecnológico.

Capítulo 3

Antecedentes

El estudio de las lenguas de signos mediante métodos de aprendizaje automático constituye un campo relativamente reciente dentro de la investigación en inteligencia artificial. La aparición de modelos de lenguaje y visión cada vez más potentes ha abierto nuevas fronteras para el análisis automático de la comunicación visual humana. Sin embargo, la mayoría de los trabajos existentes se centran en tareas monolingües, como la transcripción de glosas a texto o la traducción automática dentro de una misma lengua de signos, sin abordar de manera comparativa las diferencias y similitudes entre distintos idiomas signados.

Desde un punto de vista lingüístico, diversos estudios, como [Aronoff et al. \(2005\)](#), demuestran que, aunque existen similitudes estructurales entre distintas lenguas de signos, cada una mantiene características fonológicas, morfológicas y semánticas propias, incluso al representar conceptos equivalentes.

En el ámbito del aprendizaje profundo, trabajos como [Albanie et al. \(2020\)](#) o [Momeni et al. \(2020\)](#) han sentado las bases para la detección de glosas en vídeos de personas signantes, combinando modelos de visión tridimensional con información adicional, como las gesticulaciones faciales y los movimientos labiales que acompañan la comunicación. Estos enfoques han demostrado que la integración de información multimodal mejora significativamente la precisión del reconocimiento.

De forma más reciente, investigaciones como [Zhou et al. \(2023\)](#) proponen arquitecturas que combinan redes convolucionales para la extracción de características visuales con modelos de lenguaje preentrenados, aplicadas a la traducción de lengua de signos alemana a texto. Este tipo de trabajos constituyen una referencia fundamental para comprender cómo los modelos multimodales pueden aprender correspondencias entre representaciones visuales y lingüísticas, sirviendo como punto de partida para el presente proyecto.

Asimismo, se consideran como antecedentes las investigaciones previas y experimentos propios realizados durante la fase inicial de diseño de este trabajo de fin de grado.

3.1. Investigaciones propias previas

El desarrollo de la temática de este proyecto no ha sido trivial. Antes de llegar a la formulación actual, se exploraron distintas ideas relacionadas con el ámbito de las lenguas de signos, algunas de las cuales fueron descartadas por su complejidad técnica o por requerir recursos computacionales de gran escala. Entre ellas se encontraban propuestas como el desarrollo de una aplicación interactiva para facilitar el aprendizaje de lenguas de signos a niños, o un sistema de traducción automática de lengua de signos española (LSE) a texto.

Aunque estos proyectos resultaban muy atractivos desde el punto de vista social y tecnológico, fueron finalmente desestimados debido a la alta demanda de datos anotados y a la escasez de recursos disponibles en el dominio. Si bien existen algunos conjuntos de datos en lenguas como la BSL o la LSE, su acceso y procesamiento a gran escala sigue siendo un reto, tanto por limitaciones de licencia como por la heterogeneidad de formatos y anotaciones.

Esta limitación en la disponibilidad de datos motivó la exploración de enfoques más viables, orientados a la **extracción y análisis de características visuales y lingüísticas**, en lugar de la traducción directa. De esta reflexión surge la línea de trabajo del presente proyecto.

3.2. Modelos utilizados como base

3.2.1. Modelo visual

El modelo empleado para la obtención de características visuales en los vídeos de lengua de signos será el modelo *I3D* preentrenado descrito en los artículos [Albanie et al. \(2020\)](#) y [Momeni et al. \(2020\)](#).

A diferencia de otras redes convolucionales bidimensionales, el modelo *I3D* incorpora una dimensión temporal adicional que le permite capturar no solo las relaciones espaciales entre píxeles, sino también las relaciones temporales entre fotogramas. Esta capacidad lo hace especialmente adecuado para analizar secuencias de gestos en movimiento, como las que conforman la comunicación en lengua de signos.

El modelo fue entrenado sobre el conjunto de datos *BSLDict* y permite transformar una secuencia de vídeo de entrada ($F \times W \times H \times C$) donde F representa el número de fotogramas en un instante del video, W y H las dimensiones espaciales y C los canales de color en un vector de características de \mathbb{R}^{1064} .

Dado que cada vídeo contiene múltiples fragmentos temporales, el paso completo por el modelo produce una secuencia de embeddings:

$$[E_0, E_1, E_2, \dots, E_t]$$

donde cada E_t representa un embedding correspondiente al instante t del vídeo. Esta representación vectorial es análoga a la tokenización utilizada en los modelos de lenguaje, en los que cada token captura información contextual y semántica dentro de una secuencia.

3.2.2. Modelo de lenguaje

Como se mencionó en la sección 3.2.1, el modelo visual genera una secuencia de embeddings que representan la información temporal y espacial de un vídeo. Esta secuencia puede ser tratada de manera análoga a una secuencia de palabras en un modelo de lenguaje natural.

Los modelos *transformers*, ampliamente utilizados en procesamiento del lenguaje natural, son capaces de capturar relaciones semánticas entre tokens dentro de una secuencia, comprendiendo el significado contextual de cada elemento. Por tanto, proponemos utilizar los embeddings producidos por el modelo visual, adaptarlos dimensionalmente y procesarlos mediante un modelo transformer multilingüe preentrenado, con el objetivo de que aprenda a discriminar entre distintas lenguas de signos y temáticas.

Este planteamiento se inspira en trabajos como Zhou et al. (2023), en los que se combinan modelos visuales y de lenguaje para la traducción automática de vídeos de lengua de signos. En su caso, se emplean redes convolucionales junto con el modelo de lenguaje mBART, de arquitectura *encoder-decoder*. En nuestro caso, optamos por un modelo *encoder-only*, como XLM-RoBERTa, que resulta más ligero computacionalmente y se ajusta mejor a las necesidades de clasificación del presente proyecto.

3.2.3. Modelo visual y de lenguaje

Tomamos también inspiración de los siguientes artículos para las formulaciones anteriores Sun et al. (2019); Ging et al. (2020); Xu et al. (2021); Bain et al. (2021), los cuales emplean redes I3D o sus variantes (como S3D) para la extracción de embeddings espacio-temporales de clips de vídeo, que posteriormente son procesados mediante arquitecturas Transformer para modelar dependencias temporales o multimodales.

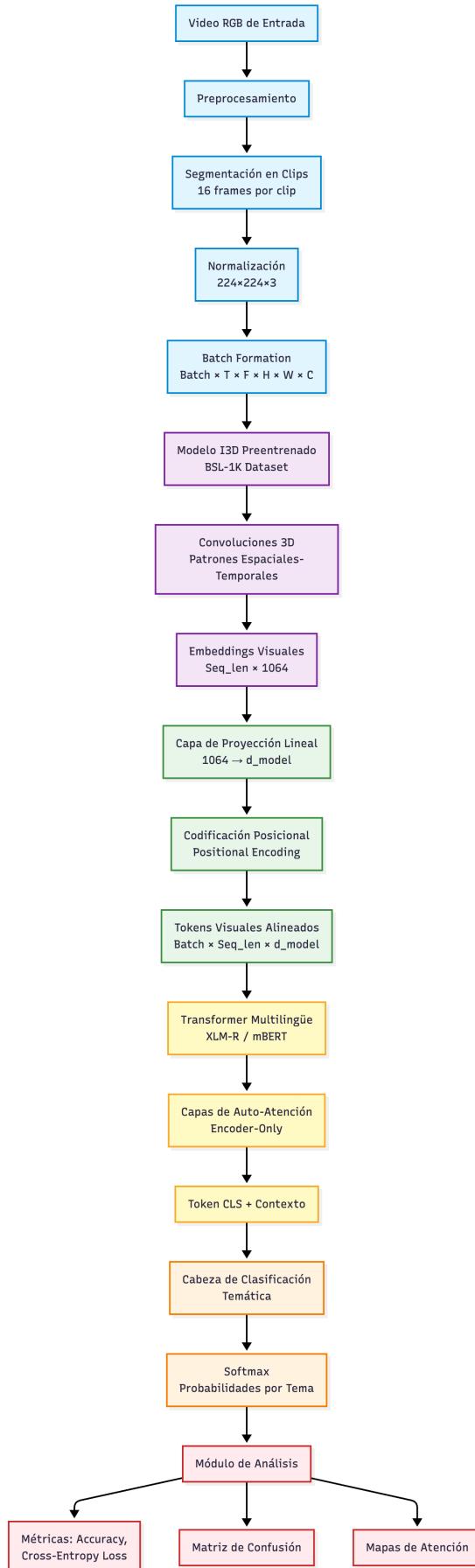


Figure 3.1: Arquitectura propuesta para la extracción de características a partir de videos de lenguas de signos, procesamiento temporal usando transformers y clasificación por temática

Capítulo 4

Fases del proyecto

El desarrollo del proyecto se estructura en una serie de fases que abarcan desde la recopilación y preparación de los datos hasta la evaluación y análisis de los resultados obtenidos. Cada fase persigue un conjunto de objetivos técnicos concretos y produce entregables que sirven de base para la siguiente etapa.

A continuación, se describen las distintas fases del proyecto:

Fase 1: Recolección y diseño del conjunto de datos

En esta primera fase se llevará a cabo la búsqueda, recopilación y organización de vídeos de lenguas de signos disponibles en plataformas abiertas. Para ello se emplearán técnicas de *web scraping* y filtrado automatizado, garantizando la diversidad temática y la calidad visual del material recopilado. El objetivo principal es conformar un conjunto de datos amplio, equilibrado y debidamente etiquetado con la siguiente información:

- **Glosa:** gesto o secuencia de gestos asociados a una palabra o frase.
- **Categoría temática:** ámbito semántico del contenido (educación, salud, alimentación, etc.).
- **Categoría gramatical:** ámbito gramatical de la glosa (verbo, nombre, adjetivo, etc.).
- **Idioma:** idioma del país con el que corresponde la glosa.

Sólo clasificaremos **Categoría temática**, las demás etiquetas se usarán como referencia. Las fuentes principales de los vídeos serán la plataforma *Spread the Sign* y el diccionario de vídeos de lengua de signos española (CNSE).

Fase 2: Procesamiento y normalización de datos

Los vídeos recopilados serán sometidos a un proceso de normalización y preprocesamiento, que incluirá:

- Reducción de ruido visual y homogeneización de formato.
- Segmentación temporal de secuencias relevantes.
- Aplicación de técnicas de aumento de datos para compensar el desequilibrio entre clases.

El resultado será un conjunto de datos coherente, balanceado y apto para el entrenamiento de modelos de aprendizaje profundo.

Fase 3: Extracción de características visuales

En esta etapa se empleará un modelo tridimensional de visión artificial (I3D o equivalente) para la extracción de características visuales de cada secuencia. El modelo generará representaciones vectoriales de alta dimensión que describen los movimientos, posiciones y patrones visuales asociados a cada gesto.

Fase 4: Integración del modelo de lenguaje

Las representaciones visuales extraídas serán proyectadas al espacio de entrada de un modelo de lenguaje preentrenado basado en transformadores (por ejemplo, *XLM-RoBERTa* o *mBART*). Esta fase busca aprovechar la capacidad de estos modelos para capturar dependencias temporales y relaciones semánticas entre las secuencias visuales, reforzando la clasificación temática.

Fase 5: Entrenamiento y ajuste del modelo multimodal

Se entrenará el sistema de forma supervisada utilizando las etiquetas temáticas definidas. Durante esta fase se ajustarán los parámetros de las capas de proyección y atención temporal, empleando como función de pérdida la entropía cruzada categórica. Se explorarán estrategias de regularización y optimización para mejorar el rendimiento del modelo en la tarea de clasificación.

Fase 6: Evaluación y análisis de resultados

El modelo final será evaluado mediante métricas estándar de clasificación, incluyendo precisión, *top-k accuracy*, matriz de confusión y análisis de errores. Adicionalmente, se visualizarán las regiones de atención del modelo para interpretar qué partes de los vídeos han contribuido más a la decisión final. Este análisis permitirá comprender mejor cómo los modelos multimodales representan la información visual de las lenguas de signos.

Fase 7: Documentación y conclusiones

Finalmente, se redactará la memoria del trabajo, incluyendo la descripción técnica del sistema, la metodología empleada, los resultados experimentales y las conclusiones derivadas del estudio. Se discutirán también las posibles líneas futuras de investigación y mejoras del modelo.

4.1. Distribución temporal

De acuerdo con las directrices de la Escuela Politécnica Superior de Córdoba, el Trabajo de Fin de Grado debe reflejar al menos 300 horas de dedicación. En base a esta estimación, se propone el siguiente cronograma de trabajo:

Fase	Descripción principal	Duración estimada
Fase 1	Recolección de vídeos, obtención de enlaces, diseño del dataset y etiquetado inicial.	Semanas 1–3 (30 h)
Fase 2	Preprocesamiento de vídeos, segmentación y aumento de datos.	Semanas 4–6 (40 h)
Fase 3	Extracción de características visuales mediante el modelo I3D preentrenado.	Semanas 7–9 (45 h)
Fase 4	Integración del modelo transformer y adaptación de <i>embeddings</i> .	Semanas 10–12 (45 h)
Fase 5	Entrenamiento conjunto y ajuste fino del modelo multimodal.	Semanas 13–17 (60 h)
Fase 6	Evaluación, análisis de resultados y visualización de atención.	Semanas 18–20 (40 h)
Fase 7	Redacción, revisión y presentación del documento final.	Semanas 21–24 (40 h)
Total		300 horas

Table 4.1: Cronograma estimado de trabajo del proyecto.

Aunque se especifica una fase para la redacción y documentación, esta se llevará a cabo de forma continua a lo largo del desarrollo del proyecto. Asimismo, no se incluye una fase específica de formación previa, dado que gran parte de los conocimientos necesarios ya se han adquirido durante el grado y el resto se consolidará durante la ejecución del trabajo.

Capítulo 5

Recursos y requerimientos

Debido a las necesidades de nuestro problema utilizaremos principalmente librerías de aprendizaje automático implementadas en el lenguaje de programación **Python**, algunas de esta librerías son:

- PyTorch
- HuggingFace
- BeautifulSoup
- ScikitLearn
- ...

Para necesidades donde de rendimiento en CPU sean altas optaremos por usar los lenguajes de programación **Rust** o **C/C++**, además que nos permite extender de manera sencilla librerías de Python, haciendo código ejecutable más rápido.

Como entornos de desarrollo se utilizaran IDEs como VSCode, Zed o NeoVim, además para realizar los entrenamientos o ajustes finos de los modelos utilizaremos el entorno en la nube que nos ofrece Google Colab, para computación con GPU.

Tambien se usará el portatil personal del autor, cuyas especificaciones son:

- Chip: Apple M4
- Memoria: 16 GB
- Almacenamiento: 500 GB
- GPU: 10 Nucleos

La redacción del documento sera realiza en **LATEX**, en el entorno de Overleaf.

Todo realizado por el autor Carlos David López Hinojosa, bajo la tutela de Francisco José Madrid Cuevas.

Bibliografía

- Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N. and Zisserman, A. (2020), BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues, *in* 'ECCV'.
- Aronoff, M., Meir, I. and Sandler, W. (2005), 'The paradox of sign language morphology', *Language* **81**(2), 301–344.
URL: <https://doi.org/10.1353/lan.2005.0043>
- Bain, M., Nagrani, A., Varol, G. and Zisserman, A. (2021), Frozen in time: A joint video and image encoder for end-to-end retrieval, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 1728–1738.
URL: <https://arxiv.org/abs/2104.00650>
- Ging, S., Zöller, M., Tsai, Y.-H. H., Nießner, M. and Fidler, S. (2020), Coot: Cooperative hierarchical transformer for video-text representation learning, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.
URL: <https://arxiv.org/abs/2011.00597>
- Momeni, L., Varol, G., Albanie, S., Afouras, T. and Zisserman, A. (2020), Watch, read and lookup: learning to spot signs from multiple supervisors, *in* 'ACCV'.
- Sun, C., Myers, A., Vondrick, C., Murphy, K. and Schmid, C. (2019), Videobert: A joint model for video and language representation learning, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 7464–7473.
URL: <https://arxiv.org/abs/1904.01766>
- Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Feichtenhofer, C., Ryoo, M., Jain, L., Malik, J. and Rohrbach, M. (2021), 'Videoclip: Contrastive pre-training for zero-shot video-text understanding', *arXiv preprint arXiv:2109.14084* .
URL: <https://arxiv.org/abs/2109.14084>
- Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z. and Zhang, D. (2023), 'Gloss-free sign language translation: Improving from visual-language pretraining'.
URL: <https://arxiv.org/abs/2307.14768>