Detección de gestos heterogéneos mediante *few-shot learning*

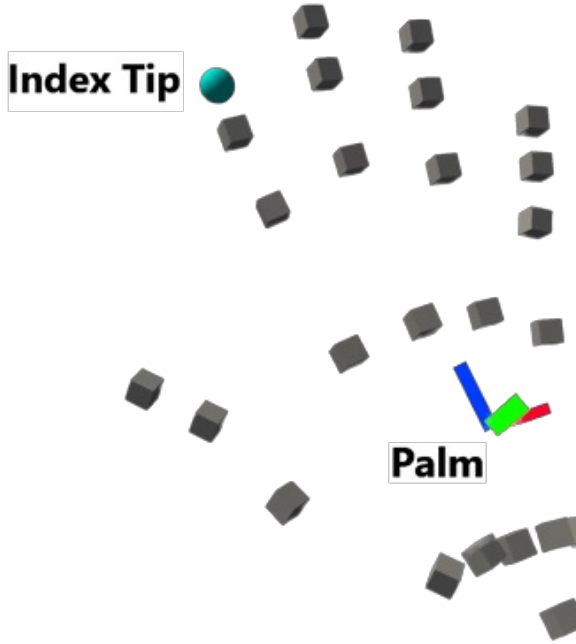
Mario Andreu Villar

Raúl Balanzá García

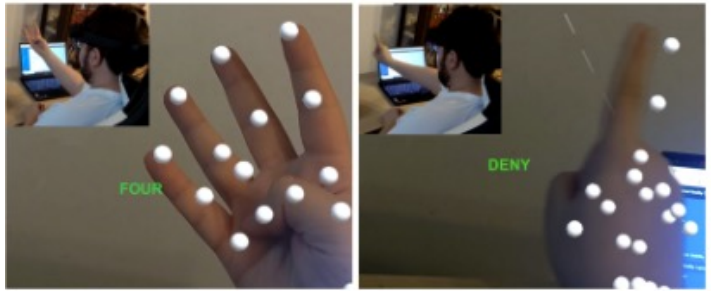
Índice

Introducción

- Tarea compartida del congreso SHREC 2022.
- **Objetivo**: detectar y clasificar gestos a partir de trayectorias 3D de las articulaciones de los dedos.
- Primera aproximación offline para luego adaptarse a un contexto online.
- *Few-shot*: modelos capaces de generalizar a partir de un conjunto de entrenamiento pequeño.



Trabajos relacionados



- Otros trabajos *[1]*: 3 equipos participantes en la tarea + **la propuesta de la organización**.

2ST-GCN

- Convolución en grafo
- Dos flujos de datos
  - Flujo **espacial**
  - Flujo **temporal**

Causal TCN

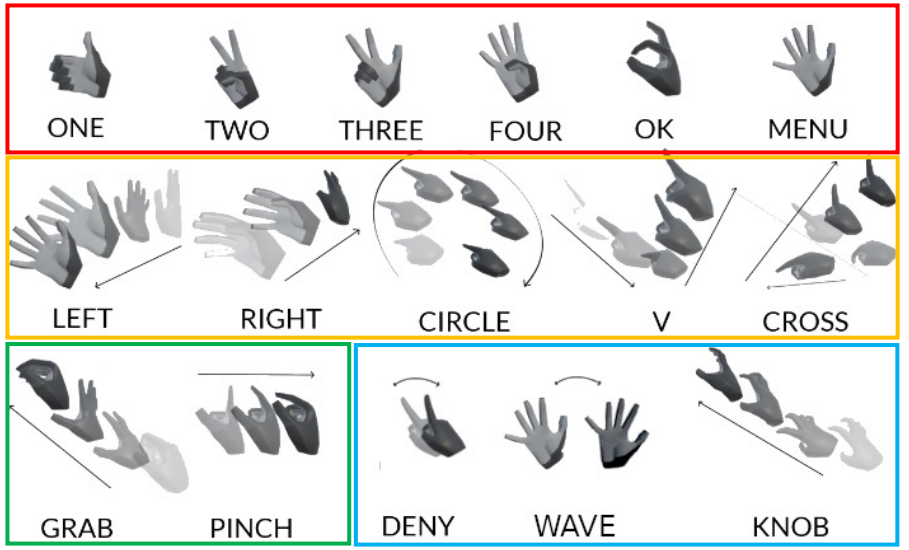
- Procesa datos secuenciales
- Capas convolucionales en cascada
- Modela dependencias temporales a largo plazo

TN-FSM

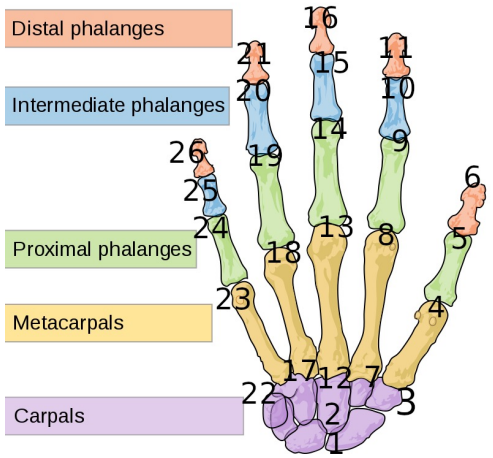
- FSM con 4 estados para detectar si el gesto ha empezado, es la parte del medio o es el final.
- Dos partes: una con Transformers y otra con una FC para clasificar.

Descripción de la tarea

- Conjunto de **288 secuencias** con varios gestos
  - 144 para **train** (*anotadas*) y 144 para **test** (*sin anotar*, descartado)
  - Separadas en 576 gestos individuales para entrenar los modelos
- 16 clases: una por cada gesto
  - Se puede añadir una clase adicional cuyo significado es **no-gesto**
- Tipos de gestos
  - **Estáticos**: una pose que se mantiene fija
  - **Dinámicos**: una única trayectoria de la mano
  - **Dinámicos de grano fino**: articulación de los dedos
  - **Dinámicos-periódicos**: el mismo patrón de movimiento de los dedos se repite
- Los datos son recogidos por un dispositivo HoloLens 2.



Extracción de características



- Formato: datos de los dedos capturados con un dispositivo *Hololens2*.
  - Secuencias temporales en archivos de texto
  - Cada fila: datos de un marco temporal con coordenadas de 26 articulaciones
  - Cada articulación se caracteriza por 3 *floats* (posición x, y, z) → total: **78 floats**

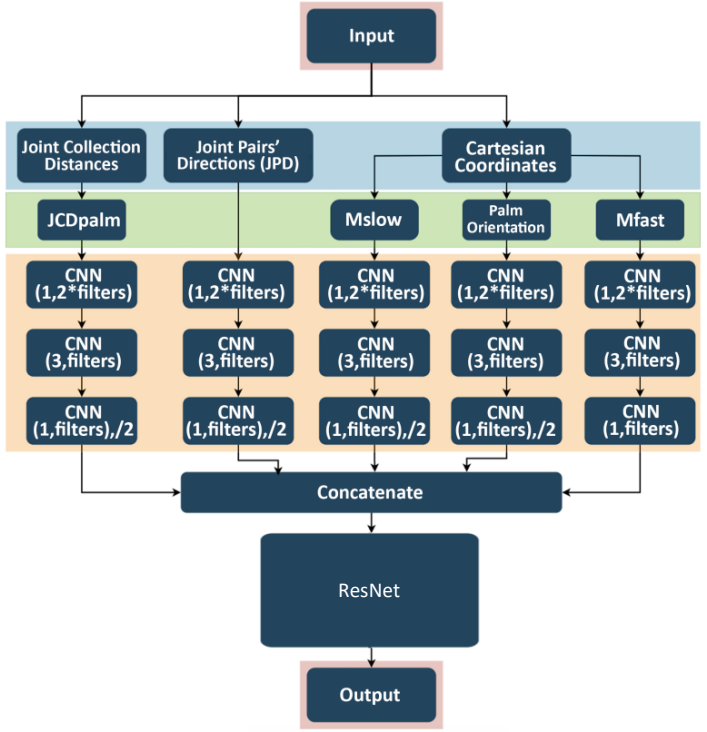
- Modelo de reconocimiento por gesto
  - **Entrada**: posición de las articulaciones en ventana de *n* frames → $(n \times 78)$
  - **Salida**: clase

- Sistema total
  - **Entrada**: secuencia de *m* frames que contienen varios gestos + timestamp → $(m \times (78 + 1))$
  - **Salida**: timestamps de inicio y final de cada gesto + sus etiquetas: *bounding ranges*

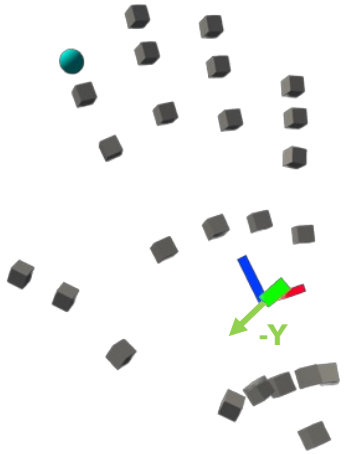
Arquitectura del modelo

- Red **STRONGER** *[2]* (modificación de las DDNet)
  - Propuesta por los organizadores de la tarea
  - *DDNet*: red para clasificar señales direccionales

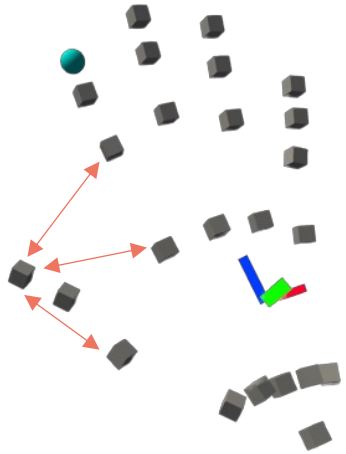
- Nuestra aproximación: versión modificada de STRONGER
  - Utiliza 1D CNNs: rápidas
  - Entrada a la red: 5 inputs → **transformaciones** de los *joints*
    - *Joint Collection of Distances* (JCD)
    - *Joint Pairs' Directions* (JPD)
    - *Palm Orientation* (PO)
    - *Slow global motion* (Mslow)
    - *Fast global motion* (Mfast)
  - Estas entradas forman un *embedding* del gesto



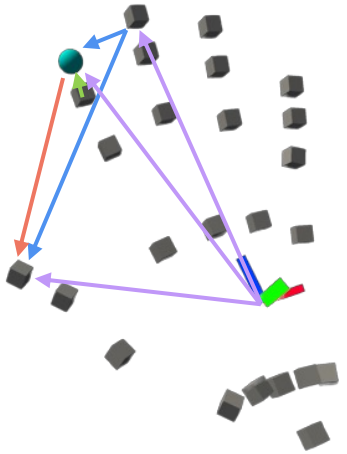
Transformaciones de los *joints*



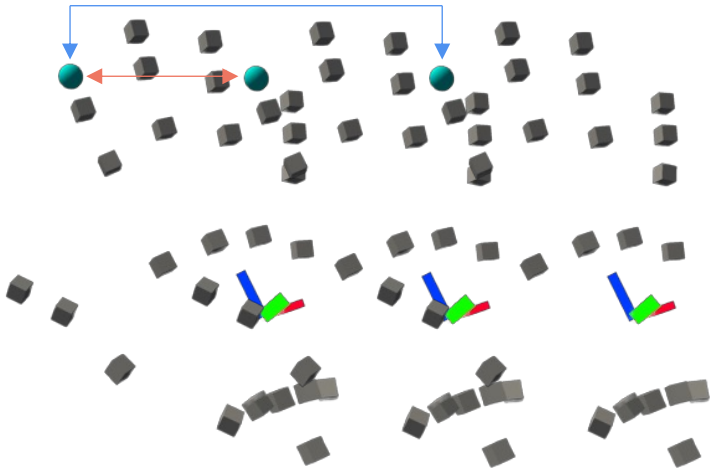
Palm Orientation



Joint Collection of Distances



Joint Pairs' Directions

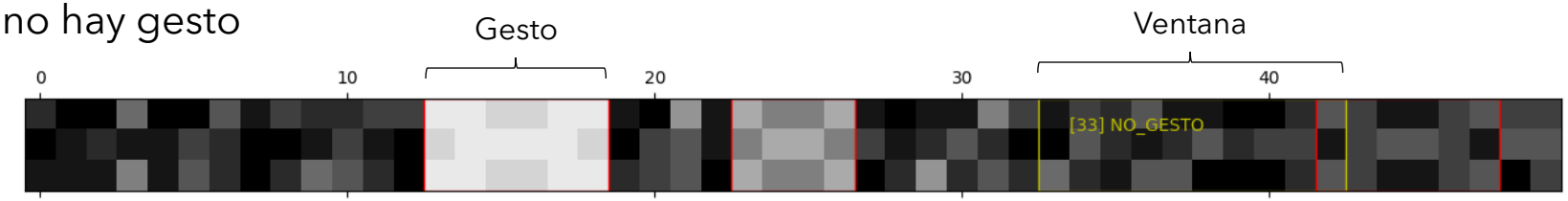


Mfast y Mslow

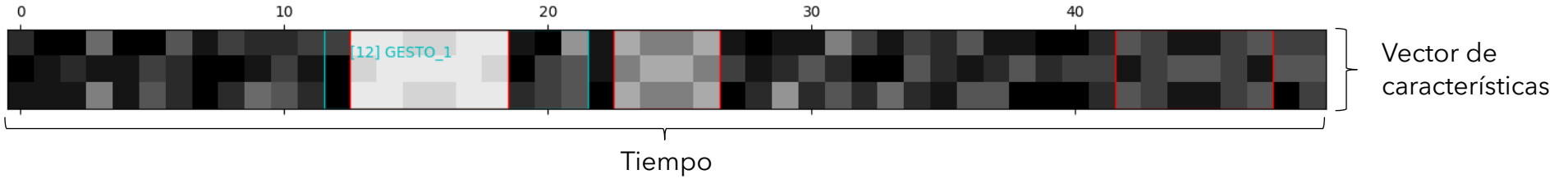
Arquitectura del sistema

- La entrada al sistema es una secuencia de *frames*
  - En la secuencia hay varios gestos (de 3 a 5)
- Se define un tamaño de ventana que se va desplazando
  - En cada desplazamiento → se obtiene con el modelo la probabilidad de haber un gesto

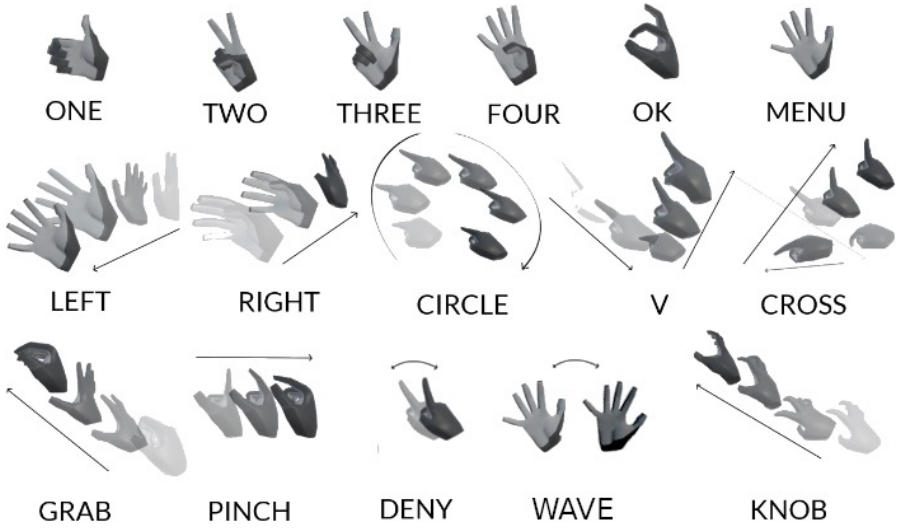
  - **Caso 1**: no hay gesto



  - **Caso 2**: se encuentra gesto → anotar *timestamp* de detección



Diseño experimental



- **Entrenamiento y desarrollo**
  - k-fold cross-validation con $k = 5$ en conjunto de *train*
  - **144 secuencias** con un total de **576 gestos**
  - Mediana de duración de cada gesto: 36 frames
- **Test**
  - Grabación de nuestro propio conjunto de datos con las gafas HoloLens 2
  - **16 secuencias** con un total de **64 gestos**
  - Mediana de duración de cada gesto: 163 frames
- Reducción de la duración de cada gesto
  - Agrupar frames en bloques de tamaño N y hacer su media
  - Hacer saltos de tamaño N y quedarse con 1 de cada N frames → **submuestreo**

Resultados

- **Modelo de reconocimiento de gestos offline**
  - Entrenado durante 100 epochs

Modelo offline	**Accuracy**		
	Desarrollo (k-fold CV)	**Test** (media)	**Test** (submuestreo)
Fully Connected	0.83 ± 0.09	0.46	0.43
ResNet	0.95 ± 0.04	0.28	0.32
STRONGER + ResNet	1.00 ± 0.01	0.85	**0.88**

- **Sistema de reconocimiento de gestos online**
  - Modelo: STRONGER + ResNet
  - Si (confianza < *threshold*) → clasificar en clase **no gesto**

Sistema online	**Desarrollo** (20% train)			**Test** (submuestreo)		
Threshold	*Accuracy*	*Recall*	*F1 score*	*Accuracy*	*Recall*	*F1 score*
0.990	0.80	0.76	**0.78**	0.52	0.58	**0.55**
0.995	0.81	0.64	0.72	0.56	0.51	0.53
0.999	0.79	0.32	0.46	0.64	0.35	0.45

Discusión

- Los **modelos offline** obtienen muy buenos resultados (alta precisión)
  - Todas las variantes: >80% de precisión en el conjunto de desarrollo
  - Nuestra propuesta obtiene máxima precisión en desarrollo y 88% en test
- El **sistema online** aprende de forma aceptable pero tiene margen de mejora
  - Es necesario ajustar el umbral de confianza
  - A veces el ruido viene del propio etiquetado y no del sistema
  - En ocasiones, se predice un gesto con alta confianza cuando en realidad hay ruido (*no-gesto*)

Real	Predicción	Confianza
CROSS	**V**	0.975373
CROSS	**V**	0.953544
CROSS	**V**	0.912978
CROSS	**NO_GESTO**	<0.9
CROSS	**NO_GESTO**	<0.9
CROSS	**NO_GESTO**	<0.9
NO_GESTO	NO_GESTO	<0.9
NO_GESTO	NO_GESTO	<0.9
NO_GESTO	NO_GESTO	<0.9

Real	Predicción	Confianza
NO_GESTO	NO_GESTO	<0.9
NO_GESTO	NO_GESTO	<0.9
NO_GESTO	NO_GESTO	<0.9
NO_GESTO	WAVE	0.991562
NO_GESTO	WAVE	0.991737
NO_GESTO	WAVE	0.992262
WAVE	WAVE	0.993068
WAVE	WAVE	0.993068
WAVE	WAVE	0.993149

Conclusiones

- Las transformaciones a las articulaciones ayudan a generalizar mejor.
- Necesario muy buen modelo offline.
- Se debe anotar correctamente el conjunto de datos de entrenamiento.
- Se debe alcanzar un balance entre precisión y *recall* a través del umbral de confianza.
- La técnica de ventana deslizante obtiene buenos resultados en sistemas online.

Trabajo futuro

- Se deben generar más datos con un mejor proceso de anotación.

- Propuestas para mejorar modelos y sistemas:
  - Probar a utilizar redes recurrentes
  - Utilizar un modelo preentrenado de gestos + proceso de *finetuning*
    - *No hay buenos modelos preentrenados de gestos disponibles (aún)*
  - Entrenar los modelos offline con una clase **no-gesto**
    - Se entrenaría con el ruido de las grabaciones entre cada gesto

Referencias

- **[1]** Caputo, A., Emporio, M., Giachetti, A., Cristani, M., Borghi, G., D'Eusanio, A., ... & von Tycowicz, C. (2022). SHREC 2022 Track on Online Detection of Heterogeneous Gestures. *arXiv preprint arXiv:2207.06706*.

- **[2]** Emporio, M., Caputo, A., & Giachetti, A. (2021). STRONGER: Simple trajectory-based online gesture recognizer. *10.2312/stag.20211481*

¡Gracias por vuestra atención!

Detección de gestos heterogéneos mediante *few-shot learning*

Mario Andreu Villar

Raúl Balanzá García