

Interacción persona-máquina

Unidad 4. Interfaces para reconocimiento de movimiento basadas en sensores y vídeo

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Unidad 4:

- Interacción mediante movimiento del cuerpo, o reconocimiento de gestos: visión por computador (bien mediante cámaras RGB, RGB+D, u otras) para videojuegos, entornos virtuales, rehabilitación, vida asistida por el entorno, teleasistencia, etc.
- Acelerómetros y giroscopios (en teléfonos móviles, mandos de videoconsolas, dispositivos ‘wearable’) para la detección de inclinación, movimientos, etc.
- Guantes hápticos, y sensores de posición (tipo Leap Motion): para entornos VR/AR, simulación del tacto, detección precisa de gestos de los dedos.
- Dispositivos de seguimiento ocular.

1. Mi evolución histórica en visión por computador e interacción persona-entorno
2. Introducción a la visión por computador para la interacción persona-máquina
3. Reconocimiento de movimientos, acciones y actividades mediante visión
4. Dispositivos ponibles, interacción mediante gestos, reconocimiento de actividades

Contenido

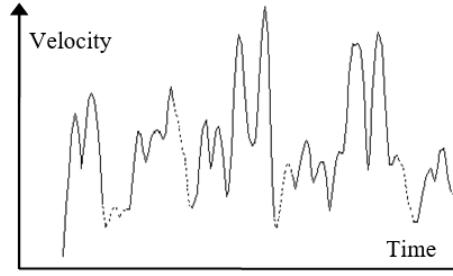
Mi evolución histórica en visión por computador para la interacción persona-entorno

Proyecto fin de carrera

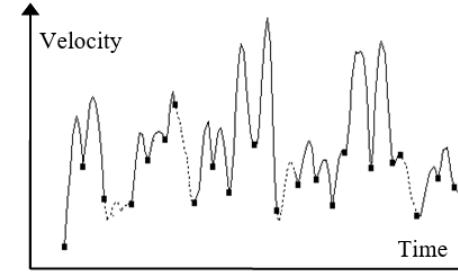
Reconocimiento de escritura en cursiva empleando redes neuronales auto-organizativas

regalo

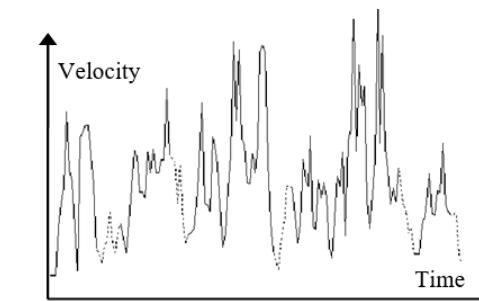
a. Original writing.



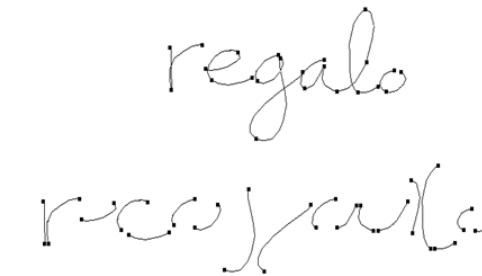
c. Velocity profile after the filtering.



d. Localization of the local minima.



b. Velocity profile before the filtering.

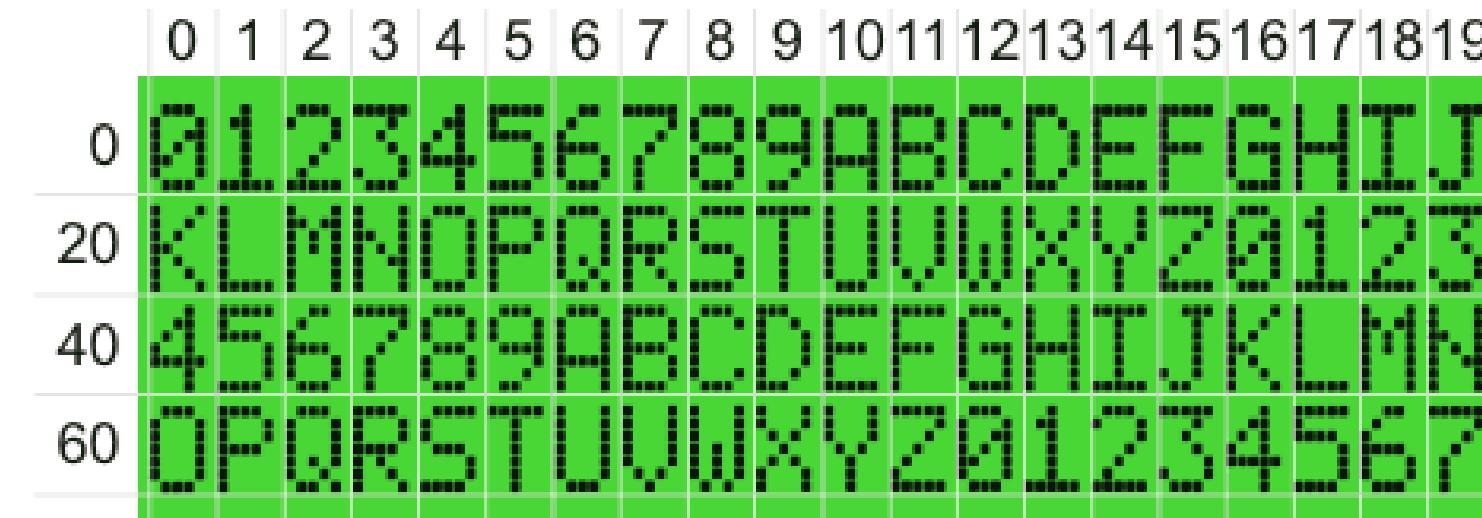


e. Segmentation into strokes.

1994

Proyecto fin de carrera – Compañero de carrera

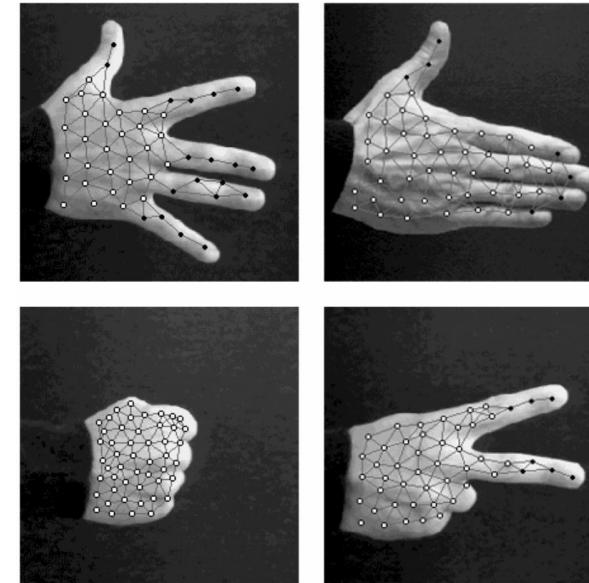
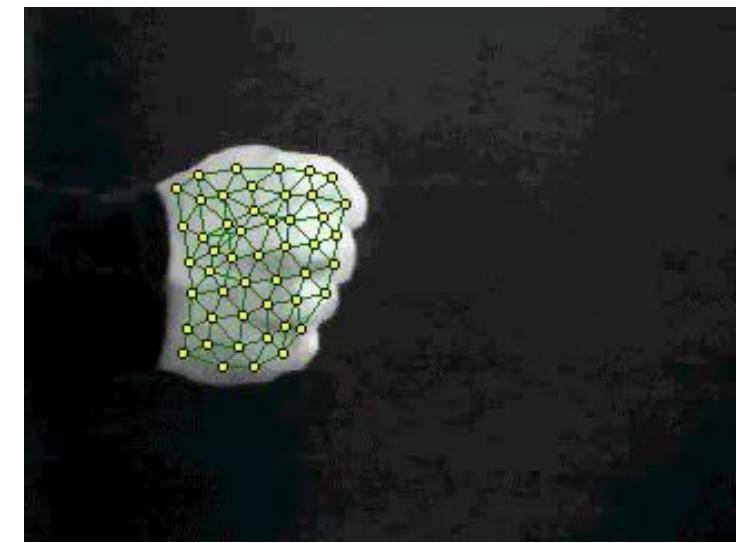
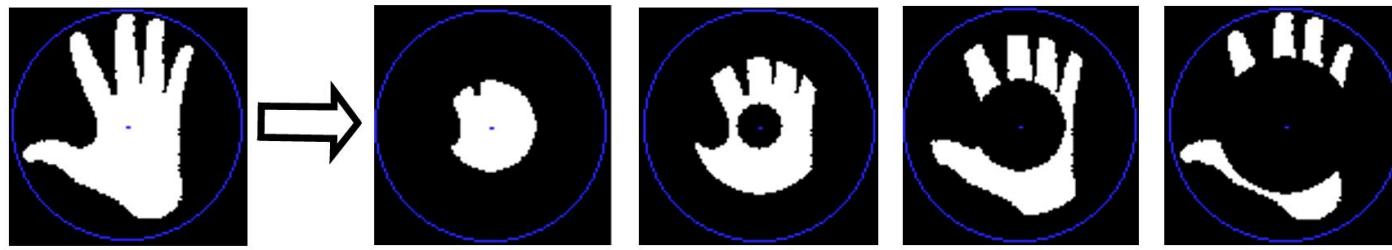
Entrenamiento de perceptrón multicapa para reconocimiento de caracteres ASCII



1994

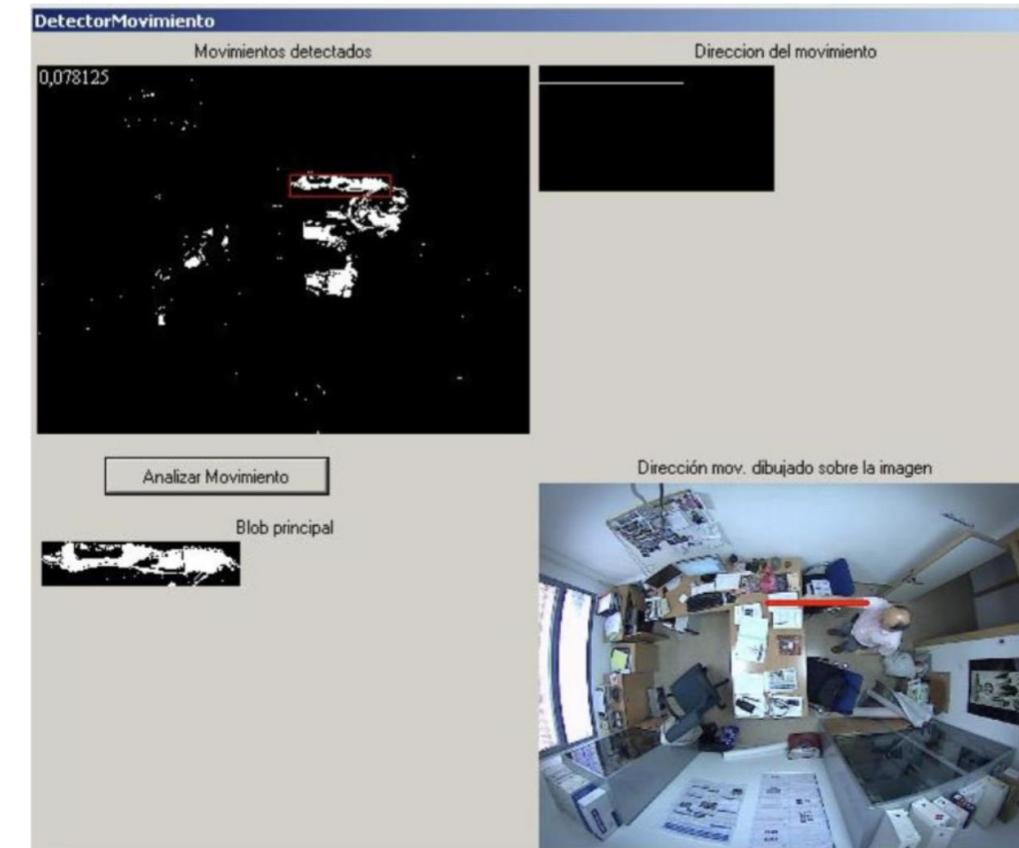
Tesis doctoral - Modelo de representación y procesamiento de movimiento para diseño de arquitecturas de tiempo real especializadas ([enlace](#))

Reconocimiento de gestos de la mano



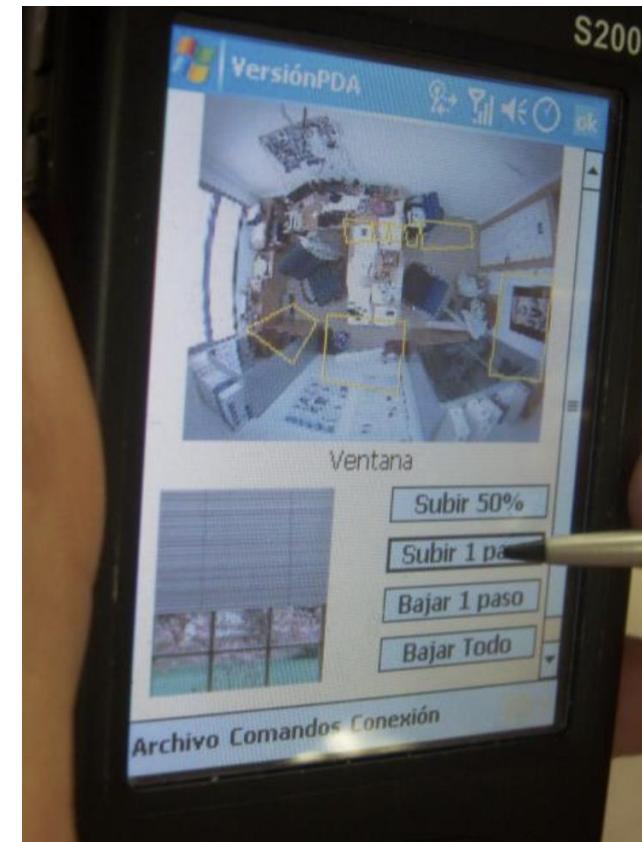
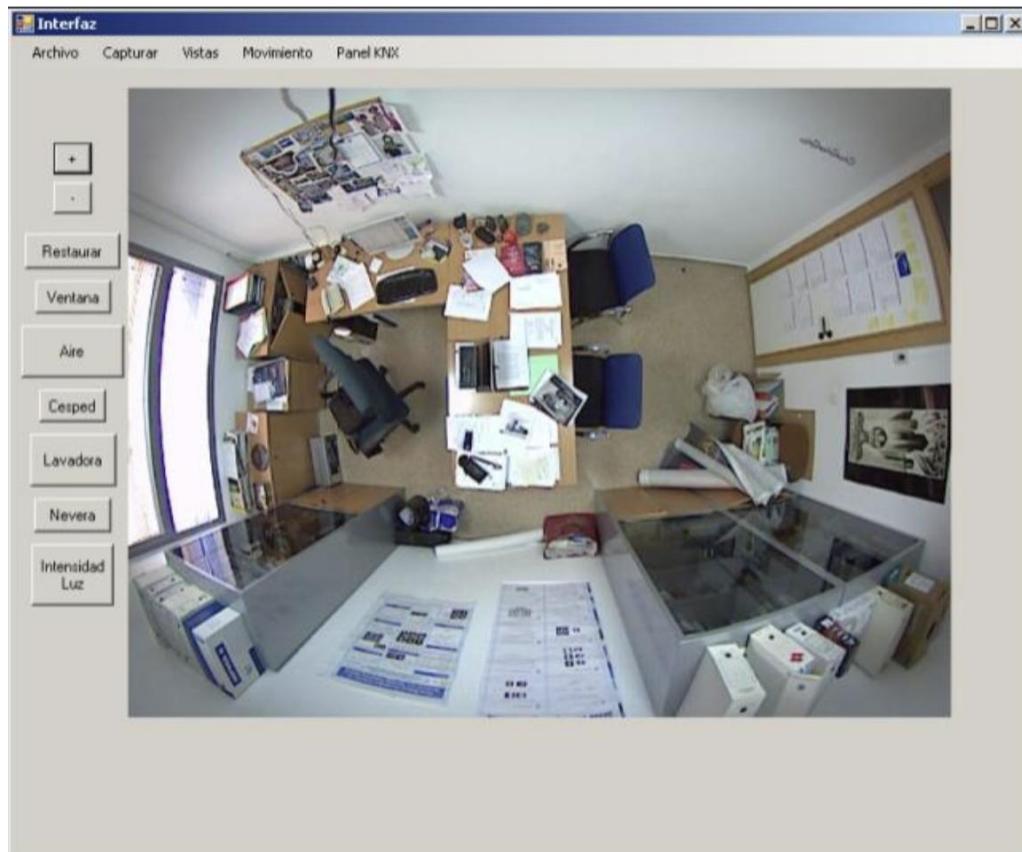
2001

Sistema de interacción con el entorno mediante gestos empleando cámaras omnidireccionales



2007

Sistema de interacción remoto con el entorno a través de imágenes panorámicas y paneles de control virtuales



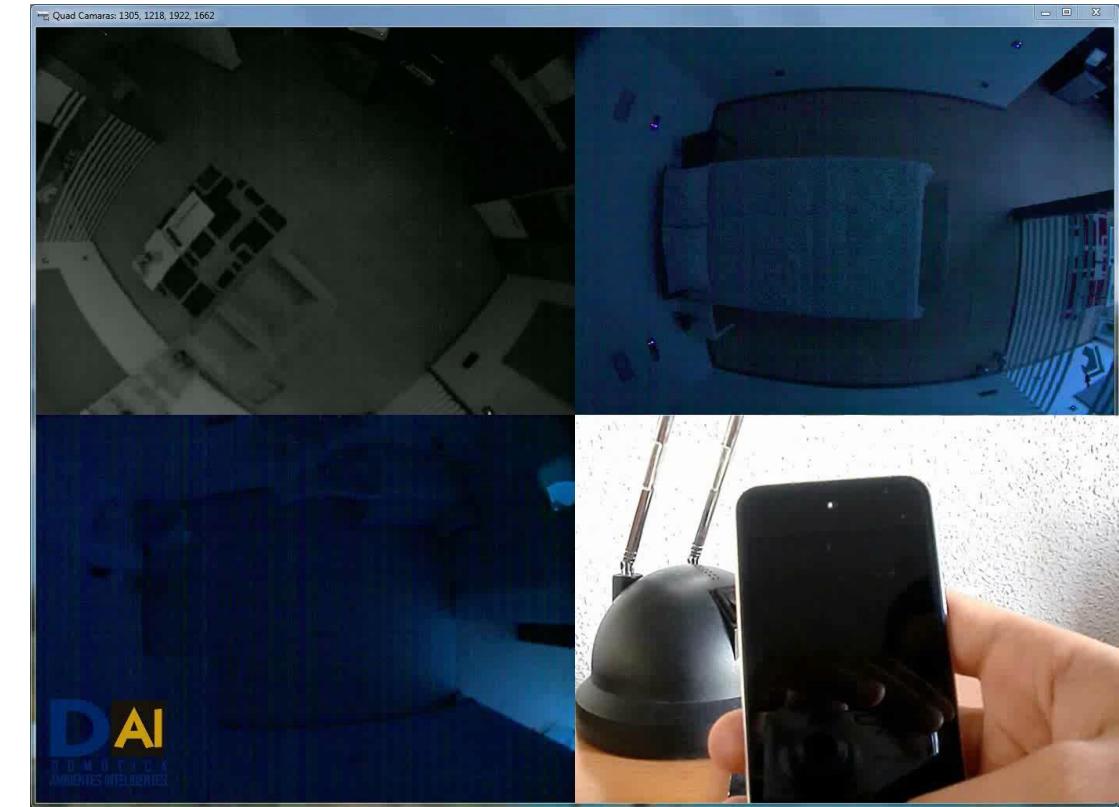
2007

metalTIC – Hogar Digital ([enlace](#))



2010

metaTIC – Hogar Digital



2010

DAI Lab – Domotics and Ambient Intelligence



2011



Desde 2012



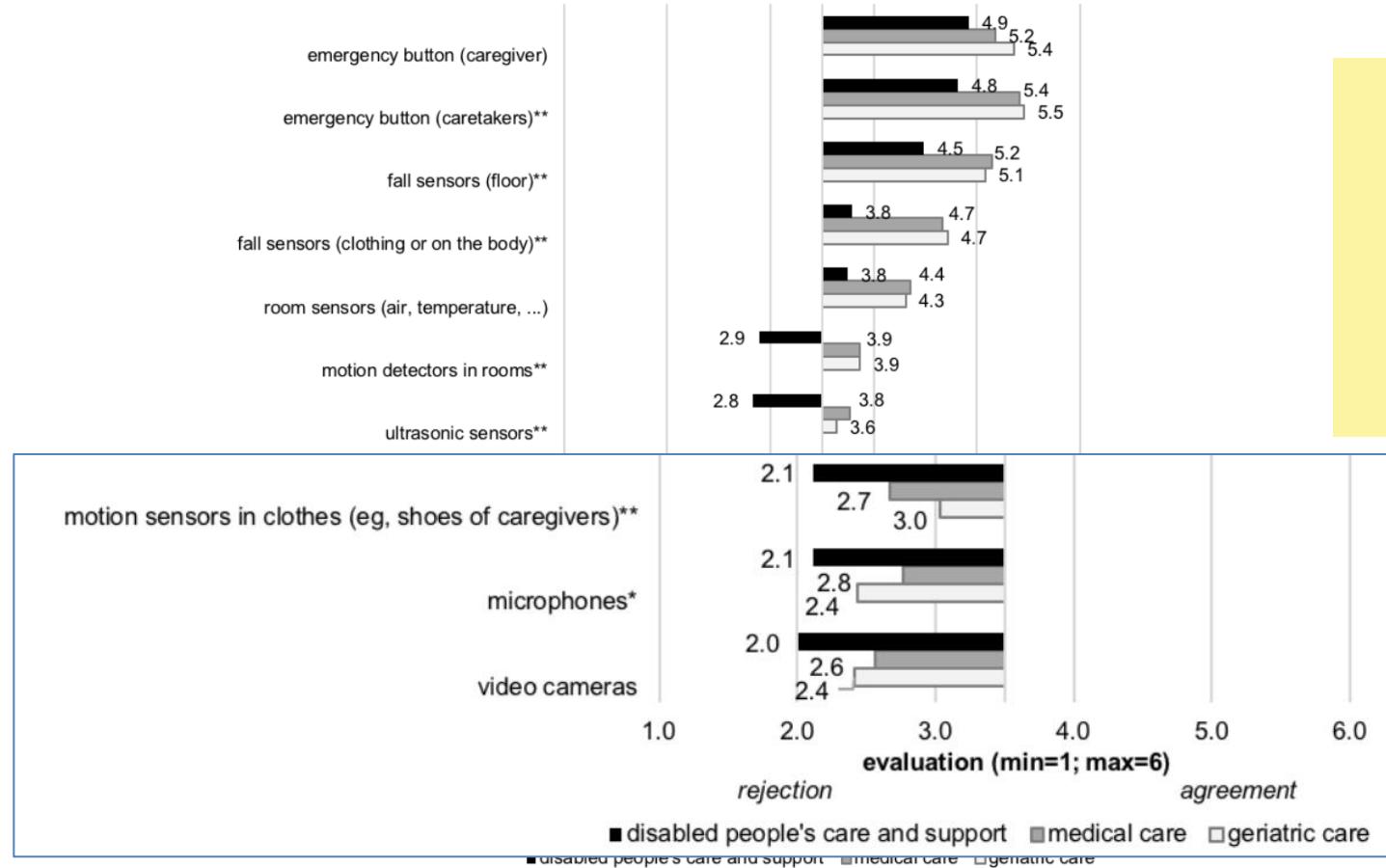
KINECT
for XBOX 360.

The image shows a split-screen view. On the left is a live video feed from a Canary security camera, showing a modern living room with a sofa, a coffee table, and a television. On the right is a summary card with the following text:
OCTOBER 28TH AT 8:15 AM
Activity detected in night mode in Living Room
[Bookmark] [Delete]
No comments

Below the video feed, there's a section titled "Related videos" with two thumbnail images: one from "4:41 PM" and another from "8:25 AM". At the bottom, there are buttons for "Previous Event" and "Next Event".



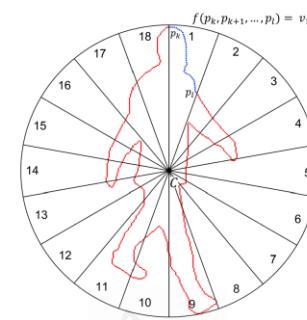
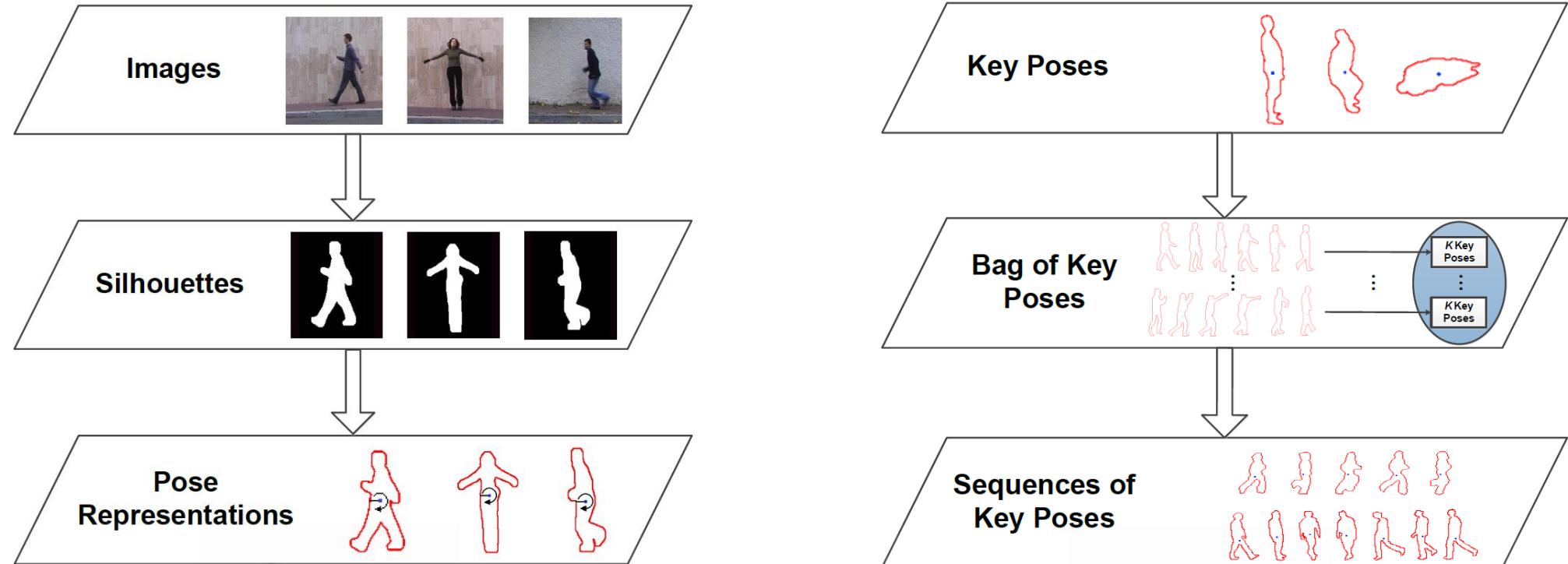




Arning, K., & Ziefle, M. (2015). "Get that Camera Out of My House!" Conjoint Measurement of Preferences for Video-Based Healthcare Monitoring Systems in Private and Public Places. In International Conference on Smart Homes and Health Telematics (pp. 152-164). Springer, Cham.

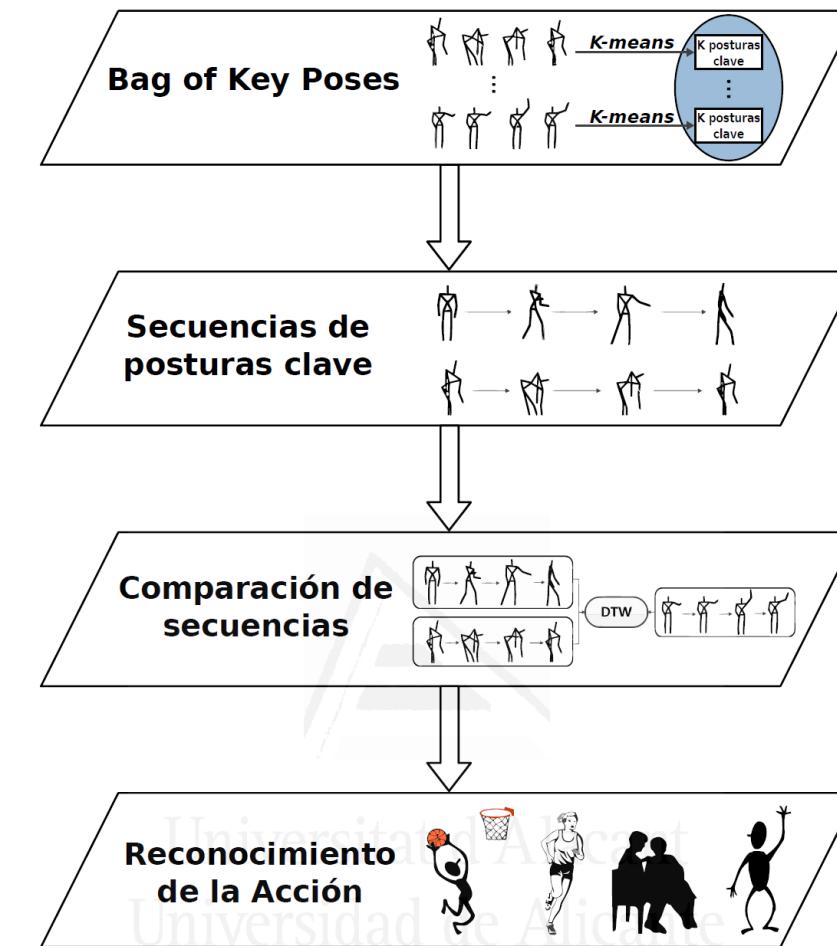
Offermann-van Heek, J., & Ziefle, M. (2018). They Don't Care About Us! Care Personnel's Perspectives on Ambient Assisted Living Technology Usage: Scenario-Based Survey Study. JMIR rehabilitation and assistive technologies, 5(2).

Tesis Alexandros Chaaraoui - Reconocimiento del comportamiento humano para entornos inteligentes ([enlace](#))



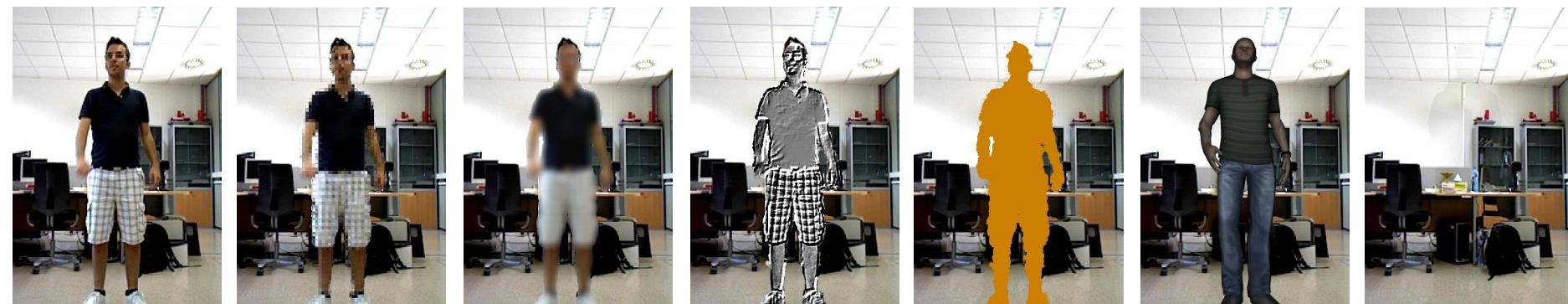
2014

Tesis Alexandros Chaaraoui - Reconocimiento del comportamiento humano para entornos inteligentes ([enlace](#))



2014

Tesis Jose Ramón Padilla – Protección de la privacidad visual basada en el reconocimiento del contexto ([enlace](#))



Original

Pixelado

Emborronado

Relieve

Silueta

Avatar

Invisibilidad

No privacy provided

Real image

Appearance partially protected

Blurred, pixelated or embossed silhouette

Appearance highly protected

Solid silhouette

Appearance totally protected

3D Avatar

High privacy provided

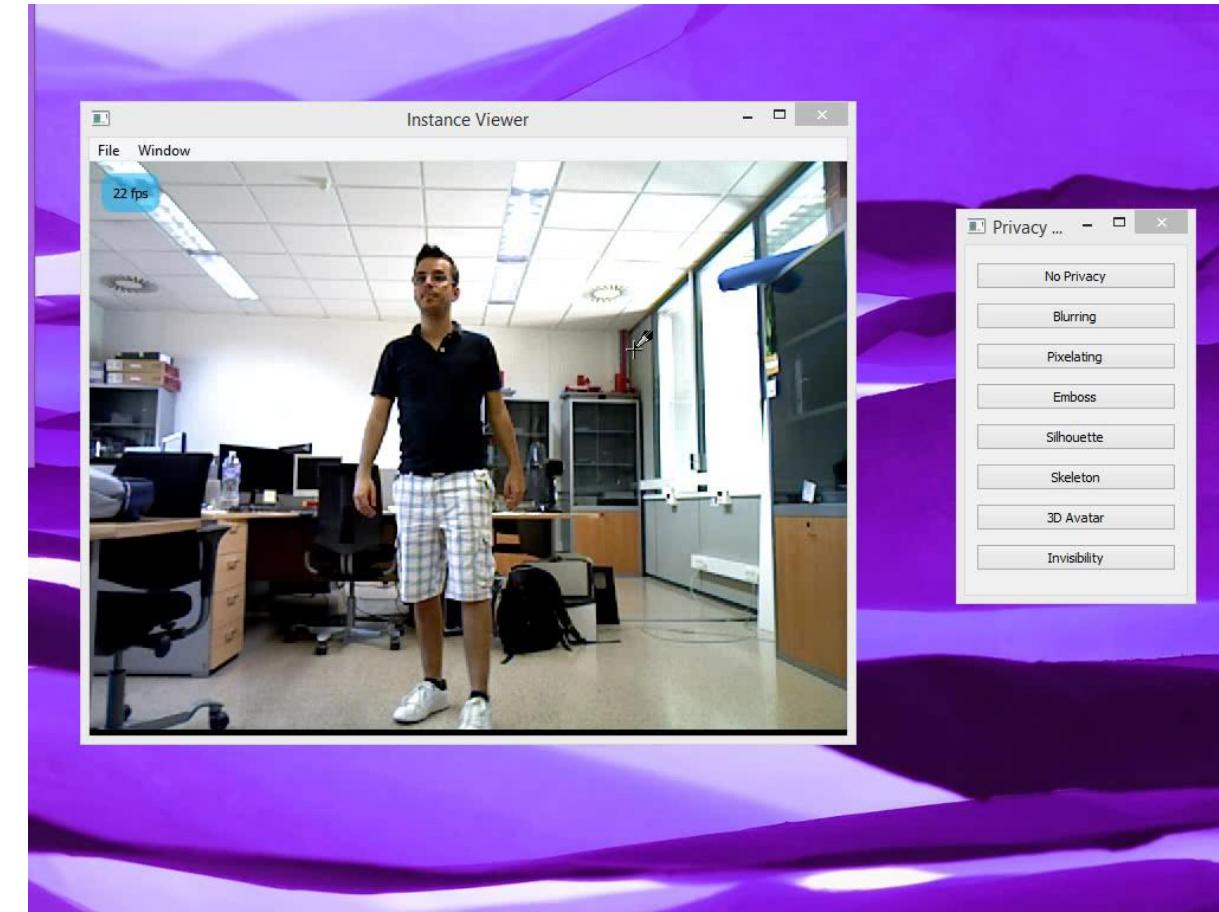
Person completely removed

Low Privacy

High Privacy

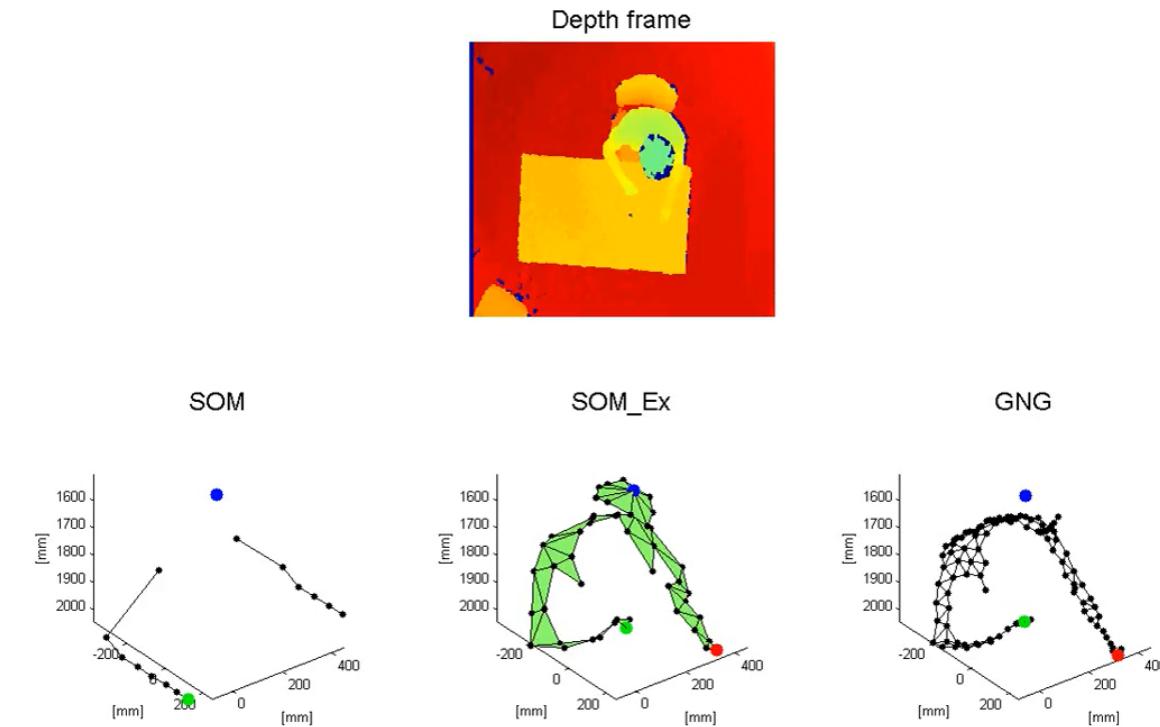
2015

Tesis Jose Ramón Padilla – Protección de la privacidad visual basada en el reconocimiento del contexto ([enlace](#))



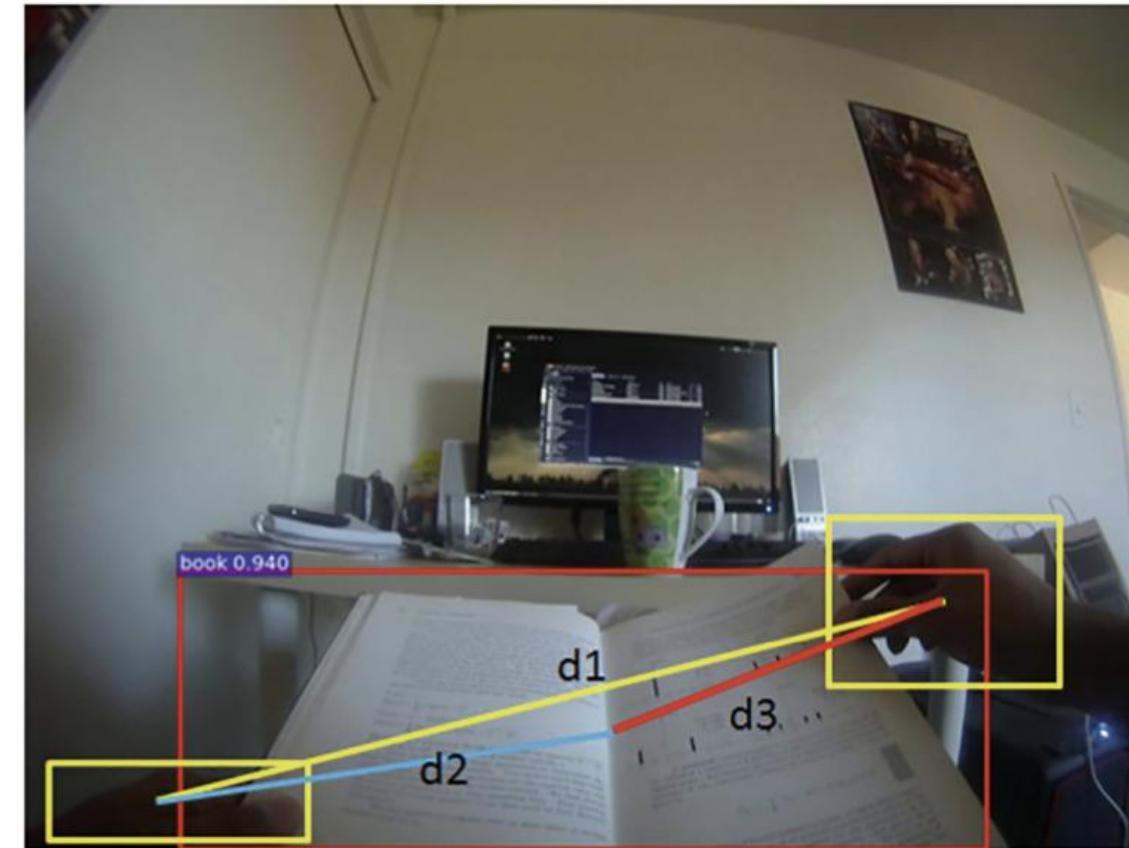
2015

Tesis Samuele Gasparrini – Monitorización de la ingesta empleando Kinect con visión cenital



Gasparrini, S., Cippitelli, E., Gambi, E., Spinsante, S., & Flórez-Revuelta, F. (2015, November). Performance analysis of self-organising neural networks tracking algorithms for intake monitoring using Kinect. In IET International Conference on Technologies for Active and Assisted Living (TechAAL) (pp. 1-6). IET.

Tesis Máster Cuc Nguyen – Reconocimiento de actividades de la vida diaria empleando visión egocéntrica



Nguyen, T. H. C., Nebel, J. C., & Florez-Revuelta, F. (2018, June). Recognition of activities of daily living from egocentric videos using hands detected by a deep convolutional network. In *International Conference Image Analysis and Recognition*(pp. 390-398). Cham: Springer International Publishing.

2018

Protección de la privacidad visual en vídeos adquiridos con cámaras RGB para aplicaciones AAL

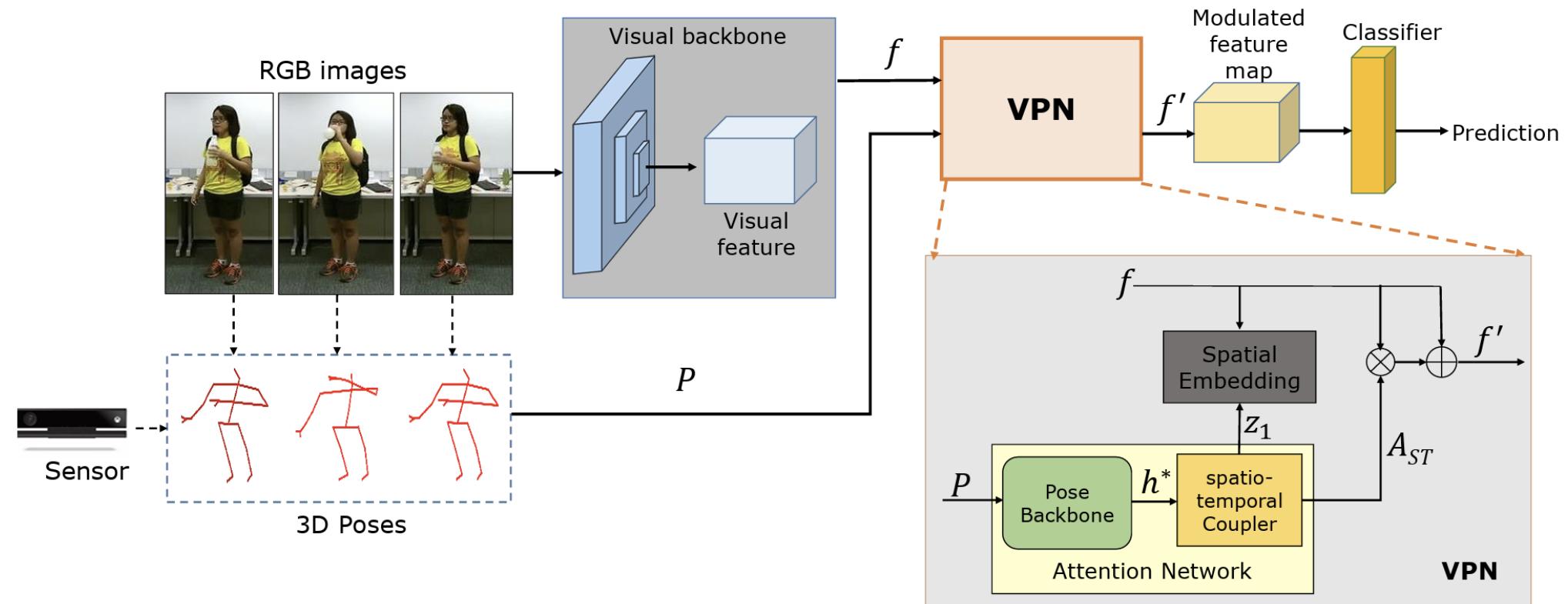


ethicam
smart ethical cameras

Climent-Pérez, P., & Florez-Revuelta, F. (2021). Protection of visual privacy in videos acquired with RGB cameras for active and assisted living applications. *Multimedia Tools and Applications*, 80(15), 23649-23664.

2021

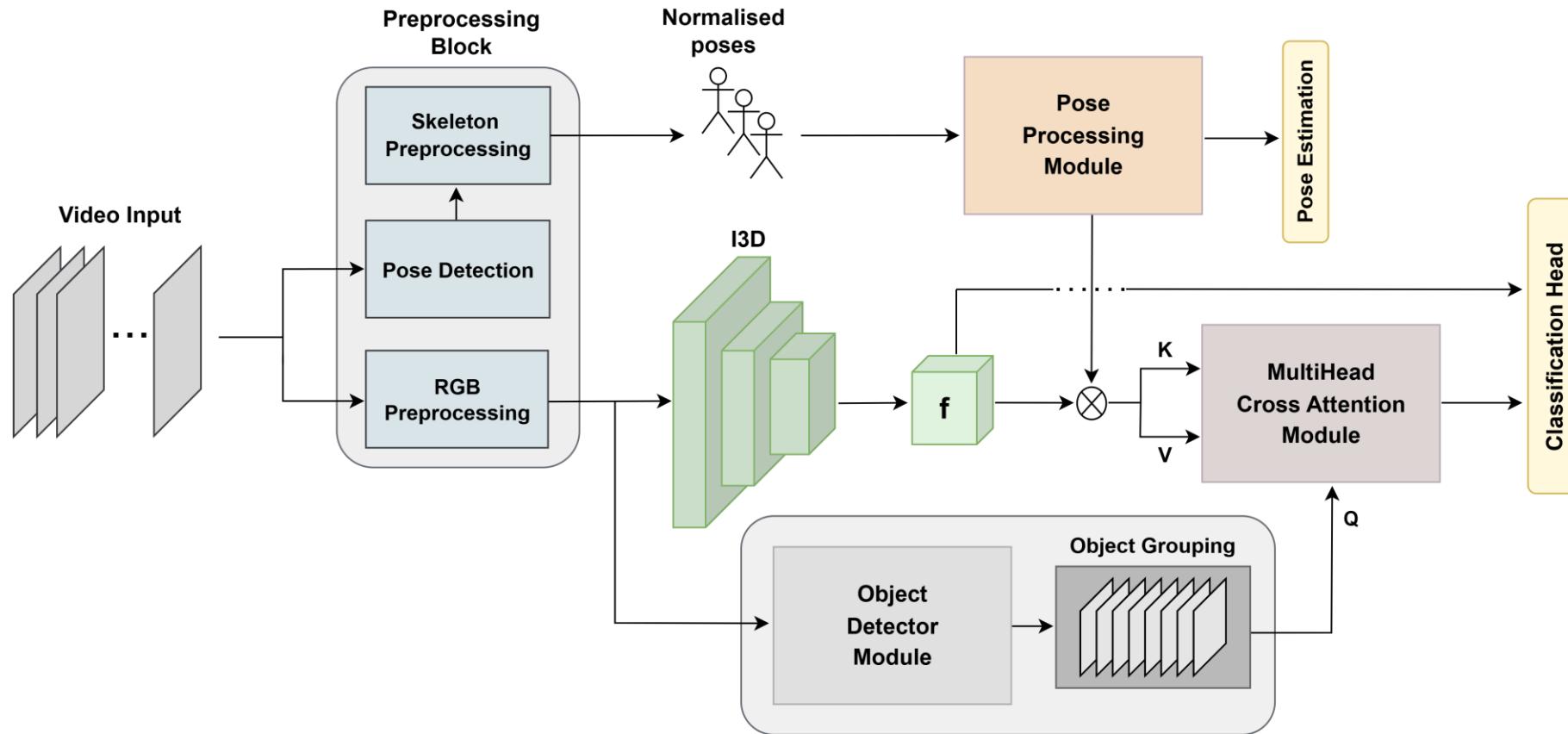
Mejora del reconocimiento de acciones con atención espaciotemporal separable mediante el preprocesamiento alternativo de esqueletos y vídeo



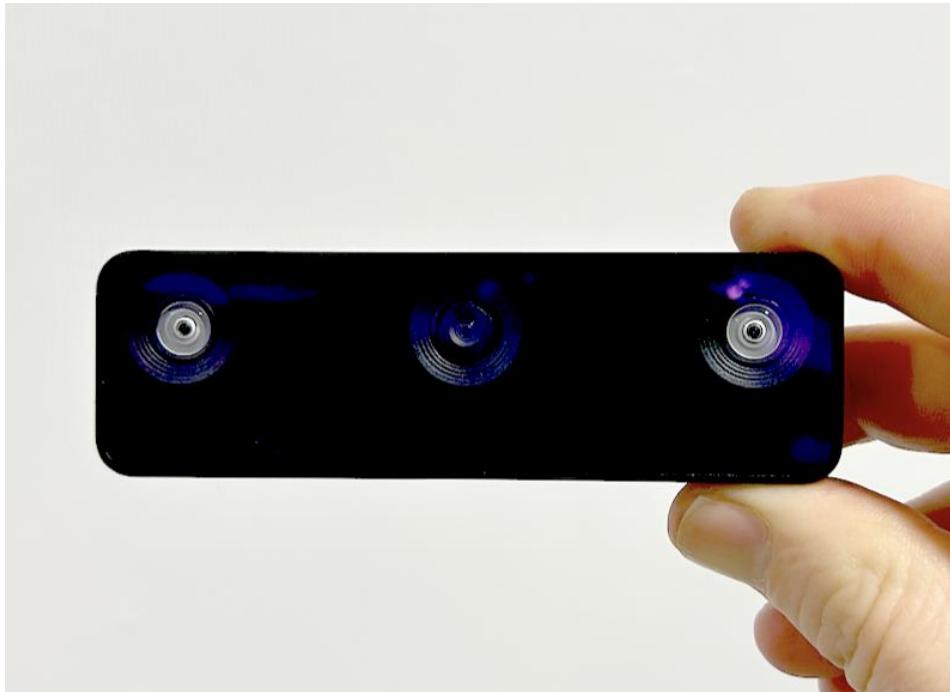
Climent-Perez, P., & Florez-Revuelta, F. (2021). Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing. Sensors, 21(3), 1005.

Original model and image: Das, S., Sharma, S., Dai, R., Bremond, F., & Thonnat, M. (2020, August). VPN: Learning video-pose embedding for activities of daily living. In European Conference on Computer Vision (pp. 72-90). Cham: Springer International Publishing.

Tesis Kooshan Hashemifard – Reconocimiento de actividades cotidianas mediante aprendizaje profundo multimodal: un enfoque basado en vídeo, postura y objetos



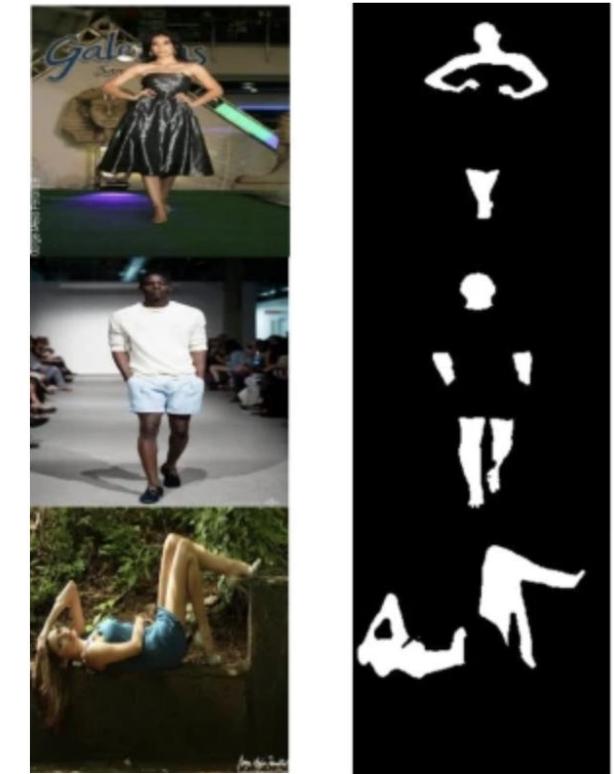
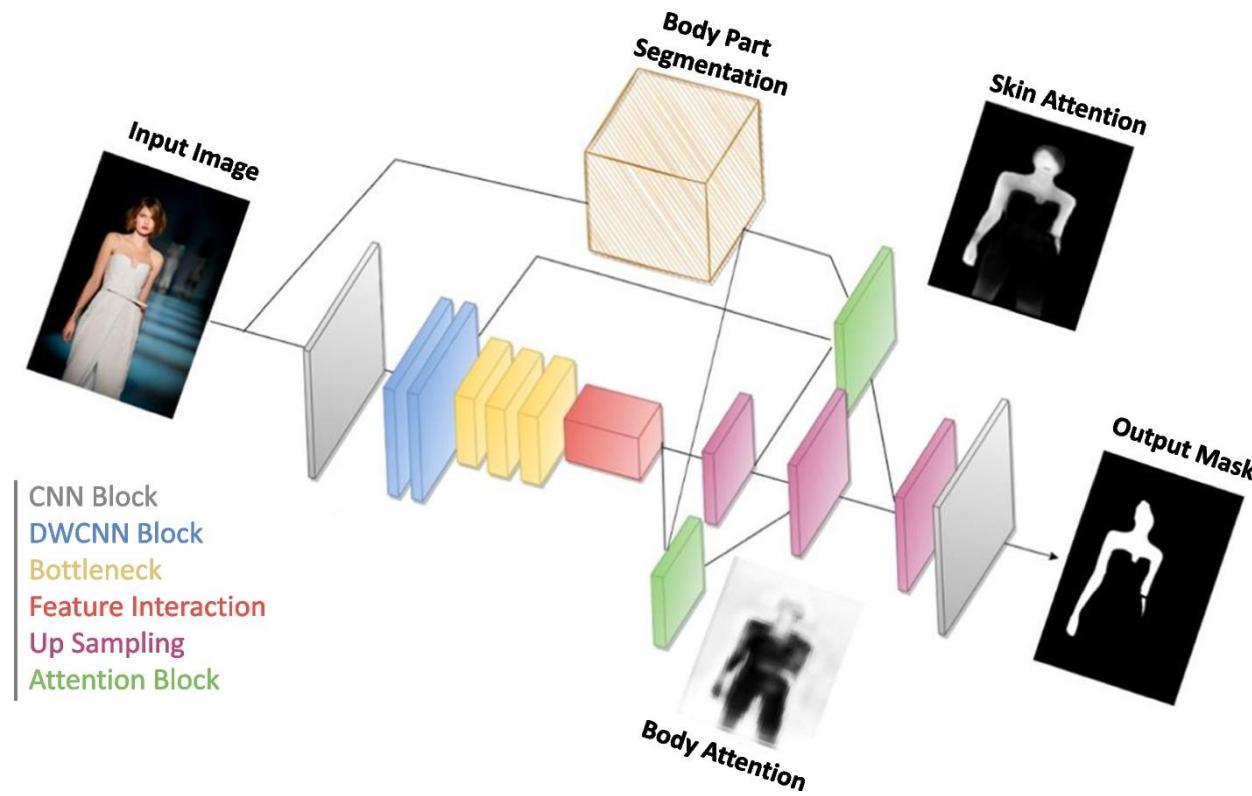
Tesis Kooshan Hashemifard – Detección de caídas empleando una cámara “edge”



Hashemifard, K., Florez-Revuelta, F., & Lacey, G. (2023). A Fallen Person Detector with a Privacy-Preserving Edge-AI Camera. 9th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE, Prague.

2023

Tesis Kooshan Hashemifard – Segmentación de la piel → Estimación del grado de desnudez



Hashemifard, K., Climent-Perez, P., & Florez-Revuelta, F. (2024). Weakly supervised human skin segmentation using guidance attention mechanisms. *Multimedia Tools and Applications*, 83(10), 31177-31194.

2024

Smart Home UA – Laboratorio de gerontotecnología



Universitat d'Alacant
Universidad de Alicante

2024

Monitorización ecológica para la evaluación de la fragilidad en personas mayores

Indice de Fragilidad de Fried

Pérdida de peso no intencionada
Sensación de cansancio
Debilidad (fuerza de agarre)
Velocidad de marcha lenta
Baja actividad física

Batería de Rendimiento Físico Corto

Equilibrio (posición lado a lado)
Equilibrio (posición semi-tándem)
Equilibrio (posición tandem)
Prueba de velocidad al caminar
Prueba de levantarse de una silla

FRAIL

Fatiga en las últimas 4 semanas
Resistencia: dificultad para subir 10 escalones
Dificultad para andar varios cientos de metros sin ayudas
Presencia de 5 o más enfermedades
Pérdida de peso

PRISMA-7

Mayor de 85 años
Hombre/mujer
Dificultad para realizar actividades
Necesidad de ayuda regularmente
Dificultad para salir de casa
Necesidad de ayuda en caso de problemas
Uso de bastón, andador o silla de ruedas

Kihon Checklist

1. ¿Sale solo en autobús o tren?
 2. ¿Va a comprar artículos básicos por sí mismo?
 3. ¿Administra sus depósitos y ahorros en el banco?
 4. ¿Visita a sus amigos ocasionalmente?
 5. ¿Busca consejo en su familia o amigos?
 6. ¿Sube escaleras sin usar barandilla o pared como soporte?
 7. ¿Se levanta normalmente de una silla sin ayuda?
 8. ¿Camina continuamente durante 15 minutos?
 9. ¿Ha sufrido una caída en el último año?
 10. ¿Tiene miedo a caerse mientras camina?
- ...

 Universitat d'Alacant
Universidad de Alicante

 Entornos Inteligentes para un Envejecimiento Activo y Saludable

Sede Electrónica  Webmail  UACloud

Grupos de investigación

Inicio Miembros Proyectos Publicaciones Actividades Mediateca Noticias Jornadas de Gerontecnología 2023

Inteligencia ambiental

Tecnología para envejecimiento activo y saludable

Vida asistida por el entorno

Privacidad y ética

Principales proyectos de investigación e innovación

 **visuAAL**
[visuAAL - Privacy-Aware and Acceptable Video-Based](#)

 **goodbrother**
[GoodBrother - Network on Privacy-Aware Audio- and Video-Based](#)

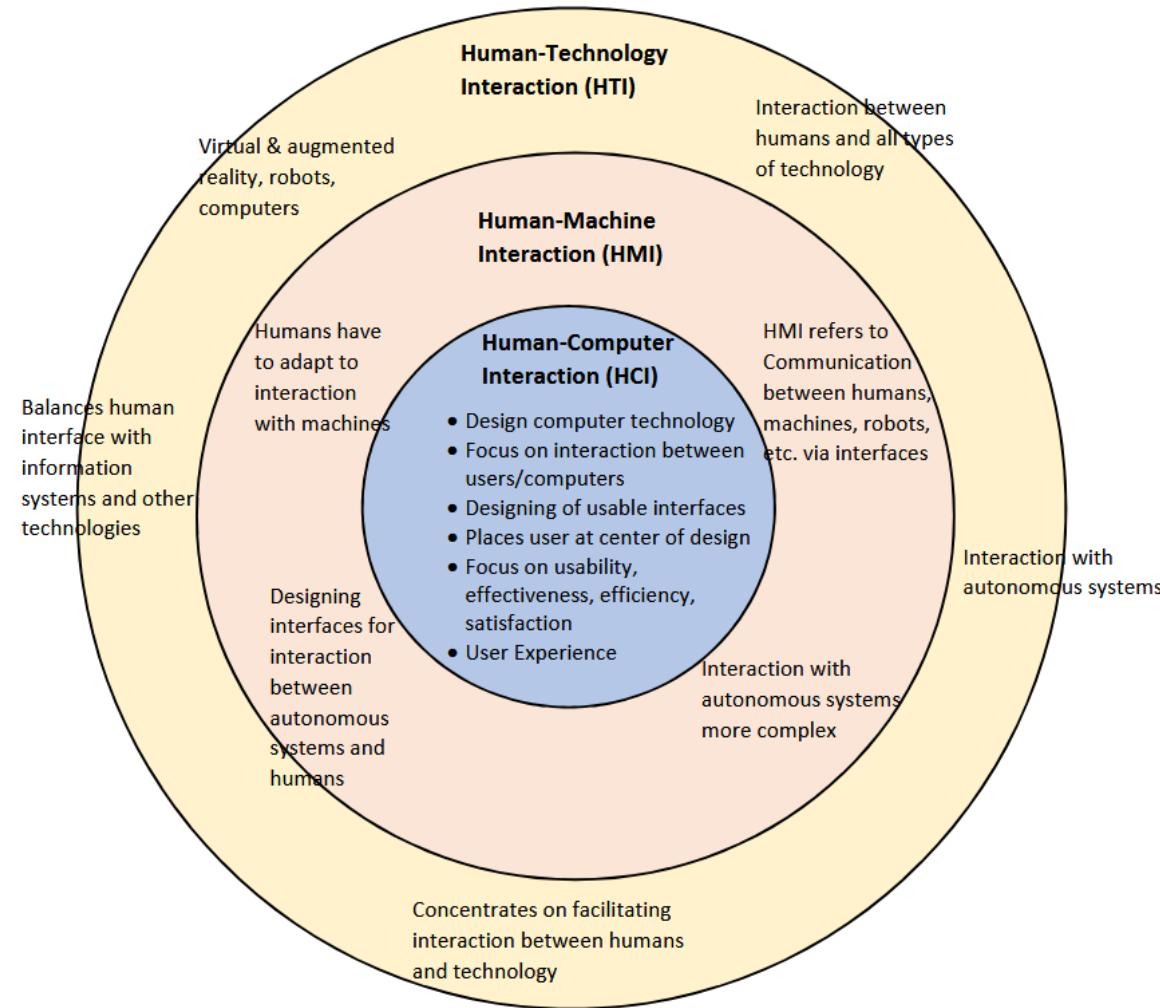
 **paal**
Privacy-Aware and Acceptable Lifelogging services for older and frail people

 **metalTIC**
hogar digital
[metalTIC - Hogar Digital](#)

<https://web.ua.es/es/ami4aha>

Grupo Aml4AHA

Introducción



Interacción persona-tecnología

La interacción debe ser natural, segura y centrada en la persona

Natural:

- Natural: Interacción intuitiva, fluida y coherente con el comportamiento humano.
- Segura: Protección de datos, privacidad, fiabilidad y control del usuario.
- Centrada en la persona: Diseño inclusivo, accesible y adaptado a las necesidades reales

“El objetivo no es que las personas se adapten a la tecnología, sino que la tecnología se adapte a las personas”

Interacción natural

Adquisición

- Sensores en el entorno
- Sensores sobre la persona (ponibles, o “wearables”)

Procesamiento

- Sincronización de datos de múltiples fuentes (si las hay)
- Selección de características (para facilitar el aprendizaje o para protección de datos)

Razonamiento

- Inteligencia Artificial (AI) > Machine Learning (ML) > Deep Learning (DL)
- Técnicas supervisadas (clasificación, regresión) o no supervisadas (ej. clustering)

Etapas

Cualquier mecanismo que convierte una magnitud física en una señal eléctrica

Analógico: Magnitud grande \Rightarrow Señal grande

- Termistor: convierte el calor en una señal eléctrica análoga
- Fotorreceptor: detecta la incidencia de fotones
- Sensor de presión: convierte la fuerza ejercida en una corriente eléctrica

Digitales: la entrada y/o la salida son binarias (1 o más bits)

¿Es una cámara digital un sensor? Bueno, de hecho, una matriz de ellos

Los sensores se pueden instalar según sea el caso:

- En el entorno: hablamos de Interacción Persona-Entorno (Entornos/Ambientes Inteligentes)
 - Por ejemplo, sensores en domótica: infrarrojo pasivo (PIR), apertura/cierre puertas y ventanas, presión
 - Cámaras, micrófonos (incluyendo altavoces inteligentes, Google Assistant, Alexa, etc.)
 - Sensores ‘conectados’ (IoT) del hogar.
- En la persona (ponibles, aka ‘wearables’)
 - IMU – Inertial Measurement Unit
 - Brazaletes inteligentes, datos de salud, cardíacos (pulso, presión), temperatura corporal, oximetría (IoHT)
 - Monitores de glucosa conectados u otros (interacción Paciente-Máquina).

Introducción a la visión por computador para la interacción persona-máquina

Introducción a la visión por computador para la interacción persona-máquina

Percepción

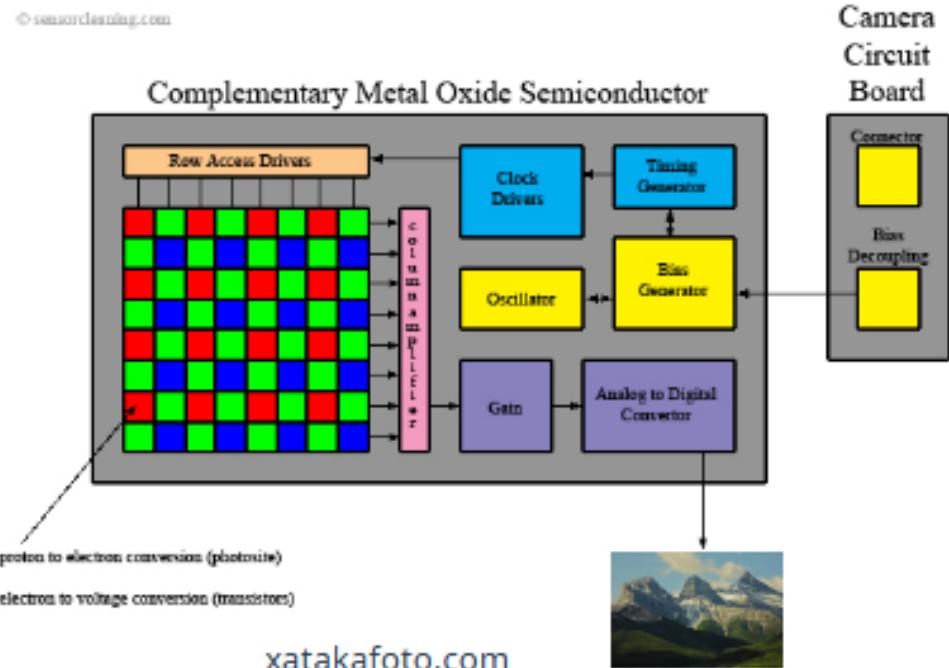
Se trata de una matriz de sensores

- Varias tecnologías: CMOS (mucho más barato, común) o CCD (más caras, más raras)

Detectan:

- Intensidad lumínica (brillo, luminosidad): tonos de blanco y negro
- Rojo, Verde, Azul (RGB); bio-inspirados en el ojo humano

Algunas detectan otros rangos (fuera del espectro de luz visible)



Cámaras como matrices de sensores

Aportan información “muy rica”

- Aunque es muy difícil de extraer de forma automática mediante algoritmos

Aunque si se puede extraer esa información ...

- ... una sola cámara se puede usar para muchas aplicaciones
- ... se pueden tener muchos servicios con una sola instalación

Esto está conectado con el concepto de “Software as a Service” (SaaS)

- Muchos algoritmos diferentes trabajando, por ejemplo, desde la misma fuente de vídeo

¿Por qué usar cámaras?

Cámaras RGB

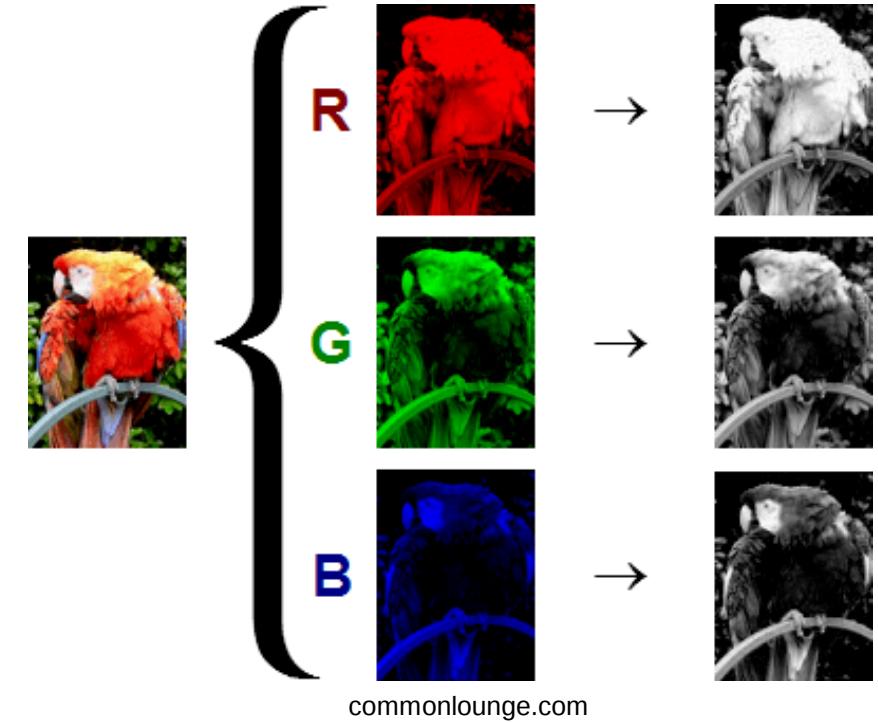
Capturan imágenes en tres canales de color:
Rojo (R), **Verde (G)** y **Azul (B)**.

Son las cámaras más comunes: teléfonos móviles, webcams, cámaras de seguridad, etc.

Cada píxel registra la intensidad de luz en cada canal (mezcla → color percibido).

Permiten extraer información bidimensional (color, forma, textura, movimiento).

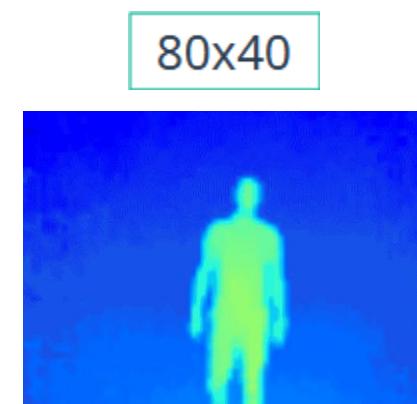
No proporcionan información de profundidad o distancia.



RGB-D – Cámaras de tiempo de vuelo (ToF)

Emiten un haz láser que “viaja” por el aire hasta alcanzar un objeto, luego rebota y se calcula el tiempo que tarda en regresar al emisor.

Suelen ser caras (precio por píxel), por lo que se comercializan en resoluciones bajas más económicas.



<https://www.terabee.com/depth-sensors-precision-personal-privacy>



Tipos de cámaras

RGB-D – Cámaras de profundidad

Se proyecta un patrón infrarrojo no visible.

Dado que el patrón es conocido, cualquier lugar en el que dicho patrón presente cambios implica que hay un obstáculo (persona, objeto).



Tipos de cámaras

RGB-D – Cámaras de profundidad

Se proyecta un patrón infrarrojo no visible.

Dado que el patrón es conocido, cualquier lugar en el que dicho patrón presente cambios implica que hay un obstáculo (persona, objeto).

Existen varias marcas y modelos: Microsoft Kinect v1, v2; ASUS Xtion, Xtion Pro; Orbbec Astra, ...



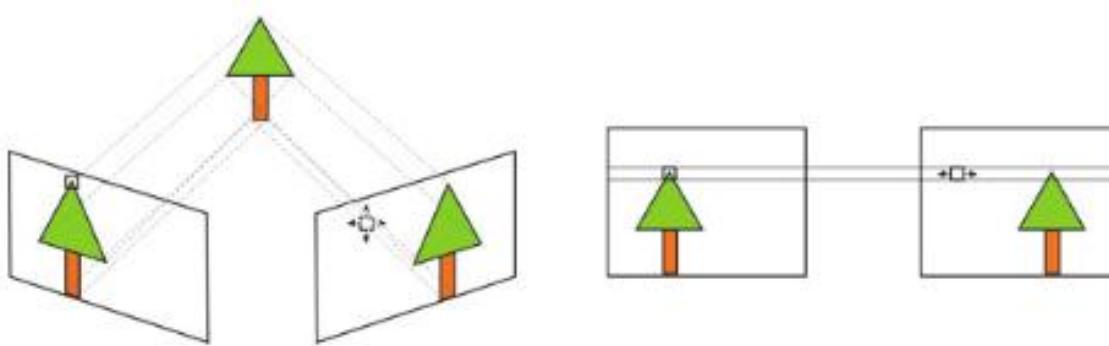
Tipos de cámaras

Cámaras estéreo

Se usan dos cámaras separadas: la diferencia entre sus imágenes permite calcular la distancia mediante triangulación

Técnica bio-inspirada en la visión con ‘dos sensores’:

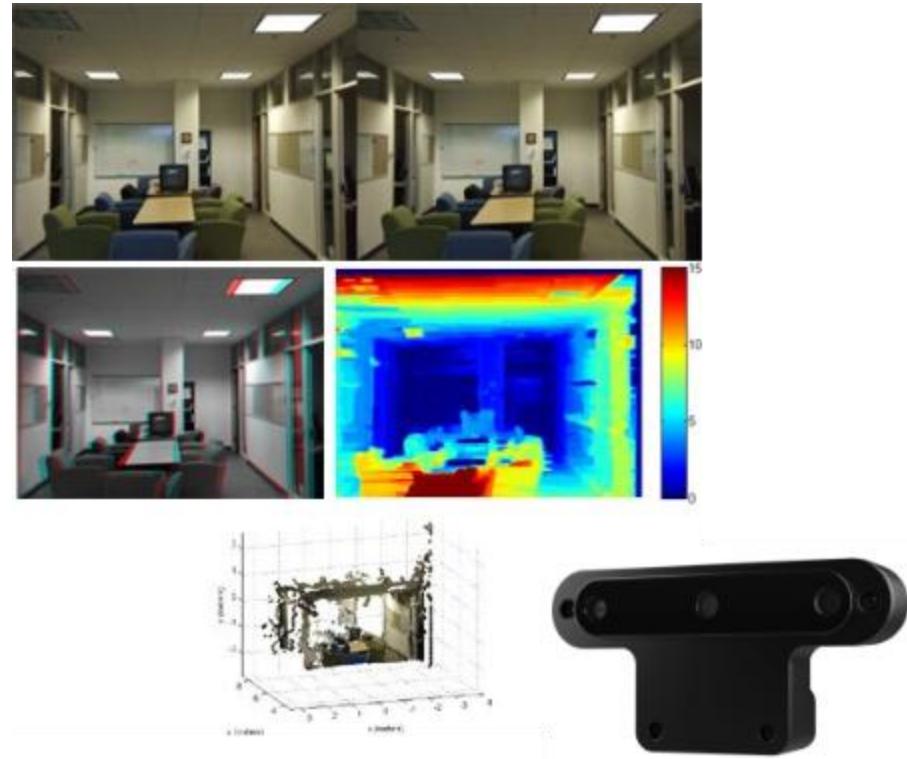
- Basado en la forma en que los animales con dos ojos calculan la distancia: se conoce la línea base (distancia entre cámaras) y luego se utiliza trigonometría para estimar los “mapas de disparidad”.
- Este método usa la información del color, por lo que las superficies “planas” (sin textura) son difíciles de emparejar.



https://www.researchgate.net/figure/Stereo-vision-disparity_fig1_236455177

Tipos de cámaras

Visión estéreo - ejemplos



Tipos de cámaras

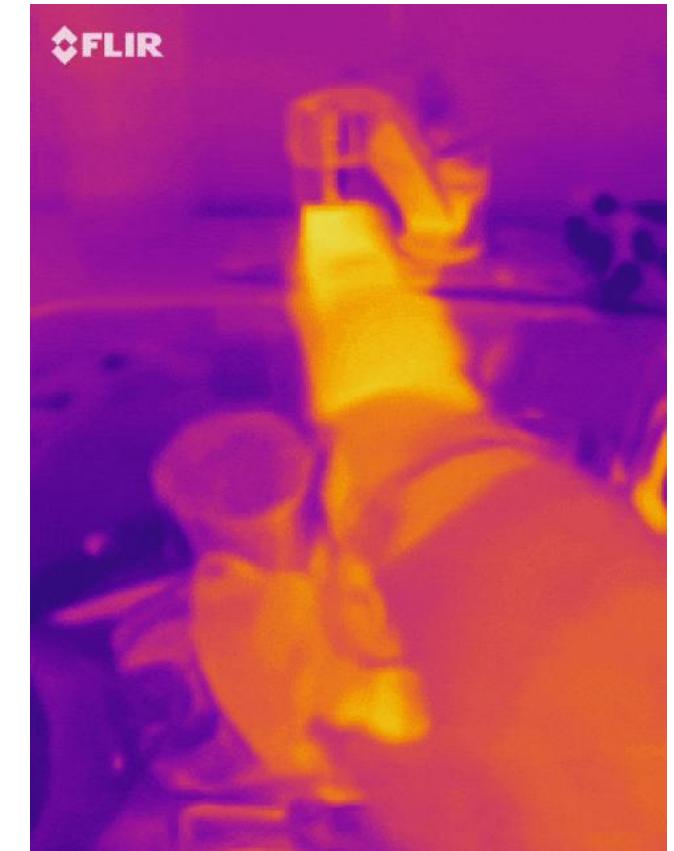
Cámaras de temperatura/calor (cámaras infrarrojas)

Este tipo de sensores recibe la radiación infrarroja emitida por los objetos

No emiten ninguna señal ni patrón de infrarrojos.

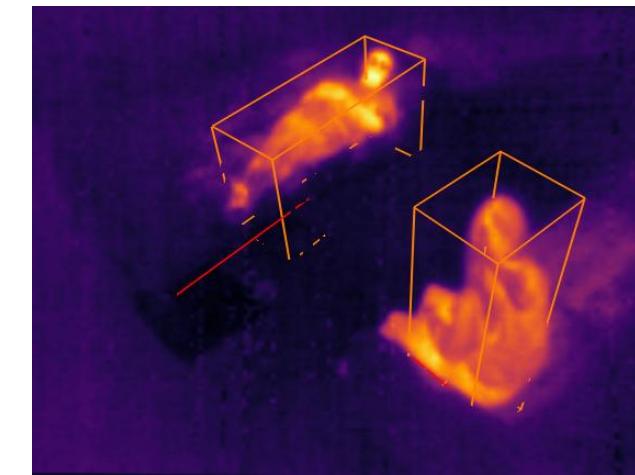
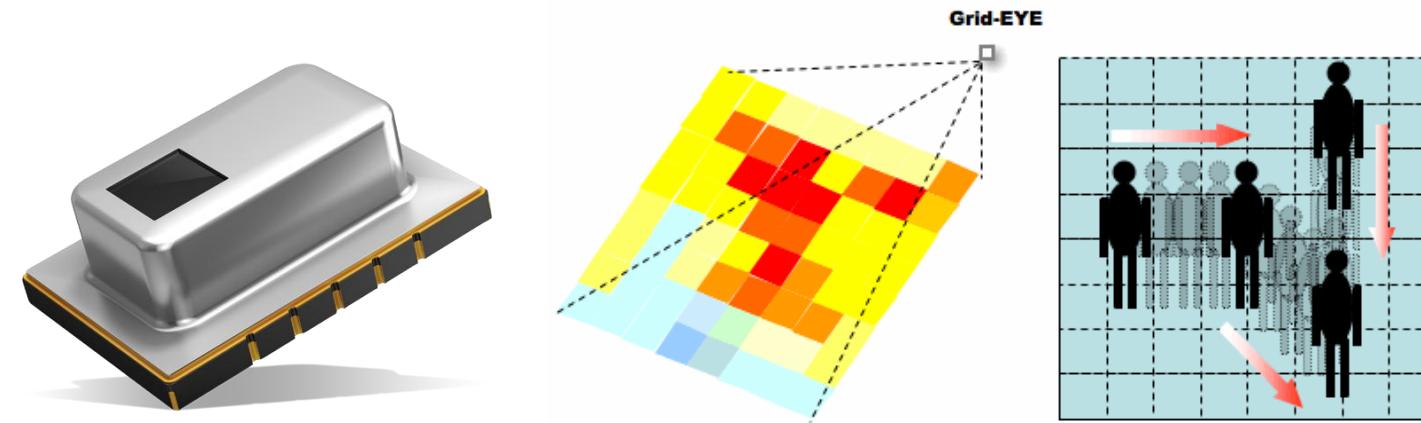
Sin embargo, reaccionan lentamente a los cambios (baja frecuencia de actualización): los objetos dejan un “rastro de calor” al retirarse.

Por ejemplo: el calor en el colchón al levantarse de la cama, o al mover una taza con una bebida caliente/fría (calor residual en las superficies), etc.



https://www.reddit.com/r/oddlysatisfying/comments/fdavxc/hot_and_cold_tap_water_as_seen_through_a_thermal

Tipos de cámaras



Tipos de cámaras

Cámaras portables o egocéntricas

Menor problema de occlusiones causadas por el tronco de la persona

Mayor dificultad de procesamiento

Problemas de “motion blur”



<https://epic-kitchens.github.io/2025>

Tipos de cámaras

Cámaras portables o egocéntricas



Google Glass

google.com/glass



Rayban | Meta

nytimes.com



SenseCam

microsoft.com



GoPro (or sportscam)

amazon.es



amazon.es

Tipos de cámaras

Cámaras omnidireccionales

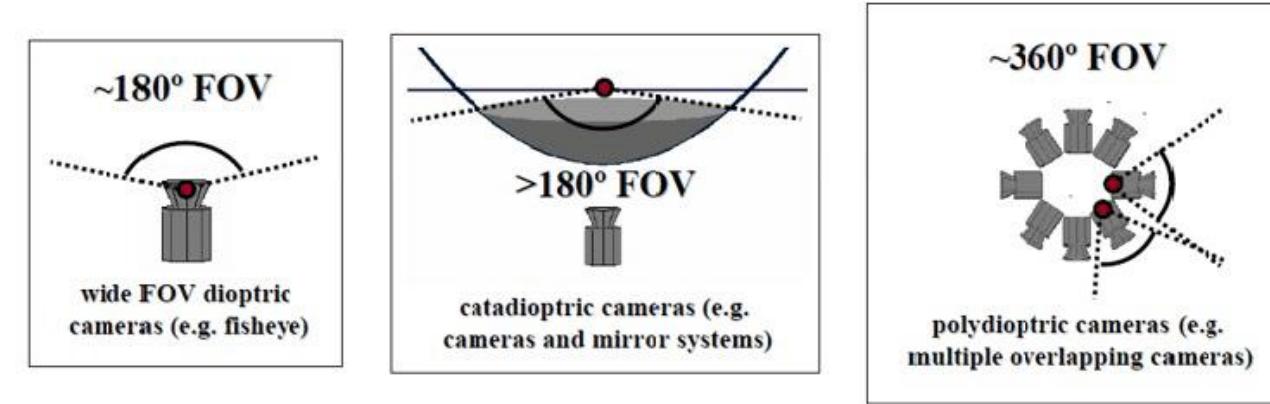
Capturan imágenes con un campo de visión de 360°, permitiendo observar todo el entorno desde un solo punto.

Características principales

- Campo de visión completo: cubren toda la escena sin necesidad de mover la cámara.
- Diseño óptico especial: pueden usar espejos cóncavos, lentes ojo de pez o sistemas multicámara.
- Salida de imagen: normalmente se obtiene una imagen esférica o panorámica que luego se “desenvuelve” mediante software.

Tipos de cámaras

Cámaras omnidireccionales



https://www.researchgate.net/figure/The-omnidirectional-camera-types-and-characterization-From-Youcef-Mezouar_fig1_309774819



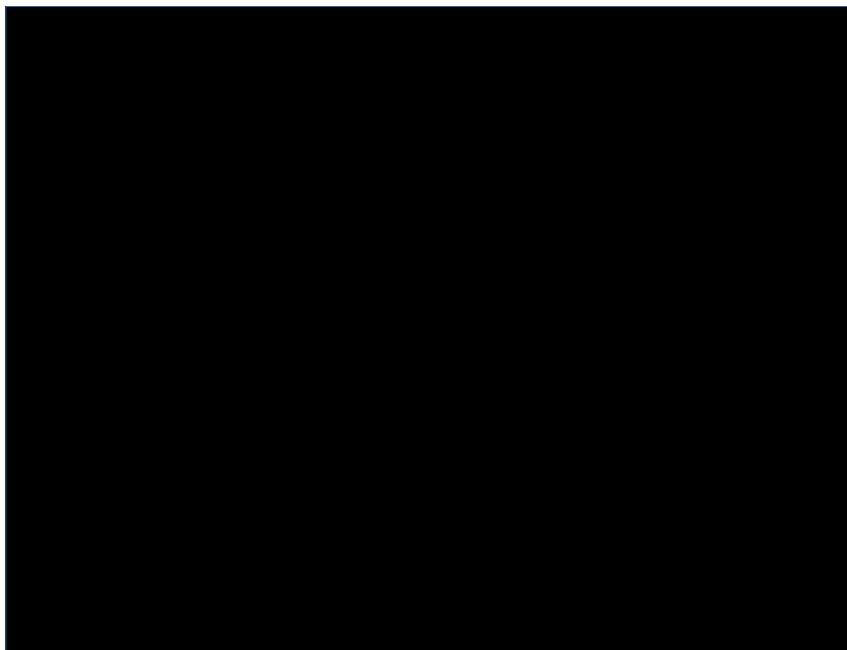
<https://www.google.com/intl/es/streetview/how-it-works>

Tipos de cámaras

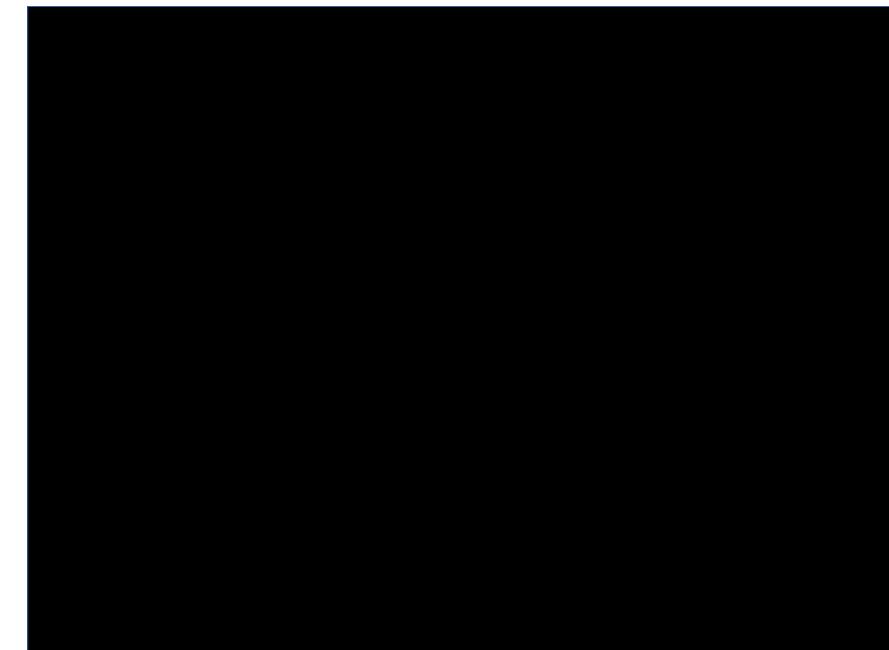
Cámaras omnidireccionales

El problema de la situación: lateral vs cenital

- Muchos más algoritmos para situación lateral de las cámaras (más común)
- Menos occlusiones en cámaras cenitales



<https://project.inria.fr/toyotasmarthome>

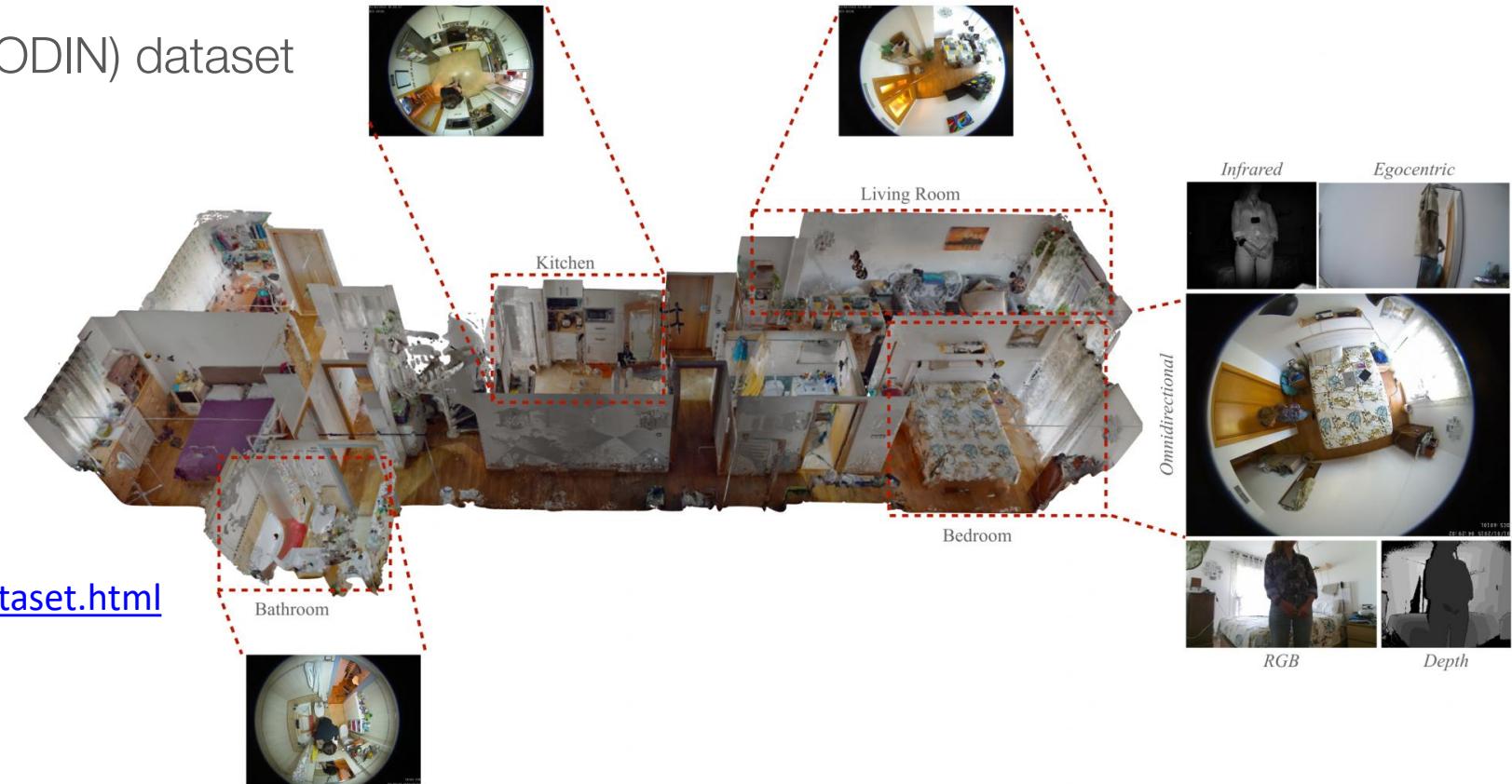


Tipos de cámaras

Cámaras omnidireccionales

El problema de la situación: lateral vs cenital

- OmniDirectional INdoor (ODIN) dataset



<https://web.ua.es/es/ami4aha/odin-dataset.html>

Tipos de cámaras

Cámaras edge

Combinan captura de imagen y procesamiento local de datos en un mismo dispositivo, reduciendo la dependencia del envío de información a la nube o a un servidor central.

Características principales

- Procesamiento en el borde (edge computing): la cámara analiza las imágenes directamente, sin necesidad de un ordenador externo.
- Hardware integrado: incluyen CPU, GPU o chips de IA (como NVIDIA Jetson o Google Coral).
- Reducción de latencia: las decisiones se toman casi en tiempo real.
- Optimización de ancho de banda: solo se envían resultados o metadatos, no las imágenes completas.
- Mayor privacidad: los datos sensibles no salen del dispositivo.

Tipos de cámaras

Cámaras neuromórficas

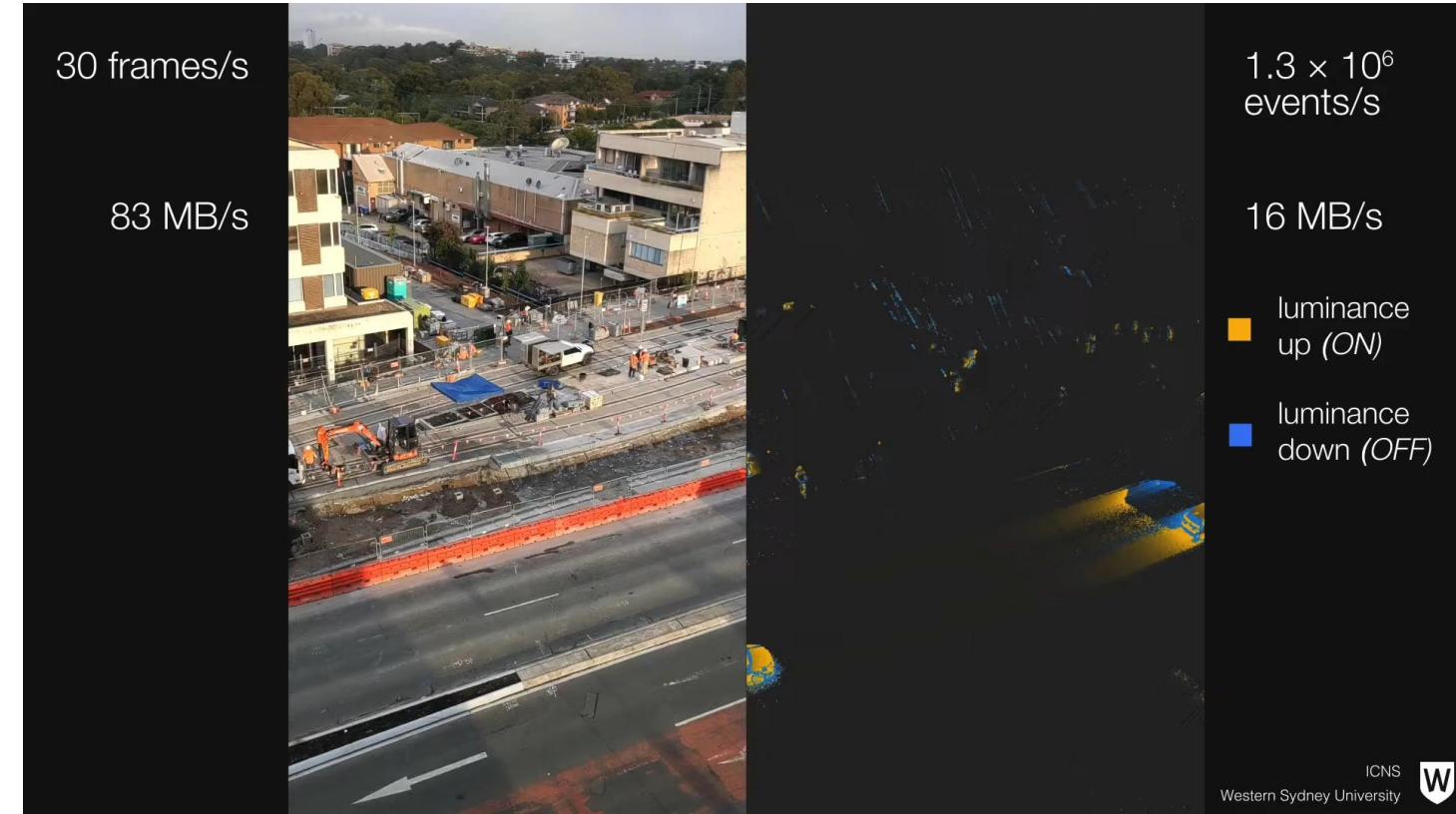
Imitar el funcionamiento del sistema visual biológico: registran solo los cambios en la escena, no fotogramas completos.

Características principales

- Inspiración biológica: basadas en cómo la retina codifica la información visual.
- Funcionamiento asíncrono: cada píxel actúa de forma independiente y emite un evento solo cuando detecta un cambio de luminosidad.
- Salida de datos tipo evento (event-based): en lugar de imágenes, generan una secuencia de eventos espaciales y temporales.
- Altísima velocidad temporal: pueden registrar micro-cambios de movimiento con latencias del orden de microsegundos.
- Bajo consumo energético: al no capturar fotogramas redundantes

Tipos de cámaras

Cámaras neuromórficas



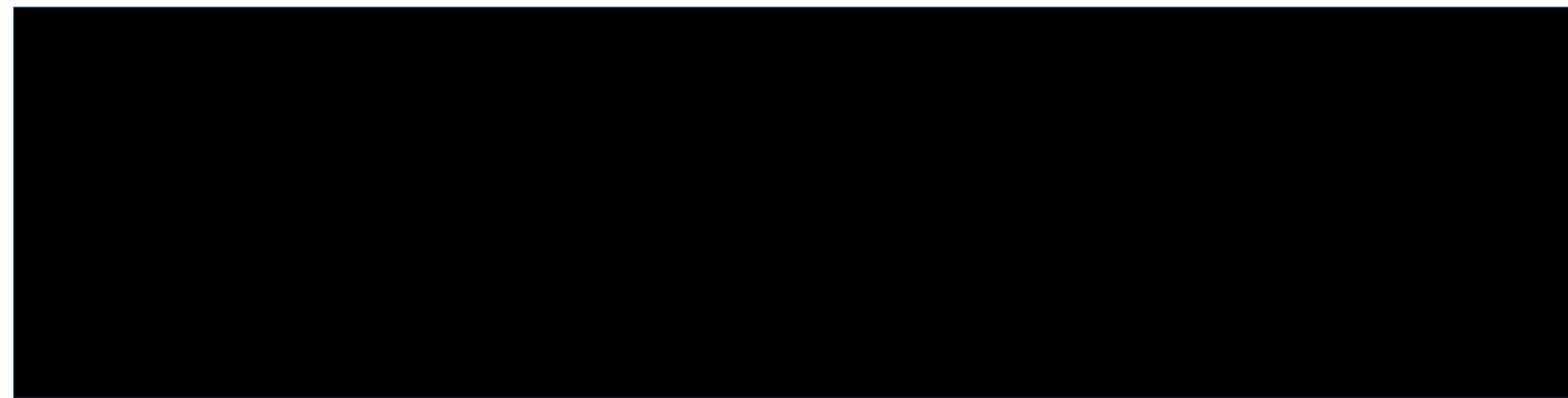
Tipos de cámaras

Combinación de modalidades

RGB, profundidad y térmica

Estas fuentes de datos pueden complementarse entre sí.

La preservación de la privacidad se logra cuando se utiliza solo D+T para el reconocimiento.



Tipos de cámaras

Introducción a la visión por computador para la interacción persona-máquina

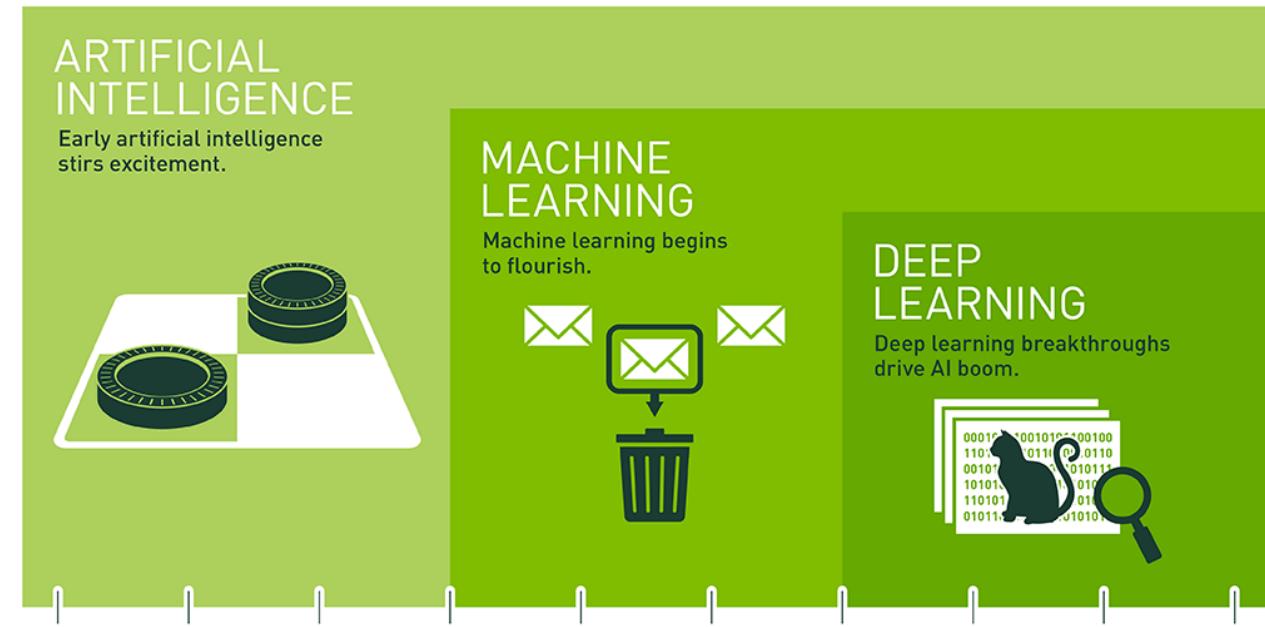
Procesamiento

Machine Learning (Aprendizaje automático)

- Supervisado
- No supervisado
- De refuerzo

Deep Learning (Aprendizaje profundo)

- Convolucionales
- Recurrentes
- Generativas
- Deep Belief



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

DL ⊂ ML ⊂ AI

Clasificación (supervisada, etiquetas discretas)

- Tenemos algunos correos electrónicos “buenos” y algunos de “spam”. Los etiquetamos ($L=2$).
- La máquina aprende los “rasgos” de ambos grupos (p. ej., presencia de palabras).
- Cuando llega un nuevo correo, se clasifica en consecuencia. Nota: las etiquetas podrían ser $L>2$.

Regresión (supervisada, salida continua)

- Ejemplo: ¿Podemos estimar el colesterol total únicamente a partir de medidas físicas?
- Peso, altura (o IMC); pero también circunferencia de cintura, cadera, piernas, etc.

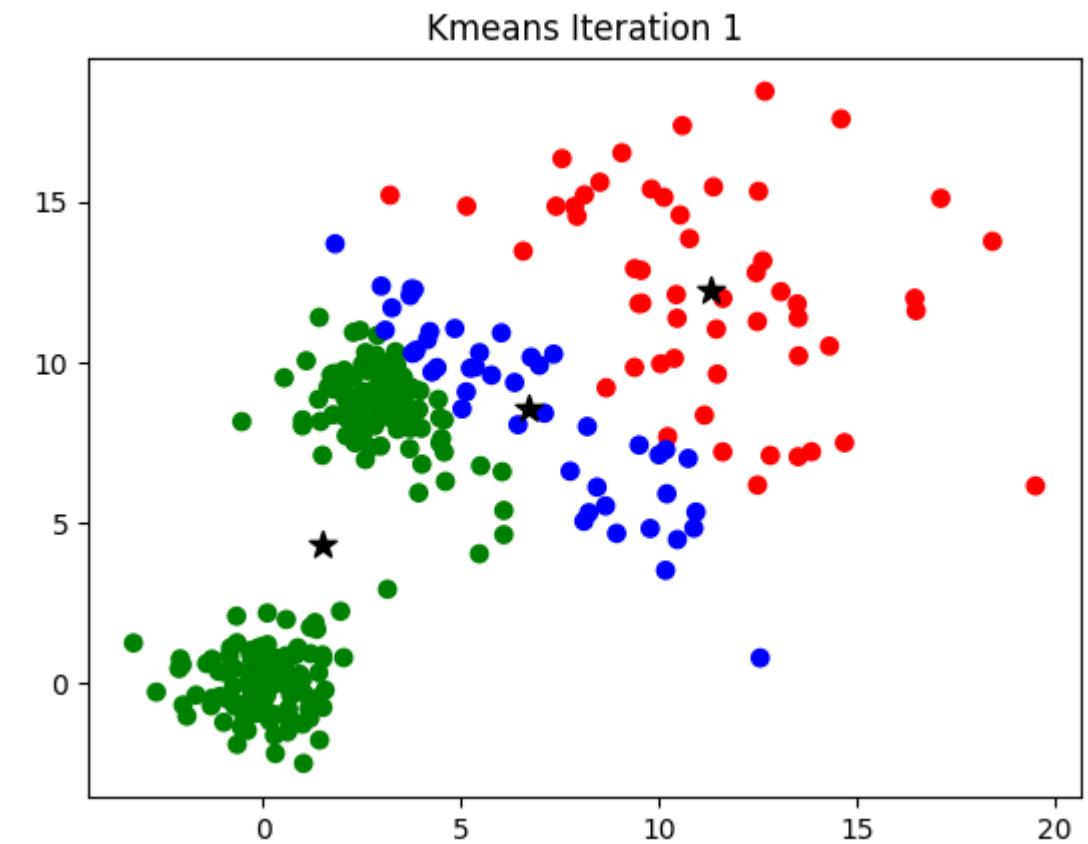
Clustering (no supervisado, genera grupos)

- Por ejemplo, ¿se pueden dividir fácilmente (de forma visual) grupos de pacientes en categorías?

Clustering k-Means

- Se inicializa aleatoriamente (por ejemplo $k=3$)
- Busca grupos de puntos por similitud
- Iterativo

Existen muchos otros métodos



Ejemplo de clustering

‘Ingeniería de características’ (ajuste de rasgos)

- El/la investigador/a o ingeniero/a “elige” descriptores y los “modifica” o “ajusta” según sea necesario.
- Por ejemplo: para detectar “orejas de panda” esperamos orejas redondeadas y peludas → características: detector de círculos, detector de texturas → ¿puntúa alto en ambos? Entonces, ¡es un panda!

Uso de descriptores locales/globales de una imagen (histogramas de color, histogramas de gradientes, detectores y descriptores de puntos de interés).

Clasificadores clásicos: p. ej., Support Vector Machines (SVM).

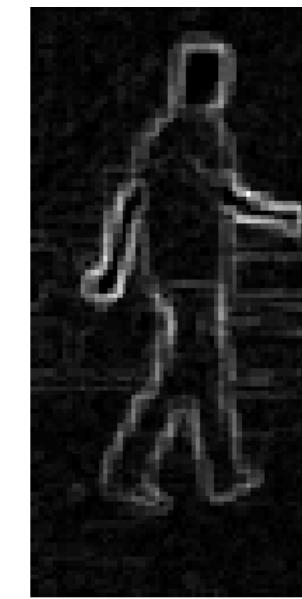
Clasificadores que utilizan descriptores locales: Bag of [visual] Words (BoW).

Pre – deep learning

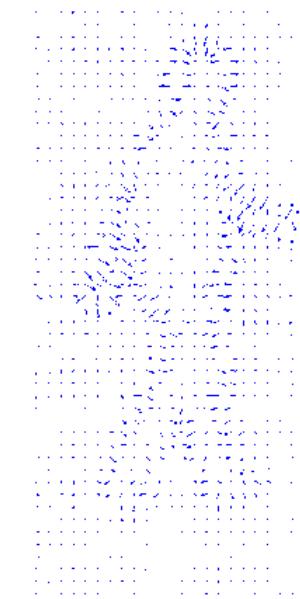
Descriptores – Histogramas de gradientes orientados (HOG)



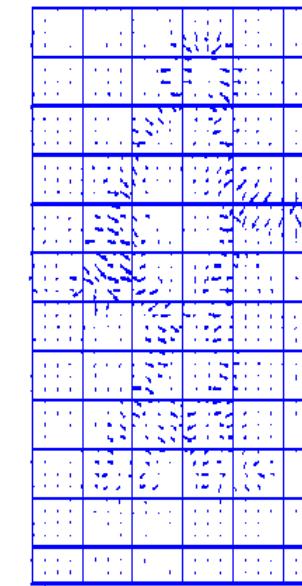
a



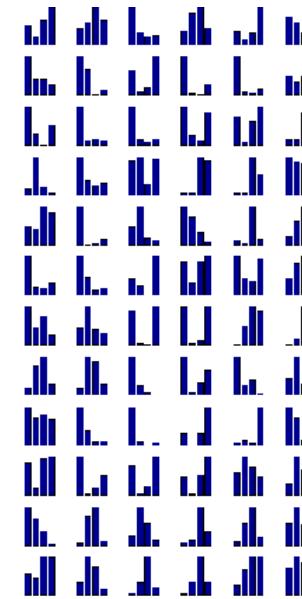
b



c



d



e

<https://ieeexplore.ieee.org/document/4357692?figureId=fig3#fig3>

Pre – deep learning

Descriptores – Extracción de descriptoros “en pirámide”

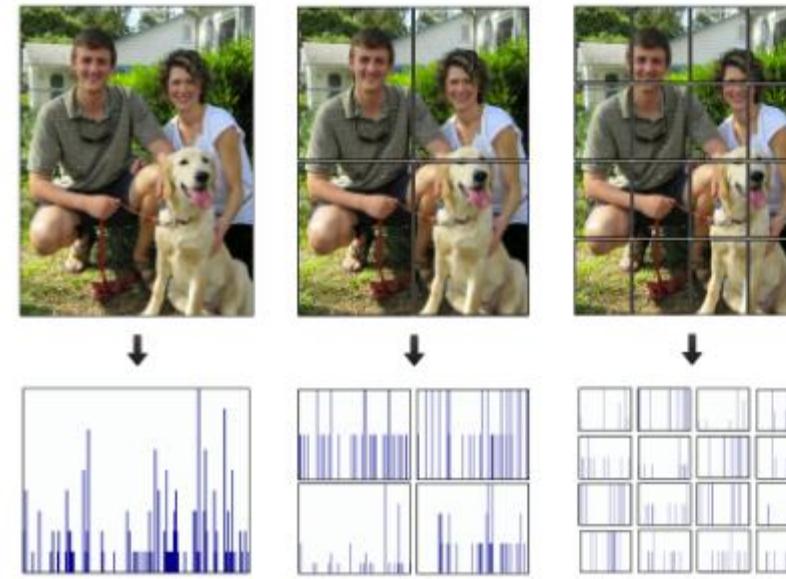


Figure 3: Visual representation of partitioning an image into sub-images and constructing the histograms.

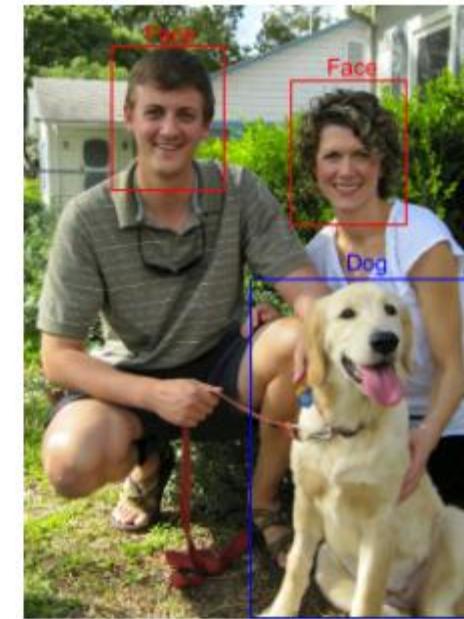


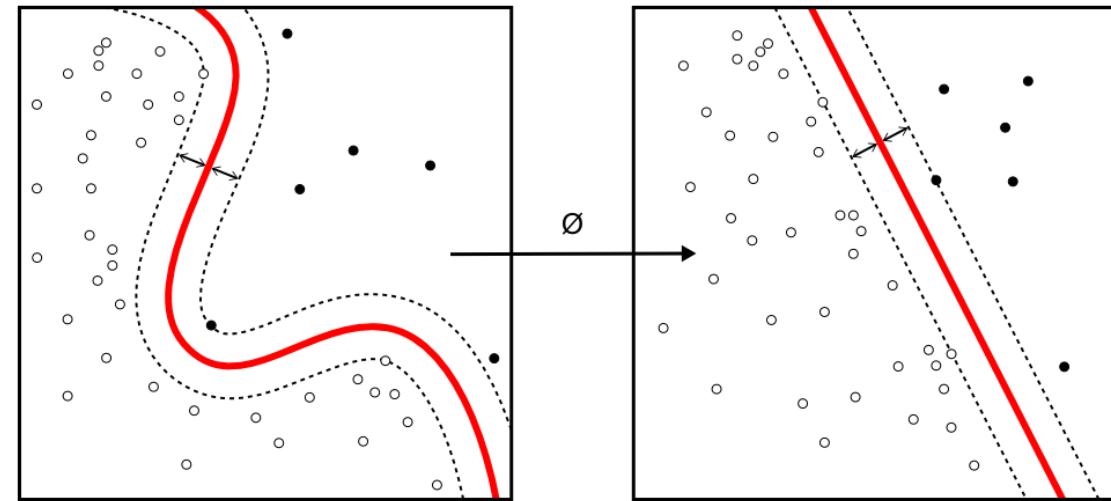
Figure 4: Results showing both image classification and localization.

<https://cs229.stanford.edu/proj2011/SchmittMcCoy-ObjectClassificationAndLocalizationUsingSURFDescriptors.pdf>

Pre – deep learning

Clasificadores clásicos – Ejemplo: Support Vector Machines (SVM)

Trata de encontrar la función que parte un [hiper]espacio



<https://www.npmjs.com/package/node-svm>

El SVM fue dominante en visión por computador hasta 2012, cuando las redes convolucionales demostraron que podían aprender automáticamente tanto los descriptores como las fronteras de decisión

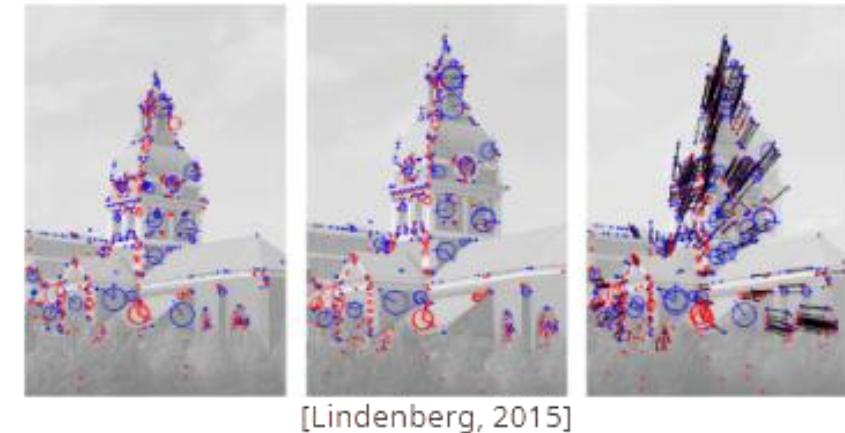
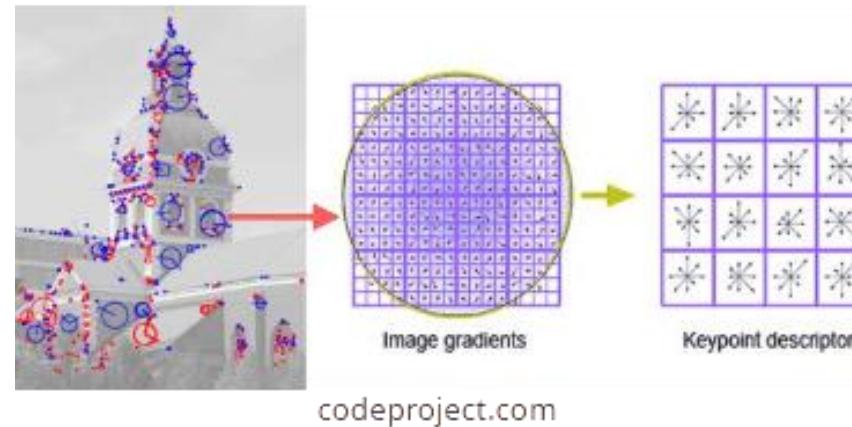
Pre – deep learning

Descriptores basados en puntos de interés

Puntos de interés → cambio abrupto en los valores de los píxeles

- Como se ha dicho, las superficies “planas” contienen poca información (baja entropía).
- Existen diferentes métodos: SIFT, SURF, FAST, Daisy, ORB, etc.

Los puntos de interés de fragmentos de imagen similares deberían ser numéricamente parecidos; los de fragmentos distintos, diferentes.



Pre – deep learning

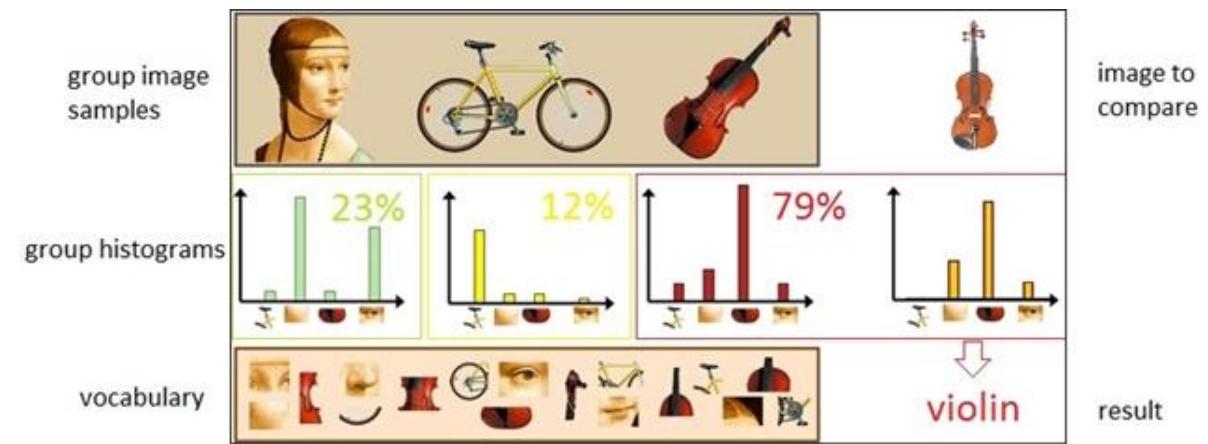
¿Cómo usar entonces descriptores ‘locales’?

Clasificador Bag-of-Words (palabras visuales)

- IDEA: “Si parece un pato, nada como un pato, y grazna como un pato, entonces probablemente sea un pato”
- Si las partes que componen un objeto están presentes, el objeto debe de estar allí.

Pasos:

1. Construir un “vocabulario” de “palabras clave” (fragmentos de imagen relevantes).
2. Evaluar la presencia de esas palabras.



Pre – deep learning

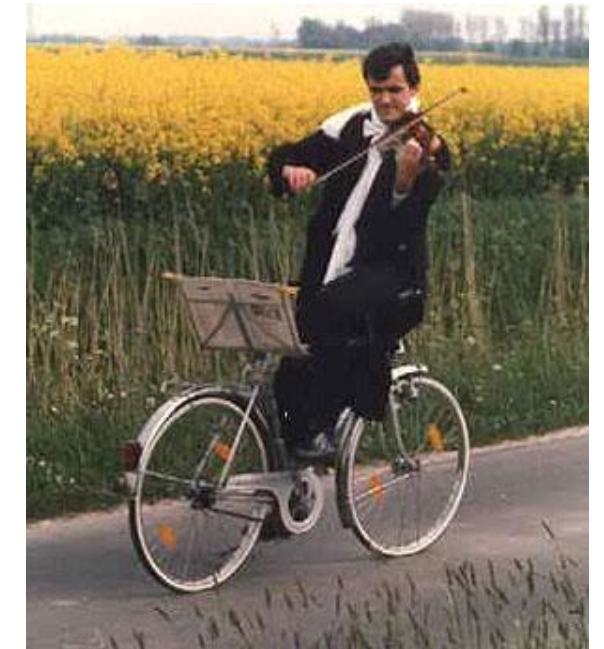
Conclusiones

¡La realidad es mucho más compleja! (por ejemplo, esto →)

- Descriptor HOG: “no es un humano”.
- Descriptores locales: “césped, bicicleta, persona, violín,
...”
- Perspectiva diferente: una bicicleta distinta a cualquiera vista antes.

Todos estos problemas deben resolverse.

- Demasiado trabajo “manual” para diseñar soluciones que no son generales.
- Y aun así, a veces resultan “inabordables”.



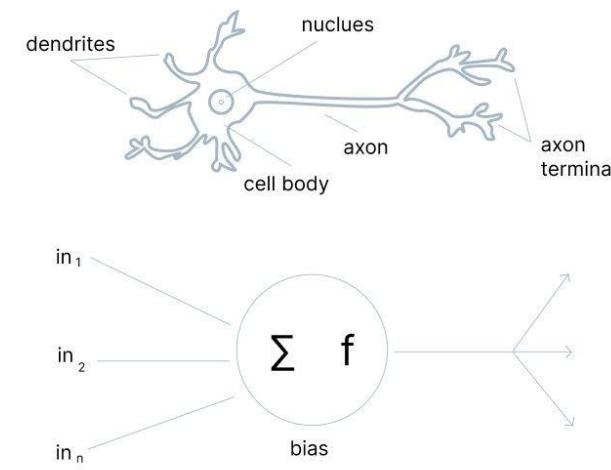
<https://es.pinterest.com/pin/383861568257586224>

Breve introducción

“Deep” porque emplea modelos neuronales con “muchas” capas.

Neurona: una simplificación matemática de la sinapsis neuronal.

- Cada neurona está en una capa, como ocurre en la corteza visual.
- Las neuronas de una capa reciben datos de las neuronas de la(s) capa(s) anterior(es).



Deep learning

Breve introducción

Algoritmos para aprender los mejores algoritmos (por ejemplo, los mejores descriptores)

Podemos tener problemas de ‘juguete’ o más complejos: ¿Muffin o chihuahua? ¿Fregona o perro?



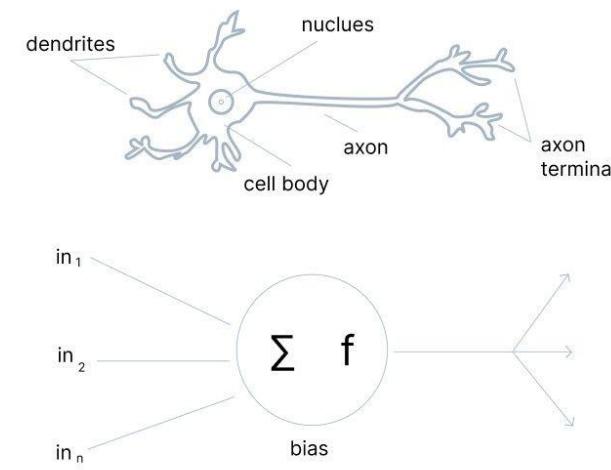
Deep learning

Breve introducción

“Deep” porque emplea modelos neuronales con “muchas” capas.

Neurona: una simplificación matemática de la sinapsis neuronal.

- Cada neurona está en una capa, como ocurre en la corteza visual.
- Las neuronas de una capa reciben datos de las neuronas de la(s) capa(s) anterior(es).



Deep learning

ML antes del deep learning

Hemos dicho: extraer características locales/globales.

- Si son globales, se introducen directamente en un clasificador (p. ej., un SVM).
- Si son locales, se usan técnicas de clustering y bag-of-words o similares (p. ej., Fisher vectors).

El algoritmo (el clasificador):

- Proceso de entrenamiento:
 - Recibe los descriptores (X) y las etiquetas esperadas (salidas; y').
 - Estas etiquetas son las “respuestas correctas”.
 - Durante el entrenamiento: se compara la salida del clasificador (y) con la etiqueta esperada (y').
 - La función interna (p. ej., un polinomio) que separa las muestras se actualiza con la “retroalimentación”.
- Una vez entrenado: se obtiene el descriptor de una imagen, se pasa al clasificador y se obtiene el resultado.

Deep learning

ML antes del deep learning

En cada paso, un ‘experto humano’ elige el extractor, modifica los descriptores, elige el clasificador, etc.



Deep learning

La idea tras las redes neuronales

Dejemos que el ordenador “aprenda” el algoritmo correcto.

Aprender conjuntamente lo mejor: algoritmo, descriptor, clasificador.

- Se utilizan “redes de neuronas”: cada neurona toma una entrada y aplica una operación matemática simple ($y = w \cdot x + b$), y devuelve una salida.
- Aunque la neurona sea muy simple, al apilarse en capas se pueden lograr operaciones mucho más complejas y matizadas.
- El aprendizaje consiste entonces en encontrar los **w** y **b** que aproximen la función que queremos aprender.

La idea tras las redes neuronales

Este tipo de sistemas requieren mucha potencia de cálculo.

Las redes neuronales existen en teoría desde los años 70 y hay ejemplos prácticos en los 90.

Sin embargo, su aplicación al mundo real era muy limitada debido al hardware disponible.

La industria de los videojuegos creó la necesidad de dispositivos capaces de generar gráficos complejos muy rápido (Graphical Processing Units, GPUs).

Hoy en día usamos las GPUs para computación masivamente paralela (incluyendo redes neuronales).

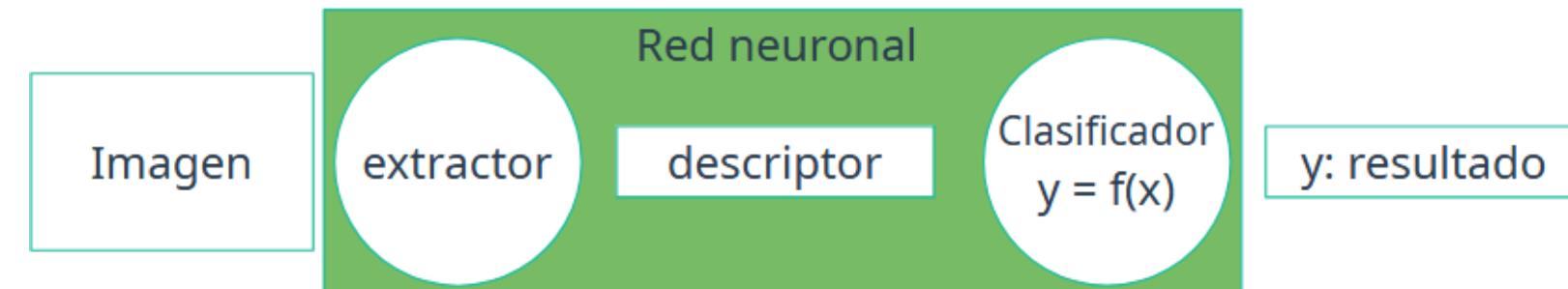
Deep learning

La idea tras las redes neuronales

En una red neuronal, los descriptores se “aprenden” a partir de los datos, junto con el mejor clasificador.

- El entrenamiento es un proceso iterativo, actualizando los w y b
- Los descriptores útiles para la tarea en cuestión se mantienen; de lo contrario, las actualizaciones los modifican.
- Los parámetros internos del clasificador también se actualizan.

El algoritmo que entrena la red neuronal recibe únicamente imágenes y etiquetas.

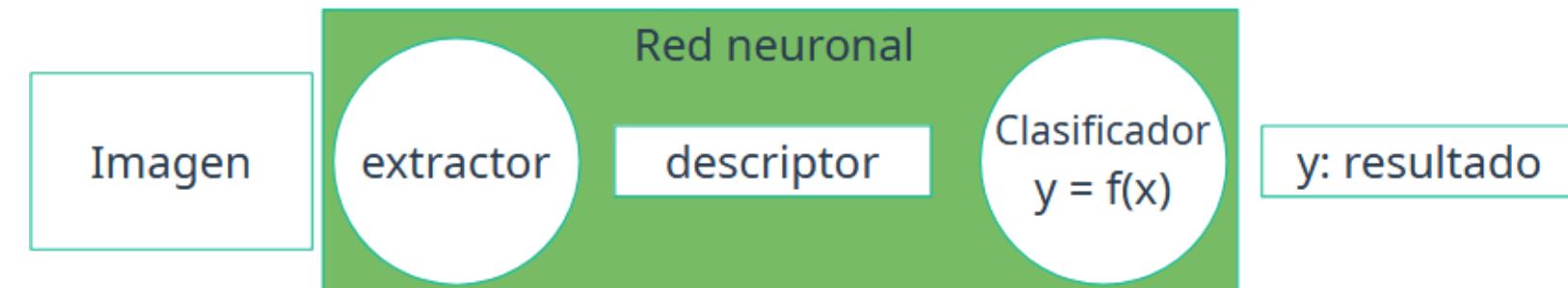


Deep learning

La idea tras las redes neuronales

De forma similar al caso sin redes neuronales:

- Las “etiquetas” conocidas y esperadas se comparan con la salida.
- Esto nos da un valor de qué tan lejos estamos del objetivo (loss o función de pérdida).
- Los pesos y sesgos (w , b) se actualizan en consecuencia.



Deep learning

Redes neuronales convolucionales (CNNs)

Este es el tipo más común cuando se trabaja con imágenes.

Las convoluciones aplican “filtros” a cada imagen.

Estos filtros, a diferencia de los descriptores/extractores de características, son aprendidos por la red.

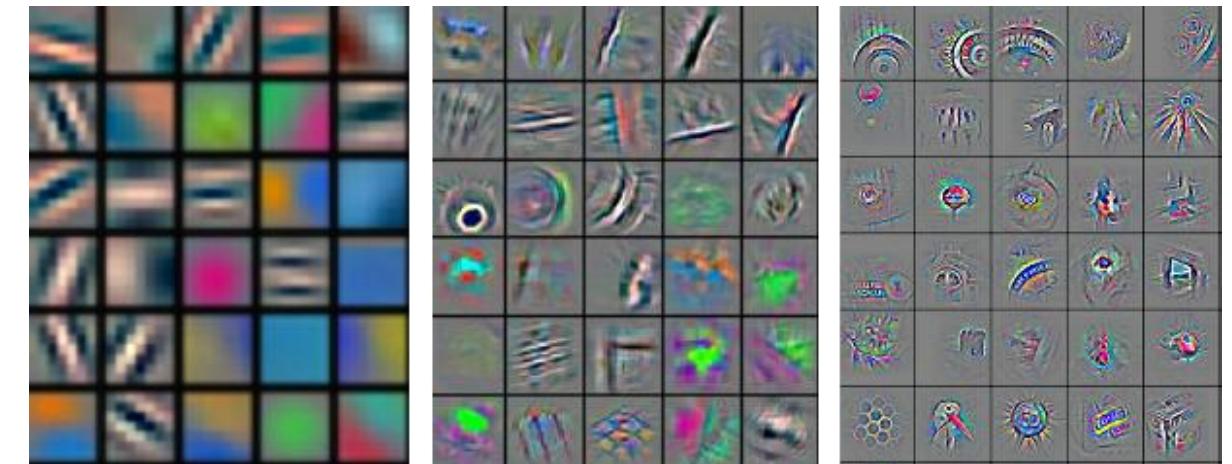
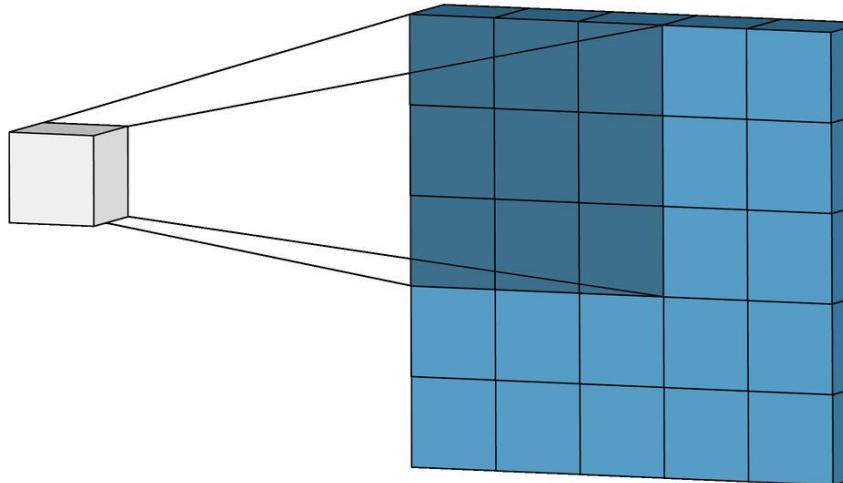
Para “resumir” los descriptores, se aplica una operación llamada pooling. Al final de la red hay neuronas que realizan la clasificación.

De este modo, se aprende la función $f(x)$ que separa correctamente los descriptores procedentes de diferentes imágenes.

Deep learning

Redes neuronales convolucionales (CNNs)

Convolución: concepto proveniente del campo del “procesamiento de señales”



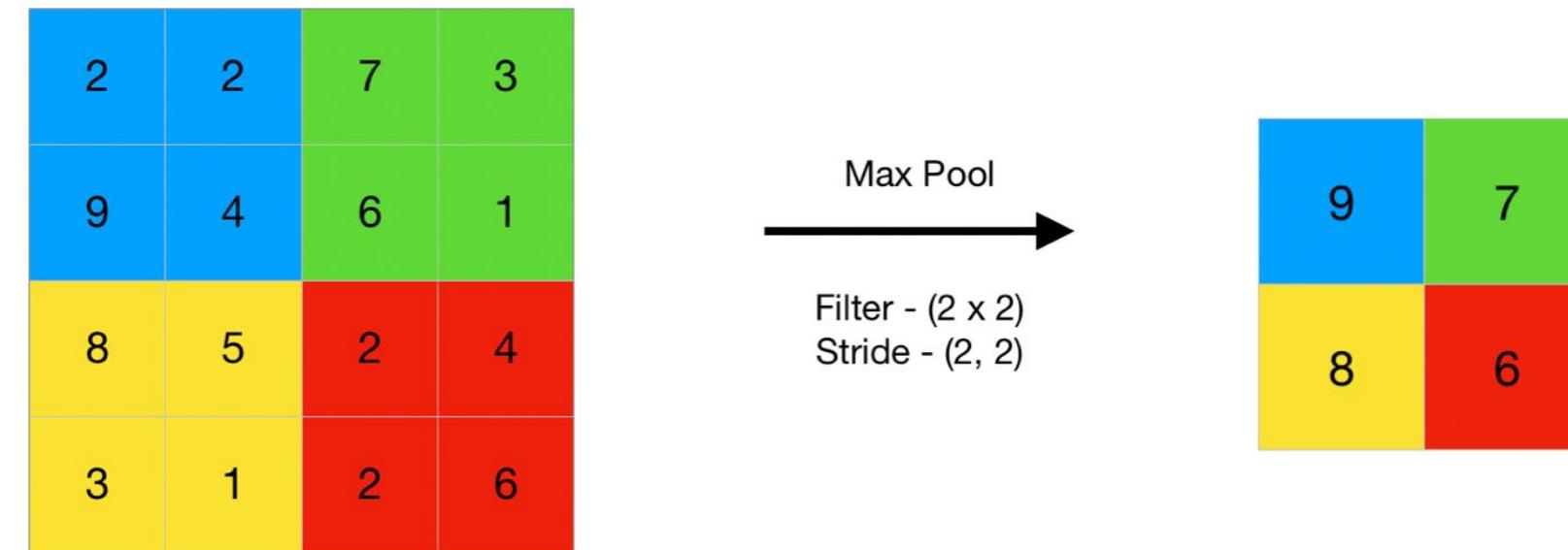
<https://medium.com/data-science/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1>
<https://stackoverflow.com/questions/41598858/how-can-filters-later-in-a-cnn-recognize-color>

Deep learning

Redes neuronales convolucionales (CNNs)

Operaciones de “pooling” – reducción de datos

Ejemplo: Max Pooling



Deep learning

Redes neuronales convolucionales (CNNs) - VGG-16

Modelo de red puramente CNN:

- Capas azules: convolución
- Capas rojas: pooling

Clasificador mediante neuronas completamente conectadas

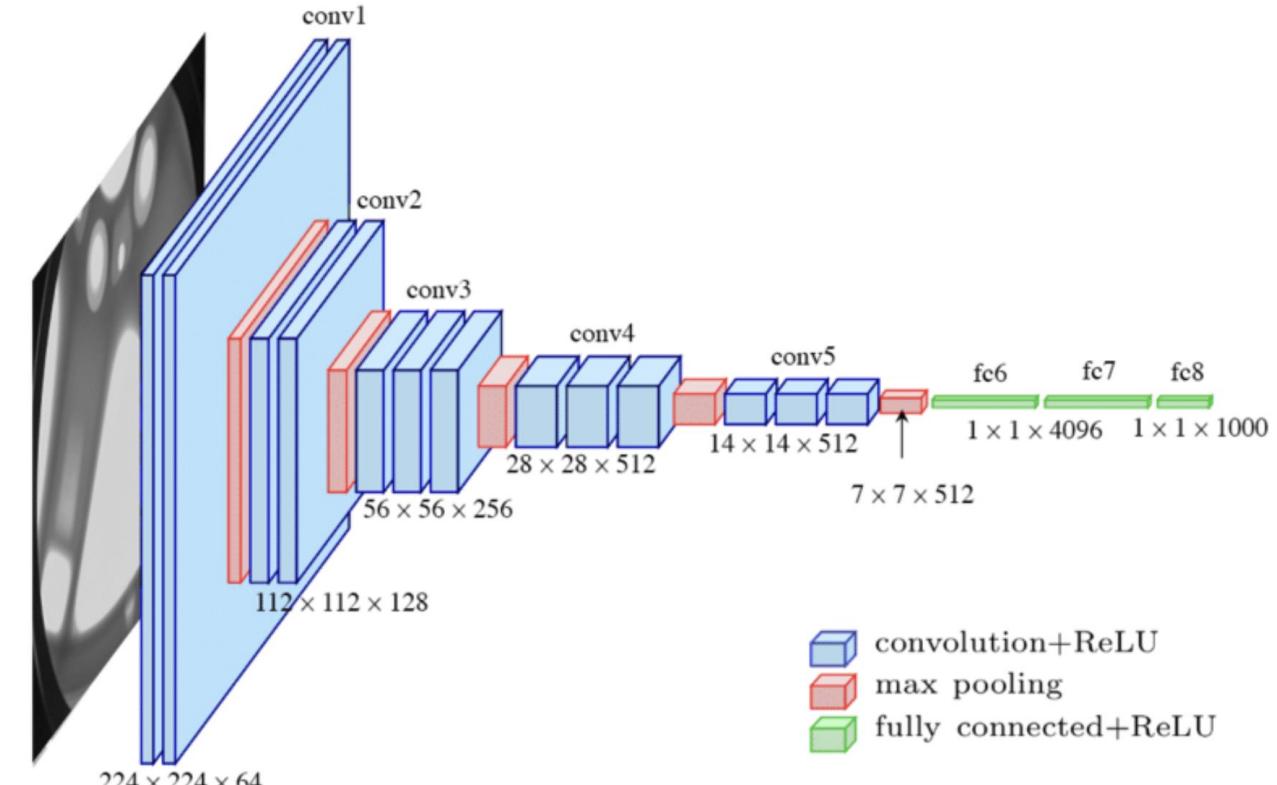


Fig. credits: Khuyen Le, medium.com

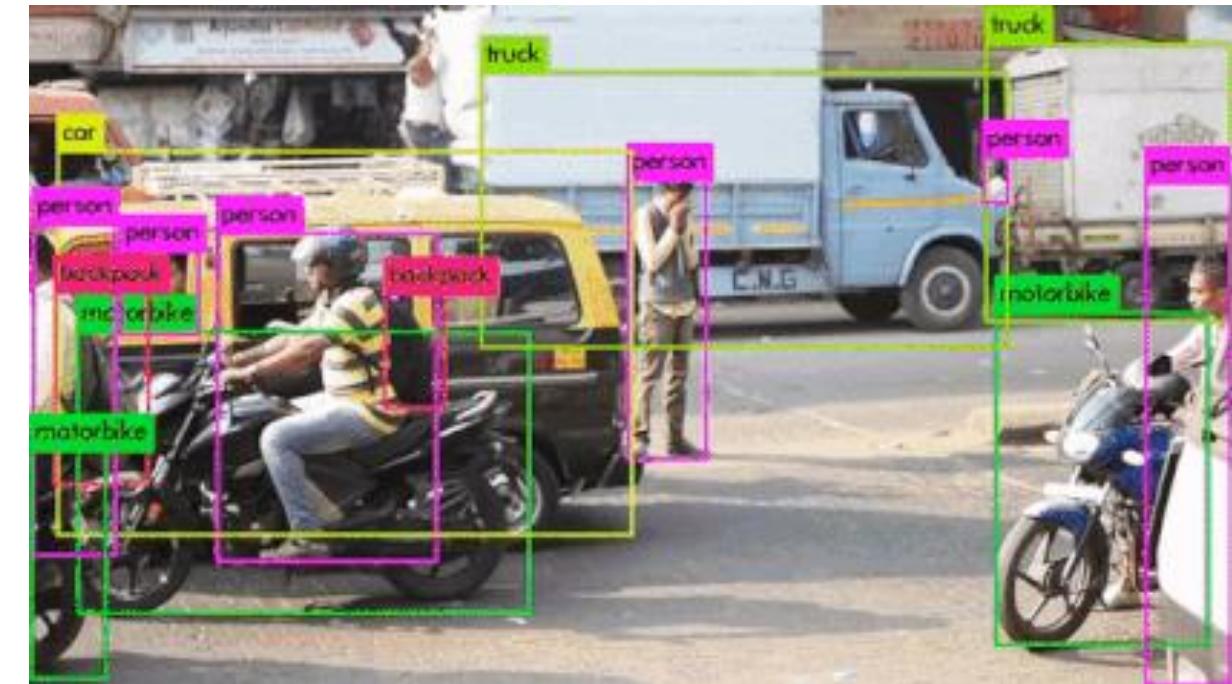
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Deep learning

Redes neuronales convolucionales (CNNs) – YOLO v3

No solo para clasificar imágenes enteras

YOLO realiza detección y
clasificación de objetos



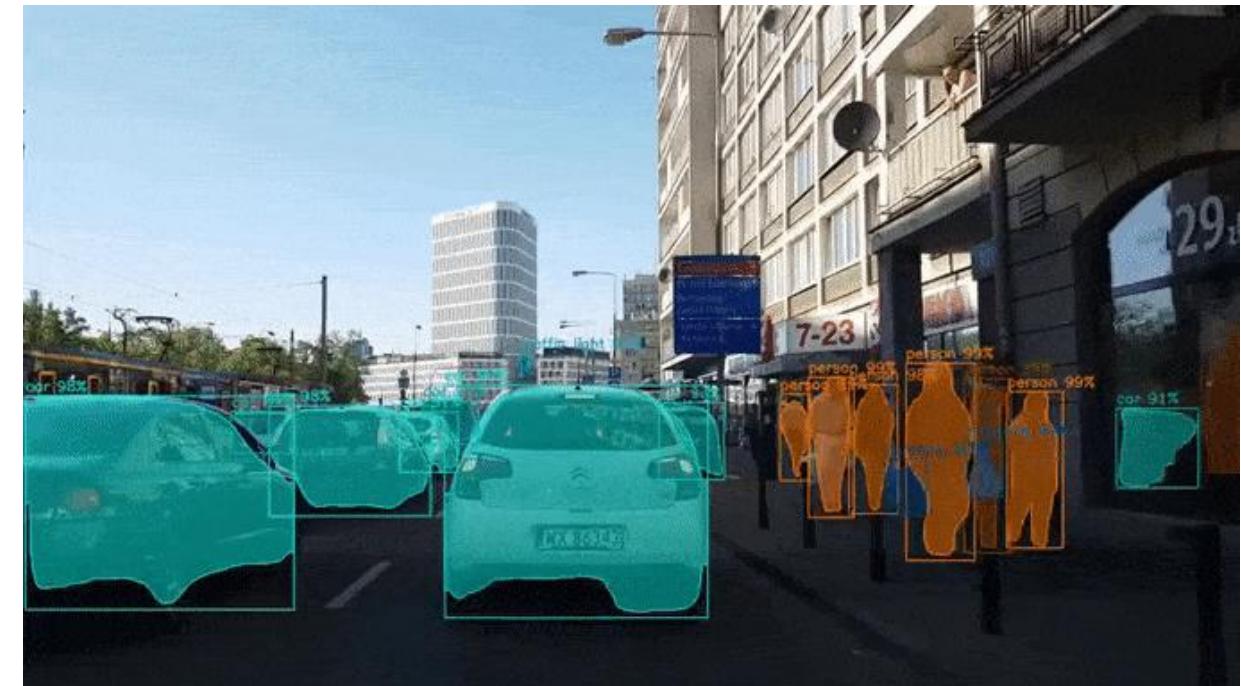
<https://medium.com/data-science/review-yolov3-you-only-look-once-object-detection-eab75d7a1ba6>

Deep learning

Redes neuronales convolucionales (CNNs) – Mask R-CNN

La “R” es de Residual, lo que significa que existen conexiones con capas no adyacentes.

“Mask” porque su salida es una silueta y no solo una bounding box (caja delimitadora).



https://github.com/karolmajek/Mask_RCNN

Deep learning

Redes neuronales convolucionales (CNNs) – OpenPose

Devuelve ‘esqueletos’

Coordenadas (x, y) de las articulaciones del cuerpo



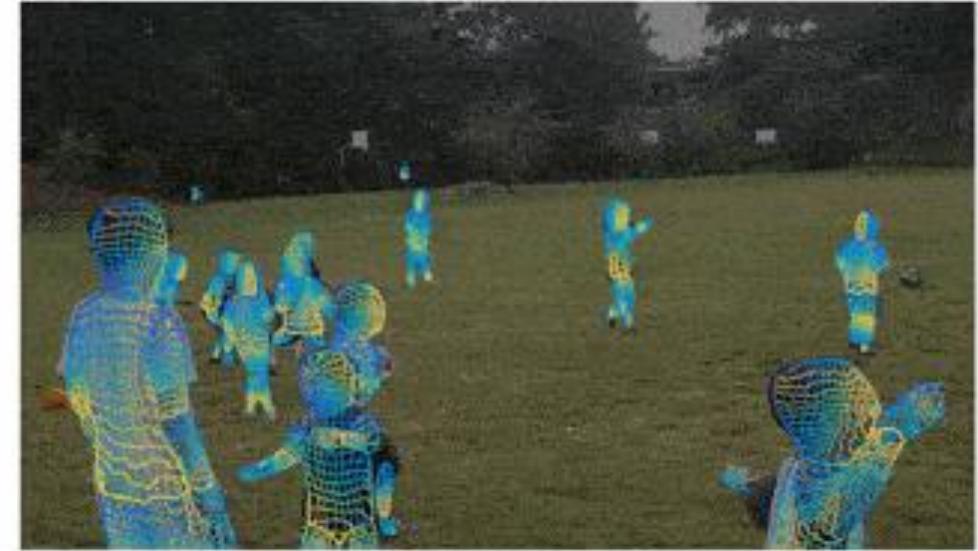
<https://github.com/CMU-Perceptual-Computing-Lab/openpose>

Deep learning

Redes neuronales convolucionales (CNNs) – DensePose

Devuelve una “malla” de puntos sobre la superficie del cuerpo

Emplea el modelo SMPL ([Skinned
Multi-Person Linear](#))



<http://densepose.org>

Deep learning

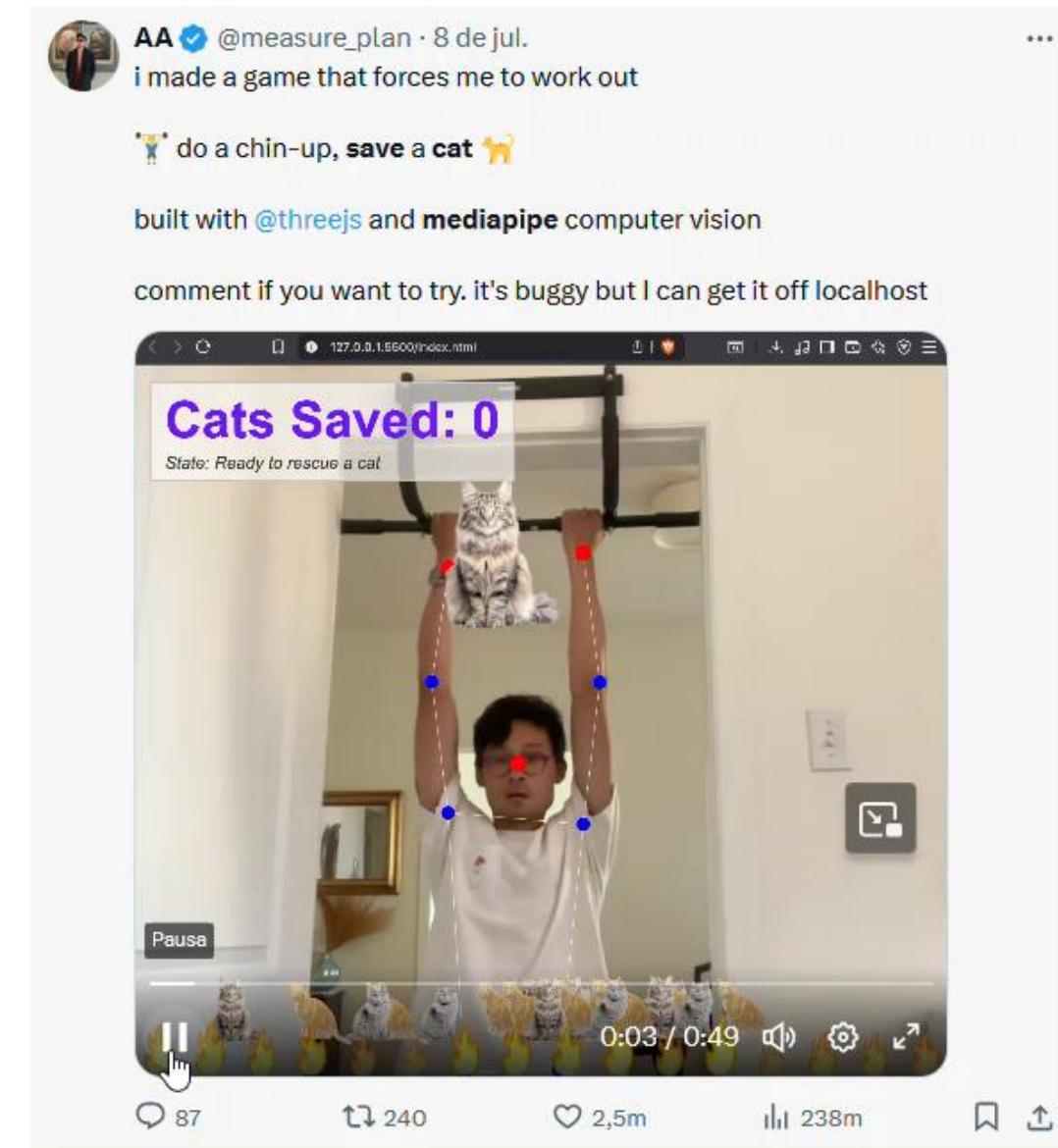
OpenPose, y redes similares nos dicen la posición de las partes del cuerpo

DensePose también da un “mapa de bits” en que cada parte del cuerpo recibe un valor numérico distinto

Esto se puede utilizar para realizar reconocimiento de gestos, acciones, actividades

Aplicaciones en interacción persona-máquina y persona-entorno:

- Interfaces mediante visión, gestos, reconocimiento fino de los dedos de la mano, etc.
- Entornos/Ambientes inteligentes, para confort, pero también para aplicaciones en envejecimiento activo y saludable



Práctica

Interacción persona-máquina

Unidad 4. Interfaces para reconocimiento de movimiento basadas en sensores y vídeo

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Interacción persona-máquina

Unidad 4. Interfaces para reconocimiento de movimiento basadas en sensores y vídeo

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Reconocimiento de movimientos,
acciones y actividades mediante
visión

‘Gesto’ o movimiento: primitiva de movimiento

- Por ejemplo: avanzar una pierna al andar, o levantar un brazo para saludar

Acción: secuencia de movimientos semánticamente coherente

- Se incluye la interacción con objetos
- Por ejemplo: levantar el brazo, palma abierta, mover el brazo lateralmente (= realizar un saludo con el brazo)

Actividad: secuencia de acciones, con interacción con objetos y/o personas

- Por ejemplo: Actividades de la Vida Diaria (AA.V.D.) cocinar, sentarse a comer, lavar los platos, etc.

Comportamiento: se derivan de la realización de actividades de forma rutinaria en el día a día

- Por ejemplo: Análisis de la tendencia de la hora de cocinar la comida, horas de comer habituales, desviaciones de patrones rutinarios, etc.

De gestos a comportamientos

Cada nivel dentro de esta taxonomía, representa un grado de complejidad semántica mayor (más información)

Cada nivel, además, representa un marco temporal más amplio (segundos, minutos, horas, días)



Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2012). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert systems with applications*, 39(12), 10873-10888.

De gestos a comportamientos

Para la interacción persona-máquina por tanto, será necesario, dependiendo del contexto de aplicación, determinar qué se quiere reconocer

Por ejemplo, para aplicaciones de interacción con computadores (interfaces de usuario), puede que sobre con el uso de gestos, movimientos “de ratón” en el aire, o similares



MOVIMIENTOS, ACCIONES

Para sistemas más complejos, en la interacción “persona-entorno”, en la aplicación de entornos inteligentes para aplicaciones concretas (por ejemplo, sistemas de apoyo a la salud, apoyo a la discapacidad, apoyo al envejecimiento activo y saludable), puede que necesitemos grados mayores de complejidad en la detección



ACTIVIDADES, COMPORTAMIENTOS

Qué reconocer según el tipo de interacción

Variaciones en perspectiva: La misma acción puede verse muy diferente según el punto de vista de la cámara o su posición relativa a la persona.

Cambios de iluminación: Sombras, reflejos o variaciones de luz afectan el aspecto de las imágenes y dificultan la detección fiable.

Diferencias de velocidad: Una misma acción (por ejemplo, caminar) puede realizarse a ritmos distintos, lo que complica la comparación temporal.

Escala y distancia: El tamaño aparente de la persona o del objeto cambia según la distancia a la cámara, afectando la precisión de los descriptores.

Oclusiones: Partes del cuerpo pueden quedar ocultas por objetos, otras personas o el propio movimiento, interrumpiendo la detección.

Principales retos

Interacción persona máquina: Interfaces alternativas al ratón/teclado o para personas con movilidad reducida

Videojuegos: Control sin dispositivos físicos, interacción natural en entornos VR/AR

Vigilancia y seguridad: Detección de eventos importantes en sistemas CCTV con múltiples cámaras

Cuidado de personas mayores: Monitorización de actividades y evaluación de la capacidad funcional

Rehabilitación: Herramientas (videojuegos “serios”) para pacientes con dificultades motrices

Resumen automático de vídeos: Identificación de escenas clave

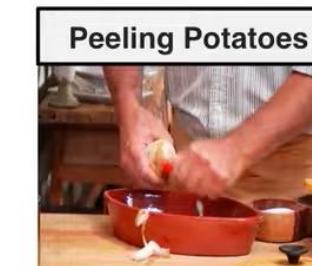
Indexación de vídeos: Búsqueda basada en contenido/semántica

Biometría conductual: Identificación basada en patrones de movimiento (p.ej., la marcha)

Áreas de aplicación

De propósito general – RGB

Ejemplos: [Kinetics](#), [AVA \(Atomic Visual Actions\)](#), [Charades](#), [ActivityNet](#)



ACTIVITYNET



Ejemplos de actividades -
datasets

De propósito general – RGB-D

Ejemplo: "[NTU RGB+D](#)" datasets



Ejemplos de actividades -
datasets

De propósito específico – RGB-D

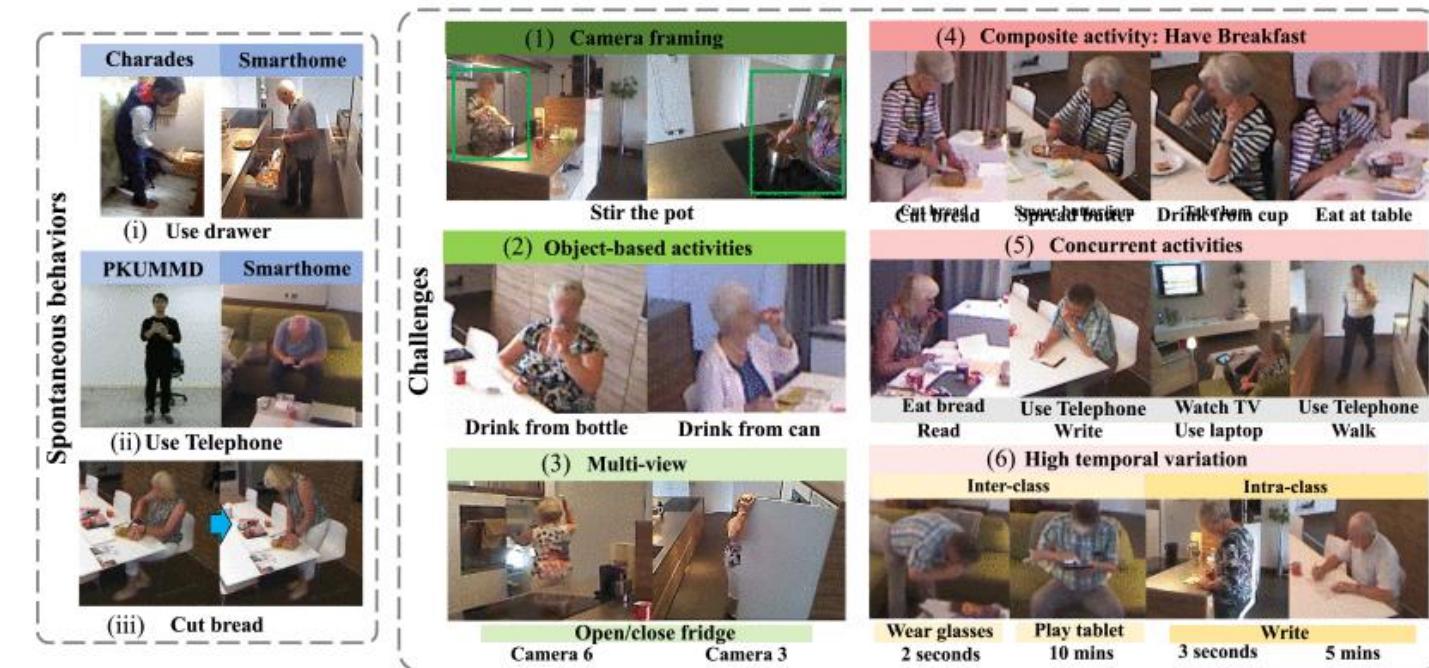
Ejemplo: [Toyota Smarthome](#)

Conjunto específico de personas mayores realizando actividades de la vida diaria (AA.V.D.)

Múltiples cámaras RGB+D

Vistas laterales (no cenital)

Natural (sin guion), los participantes ‘vivieron’ en el entorno durante 24-48 horas



Ejemplos de actividades - datasets

De propósito específico – RGB-D

Ejemplo: [Toyota Smarthome](#)



Ejemplos de actividades - datasets

De propósito específico – RGB-D

Ejemplo: [Toyota Smarthouse](#)

¿Cómo se valida?

- Protocolo fijo definido por los autores
- Métrica: mean per-class accuracy
- Cross-Subject (CS): 11 sujetos para train, los restantes para validación
- Cross-View (CV)
 - CV1: Train (cámara 1), Validación (cámara 5), Test (cámara 2)
 - CV2: Train (cámaras 1, 3, 4, 6, 7), Validación (cámara 5), Test (cámara 2)

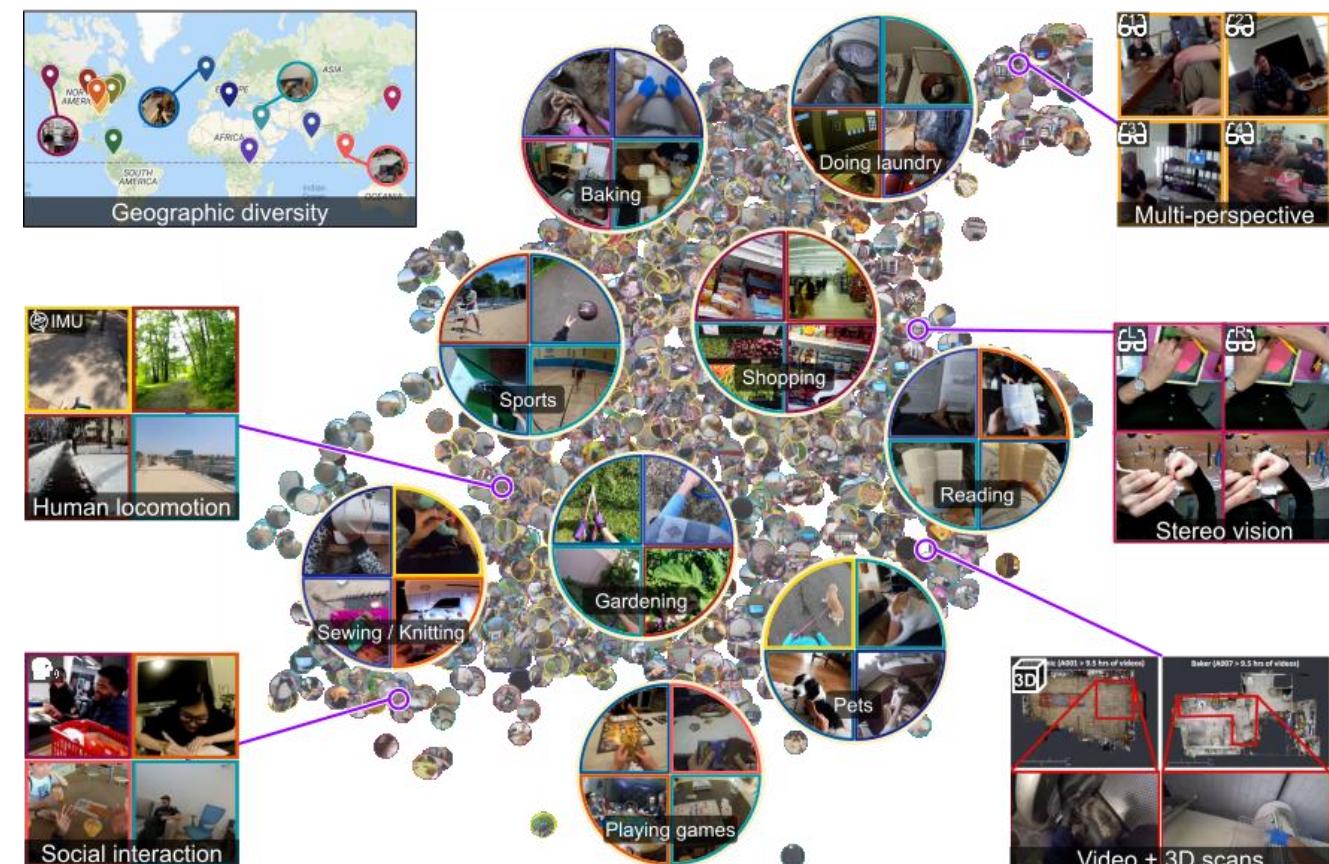
Ejemplos de actividades -
datasets

De propósito general – Visión egocéntrica

Ejemplo: [Ego4D](#)

Gran conjunto de datos desde perspectiva de visión egocéntrica

Gran variedad geográfica y cultural en la forma de realizar las actividades



Ejemplos de actividades - datasets

De propósito específico – Visión egocéntrica

Ejemplo: [Epic Kitchens](#)



Ejemplos de actividades -
datasets

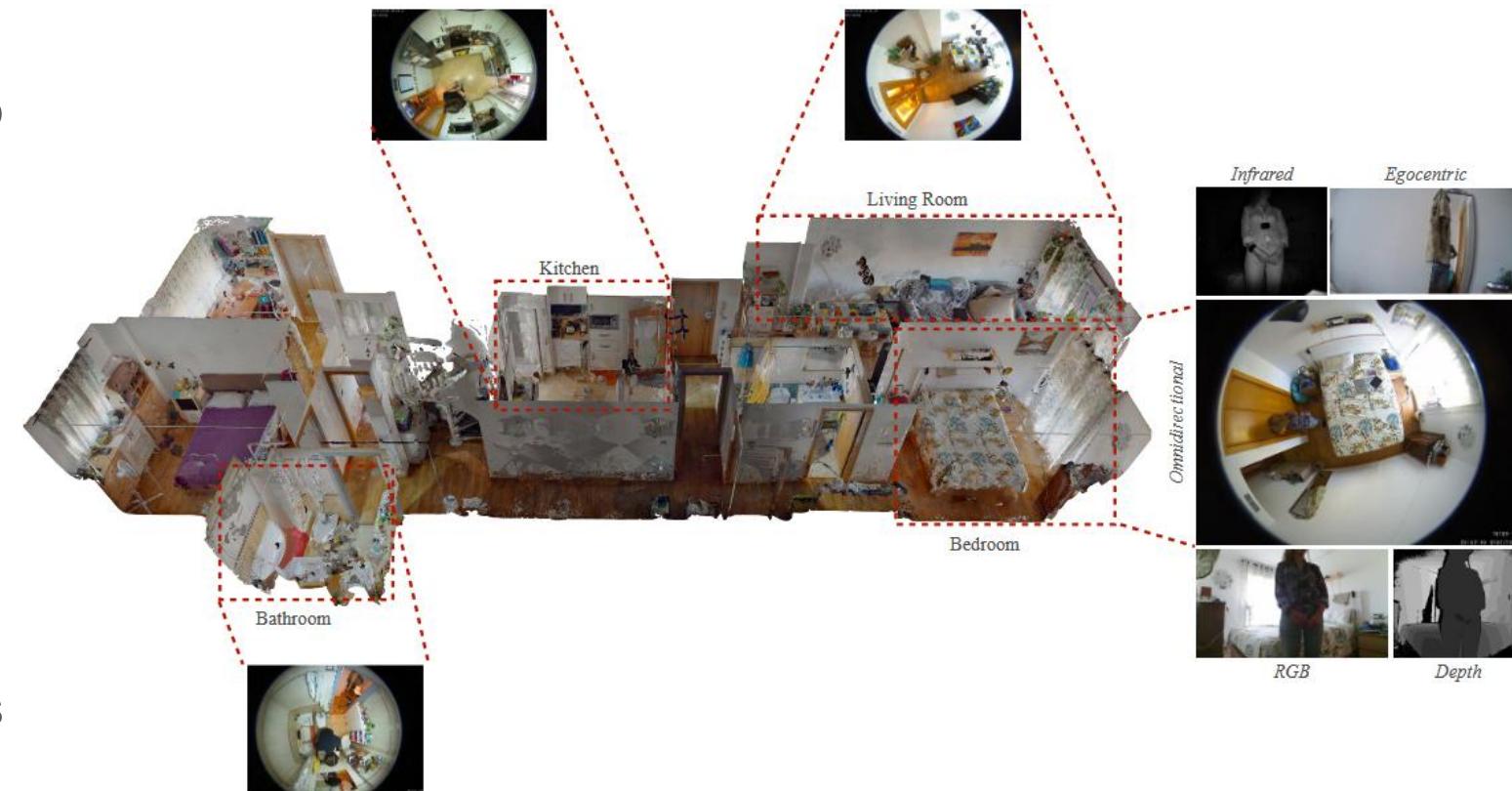
De propósito específico – Múltiples modalidades

Ejemplo: [OmniDirectional Indoor \(ODIN\) dataset](#)

Capturado por nuestro grupo
de investigación

Múltiples vistas y modalidades:

- Omnidireccional (cenital)
- Laterales (RGB+D)
- Egocéntrica
- Escaneos 3D de las viviendas

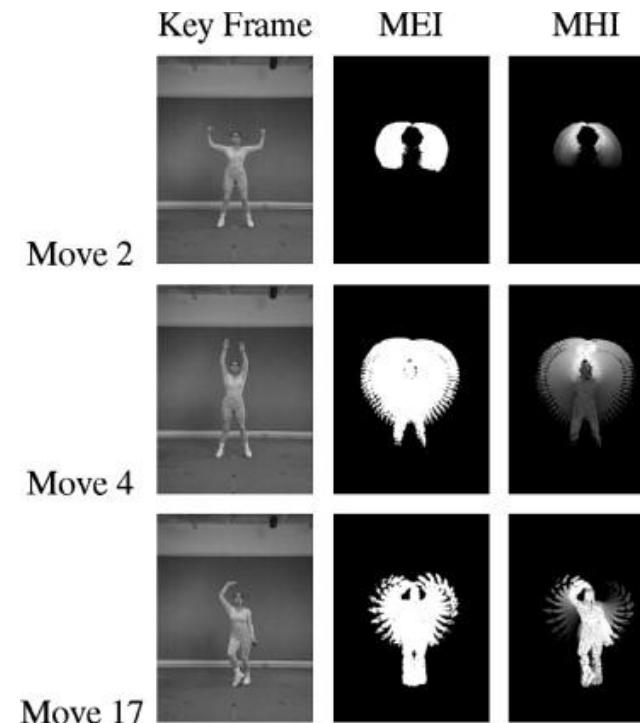


Ejemplos de actividades -
datasets

Motion history/energy image (MEI/MHI)

Reconocimiento a partir de extracción de siluetas basada en eliminación del fondo (*background subtraction*)

Cada movimiento ‘deja’ un rastro de patrones distintos, se analizan esas ‘huellas’ de cada acción

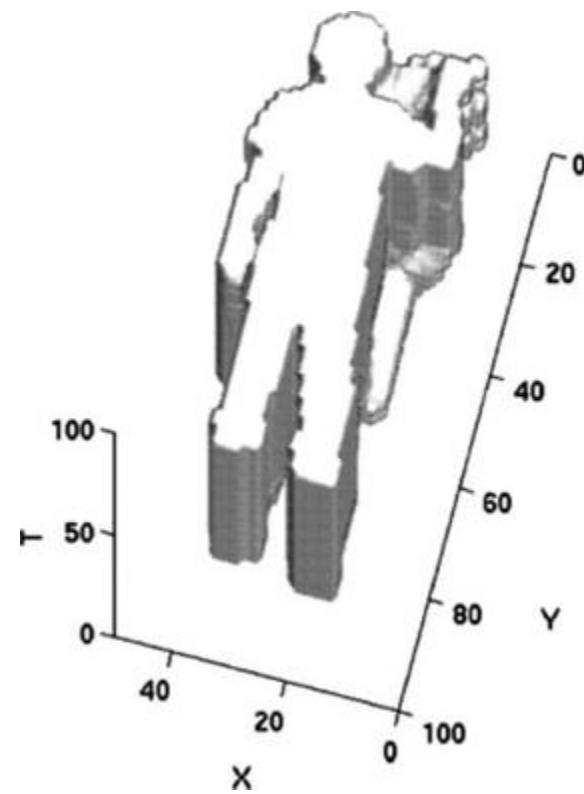


Bobick, A. F., & Davis, J. W. (2002). The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3), 257-267.

Pre-deep learning

Volúmenes XYT

Se genera un volumen a partir de la evolución de la silueta de la persona a lo largo del tiempo



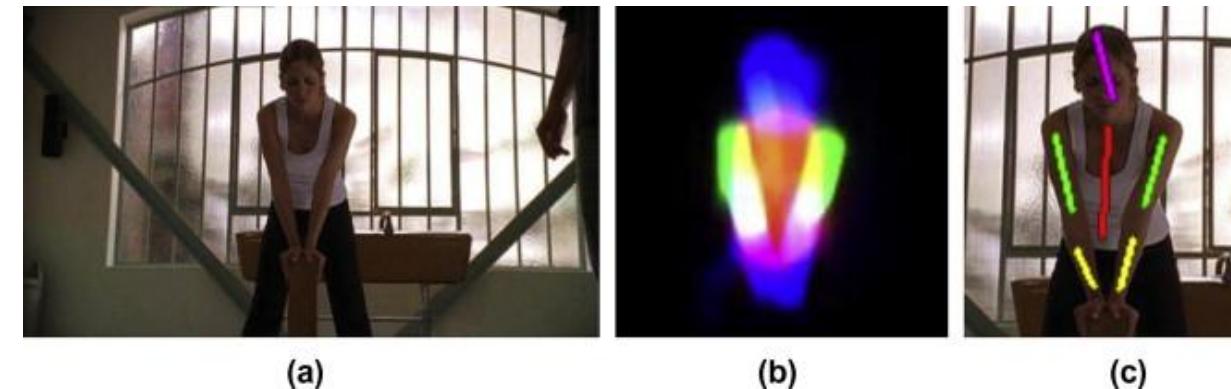
Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11), 1473-1488.

Pre-deep learning

Aunque el reconocimiento de acciones se puede realizar directamente desde información visual

Habitualmente, se usan técnicas de detección de las partes del cuerpo (body part detection), o extracción de la pose corporal (human pose estimation)

En algunos casos, solo nos interesa la detección de la parte superior del tronco, el cuerpo completo, o las manos



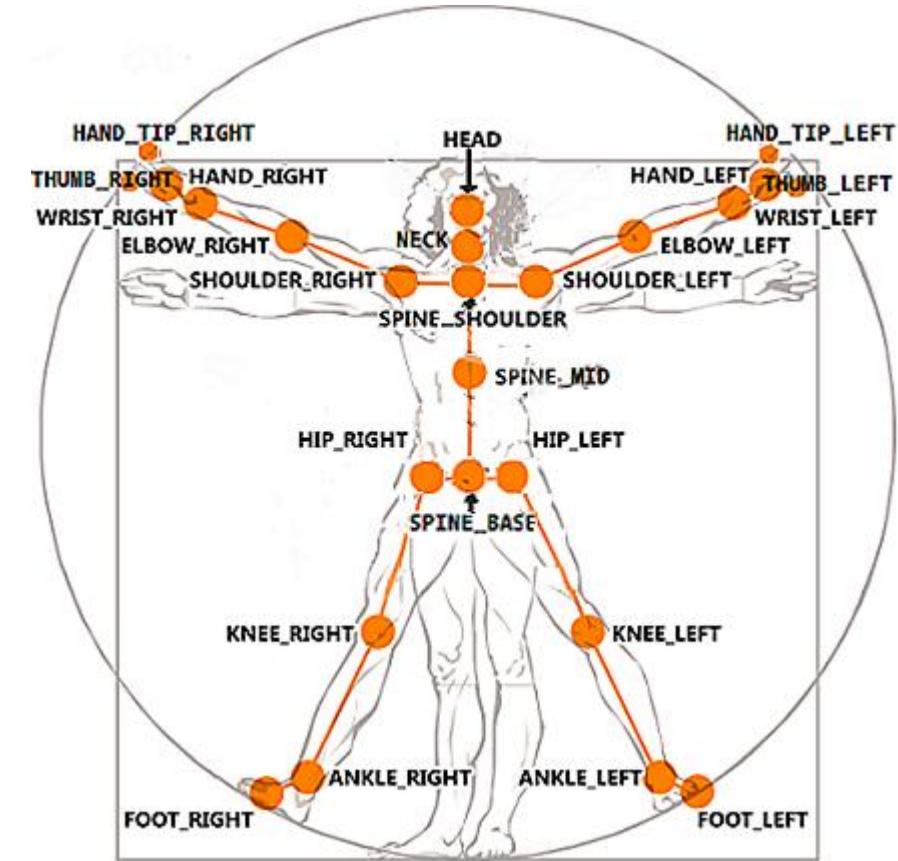
Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008, June). Progressive search space reduction for human pose estimation. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

Estimación de pose

Aunque los llamemos ‘esqueletos’ no tienen porque ser una representación anatómicamente fiel

En realidad, son grafos (árboles) formados por las posiciones 2D o 3D de las articulaciones del cuerpo

Para extraerlos, se usan detectores de partes del cuerpo entrenados específicamente para cada parte, y se unen en representaciones ‘plausibles’ estadísticamente



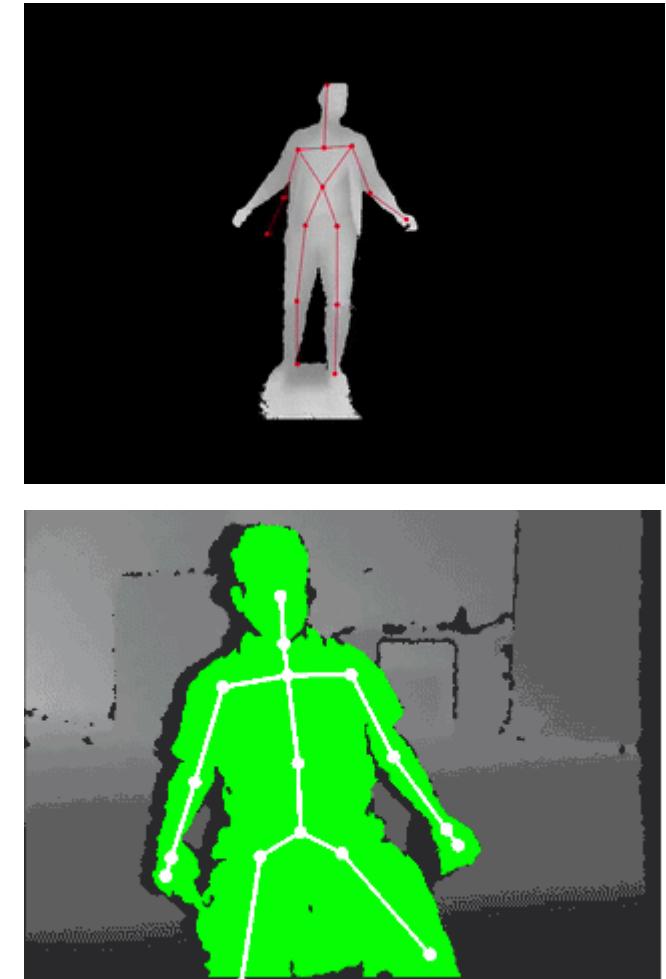
Extracción de pose - esqueletos

Las cámaras RGB+D fueron una de las primeras opciones para obtener datos de pose 3D fiables

Anteriormente, mediante visión monocular no era posible extraer datos de pose 3D de forma fiable o eficiente

Sin embargo, debido a la finalidad con la que se desarrollaron, estos algoritmos fallan para actividades ‘naturales’ no relacionadas con videojuegos. Fallan, por ejemplo:

- Si la persona no está en vista frontal
- Si no se ve completamente a la persona



Extracción de pose con RGB-D

Detectores que extraen esqueletos 2D directamente de una imagen

Pero también detectores que extraen poses 3D

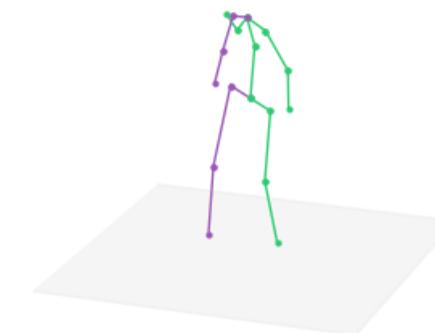
Sin embargo, no disponemos de la pose de la cámara (p.ej., distancia = tamaño de las personas)

Más rápidos, pueden ejecutarse (actualmente) en dispositivos móviles

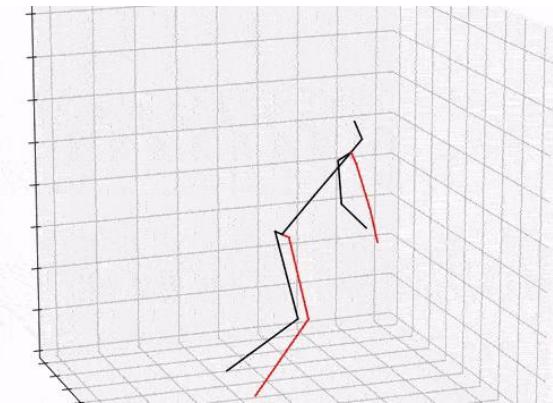
Ejemplos: OpenPose, MediaPipe, LCRNet++



<https://forknwork.wordpress.com/2018/02/14/openpose-3d-pose-baseline>



<https://victoriamazo.medium.com/3d-human-pose-estimation-ce1259979306>



Extracción de pose con RGB

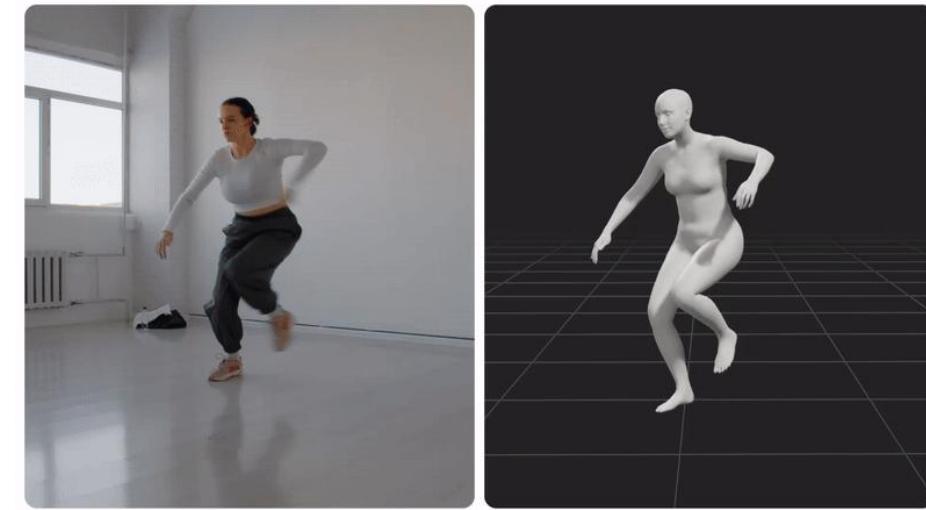
No solo extraen el esqueleto, sino una malla 3D que representa un modelo humano ([SMPL – Skinned Multi-Person Linear Model](#))

El modelo se adapta a la pose (parámetros alfa) y a la forma del cuerpo (parámetros beta)

Además, muchos de estos modelos aprenden los parámetros de la cámara a la vez

Más lentos (muchos más parámetros)

Ejemplos: HuMoR, WHAM, MoCapade



<https://medium.com/meshcapade/introducing-mocapade-3-0-1f9d8af42098>

Extracción de modelos 3D

Una vez extraídos los esqueletos, se pueden aprender de varias formas las acciones/actividades a partir de los datos de las articulaciones

Veamos ahora algunos ejemplos de reconocimiento de actividades mediante esqueleto:

- Anteriores al aprendizaje profundo
- Con aprendizaje profundo
- Ejemplos híbridos: vídeo-pose o vídeo-pose-objetos.

Reconocimiento de actividades mediante esqueletos

Una vez extraídos los esqueletos, se pueden aprender de varias formas las acciones/actividades a partir de los datos de las articulaciones

Veamos ahora algunos ejemplos de reconocimiento de actividades mediante esqueleto:

- Anteriores al aprendizaje profundo
- Con aprendizaje profundo
- Ejemplos híbridos: vídeo-pose o vídeo-pose-objetos.

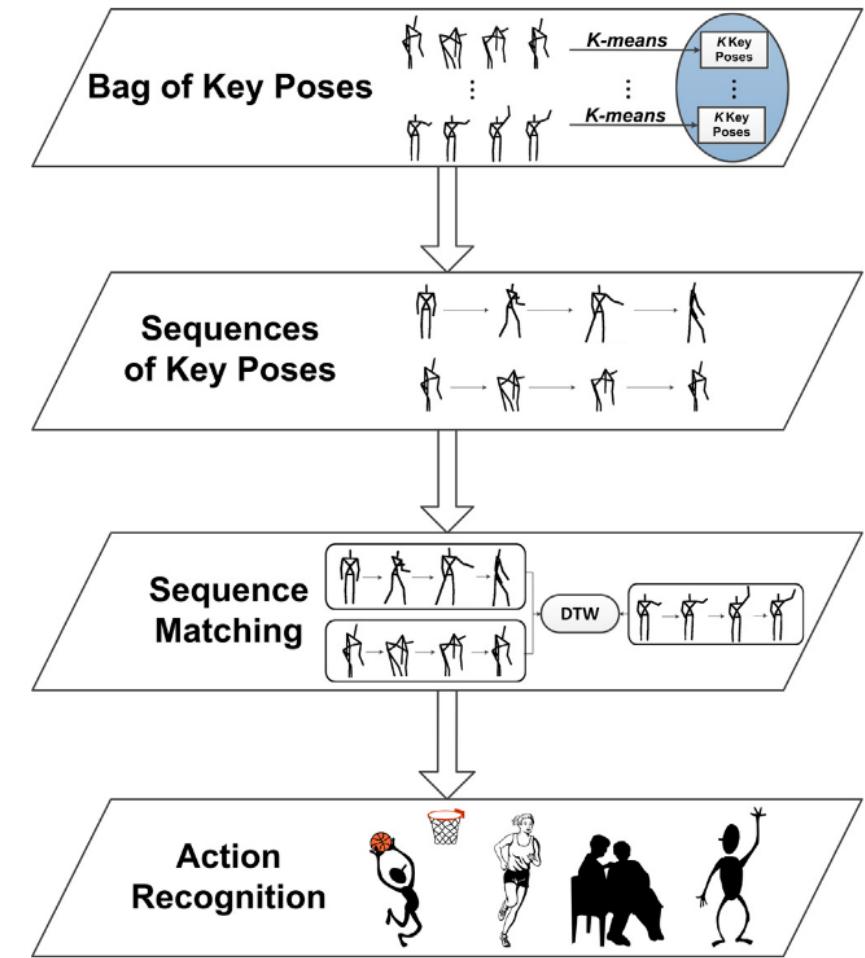
Reconocimiento de actividades mediante esqueletos

Anteriores al aprendizaje profundo

Dado que los datos de esqueleto tienen una dimensionalidad mucho más baja que los datos de vídeo, es posible aplicar modelos más sencillos para el reconocimiento de acciones/actividades

Por ejemplo, se pueden extraer, para cada acción las poses más representativas mediante clustering con k-means

Posteriormente, se puede aplicar la técnica de Bag-of-Words sobre los esqueletos de la acción a reconocer



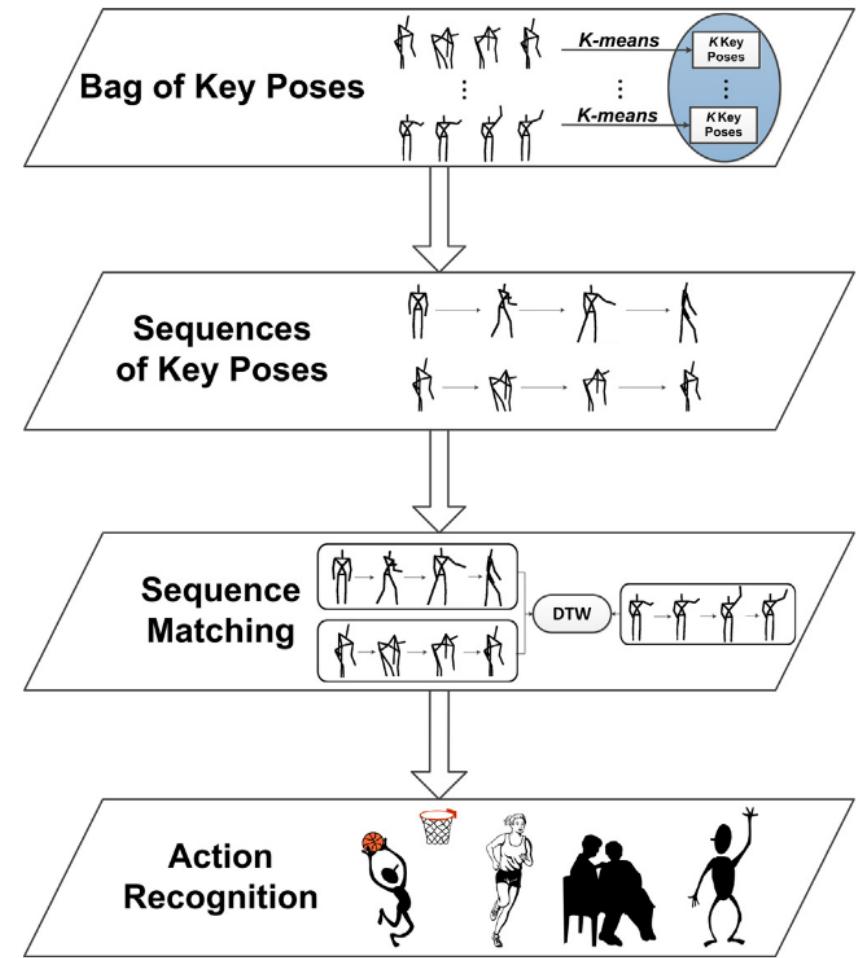
Chaaraoui, A. A., Padilla-López, J. R., Ferrández-Pastor, F. J., Nieto-Hidalgo, M., & Flórez-Revuelta, F. (2014). A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. *Sensors*, 14(5), 8895-8925.

Reconocimiento de actividades mediante esqueletos

Anteriores al aprendizaje profundo

Validación:

- Leave-One-Sequence-Out (LOSO)
- Leave-One-Actor-Out (LOAO)
- Podría haber sido con múltiples folds:
 - División del dataset en k subconjuntos del mismo tamaño (folds)
 - Entrenamiento y validación iterativos:
 - En cada iteración, se usa $k-1$ folds para entrenar y el fold restante para validar.
 - Se repite el proceso k veces.
 - Promedio de resultados.



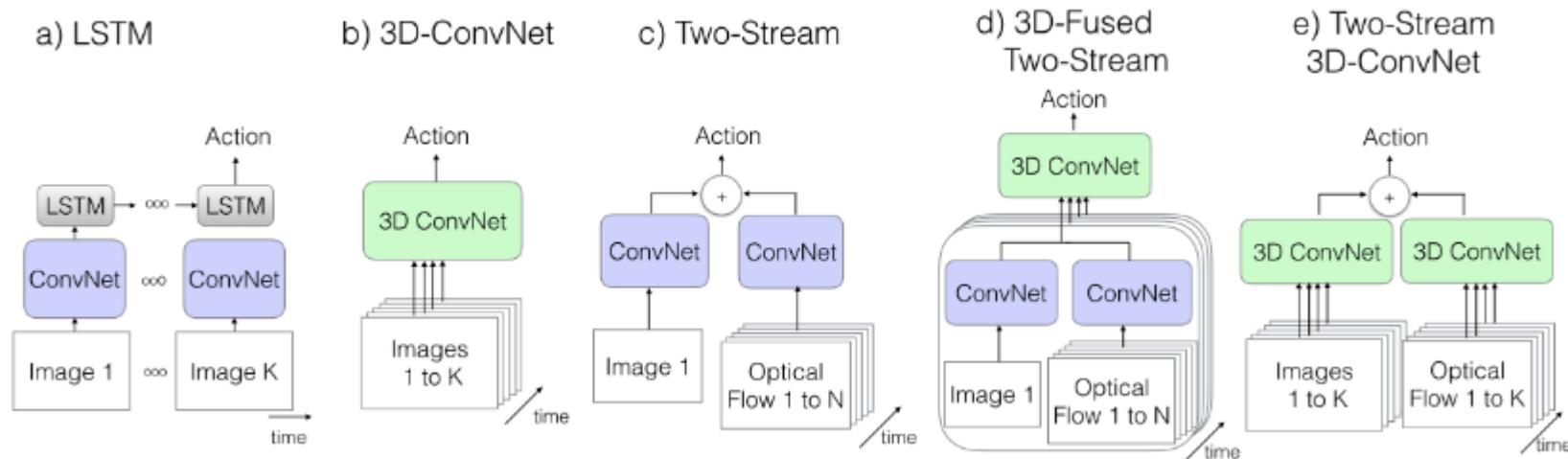
Chaaraoui, A. A., Padilla-López, J. R., Ferrández-Pastor, F. J., Nieto-Hidalgo, M., & Flórez-Revuelta, F. (2014). A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. Sensors, 14(5), 8895-8925.

Reconocimiento de actividades mediante esqueletos

Emplean imágenes RGB

La información de vídeo es más rica, contiene los objetos, y otros datos que se pierden en los esqueletos

Las CNN (2D) se pueden “ensanchar” hacia la tercera dimensión:



Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

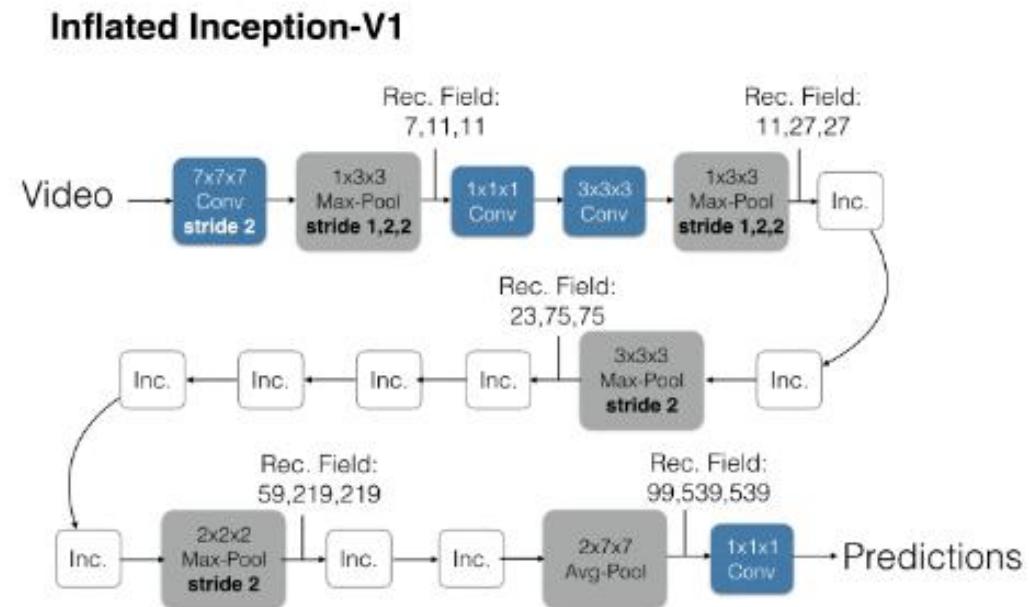
Deep learning – Redes convolucionales 3D

Las convoluciones y el resto de la red pasan de usar filtros 2D a 3D

La entrada, en vez de imágenes, ahora son volúmenes de vídeo

Sin embargo, el vídeo de entrada está limitado en su tamaño, ya que el volumen 3D está limitado por dimensiones XYT

Ejemplo: I3D



Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

Inspiración en los Transformers de lenguaje: Procesan secuencias de datos (como palabras o fotogramas) para capturar dependencias a largo plazo.

El vídeo se divide en pequeños fragmentos (espaciales y temporales), llamados tokens, que representan regiones en distintos momentos.

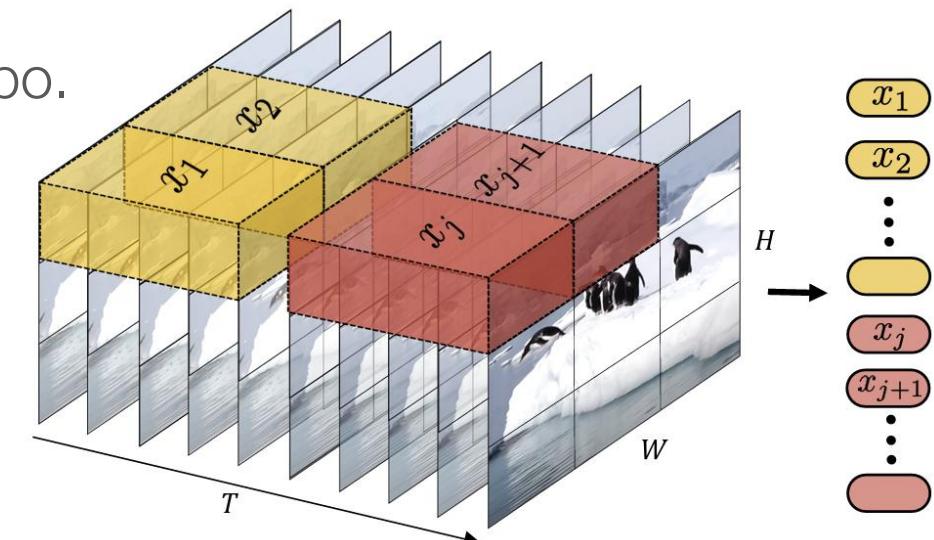
Self-Attention: Permite que el modelo aprenda qué partes del vídeo son más relevantes para una tarea (por ejemplo, reconocer una acción).

Capturan relaciones complejas entre espacio y tiempo.

Alto coste computacional y de memoria.

Requieren gran cantidad de datos para entrenarse.

Ejemplo: ViViT (Vision Video Transformer)



Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6836-6846).

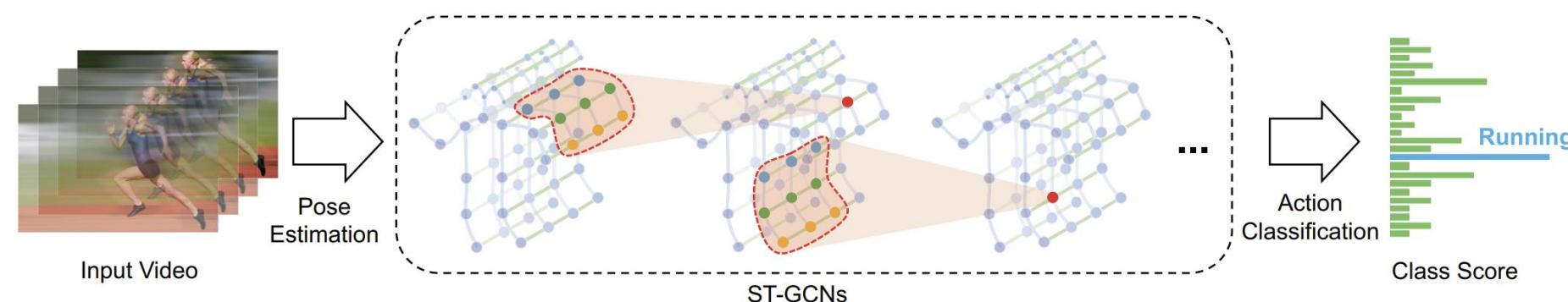
Deep learning – Video
Transformers

Permiten modelar relaciones espaciales y temporales entre articulaciones, capturando cómo se mueven de forma coordinada.

Entrada: Secuencias de esqueletos obtenidos de modelos como OpenPose o DensePose.

Cada capa de la GNN propaga información entre articulaciones conectadas, aprendiendo patrones estructurales del movimiento.

Ejemplo: ST-GCN



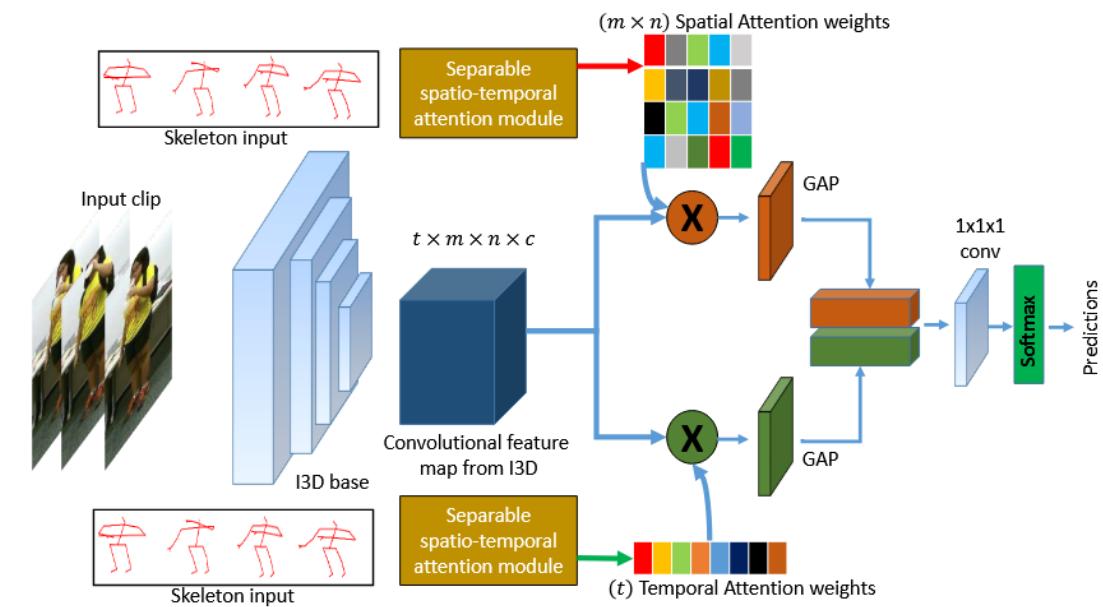
Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

Deep learning – Graph Neural Networks (GNNs)

Los esqueletos presentan poca información del contexto, por su simplificación

El procesamiento de vídeo se ve limitado por las dimensiones XYT de los clips

¿Por qué no proponer redes híbridas que combinen información de varias fuentes?



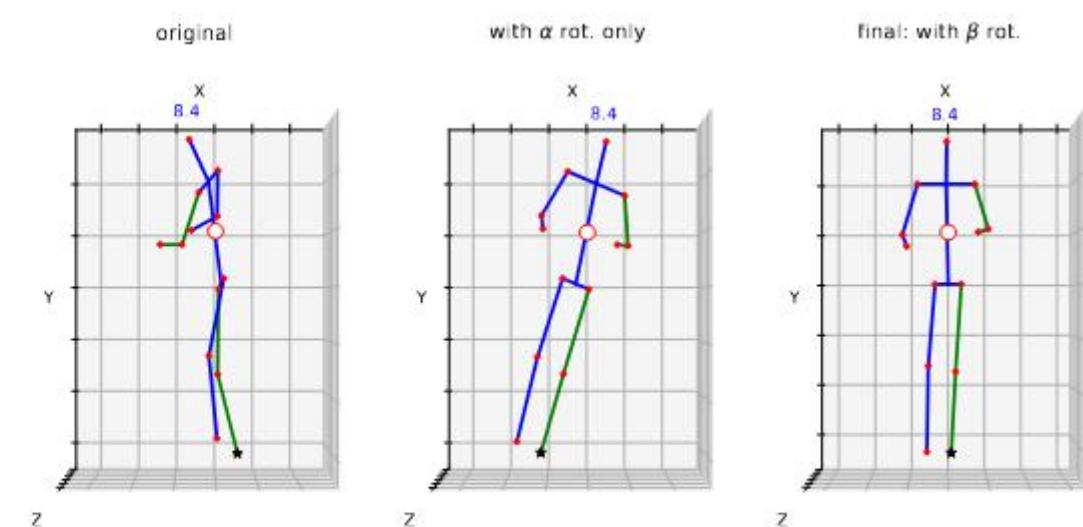
Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., & Francesca, G. (2019). Toyota smarthome: Real-world activities of daily living. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 833-842).

Redes híbridas vídeo-pose

Una forma de simplificar el aprendizaje para los modelos es hacer los esqueletos invariantes a la pose en el vídeo concreto



(a) Original RGB frame

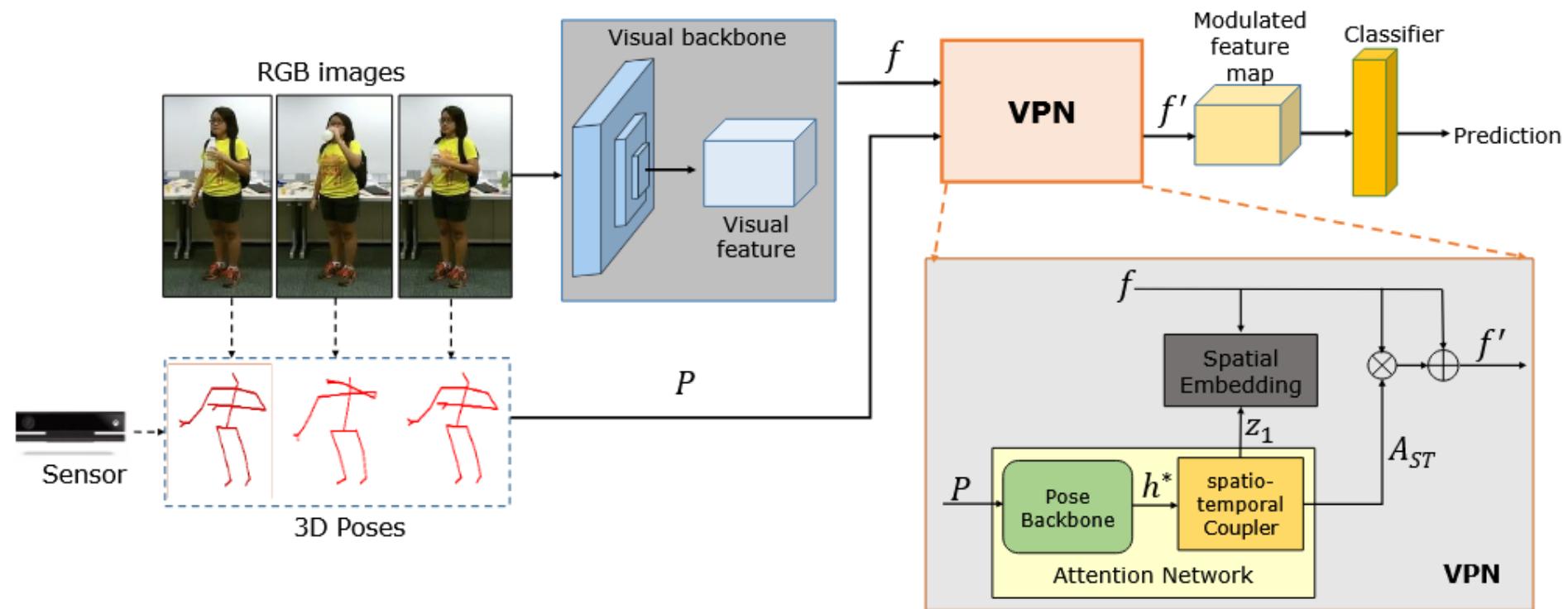


(b) Skeleton plots

Climent-Perez, P., & Florez-Revuelta, F. (2021). Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing. Sensors, 21(3), 1005.

Redes híbridas vídeo-pose

Igualmente, si los esqueletos están alineados con los fotogramas del vídeo, se puede explotar de mejor manera la combinación de datos aportados al modelo neuronal



Das, S., Sharma, S., Dai, R., Bremond, F., & Thonnat, M. (2020, August). Vpn: Learning video-pose embedding for activities of daily living. In European conference on computer vision (pp. 72-90). Cham: Springer International Publishing.

Redes híbridas vídeo-pose

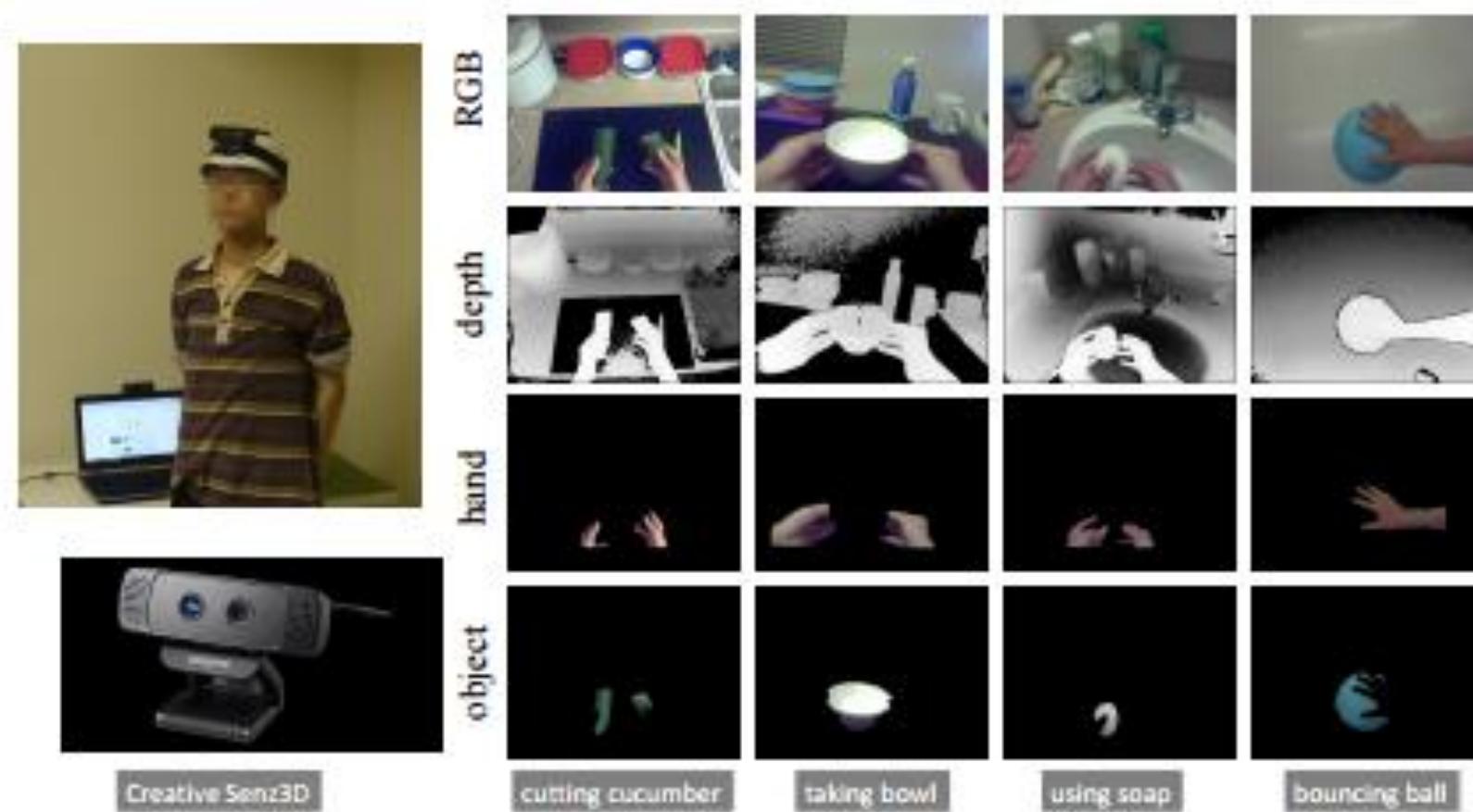
La interacción con objetos concretos puede marcar la diferencia en la clasificación de actividades similares

- Por ejemplo “beber de una taza” vs “beber de una botella”

Prestar atención a los objetos manipulados puede ser determinante para mejorar el reconocimiento de la actividad concreta

La información del contexto, y en concreto la atención a los objetos circundantes puede ser crucial

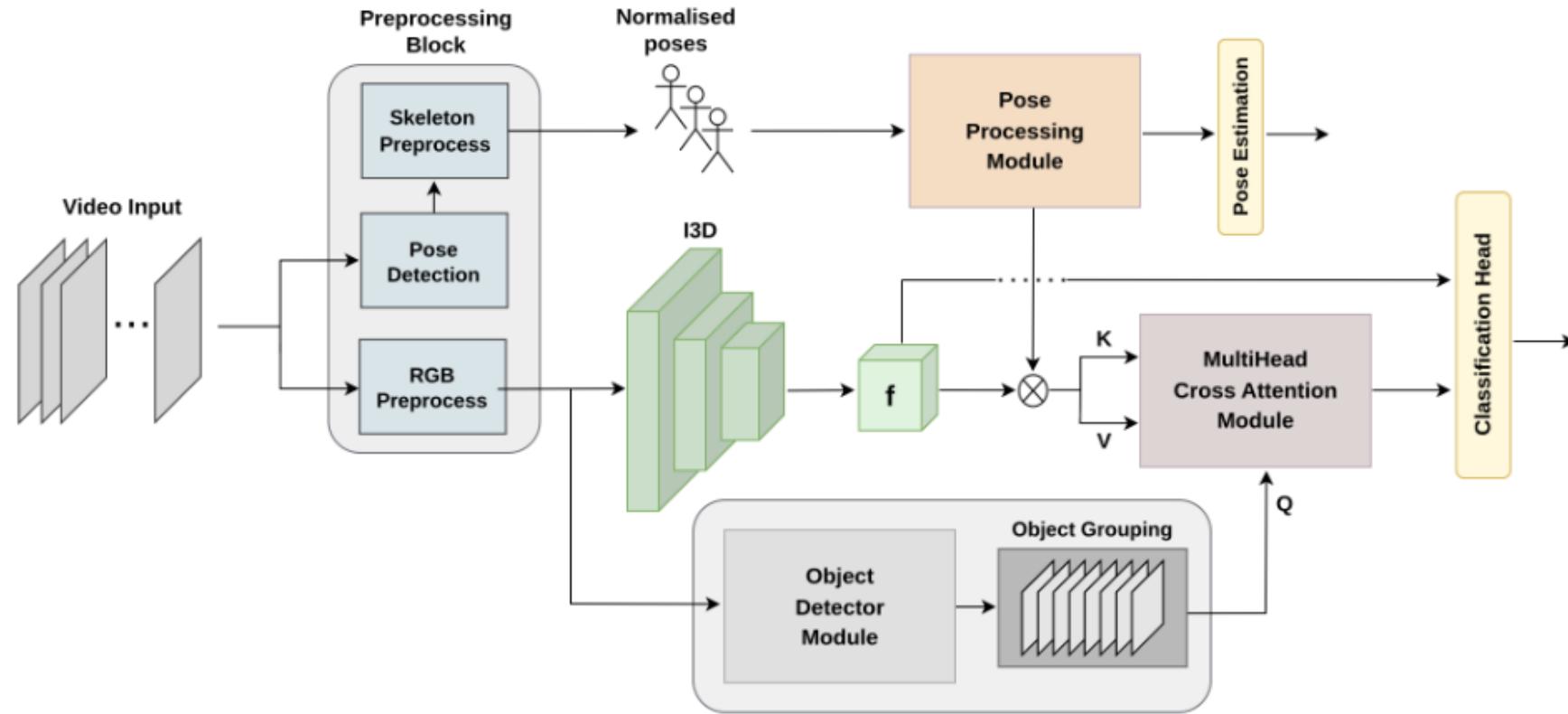
Esto ya se viene haciendo en visión egocéntrica



Wan, S., & Aggarwal, J. K. (2015). Mining discriminative states of hands and objects to recognize egocentric actions with a wearable RGBD camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 36-43).

Redes híbridas
vídeo-pose-objeto

En esta red de ejemplo, se incorpora un módulo de atención a los objetos.



Hashemifard, K., Climent-Perez, P., & Florez-Revuelta, F. (2025). Recognition of Daily Activities through Multi-Modal Deep Learning: A Video, Pose, and Object-Aware Approach for Ambient Assisted Living – to be submitted

Redes híbridas
vídeo-pose-objeto

Las vistas laterales son la forma más común de captura de imágenes

Las vistas alternativas, sin embargo, cuentan con varias ventajas:

- Las vistas cenitales especialmente con lentes fisheye (ojo de pez) permiten capturar una habitación entera con una sola cámara, y evitan en gran medida las occlusiones causadas por muebles u otras personas
- Las vistas egocéntricas, permiten capturar lo que se está haciendo delante de las manos, que son, al fin y al cabo, con las que realizamos muchas de las tareas de la vida diaria.

Sin embargo, la mayoría de los modelos no funcionan con estas vistas.



(a) Original image



(b) Textured avatar output (DensePose)

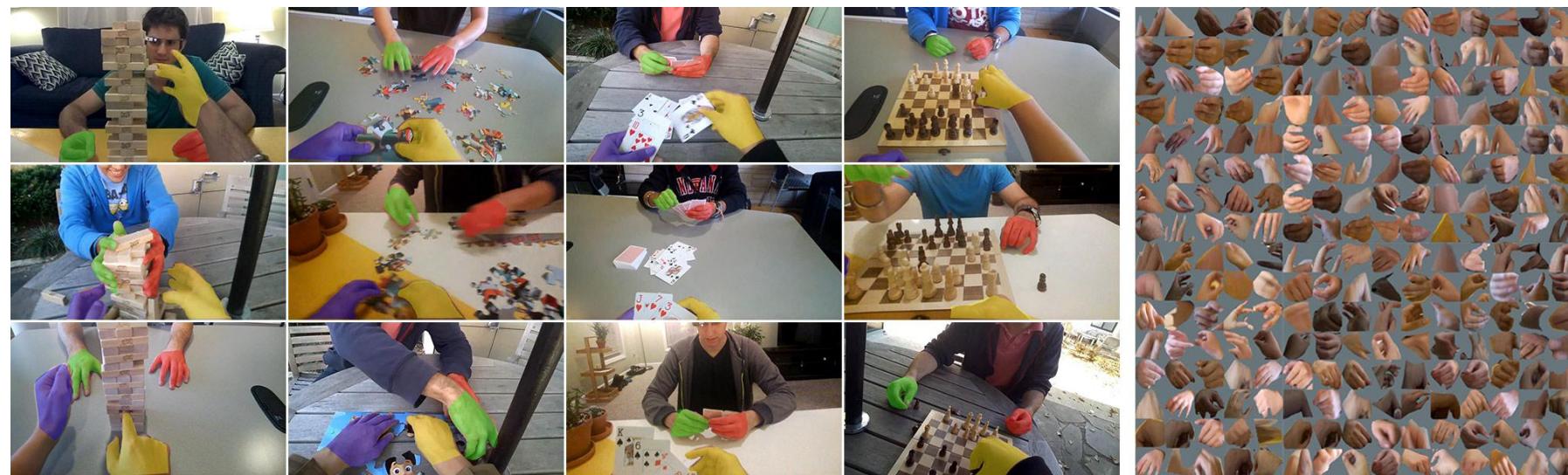


(c) Segmentation provided by Mask R-CNN

Son necesarios modelos específicos para estas vistas alternativas

En visión egocéntrica, se requiere la detección de las manos (y los antebrazos), pues con ellas se realizan la mayoría de las tareas que se observan frente a este tipo de cámara

Existen conjuntos de datos específicos para estimar la pose de las manos en visión egocéntrica (ej. [EgoHands](#))



Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the IEEE international conference on computer vision (pp. 1949-1957).

Visión egocéntrica

Las dificultades mencionadas, han llevado a crear grandes repositorios y conjuntos de datos para el análisis desde esta modalidad de vídeo

Por ejemplo, Ego4D y Epic Kitchens



Visión egocéntrica

Epic Kitchens - características

Grabaciones en primera persona en cocinas reales

Actividades no guionadas: los participantes graban su día a día

55 h de vídeo (~11.5 M de fotogramas)

39.6k segmentos de acción anotados

454k fotogramas con objetos anotados

Anotaciones narradas por los propios participantes → reflejan la intención real

Retos:

- Reconocimiento de acciones (verbo + objeto)
- Anticipación de acciones futuras
- Detección de objetos interactivos

Versiones ampliadas: EPIC-Kitchens-100 y EPIC Fields (con info 3D)

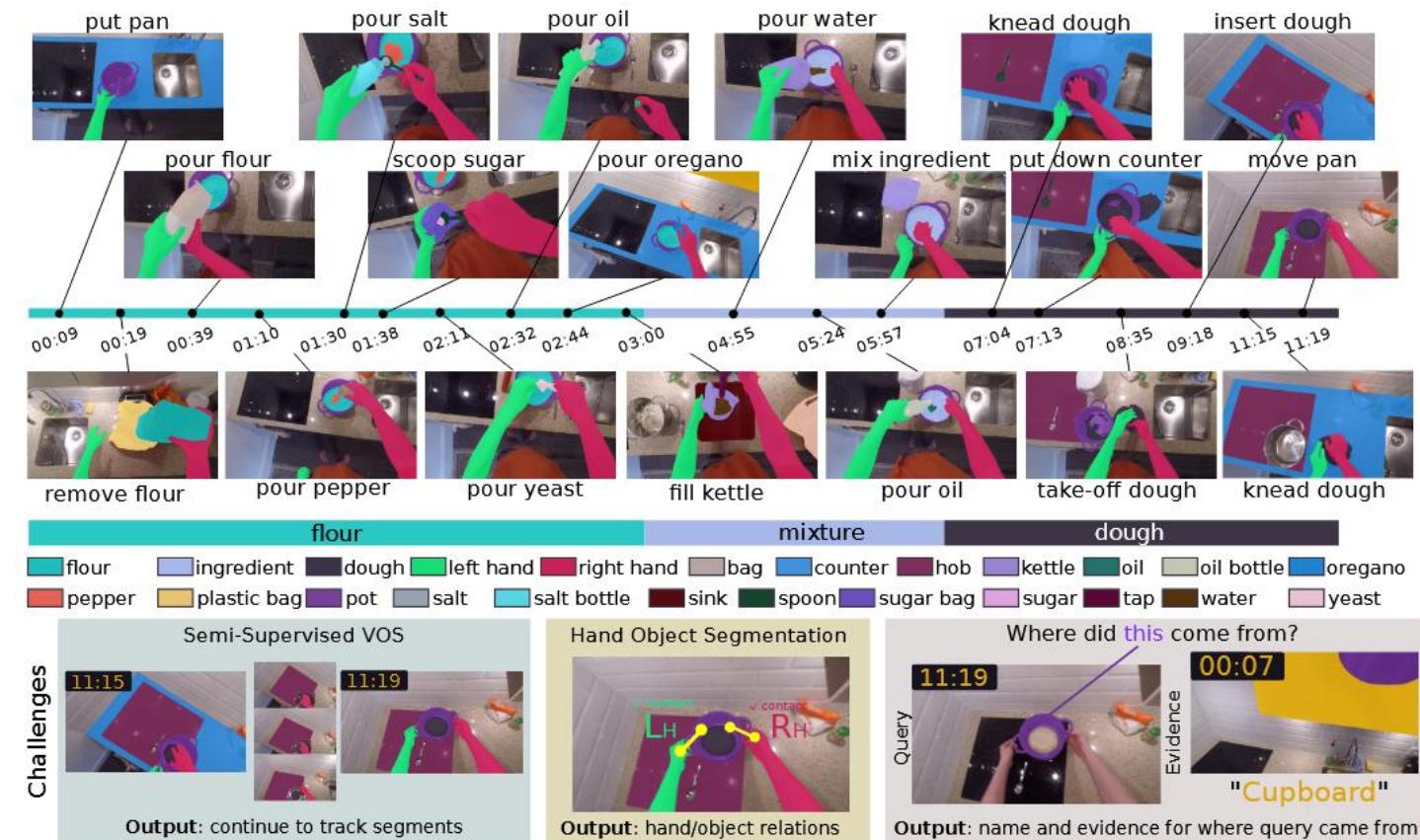
Visión egocéntrica

Epic Kitchens - ¿qué se puede hacer?

Se pueden elaborar modelos de reconocimiento de actividades complejas a lo largo de, por ejemplo, 11 minutos

Esto gracias a haber pedido a los participantes una narración

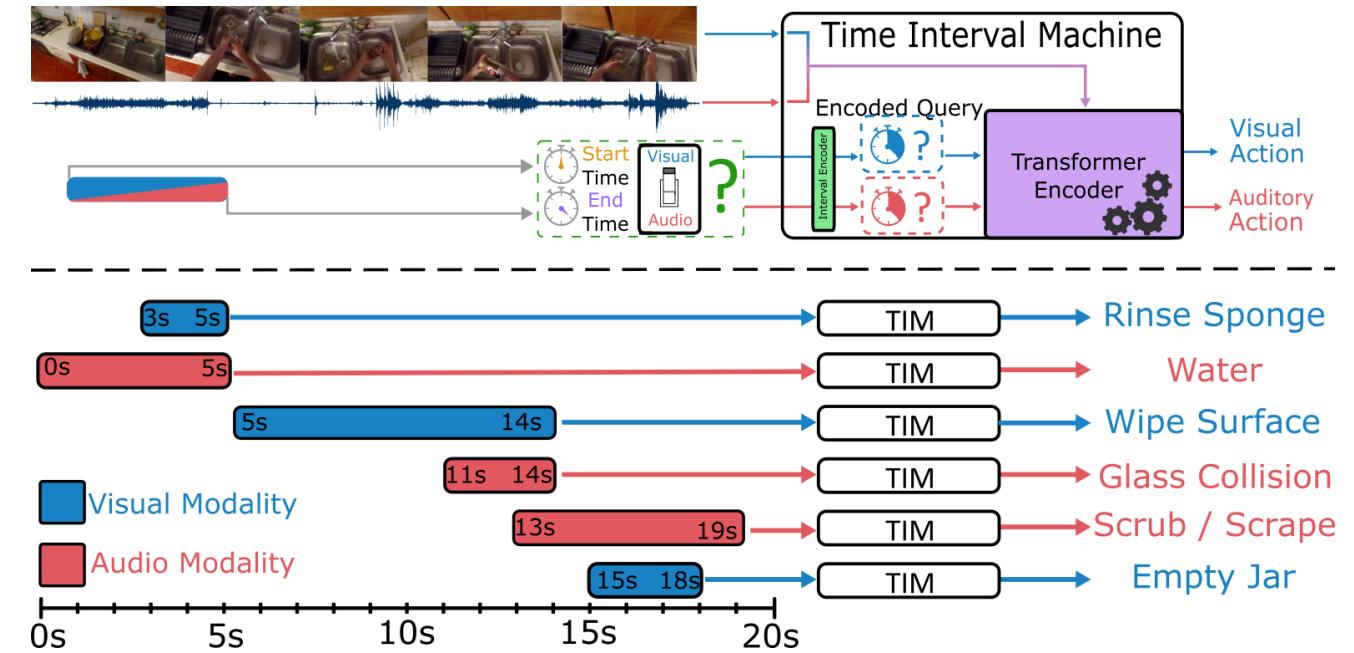
Las narraciones se emplean como fuente de ground truth o etiquetado semi-supervisado



Además de lo ‘visible’ en el vídeo (no siempre la cámara enfoca hacia la actividad) ...

... está lo invisible, pero audible,

Existen también modelos en visión egocéntrica que combinan ambas modalidades



Chalk, J., Huh, J., Kazakos, E., Zisserman, A., & Damen, D. (2024). Tim: A time interval machine for audio-visual action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18153-18163).

Visión egocéntrica + audio

Interacción persona-máquina

Unidad 4. Interfaces para reconocimiento de movimiento basadas en sensores y vídeo

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Interacción persona-máquina

Unidad 4. Interfaces para reconocimiento de movimiento basadas en sensores y vídeo

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Reconocimiento de movimientos,
acciones y actividades mediante con
otros sensores

Reconocimiento de movimientos,
acciones y actividades mediante con
otros sensores

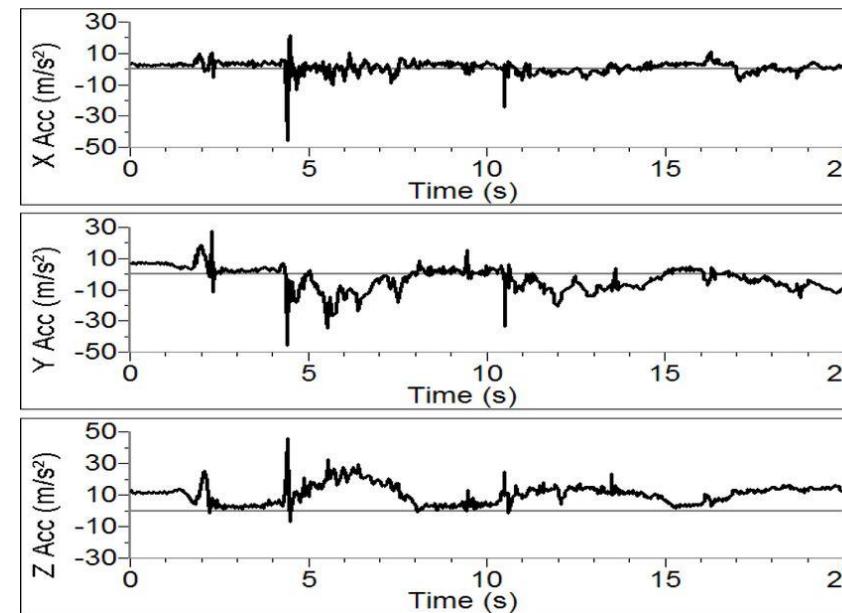
Unidades de Medición Inercial

Sensores que miden aceleración lineal en uno o varios ejes (X, Y, Z).

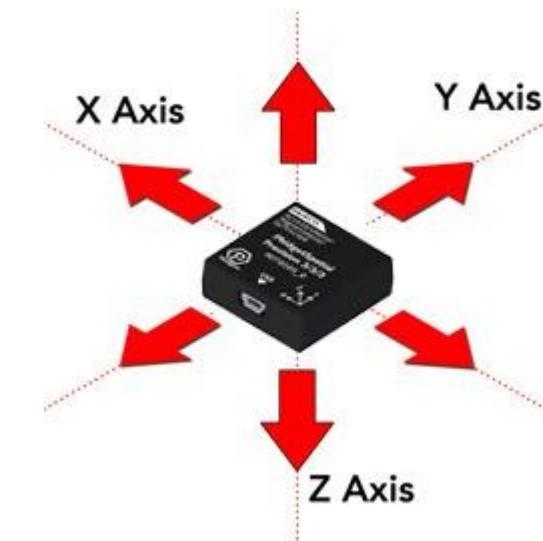
Detectan cambios de velocidad y dirección en el espacio.

Permiten inferir orientación relativa respecto a la gravedad.

Usos típicos: contar pasos, detectar caídas, orientación de pantallas.



<https://www.verniercanada.ca/product/sensors/accelerometers/3-axis-accelerometer>



https://www.phidgets.com/docs/Accelerometer_Guide

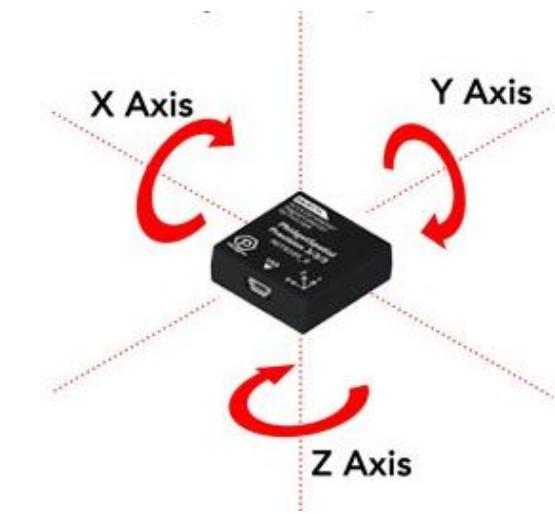
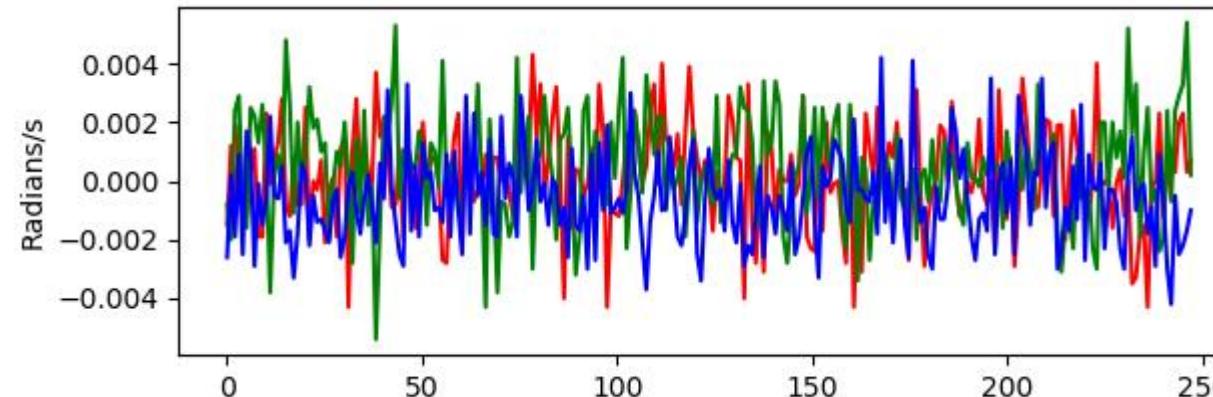
Acelerómetros

Sensores que miden la velocidad angular (rotación) en torno a los ejes.

Capturan movimientos de giro y orientación dinámica.

Complementan al acelerómetro para detectar rotaciones rápidas.

Usos típicos: estabilización de cámaras, control de drones, VR/AR.



https://www.phidgets.com/docs/Gyroscope_Guide

<https://learn.adafruit.com/adafruit-sensorlab-gyroscope-calibration/gyroscope-calibration>

Giroscopios

Acelerómetros: sensibles al ruido y vibraciones, no distinguen bien entre inclinación y movimiento lineal.

Giroscopios: acumulan error con el tiempo (drift).

Ninguno por sí solo ofrece una medida completa y estable de la orientación.

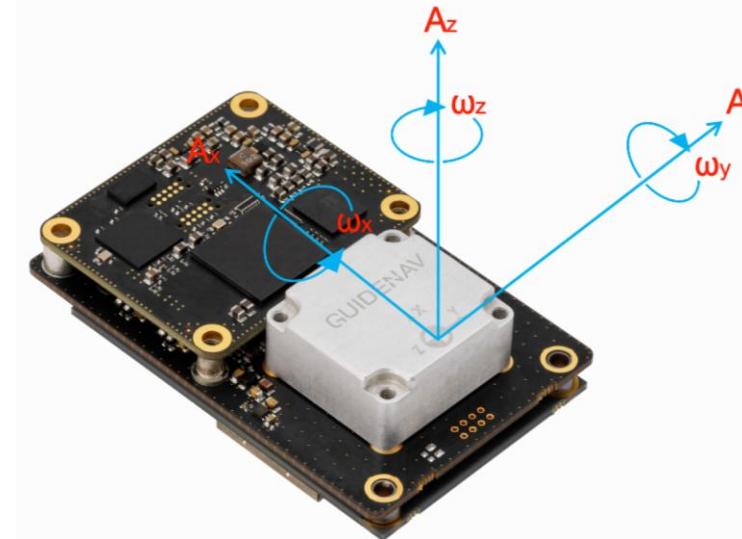
Limitaciones

Unidad de Medición Inercial (IMU, Inertial Measurement Unit).

Integran acelerómetros + giroscopios (y a veces magnetómetros).

Permiten obtener posición y orientación más precisas.

Reducen las limitaciones de cada sensor por separado.



[https://guidenav.com/es/comprendiendo-las-aplicaciones-basicas-de-la-unidad-de-medida-inercial-imu-y-tendencias-futuras./](https://guidenav.com/es/comprendiendo-las-aplicaciones-basicas-de-la-unidad-de-medida-inercial-imu-y-tendencias-futuras/)

Unidad de Medición Inercial

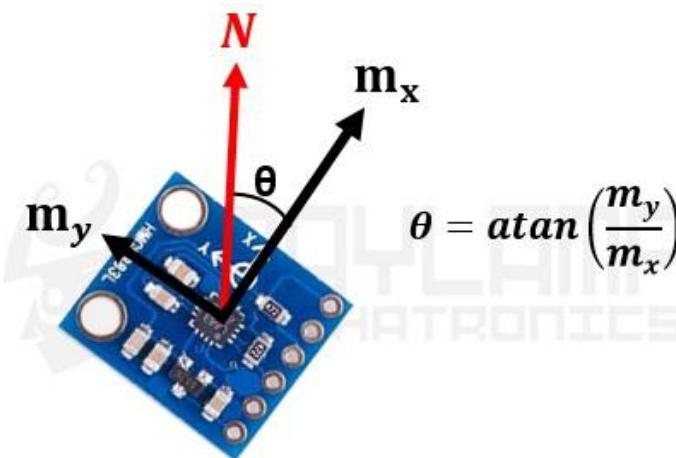
Magnetómetro en las IMUs

Mide el campo magnético terrestre, como una brújula.

Proporciona una referencia absoluta de orientación.

Al fusionarse con acelerómetro y giroscopio:

- Corrige errores acumulados (drift).
- Ofrece datos de posición y orientación más estables.



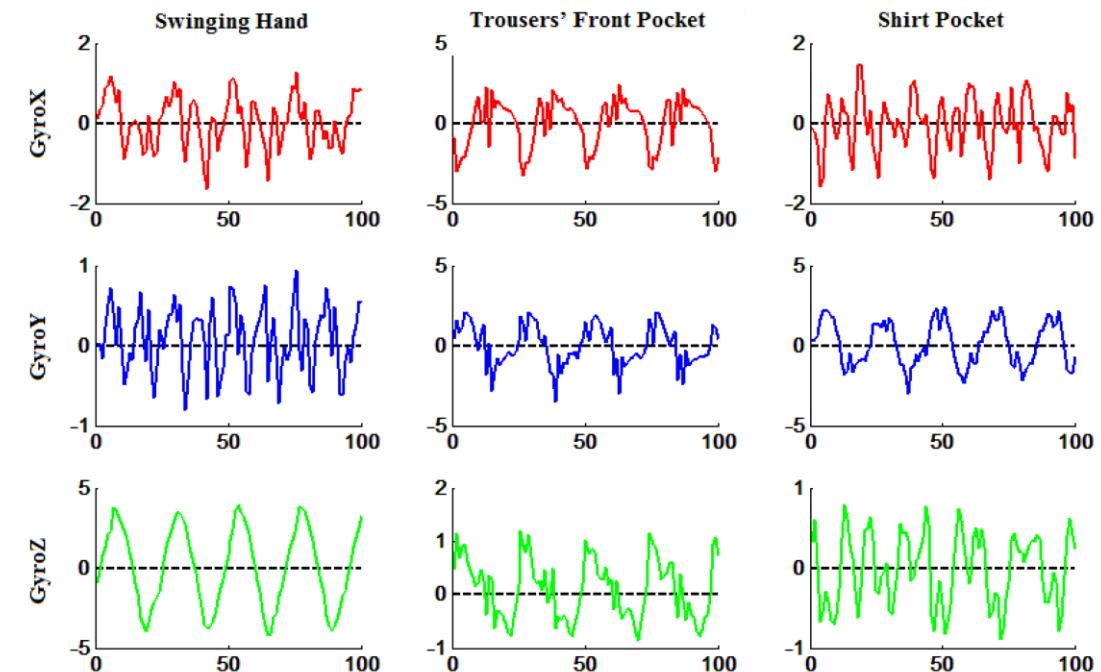
https://naylampmechatronics.com/blog/49_tutorial-magnetometro-hmc5883l.html

Acelerómetro → mide aceleración lineal.

Giroscopio → mide rotación (velocidad angular).

IMU → combinación de ambos (y magnetómetro) para datos más robustos.

Aplicaciones: smartphones, wearables, videojuegos, rehabilitación, realidad virtual, robótica.



Kang, X., Huang, B., & Qi, G. (2018). A novel walking detection and step counting algorithm using unconstrained smartphones. Sensors, 18(1), 297.

Aplicaciones

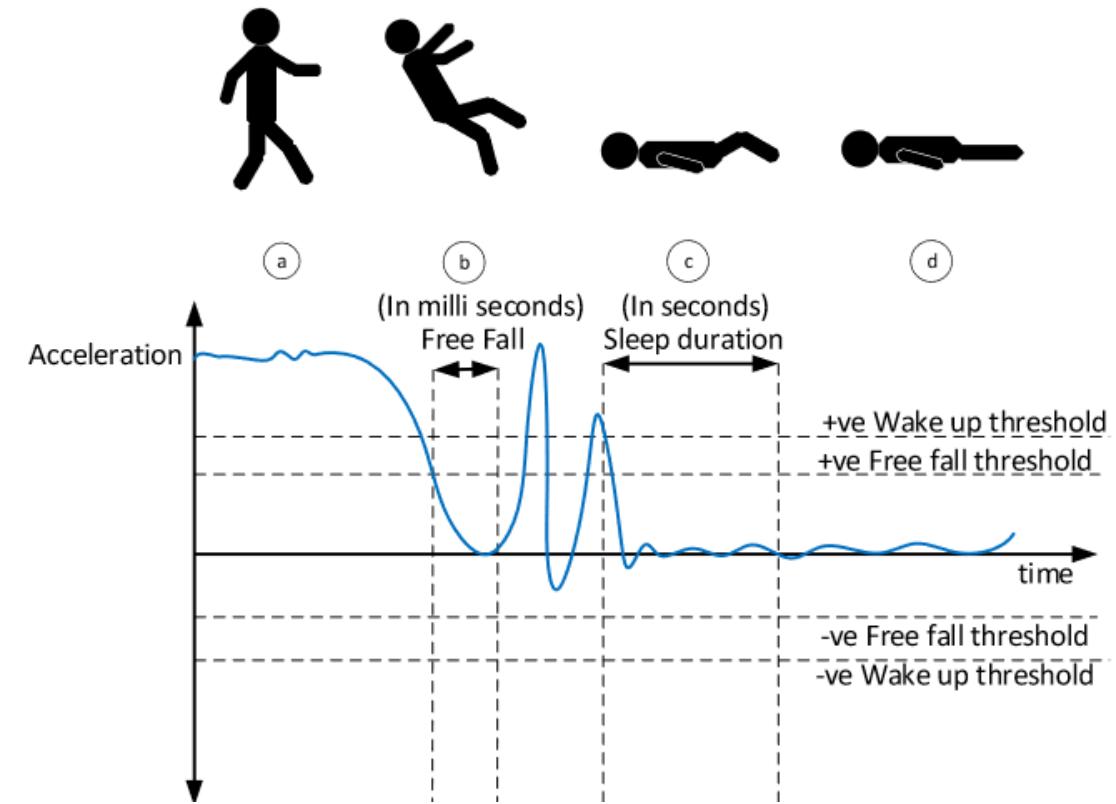
Detección de caídas

Una de las aplicaciones más comunes

Integrada en productos de consumo

Se puede detectar el momento de la “caída libre”

Problema: falsos positivos (caída de sensor, sin el usuario)



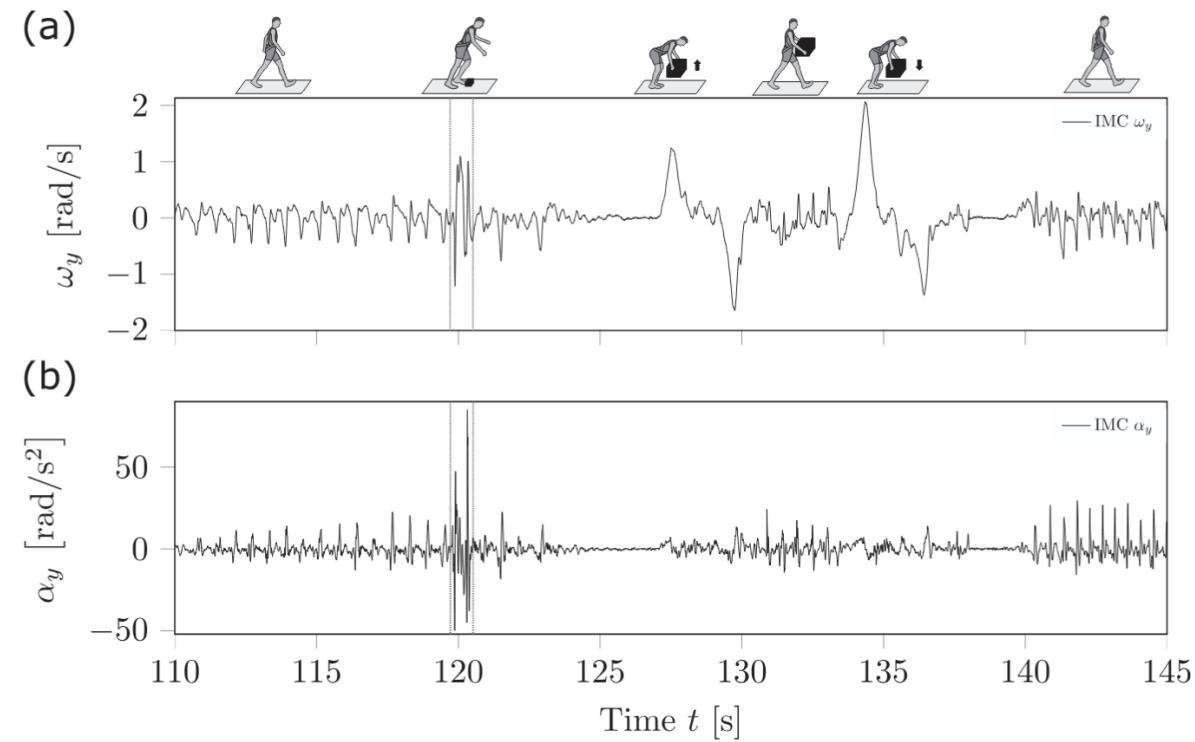
<https://www.linkedin.com/pulse/human-fall-detection-using-3-axis-mems-acceleration-sensor-ramu>

Detección de tropiezos y (casi) caídas

Con las mejoras en los métodos, se pretende no solo detectar las caídas, sino diferenciar las actividades normales que suponen agacharse (más espaciadas en el tiempo) de las caídas o los tropiezos que no causan caídas (casi-caídas)

Camino hacia la PREVENCIÓN de caídas

Gießler, M., Werth, J., Waltersberger, B., & Karamanidis, K. (2024). A framework to automatically detect near-falls using a wearable inertial measurement cluster. Communications Engineering, 3(1), 181.

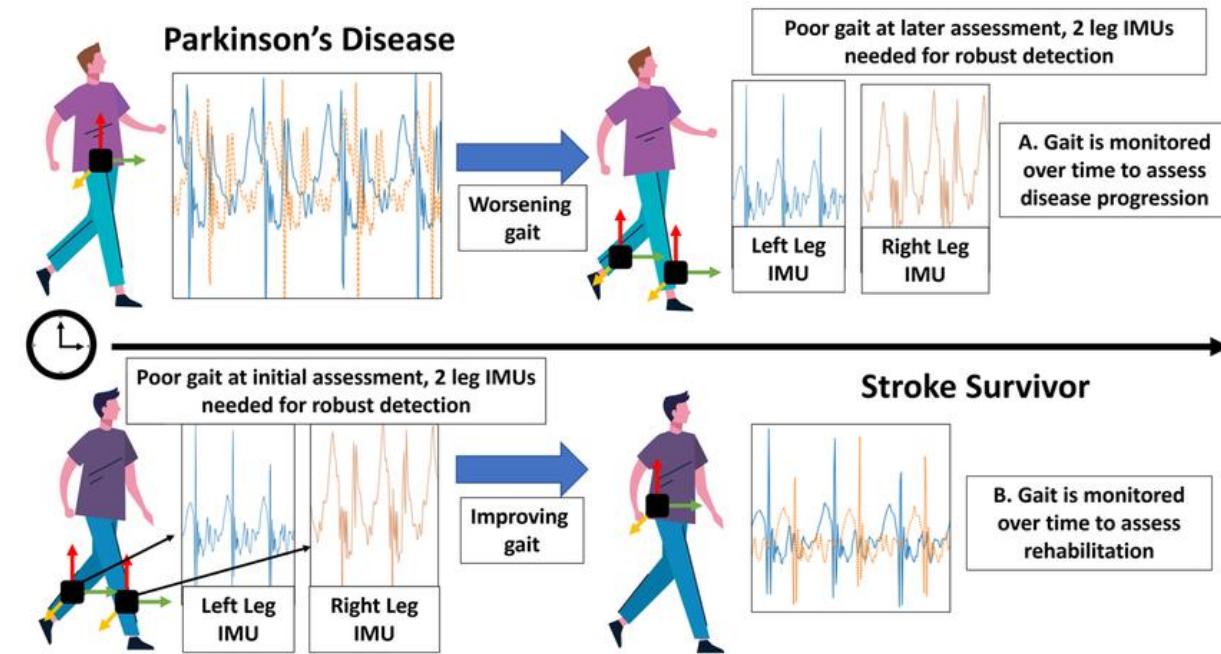


Aplicaciones

Monitorización continua o rehabilitación

El lugar en el que se coloca o número de IMUs necesarios es relevante en función de la aplicación

Localización o número puede variar a lo largo del tiempo



Young, F., Mason, R., Morris, R. E., Stuart, S., & Godfrey, A. (2023). IoT-enabled gait assessment: The next step for habitual monitoring. Sensors, 23(8), 4100.

Detección del “freezing of gait” en Parkinson

Síntoma motor característico de la enfermedad de Parkinson.

Episodios breves en los que la persona no puede iniciar o continuar la marcha, como si los pies “se pegaran al suelo”.

Se desencadena con frecuencia en situaciones concretas:

Al comenzar a andar.

Al girar o atravesar espacios estrechos (puertas, pasillos).

Bajo estrés o presión temporal.

Aumenta el riesgo de caídas y la pérdida de autonomía.

Es un foco clave de estudio de la evolución del Parkinson



<https://www.youtube.com/watch?v=CCD9ncoDf90>

Detección del “freezing of gait” en Parkinson

Las IMUs (acelerómetros + giroscopios + magnetómetros) permiten medir movimiento en tiempo real.

Se colocan habitualmente en piernas, pies o cintura para registrar la marcha.

Detectan patrones anómalos de aceleración y rotación que caracterizan el “freezing of gait”

Posibilitan algoritmos automáticos de:

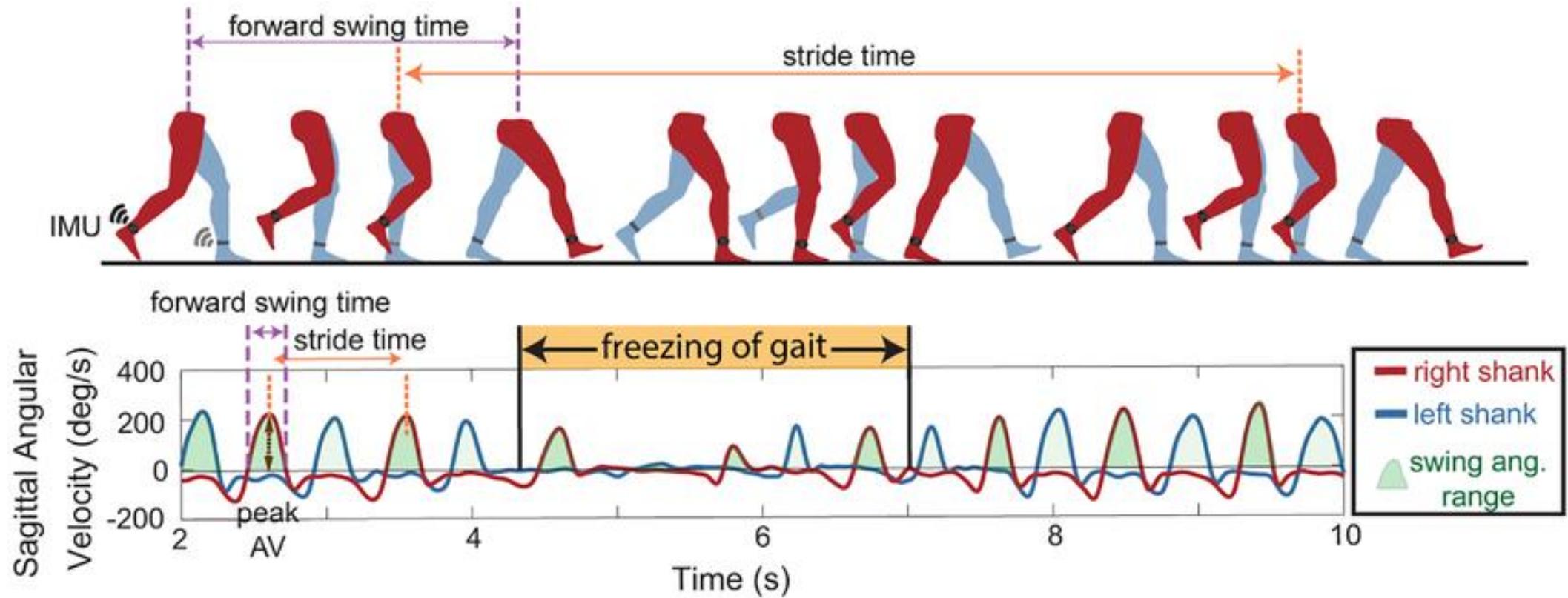
- Detección de episodios de bloqueo.
- Cuantificación de la severidad y frecuencia.
- Seguimiento remoto de pacientes en su vida diaria.

Herramienta clave para desarrollar sistemas de asistencia y rehabilitación personalizados.



<https://www.youtube.com/watch?v=CCD9ncoDf90>

Detección del “freezing of gait” en Parkinson



O'Day, J., Syrkin-Nikolau, J., Anidi, C., Kidzinski, L., Delp, S., & Bronte-Stewart, H. (2020). The turning and barrier course reveals gait parameters for detecting freezing of gait and measuring the efficacy of deep brain stimulation. PLoS One, 15(4), e0231984.

Aplicaciones

Reconocimiento de actividades

Mediante pulseras multi-sensor como la Empatica E4

Se trata de una pulsera para uso experimental

Existen otras con formato y tamaño más reducido, así como productos de consumo

La persona puede llevarla en la mano dominante



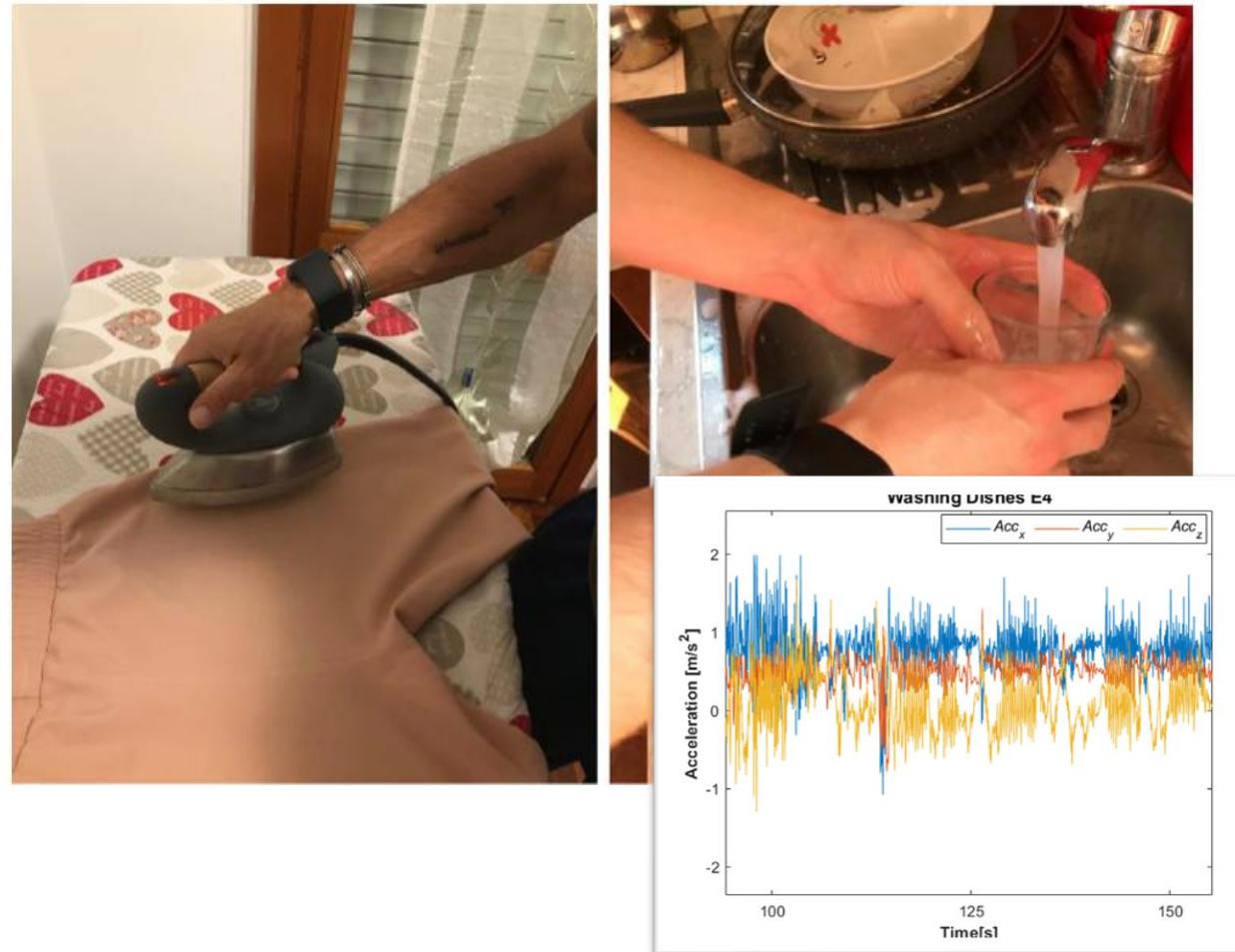
Aplicaciones

Reconocimiento de actividades

En los experimentos se les pidió a 52 participantes realizar una serie de tareas de la vida diaria

Se realizaron 24 actividades diferentes

Para cada una, los participantes realizaron una media de 5 repeticiones de cada actividad



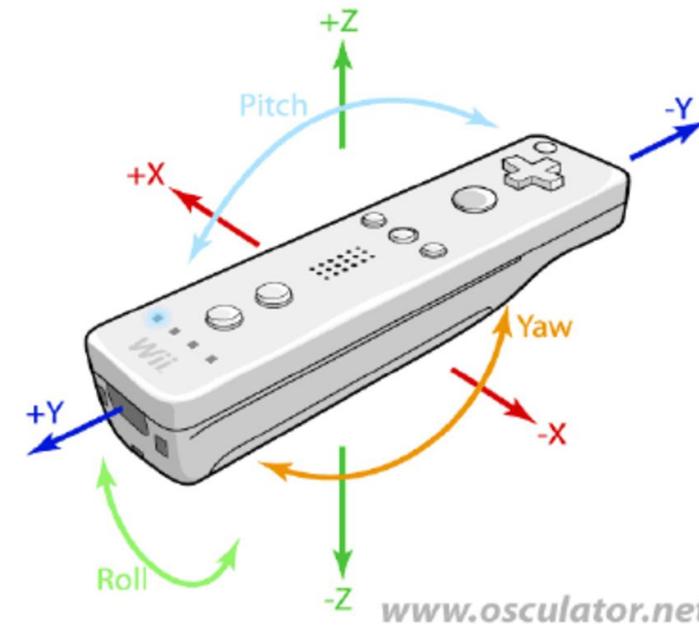
Climent-Pérez, P., Muñoz-Antón, Á. M., Poli, A., Spinsante, S., & Florez-Revuelta, F. (2022). Dataset of acceleration signals recorded while performing activities of daily living. Data in Brief, 41, 107896.

Aplicaciones

Juegos serios para rehabilitación

Los mandos de algunas consolas permiten también la interacción mediante acelerómetros

Permiten comprobar la realización de ejercicios pautados por un/a fisioterapeuta para la rehabilitación



Aplicaciones

Juegos serios para rehabilitación

Ejemplos:

- Rehabilitación tras un ictus
- Rehabilitación en la hemiparesia
- Estimulación cognitiva en demencia

<https://www.neurorehabdirectory.com/virtual-rehabilitation-via-nintendo-wii-and-conventional-physical-therapy-effectively-treat-post-stroke-hemiparetic-patients>

<https://neura.edu.au/news-media/media-releases/rehabilitate-wii-help-stroke-survivors>

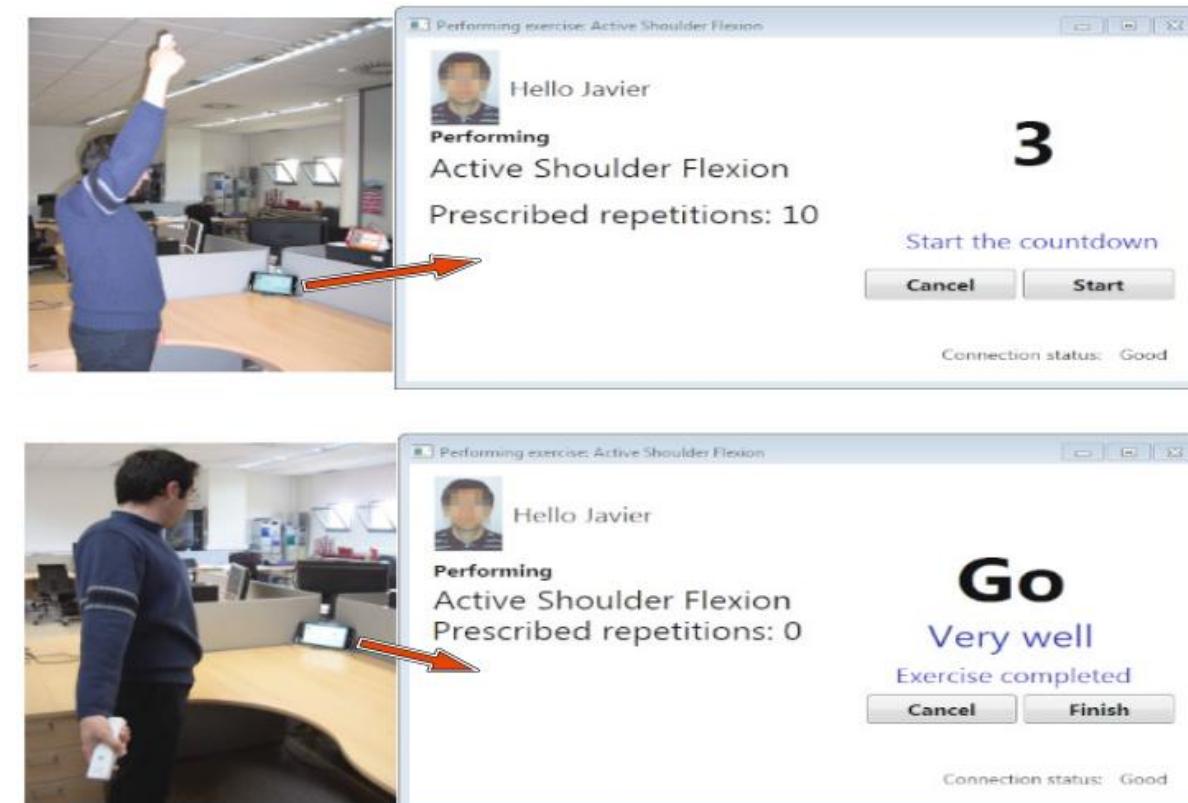
<https://www.capecodtimes.com/story/lifestyle/health-fitness/2007/10/18/wii-game-system-becomes-new/52767545007>



Aplicaciones

Juegos serios para rehabilitación

Ejemplo: detección de la flexión de hombro realizada de forma correcta según lo prescrito



García, J., Ruiz, D., Soriano, A., Marín, O., Hernández, S., & Ferrairó, S. (2012, December). eFisioTrack: A Telerehabilitation Platform for Monitoring Prescribed Therapeutic Exercises in Orthopaedic Injuries. In International Workshop on Ambient Assisted Living (pp. 423-430). Berlin, Heidelberg: Springer Berlin Heidelberg.

Aplicaciones

Reconocimiento de movimientos,
acciones y actividades mediante con
otros sensores

Sensores en el entorno

Espacios equipados con sensores y actuadores.

IPM ya no solo persona-máquina → también persona-entorno.

El entorno interpreta acciones y se adapta automáticamente.

Ejemplo básico: luces que se encienden al entrar en la habitación.

Ámbitos de aplicación:

- Domótica residencial: confort y ahorro energético.
- Salud y envejecimiento activo: monitorización de AA.V.D.
- Entornos laborales: ergonomía y seguridad, mejora de la gestión de espacios.
- Educación y ocio: espacios interactivos.

Sensores:

- Portables (wearables): viajan con el usuario (relojes, pulseras, IMUs).
- Desplegados: instalados en el entorno (PIR, contactos magnéticos, luz).

Combinación con IA e IoT para aprendizaje de hábitos

Estimación de comportamientos a largo plazo.

Mayor naturalidad de la interacción.

Importancia de la **computación ubicua**.

Sensores de movimiento (PIR)

Detectan presencia mediante infrarrojos pasivos.

Muy usados en iluminación automática.

Limitación: no detectan presencia estática.



<https://www.aqara.com/en/product/human-motion-sensor>

Tipos de sensores

Sensores de puertas y ventanas

Contactos magnéticos que detectan aperturas y cierres.

Permiten inferir actividades: cocinar, salir de casa, etc.

Complemento para sistemas de seguridad.



<https://www.amazon.ca/Aqara-11LM-MOCGQ-Window-Sensor/dp/B07D37VDM3>

Tipos de sensores

Sensores ambientales

Miden temperatura, humedad, presión atmosférica.

Relevantes para confort, climatización y salud.

También usados en sistemas de alerta (incendios, fugas).



<https://eu.aqara.com/es-eu/products/aqara-detector-de-humo>



<https://www.ikea.com/es/es/p/vindstyrka-sensor-calidad-aire-inteligente-00498231>

Tipos de sensores

Sensores de luminosidad

Detectan intensidad lumínica en cada estancia.

Permiten ajustar persianas, iluminación y consumo energético.

Suelen combinarse con sensores de movimiento PIR (PIR+luz)



<https://www.aqara.com/en/product/human-motion-sensor>

Tipos de sensores

Sensores acústicos

Micrófonos para detectar patrones de sonido.

Ejemplos: caída de un objeto, grito de ayuda, llanto de bebé.

Riesgos de privacidad → anonimización del audio necesaria

Tipos de sensores

Sensores de consumo eléctrico

Medición de corriente en electrodomésticos.

Permite inferir qué aparato se usa → actividad asociada.

Ejemplo: hervidor encendido = preparar té.



<https://www.ikea.com/es/es/p/tretak-enchufe-macho-inteligente-80540349>

Tipos de sensores

Sensores conectados en red.

Comunicación mediante Wi-Fi, Zigbee, Bluetooth, LoRa.

Hacia la visión de un hogar conectado.

Limitaciones actuales:

- Interoperabilidad entre marcas y protocolos.
- Dependencia de conexión a internet.
- Problemas de seguridad y hackeo: Bluetooth se ha demostrado vulnerable, ...

Interacción persona-máquina

Unidad 4. Interfaces para reconocimiento de movimiento basadas en sensores y vídeo

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Interacción persona-máquina

Unidad 5.1 – Interfaces hápticas

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Interacción mediante interfaces
hápticos

Definición: interacción con máquinas a través del sentido del tacto.

Informa mediante vibraciones, fuerzas, texturas simuladas.

Permite enriquecer la experiencia de usuario más allá de la vista y el oído.

Tiene aplicaciones en robótica, VR/AR y accesibilidad.

Tipos de feedback haptico:

- **Vibratorio:** motores que generan patrones de vibración (móviles, mandos).
- **De presión:** actuadores que simulan contacto o resistencia.
- **De fuerza:** dispositivos que aplican fuerzas físicas (joysticks, exoesqueletos).
- **De textura:** ultrasonidos o microactuadores para simular superficies.

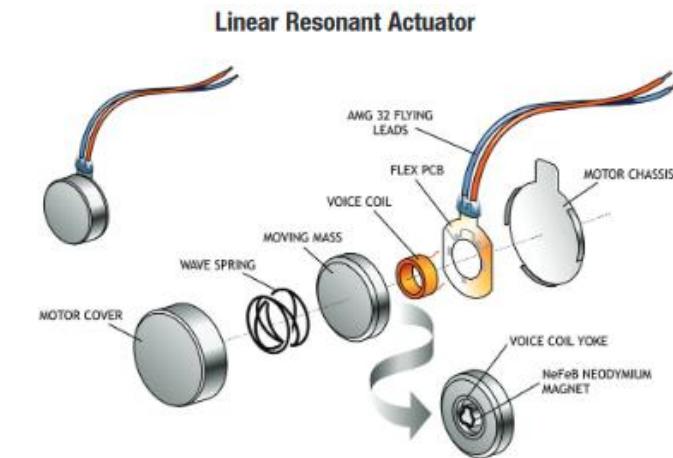
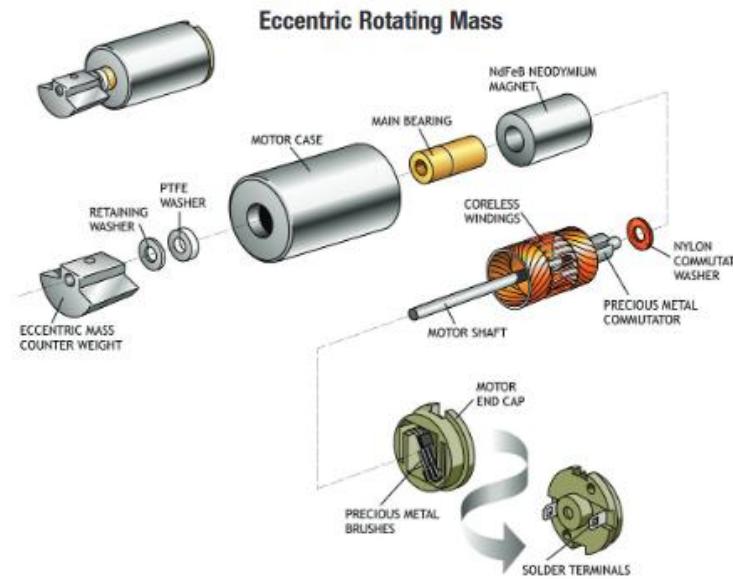
Retroalimentación haptica

Motores de vibración (ERM, LRA): los más comunes.

Piezoeléctricos: precisos, de respuesta rápida.

Ultrasónicos: permiten feedback en el aire, sin contacto físico.

Guantes hápticos: recrean presión y movimiento en dedos/manos.



Images courtesy of Precision Microdrives

Ejemplos de actuadores
hápticos

Smartphones: vibraciones diferenciadas para notificaciones.

Relojes inteligentes: feedback discreto mediante golpecitos.

Pulseras de fitness: alertas hápticas para objetivos de actividad.

Ventaja: alta aceptación social y disponibilidad masiva.



<https://www.wsj.com/articles/feel-the-phone-buzz-in-your-pocket-it-could-be-all-in-your-head-11576773050>

Interfaces hapticos en móviles y wearables

Guantes y trajes hapticos: añaden inmersión en VR/AR.

Simulan contacto con objetos virtuales.

Aplicaciones: videojuegos, formación, medicina.

Importante en la IPM: reduce la barrera entre lo virtual y lo físico.



<https://skarredghost.com/2021/02/02/haptx-gloves-price-availability-future>

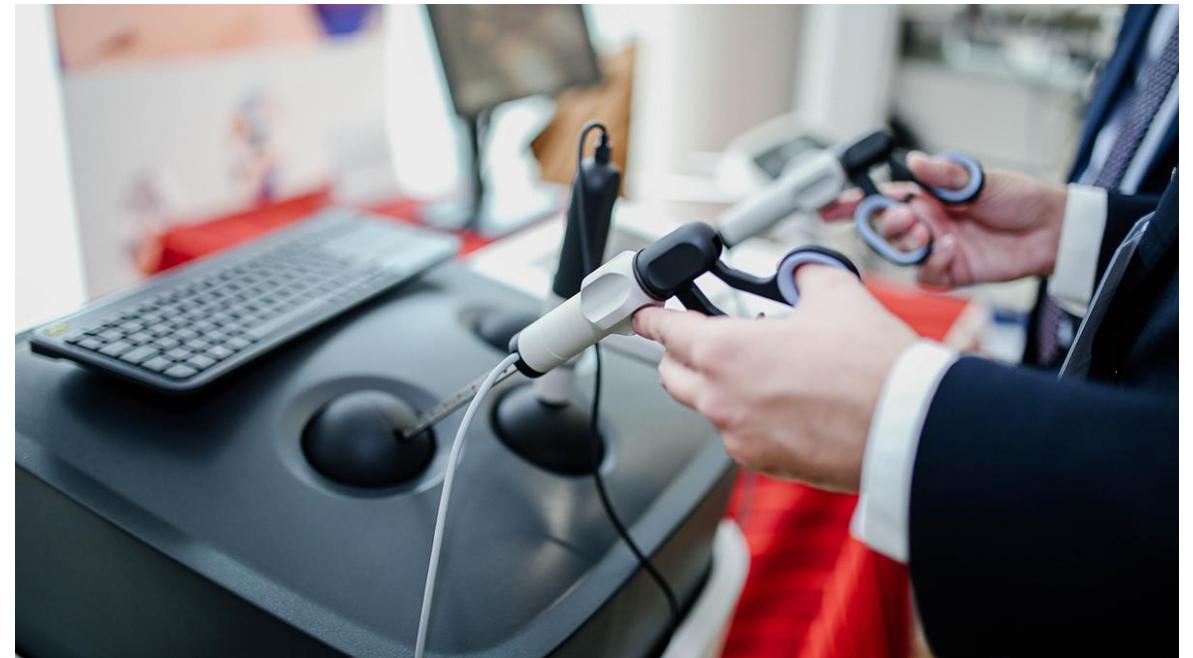
Interfaces hapticos para VR/AR

Telecirugía: feedback táctil para operaciones a distancia.

Teleoperación industrial: control remoto con precisión.

Robótica colaborativa (interacción persona-robot): seguridad mediante detección táctil.

<https://www.automate.org/robotics/blogs/improvements-in-robot-assisted-surgery-driven-by-haptic-feedback-systems>



Aplicaciones hápticas emergentes

El tacto humano tiene umbrales de detección específicos (fuerza, vibración).

Es sensible a frecuencia, intensidad y localización del estímulo.

Los dispositivos deben adaptarse a estos límites.

La percepción háptica está ligada a procesos cognitivos (atención, memoria).

Limitaciones de la tecnología:

- Resolución táctil baja → no siempre simula texturas reales.
- Latencia: retrasos afectan a la sensación de naturalidad.
- Consumo energético: limita portabilidad.
- Coste elevado en sistemas avanzados (guantes, trajes).

Limitaciones

Interacción persona-máquina

Unidad 5.1 – Interfaces hápticas

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Interacción persona-máquina

Unidad 5.2 – Tecnologías asistivas

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Sustituye el concepto de “discapacidad” → foco en capacidades y formas diversas de interactuar.

Reconoce diferencias en visión, audición, movilidad y cognición.

Objetivo en ingeniería: diseñar tecnologías que potencien la autonomía.

Plantea un marco más positivo e inclusivo.

Diversidad funcional

Principio: productos y servicios accesibles a todas las personas.

Relacionado con el concepto de Diseño para todos.

Implica interfaces intuitivas, adaptables y compatibles con distintos usuarios.

Evita barreras tecnológicas y fomenta la inclusión digital.

Accesibilidad universal

Recursos que facilitan la vida diaria de personas con diversidad funcional.

Apoyos tradicionales: muletas, sillas de ruedas, lupas.

Apoyos tecnológicos avanzados:

- Exoesqueletos para movilidad y rehabilitación.
- Sistemas hápticos para interacción táctil.
- Seguimiento ocular para comunicación.
- Prótesis inteligentes con sensores.

Representan un salto cualitativo en accesibilidad.

Productos de apoyo

Tecnologías específicas para personas con discapacidad visual.

Ejemplos:

- Lectores de pantalla con síntesis de voz.
- Líneas braille electrónicas dinámicas.
- Sistemas hápticos y de orientación espacial.
- Aplicaciones móviles con reconocimiento de objetos/texto.

La tecnología transforma el acceso a la información.

Tiflotecnología

La IA (ML/DL) permite apoyos que se adapten al usuario en tiempo real.

Ejemplos:

- Exoesqueletos que aprenden el patrón de marcha.
- Sistemas de seguimiento ocular con predicción de palabras.

Mejora precisión y eficiencia en los dispositivos de apoyo (mejor usabilidad).

Abre la puerta a tecnologías personalizadas y evolutivas (más adaptabilidad).

Fuertemente relacionados con el concepto de la interacción persona-máquina y persona-entorno.

La IPM en la discapacidad

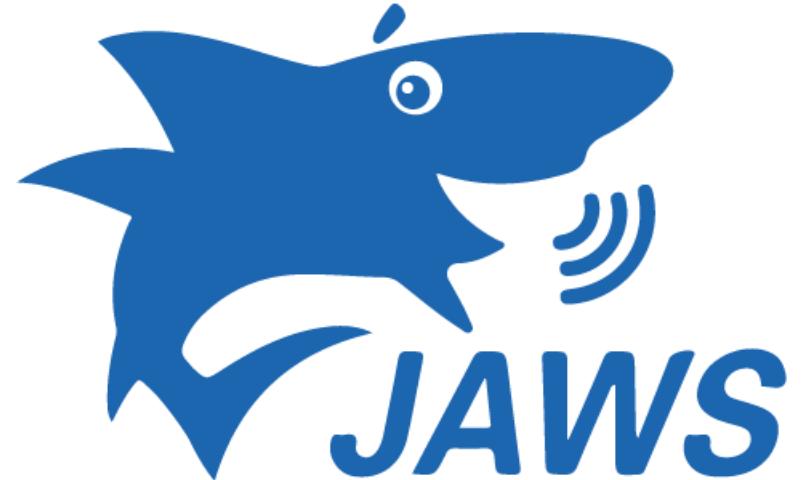
Tiflotecnología

Software que convierte texto en síntesis de voz.

Ejemplos: NVDA, JAWS, VoiceOver.

Uso extendido en ordenadores, móviles y tablets.

IPM: Facilitan navegación web y uso de aplicaciones.



Lectores de pantalla

Dispositivos que traducen texto digital a braille dinámico.

Se conectan por USB o Bluetooth a ordenadores/móviles.

Vuelta a la Interacción persona-máquina “clásica”
(terminal de líneas de texto)



Lineas Braille electrónicas

“Aunque el Blitab se parece a otros dispositivos tipo tableta, su superficie de vidrio superior está perforada en una cuadrícula con orificios que permiten a la tecnología basada en líquido presentar el contenido en braille.”



<https://www.nytimes.com/2018/09/03/arts/tablet-devices-blind-braille.html>

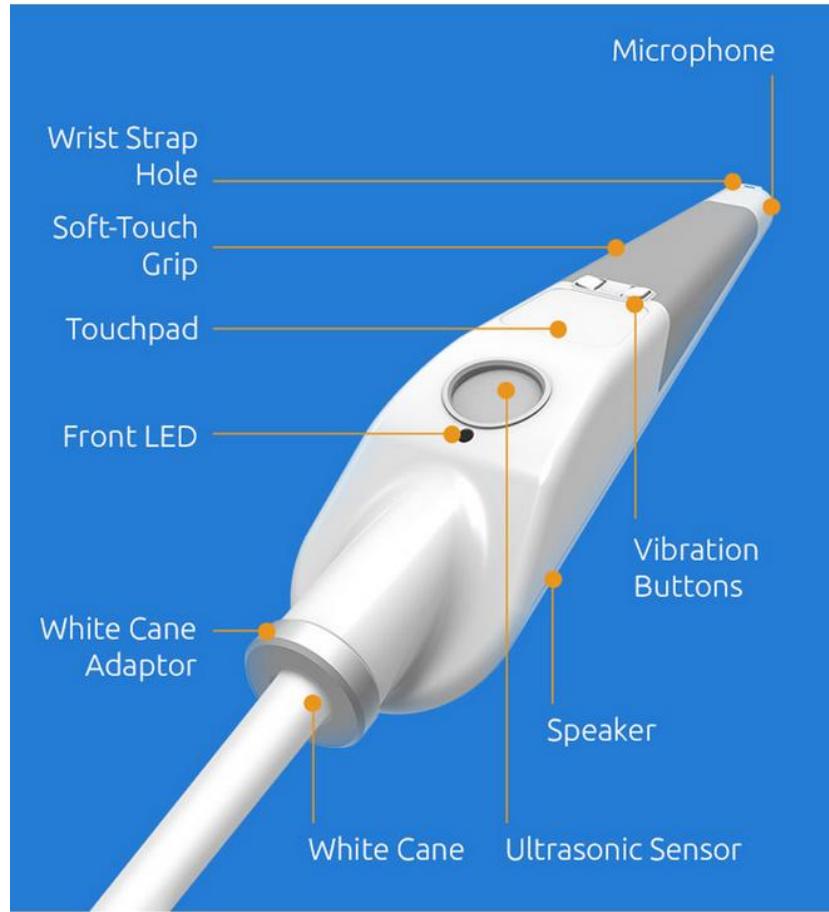
Tabletas Braille

Evolución del bastón tradicional → integran sensores ultrasónicos o láser.

Detectan obstáculos aéreos y en el suelo.

Alertan con vibración o sonidos.

Ejemplo: SmartCane, WeWALK.



<https://wewalk.io>

Bastones electrónicos

Gafas que permiten apoyar a personas con problemas de visión

Algunas permiten leer textos, reconocer personas, identificar productos, ...



<https://www.orcam.com/en-gb/orcam-myeye-3-pro>



<https://www.eyenews.uk.com/reviews/tech-reviews/post/oxsight-onyx-a-sight-enhancing-wearable>

Gafas inteligentes

QRs de lectura ultrarrápida a distancia

Desarrollado en la UA



<https://www.murciadiario.com/articulo/empresas/los-cereales-kellogs-primer-a-empresa-del-mundo-en-incorporar-la-tecnologia-murciana-de-naviglens/20210705235702053660.html>



<https://museos.navilens.com>

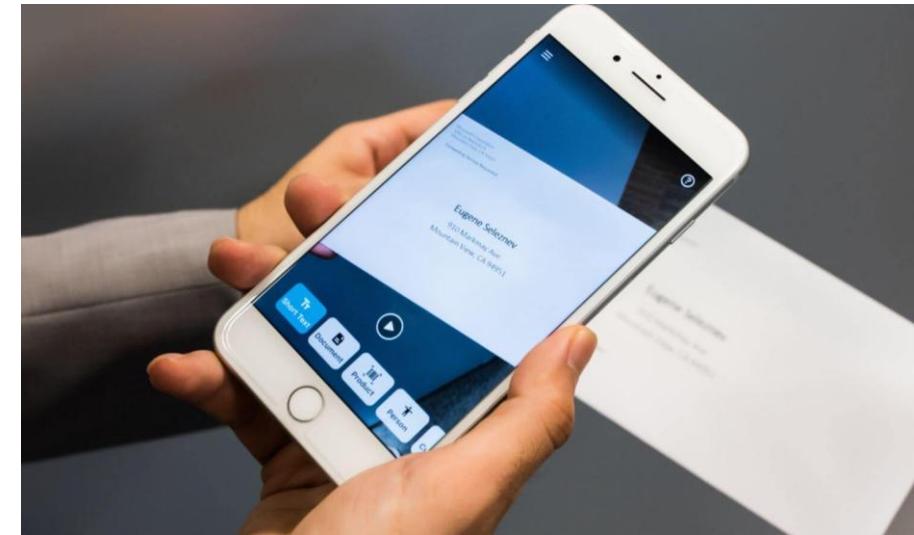
Navilens

Uso de visión artificial para identificar objetos del entorno.

Feedback mediante voz o vibración.

Ayuda en tareas cotidianas: leer etiquetas, medicaciones, identificar personas.

Incorporado en gafas inteligentes. Por ejemplo, Envision (<https://www.letsenvision.com>)



<https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai>

Reconocimiento de objetos

Cinturones o chalecos con vibraciones direccionales.

Guían al usuario en navegación sin necesidad de audio.

Ventaja: discreción y manos libres.

Complemento a GPS y bastones electrónicos.



<https://elitacwearables.com/haptic-feedback-wearables>

Dispositivos vestibles hápticos

Tecnologías para movilidad
reducida

Necesarias cuando no es posible usar teclado, ratón o pantallas táctiles.

Basadas en: voz, movimientos oculares, gestos faciales o señales cerebrales.

Objetivo: permitir comunicación y control del entorno digital.



Con la mano Ver productos →	Con cabeza o boca Ver productos →
Con la mirada Ver productos →	Con el joystick de la silla de ruedas Ver productos →

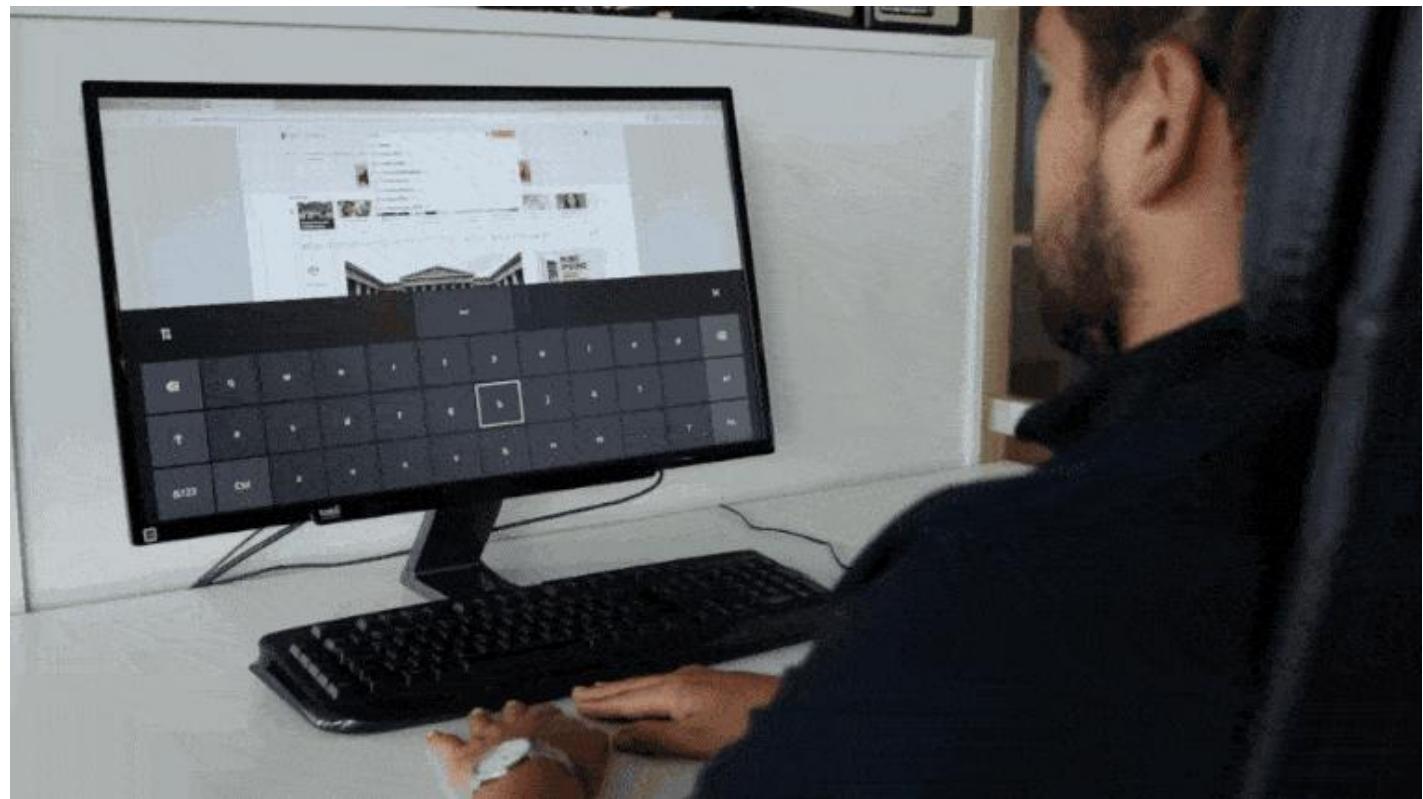
<https://qinera.com/es/tienda>

Interfaces alternativas de entrada

Utiliza cámaras de infrarrojos que detectan la posición de la pupila.

Convierte la mirada en un cursor en pantalla.

Ejemplos: Tobii Eye Tracker



<https://newmobility.com/tobii-microsoft-collaborates-to-add-eye-tracking-to-windows-10>

Seguimiento ocular (eye-tracking)

Interfaces diseñadas para escribir y comunicarse mediante la mirada.

Incluyen teclados en pantalla, pictogramas y frases predefinidas.

Permiten construir mensajes con rapidez.

Ejemplo: Tobii Dynavox

Ampliamente usado por usuarios con ELA, tetraplejia o parálisis cerebral.

Personalizable para diferentes niveles de movilidad y lenguaje.

<https://www.youtube.com/watch?v=04Rhnt4J2b4>



<https://uk.tobiidynavox.com/pages/td-snap-scanning>

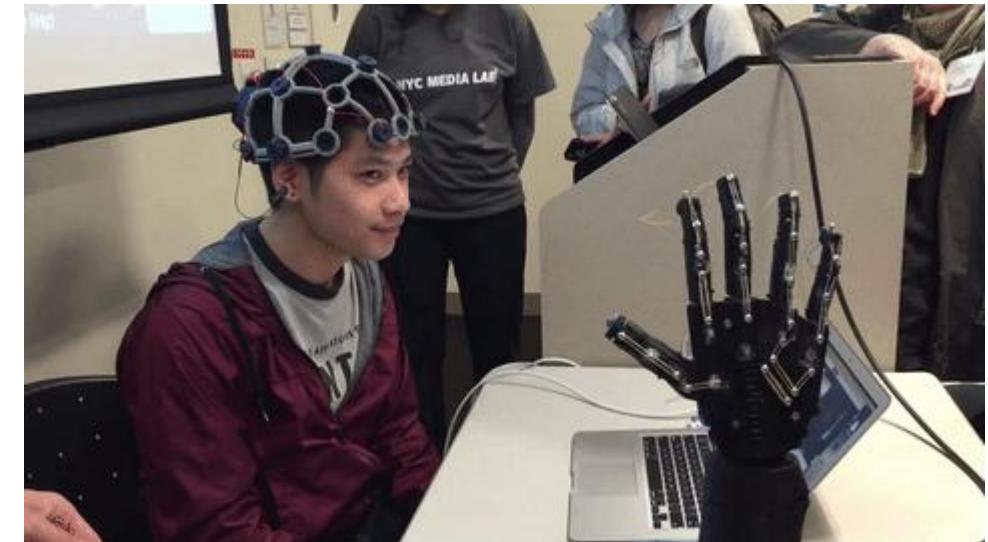
Comunicación aumentativa y alternativa

Uso de EEG (electroencefalografía) para detectar actividad cerebral.

Permite controlar cursores, sillas de ruedas o prótesis.

Ventaja: no requiere movimiento físico.

Limitaciones: precisión y ruido en la señal.



Interfaces cerebrales no
invasivas (BCI)

Dispositivo robótico que se coloca sobre el cuerpo humano.

Proporciona soporte, fuerza adicional o rehabilitación.

Tipos: parciales (miembros específicos) o totales (cuerpo completo).

Usados en medicina, rehabilitación y también en industria.



<https://www.chiuhome.com/en/products/neuroscience/cyberdyne/hal>

Exoesqueletos

Interacción persona-máquina

Unidad 5.2 – Tecnologías asistivas

Francisco Flórez

<https://personal.ua.es/francisco.florez>

<https://web.ua.es/ami4aha>

Interacción Persona-Máquina

U6: Interfaces basadas en reconocimiento de voz y sonido (I)

Javier Rodríguez Juan
j.rodriguezjuan@ua.es

José García Rodríguez
jgr@ua.es

Universidad de Alicante

2025-2026

Grado en Ingeniería en Inteligencia Artificial



Universitat d'Alacant
Universidad de Alicante

— ○ • ○ ○ —
3DPERCEPTIONLAB

distic

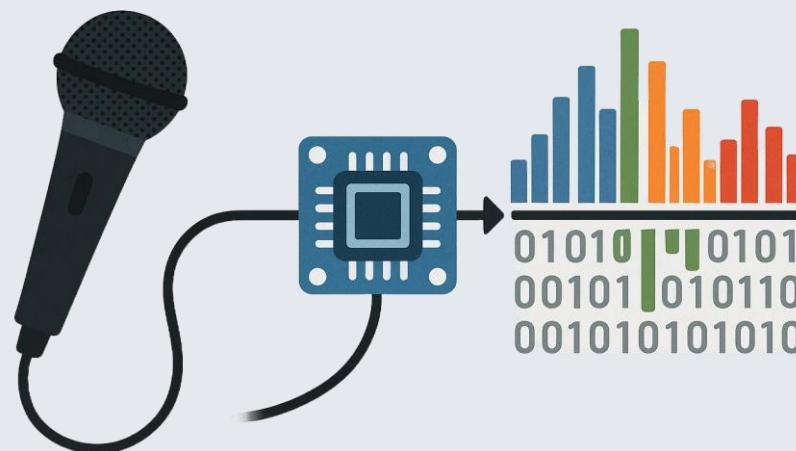
Contenidos

- Micrófonos
- Preprocesamiento
- Asistentes de voz
- Reconocimiento de sonidos ambientales

Micrófonos

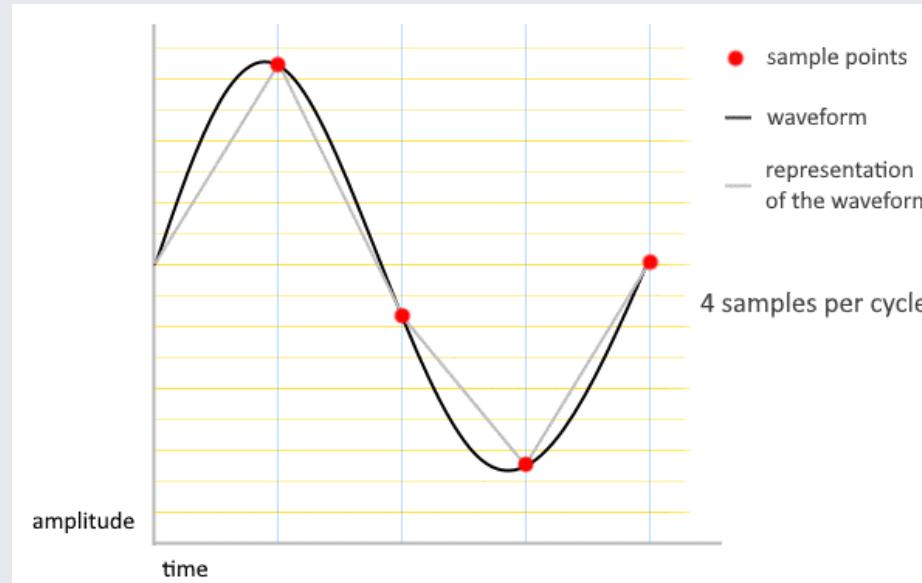
Micrófonos: Introducción

- Un micrófono es un transductor que convierte las variaciones de presión del sonido (ondas sonoras) en señales eléctricas. Estas señales luego se digitalizan mediante una tarjeta de sonido. Así se obtiene una representación numérica de la onda.



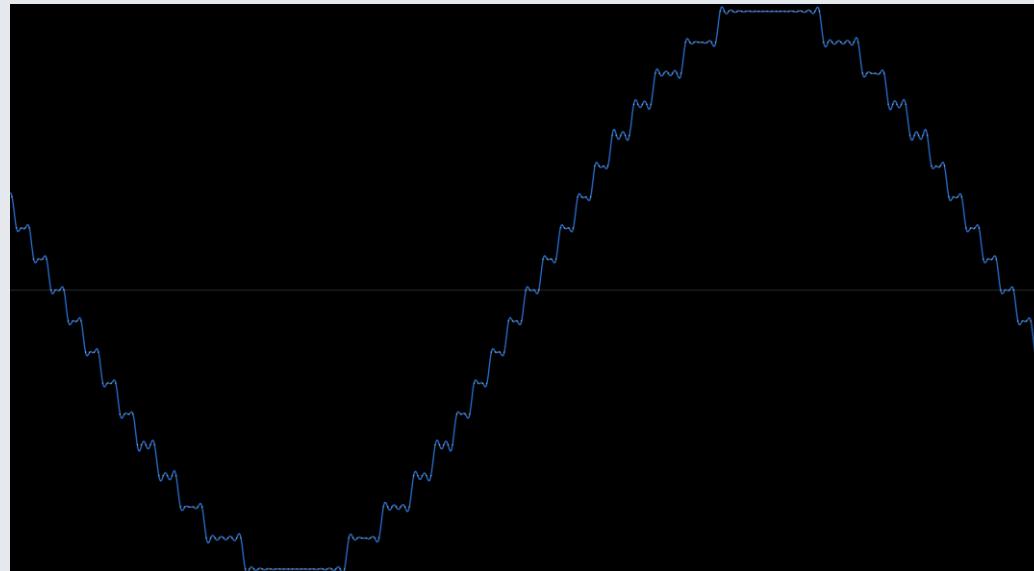
Micrófonos: Características I

- **Frecuencia de muestreo.** Son las muestras (“fotografías”) por segundo tomadas. Cuanto mayor la frecuencia de muestreo, más detalles del sonido se capturan. *16 kHz es suficiente para la voz y con 44.1 kHz se incluye todo el espectro de sonidos.*



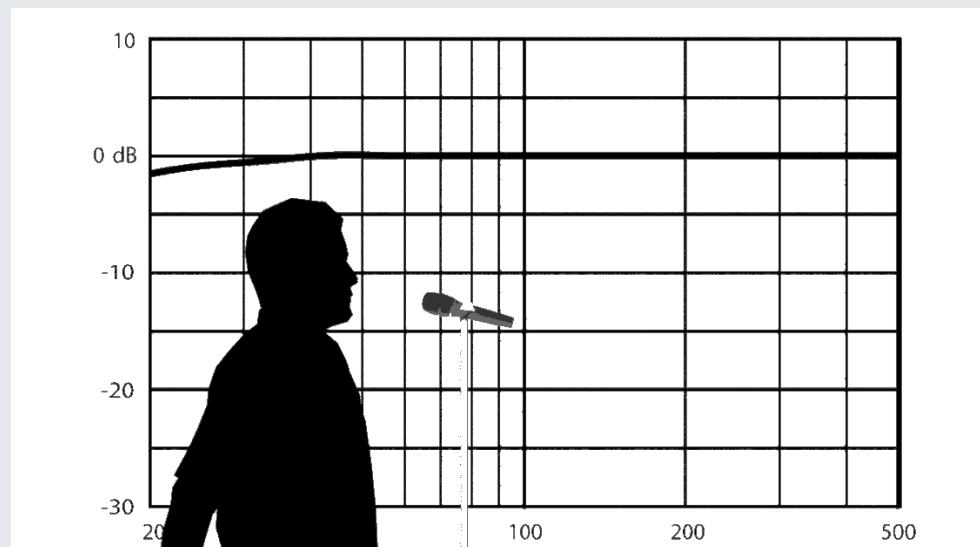
Micrófonos: Características II

- **Resolución en bits.** Es la precisión con la que se representa (codifica) cada muestra de audio, medida en bits. A más bits, mejor calidad y, menos ruido en rangos de voz elevados. Una gran cantidad de bits también crea archivos de audios más pesados. *Los CD utilizan una resolución de 16 bits.*



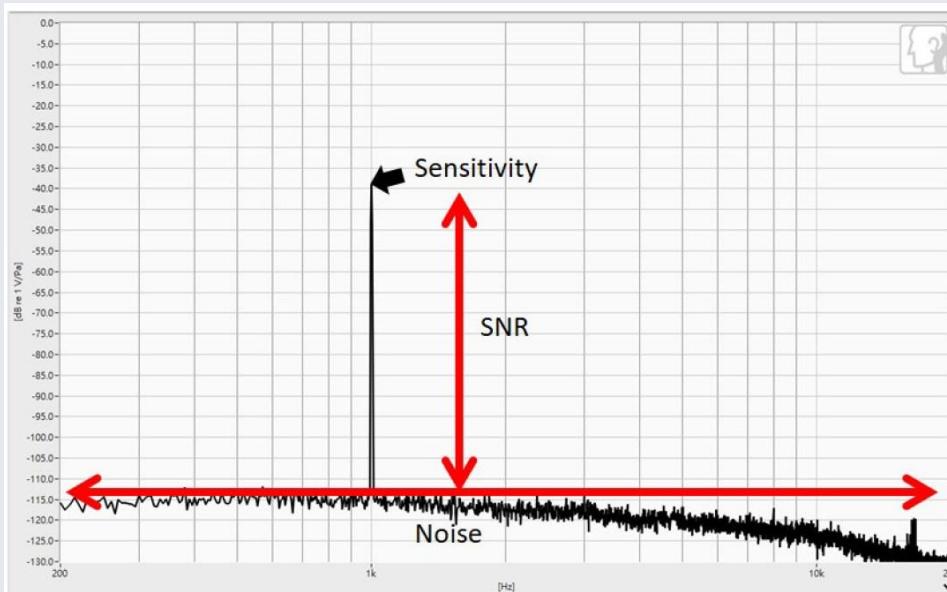
Micrófonos: Características III

- **Sensibilidad.** Es la cantidad de señal que genera el micrófono ante un sonido. Ayuda a determinar si un micro es adecuado para voces cercanas, entornos silenciosos o grabaciones a distancia. En asistentes de voz, la sensibilidad debe compensar la distancia del usuario. *Un micro muy sensible detectará voces suaves con claridad, pero también captará más ruido de fondo.*



Micrófonos: Características IV

- **Relación señal-ruido (SNR).** Es la diferencia entre el volumen de la señal útil (voz) y el ruido que añade el propio micrófono, medida en decibelios (dB). Cuanto más alto sea el SNR, más claro será el audio. *Un SNR de 70 dB significa que la voz está 70 dB por encima del ruido de fondo generado por el micro.*



Micrófonos: Tipos

De array (Campo lejano)

Definición

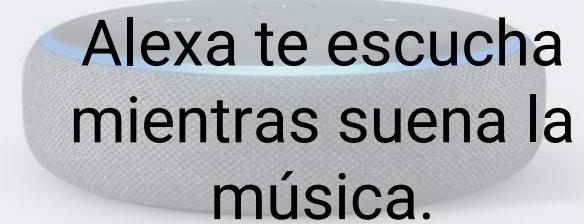
Varios micrófonos distribuidos por el dispositivo de forma circular/lineal.

Utilidad

Permiten detectar desde qué dirección viene la voz y enfocarse en ella (beamforming).

Ejemplo

Alexa te escucha mientras suena la música.



Micrófonos: Tipos

De diadema

Definición

Van integrados en unos auriculares, muy cerca de la boca del usuario

Utilidad

Muy robustos al ruido. Se usan en interacción dirigida por voz de alta precisión.

Ejemplo



Micrófonos: Tipos

Direccionales

Definición

Son alargados y focalizan el sonido en una dirección estrecha, rechazando el ruido lateral.

Utilidad

Se usan para grabar voces desde cierta distancia sin captar el ruido del entorno.

Ejemplo



Micrófonos: Tipos

MEMS (Micro-Electro-Mechanical Systems)

Definición

Micros en miniatura fabricados con semiconductores. Integran todo en un solo chip de silicio.

Utilidad

Permiten incorporar captura de voz o sonido ambiental en dispositivos compactos (móviles, wearables, etc).

Ejemplo



Preprocesamiento

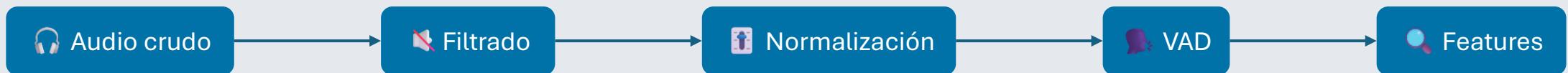
Preprocesamiento: Introducción

- Para que el audio capturado por un micrófono pueda ser usado por una arquitectura de IA, normalmente necesitaremos un paso previo donde este audio se adapta al tipo de entrada esperada por la arquitectura. Este paso previo para preparar el audio es lo que llamamos **preprocesamiento del audio**.



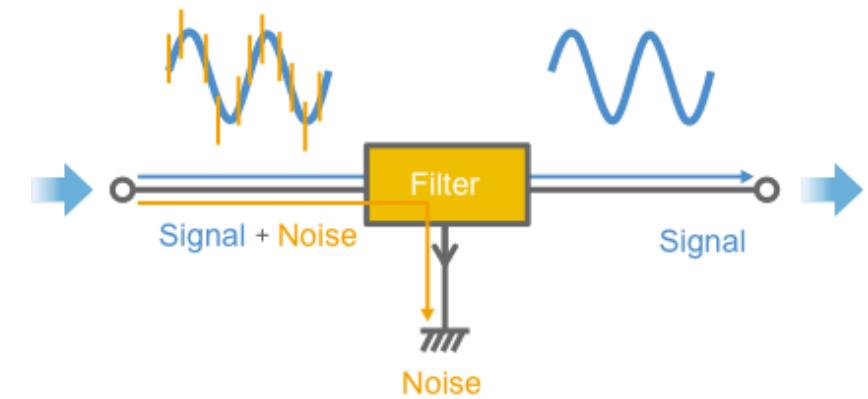
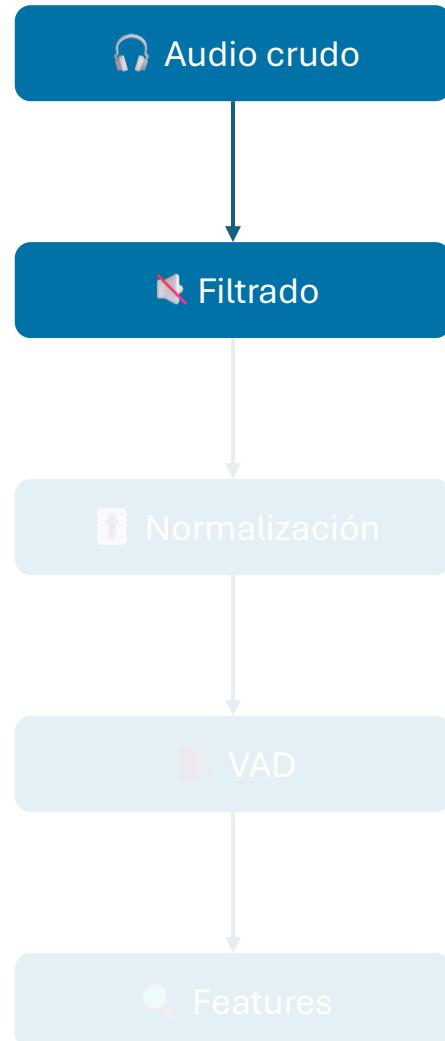
Preprocesamiento: Etapas

- El número de etapas a la hora de preprocesar un audio dependerá mucho de la arquitectura de IA específica a usar, así como de las necesidades de nuestro caso de uso. A continuación, se presenta un esquema típico de pipeline de preprocesamiento. En la práctica algunas etapas podrían añadirse o eliminarse según necesidades.



Preprocesamiento: Etapas

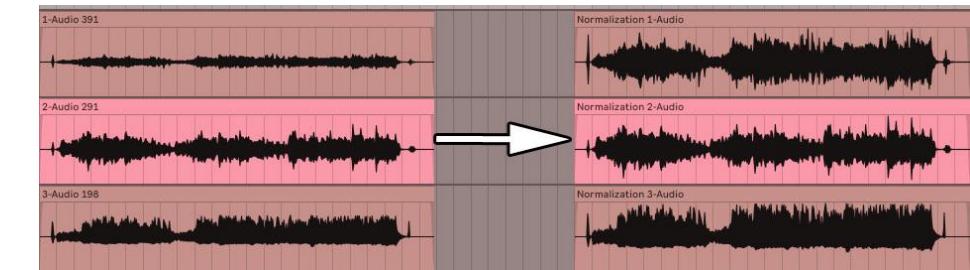
1. **Captura:** Obtención del audio a través de un micrófono.
2. **Filtrado:** Eliminar información no deseada. Técnicas comunes:
 - Filtro pasa banda: Mantiene frecuencias del rango del habla
 - *Spectral subtraction:* Resta el espectro de un ruido estimado
 - Aprendizaje profundo: Existen redes neuronales específicas para el *Noise reduction* (usualmente usado en asistentes de voz modernos).



Filtro de reducción de ruido. Extraído de [enlace](#).

Preprocesamiento: Etapas

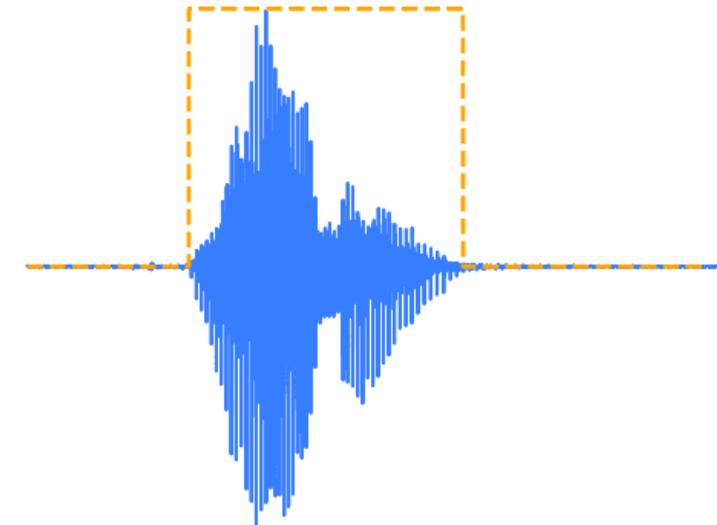
3. **Normalización:** Ajusta el volumen/energía para que el modelo no dependa del micrófono o del usuario. Una práctica común es escalar la amplitud de la señal al rango [-1,1]



Normalización de audio. Extraído de [enlace](#).

Preprocesamiento: Etapas

- 4. VAD (Voice Activity Detection):**
Separar segmentos donde hay habla de los silencios/ruido.
Este componente detecta cuándo empieza y termina de hablar la persona para evitar el procesamiento innecesario de intervalos vacíos. Se puede hacer de varias formas:
- Basado en energía (mide amplitudes del sonido)
 - Basado en modelos (RNNs, CNNs de detección de voz)



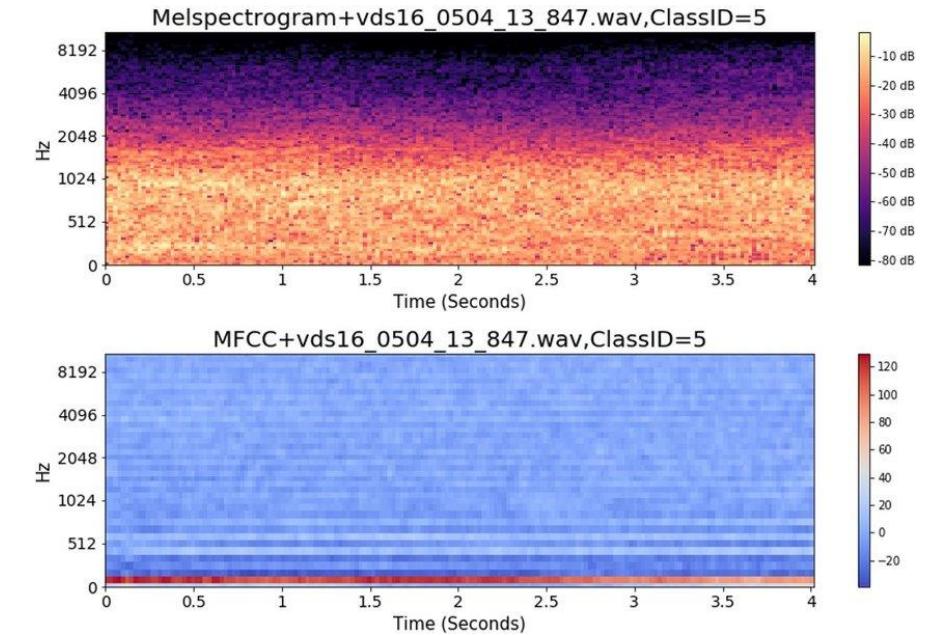
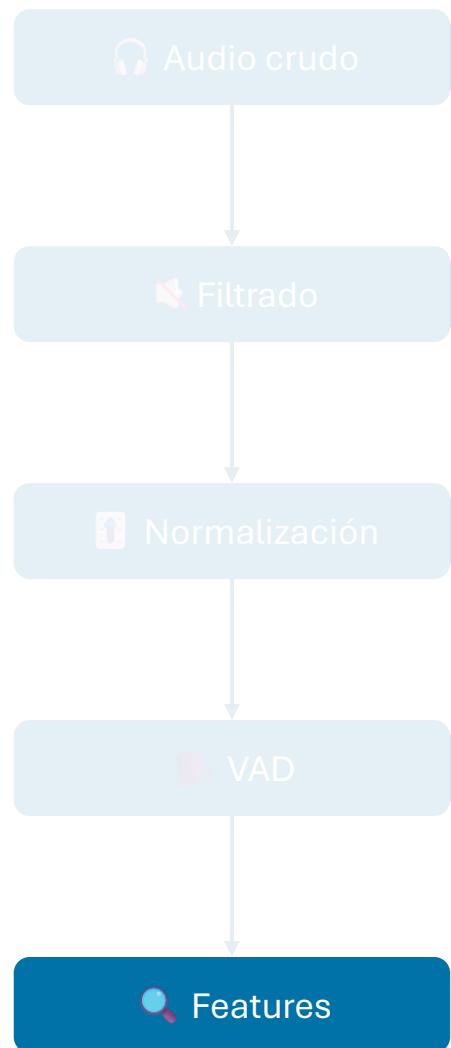
Normalización de audio. Extraído de la documentación de [Picovoice](#).

Preprocesamiento: Etapas

5. Extracción de features:

Convertir el audio en una representación numérica informativa para redes neuronales. Este paso se compone de dos pasos principales: Primero se divide el audio en ventanas de 20-40ms (*windowing*) y luego se obtiene la representación. A continuación, las principales:

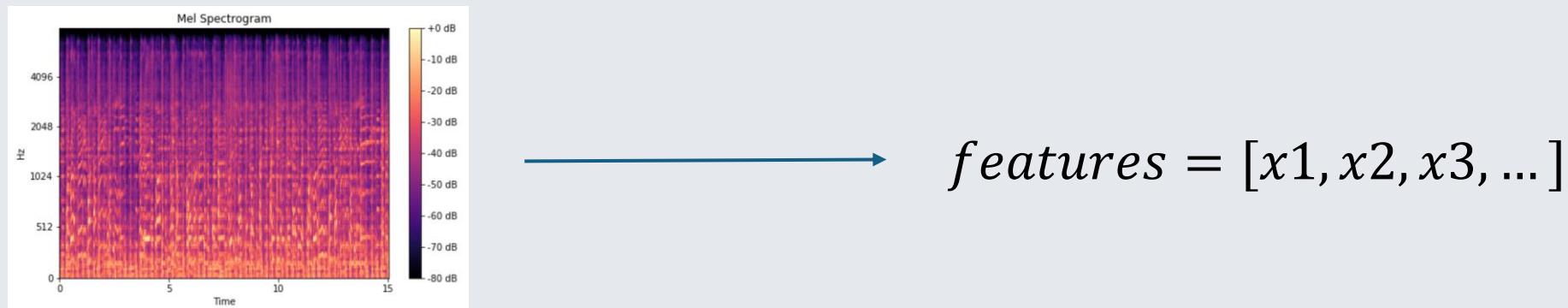
- Espectrograma de corto plazo (STFT)
- Mel Spectrogram
- Mel-Frequency Cepstral Coefficients (MFCC)



Visualización de representaciones. Extraído de *Nam et al. (2020). Traffic Density Classification Using Sound Datasets: An Empirical Study on Traffic Flow at Asymmetric Roads.*

Preprocesamiento: Embeddings

- Representaciones como el espectrograma de Mel o los coeficientes MFCC ya pueden ser usados como input de modelos de *Deep learning*, sin embargo, estas representaciones pueden convertirse en *embeddings*, que enriquecen la representación incluyendo información semántica/fonética.
- Estos *embeddings* los generan arquitecturas tipo CNN/Transformer como Wav2Vec [7], YAMNet [12] o el encoder de Whisper [15].



Asistentes de voz

Asistentes de voz: Definición y hardware

Sistemas de IA diseñados para interactuar mediante diálogo hablado. Sus micrófonos tienen las siguientes características:

- Diseñados para captura de campo lejano (*far-field*): Para que el asistente pueda escucharte desde la lejanía, el usuario no tiene porque estar cerca del dispositivo.
- Usan micrófonos de array. Normalmente utilizan arrays de múltiples micrófonos (ej. 4 u 8 micrófonos dispuestos circularmente)

Asistentes de voz: Software

Tecnologías para mejorar la calidad de la escucha:

- **Beamforming:** Orientar la escucha hacia la voz del usuario y filtrar ruidos de fondo
- **Reducción de ruido**
- **Cancelación de eco:** Eco del propio sonido del asistente si este está reproduciendo audio (p. ej. música)
- **Wake Word Spotting:** Detección de palabras clave (p. ej. Detectar cuándo se dice “Alexa”)

Algunos ejemplos de asistentes pueden ser: *Siri* (Apple), *Google Assistant* (Google), *Alexa* (Amazon) o *Cortana* (Microsoft)

Asistentes de voz: Software

Tecnologías para:

- Beamforming de fondo
- Reducción de ruido
- Cancelación de reproducción
- Wake Word Sensors que se dice "Alexa"



Siri



Google



amazon alexa



Algunos ejemplos:
(Google), Alexa

filtrar ruidos

este está

ectar cuándo

le Assistant

Asistentes de voz: Pipeline

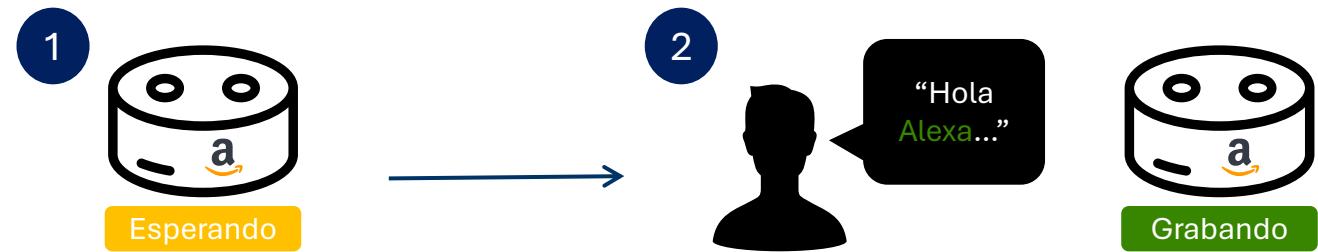
Cada asistente tiene sus peculiaridades, pero en general todos siguen un pipeline similar:

- 1. Reposo:** Dispositivo en reposo hasta la escucha de la palabra de activación.



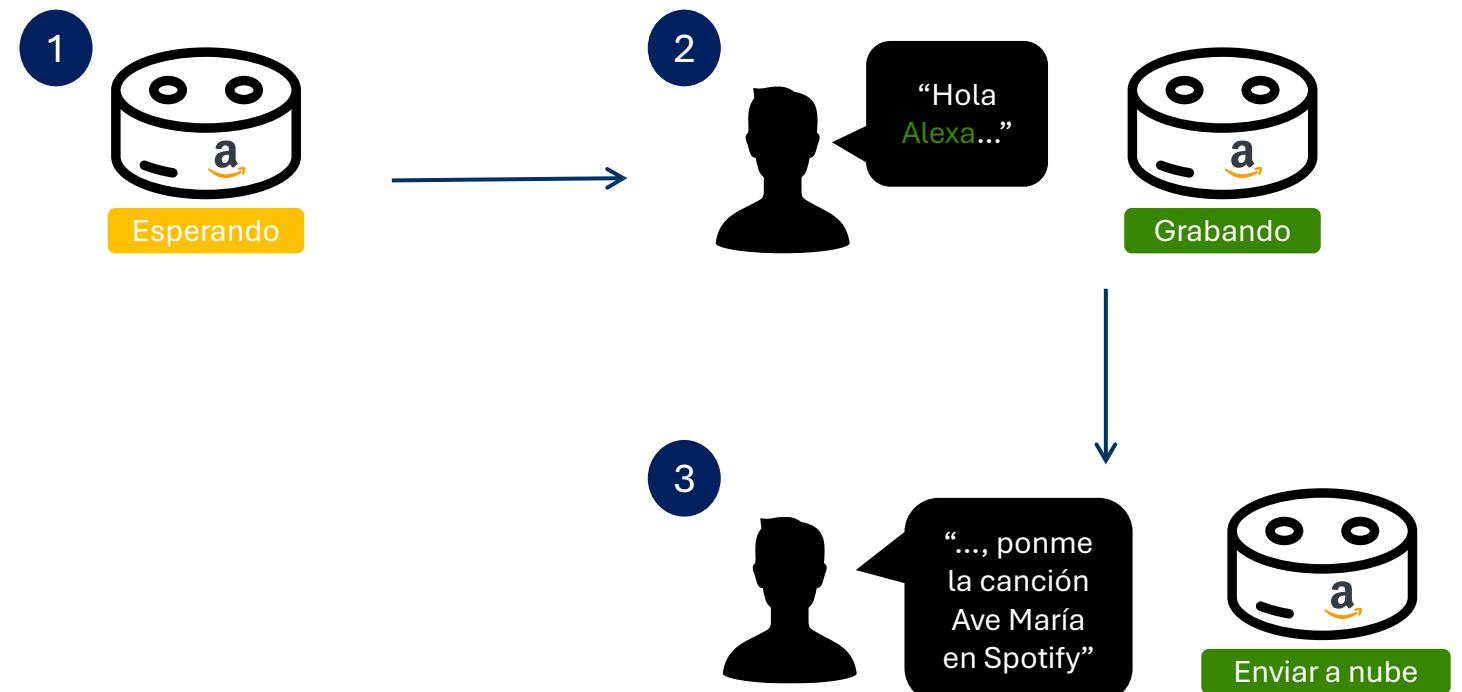
Asistentes de voz: Pipeline

2. **Activación:** Dispositivo detecta la palabra de activación ("Hey Siri", "Ok Google", "Alexa", etc). Esto suele hacerse con un modelo ligero de reconocimiento de palabras clave (*Wake Word Spotting*) que corre localmente en el dispositivo.



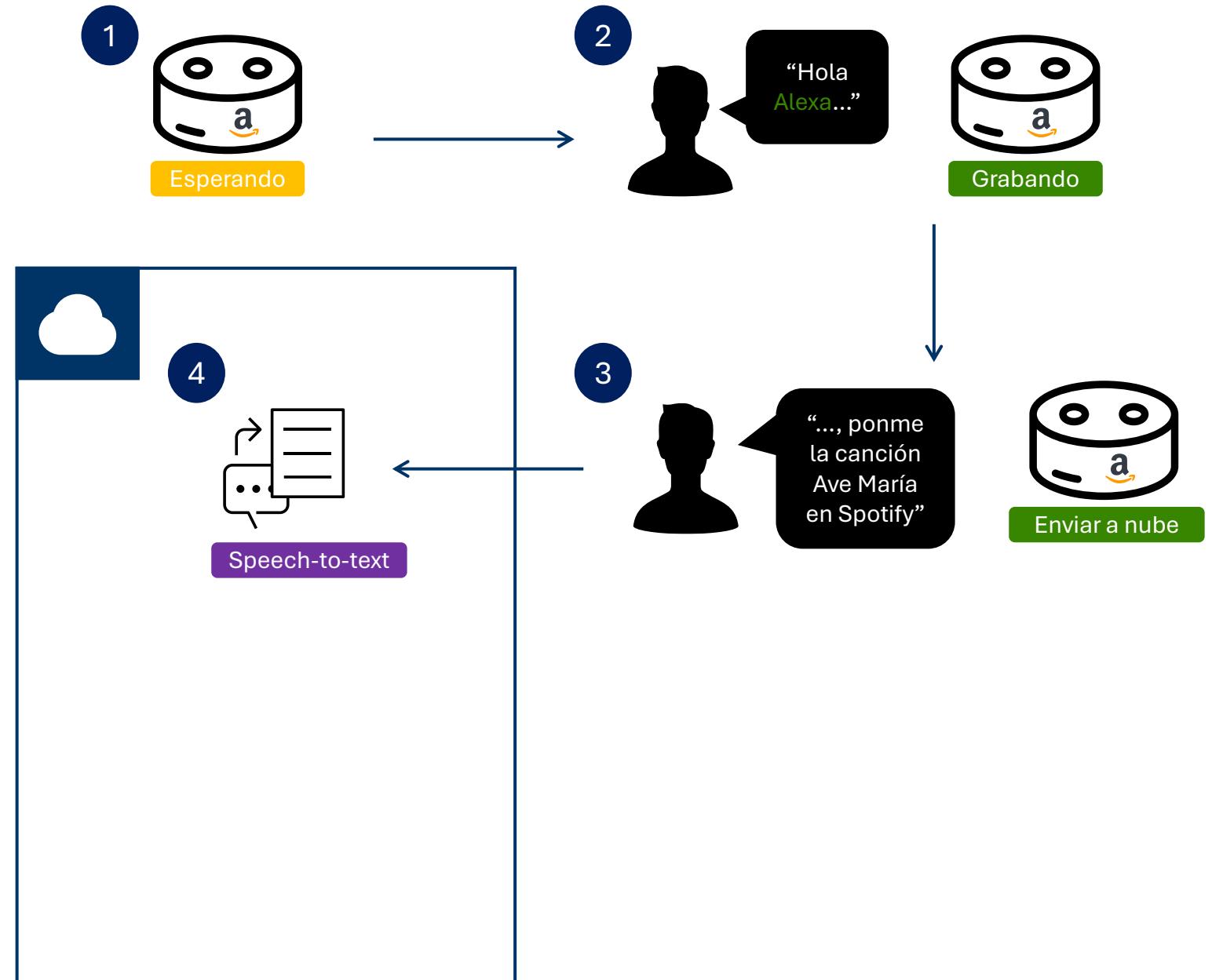
Asistentes de voz: Pipeline

3. **Captura y envío de audio:**
Una vez activado, el asistente graba la voz hasta detectar final de frase (modelo *end-of-turn*) y envía audio a la nube para su procesamiento.



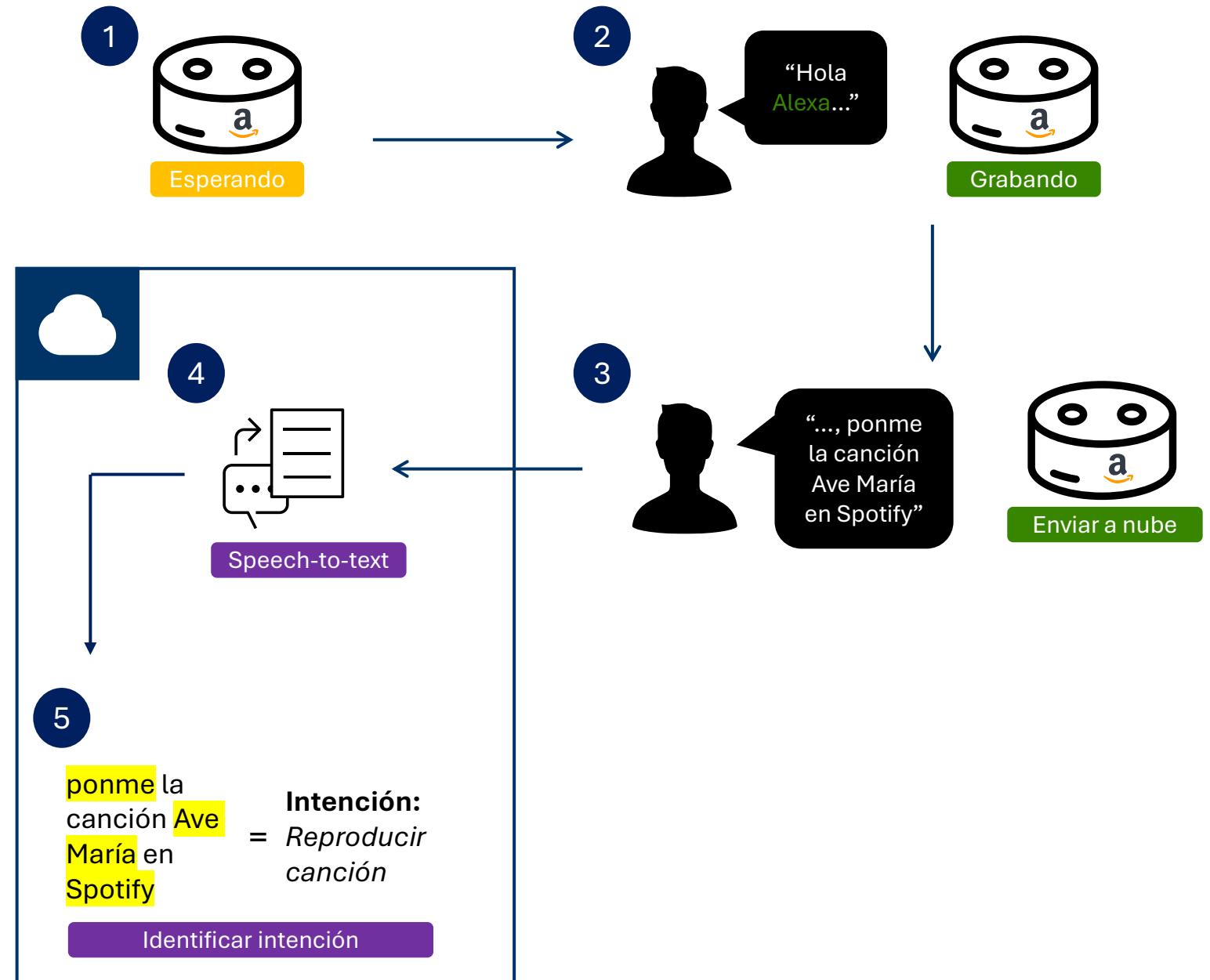
Asistentes de voz: Pipeline

4. **Reconocimiento de voz (ASR):** En la nube, un modelo de reconocimiento automático del habla convierte el audio en texto. Antiguamente se usaban modelos clásicos de *machine learning* como los Modelos Ocultos de Markov (HMM). Hoy en día lo más común es el uso del *Deep Learning*, dónde las redes neuronales son capaces de realizar el reconocimiento con mucha precisión.



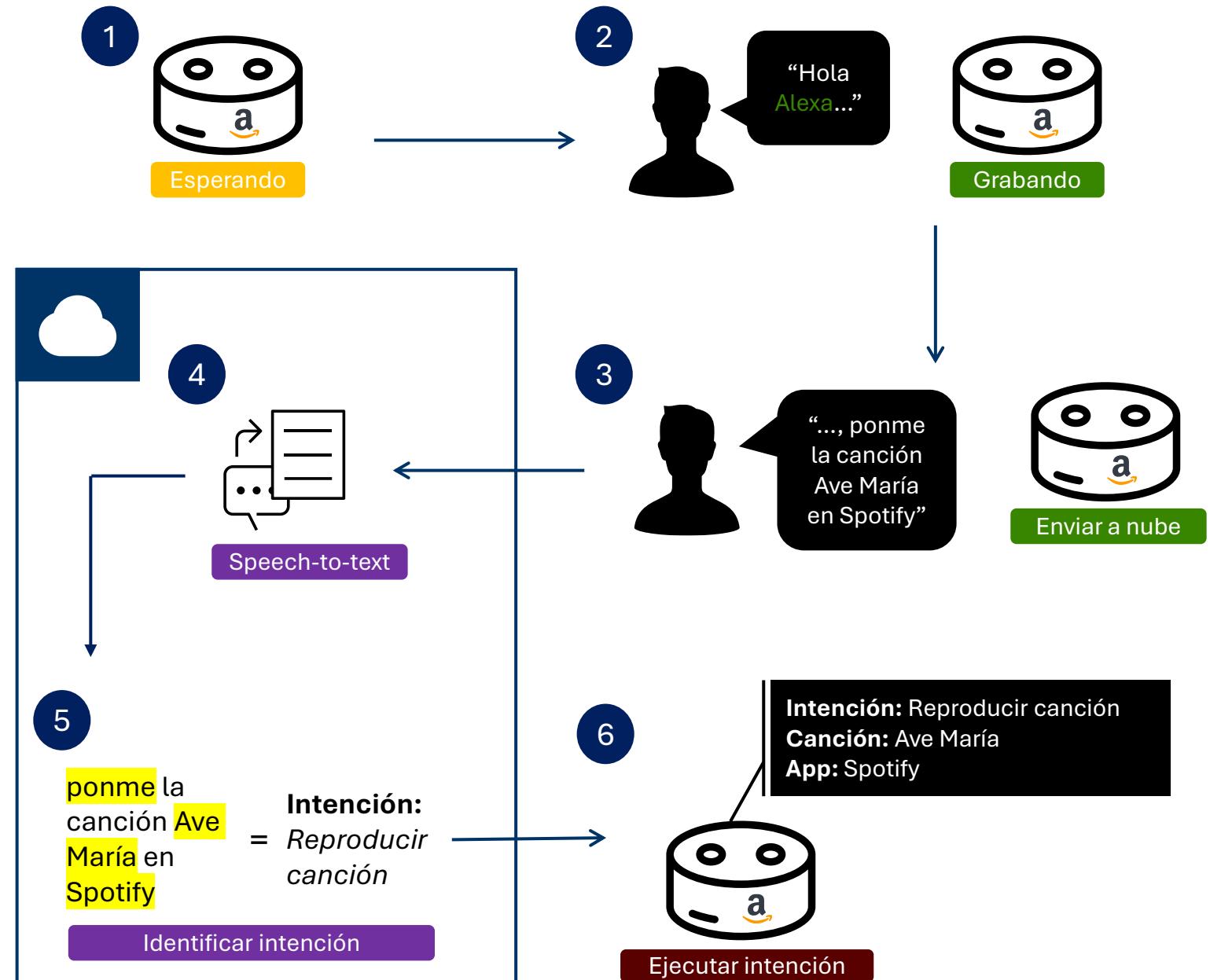
Asistentes de voz: Pipeline

5. **Compresión del lenguaje natural (NLU):** Una vez el texto se ha transscrito, el asistente debe entender la intención del usuario. Es decir, identificar que acción y respuesta se requiere. Este suele ser un módulo más complejo donde se realizan distintas tareas, como la tokenización, la clasificación de la intención y la extracción de entidades. Los asistentes modernos usan LLMs (p. ej. GPT-4o) para mejorar la comprensión.



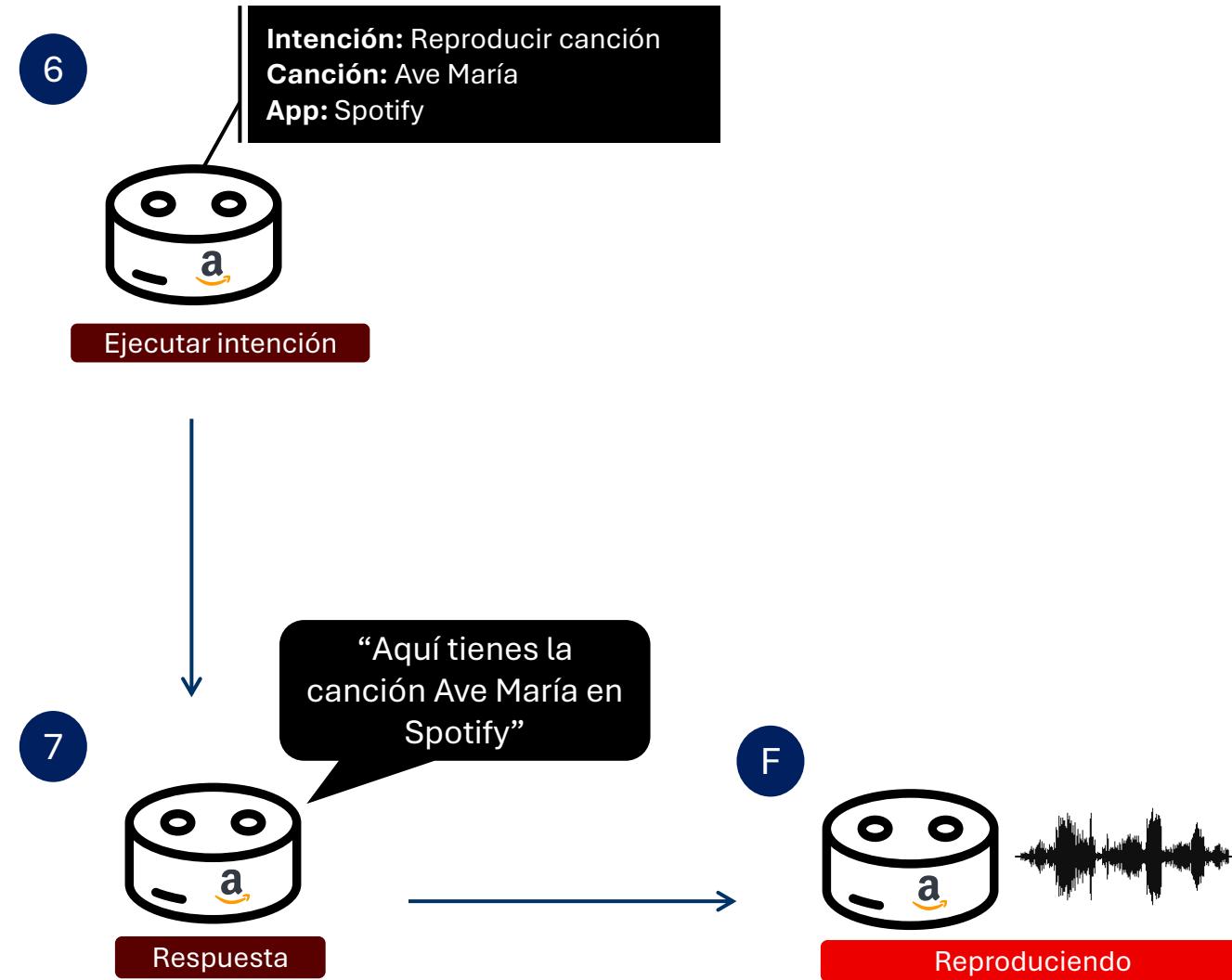
Asistentes de voz: Pipeline

6. **Gestión del diálogo y acción:** La información viaja desde la nube al asistente, el cuál determina cómo responder. Puede implicar consultar alguna base de conocimiento, ejecutar una acción o simplemente elaborar una respuesta.



Asistentes de voz: Pipeline

7. **Síntesis del habla (TTS):**
Finalmente, los asistentes generan una respuesta conversacional usando modelos *Text-To-Speech*. Al igual que en el ASR, en el TTS también domina el uso de redes neuronales para generar voces hiper realistas y respuestas naturales.



Asistentes de voz: Retos I

Y esto no es todo! Los asistentes son sistemas de IA complejos que usan una gran variedad de **submódulos** para mejorar la precisión, usabilidad y naturalidad en las interacciones. Estos submódulos buscan solucionar distintos desafíos a los que se enfrenta un sistema inteligente cuando se despliega en entornos reales. Pongamos el ejemplo de Alexa...



Voice-adapted

Alexa usa propiedades acústicas de la voz de la persona que realiza la petición de activación “Alexa”, para filtrar las voces de fondo que producen ruido en las siguientes interacciones.



Context-aware

Alexa también usa el contexto del dispositivo, que puede incluir el día que es, las interacciones anteriores del usuario, sus preferencias o el lugar en el que está colocado el dispositivo.



Error handling

Al ejecutar ciertas acciones, se pide confirmación repitiendo alguno de los datos descritos por el usuario. Esto también ocurre cuando la confianza del asistente es baja.

Asistentes de voz: Retos I

Y esto no es todo! Los asistentes son sistemas de IA complejos que usan una gran variedad de **submódulos** para mejorar la precisión, usabilidad y naturalidad en las interacciones. Estos submódulos buscan solucionar distintos desafíos a los que se enfrenta un sistema inteligente cuando se despliega en entornos reales. Pongamos el

Esto junto a otros componentes (como los modelos *wake word spotting* o *end-of-turn*) es lo que componen lo que conocemos hoy en día como **asistentes de voz**.

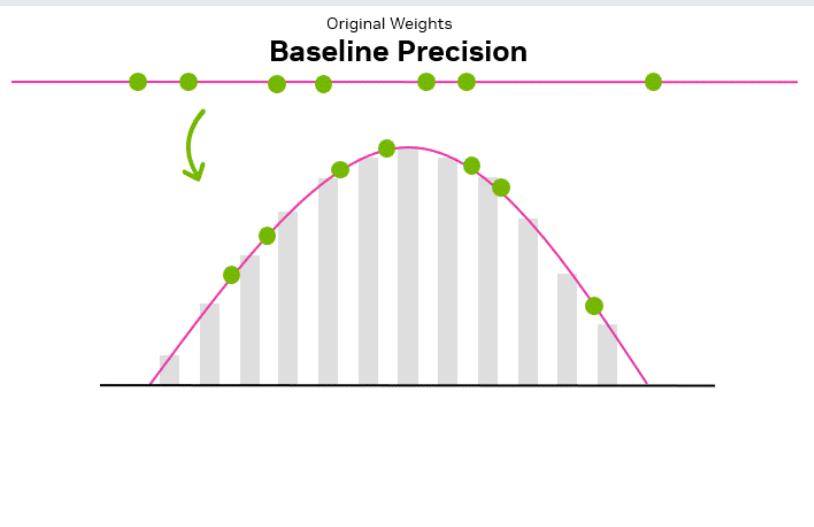
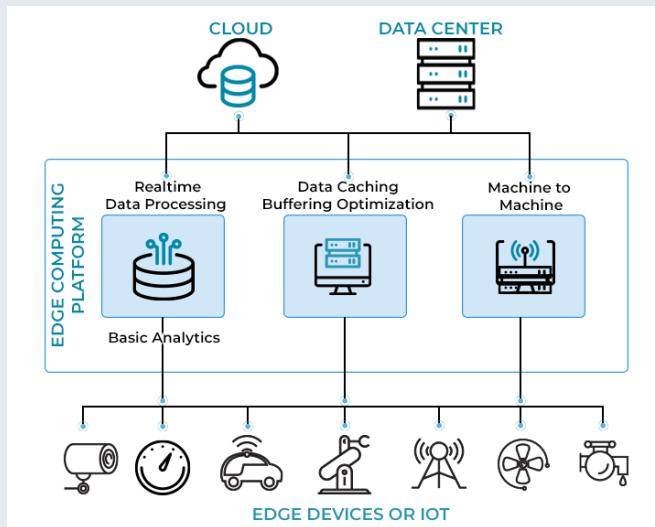
Alexa usa propiedades acústicas de la voz de la persona que realiza la petición de activación “Alexa”, para filtrar las voces de fondo que producen ruido en las siguientes interacciones.

Alexa también usa el contexto del dispositivo, que puede incluir el día que es, las interacciones anteriores del usuario, sus preferencias o el lugar en el que está colocado el dispositivo.

Al ejecutar ciertas acciones, se pide confirmación repitiendo alguno de los datos descritos por el usuario. Esto también ocurre cuando la confianza del asistente es baja.

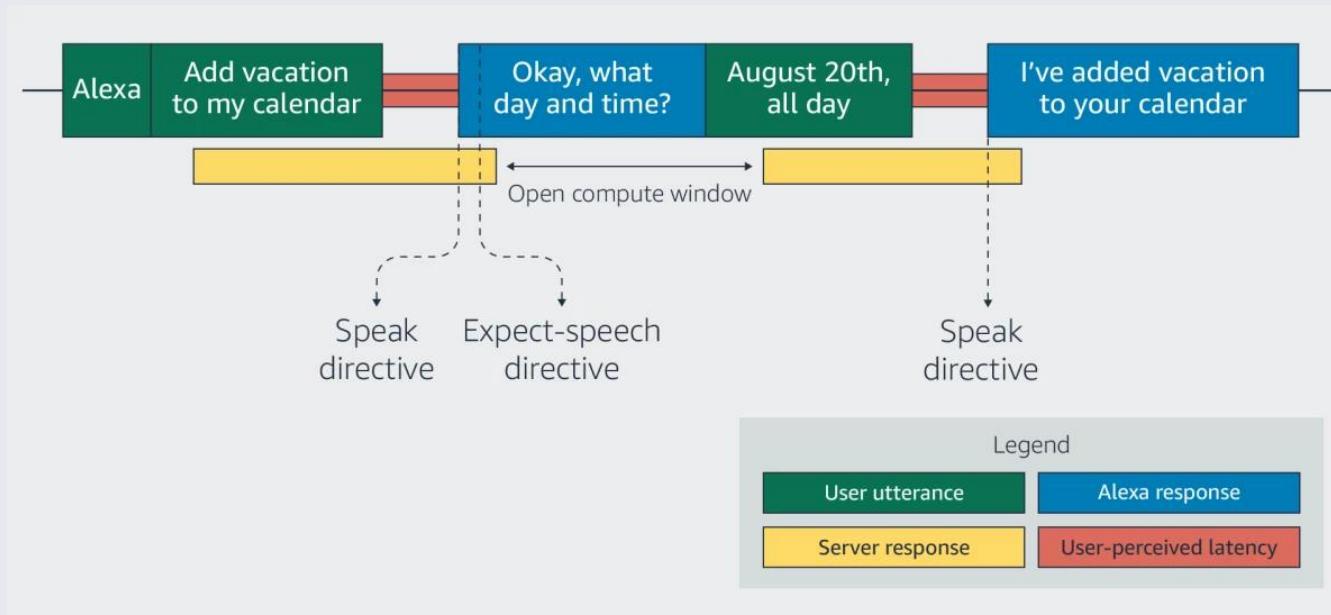
Asistentes de voz: Retos II

Y ojo! Que todo esto debe ocurrir en tiempo real para que la interacción sea fluida. Las grandes empresas optimizan la latencia llevando ciertos cálculos al dispositivo (p. ej. *on-the-edge computing*), usando compresión de modelos (p. ej. *model quantization*) y servidores muy escalables (p. ej. AWS EC2 Autoscaling).



Asistentes de voz: Retos III

Pero aún hay más! Los asistentes de voz no solo presentan retos a nivel de IA, sino también a nivel de arquitectura software. Estos son obras de ingeniería sofisticadas que han pasado por un largo proceso iterativo de desarrollo para adaptarse perfectamente al uso en la vida diaria.





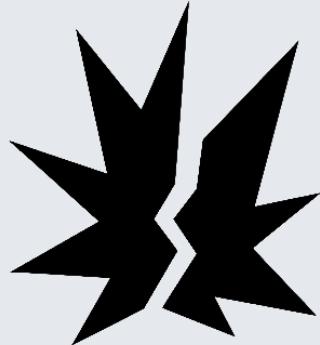
"This work is the culmination of very close collaboration between scientists and engineers to design contextual machine learning to operate at Alexa scale"

Debprakash Patnaik, Ph.D.
Senior ML Engineering Manager @ Amazon AGI

Reconocimiento de sonidos ambientales (ESR)

ESR: Introducción

Nuestros entornos están llenos de **sonidos** relevantes: timbres, pasos, aplausos, sirenas, ladridos, rotura de cristales, entre muchos más. El **reconocimiento de sonidos ambientales** (también llamado *Environmental Sound Recognition*, ESR) es la capacidad de un sistema para identificar y clasificar estos sonidos no verbales.



ASR vs ESR

Similitudes	Diferencias
En ambos campos predomina el uso de redes neuronales como CNNs o Transformers	Mientras que el análisis del habla tiene una estructura sintáctica clara, los sonidos son más diversos.
	Escasez de datos para sonidos raros o eventos de corta duración.
	Mientras que dentro del ASR hay una gran cantidad de subtareas (transcripción, subtulado en streaming, clasificación), el ESR es más limitado y suele trabajarse como una tarea de clasificación de audio

ESR: Limitación

En aplicaciones reales (asistentes de voz, wearables, sensores IoT...), necesitamos detectar eventos acústicos como pasos, alarmas, conversaciones o caídas directamente en el dispositivo:

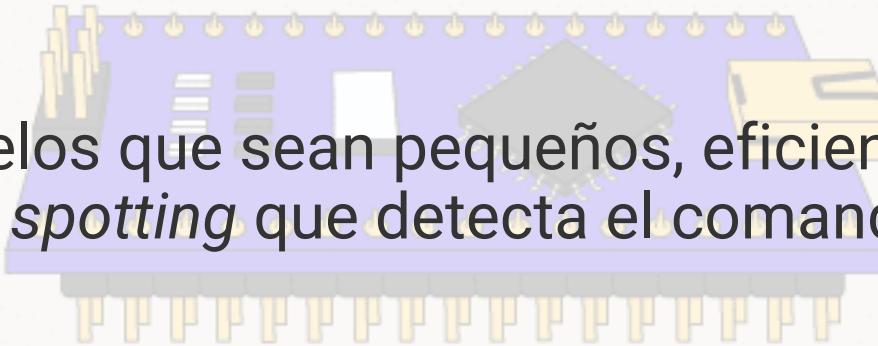
-  Sin enviar audio a la nube → Mejor privacidad
-  Con baja latencia → Respuestas en tiempo real
-  Con bajo consumo energético → Adecuado para dispositivos ligeros

Esto propició el desarrollo del **TinyML**.

ESR: TinyML

TinyML (Tiny Machine Learning) es una rama de la inteligencia artificial que se centra en ejecutar modelos de aprendizaje automático directamente en dispositivos muy pequeños, con recursos muy limitados (microcontroladores, sensores, wearables, IoT).

El TinyML requiere modelos que sean pequeños, eficientes y rápidos (p. ej. El modelo de *wake word spotting* que detecta el comando “Hey Google” en el *Google Assistant*).

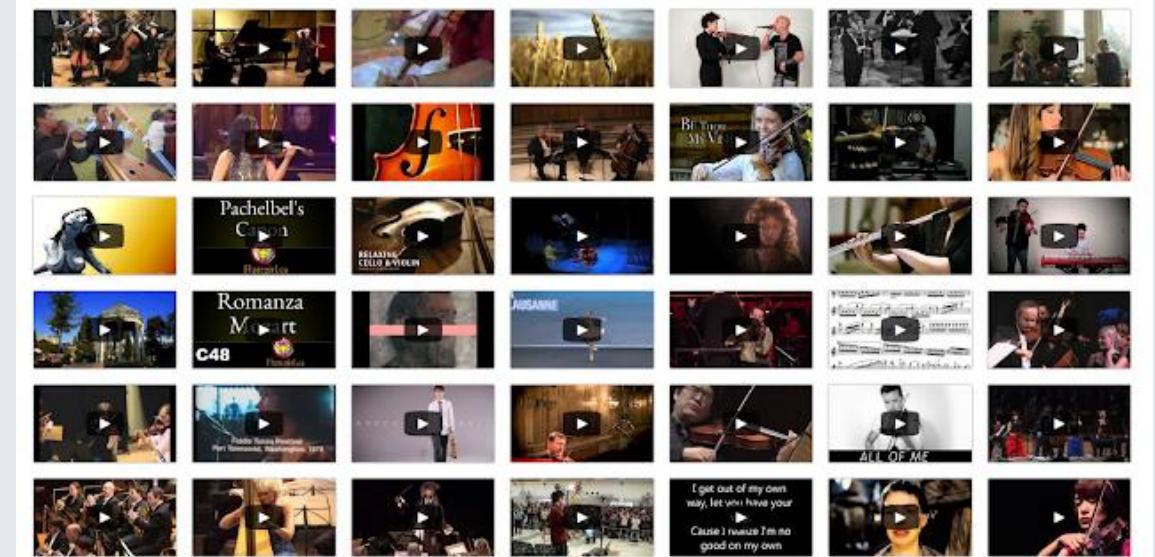


ESR: Modelos

Modelo	Arquitectura Interna	Ideal para...	¿TinyML?
YAMNet [12]	Red ligera basada en una CNN (MobileNet). Utiliza capas convolucionales separables para extraer características del espectrograma log-Mel.	Clasificación de sonidos en tiempo real. Detección simple o embedded (dispositivos móviles).	 Si (muy ligero)
PANNs [13]	Conjunto de modelos CNN con atención global. Procesan espectrogramas Mel y permiten interpretar qué parte del audio activa cada clase.	Clasificación robusta de sonido, tagging, localización de eventos.	 Parcial, dependiendo de la versión del modelo
AST [14]	Modelo basado en Transformer tipo ViT . Toma espectrogramas log-Mel y los divide en patches como si fueran imágenes	Clasificación avanzada, transfer learning, alto rendimiento en tareas complejas.	 No (modelo pesado, necesita GPU o CPU potente)

ESR: Datasets

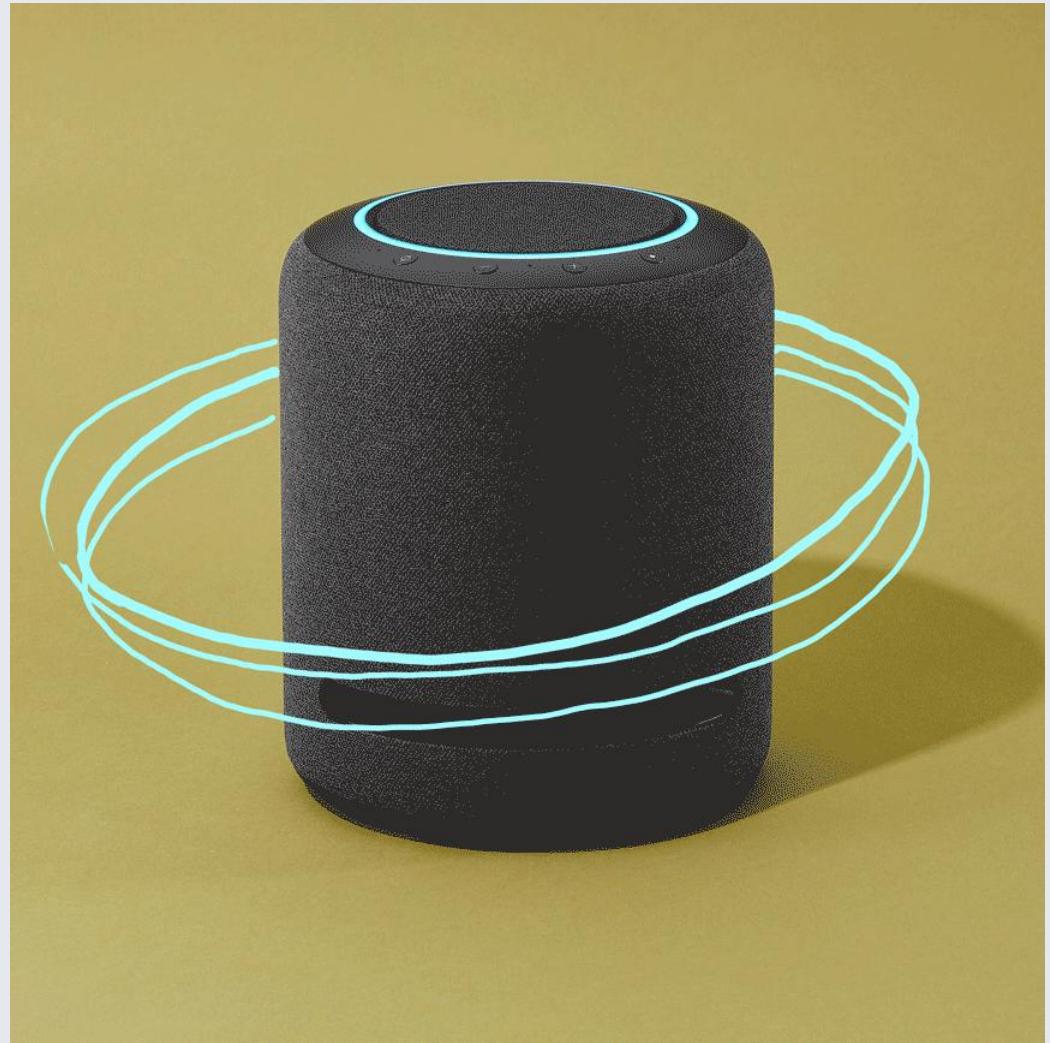
- **UrbanSound8k [9]:** 8,732 sonidos urbanos variados en 10 clases
- **ESC-50 [10]:** 2,000 clips cortos en 50 clases
- **AudioSet [11]:** Dataset de gran escala creado por Google. Alberga más de 2M de clips y tiene 527 clases de sonido distintos



Segmentos representando la clase *Violin*, en el dataset *AudioSet*. Extraído de <https://research.google/blog/announcing-audioset-a-dataset-for-audio-event-research/>

ESR: Alexa Guard

En una **casa inteligente**, un asistente podría escuchar eventos importantes. Amazon Alexa cuenta con la función Alexa Guard, donde los altavoces Echo detectan automáticamente sonidos de alarma de humo o cristal rompiéndose cuando el usuario no está en casa, y envían una alerta al móvil.



ESR: Seguridad

En entornos de seguridad, el análisis de audio puede detectar disparos, rompimiento de puertas o cristales en intrusiones. Un ejemplo de esto es [ShotSpotter](#), un sistema de la empresa SoundThinking que detecta disparos y triangula su posición.

Este tipo de sistemas es crítico en entornos con altas tasas de criminalidad, como es el caso de Estados Unidos, que tiene una amplia red de micrófonos alrededor de más de 170 ciudades para tratar de reducir la violencia armada. Recientemente, este sistema se ha lanzado también en América Latina (Uruguay).



ESR: Vehículos autónomos

En el campo de la conducción autónoma [9], se investiga usar micrófonos para detectar sirenas de ambulancias, bocinas u otros sonidos de tráfico que un vehículo debería tener en cuenta.

Esto complementaría la visión (cámaras, LIDAR) para una percepción del entorno más completa.



Referencias

- [1] Hidden Markov Models (2025). ScienceDirect.
<https://www.sciencedirect.com/topics/neuroscience/hidden-markov-model>
- [2] Mael Fabien (2020, Mayo 26) Introduction to Automatic Speech Recognition (ASR).
https://maelfabien.github.io/machinelearning/speech_reco/#
- [3] Jurafsky, D., & Martin, J. H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/>
- [4] Matt Payne (2021, Sept 29). What is Beam Search? Explaining The Beam Search Algorithm. Width.ai.
<https://www.width.ai/post/what-is-beam-search>
- [5] Andrew Maas (2025). Spoken Language Processing (CS224S). Stanford.
<https://web.stanford.edu/class/cs224s/semesters/2025-spring/syllabus>
- [6] Mamyrbayev Orken et al. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. Scientific Reports. <https://doi.org/10.1038/s41598-022-12260-y>
- [7] Alexei Baevski et al. (2020). Wav2vec 2.0: A framework for Self-Supevised Learning of Speech Representations. <https://arxiv.org/pdf/2006.11477>
- [8] Loren Lugosch (2020). Sequence-to-sequence learning with Transducers.
<https://lorenlugosch.github.io/posts/2020/11/transducer/>

Referencias

- [9] André Luiz Florentino (2025). A dataset for environmental sound recognition in embedded systems for autonomous vehicles. *Scientific Data*. <https://doi.org/10.1038/s41597-025-05446-2>
- [10] Karol J. Piczak (2015). ESC: Dataset for Environmental Sound classification. *Proceedings of the 23rd Annual ACM Conference on Multimedia*. <http://dx.doi.org/10.1145/2733373.2806390>
- [11] Jort F. Gemmeke (2017). Audio Set: An ontology and human-labeled dataset for audio events. *IEEE ICASSP*. <https://ieeexplore.ieee.org/abstract/document/7952261>
- [12] Konstantinos Drossos (2020). Sound Event Detection with Depthwise Separable and Dilated Convolutions. <https://arxiv.org/abs/2002.00476>
- [13] Qiuqiang Kong (2019). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. <https://arxiv.org/abs/1912.10211>
- [14] Yuan Gong (2021). AST: Audio Spectrogram Transformer. <https://arxiv.org/abs/2104.01778>
- [15] Radford et al. (2022). Robust speech recognition via large-scale weak supervision. <https://doi.org/10.48550/arXiv.2212.04356>

Interacción Persona-Máquina

U6: Interfaces basadas en reconocimiento de voz y sonido (II)

Javier Rodríguez Juan
j.rodriguezjuan@ua.es

José García Rodríguez
jgr@ua.es

Universidad de Alicante

2025-2026

Grado en Ingeniería en Inteligencia Artificial



Universitat d'Alacant
Universidad de Alicante

— ○ • ○ ○ —
3DPERCEPTIONLAB

distic

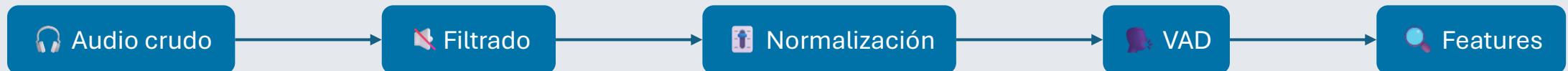
Contenidos

- Repaso
- Introducción al ASR
- El progreso
- El estado del arte
- ASR en la práctica

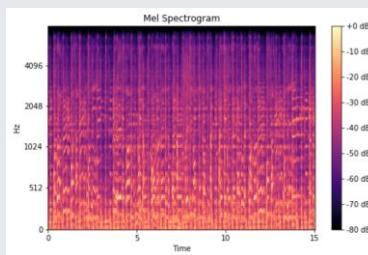
Repaso

Repaso: Preprocesamiento

- Al capturar un audio se deben realizar varios pasos de procesamiento antes de introducir este como entrada de la arquitectura de IA.



- Es común usar CNNs/Transformer encoders para generar representaciones más informativas. A estas las llamamos *embeddings*



$$features = [x_1, x_2, x_3, \dots]$$

Repaso: ESR

Nuestros entornos están llenos de **sonidos** relevantes: timbres, pasos, aplausos, sirenas, ladridos, rotura de cristales, entre muchos más. El **reconocimiento de sonidos ambientales** (también llamado *Environmental Sound Recognition*, ESR) es la capacidad de un sistema para identificar y clasificar estos sonidos no verbales.



Repaso: ASR vs ESR

Similitudes	Diferencias
En ambos campos predomina el uso de redes neuronales como CNNs o Transformers	Mientras que el análisis del habla tiene una estructura sintáctica clara , los sonidos son más diversos.
	Escasez de datos para sonidos raros o eventos de corta duración.
	Mientras que dentro del ASR hay una gran cantidad de subtareas (transcripción, subtitulado en streaming, clasificación) , el ESR es más limitado y suele trabajarse como una tarea de clasificación de audio

Introducción al ASR

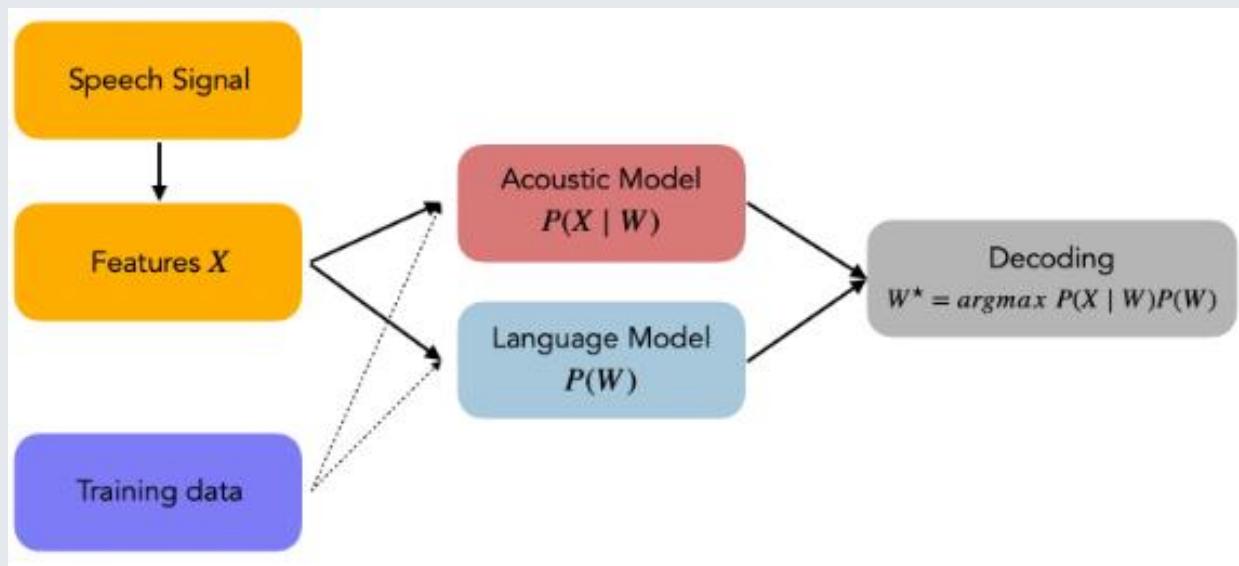
ASR: ¿Qué es?

El *Automatic Speech Recognition* (ASR) es la capacidad de una máquina para transformar el lenguaje natural en un formato interpretable para ella. Dentro de este se encuentra la tarea del *Speech-To-Text* (STT) donde a partir de un audio de entrada, se obtiene su transcripción asociada.

 No hay que confundir el ASR con el STT. El STT es la tarea de pasar de voz a texto mientras que el ASR abarca todas las técnicas involucradas con el análisis de audio, como el análisis de emociones, la diarización de locutores o incluso la predicción del deterioro cognitivo.

ASR: Los inicios

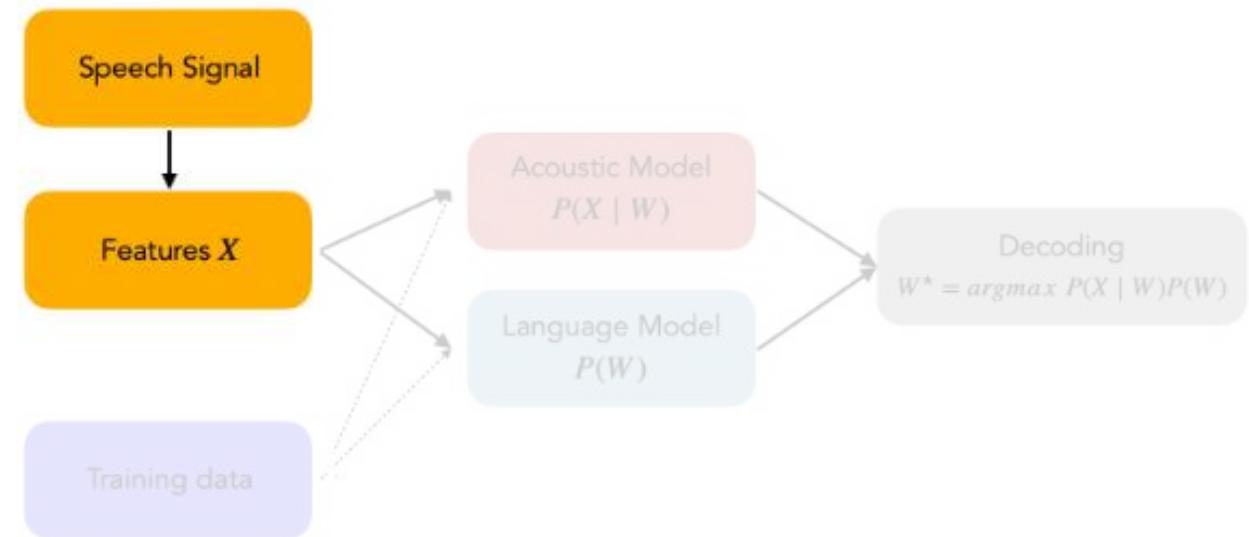
Inicialmente, los motores de ASR se componían de los siguientes módulos:



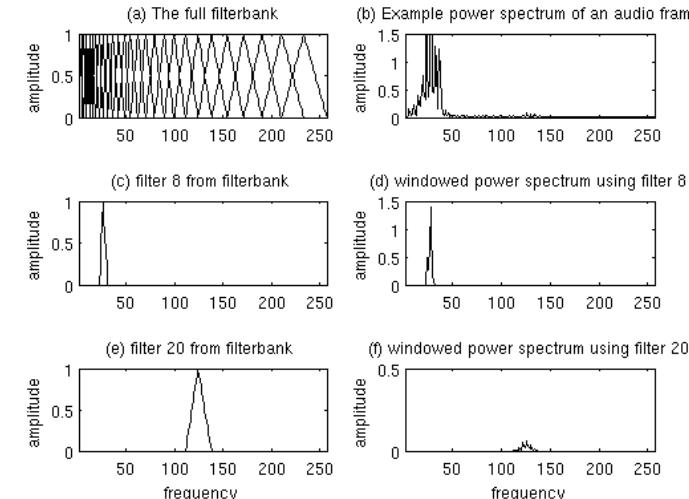
Componentes de un sistema de ASR. Extraído de [2].

ASR: Los inicios

1. **Captura de señal de audio:**
Micrófono codifica señal
2. **Extracción de *features*:** A partir de la codificación se obtiene vector de *features* ([MFCC](#), PLP)



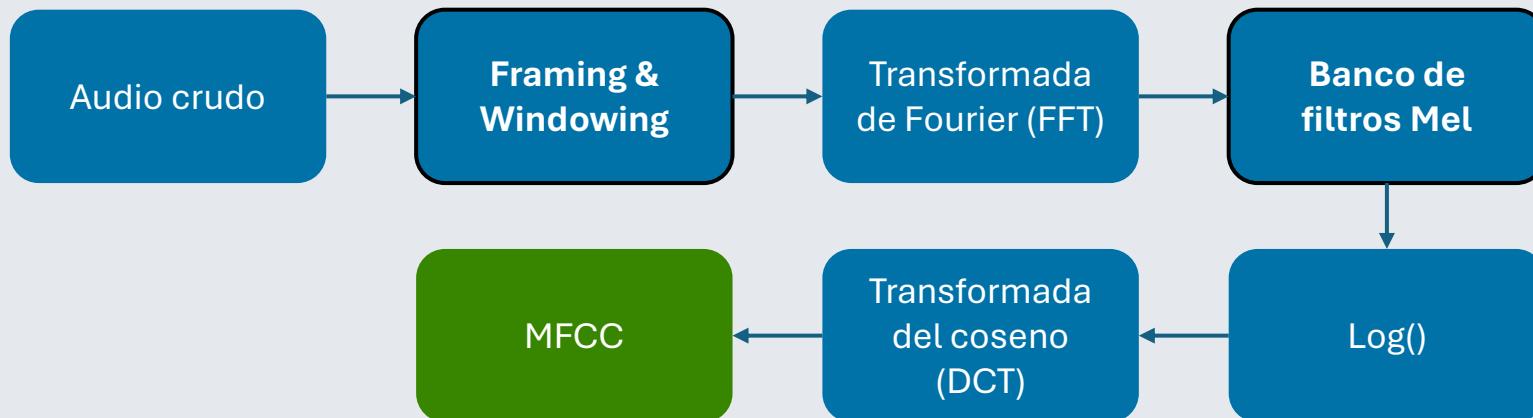
Componentes de un sistema de ASR. Extraído de [2].



Mel Filterbanks usados para calcular las features MFCC.
Extraído de [2].

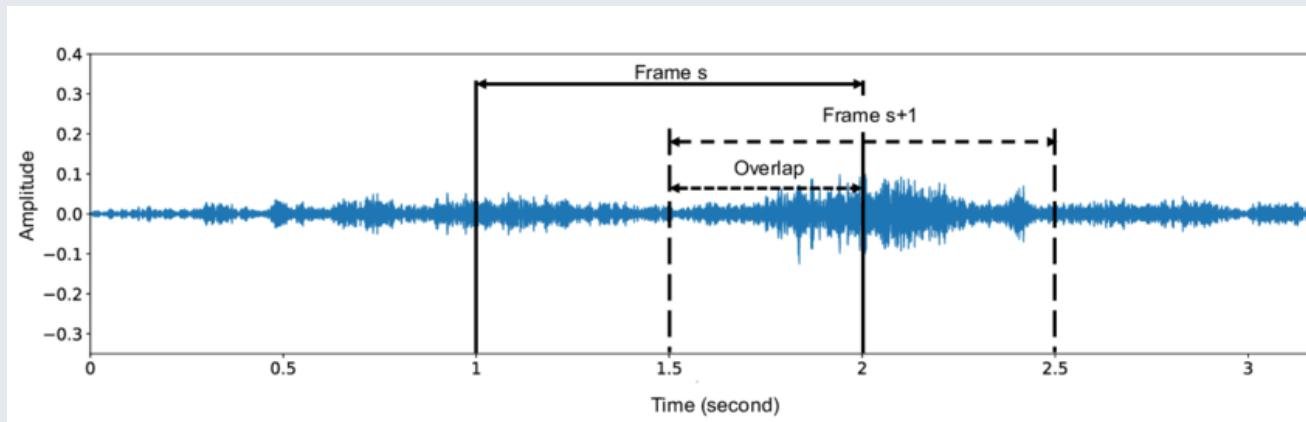
ASR: MFCC features

- Las features MFCC son una forma común de representar el audio. Su uso es muy común a la hora de realizar tareas de reconocimiento de audio, como la detección de enfermedades o la identificación de hablantes.
- Para extraer estas *features*, el proceso se compone de varios pasos. Por sencillez destacaremos los dos más importantes: el *windowing* y la aplicación de los filtros de Mel.



ASR: MFCC, Extracción

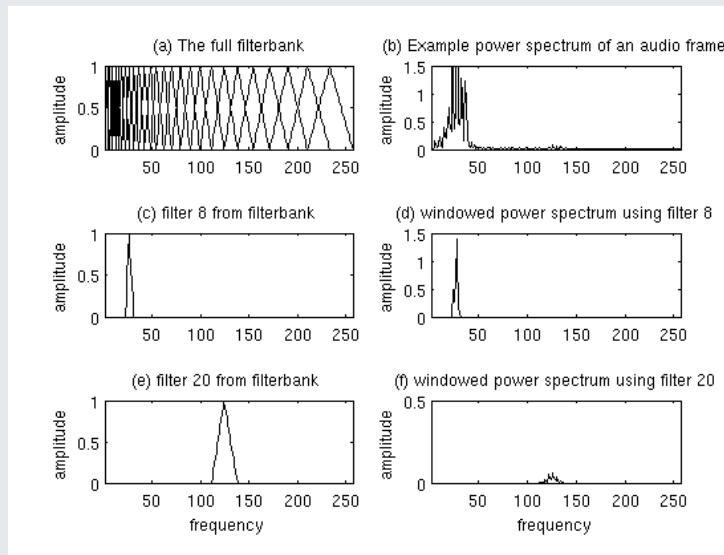
- **Windowing:** Un audio no se procesa como una unidad, sino que se separa en *frames*, mediante el proceso de *framing*. Después a cada *frame* se le aplica una ventana para reducir las discontinuidades en los bordes. Cada unidad del audio a analizar son las ventanas creadas tras el *windowing*.



Proceso de windowing. Extraído de [9]

ASR: MFCC, Extracción

- **Banco de filtros Mel:** Despues del *windowing*, se le aplica a cada ventana la transformada de Fourier para obtener el espectro de magnitud del *frame*. Este espectro es el que se pasa por un conjunto de filtros triangulares distribuidos en la escala *Mel* (escala perceptual del oido humano). Cada filtro extrae la energia en una banda de frecuencias.



Mel Filterbanks usados para calcular las features MFCC.
Extraido de [2].

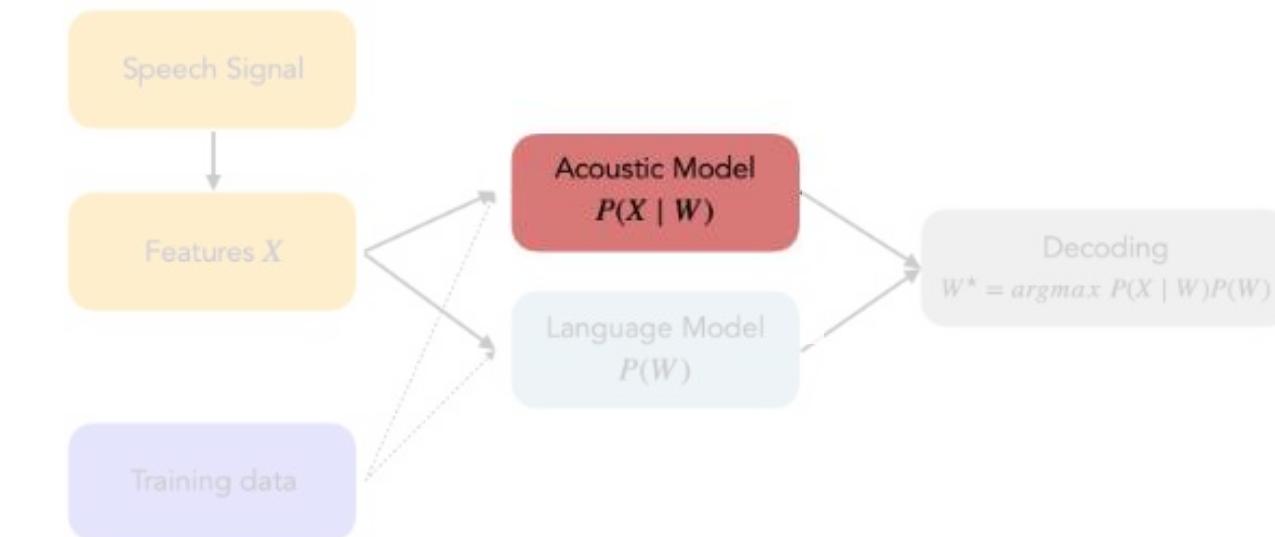
ASR: Los inicios

3. Modelo acústico (AM):

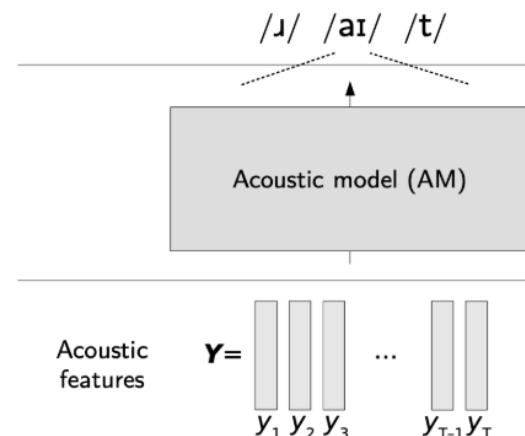
Clasifica *features* en unidades fonéticas (HMM [1], ANN). Fonemas se convierten en palabras con diccionario fonético.

Ejemplo de *diccionario fonético*:

- “cat” → K AE T
- “cut” → K AH T
- “saw” → S AO



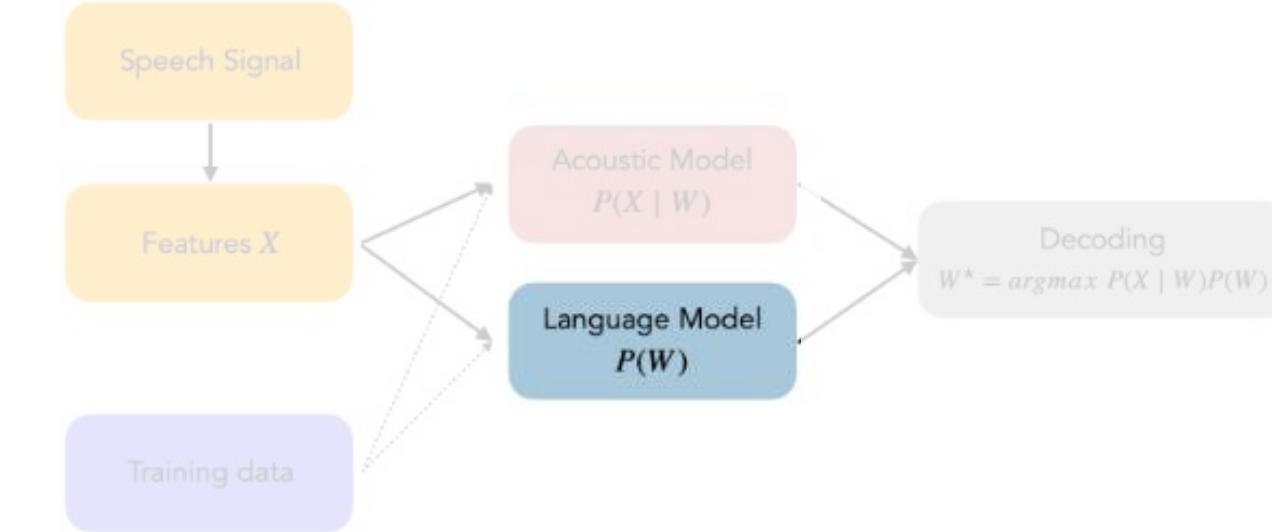
Componentes de un sistema de ASR. Extraído de [2].



Representación de un modelo acústico. Extraído de [2].

ASR: Los inicios

4. **Modelo de lenguaje (LM):** A partir de secuencias textuales del AM el LM favorece opción más probable según contexto. Una práctica común es el uso de un [n-gram model](#).



Componentes de un sistema de ASR. Extraído de [2].

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

Here are the calculations for some of the bigram probabilities from this corpus

$$\begin{aligned} P(I|<s>) &= \frac{2}{3} = 0.67 & P(Sam|<s>) &= \frac{1}{3} = 0.33 & P(am|I) &= \frac{2}{3} = 0.67 \\ P(</s>|Sam) &= \frac{1}{2} = 0.5 & P(Sam|am) &= \frac{1}{2} = 0.5 & P(do|I) &= \frac{1}{3} = 0.33 \end{aligned}$$

Ejemplos de probabilidades de un modelo de bigramas (2-gram model). Extraído de [3].

ASR: ¿Es suficiente el modelo acústico?

- NO! El modelo acústico puede confundir sonidos similares.
- El modelo de lenguaje (LM) usa el contexto para elegir la palabra más probable.

Ejemplo:

Fonema → /aɪ skju:m/

¿Cuál es su transcripción?

Audio: /aɪ skju:m/
Possible transcripción 1: “I scream”
Possible transcripción 2: “Ice cream”

Contexto 1:
“Every time I see a spider,...”

El LM favorece:
“I scream”

Contexto 2:
“After dinner, I want some...”

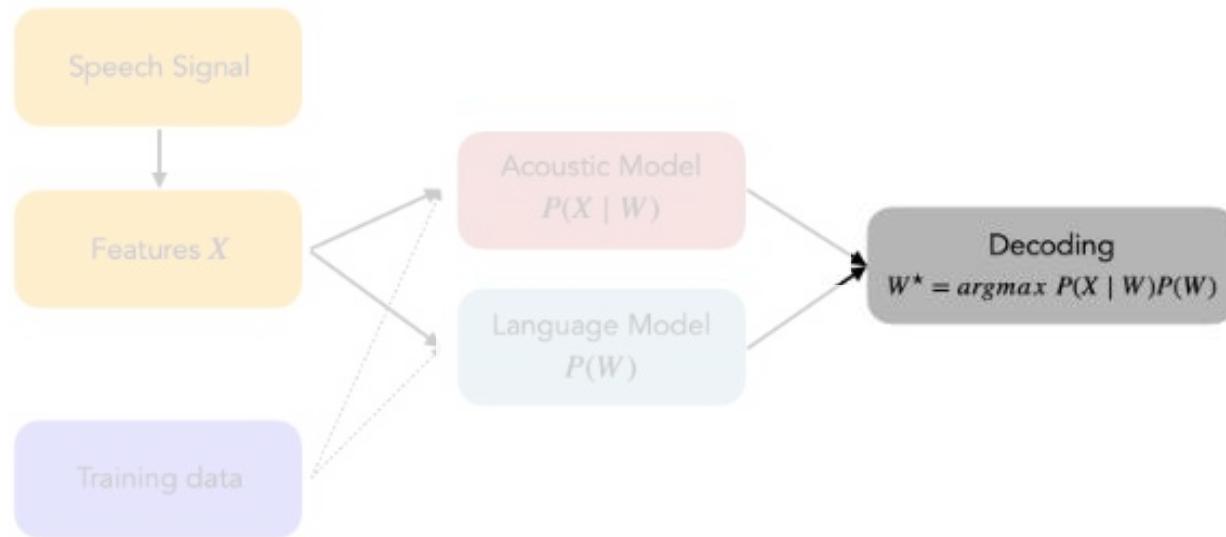
El LM favorece:
“Ice cream”

ASR: Los inicios

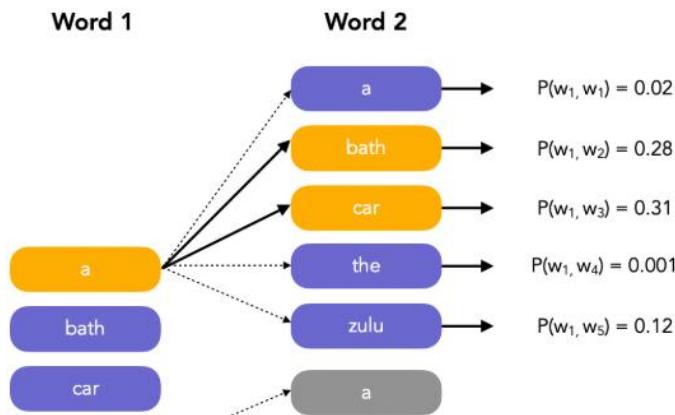
5. Decodificación: Beam Search [4] sobre el vocabulario de nuestro modelo maximizando la probabilidad:

$$\operatorname{argmax} P(X|W) * P(W)$$

$$P(X|W) = \text{prob. del AM}$$
$$P(W) = \text{prob. del LM}$$



Componentes de un sistema de ASR. Extraído de [2].



Ejemplo de *beam search*. Las cajas amarillas son las palabras escogidas para continuar la búsqueda. Extraído de [2].

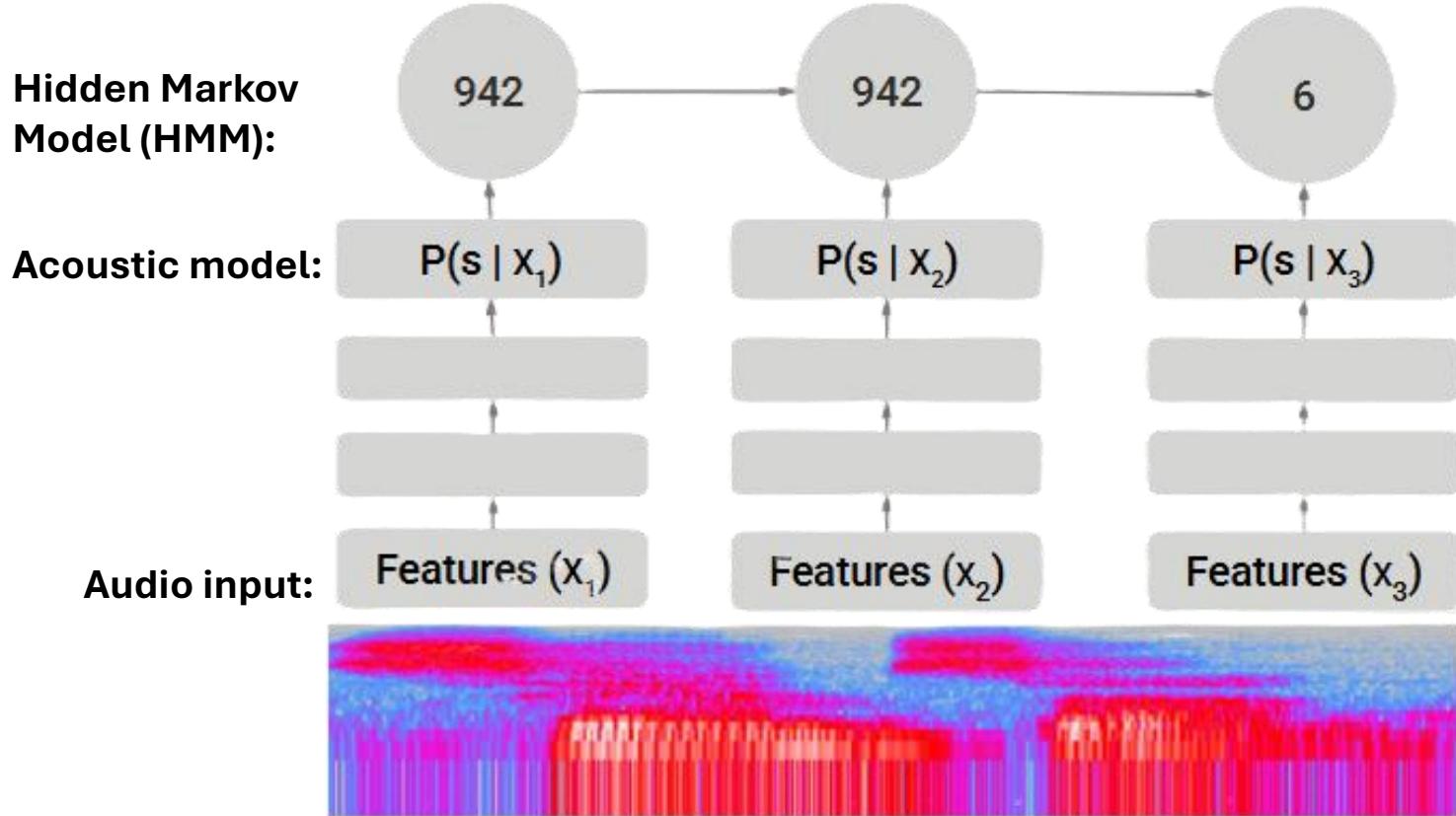
El progreso del ASR

El progreso: HMM-DNN

Se avanza hacia el uso de redes neuronales. La primera aproximación que incluyeron estas fueron los modelos HMM-DNN [5], dónde el modelo acústico es una *deep neural network* en vez de una GMM (mezcla gaussiana).

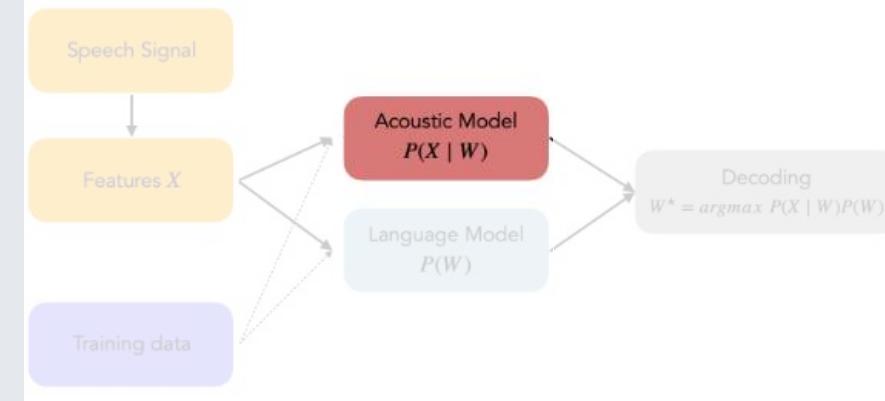
Posteriormente se fueron introduciendo en los modelos de ASR las RNNs, LSTMs y GRUs.

Pronunciation: Samson
Pronunciation: S – AE – M – S – AH – N
Subphones: 942 – 6 – 37 – 8006 – 4422 - ...



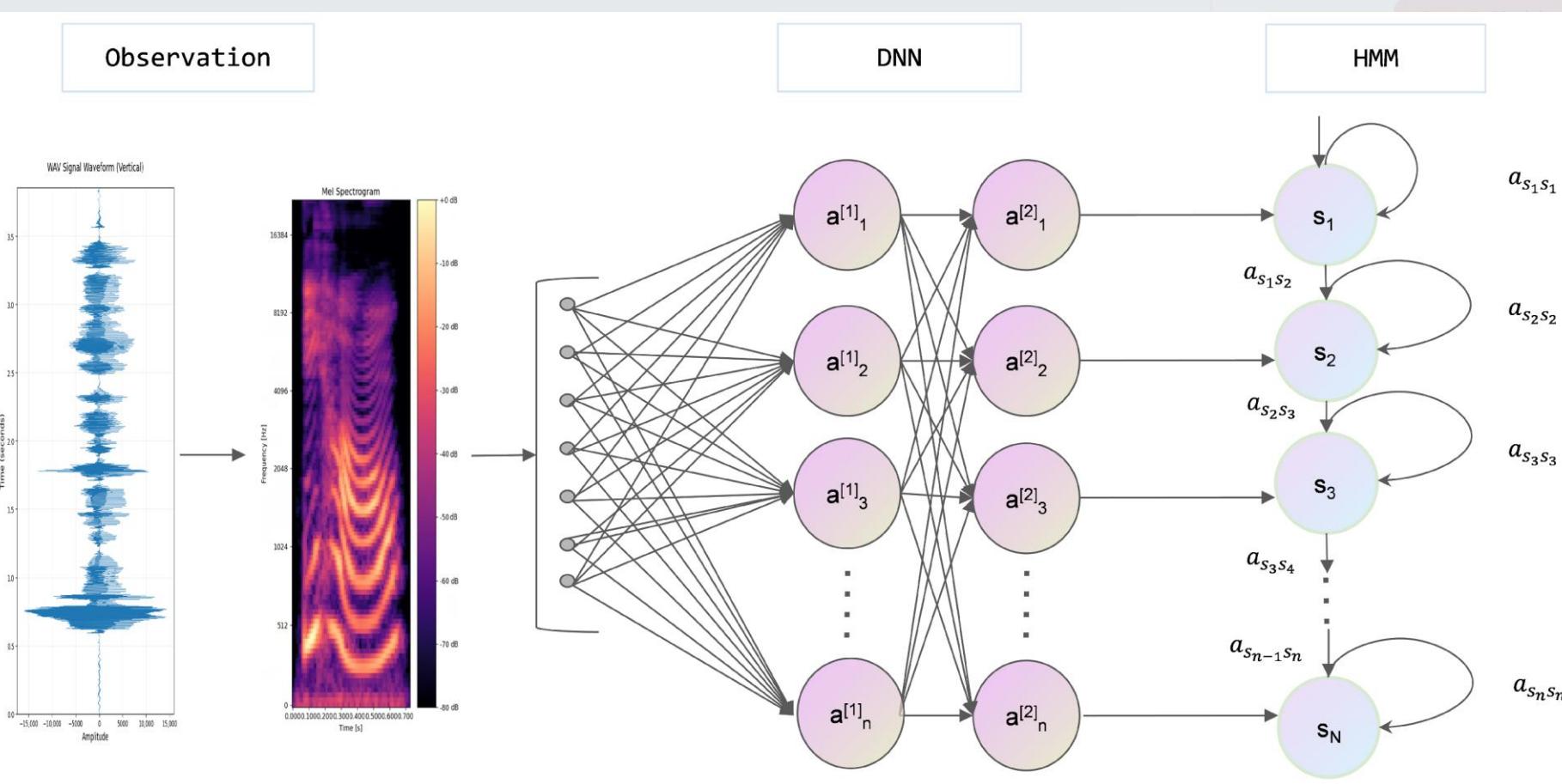
Arquitectura HMM-DNN. Extraído de [5].

El progreso: HMM-DNN



- La idea es introducir como entrada al modelo acústico (DNN) un vector de features MFCC, y este genera como salida no fonemas, sino probabilidades posteriores de estados HMM. Es decir, la DNN estima la **probabilidad de que un frame corresponda a cierto estado fonético**.
- El HMM define como de probable es pasar de un estado a otro. Cada fonema tiene tres estados: inicio, medio, final.

Lo que predice la DNN es la parte del fonema (inicio, medio, fin) que se está pronunciando.



- La idea de fechas probadas probadas
- El HMM foneo

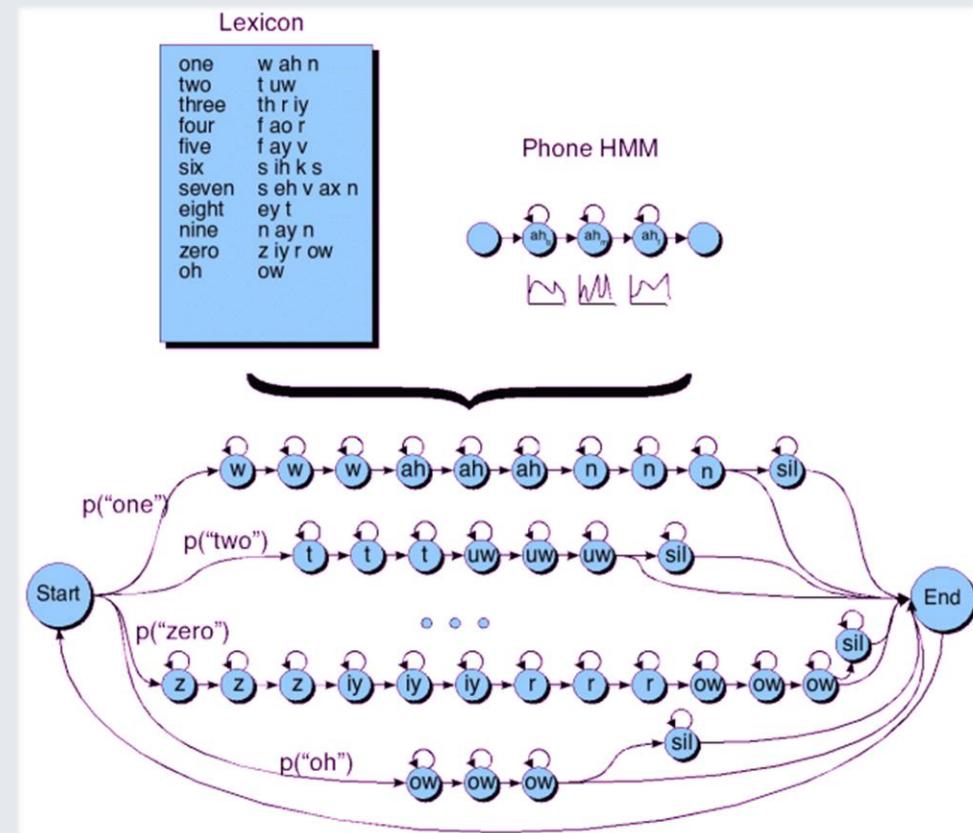
Estructura interna de la DNN y del HMM que componen una arquitectura HMM-DNN. Extraído de Wu et al. (2025). *Integrating Speech Recognition into Intelligent Information Systems: From Statistical Models to Deep Learning. Informatics.*

El progreso: HMM-DNN

i ¿Qué es un HMM?

Modelo probabilístico que representa procesos que evolucionan en el tiempo con estados ocultos y observaciones visibles.

En ASR, modela el orden de los sonidos, decide cuanto dura cada fonema, nos ayuda a alinear frames con fonemas, y permite buscar la secuencia más probable de fonemas que corresponde al audio.



Ejemplo de HMM para reconocimiento de dígitos.
Extraído de [5].

El progreso: HMM-DNN

Limitaciones

1. **Dependencia de un léxico diseñado manualmente.** Requiere ingeniería lingüística, lo que introduce suposiciones rígidas sobre la pronunciación.
2. **Pipeline de entrenamiento en múltiples etapas.** Primero se debe comenzar con un modelo HMM-GMM para obtener alineaciones iniciales, y luego usar esas alineaciones para entrenar el modelo acústico DNN.
3. **La modularidad limita la optimalidad.** Los sistemas HMM-DNN dividen el pipeline en distintos módulos que se entranan por separado, lo que conduce a una solución globalmente subóptima.

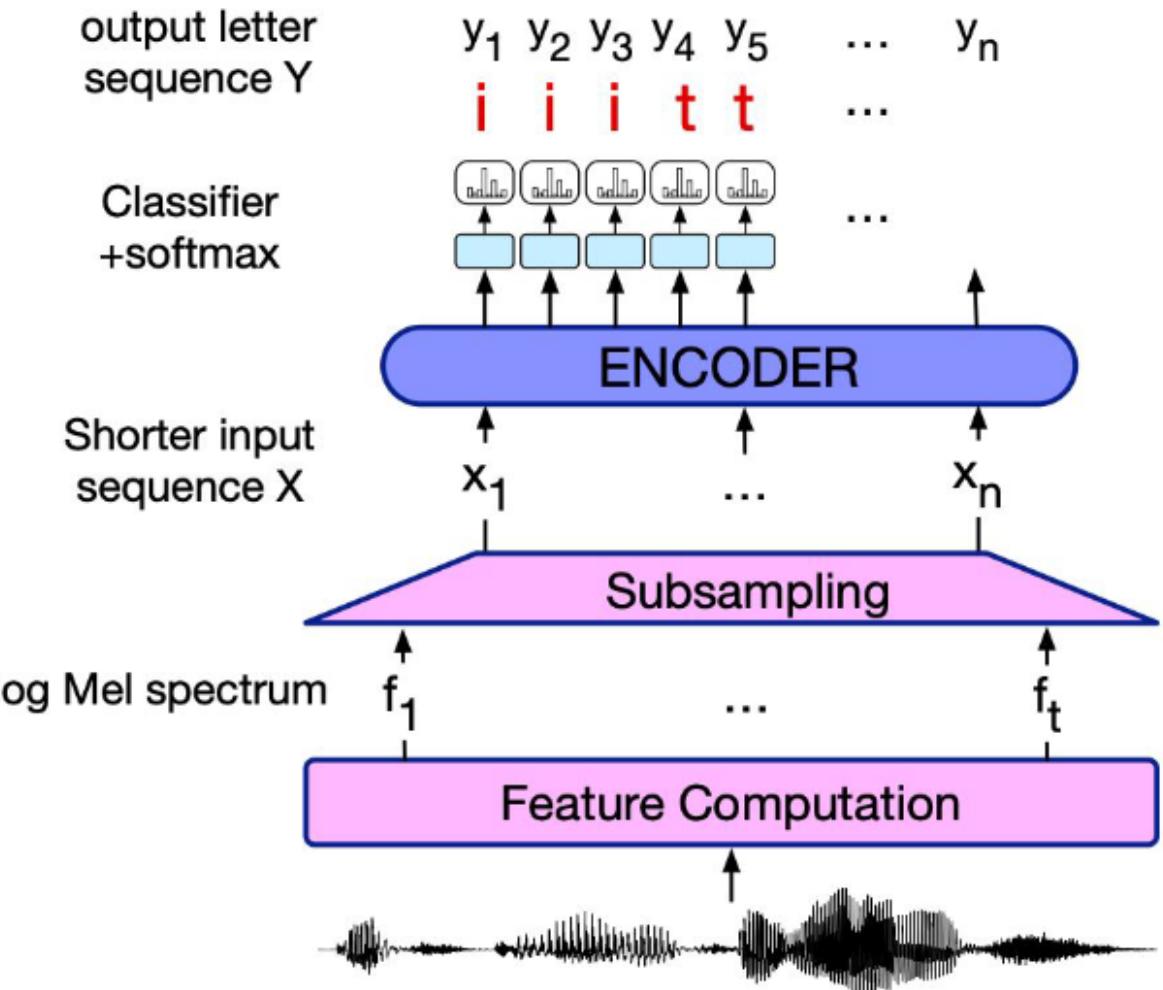
¿Puede el Deep Learning reemplazar completamente el uso de HMMs?

El progreso: CTC

El ASR progresó hacia las arquitecturas [CTC](#) [5], que predecían outputs textuales directamente de las entradas de audio.

Estos modelos predicen un carácter para cada timestamp t . Cada t es independiente del resto, una predicción no depende de la anterior.

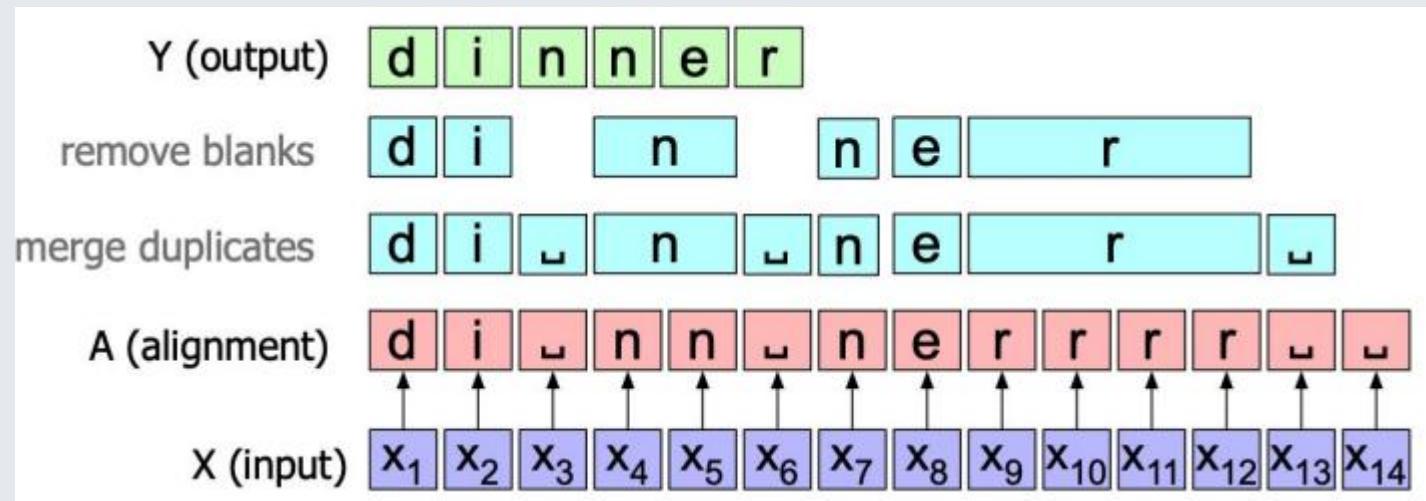
Aquí ya se puede observar la entrada de los mecanismos de Atención y de la arquitectura [Transformer](#) [3].



Arquitectura CTC. El módulo ENCODER es un Transformer Encoder. Extraído de [5].

El progreso: CTC

- Primero, en cada timestamp t el modelo produce un token de salida que puede ser un carácter o un “blank” (vacío)
- Después se aplica la regla de colapsamiento para generar la salida → Ignorar todos los “blanks”, e ignorar todas las salidas repetidas del mismo token



Regla del colapsamiento en CTC. Extraído de [3].

El progreso: CTC

Limitaciones

- **Asunción de independencia.** Asumir que la salida t es completamente independiente de la $t-1$ limita la precisión de estos sistemas, ya que no se apoyan del contexto predicho hasta el momento para refinar las predicciones.

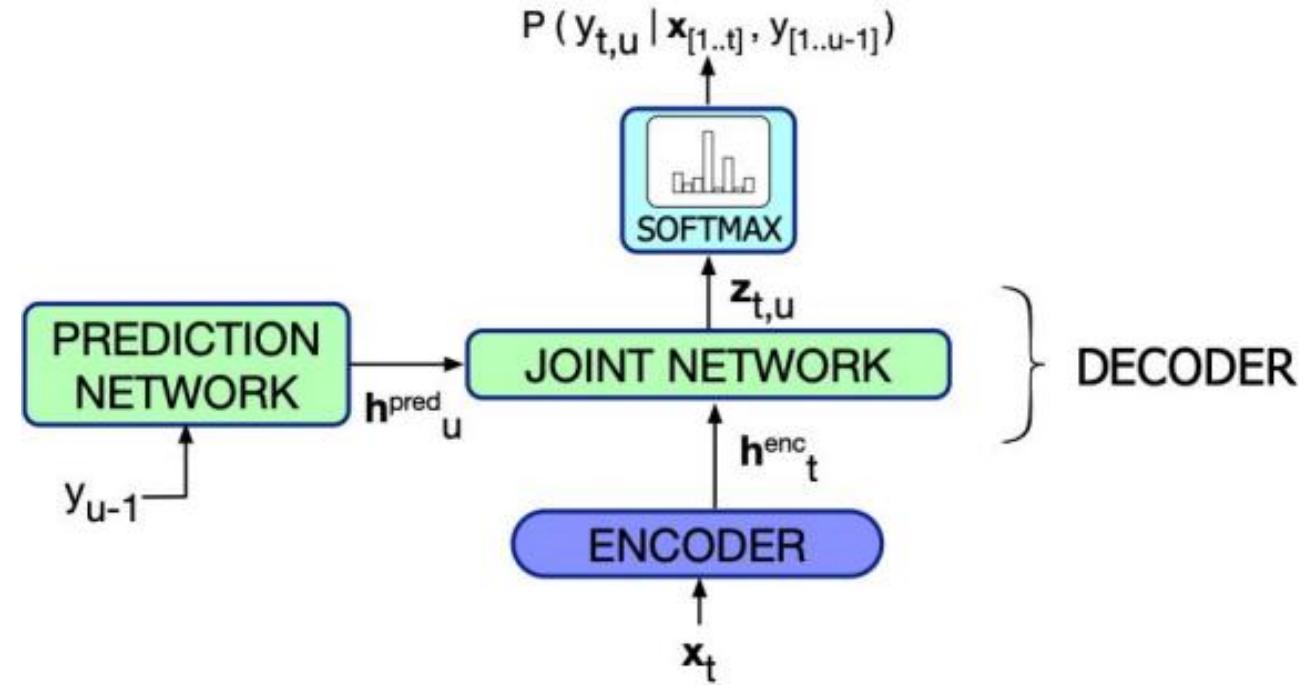
Ventajas

- **Streaming.** La asunción de independencia lo hace muy conveniente para usarse en *streaming*. *Streaming* significa reconocer las palabras online, en vez de esperar a que acabe toda la secuencia. Esto es crucial en aplicaciones donde queremos empezar el reconocimiento mientras que el usuario está hablando.

El progreso: RNN-Transducer

Para eliminar la asunción de independencia de las CTC, se introdujo el [RNN-Transducer](#) [8]. Este se compone de dos partes: un modelo acústico CTC y un modelo de lengua separado llamado *predictor*.

En cada timestamp t , el CTC produce una representación h_t dada la entrada $x_1 \dots x_n$. Después el *predictor* produce otro h_u a partir de los tokens de salida anteriores (contexto). Finalmente, h_t y h_u se introducen en una tercera red a cuya salida se le aplica una softmax para predecir el siguiente carácter.



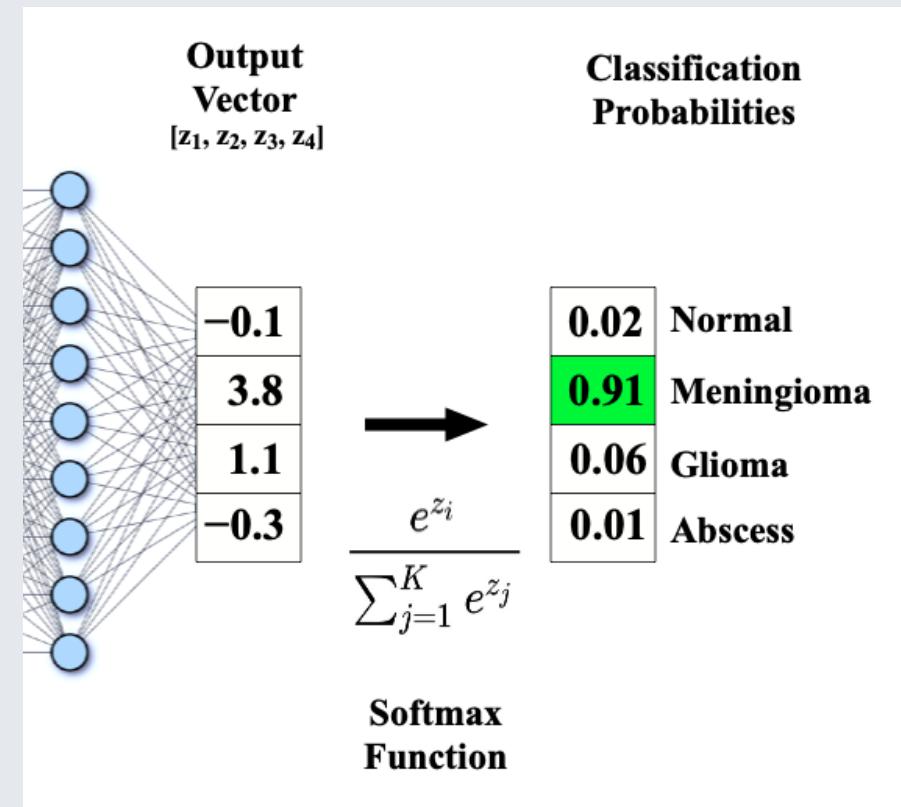
Arquitectura del RNN-Transducer. Extraído de [3].

El progreso: RNN-Transducer

¿Qué es una softmax?

La capa softmax normaliza los valores de salida obtenidos por una red neuronal (comúnmente llamado *logits*) dentro del rango [0,1].

Esto convierte los *logits* en probabilidades por clase fácilmente interpretables.

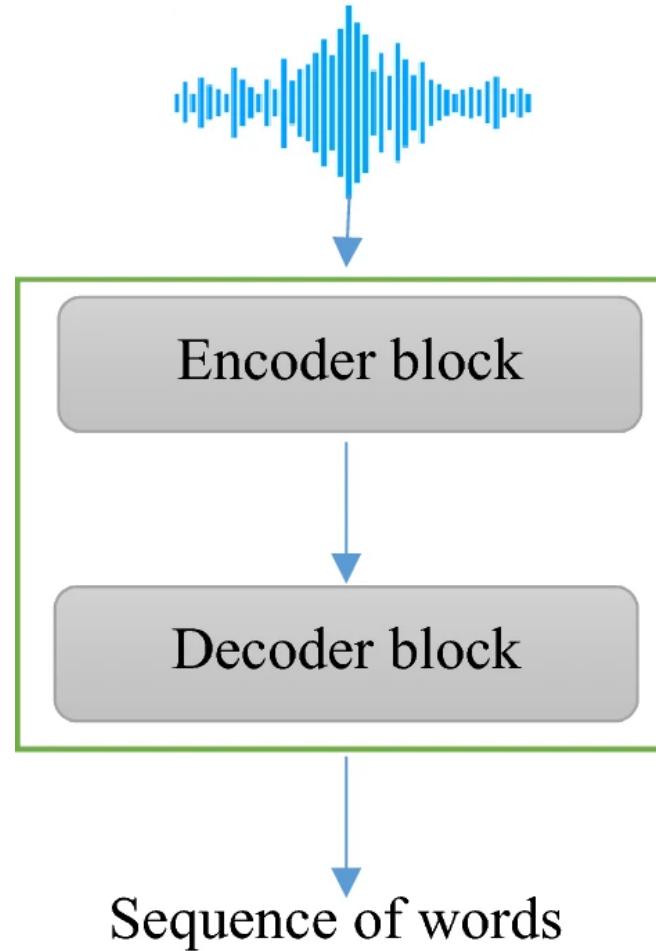


Ejemplo de clasificación con softmax. Extraído de <https://mriquestions.com/softmax.html>.

El progreso: Sequence-to- sequence

Y llegamos hasta los modelos sequence-to-sequence (S2S), donde el modelo es capaz de predecir secuencias enteras a partir del audio de entrada.

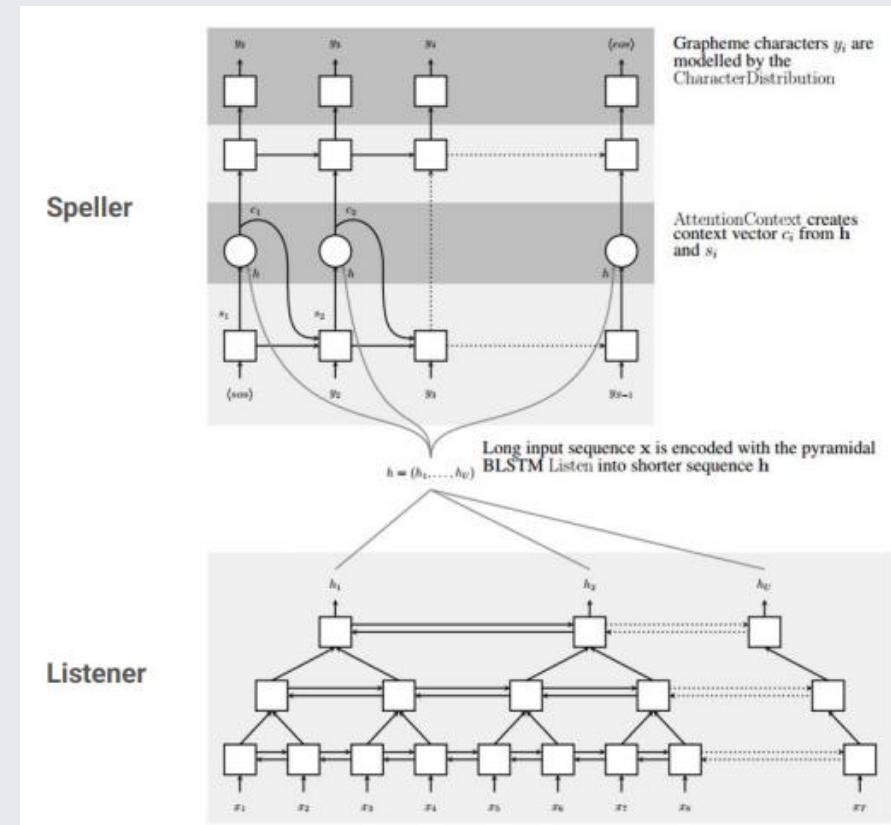
Los mejores resultados se logran con modelos Transformer o modelos híbridos CNN/Transformer. Con estos se consiguió sobre el 2016-2017 llegar a un nivel de transcripción de “paridad con humanos” con tasas de error de palabra ~5%.



Arquitectura genérica de un modelo S2S para ASR. Extraído de [6]

El progreso: Sequence-to-sequence

- El decoder suele incluir mecanismos de atención. Esto marcó el inicio de las arquitecturas encoder-decoder modernas.
- A diferencia de CTC, las S2S solo producen caracteres (no “blanks”).
- Una salida t , está condicionada por las salidas anteriores, es decir, por su contexto.



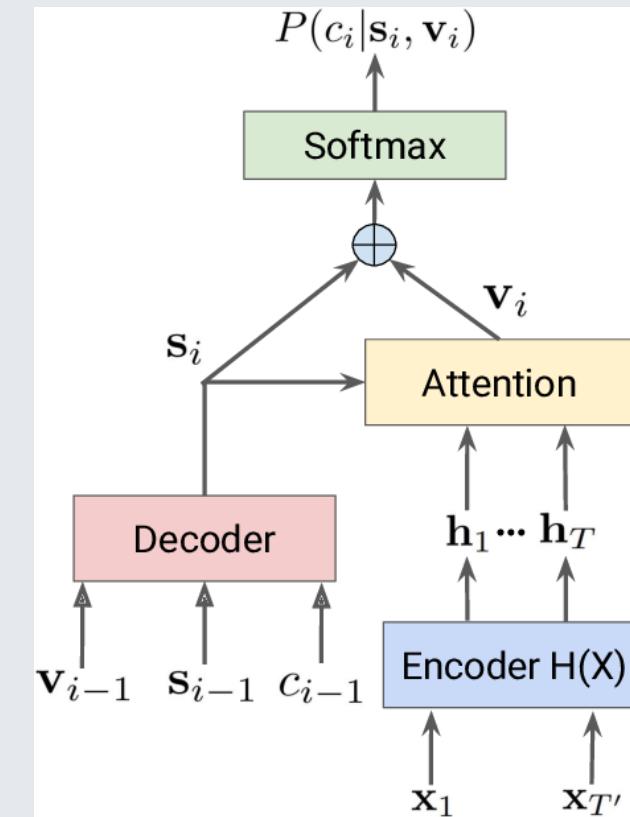
Arquitectura interna de un modelo S2S. En ASR, al encoder se le puede llamar Listener y al decoder Speller. Extraído de [6]

El progreso: Sequence-to-sequence

¿Qué es la atención?

El mecanismo de atención permite que el modelo, en vez de mirar todo el audio a la vez, se enfoque solo en las partes más relevantes en cada momento.

Dentro de una arquitectura *encoder-decoder* para ASR, el encoder transforma el audio en una serie de representaciones, y el decoder las va utilizando para generar el texto. La atención actúa como una “linterna” que, en cada paso, ilumina la parte del audio que es más útil para predecir la siguiente letra o palabra.



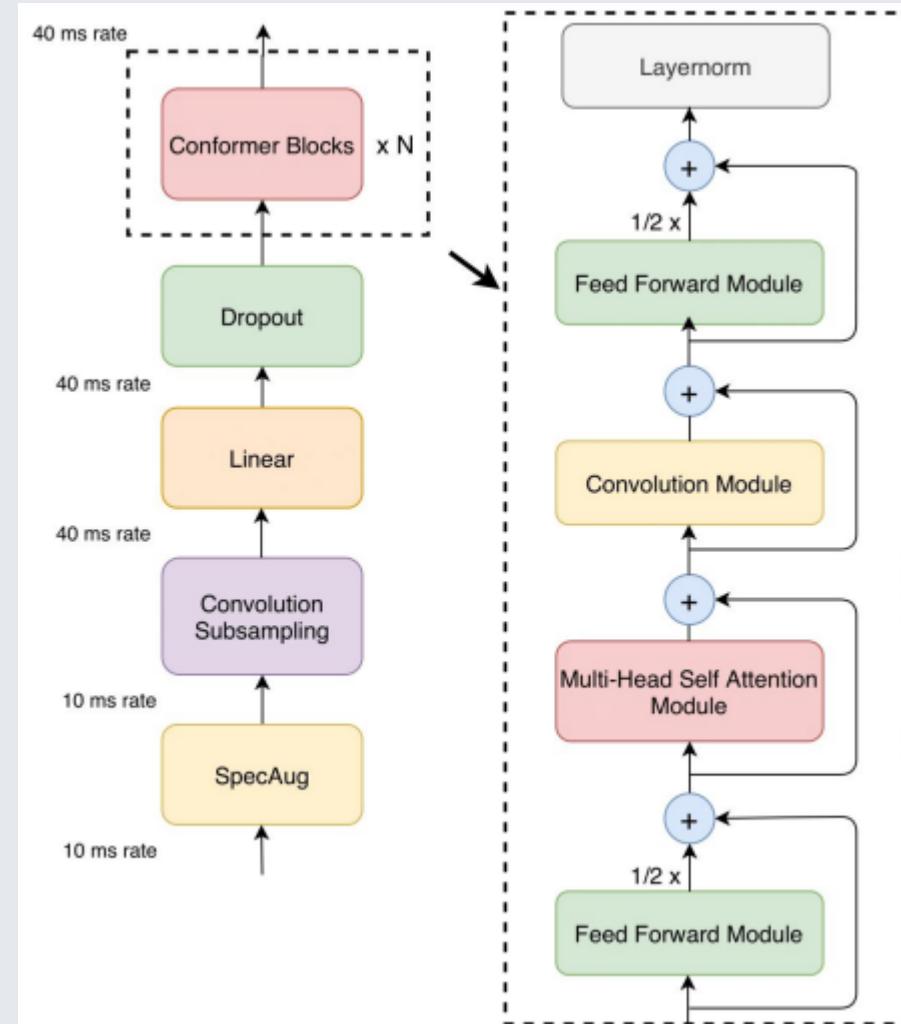
Arquitectura encoder-decoder con uso de atención. Extraído de Prabhavalkar et al. (2023). *End-to-end speech recognition: A survey*

Estado del arte del ASR

Estado del arte: Conformer (Google)

La arquitectura **Conformer** (Convolution-augmented Transformer) nació para mejorar el rendimiento en ASR combinando lo mejor de dos mundos:

- La capacidad del **Transformer** para capturar dependencias globales en secuencias largas
- La fortaleza de las **convoluciones** para modelar patrones locales y variaciones temporales del habla.

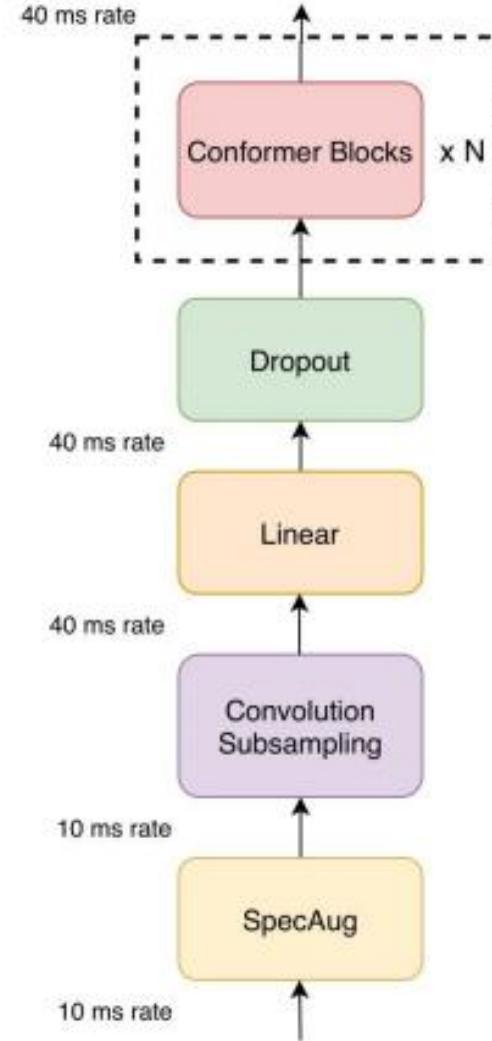


Arquitectura del modelo Conformer. Extraído de [5]

Estado del arte: Conformer, Encoder

Se compone de diferentes partes:

- **SpecAug:** técnica de *data augmentation* aplicada directamente a los espectrogramas del audio. En lugar de modificar la señal de audio original, se distorsionan ciertas partes del espectrograma.
- **ConvSubsampling:** Es una forma de *reducir la longitud de la secuencia de entrada*. En vez de procesar cada frame de audio uno por uno (lo cual es costoso), se aplican convoluciones con stride (saltos).

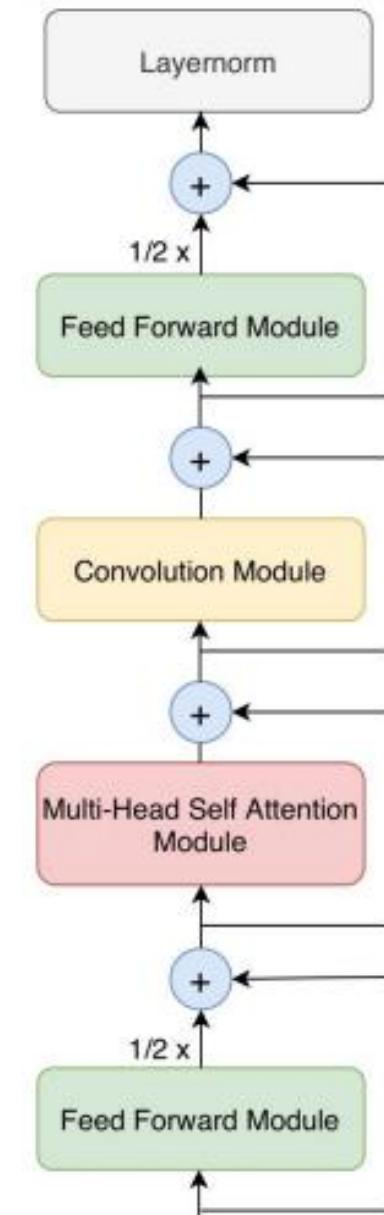


Arquitectura general del modelo Conformer. Extraído de [5]

Estado del arte: Conformer, Encoder

Se compone de diferentes partes:

- **Bloque Conformer (Conformer encoder):** Evolución del bloque Transformer diseñado específicamente para audio. Añade un bloque convolucional para capturar contexto local e introduce dos redes feedforward en lugar de una.



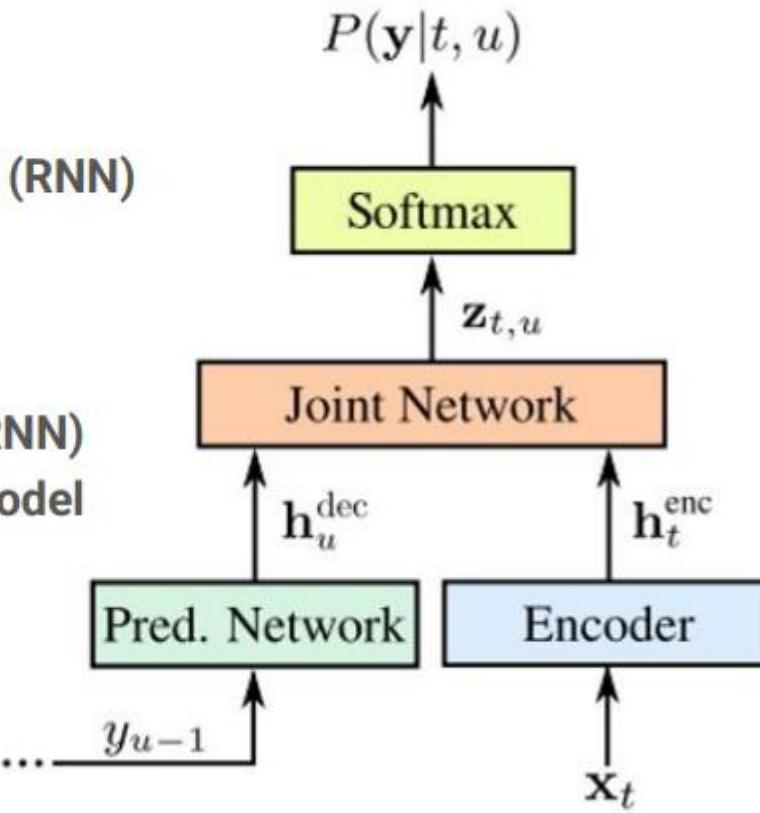
Arquitectura del bloque Conformer. Extraído de [5]

Estado del arte: Conformer, Decoder

¡Conformer no incluye ningún bloque decoder en su publicación original! El Conformer es una arquitectura que solo abarca la parte del encoder.

Para poder completar pipelines de ASR usando el Conformer, es común acompañar a este modelo de un Transformer decoder como el de Whisper o de una arquitectura RNN-T compuesta de redes recurrentes.

1 hidden layer LSTM (RNN)
3 hidden layer LSTM (RNN)
Pre-trained as language model



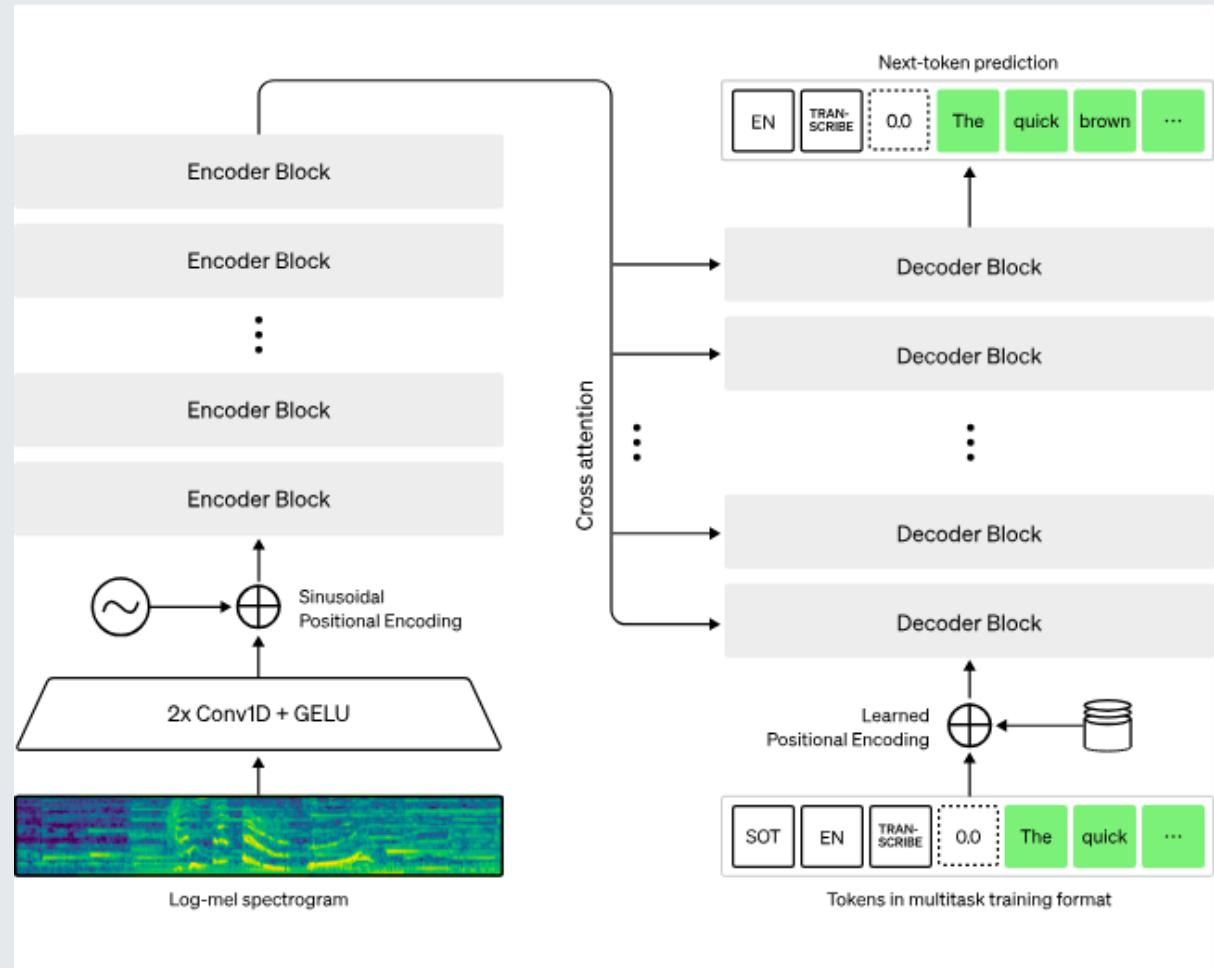
Possible decoder RNN-T para el modelo Conformer. Extraído de [5]

Estado del arte: Whisper (OpenAI)

Modelo *open-source* entrenado con 680.000 horas de audio multitarea y multilingüe. Es muy robusto a acentos, ruido de fondo y lenguaje técnico.

Además de transcribir en múltiples idiomas, puede también traducir del idioma hablado al inglés y generar marcas de tiempo por fragmento de frase.

Su arquitectura es la de un Transformer *encoder-decoder* estándar. Destaca el uso de una CNN para extraer *features* acústicas a partir de un espectrograma de Mel.



Arquitectura del modelo Whisper. Extraído de <https://openai.com/index/whisper/>

Estado del arte: Whisper

¿Qué puede hacer Whisper?

- **Transcripción inglés**

 “Ask not what your country...” →  “Ask not what your country...”

- **Transcripción y traducción al inglés**

 “El rápido zorro marrón...” →  “The quick brown fox...”

- **Transcripción de cualquier otro idioma**

 “Un caffè e una torta per favore...” →  “Un caffè e una torta per favore...”

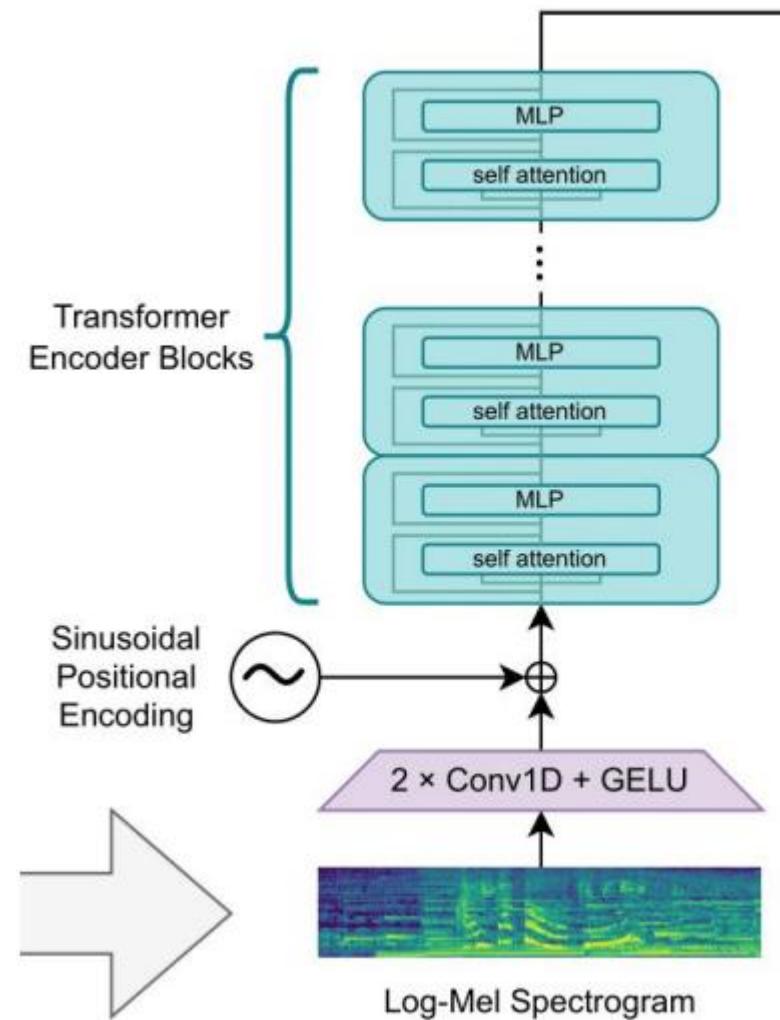
- **Identificar cuando no hay hablantes**

 (música de fondo) →  Ø

Estado del arte: Whisper, Encoder

El encoder se compone de tres partes principales

- **Entrada:** Es un spectrograma log-mel acompañado por dos capas convolucionales
- **Codificación posicional:** Es común que a la entrada de un Transformer se le aplique un embedding posicional para que cada parte del input mantenga información posicional.
- **Capa Transformer**

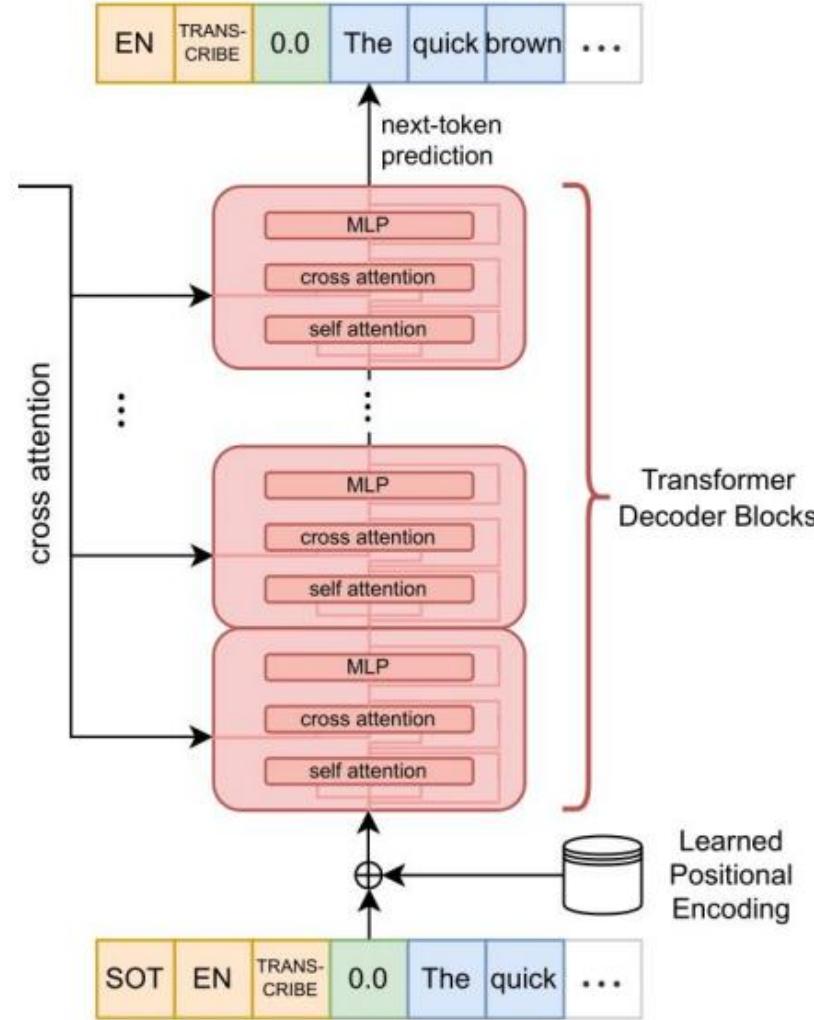


Arquitectura del Transformer Encoder de Whisper.
Extraído de [5]

Estado del arte: Whisper, Decoder

En el decoder destaca lo siguiente

- **Estilo GPT-2:** El decoder es el mismo que se utiliza como decoder en la arquitectura GPT-2
- **Vocabulario/tags personalizados:** Para establecer la tarea a realizar Whisper añade tokens que “configuran” la tarea que se va a realizar en la inferencia (p. ej. Etiquetas de lenguaje como <en> o <fr>, etiquetas de tarea como <transcribe> o <translate>, etiquetas de habla como <nospeech>)



Arquitectura del Transformer Decoder de Whisper.
Extraído de [5]

Estado del arte: Whisper

Ejemplo de salida

Ground truth: Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop.

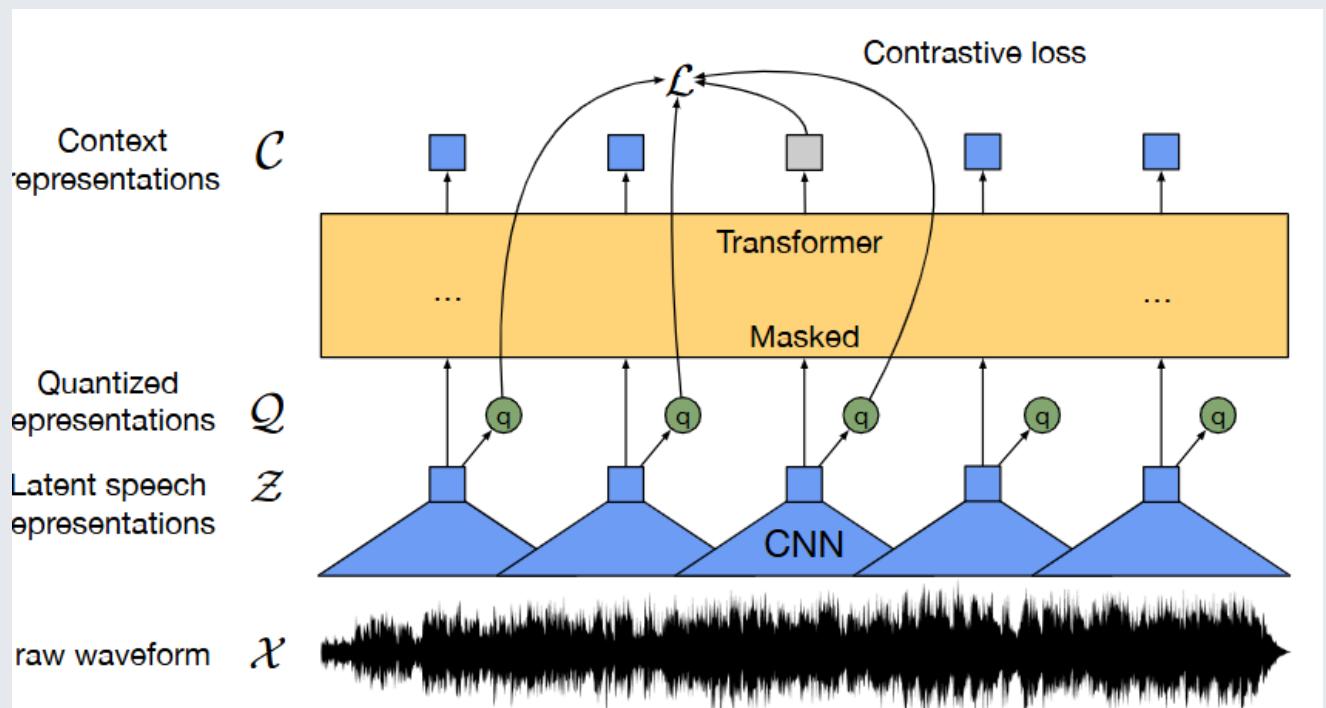
Whisper (French): ['<|startoftranscript|>', '<|fr|>', '<|transcribe|>', '<|notimestamps|>', 'Comme', 'moi', ',', 'vous', 'avez', 'peut', '-', 'être', 'déjà', 'v', 'éc', 'u', 'cette', 'scène', '..', 'Qu', "", 'est', '-', 'ce', 'que', 'tu', 'fais', 'cet', 'été', 'toi', '?', 'É', 'c', 'oute', '..', 'je', 'ne', 'sais', 'pas', 'encore', 'trop', '<|endoftext|>']

Whisper (Translation): ['<|startoftranscript|>', '<|fr|>', '<|translate|>', '<|notimestamps|>', 'Like', 'me', ',', 'you', 'may', 'have', 'already', 'experienced', 'this', 'scene', '..', 'What', 'are', 'you', 'doing', 'this', 'summer', '?', 'I', 'don', '"t', 'know', 'yet', '<|endoftext|>']

Estado del arte: Wav2Vec (Meta)

Modelo que aprende representaciones del habla de forma auto-supervisada, es decir, aprende a predecir unidades enmascaradas del habla, en vez de aprender a partir de datos etiquetados.

Es especialmente útil para la transcripción de lenguajes minoritarios donde hay una cantidad muy limitada de datos etiquetados.



Arquitectura del modelo Wav2vec. Extraído de <https://huggingface.co/blog/fine-tune-wav2vec2-english>

ASR en la práctica

ASR en la práctica: Herramientas

- **Servicios en la nube:** APIs web donde se envía el audio y se recibe texto transcritto de alta calidad. Implican costo y hay que considerar demoras por latencia (*Google Speech-to-text*, *Amazon Transcribe*, *Azure Speech*).
- **Herramientas open-source:** Repositorios de GitHub o frameworks que proveen de los modelos necesarios para realizar la tarea de STT (*DeepSpeech*, *SpeechBrain*, *HuggingFace Transformers*)
- **Whisper de OpenAI:** Mención especial a Whisper por su facilidad de uso gracias a que han desarrollado una biblioteca nativa de Python (*openai-whisper*).

ASR en la práctica: Herramientas

Google

- Servicio de transcripción por lenguaje

Speech-to-Text TRANSCRIBE



DeepSpeech

amazon



texto
horas
Azure Spech).
que

SpeechBrain



Hugging Face



ASR en la práctica: Consideraciones I

-  **Calidad de audio:** Un audio claro, con buen volumen y poco ruido, dará mucho mejor resultado. Muchos sistemas implementan filtrado de ruido antes de la transcripción.
-  **Tiempo real vs batch:** Para interfaces interactivas, a veces se necesita transcripción palabra por palabra mientras el usuario habla (p.ej., subtítulos en directo). Técnicas como *endpointing* (detectar pausas para saber cuándo terminar de transcribir) y *streaming ASR* (emitir resultados parciales a medida que llega audio) son importantes.
-  **Puntuación y formato:** La salida de un ASR suele ser texto sin puntuación ni mayúsculas. Algunos avanzados agregan comas y puntos automáticamente usando modelos de puntuación. En otros casos se necesita post-procesar la transcripción (por ejemplo, convertir “siete a m” a “7:00 AM”).

ASR en la práctica: Consideraciones II

-  **Métricas:** El estándar es la Word Error Rate (WER), que calcula porcentaje de palabras equivocadas en la transcripción comparado con un “ground truth”. Más allá de WER, también es importante la **latencia** (¿tarda medio segundo o 5 segundos en responder el asistente?), y la **robustez** ante diferentes locutores.
-  **Entrenamiento personalizado:** Si tu aplicación es muy específica (ej: reconocimiento de comandos en entornos industriales con ruido muy particular), podrías necesitar entrenar o afinar un modelo con datos propios.

Referencias

- [1] Hidden Markov Models (2025). ScienceDirect.
<https://www.sciencedirect.com/topics/neuroscience/hidden-markov-model>
- [2] Mael Fabien (2020, Mayo 26) Introduction to Automatic Speech Recognition (ASR).
https://maelfabien.github.io/machinelearning/speech_reco/#
- [3] Jurafsky, D., & Martin, J. H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/>
- [4] Matt Payne (2021, Sept 29). What is Beam Search? Explaining The Beam Search Algorithm. Width.ai.
<https://www.width.ai/post/what-is-beam-search>
- [5] Andrew Maas (2025). Spoken Language Processing (CS224S). Stanford.
<https://web.stanford.edu/class/cs224s/semesters/2025-spring/syllabus>
- [6] Mamyrbayev Orken et al. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. Scientific Reports. <https://doi.org/10.1038/s41598-022-12260-y>
- [7] Alexei Baevski et al. (2020). Wav2vec 2.0: A framework for Self-Supevised Learning of Speech Representations. <https://arxiv.org/pdf/2006.11477>
- [8] Loren Lugosch (2020). Sequence-to-sequence learning with Transducers.
<https://lorenlugosch.github.io/posts/2020/11/transducer/>

Referencias

- [9] Abdoli, Sajjad & Cardinal, Patrick & Koerich, Alessandro. (2019). End-to-End Environmental Sound Classification using a 1D Convolutional Neural Network. *Expert Systems with Applications*.
<https://www.sciencedirect.com/science/article/pii/S0957417419304403>

Interacción Persona-Máquina

U7: Interfaces basadas en VR/AR

David Mulero Pérez
Pablo Ruiz Ponce

Universidad de Alicante

2025-2026

Grado en Ingeniería en Inteligencia Artificial



Universitat d'Alacant
Universidad de Alicante

— ○ • ○ ○ —
3DPERCEPTIONLAB

distic

Contenidos

- Realidad Virtual
- Realidad Aumentada
- Realidad Mixta
- Otras interfaces inmersivas
- Conclusiones

VR, MR, AR?

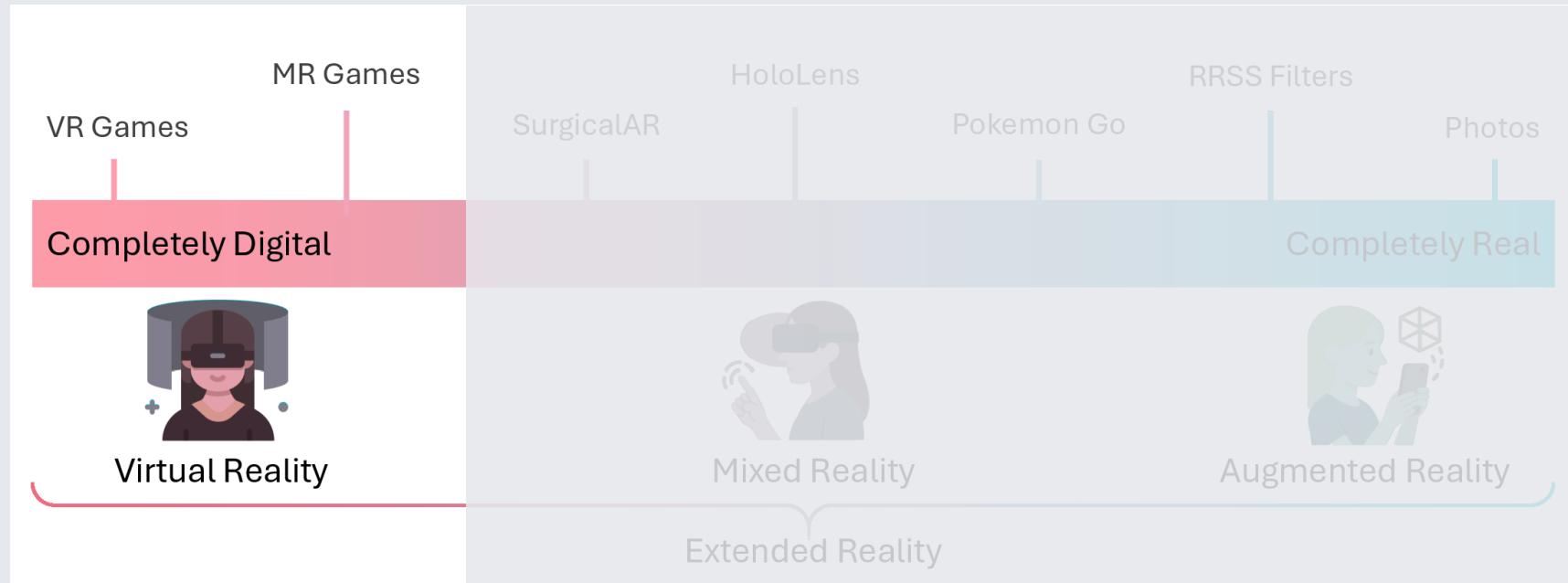
Espectro de Realidad Extendida

Realidad Extendida



Realidad Virtual:

Entorno completamente digital e inmersivo que reemplaza el mundo real.



VR, MR, AR?

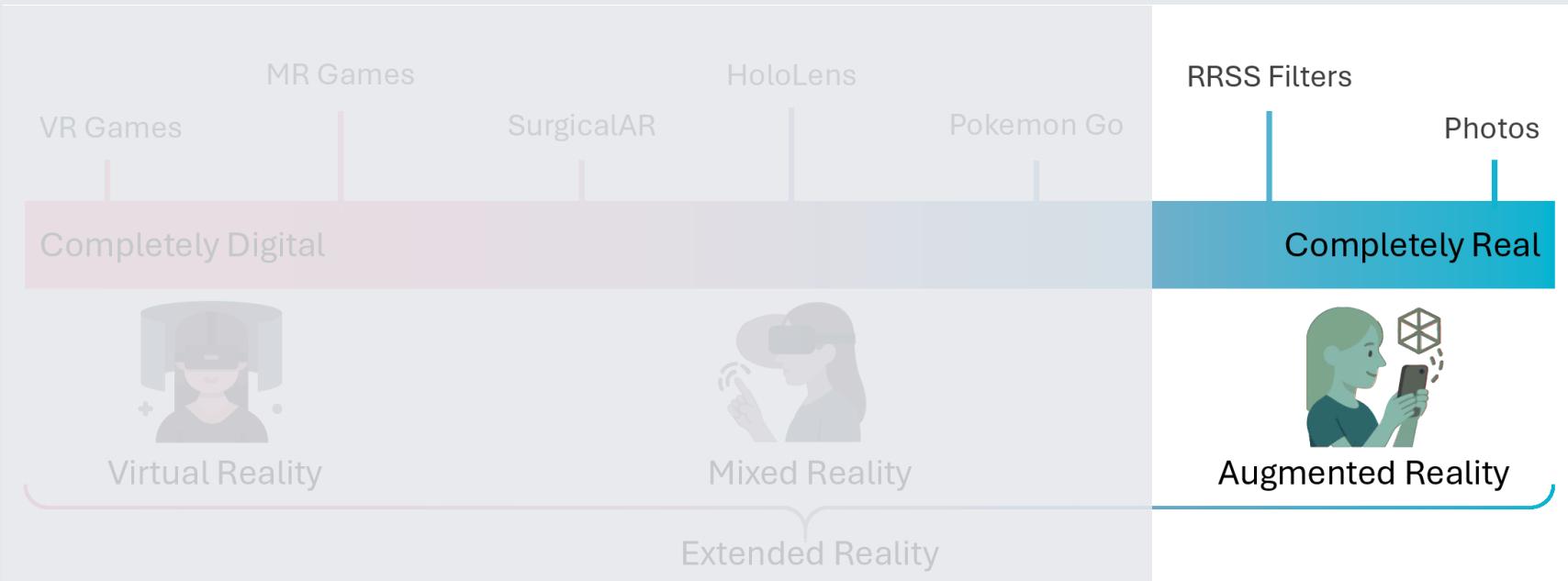
Espectro de Realidad Extendida

Realidad Extendida



Realidad Aumentada:

Superpone información digital sobre el entorno físico.



VR, MR, AR?

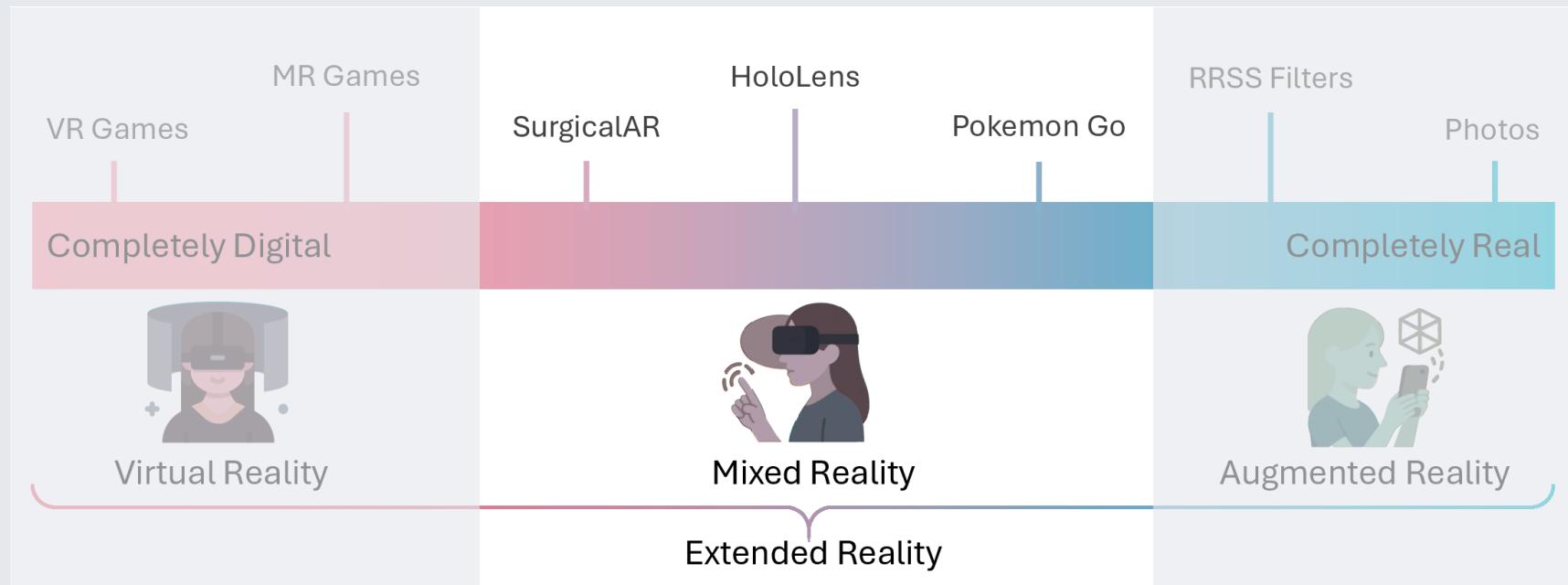
Espectro de Realidad Extendida

Realidad Extendida



Realidad Mixta:

Combina elementos reales
y virtuales que interactúan
en tiempo real.



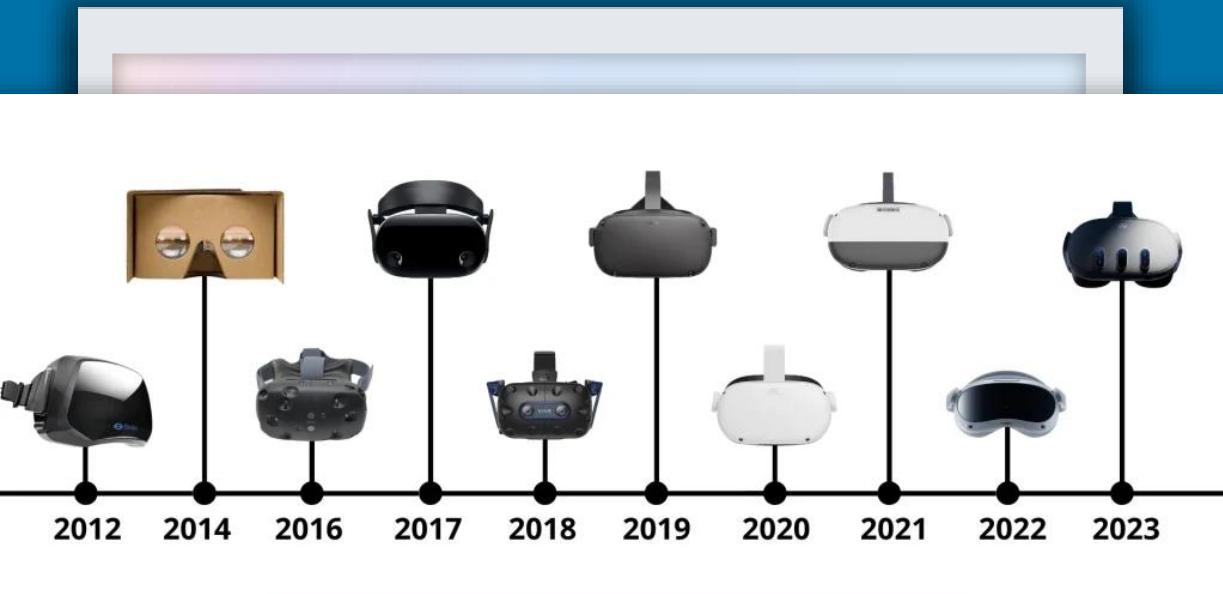
REALIDAD



VIRTUAL

Realidad Virtual

- Se crea un entorno digital completamente inmersivo.
- El usuario deja de percibir el mundo físico y se sumerge en un espacio 100% virtual.
- Requiere dispositivos como gafas o cascos de VR, mandos de control y sensores de movimiento.



Meta Quest 3(2023)

El avance tecnológico está
en marcha en la industria VR
Funciona de manera independiente y
combina VR y MR gracias a usar
cámaras RGB de alta resolución.

Interacción en entornos VR

La interacción es fundamental para crear experiencias inmersivas y naturales.

Principales métodos de interacción:



Controladores



Gestos



Seguimiento ocular



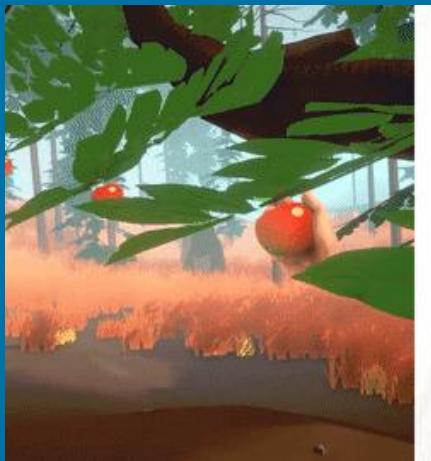
Voz



Interacción en entornos VR

🎮 Controladores

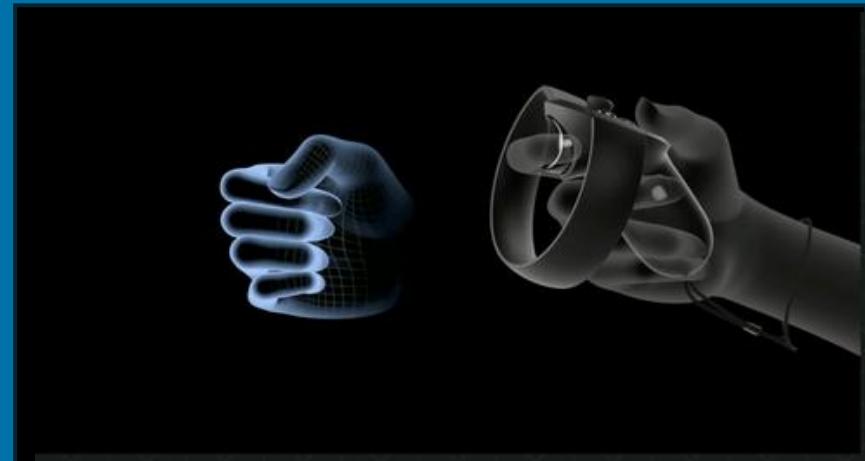
- Proporcionan control preciso de las interacciones en el entorno virtual.
- Utilizan sensores de movimiento para rastrear la posición y orientación, permitiendo acciones como agarrar, soltar o mover objetos virtuales.



Haptic PIVOT (2020)



Controladores

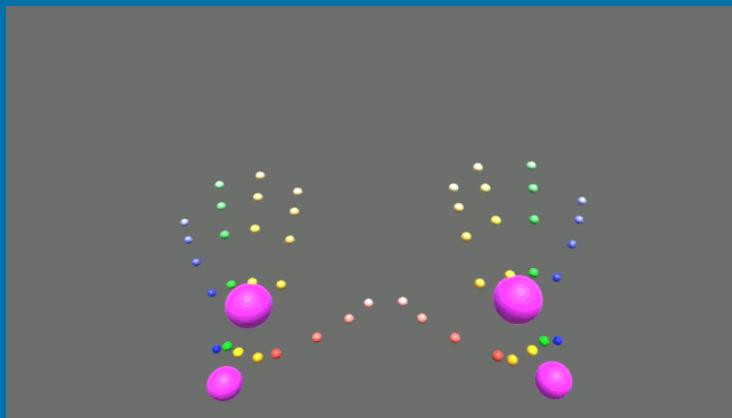


Controlador Meta Quest 2

Interacción en entornos VR

👉 Gestos

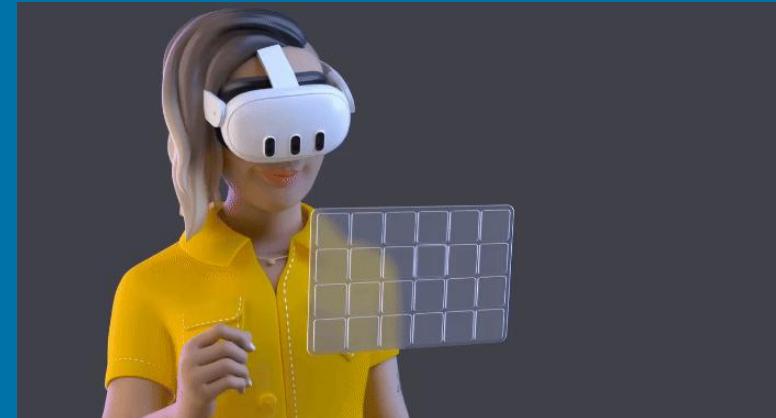
- Permiten a los usuarios interactuar con el entorno de manera natural, utilizando las manos o el cuerpo.
- Utilizan sensores de movimiento y cámaras para detectar los movimientos del cuerpo



Open XR Hand-tracking



Nintendo Power Glove

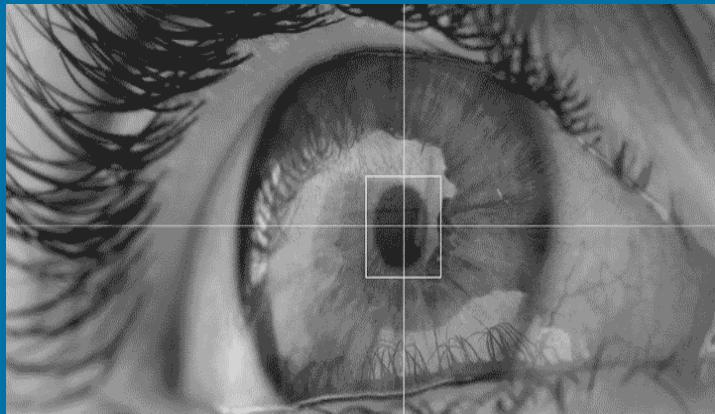


Interacción con objetos virtuales

Interacción en entornos VR

👁️ Seguimiento ocular

- Permite detectar la posición hacia la que mira el usuario, permitiendo el control de objetos o menús simplemente con la mirada.



Varjo VR Eye-tracking



Controladores



Gafas Pupil Labs Eye-tracking

Interacción en entornos VR

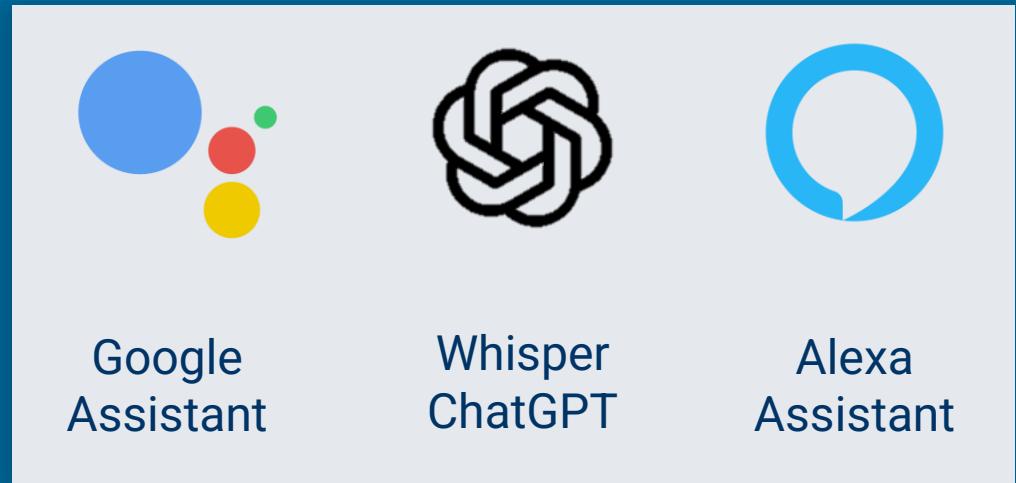


Voz

- El reconocimiento de voz permite que el usuario controle el entorno virtual con comandos verbales.



Asistente de voz en dispositivos VR



Modelos de procesamiento de voz

Interacción en entornos VR

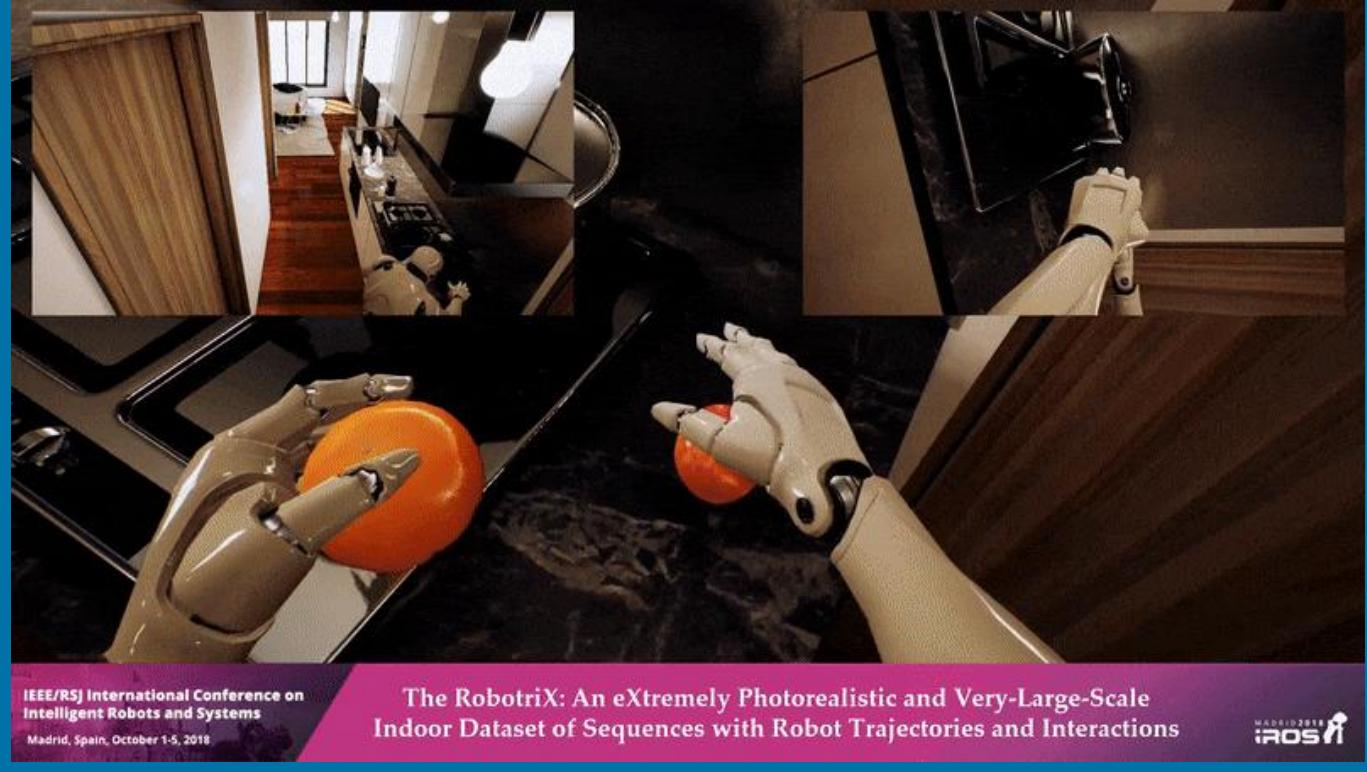
- Los espacios 3D permiten la interacción física del usuario con objetos y entornos en VR
- Navegación en 3D: El usuario puede moverse libremente por el entorno, ya sea caminando o mediante interfaces como los controladores o las esferas omnidireccionales.



Aplicaciones prácticas

UnrealRox: Generación de entornos realistas con interacción humano-objeto.

RobotriX: Dataset de agarres objetos usando entornos VR sintético.



Aplicaciones en salud

Rehabilitación física

- Entrenamiento motor y equilibrio
- Ejercicios personalizados con feedback visual y auditivo
- Mayor motivación y adherencia que la terapia tradicional
- Seguimiento del progreso en tiempo real mediante sensores y datos.



Using virtual reality for stroke rehabilitation

Aplicaciones en salud

Terapias psicológicas

- Exposición controlada a estímulos que generan miedo o ansiedad.
- Entrenamiento en afrontamiento emocional y técnicas de relajación.
- Simulación de situaciones sociales para terapia de ansiedad o autismo.
- Integración con IA para adaptar escenarios a la respuesta del paciente.



Richie's Plank Experience

REALIDAD



AUMENTADA

Realidad Aumentada

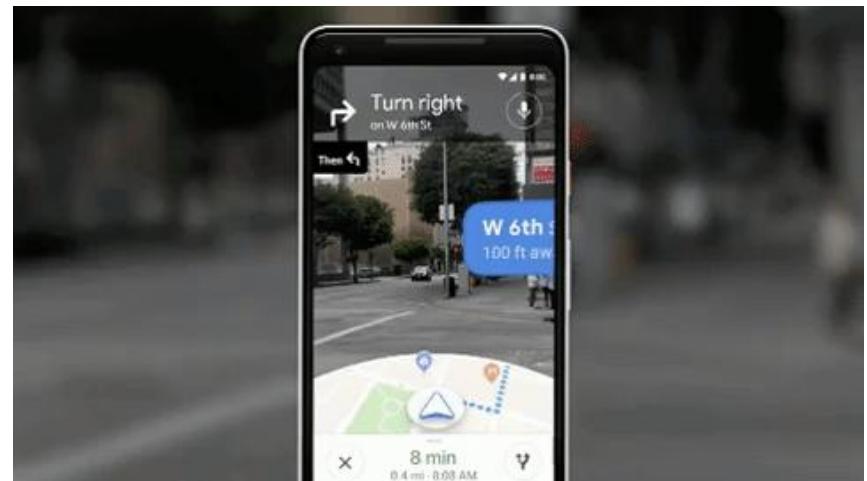
- Es la tecnología que superpone elementos digitales sobre el mundo físico en tiempo real, mejorando la percepción del entorno.
- A diferencia de la Realidad Virtual, que crea entornos completamente virtuales, AR mezcla el mundo real con capas digitales.



Microsoft Hololens



IKEA Geomagical Labs



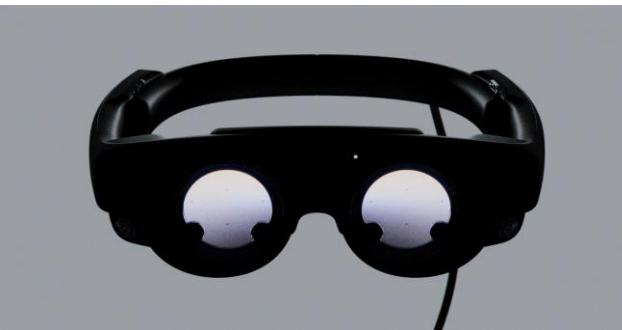
Google Maps Navigation

Dispositivos AR

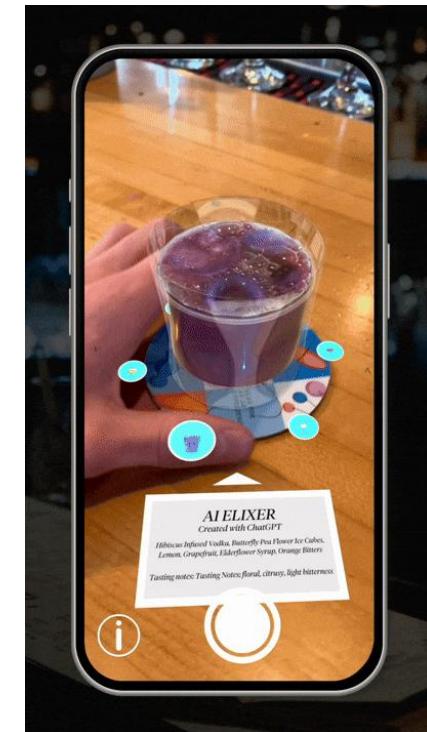
- Los dispositivos AR permiten al usuario interactuar con el entorno físico
- Gafas y cascos AR como HoloLens y Magic Leap ofrecen una experiencia inmersiva sin necesidad de pantallas externas.
- Móviles y tablets también sirven como dispositivos AR mediante cámaras y pantallas para mostrar contenido virtual superpuesto al mundo real.



HoloLens 2



Magic Leap



AR Cocktail Coaster

Interacción en entornos AR

- Interacciones más limitadas debido a la falta de sensores y controladores.
- Principales métodos de interacción:



Interacción física



Gestos



Seguimiento ocular



Voz

Interacción en entornos AR

👉 Interacción física

- A través de pantallas táctiles o dispositivos AR, el usuario puede interactuar con objetos virtuales proyectados sobre superficies físicas.



Ejemplo de Google ARCore

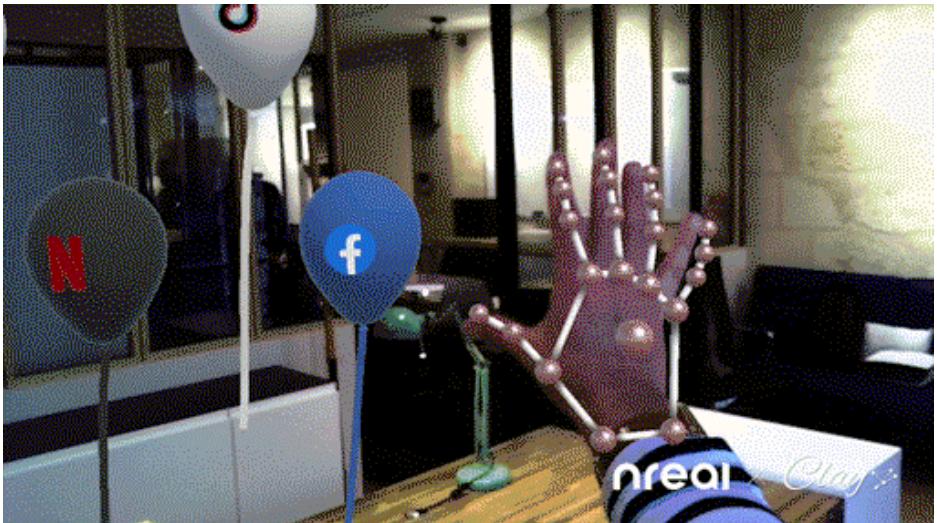


Product Demo

Interacción en entornos AR

👉 Gestos

- La detección de movimientos de las manos (con sensores de gestos o cámaras) permite al usuario manipular objetos virtuales.



Gafas AR Nreal Light

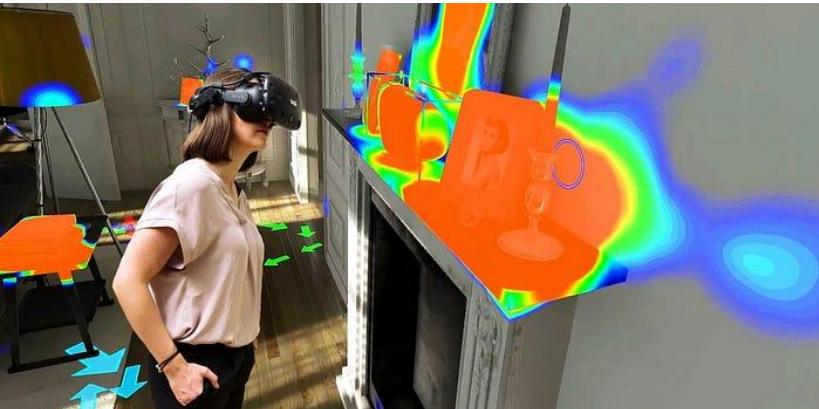


A-Frame Hand Tracking

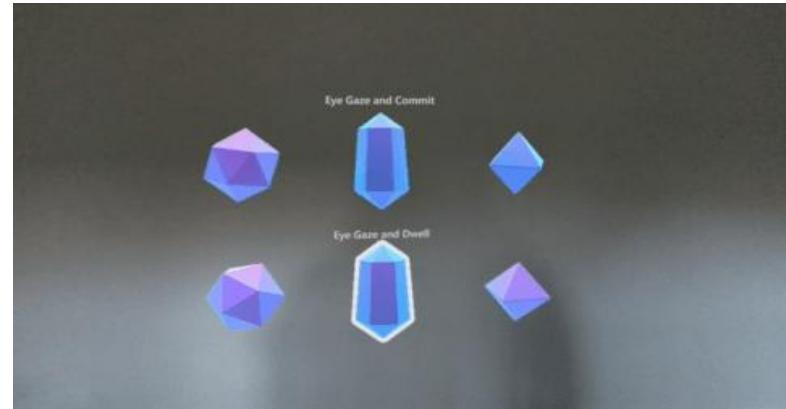
Interacción en entornos AR

👁️ Seguimiento ocular

- Algunos dispositivos AR, como HoloLens, permiten interactuar con el entorno solo con la mirada, utilizando sensores que detectan hacia dónde está mirando el usuario.



Mapa de calor de la mirada en AR



Hololens 2 Eye-tracking

Interacción en entornos AR



Voz

- En algunos sistemas AR, como Google Glass, los usuarios pueden dar comandos verbalmente para controlar las aplicaciones AR.



Meta Aira Gen 2



Google Glass

Ejemplos de Interfaces AR

Las interfaces AR se utilizan en diversos sectores, ofreciendo soluciones innovadoras en distintas áreas.

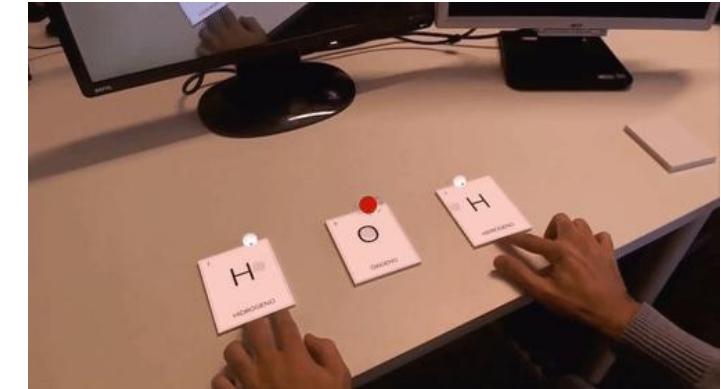
- Industria: Permite a los trabajadores visualizar instrucciones o diagramas sobre el objeto con el que están trabajando en tiempo real, facilitando la reparación de maquinaria, por ejemplo.
- Medicina: En la cirugía asistida por AR, los médicos pueden superponer imágenes de resonancia magnética (RM) o tomografías sobre el cuerpo del paciente.
- Educación: Ofrece una experiencia de aprendizaje más interactiva, como la visualización de modelos 3D de humanos o animales, lo que permite a los estudiantes estudiar de manera más visual e inmersiva.



Ford VR Gravity Sketch



AugMedicine LUMC

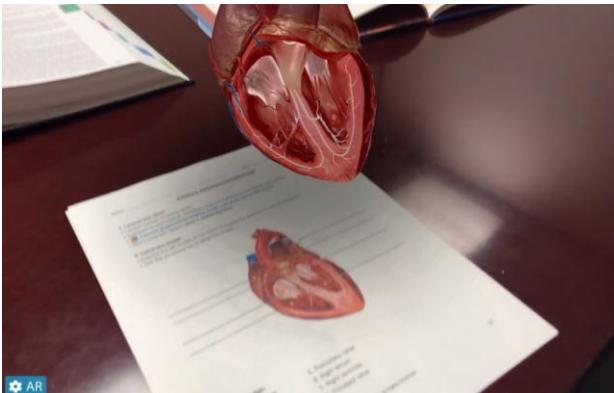


Aplicación AR Tabla Periódica

Aplicaciones AR en Salud

Se aplica en medicina para apoyar la cirugía, la formación y la atención al paciente, combinando imágenes clínicas con el mundo real.

- Cirugía guiada por AR: Se puede superponer imágenes de resonancia o TAC sobre el cuerpo del paciente durante una operación.
- Formación anatómica interactiva: Los estudiantes visualizan modelos 3D de órganos o sistemas corporales.
- Apoyo a pacientes y rehabilitación visual: Guiar ejercicios de recuperación motora o mejorar la orientación de personas con baja visión.



Visible Body AR



AugMedicine LUMC



Mirror AR

REALIDAD



MIXTA

Realidad Mixta

- Combina elementos de AR y VR para crear un entorno en el que objetos físicos y virtuales interactúan de manera dinámica y en tiempo real.
- A diferencia de AR, que solo superpone información digital al mundo físico, y de VR, que crea un entorno completamente virtual, MR permite que los elementos digitales y físicos coexistan e interactúen de forma fluida.



Dispositivos de Realidad Mixta

- Mismas posibilidades de interacción que RV o AR
- Mayor requerimiento de sensores y cómputo



Nreal Light (2020)
Diseñado para superponer
elementos virtuales en el mundo
real.



Vuzix Blade 2 (2021)
Principalmente para el
mercado empresarial, estilo
gafas de sol.



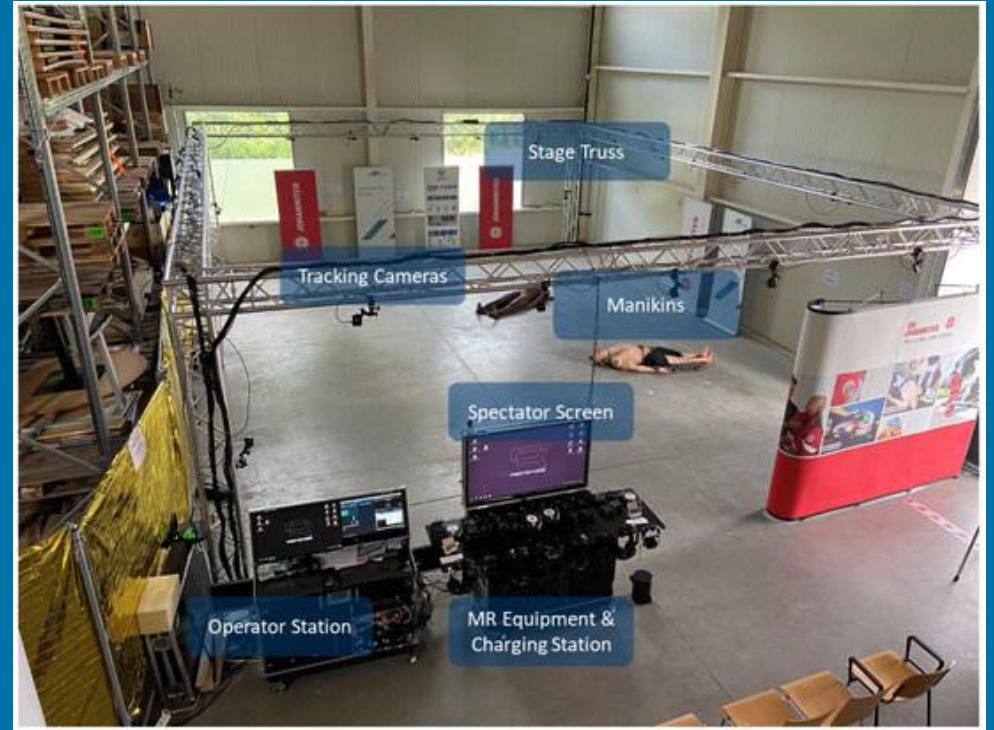
Microsoft HoloLens 2 (2019)
Dispositivo más conocido.
Usado en industria, medicina y
educación.

Aplicaciones MR

- Planificación quirúrgica 3D interactiva.
- Colaboración remota médico-especialista.
- Simulación de emergencias hospitalarias.

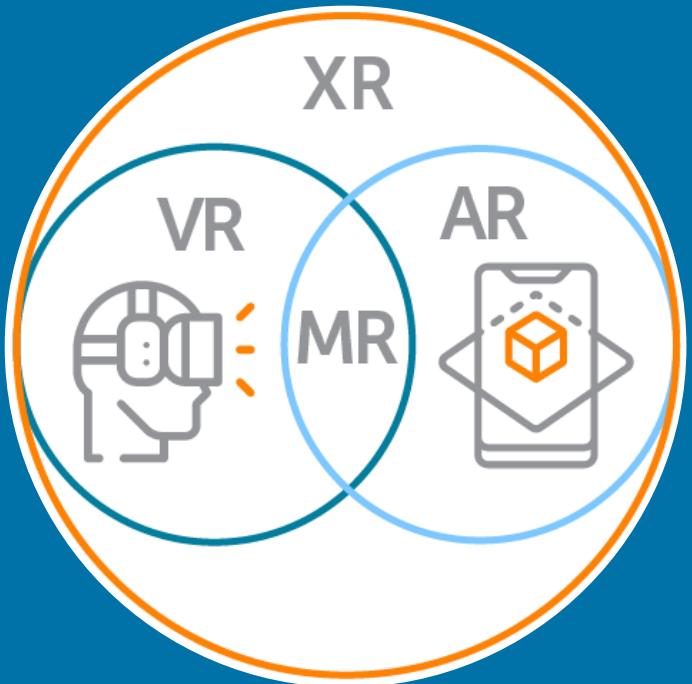


MR-5G



NextGen Training for Medical First Responders

Comparativa VR vs AR vs MR



Características	VR	AR	MR
Tipo de experiencia	Entornos completamente digitales e inmersivos	Superposición de objetos digitales sobre el mundo real	Mezcla de entornos físicos y digitales, con interacción en tiempo real
Interacción con el mundo físico	No interactúa con el mundo físico	Interacción limitada con el mundo físico (solo visual)	Interacción fluida entre lo físico y lo virtual
Uso de dispositivos	Cascos VR, mandos, sensores de movimiento	Móviles, tablets, gafas AR como HoloLens	Cascos MR como HoloLens, Magic Leap
Inmersión del usuario	Alta inmersión (aislado del mundo real)	Baja inmersión (mantiene la percepción del entorno físico)	Inmersión parcial (combinando lo físico y lo virtual)
Aplicaciones comunes	Juegos, simuladores, entrenamientos, entretenimiento	Juegos, educación, marketing, diseño	Aplicaciones industriales, medicina, educación, diseño
Ejemplos de dispositivos	Oculus Quest, HTC Vive, PlayStation VR	HoloLens, Magic Leap, Pokémon Go (móvil)	HoloLens 2, Magic Leap 2, Nreal Light

ENTORNOS DE DESARROLLO PARA EXPERIENCIAS INMERSIVAS



Tecnologías

Dependen del hardware especializado

HARDWARE

Cámaras de profundidad

Acelerómetros

Giroscopios

Pantallas UHR

SOFTWARE

VR



unity



UNREAL
ENGINE



VR



VRTK
virtual reality toolkit

AR



ARKit



vuforia™



ARCore

MR



M R T K
MIXED REALITY TOOLKIT



magic
leap



UNREAL
ENGINE

Integración con IA

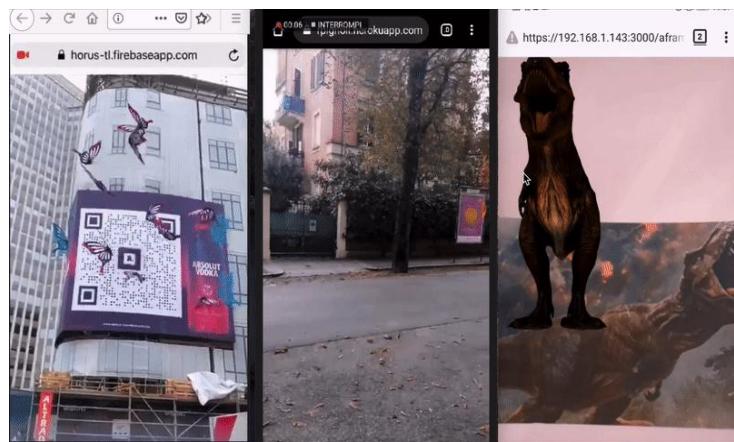
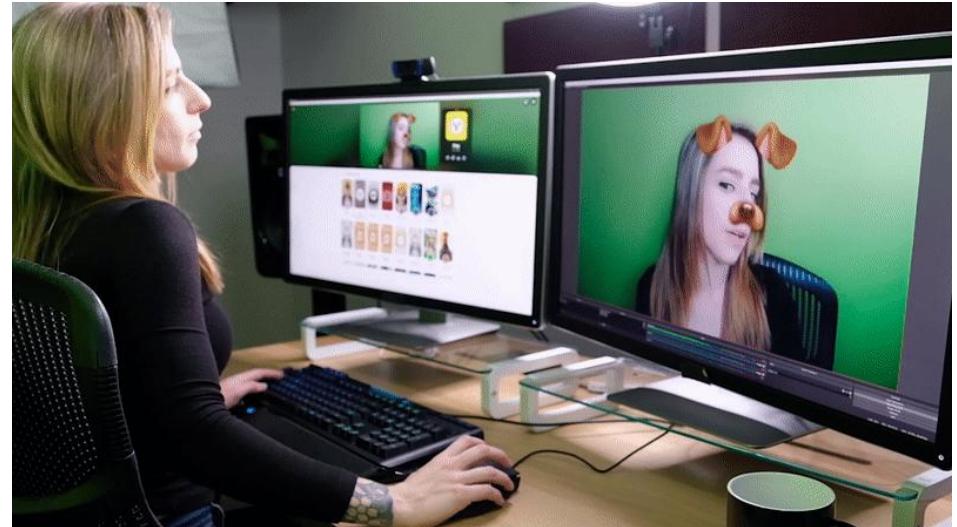
La Inteligencia Artificial puede complementar las experiencias inmersivas en VR, AR y MR.

Existen varias modalidades y modelos de IA que se pueden integrar para enriquecer estas aplicaciones.



- Procesamiento de Lenguaje Natural
Asistentes virtuales y comandos de voz
- Visión por Computador
Permite reconocer el entorno físico y objetos
- Reconocimiento de Gestos
Para interactuar con el entorno y contenido virtual
- IA Predictiva
Utiliza datos previos del usuario para predecir sus acciones futuras y ofrecer contenido o experiencias personalizadas.
- IA Generativa
Creación de contenido virtual realista y dinámico

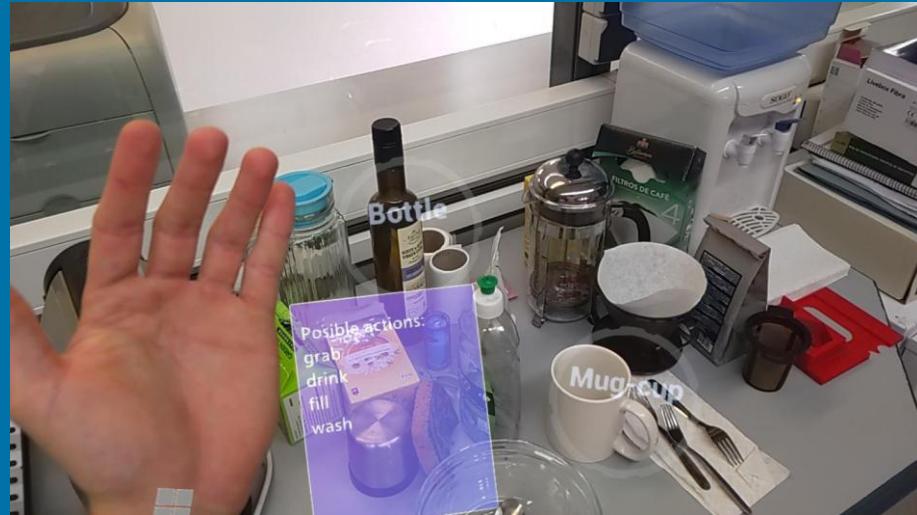
Ejemplos de aplicaciones



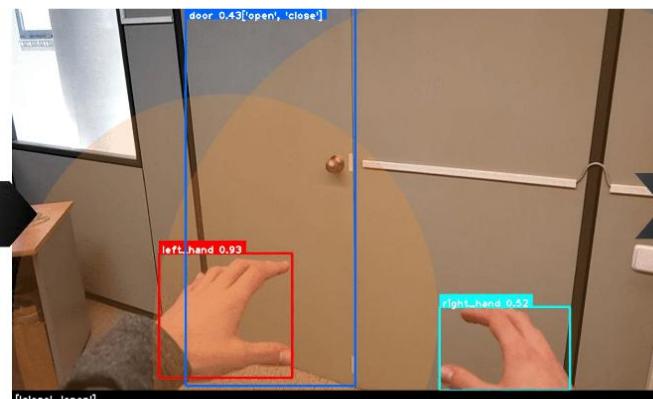
Ejemplos de Aplicaciones

HoloYolo

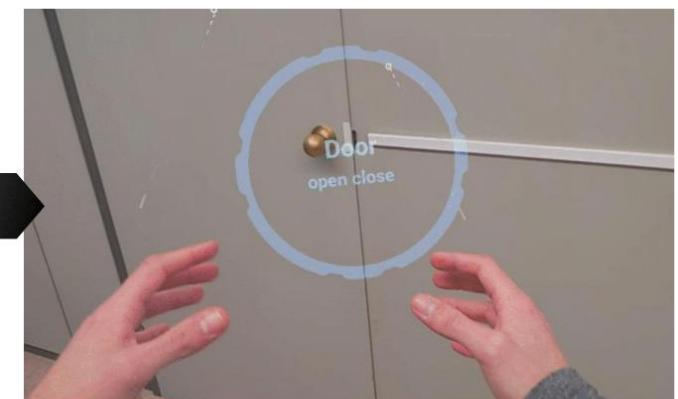
Detección de objetos y acciones con modelos de AI (YOLO), mediante dispositivos de Realidad Mixta (Hololens)



Egocentric video capture



Object and action prediction



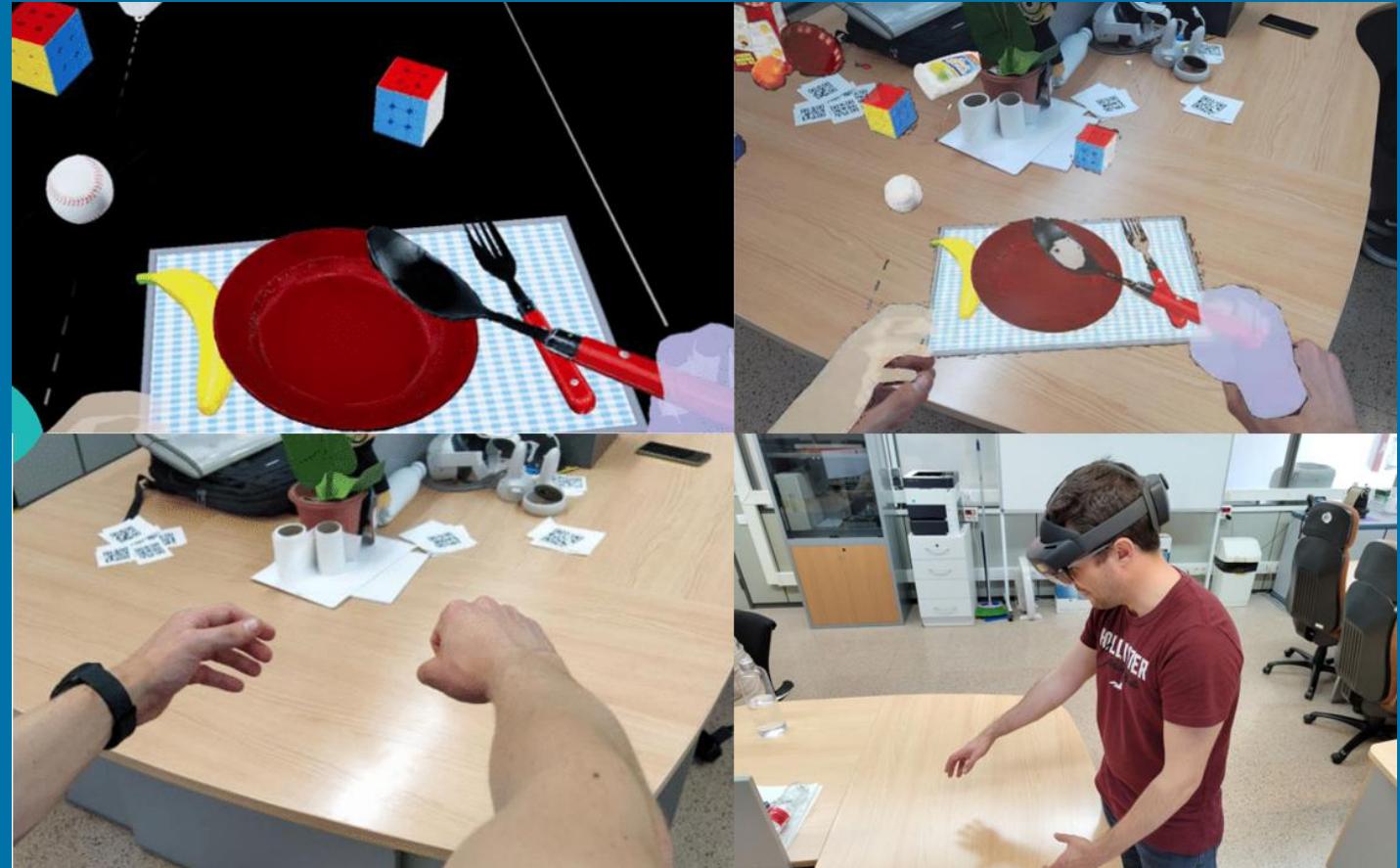
Visualization and interaction with objects

Ejemplos de Aplicaciones

HoloDemtect

Detección y estimulación cognitiva
con entornos AR

Virtualización de tareas del día a día



Ejemplos de Aplicaciones

HoloDemtect

Detección y estimulación cognitiva
con entornos AR

Virtualización de tareas del día a día

- Hacer la compra
- Emparejar objetos
- Poner la mesa



Ejemplos de Aplicaciones

Rehabilitación en grupo para pacientes con esclerosis múltiple mediante Realidad Virtual y 5G



Ejemplos de Aplicaciones

AccuVein: Dispositivo de AR que proyecta en tiempo real el mapa de las venas sobre la piel del paciente.



Ejemplos de Aplicaciones

Medi-AR: Aplicación de AR con inteligencia artificial orientada a la educación médica. Transforma imágenes médicas estáticas (ej. radiografías, TAC) en visuales 3D interactivos sobre el paciente o en el entorno.

github.com/pradeepmisl/MEDI_AR.git



2. Nasal corticosteroids: These sprays, like fluticasone or mometasone, can reduce inflammation in your nasal passages.

3. Decongestants: Oral medications like pseudoephedrine or nasal sprays like oxymetazoline can help relieve congestion.

Natural Remedies:

1. Saline Nasal Irrigation: Using a saline solution to rinse your nasal passages can help remove allergens.
2. Butterbur: Some studies suggest butterbur, a natural extract, might reduce allergy symptoms.
3. Quercetin: This natural flavonoid found in foods like apples and onions might help stabilize histamine release.
- 4.

Aplicaciones gratuitas disponibles



INTERACTIVE
Anatomy



INSIGHT
LUNG



INSIGHT
HEART



Visutate

Bibliografía

- Viglialoro, R. M., Condino, S., Turini, G., Carbone, M., Ferrari, V., & Gesi, M. (2021). Augmented reality, mixed reality, and hybrid approach in healthcare simulation: a systematic review. *Applied Sciences*, 11(5), 2338.
- Fu, Y., Hu, Y., & Sundstedt, V. (2022). A systematic literature review of virtual, augmented, and mixed reality game applications in healthcare. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(2), 1-27.
- Sevcenko, K., et. al. (2022). The effects of virtual reality training in stroke and Parkinson's disease rehabilitation: a systematic review and a perspective on usability. *Eur Rev Aging Phys Act* 19, 4
- Mulero-Pérez, D., Benavent-Lledo, et. al. (2023). HoloDemtect: A Mixed Reality Framework for Cognitive Stimulation Through Interaction with Objects. *18th International Conference on Soft Computing Models in Industrial and Environmental Applications*
- Navas-Medrano, S., Soler-Dominguez, J. L., & Pons, P. (2024). Mixed Reality for a collective and adaptive mental health metaverse. *Frontiers in psychiatry*, 14, 1272783.
- Norris, T.A., et al. (2024). Shaping corticospinal pathways in virtual reality: effects of task complexity and sensory feedback during mirror therapy in neurologically intact individuals. *J NeuroEngineering Rehabil*,
- Virtual Reality Exposure Therapy. (2025). XRHealth. En <https://www.xr.health/us/products/virtual-reality-exposure-therapy/>,

Bibliografía

- Gerup, J., Soerensen, C. B., & Dieckmann, P. (2020). Augmented reality and mixed reality for healthcare education beyond surgery: An integrative review. *International Journal of Medical Education*, 11, 1–18.
- Thomas, R. B., et al. (2022). Use of mixed reality for surgery planning: Assessment and workflow integration. *Journal of Biomedical Informatics*, 127, 104017.
- García, Francisco, Moraleda, Rubén, Schez-Sobrino, Santiago, Monekosso, Dorothy, Vallejo, David, Glez-Morcillo, Carlos. (2023). Health-5G: A Mixed Reality-Based System for Remote Medical Assistance in Emergency Situations. *IEEE Access*. PP. 1-1.
- Zechner, O., García Guirao, D., Schrom-Feiertag, H., Regal, G., Uhl, J. C., Gyllencreutz, L., Sjöberg, D., & Tscheligi, M. (2023). NextGen Training for Medical First Responders: Advancing Mass-Casualty Incident Preparedness through Mixed Reality Technology. *Multimodal Technologies and Interaction*, 7(12), 113.
- Magalhães, R., Oliveira, A., Terroso, D. et al. (2024). Mixed Reality in the Operating Room: A Systematic Review. *J Med Syst* 48, 76

Interacción Persona-Máquina

U8: Interfaces basadas en bioseñales

Manuel Benavent Lledó
Jose García Rodríguez

Universidad de Alicante

2025-2026

Grado en Ingeniería en Inteligencia Artificial



Universitat d'Alacant
Universidad de Alicante

Grado en Ingeniería en IA
Interacción Persona-Máquina

Unidad 8:
Interfaces basadas en bioseñales

Contenidos

1. Introducción
2. Tipos de bioseñales y sensores
3. Adquisición y procesamiento de bioseñales
4. Aplicaciones en accesibilidad y neurotecnología
5. Bibliografía

1. Introducción

Algunas Definiciones

- **¿Qué son las bioseñales?**

Señales eléctricas, mecánicas o bioquímicas generadas por sistemas fisiológicos del cuerpo humano.

- **¿Para qué sirven?**

Proporcionan información sobre el funcionamiento de órganos y tejidos, permitiendo diagnósticos médicos, monitorización de estados fisiológicos y desarrollo de tecnologías interactivas, como sistemas HCI.

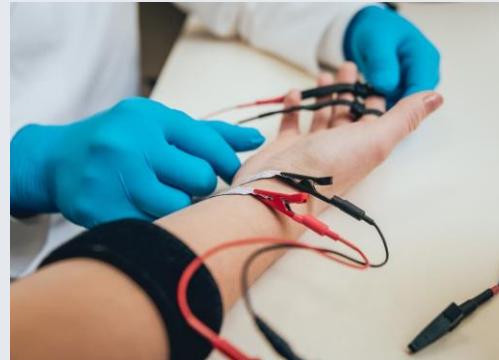
1. Introducción Ejemplos

Características: Actividad que registran, Amplitud y Frecuencia.

En función de las características seleccionaremos métodos de **adquisición y procesamiento**.



Electrocardiograma o ECG: Mide la actividad eléctrica del corazón



Electromiograma o EMG: Mide la actividad eléctrica de los músculos



Electroencefalograma o EEG: Mide la actividad eléctrica del cerebro



1. Introducción

Importancia de las señales (en HCI)

En general:

Permiten múltiples aplicaciones en inteligencia artificial, como la monitorización de actividad, estrés o sueño mediante pulseras inteligentes.

En sistemas HCI:

- Controlar dispositivos sin movimiento físico, mediante señales cerebrales.
- Personalizar experiencias de usuario según estados emocionales o nivel de atención
- Inclusión tecnológica para personas con discapacidad motora o de comunicación.

1. Introducción

Importancia de las señales (en HCI)

En general:

Permiten monitoreo y
monitorización.

En sistemas:

- Controlar dispositivos
- Personalizar la atención de atención
- Inclusión en la comunicación



o la
s inteligentes.

les cerebrales.

nales o nivel

ora o de

<https://youtu.be/F30cbHlaQF0>

2. Tipos de bioseñales y sensores

Según el tipo de adquisición (en medicina)

- **Invasivas:**

Implican el uso de sensores implantados en el cuerpo. Aunque mejora la precisión, hay un riesgo más elevado y la implantación suele requerir cirugía.

Ejemplo: Electrodos implantados en el cerebro.

- **No invasivas:**

Suele ser suficiente con sensores de superficie (ECG), electrodos adhesivos (ECG) o sensores en la piel.

2. Tipos de bioseñales y sensores

Según el tipo de adquisición (en aplicaciones)

OJO!

- Aunque las bioseñales "no invasivas" no requieren intervención médica directa, **pueden ser intrusivas o incómodas** para el usuario final debido a la necesidad de electrodos, geles conductores o bandas ajustadas.
- Algunos dispositivos como pulseras inteligentes permiten medir señales de forma no invasiva, y se pueden usar técnicas como visión por computador para analizar rasgos faciales.
- Sin embargo, la **precisión** de estos dispositivos es **más baja**, y la elección del método dependerá de la aplicación específica.

2. Tipos de bioseñales y sensores Según su origen fisiológico – EEG

Electroencefalograma (EEG)

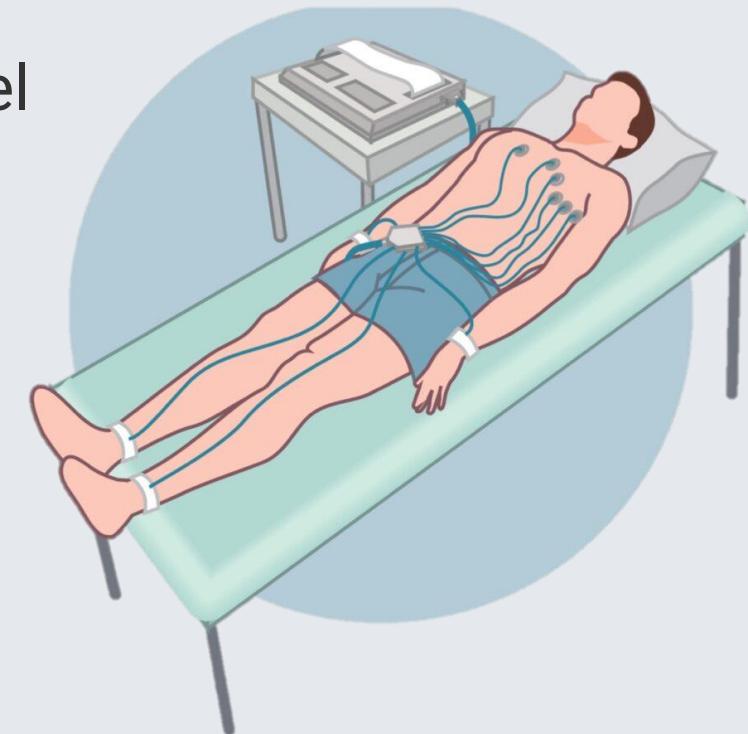
- Registra la actividad eléctrica del cerebro, reflejando la actividad de grupos neuronales.
- **Sensores y métricas:** Electrodos en el cuero cabelludo miden frecuencias cerebrales (delta, theta, alfa, beta, gamma).
- **Aplicación:** estudio de procesos cognitivos, interfaces cerebro-computador (BCI) y diagnóstico de trastornos neurológicos.



2. Tipos de bioseñales y sensores Según su origen fisiológico – ECG

Electrocardiograma

- Mide la actividad eléctrica del corazón, reflejando el ritmo y la frecuencia cardíaca.
- **Sensores:** Electrodos en el pecho miden la frecuencia cardíaca (bpm), intervalos R-R (contracciones ventriculares) y la variabilidad de la frecuencia cardíaca (HRV).
- **Aplicaciones:** Diagnóstico, control cardiovascular, monitorización de estrés y control de interfaces de usuario.



2. Tipos de bioseñales y sensores Según su origen fisiológico – EDA

Actividad electrodérmica

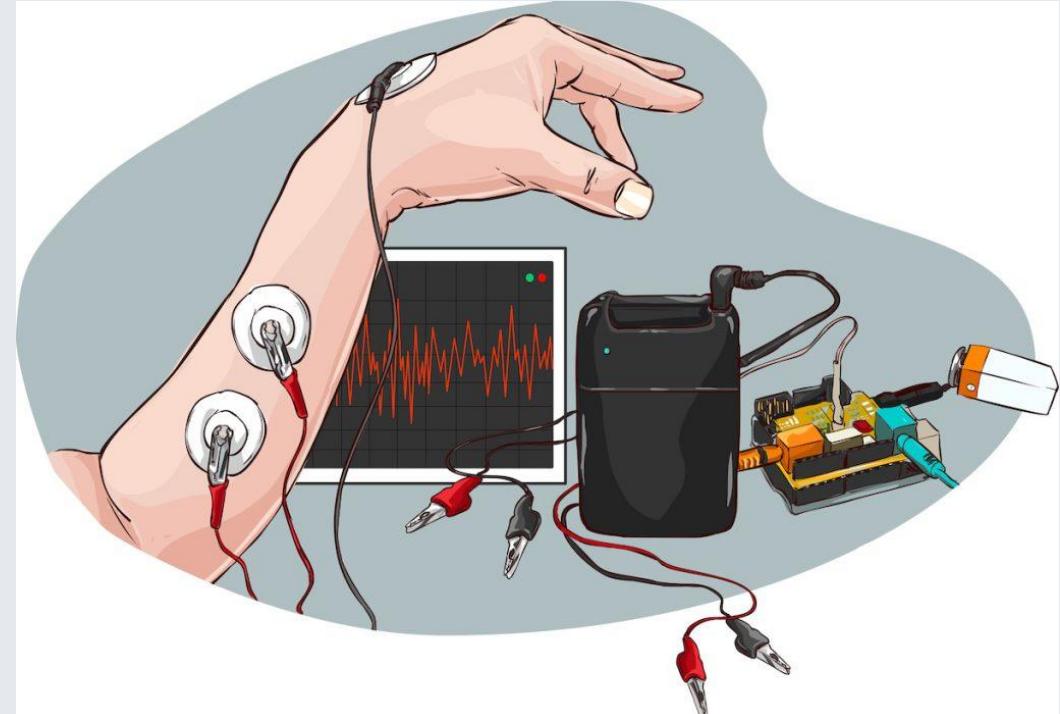
- Mide la conductividad de la piel, que varía con la sudoración y refleja respuestas emocionales o de estrés.
- **Sensores:** de conductividad en los dedos o la palma de la mano. Miden la conductancia (S) y la respuesta galvánica de la piel (GSR).
- **Aplicaciones:** Monitorización de estrés, interfaces adaptativas (ajuste de experiencia basado en emociones), investigación de emociones.



2. Tipos de bioseñales y sensores Según su origen fisiológico – EMG

Electromiograma

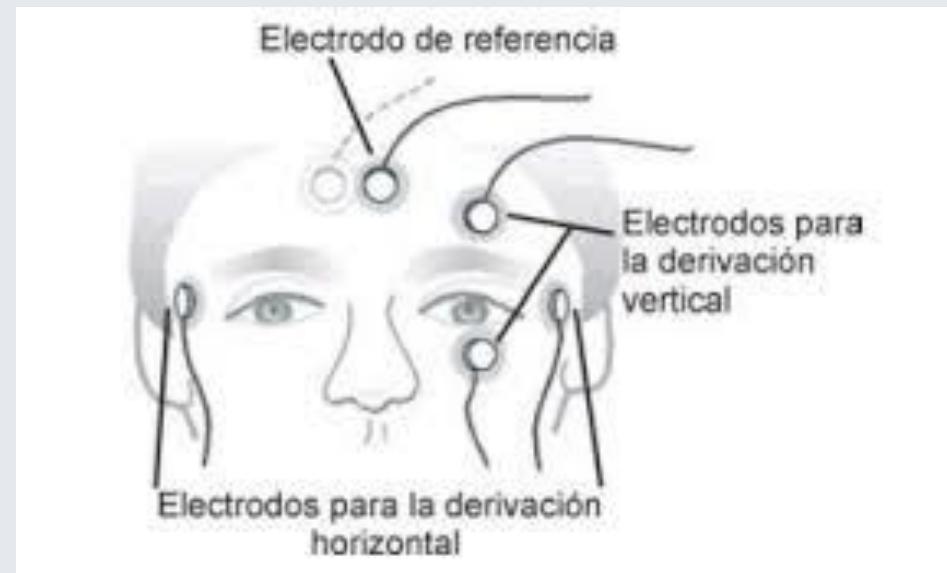
- Mide la actividad eléctrica generada por los músculos durante la contracción.
- **Sensores:** Electrodos de superficie o agujas insertadas en los músculos. Se mide la amplitud del potencial muscular y frecuencia de disparo de las fibras musculares.
- **Aplicaciones:** Control de prótesis, rehabilitación, estudios de fatiga muscular o interfaces de control basadas en movimiento.



2. Tipos de bioseñales y sensores Según su origen fisiológico – EOG

Electrooculograma

- Mide los movimientos oculares detectando la diferencia de potencial entre la retina y la córnea.
- **Sensores:** Electrodos alrededor de los ojos (cerca de los párpados) miden los movimientos horizontales, verticales y parpádeos.
- **Aplicaciones:** Seguimiento ocular, interfaces controladas por la mirada, estudios de fatiga y sueño.



2. Tipos de bioseñales y sensores Según su origen fisiológico – Otros

Otras señales: Útiles para monitorización pero limitadas para HCI

- Temperatura corporal
- Presión arterial
- Respiración
- Volumen sanguíneo: a partir de este se puede determinar la frecuencia cardíaca y saturación de oxígeno en pulseras inteligentes

3. Adquisición y Procesamiento Equipos y Métodos de Registro

Sensores fisiológicos

- Colocados sobre la piel o en contacto cercano al cuerpo
- Hechos de materiales conductores (metales, geles, adhesivos)
- Detectan las señales biológicas previas: EEG, ECG, EDA, EMG, etc.

Dispositivos de adquisición

- Amplifican, filtran y digitalizan las señales
- Pueden ser profesionales (normalmente en entornos clínicos) o wearables/kits portátiles

3. Adquisición y Procesamiento Equipos y Métodos de Registro

Dispositivos de adquisición – Wearables

Ventajas:

- Bajo coste y fácil transporte
- Medición en tiempo real
- Uso cómodo con usuarios finales

Ejemplos

- OpenBCI: plataforma open-source para EEG/EMG/ECG
- Empatica E4: pulsera multisensor
- Noraxon: sensores EEG



3. Adquisición y Procesamiento Equipos y Métodos de Registro

Dispositivos de adquisición – Wearables

Ventajas:

- Bajo coste y fácil transporte
- Medición en tiempo real
- Uso cómodo con usuarios finales



Ejemplos

- OpenBCI: plataforma open-source para EEG/EMG/ECG
- Empatica E4: pulsera multisensor
- Noraxon: sensores EEG

3. Adquisición y Procesamiento Equipos y Métodos de Registro

Condiciones de registro

- **Preparación del usuario:** limpieza y uso de geles de contacto para mejorar la conductividad
- **Calibración y colocación de sensores:** Usar sistemas estandarizados y asegurar una correcta colocación calibración de los sensores
- **Ambiente y Movimiento:** minimizar ruido eléctrico y evitar posibles factores externos (distracciones, temperatura o luces)



3. Adquisición y Procesamiento Comunicación y Transmisión de Datos

Tipo de conexión

- Cable: estable pero limita movilidad
- Inalámbrica: Bluetooth o WiFi favorecen la movilidad pero aumento de ruido o interferencias

Protocolos y herramientas útiles

- BLE (Bluetooth Low Energy): usado en la mayoría de wearables por su bajo consumo
- LSL (Lab Streaming Layer): para sincronizar flujos de datos fisiológicos

3. Adquisición y Procesamiento Comunicación y Transmisión de Datos

Tipo de conexión

- Cable: estable pero limita movilidad

- Inalámbrica:
ruido o intere

OJO!

La falta de sincronización puede llevar a malas interpretaciones, afectar a modelos de IA o generar fallos en las interfaces

Protocolos y

- BLE (Bluetooth): bajo consumo

- LSL (Lab Streaming Layer): para sincronizar flujos de datos fisiológicos

3. Adquisición y Procesamiento

Preprocesamiento y Filtrado

Ruido Generado por:

- Ambiente: interferencias eléctricas de cables y enchufes
- Movimientos del usuario
- Estímulos nerviosos como contracciones musculares

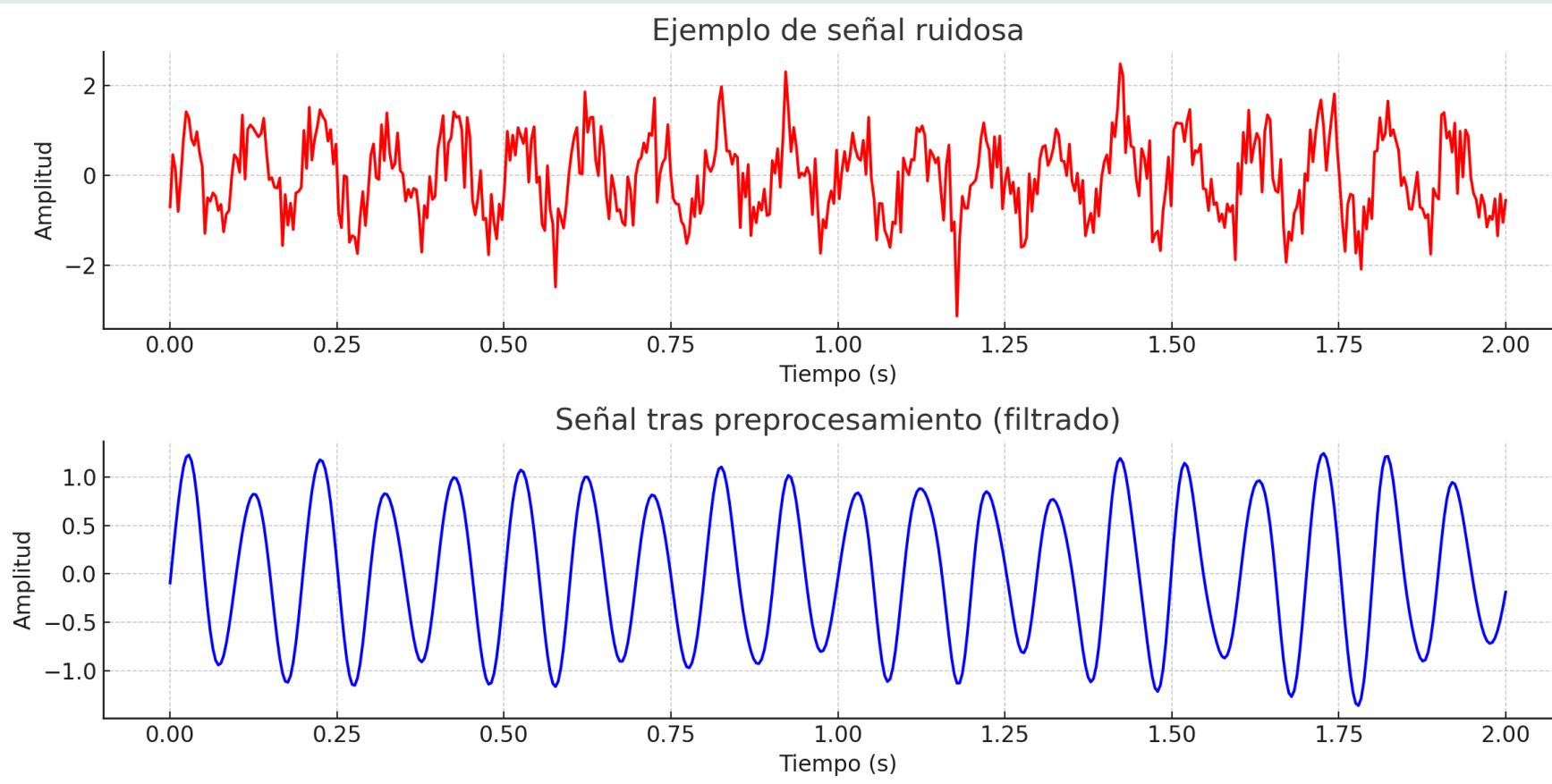
Es imprescindible preprocesar la señal para limpiar el ruido y descartar partes que no son útiles, como frecuencias con menos información

3. Adquisición y Procesamiento

Preprocesamiento y Filtrado

- Ruido (noise)
- Ambiente
 - Movimiento
 - Estímulo

Es importante separar las partes



3. Adquisición y Procesamiento

Preprocesamiento y Filtrado

Técnicas

Filtrado:

- Eliminamos las frecuencias menos relevantes
- Reducimos la dimensionalidad, mejorando la eficiencia

Corrección de artefactos

- Algoritmos para identificar y eliminar partes afectadas por movimiento, respiración, etc.

Normalización

- Ajustamos la señal a una escala común para poder comparar entre equipos, personas o sesiones.

3. Adquisición y Procesamiento

Procesamiento – Extracción de Características

Técnicas tradicionales

- Dependen del tipo de señal
- Suelen requerir intervención manual
- Ejemplos: máximo, mínimo, mediana, curtosis pero también específicos analizando segmentos de la señal para detectar movimientos concretos.

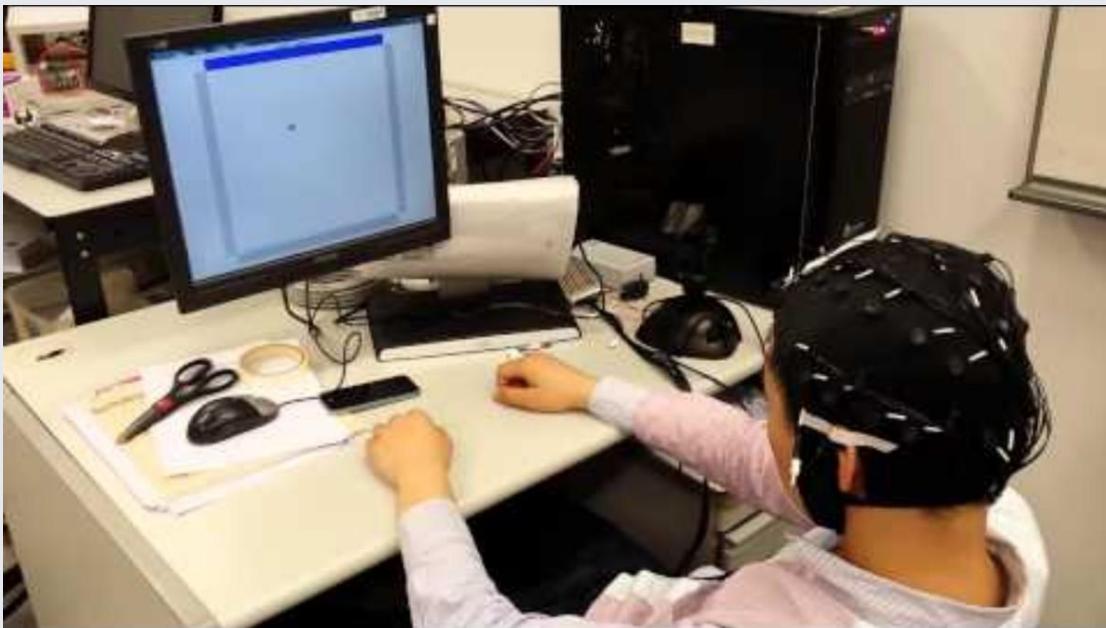
Técnicas mediante IA

- Más complejas, requieren más datos pero requieren de menos análisis manual
- Métodos de Machine Learning (SVM, Random Forests, kNN), CNNs (espectogramas 2D), RNNs/LSTMs/Transformers para series temporales
- Gran volumen de datos: necesario hacer *downsampling*, es decir, reducir la señal analizando ventanas más pequeñas

4. Aplicaciones

BCI para personas con discapacidad motora

- Útiles para personas con movilidad reducida o sin movilidad debido a parálisis cerebral, lesiones medulares o ELA
- **Aplicaciones:** escribir o controlar el cursor con señales EOG o EEG



<https://youtu.be/NmZXruqbTVs>

4. Aplicaciones

BCI para personas con discapacidad motora

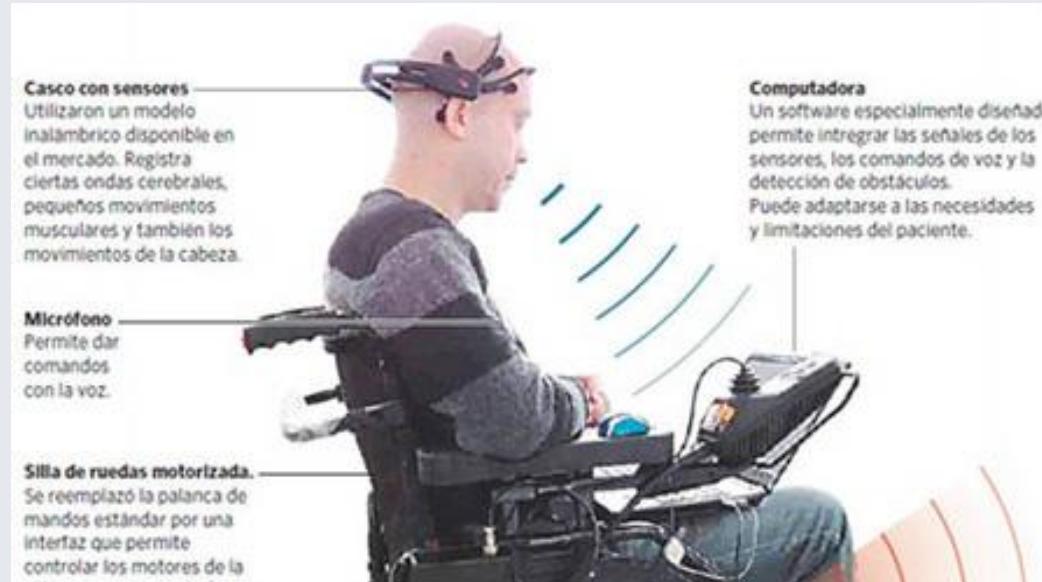
- Útiles para personas con movilidad reducida o sin movilidad debido a parálisis cerebral, lesiones medulares o ELA
- **Aplicaciones:** detección de intenciones basadas en respuestas cerebrales a estímulos visuales (P300 speller)



4. Aplicaciones

BCI para personas con discapacidad motora

- Útiles para personas con movilidad reducida o sin movilidad debido a parálisis cerebral, lesiones medulares o ELA
- **Aplicaciones:** control de sillas de ruedas o prótesis robóticas mediante señales EEG



Bibliografía

- Libenson, M. H. (2016). Practical approach to electroencephalography (2nd ed.). Saunders.
- Niedermeyer, E., & da Silva, F. L. (2004). Electroencephalography: Basic principles, clinical applications, and related fields (5th ed.). Lippincott Williams & Wilkins.
- He, B. (2013). Neural engineering (2nd ed.). Springer.
- Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2013). Principios de neurociencia (5^a ed.). McGraw-Hill Interamericana.
- Bronzino, J. D., & Peterson, D. R. (Eds.). (2015). Fundamentals of Biosignals. In Biomedical Signals and Sensors I (pp. 1-25). Springer.