

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

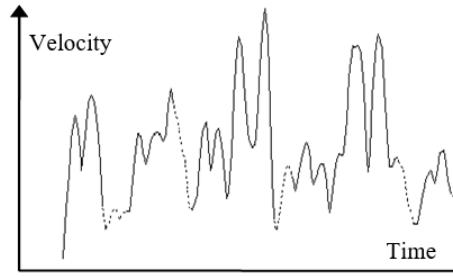
Mi evolución histórica en visión por computador para la interacción persona-entorno

Proyecto fin de carrera

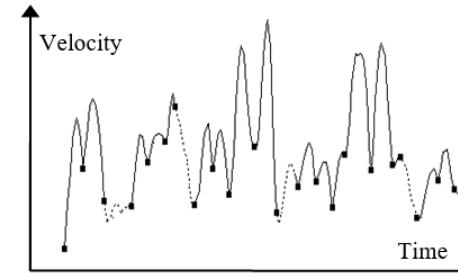
Reconocimiento de escritura en cursiva empleando redes neuronales auto-organizativas

regalo

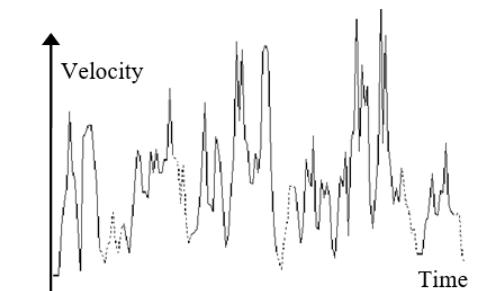
a. Original writing.



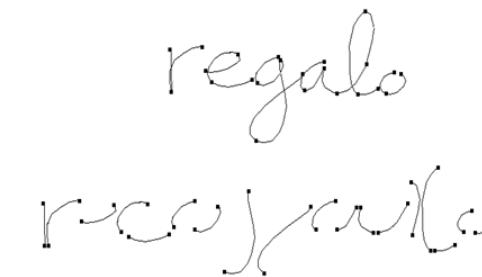
c. Velocity profile after the filtering.



d. Localization of the local minima.



b. Velocity profile before the filtering.

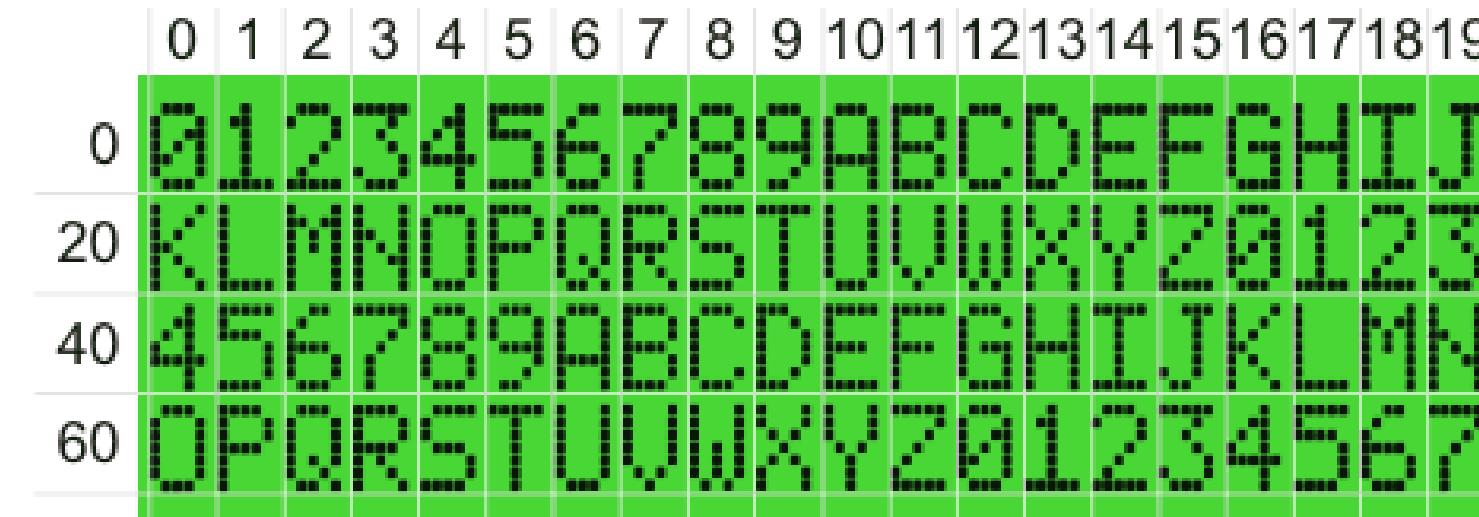


e. Segmentation into strokes.

1994

Proyecto fin de carrera – Compañero de carrera

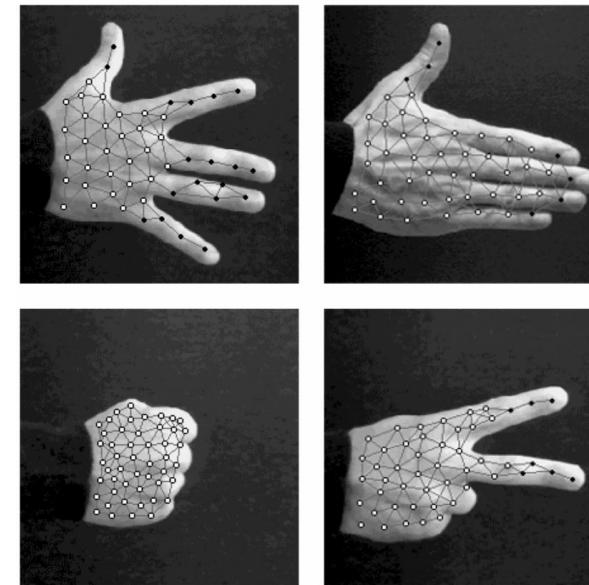
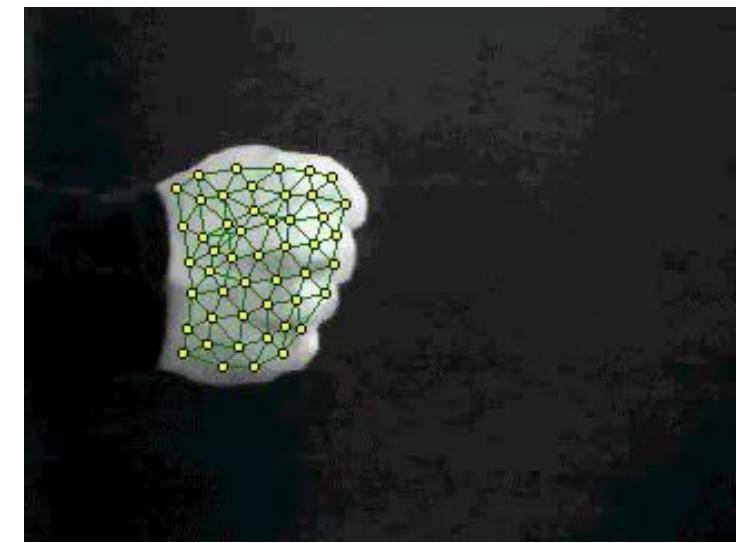
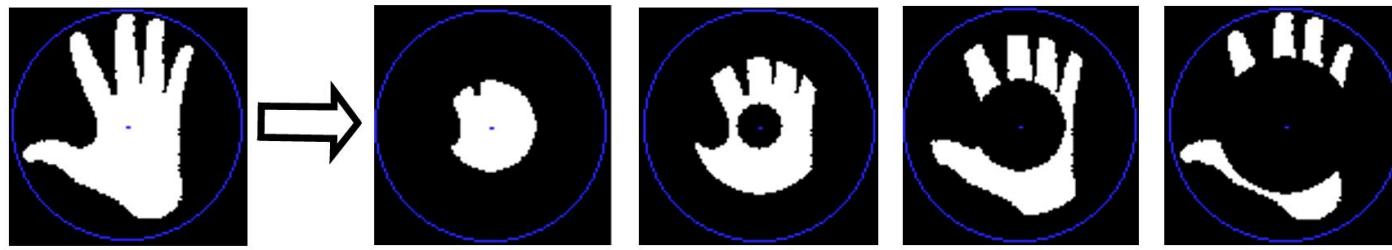
Entrenamiento de perceptrón multicapa para reconocimiento de caracteres ASCII



1994

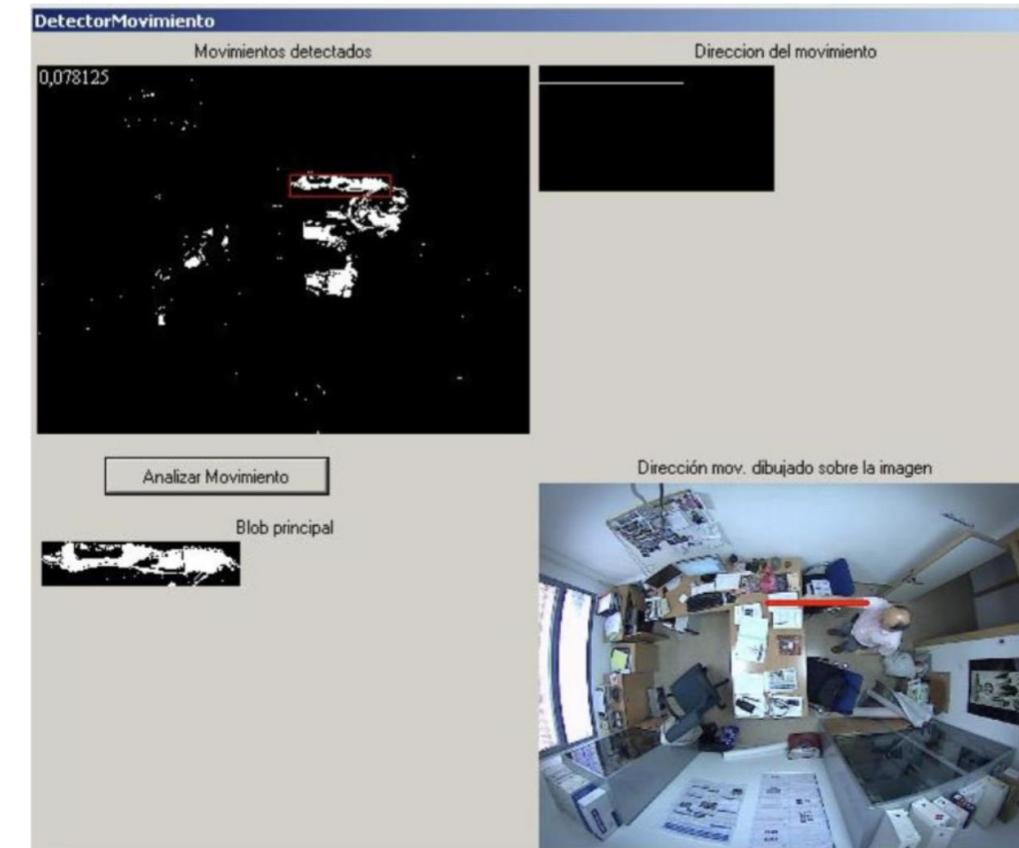
Tesis doctoral - Modelo de representación y procesamiento de movimiento para diseño de arquitecturas de tiempo real especializadas ([enlace](#))

Reconocimiento de gestos de la mano



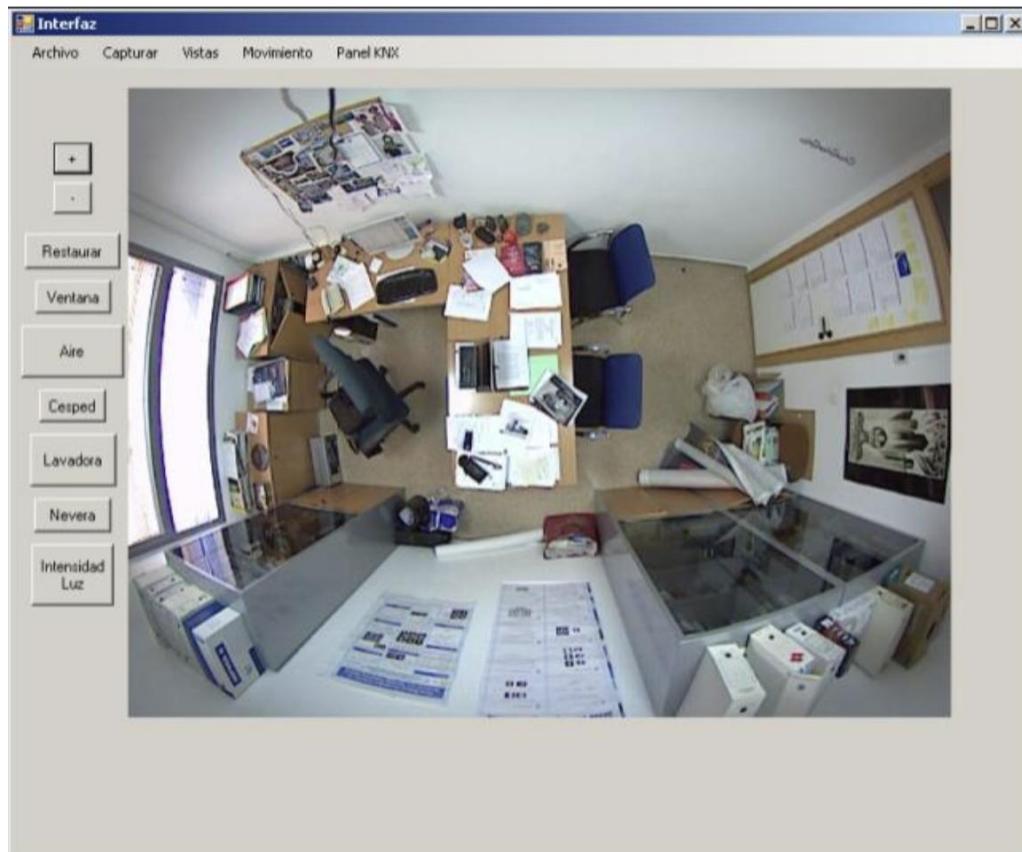
2001

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



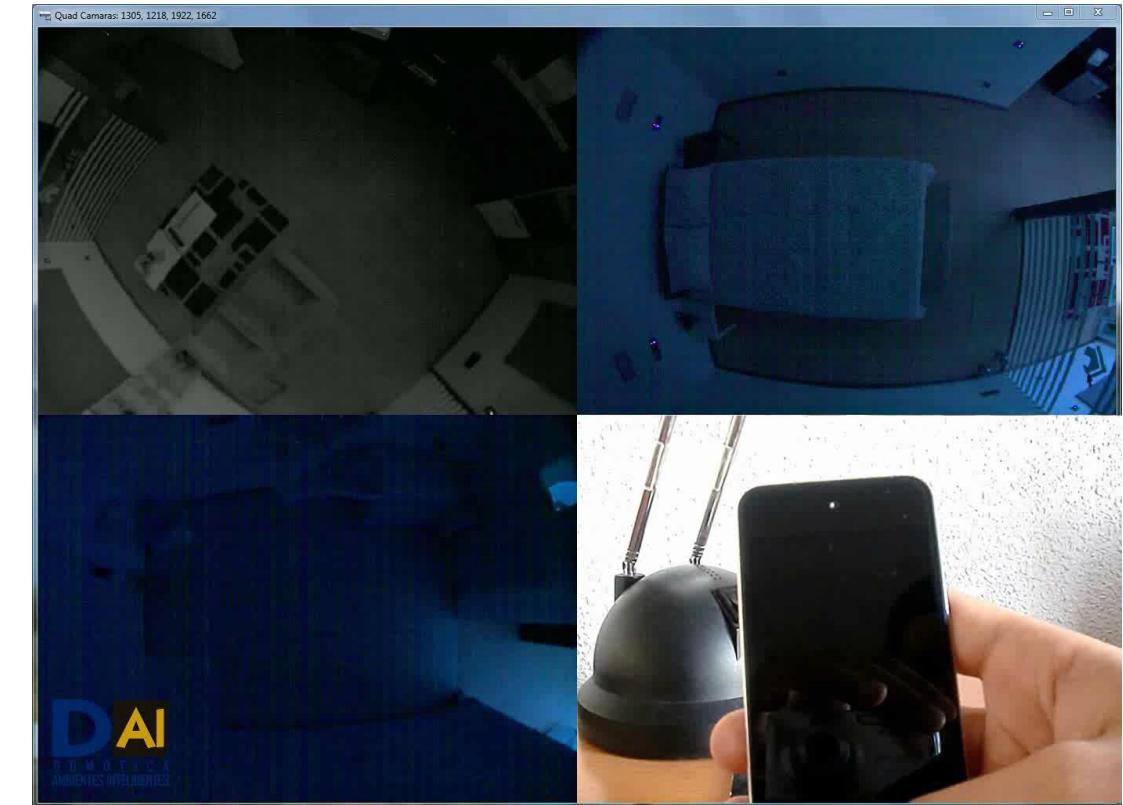
2007

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



2010

metaTIC – Hogar Digital



2010

DAI Lab – Domotics and Ambient Intelligence



2011



Desde 2012



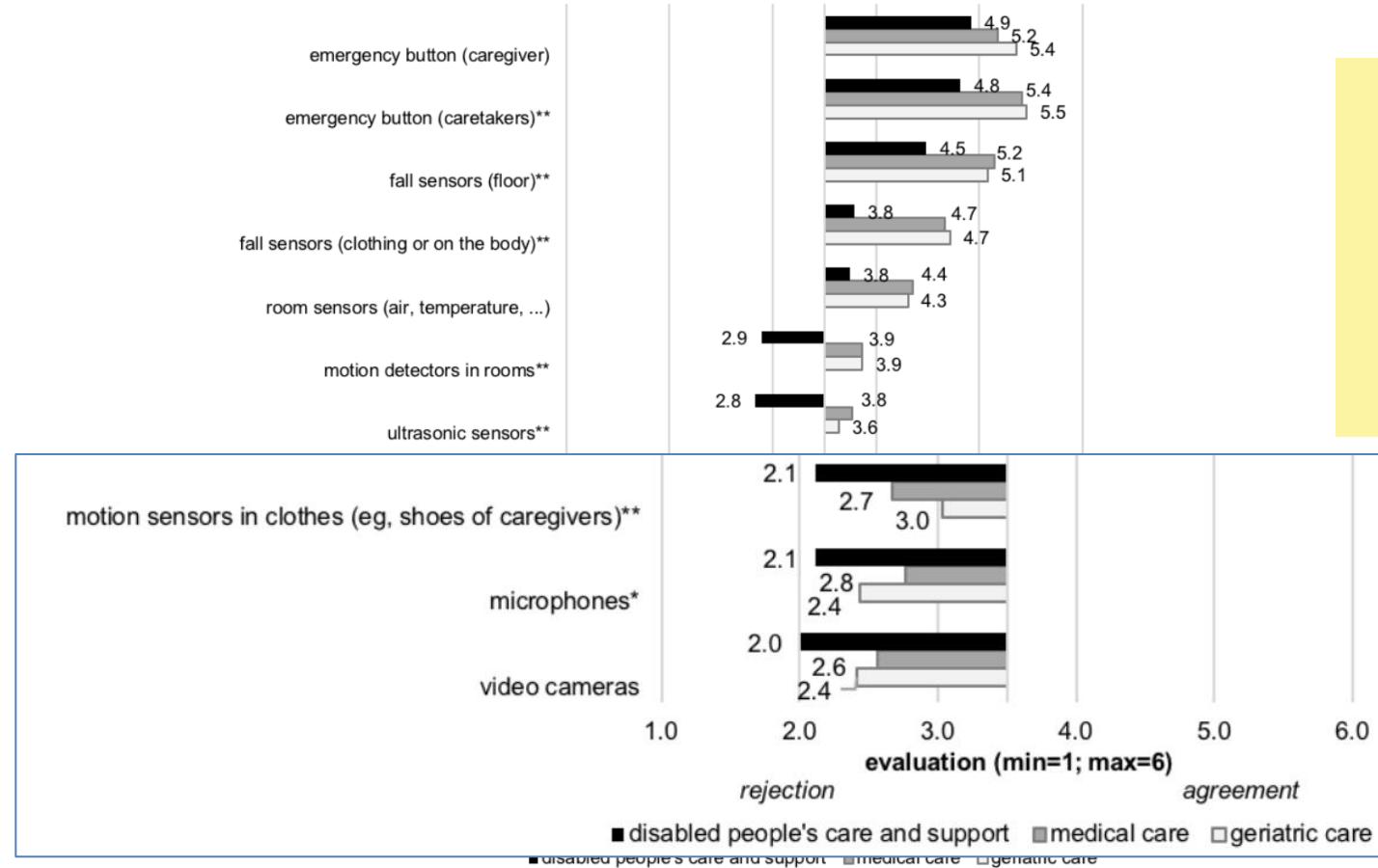
KINECT
for XBOX 360.

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".





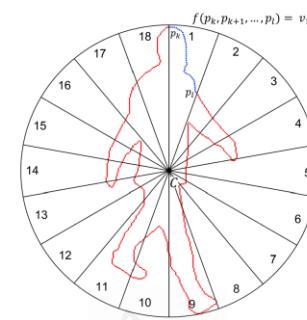
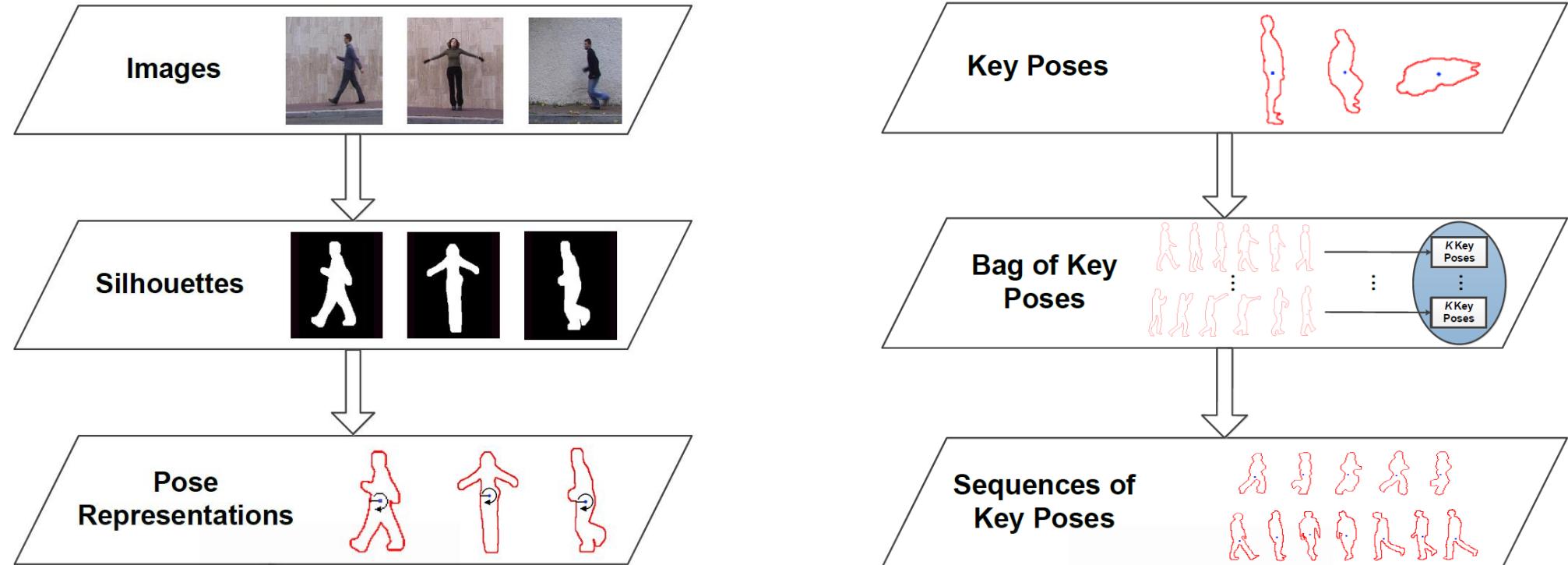


Arning, K., & Ziefle, M. (2015). "Get that Camera Out of My House!" Conjoint Measurement of Preferences for Video-Based Healthcare Monitoring Systems in Private and Public Places. In International Conference on Smart Homes and Health Telematics (pp. 152-164). Springer, Cham.

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).

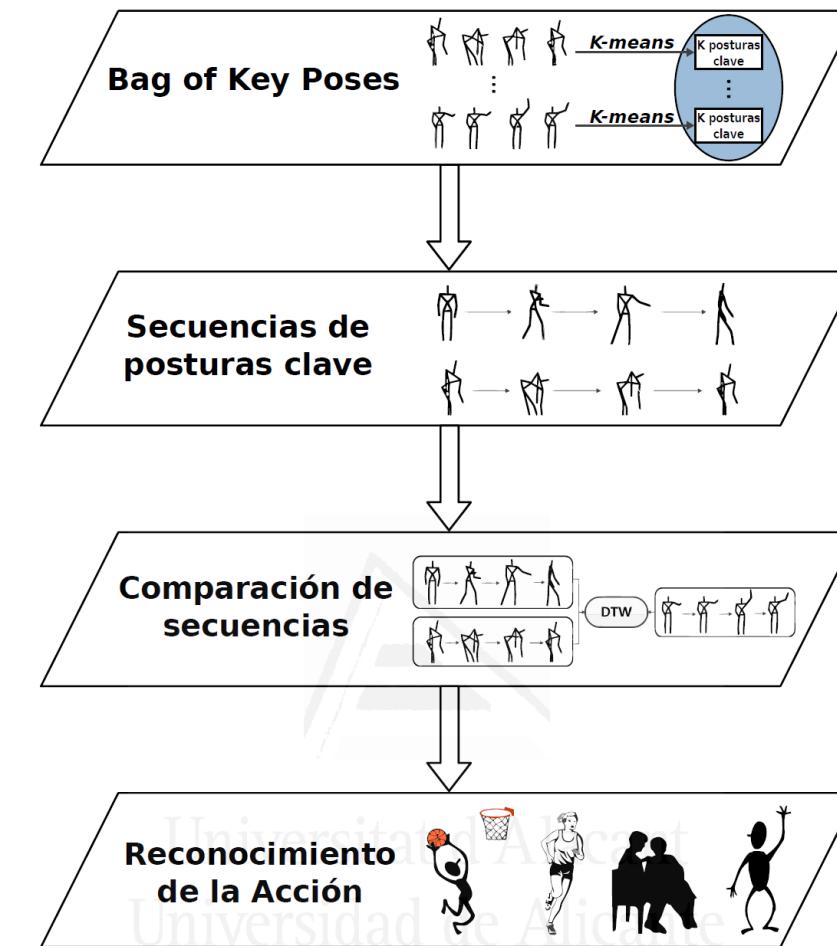
Video-monitorización ética

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



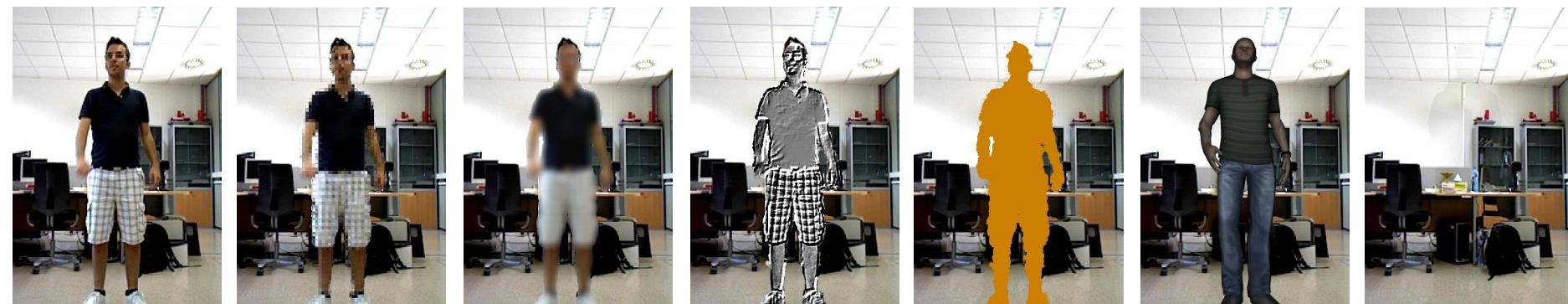
2014

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

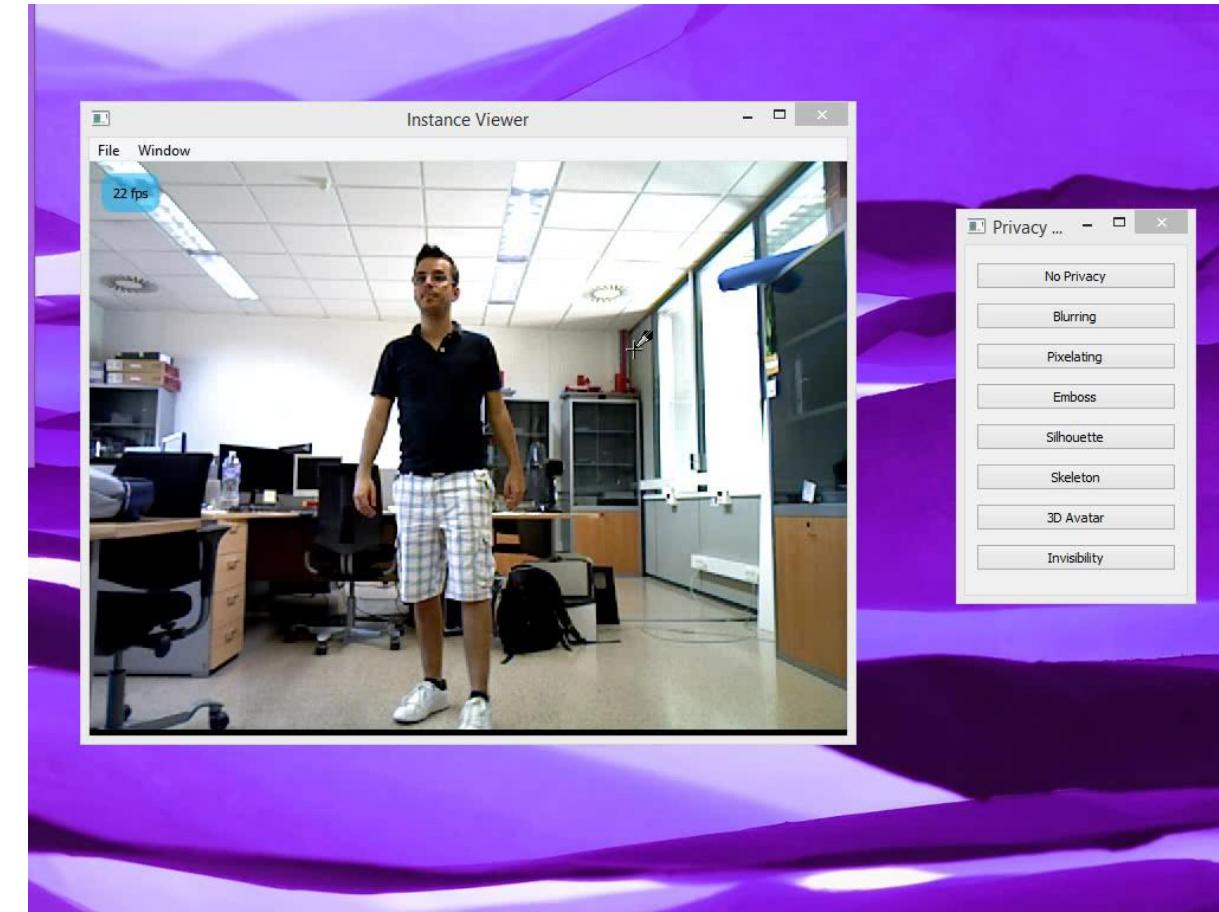


Low Privacy

High Privacy

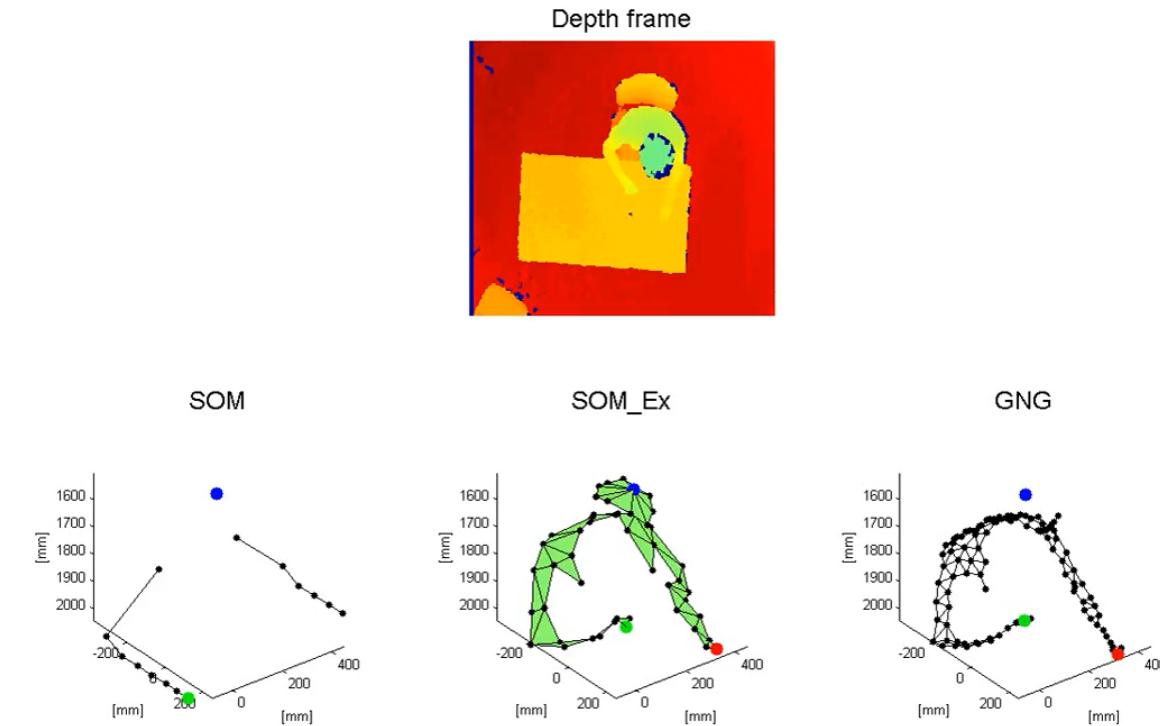
2015

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

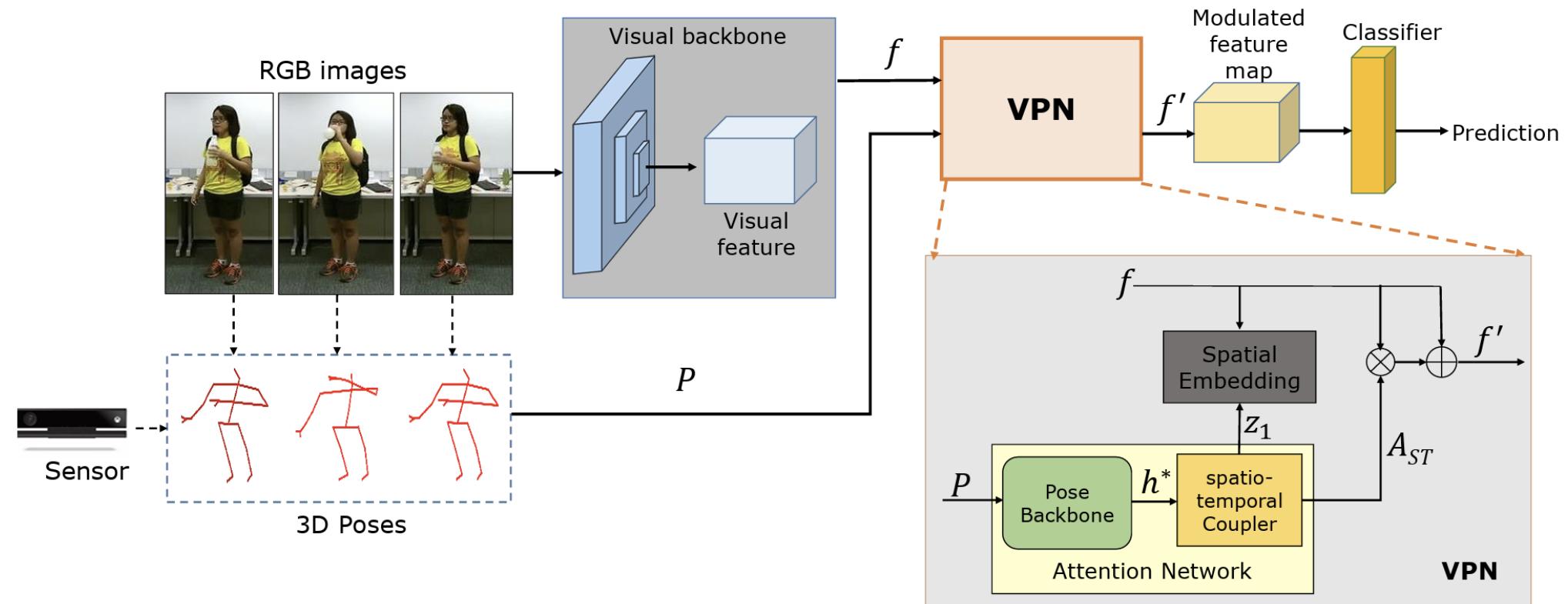


ethicam
smart ethical cameras

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

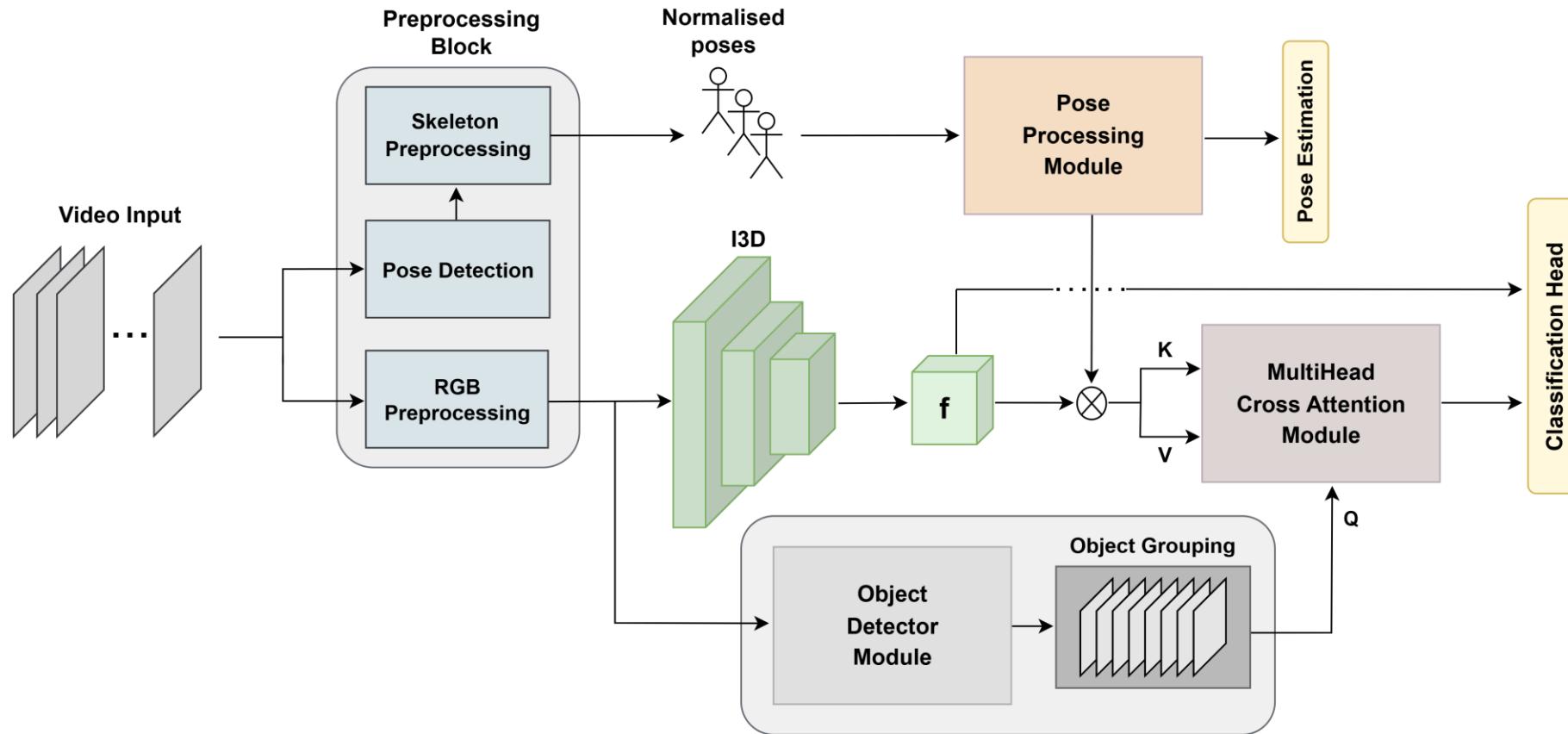
Mejora del reconocimiento de acciones con atención espaciotemporal separable mediante el preprocesamiento alternativo de esqueletos y vídeo



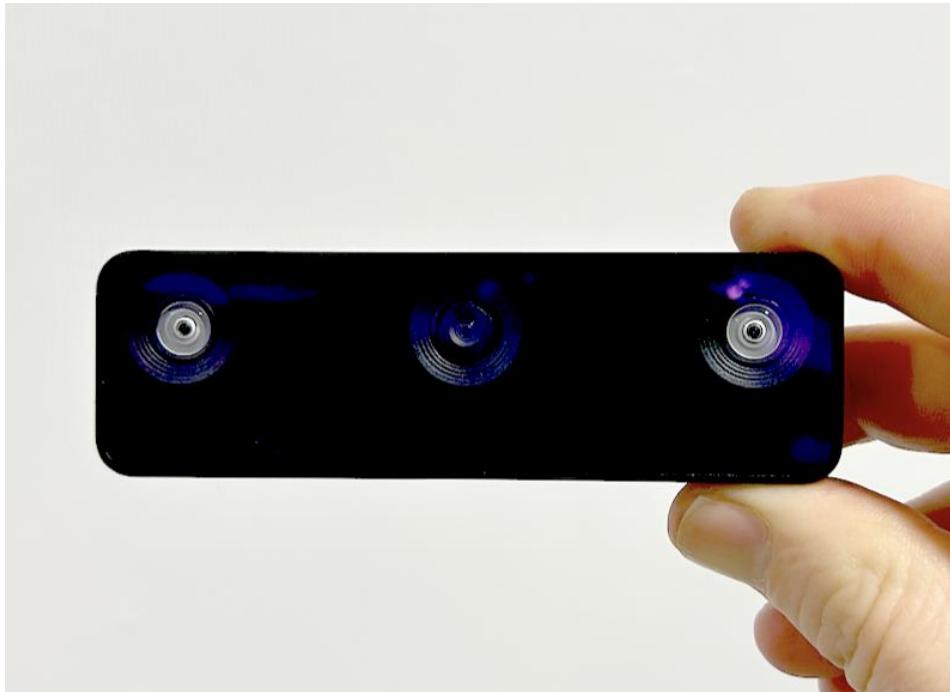
Climent-Perez, P., & Florez-Revuelta, F. (2021). Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing. Sensors, 21(3), 1005.

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



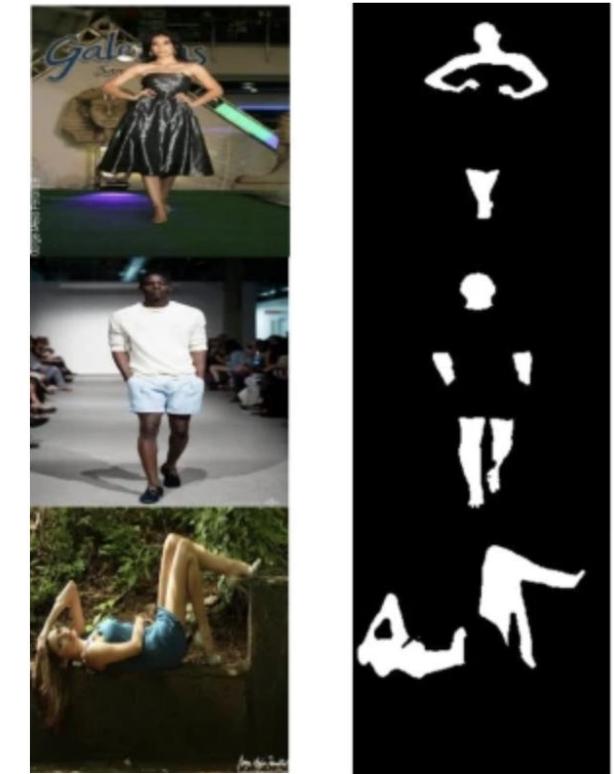
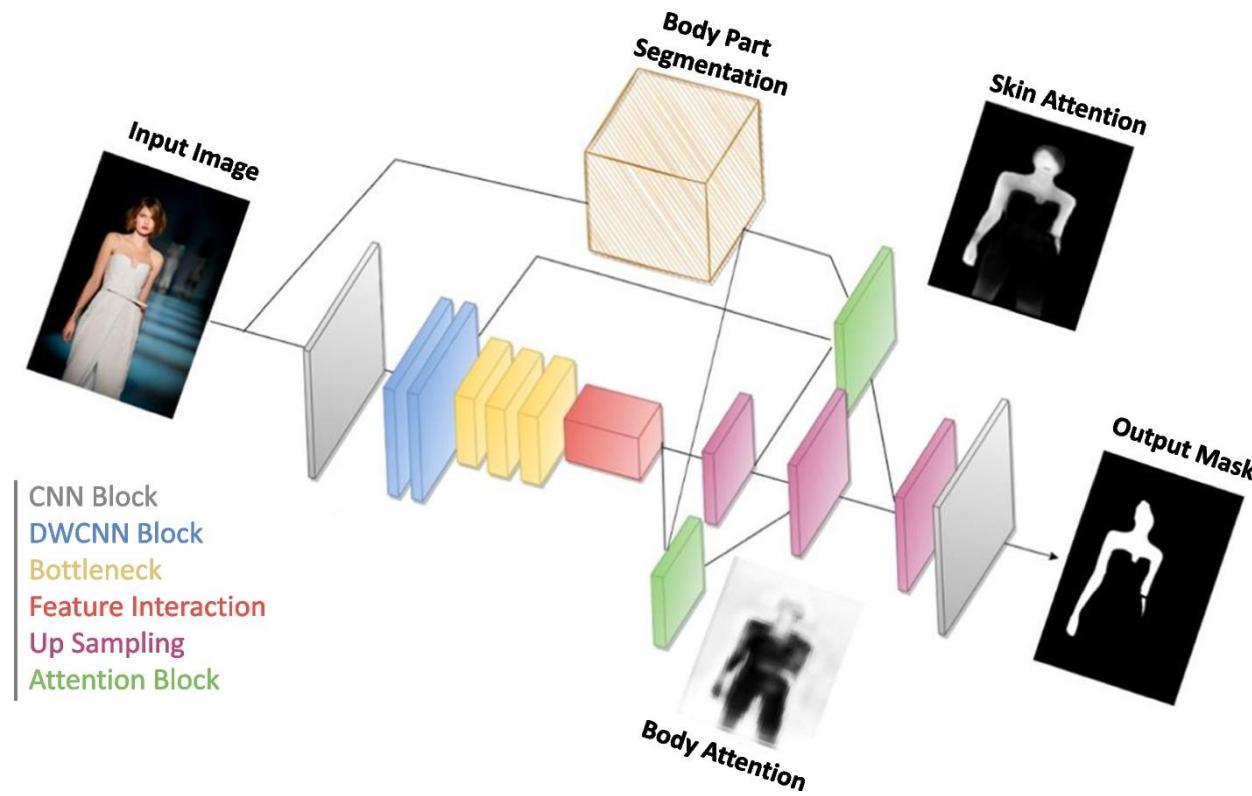
Tesis Kooshan Hashemifard – Detección de caídas empleando una cámara “edge”



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](#)

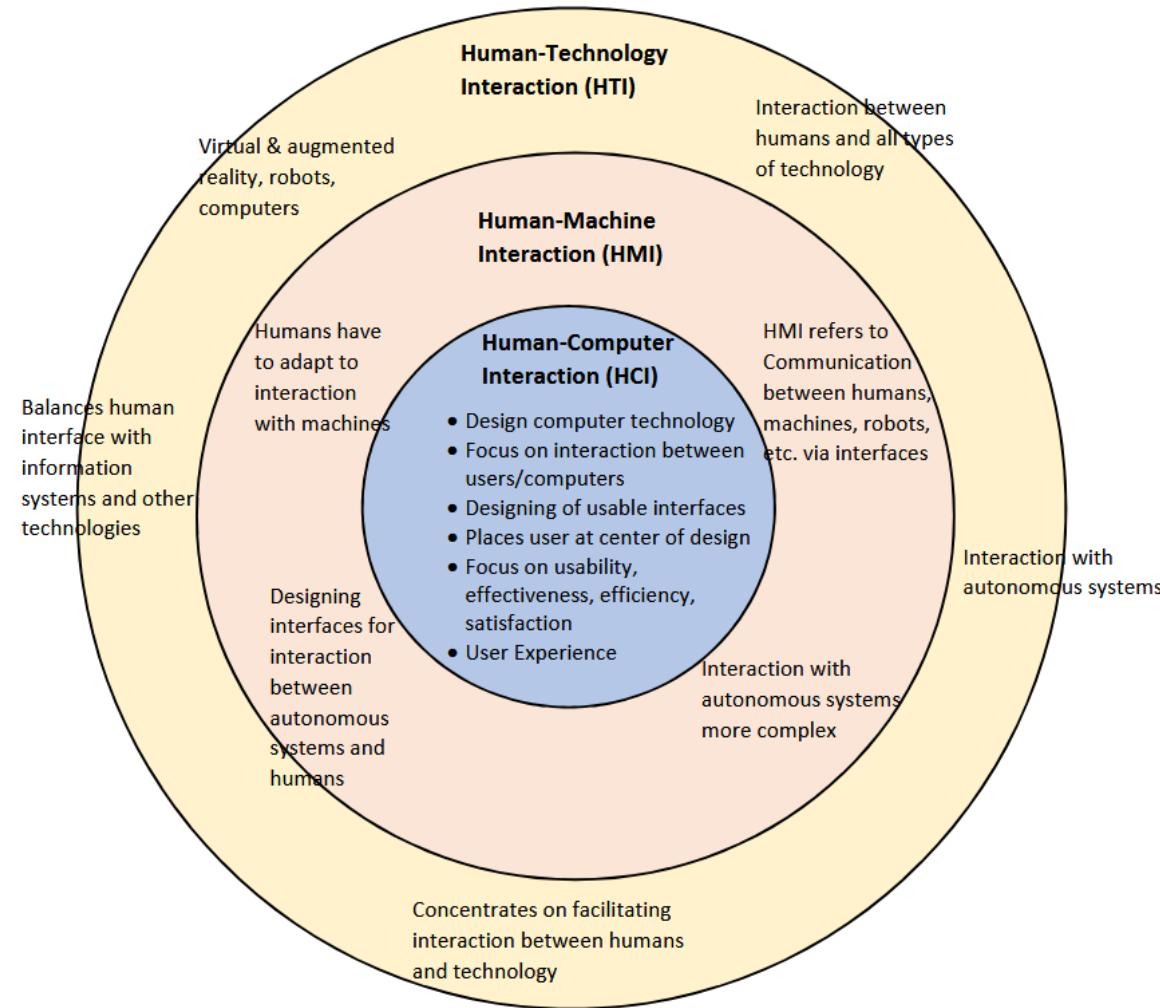
 **paal**
Privacy-Aware and Acceptable Lifelogging services for older and frail people

 **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

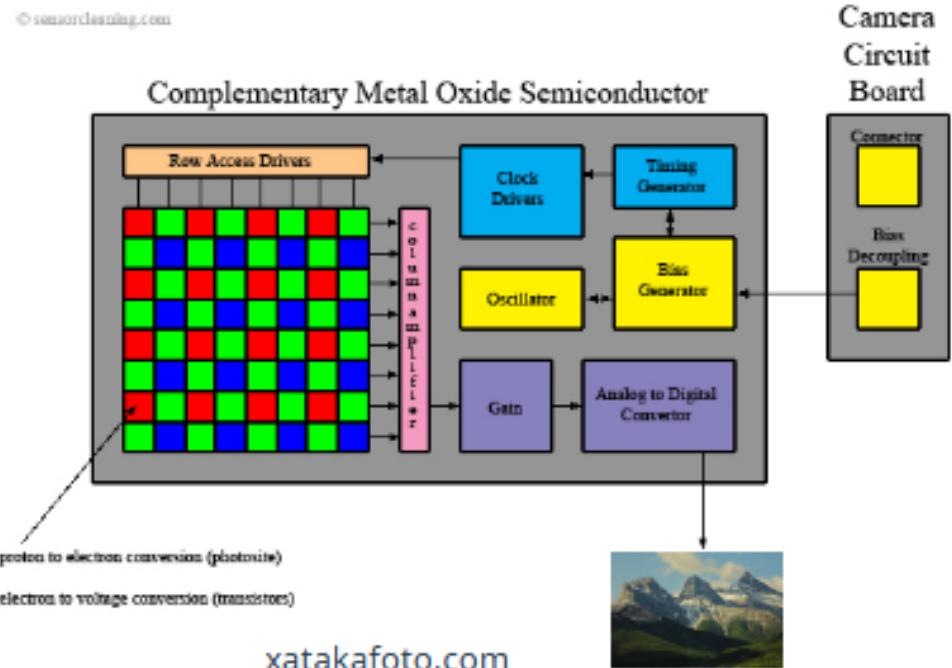
Se trata de una matriz de sensores

- 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

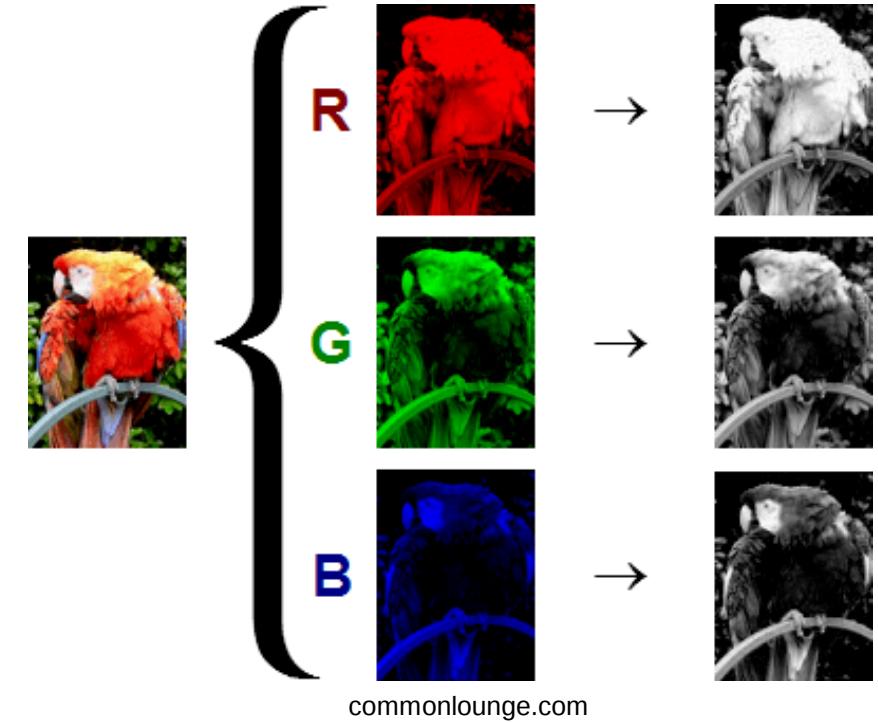
Capturan imágenes en tres canales de color:
Rojo (R), **Verde (G)** y **Azul (B)**.

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).

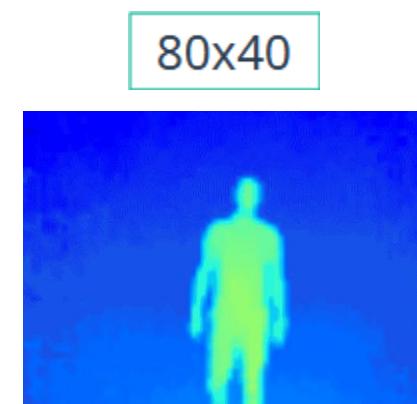
No proporcionan información de profundidad o distancia.



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, ...



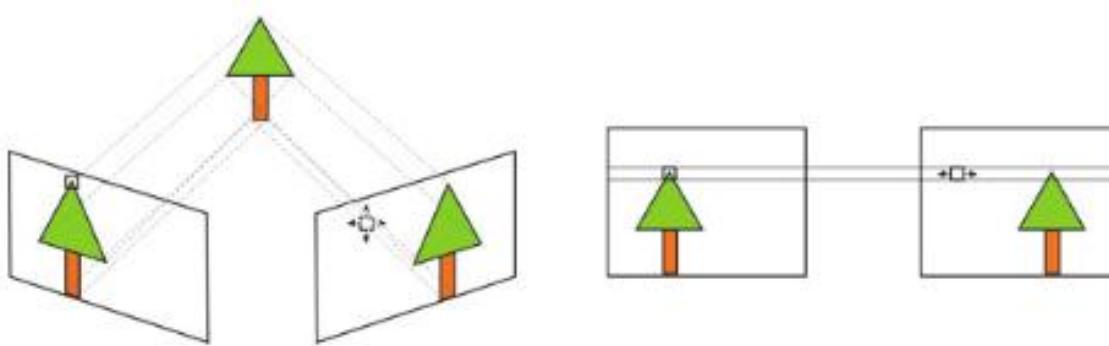
Tipos de cámaras

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’:

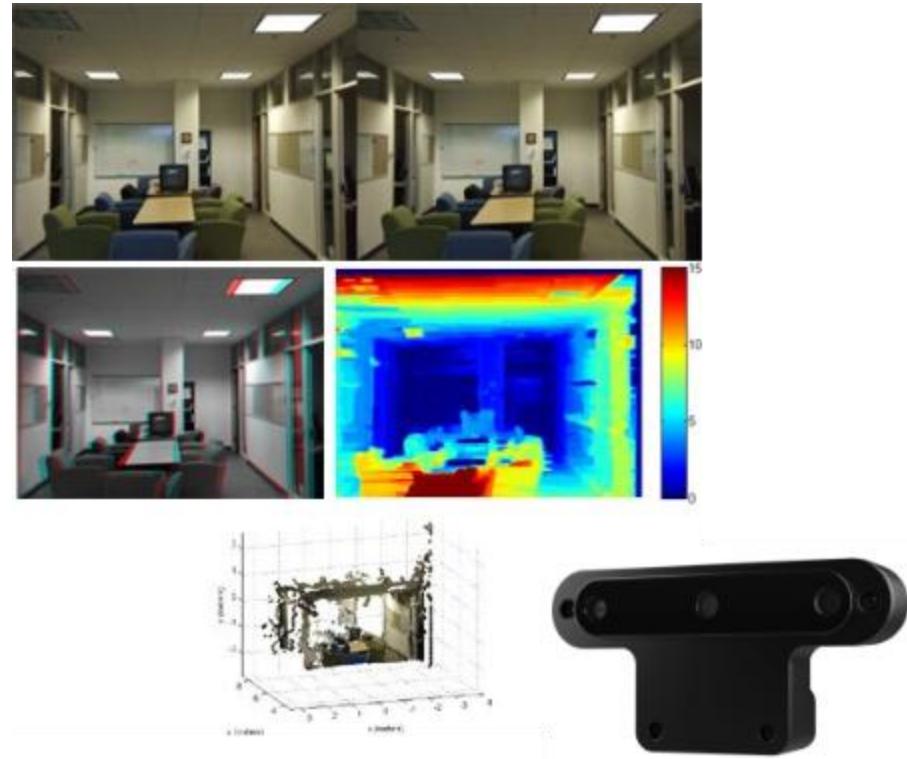
- Basado en la forma en que los animales con dos ojos calculan la distancia: se conoce la línea base (distancia entre cámaras) y luego se utiliza trigonometría para estimar los “mapas de disparidad”.
- Este método usa la información del color, por lo que las superficies “planas” (sin textura) son difíciles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



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

Tipos de cámaras

Visión estéreo - ejemplos



Tipos de cámaras

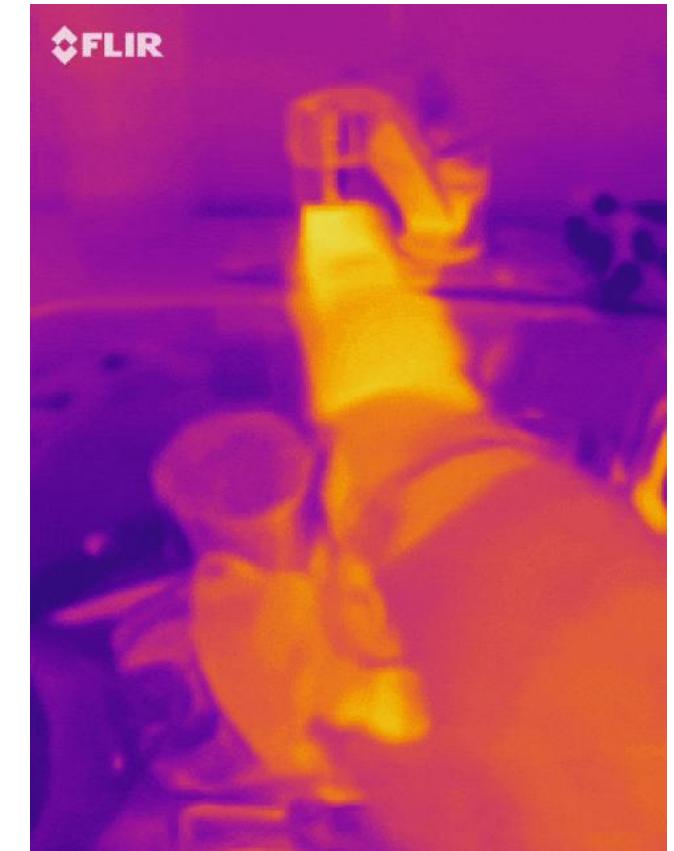
Cámaras de temperatura/calor (cámaras infrarrojas)

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

No emiten ninguna señal ni patrón de infrarrojos.

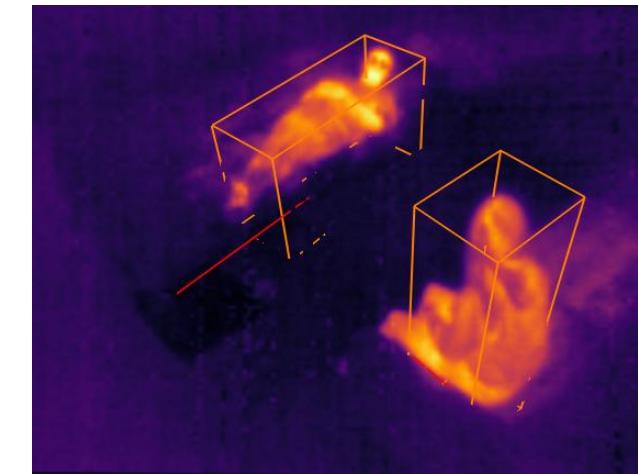
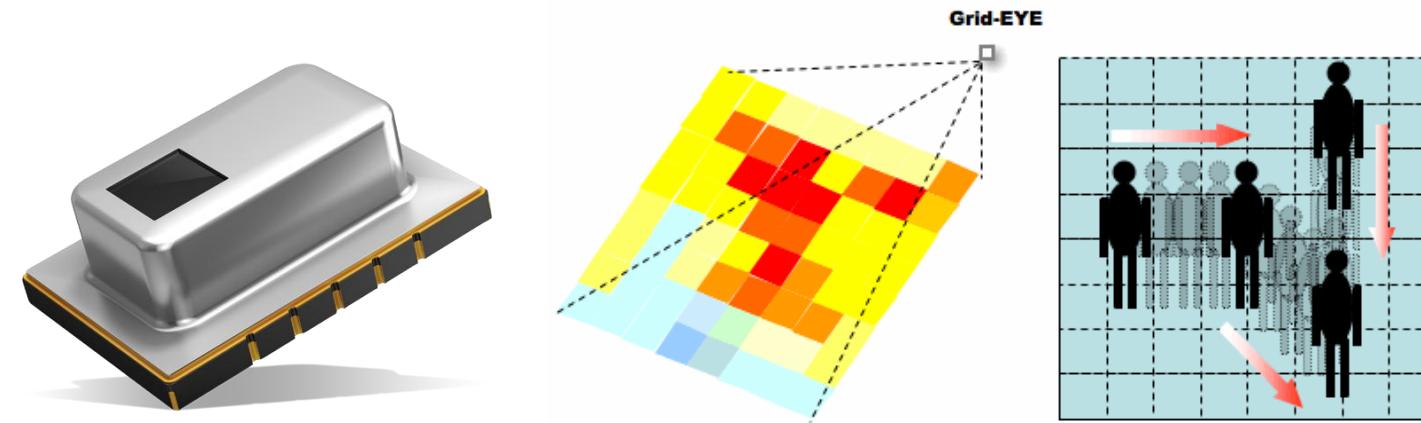
Sin embargo, reaccionan lentamente a los cambios (baja frecuencia de actualización): los objetos dejan un “rastro de calor” al retirarse.

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.com/glass



nytimes.com



microsoft.com



amazon.es



Tipos de cámaras

Cámaras omnidireccionales

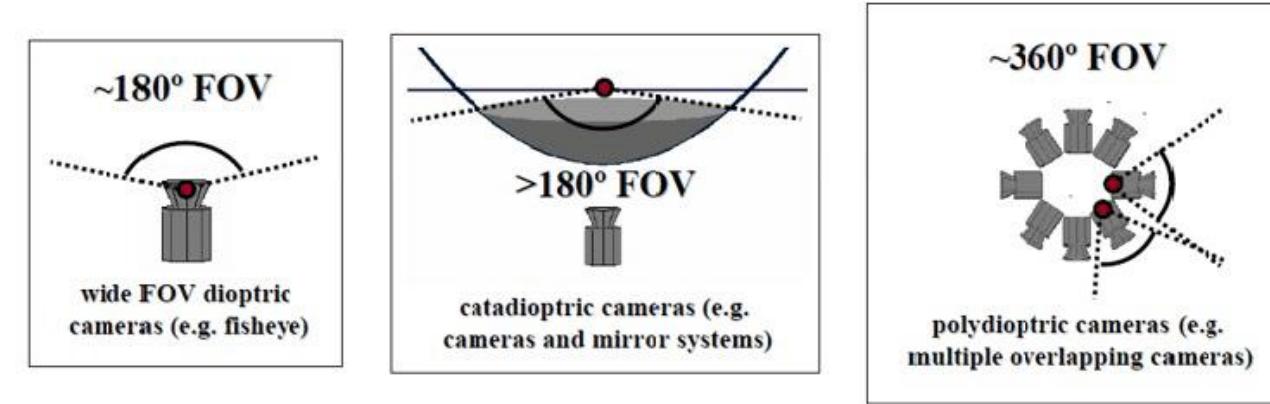
Capturan imágenes con un campo de visión de 360°, permitiendo observar todo el entorno desde un solo punto.

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



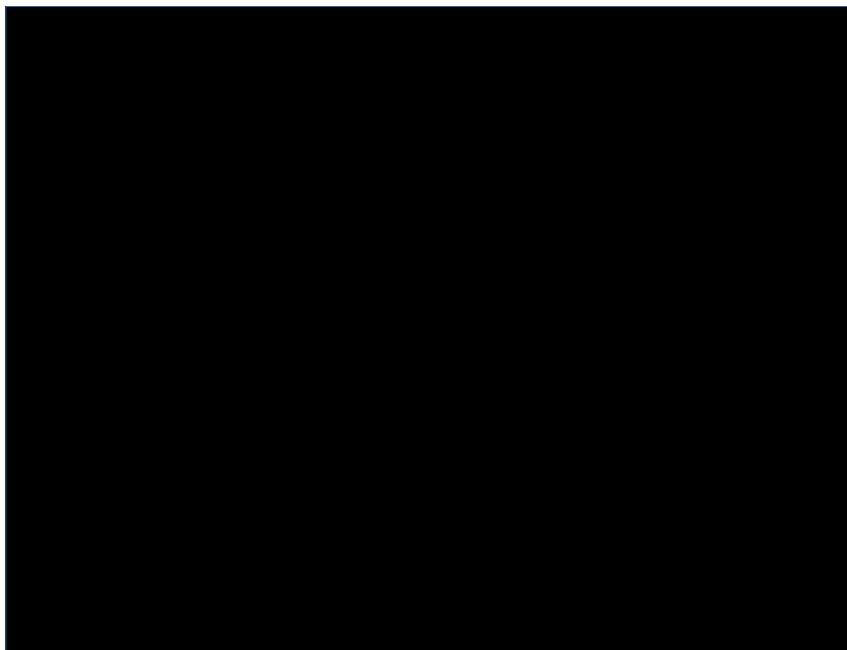
<https://www.google.com/intl/es/streetview/how-it-works>

Tipos de cámaras

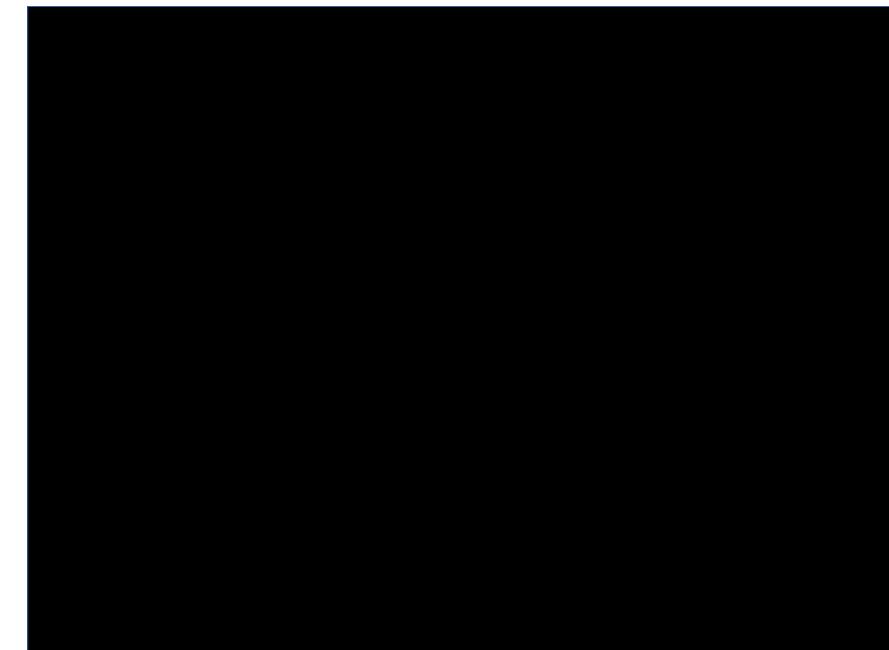
Cámaras omnidireccionales

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>

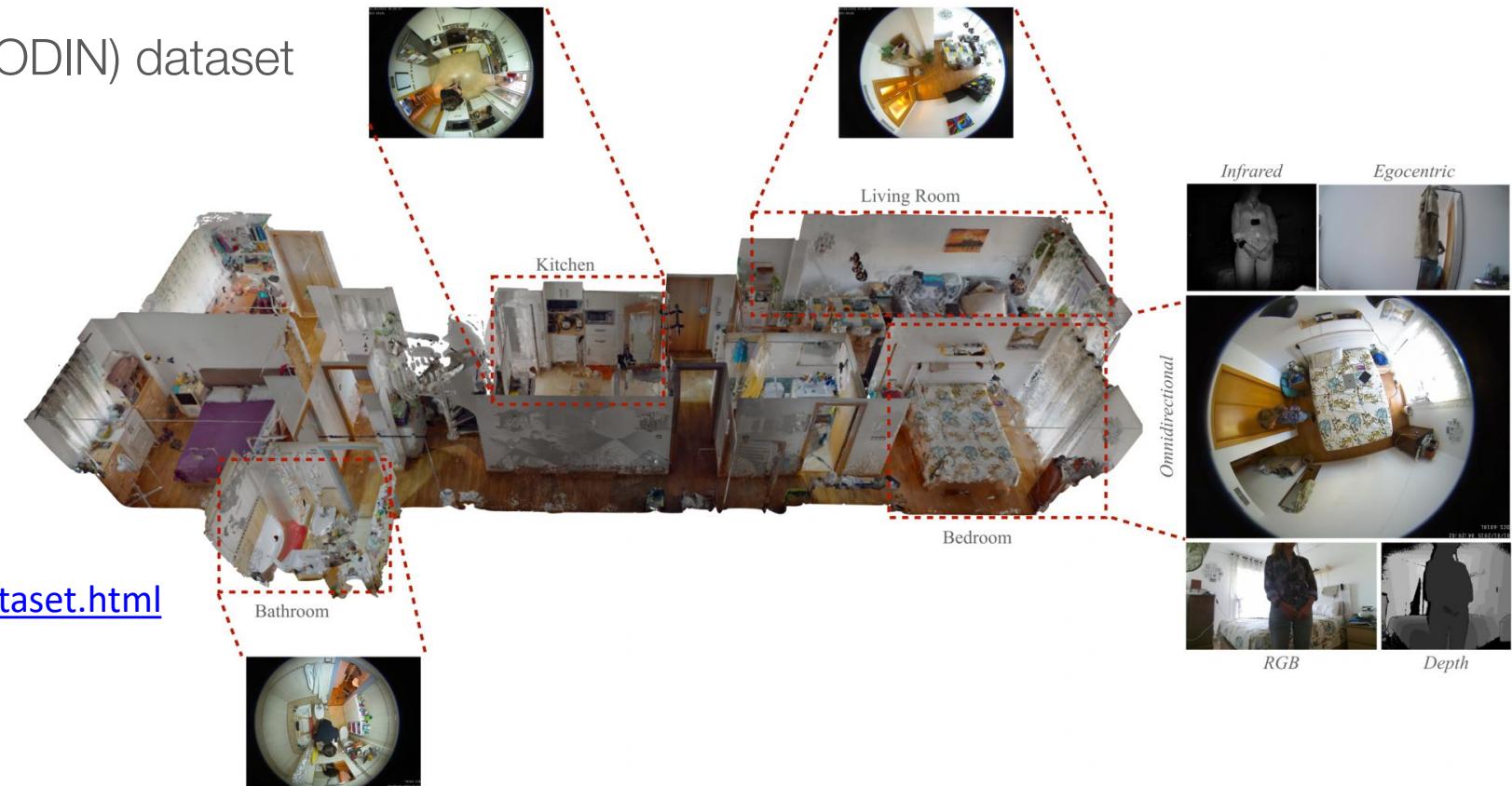


Tipos de cámaras

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

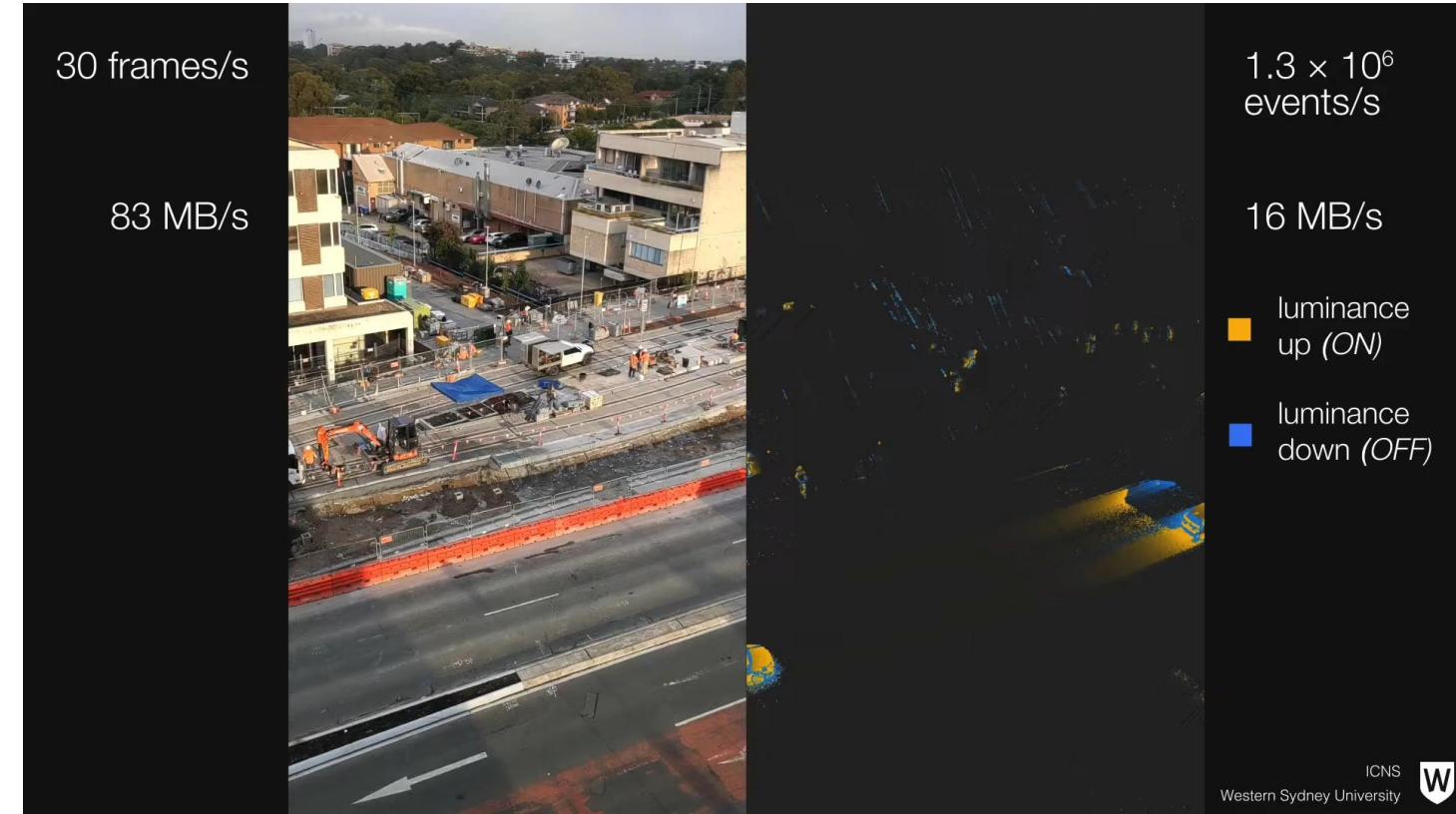
Imitar el funcionamiento del sistema visual biológico: registran solo los cambios en la escena, no fotogramas completos.

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



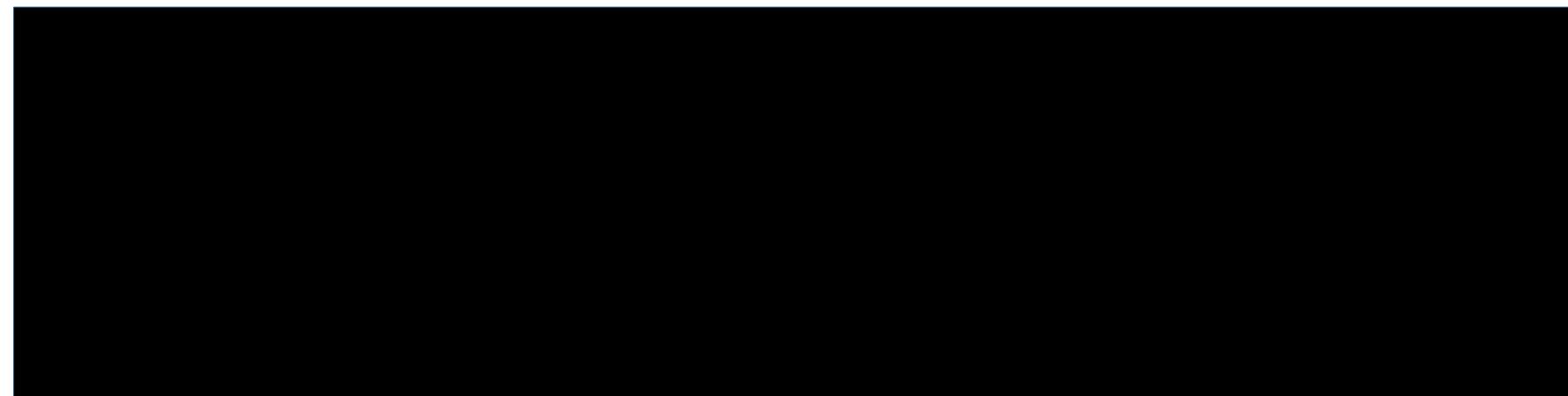
Tipos de cámaras

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

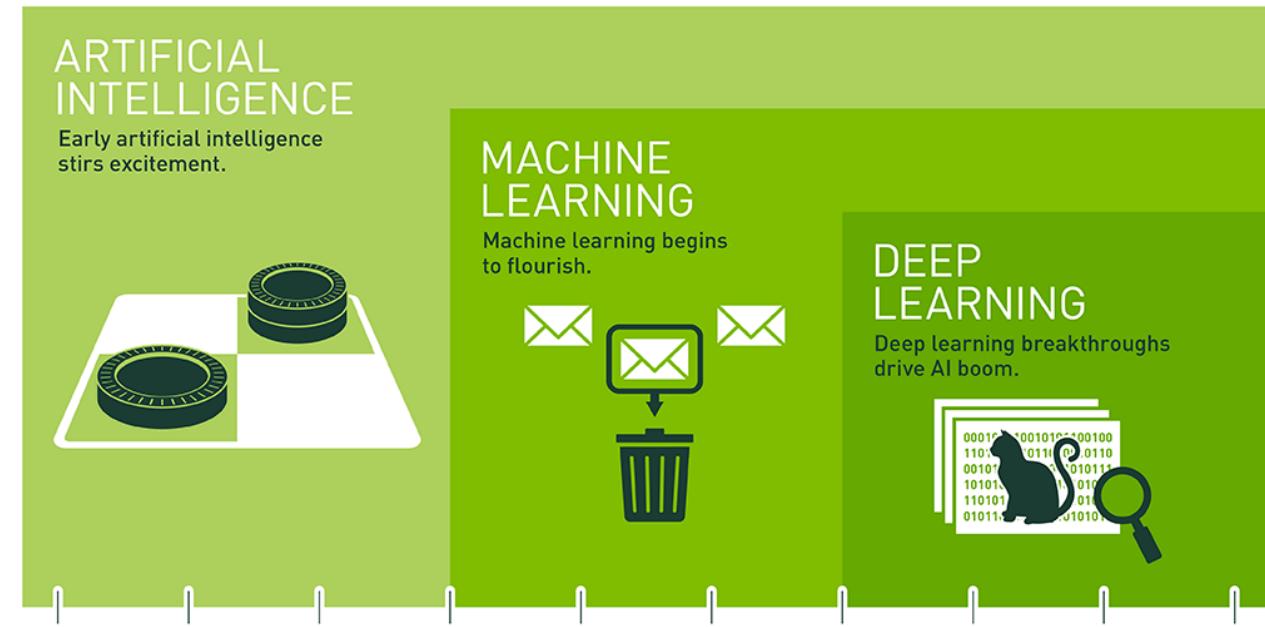
Procesamiento

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.

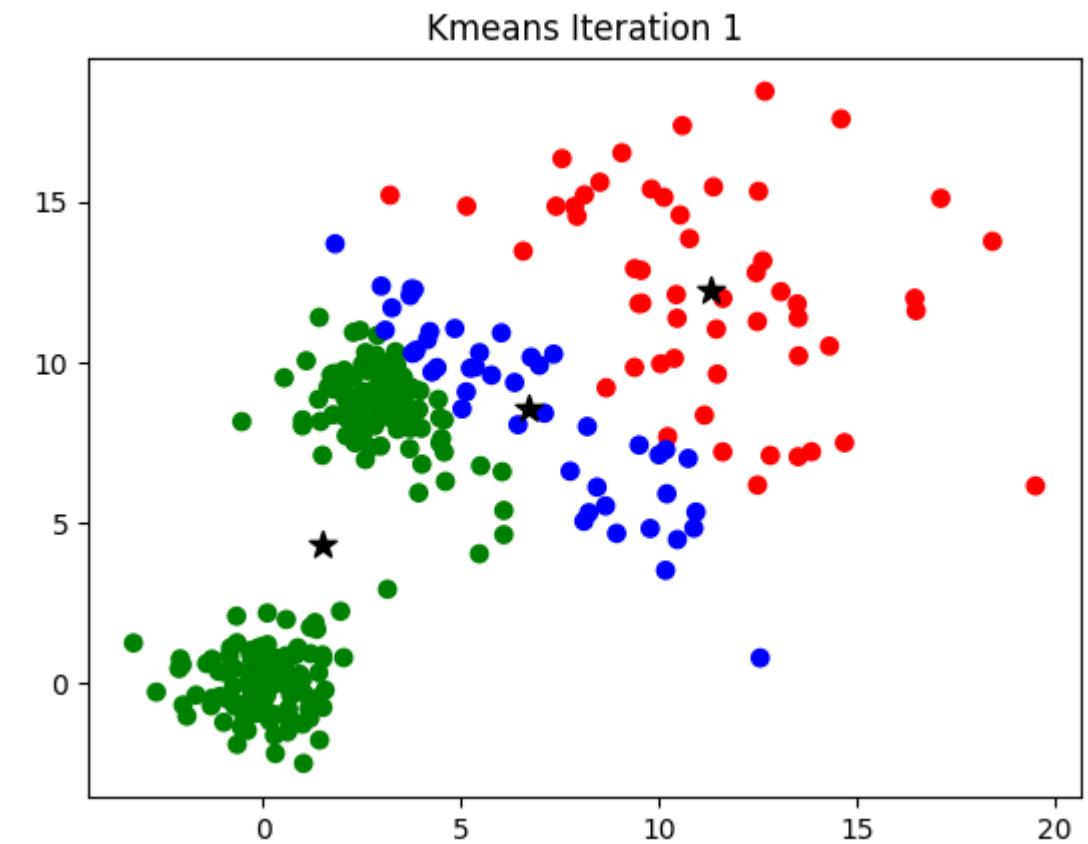
Clustering (no supervisado, genera grupos)

- 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).

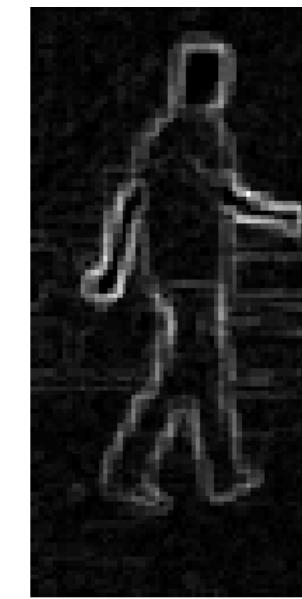
Clasificadores que utilizan descriptores locales: Bag of [visual] Words (BoW).

Pre – deep learning

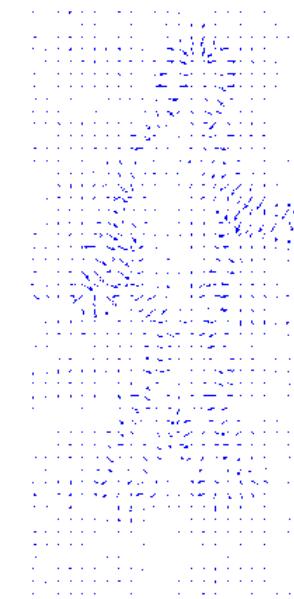
Descriptores – Histogramas de gradientes orientados (HOG)



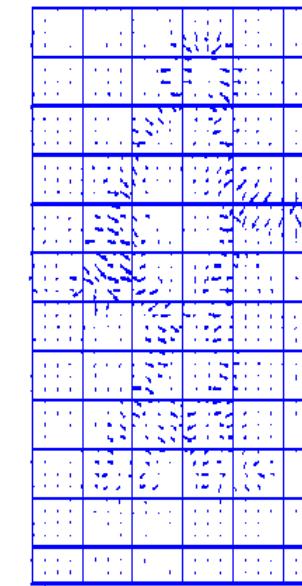
a



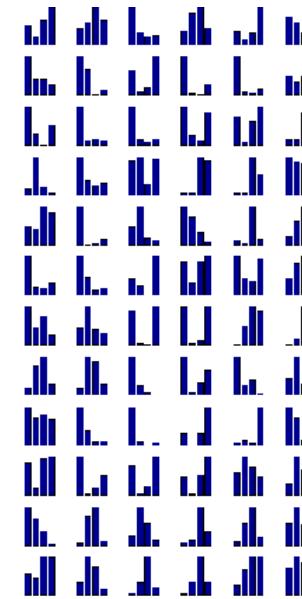
b



c



d



e

Pre – deep learning

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

Descriptores – Extracción de descriptoros “en pirámide”

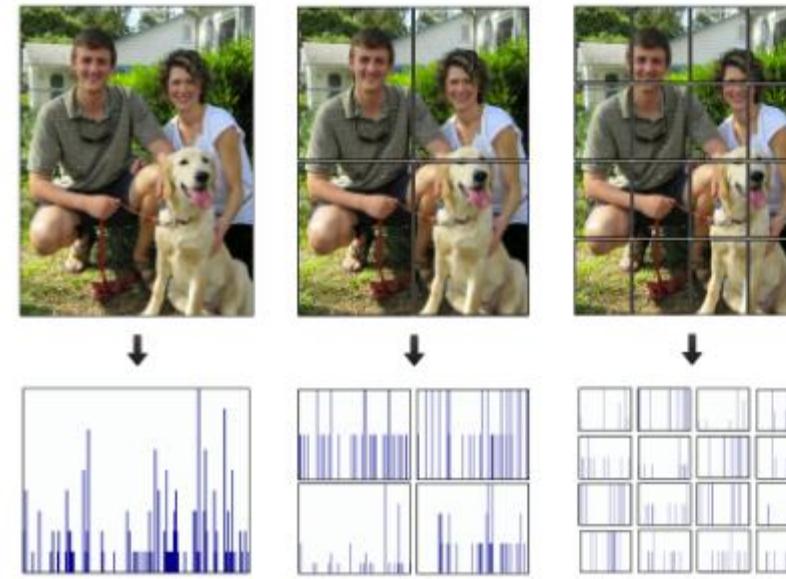


Figure 3: Visual representation of partitioning an image into sub-images and constructing the histograms.

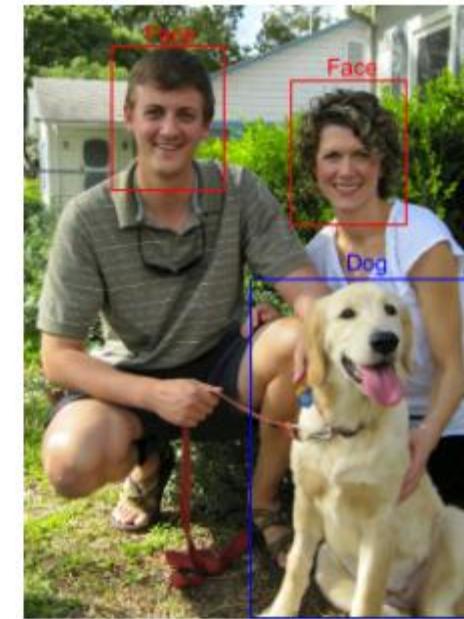


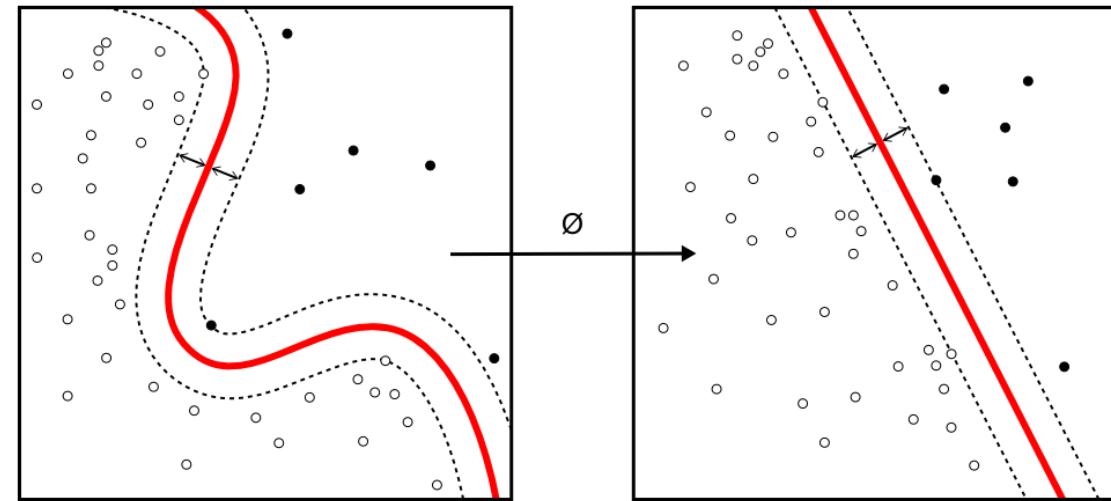
Figure 4: Results showing both image classification and localization.

<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

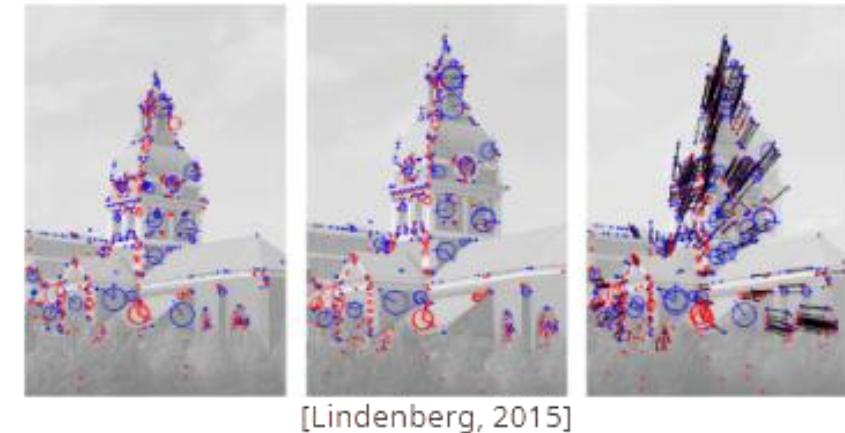
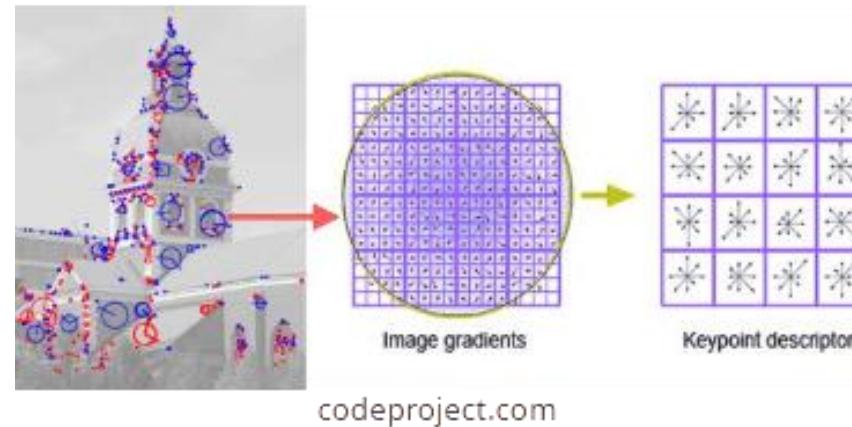
Pre – deep learning

Descriptoros 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

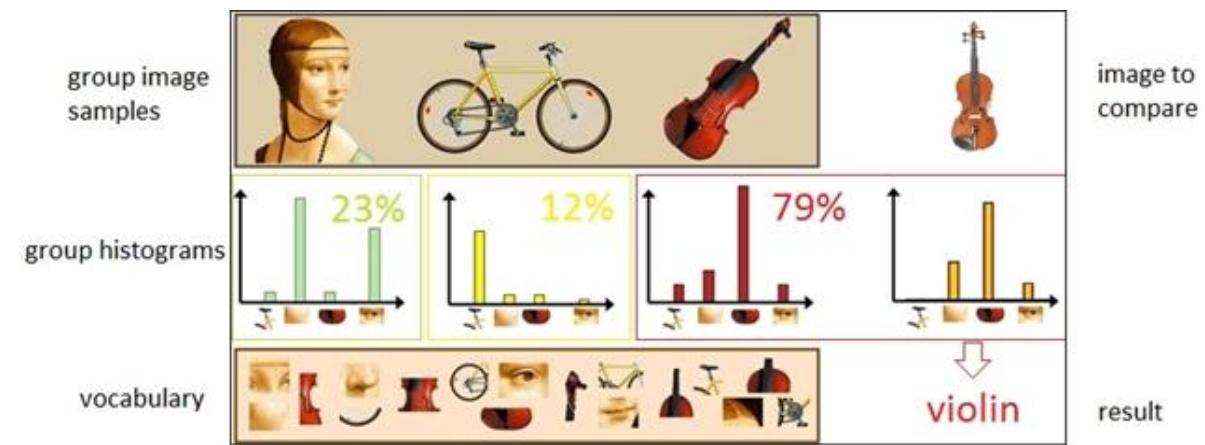
¿Cómo usar entonces descriptores ‘locales’?

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

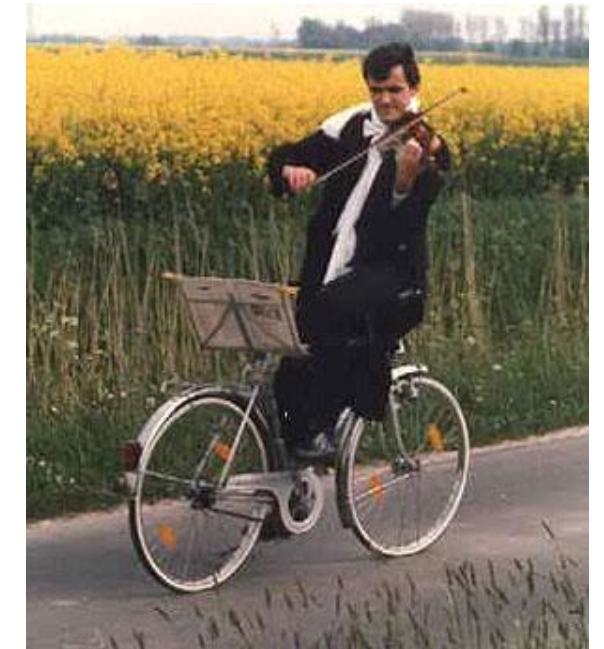
Conclusiones

¡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”.



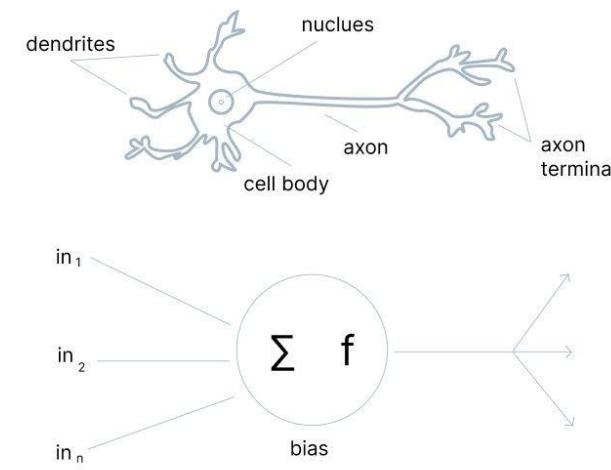
<https://es.pinterest.com/pin/383861568257586224>

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?



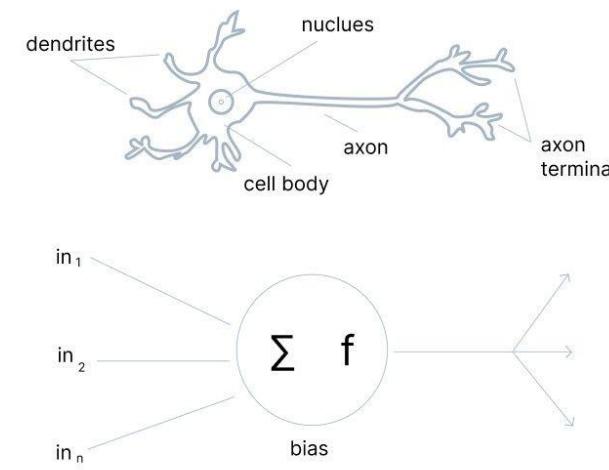
Deep learning

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).

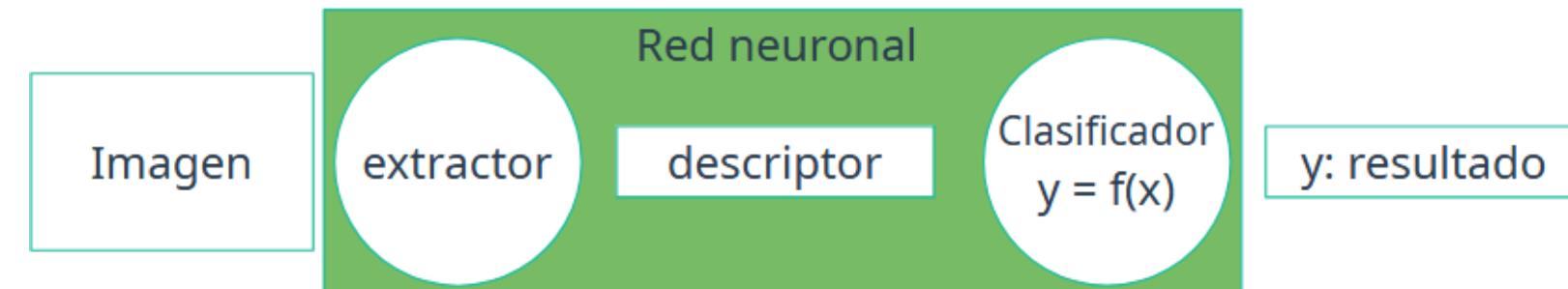
Deep learning

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.

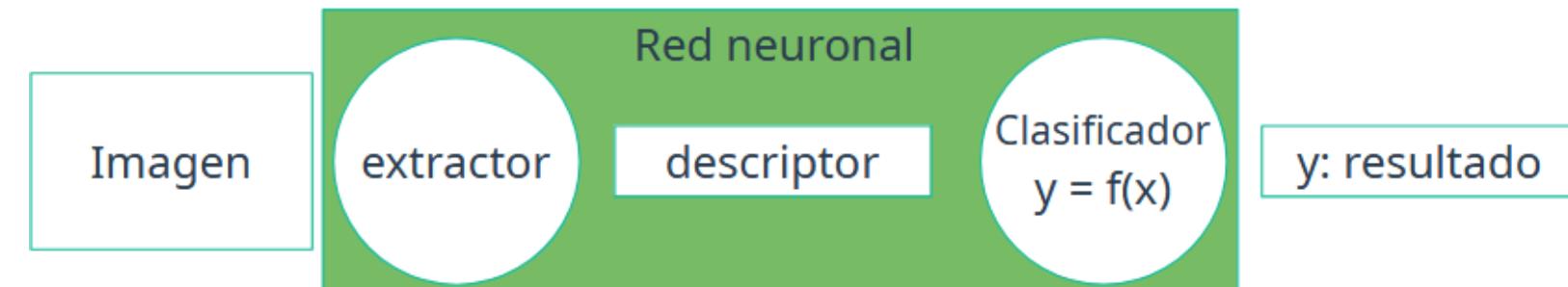


Deep learning

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.

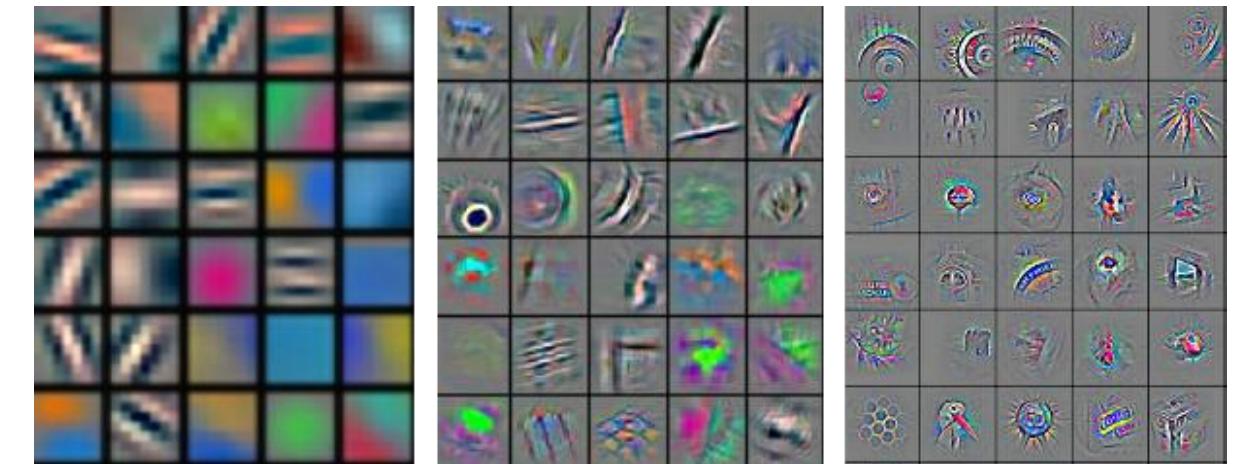
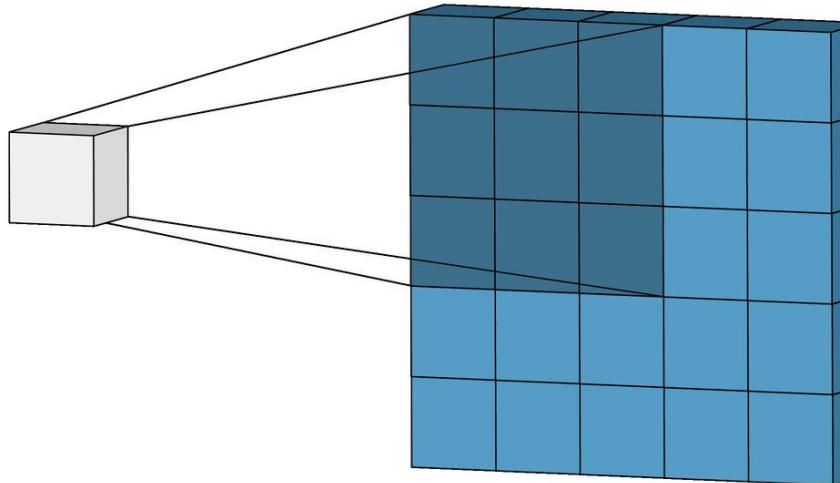
Para “resumir” los descriptores, se aplica una operación llamada pooling. Al final de la red hay neuronas que realizan la clasificación.

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”



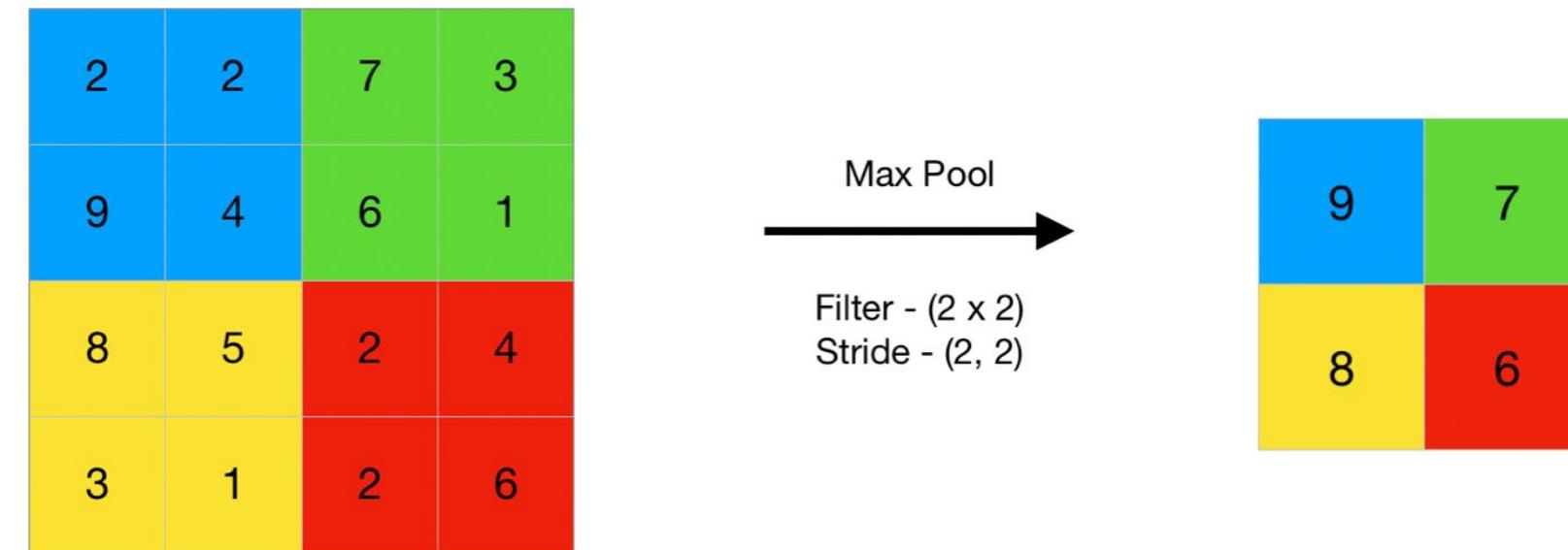
<https://medium.com/data-science/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1>
<https://stackoverflow.com/questions/41598858/how-can-filters-later-in-a-cnn-recognize-color>

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

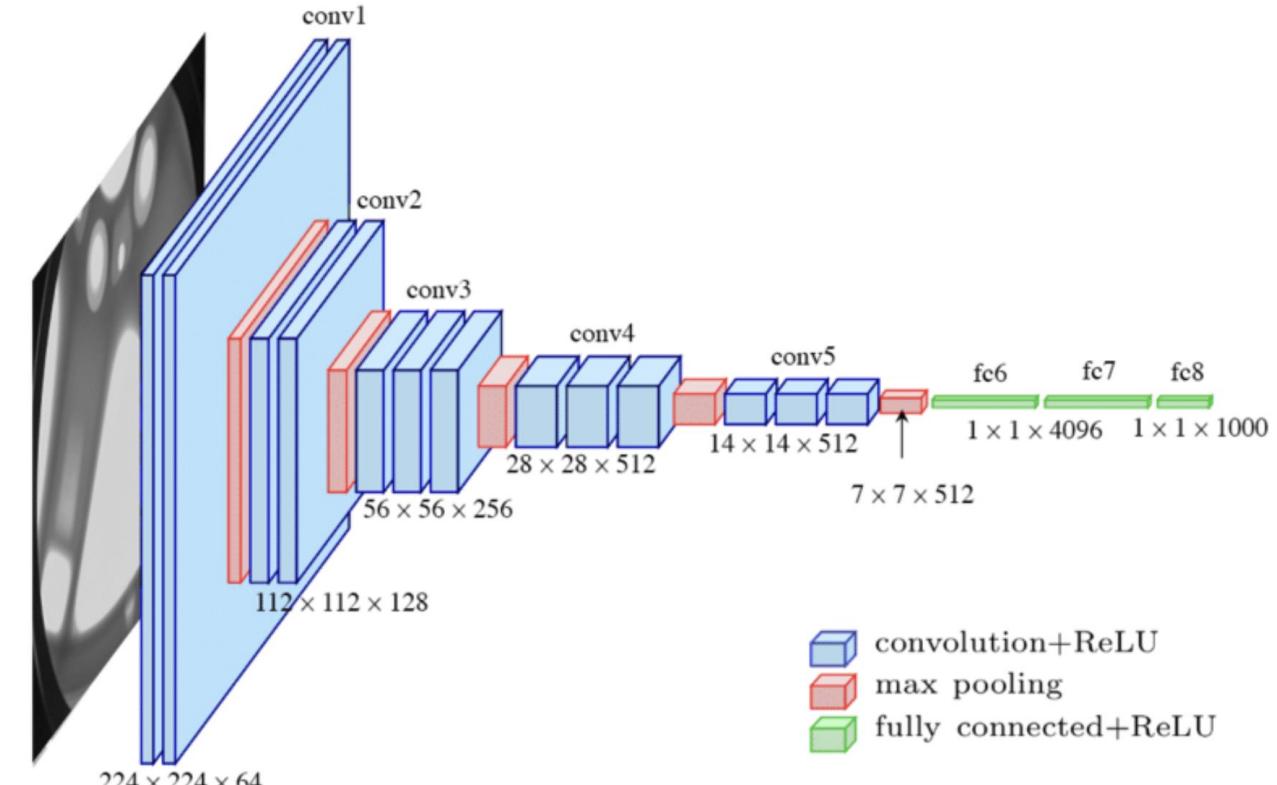


Fig. credits: Khuyen Le, medium.com

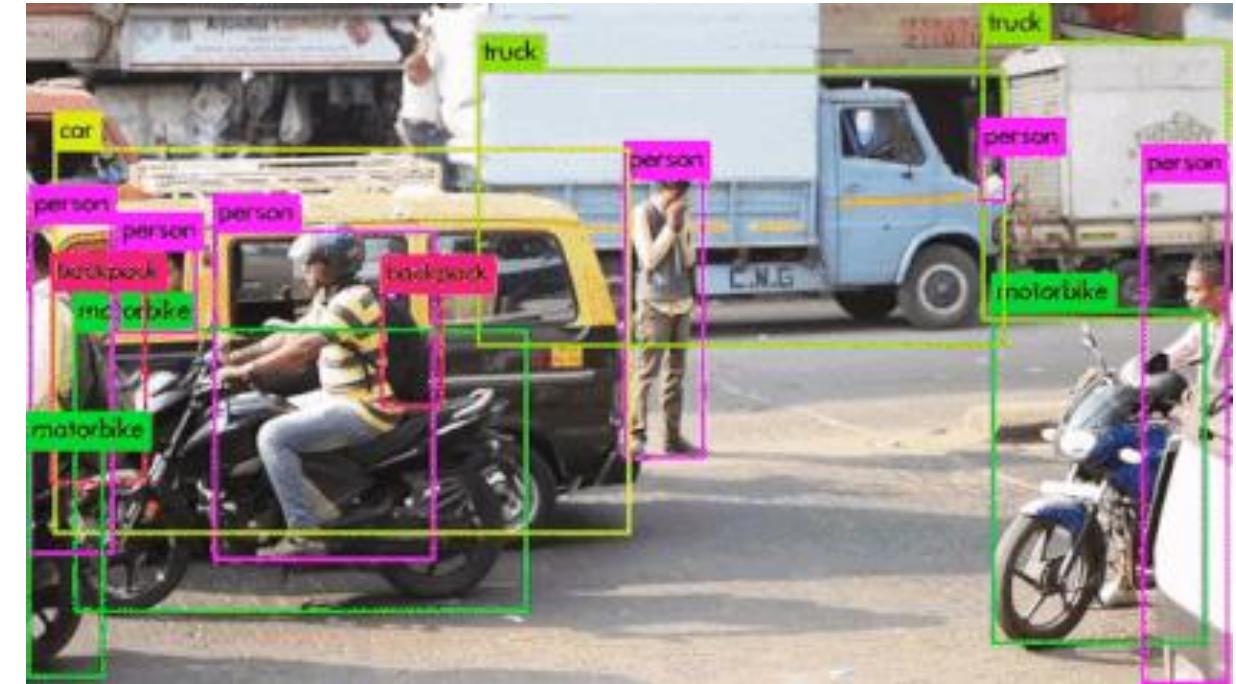
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Deep learning

Redes neuronales convolucionales (CNNs) – YOLO v3

No solo para clasificar imágenes enteras

YOLO realiza detección y
clasificación de objetos



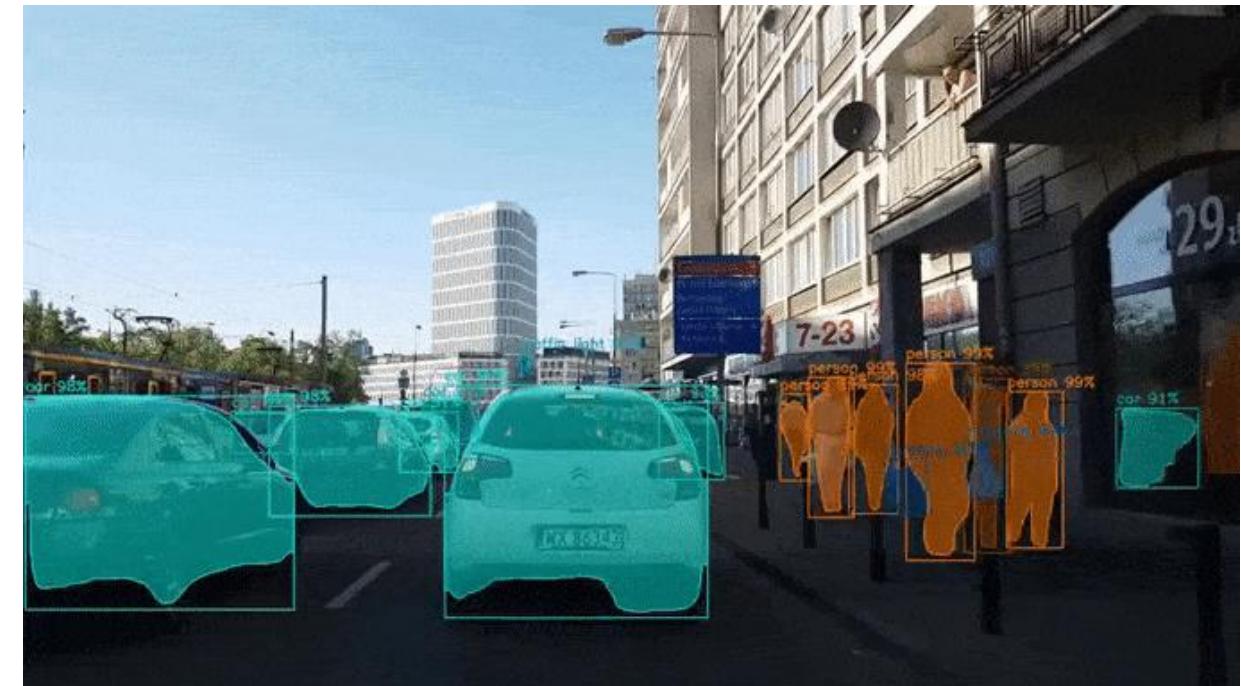
<https://medium.com/data-science/review-yolov3-you-only-look-once-object-detection-eab75d7a1ba6>

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



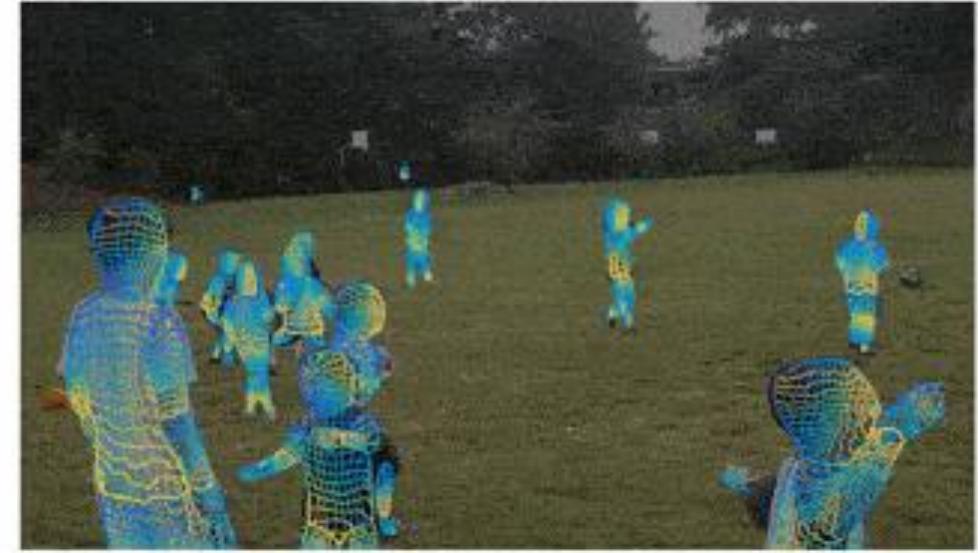
<https://github.com/CMU-Perceptual-Computing-Lab/openpose>

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

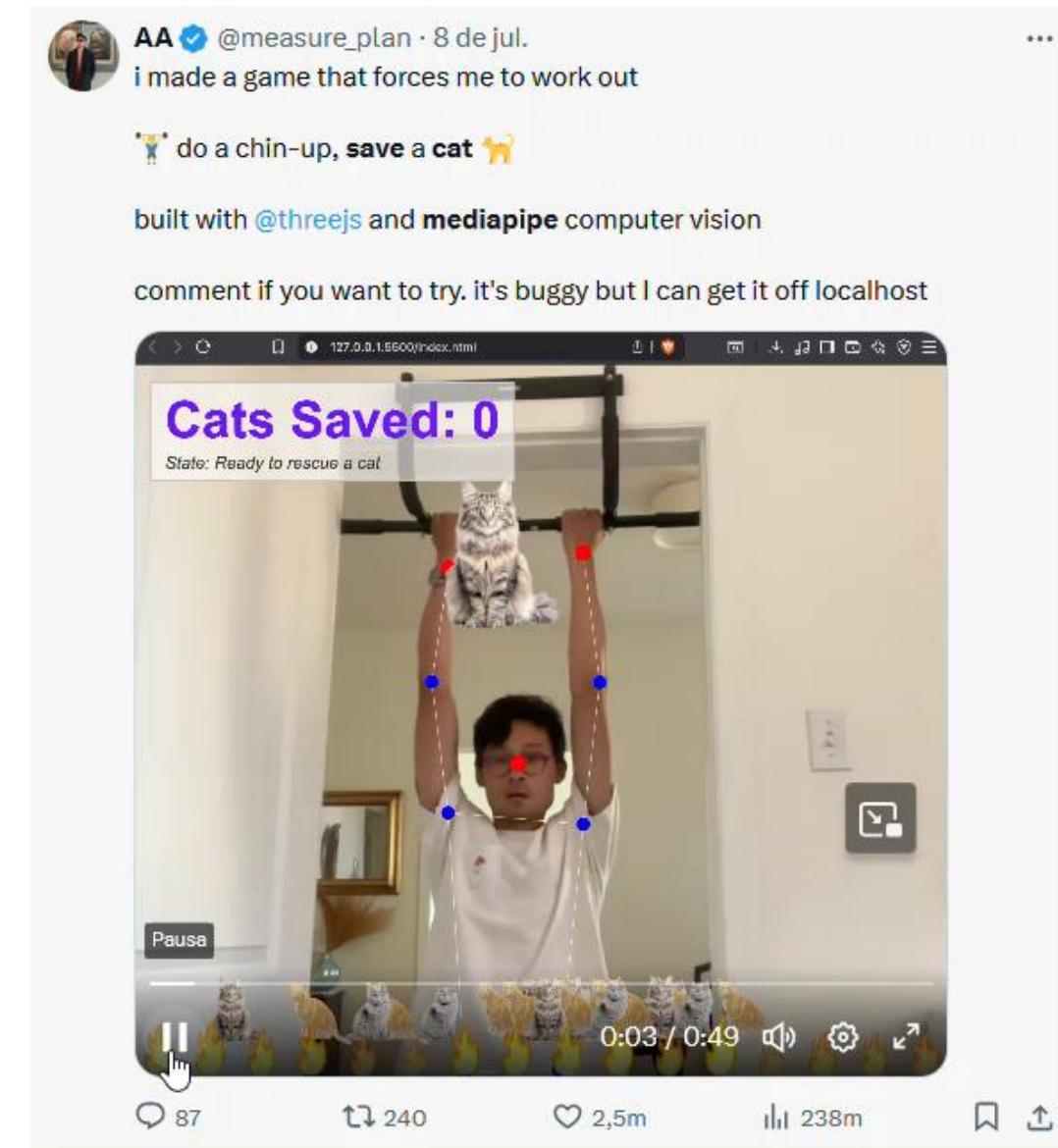
OpenPose, y redes similares nos dicen la posición de las partes del cuerpo

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>