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

AmI4AHA Ambient Intelligence for Active and Healthy Ageing

DEPARTAMENTO DE TECNOLOGÍA INFORMÁTICA Y COMPUTACIÓN

Universitat d'Alacant
Universidad de Alicante

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

Áreas de aplicación

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)

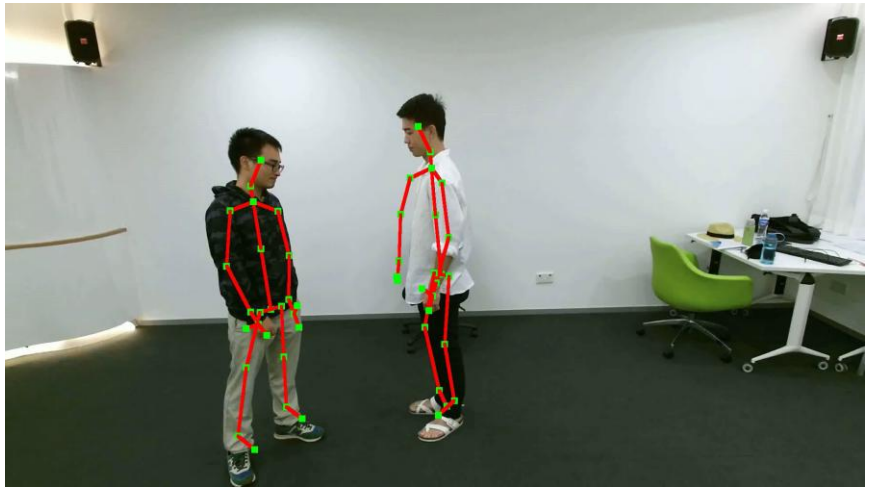
De propósito general – RGB

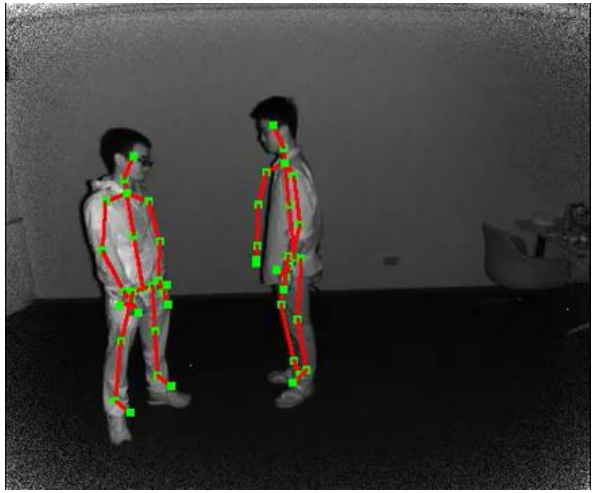
Ejemplos: Kinetics, AVA (Atomic Visual Actions), Charades, ActivityNet

Ejemplos de actividades - datasets

De propósito general – RGB-D

Ejemplo: "NTU RGB+D“ datasets





Ejemplos de actividades - datasets

De propósito específico – RGB-D

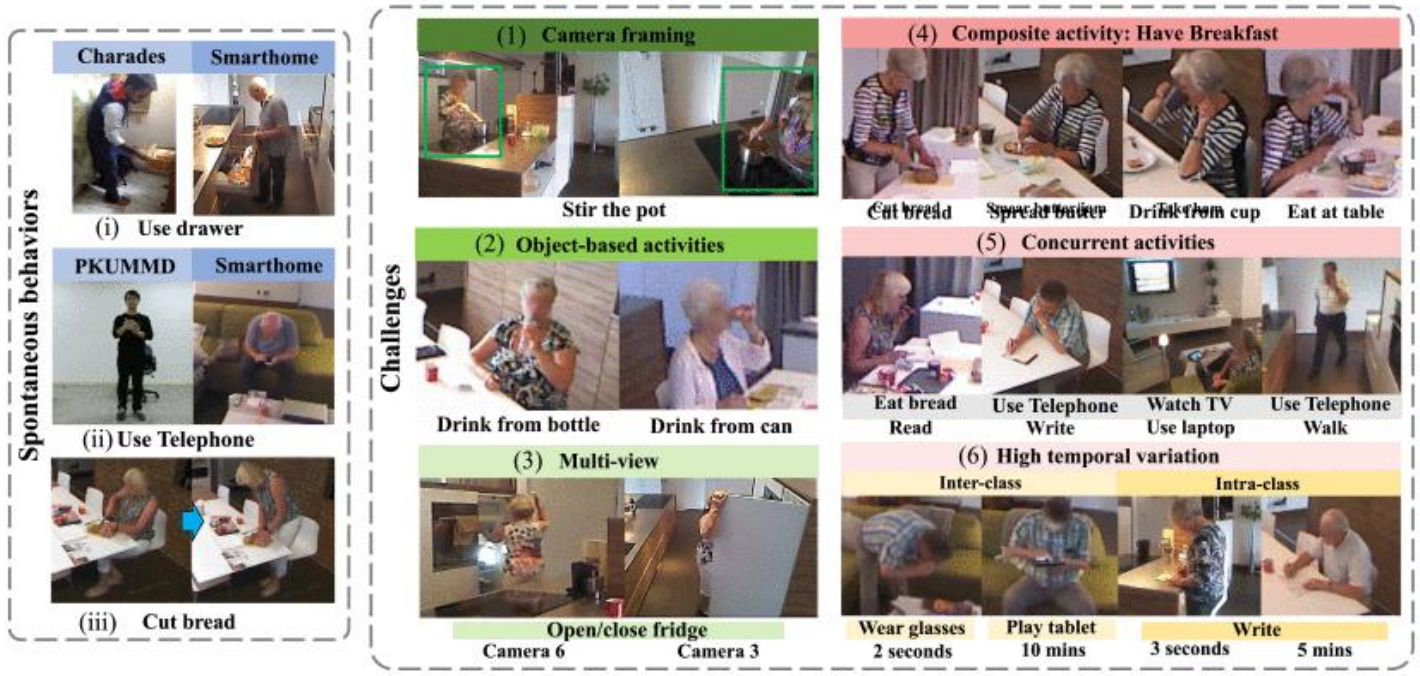
Ejemplo: [Toyota Smarthome]

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


Cook (clean dishes)


Cook (clean up)


Cook (cut)


Cook (stir)


Cook (use stove)


Take pills


Eat at table


Cut bread


Drink from bottle


Drink from can


Drink from cup


Drink from glass


Get up


Lay down


Sit down


Walk


Enter


Leave

Ejemplos de actividades - datasets

De propósito específico – RGB-D

Ejemplo: Toyota Smarthome

¿Cómo se valida?

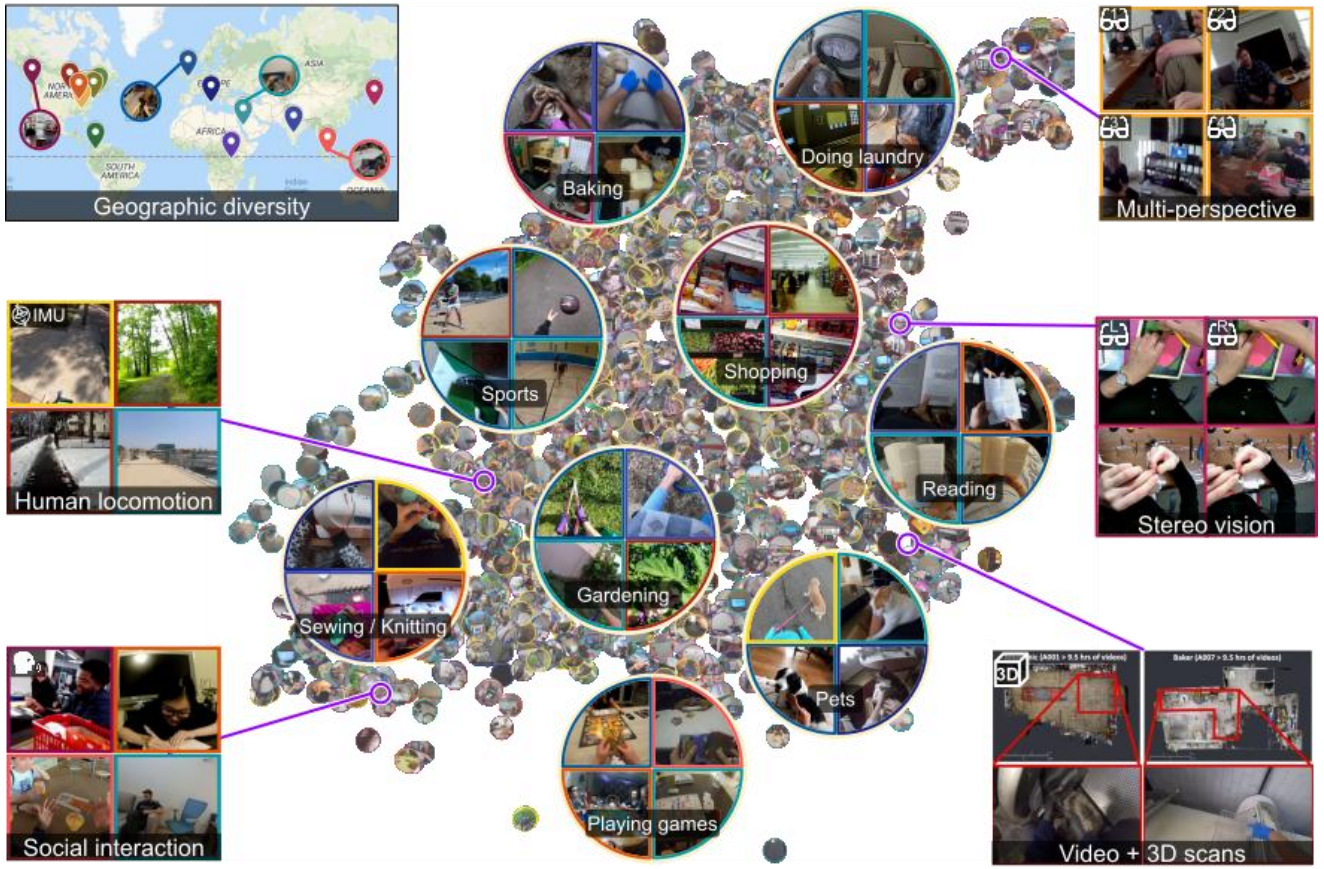
- Protocolo fijo definido por los autores
- Métrica: mean per-class accuracy
- Cross-Subject (CS): 11 sujetos para train, los restantes para validación
- Cross-View (CV)
  - CV1: Train (cámara 1), Validación (cámara 5), Test (cámara 2)
  - CV2: Train (cámaras 1, 3, 4, 6, 7), Validación (cámara 5), Test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

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

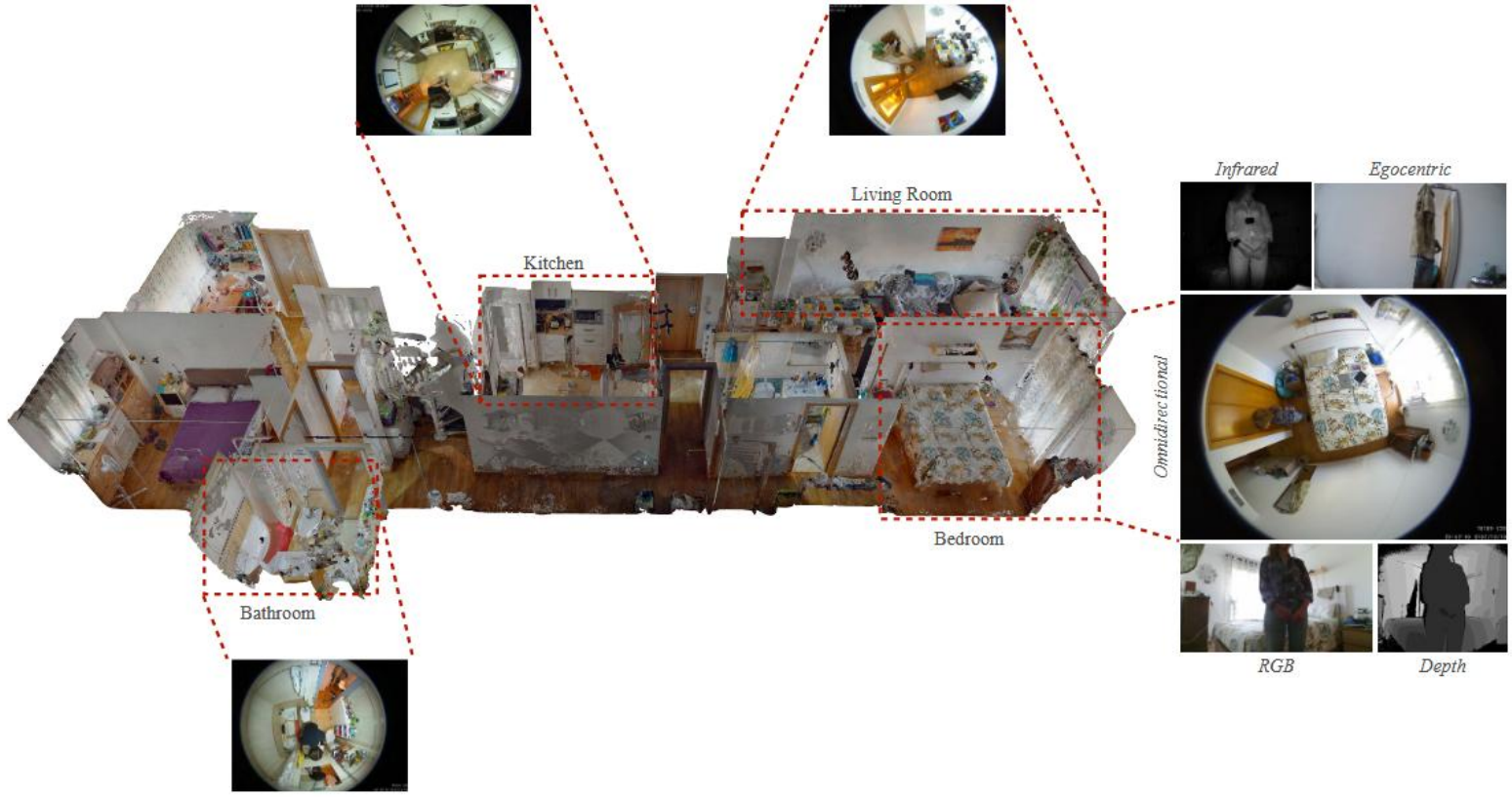
De propósito específico – Múltiples modalidades

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

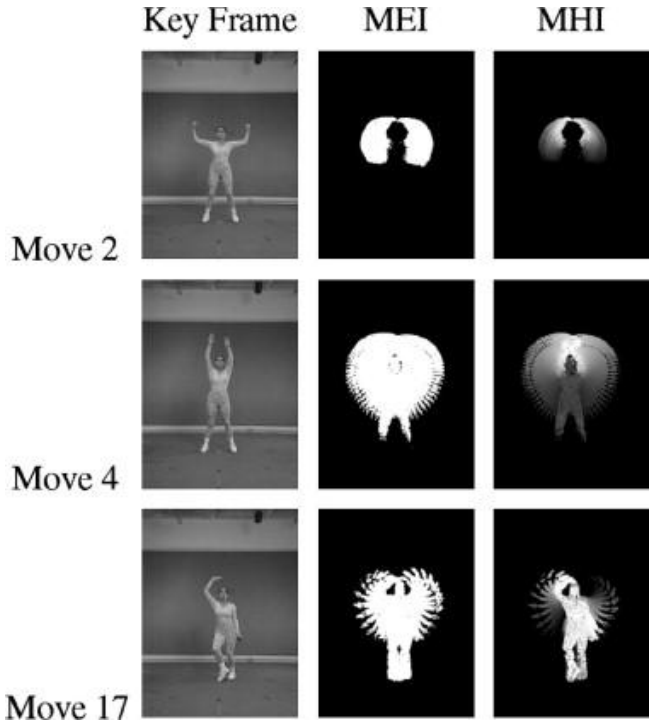


Ejemplos de actividades - datasets

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 distinto, se analizan esas 'huellas' de cada acción

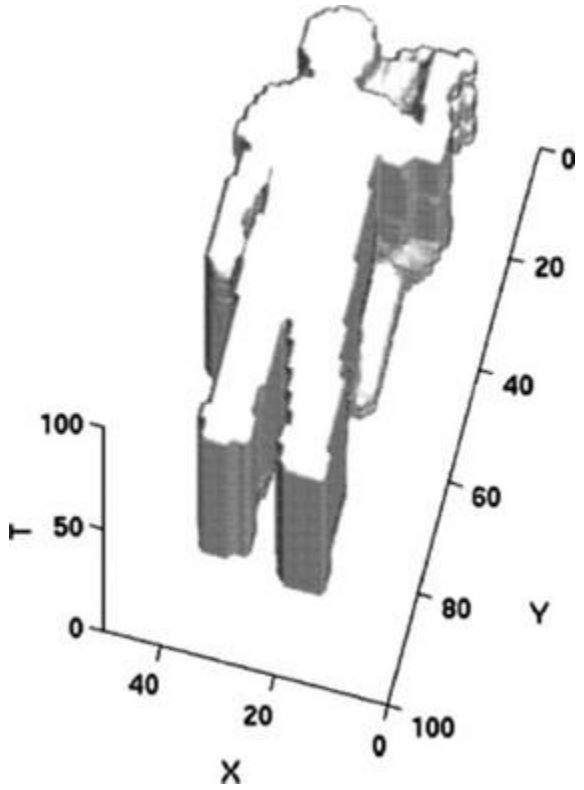


Bobick, A. F., & Davis, J. W. (2002). The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence, 23(3), 257-267.

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.

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

(a) (b) (c)

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.

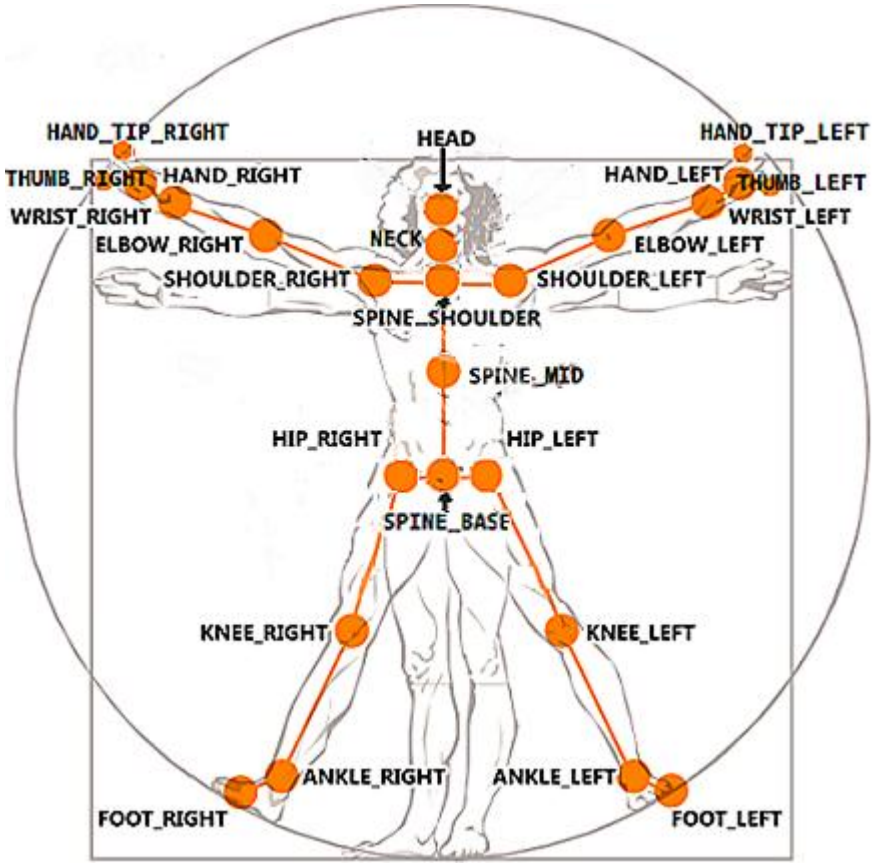
Estimación de pose

Extracción de pose - esqueletos

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



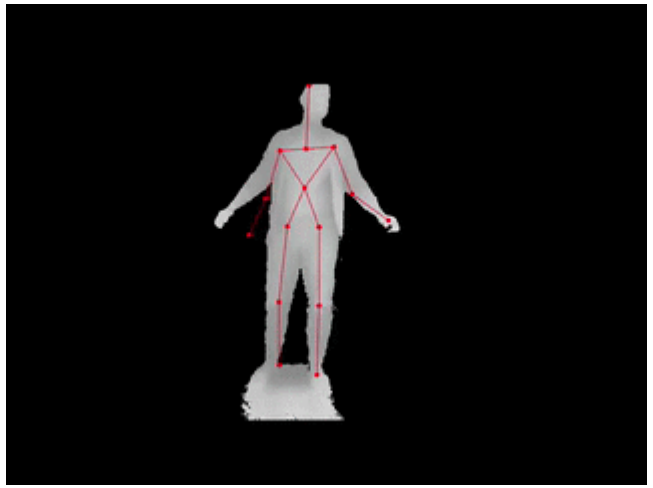
Extracción de pose con RGB-D

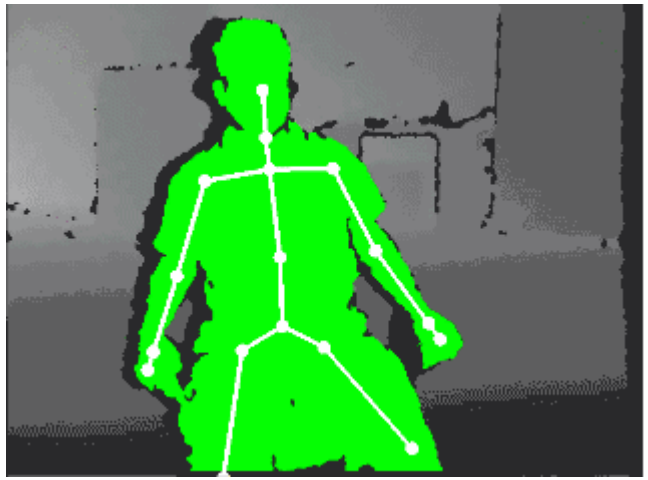
Las cámaras RGB+D fueron una de las primeras opciones para obtener datos de pose 3D fiables

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





Detectores que extraen esqueletos 2D directamente de una imagen

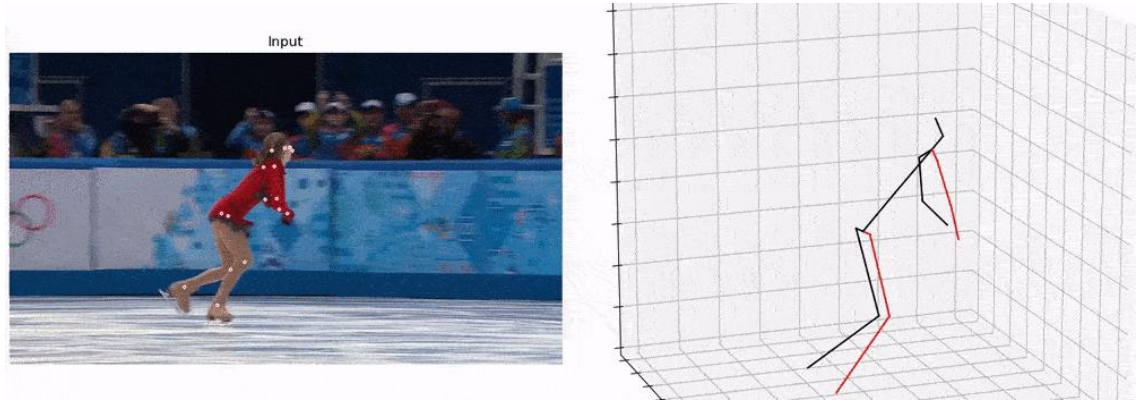
Pero también detectores que extraen poses 3D

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/openpose3d-pose-baseline



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

Extracción de pose con RGB

Extracción de modelos 3D

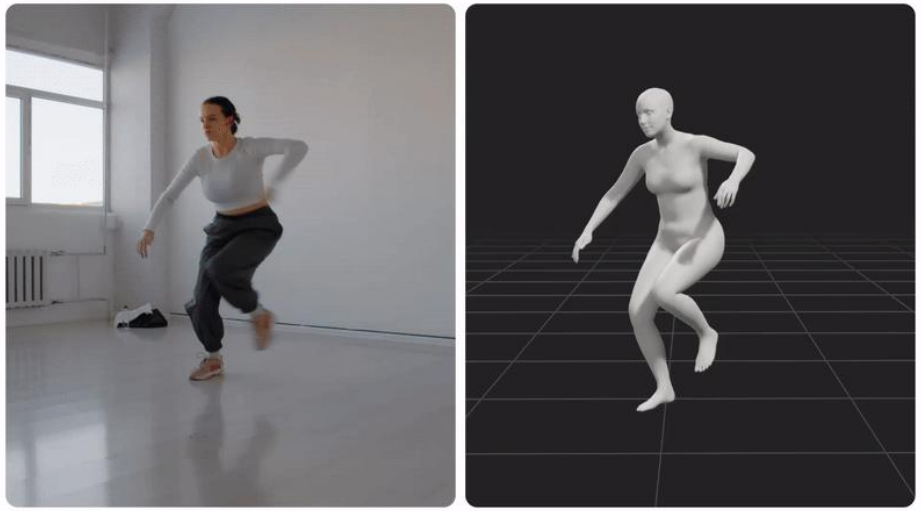
No solo extraen el esqueleto, sino una malla 3D que representa un modelo humano (SMPL – Skinned Multi-Person Linear Model)

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

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.

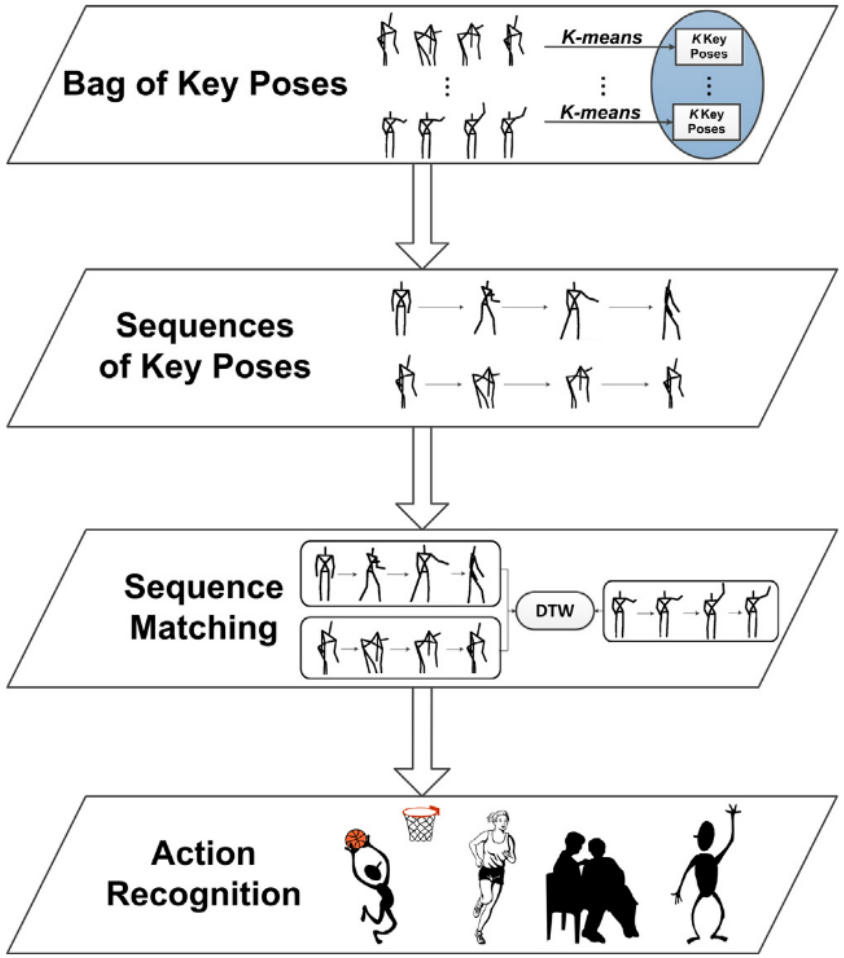
Reconocimiento de actividades mediante esqueletos

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



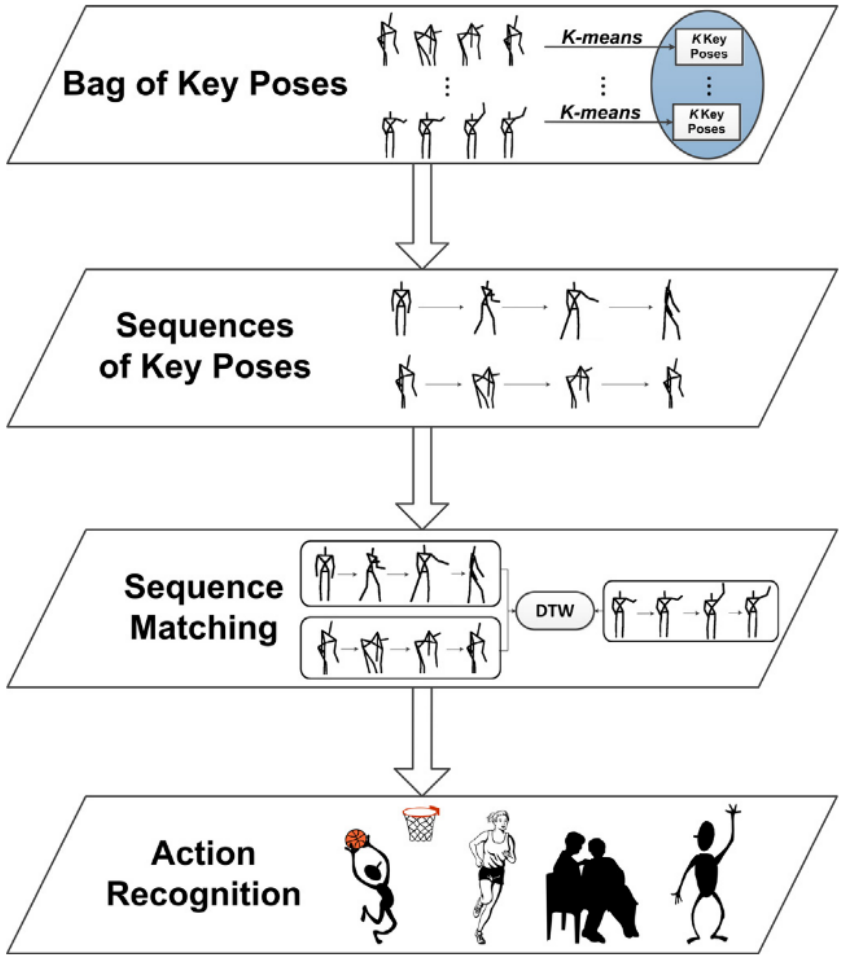
Chaaraoui, A. A., Padilla-López, J. R., Ferrández-Pastor, F. J., Nieto-Hidalgo, M., & Flórez-Revuelta, F. (2014). A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. Sensors, 14(5), 8895-8925.

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.

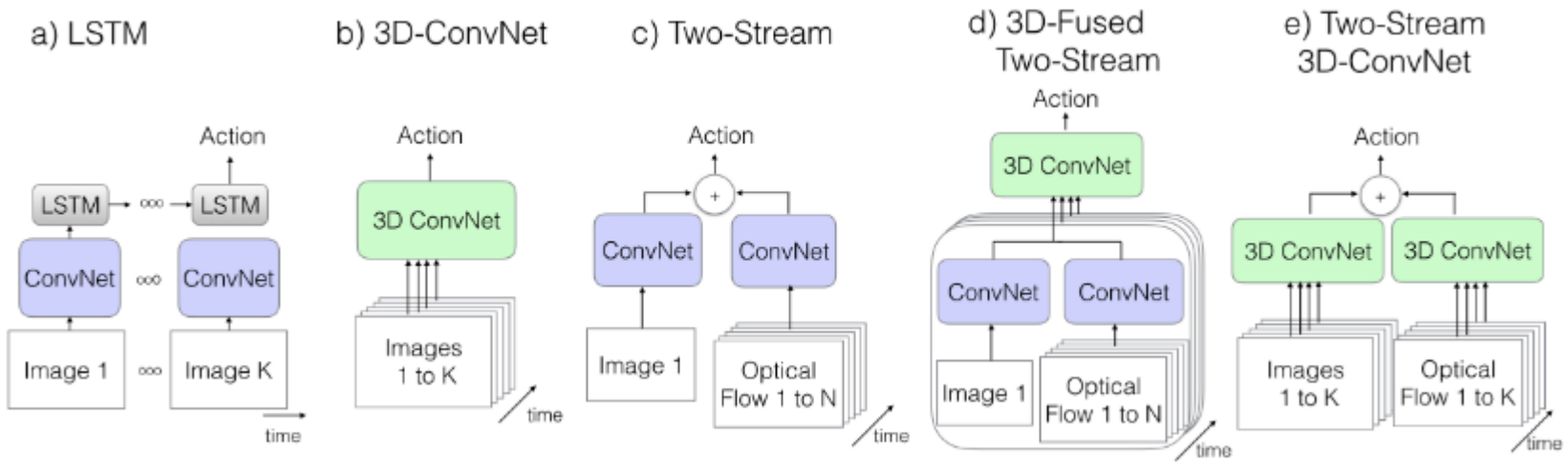


Chaaraoui, A. A., Padilla-López, J. R., Ferrández-Pastor, F. J., Nieto-Hidalgo, M., & Flórez-Revuelta, F. (2014). A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. Sensors, 14(5), 8895-8925.

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).

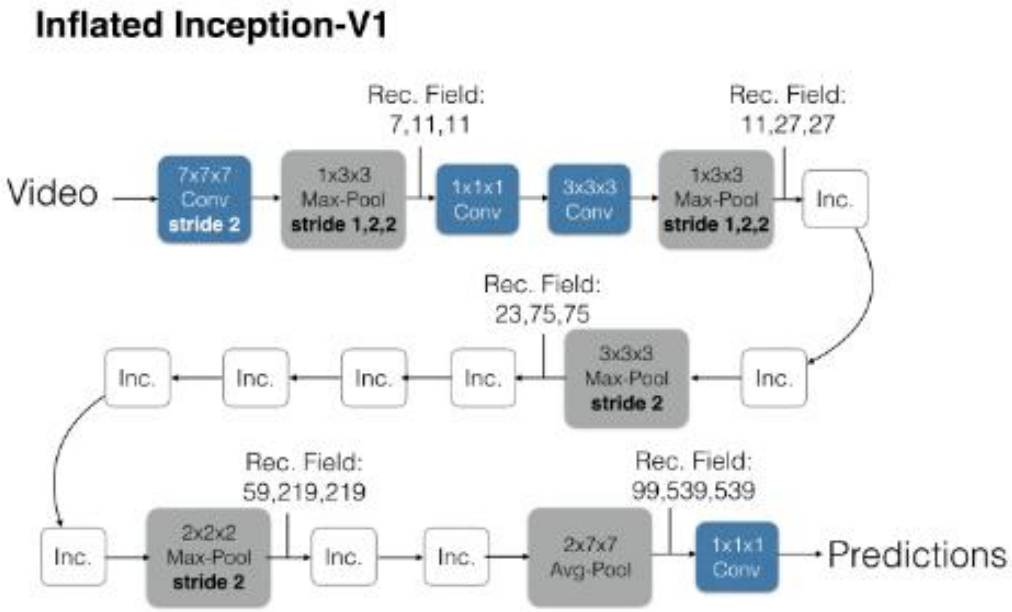
Deep learning – Redes convolucionales 3D

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).

Deep learning – Redes convolucionales 3D

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.

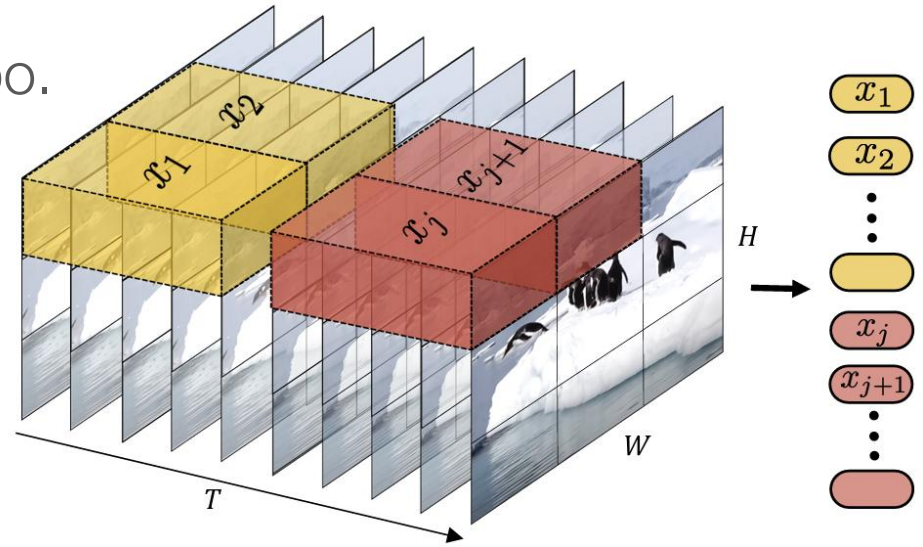
Self-Attention: Permite que el modelo aprenda qué partes del vídeo son más relevantes para una tarea (por ejemplo, reconocer una acción).

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).

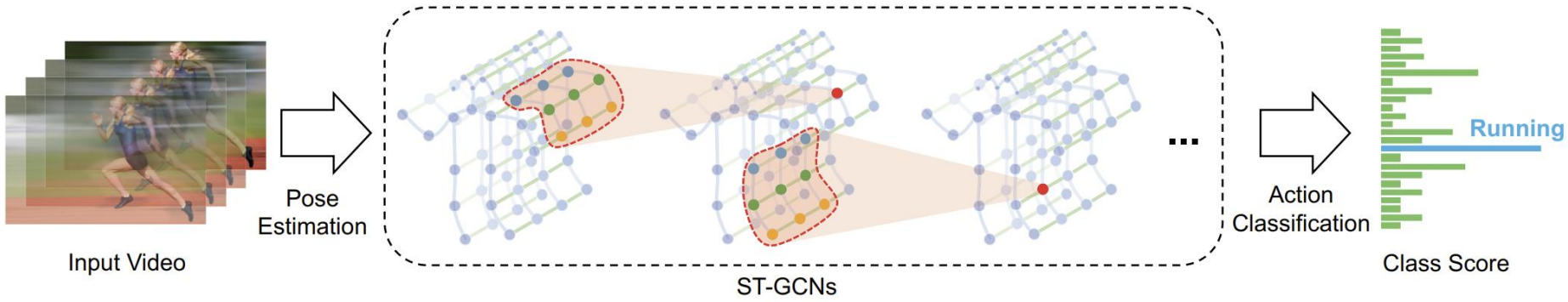
Deep learning – Video Transformers

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



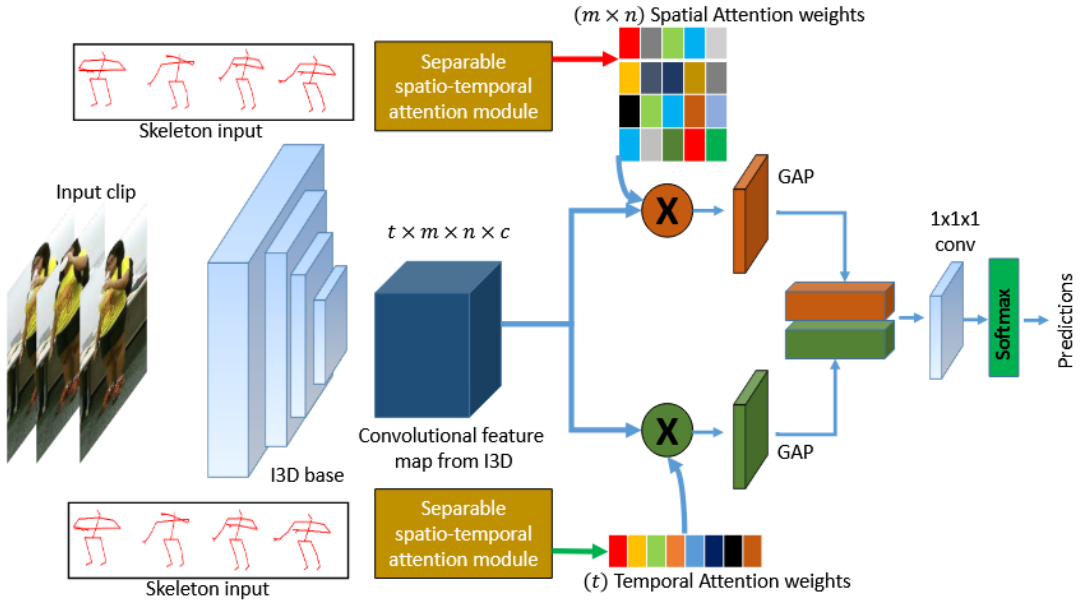
Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

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?

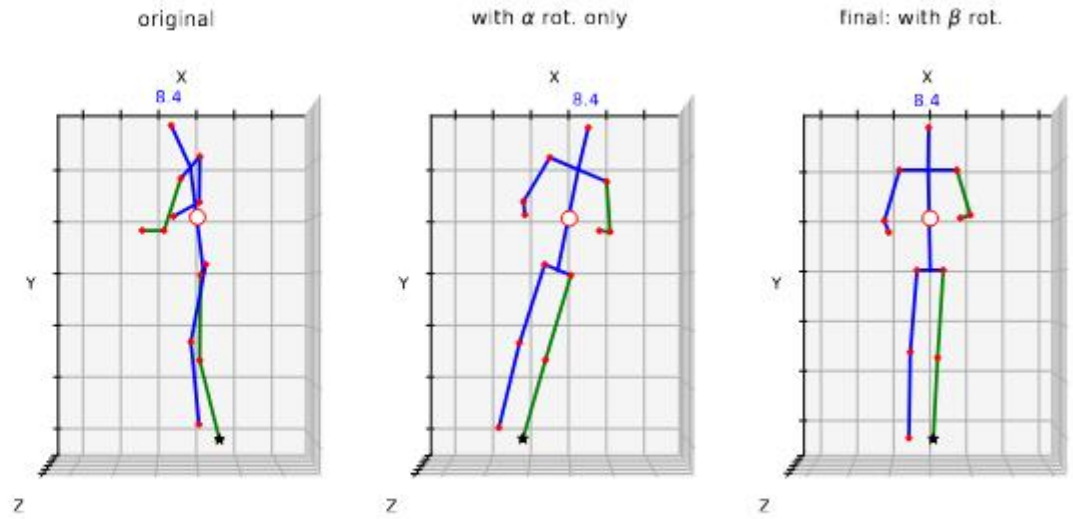


Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., & Francesca, G. (2019). Toyota smarthome: Real-world activities of daily living. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 833-842).

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

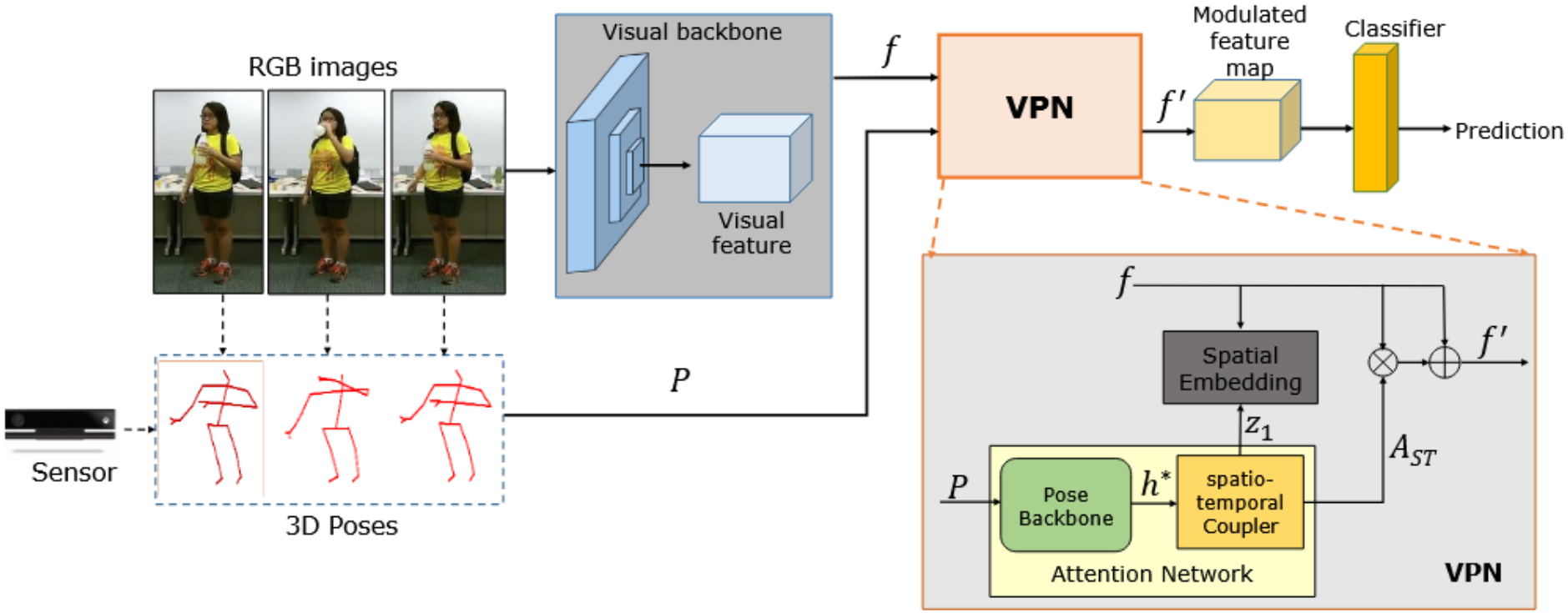


(b) Skeleton plots

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

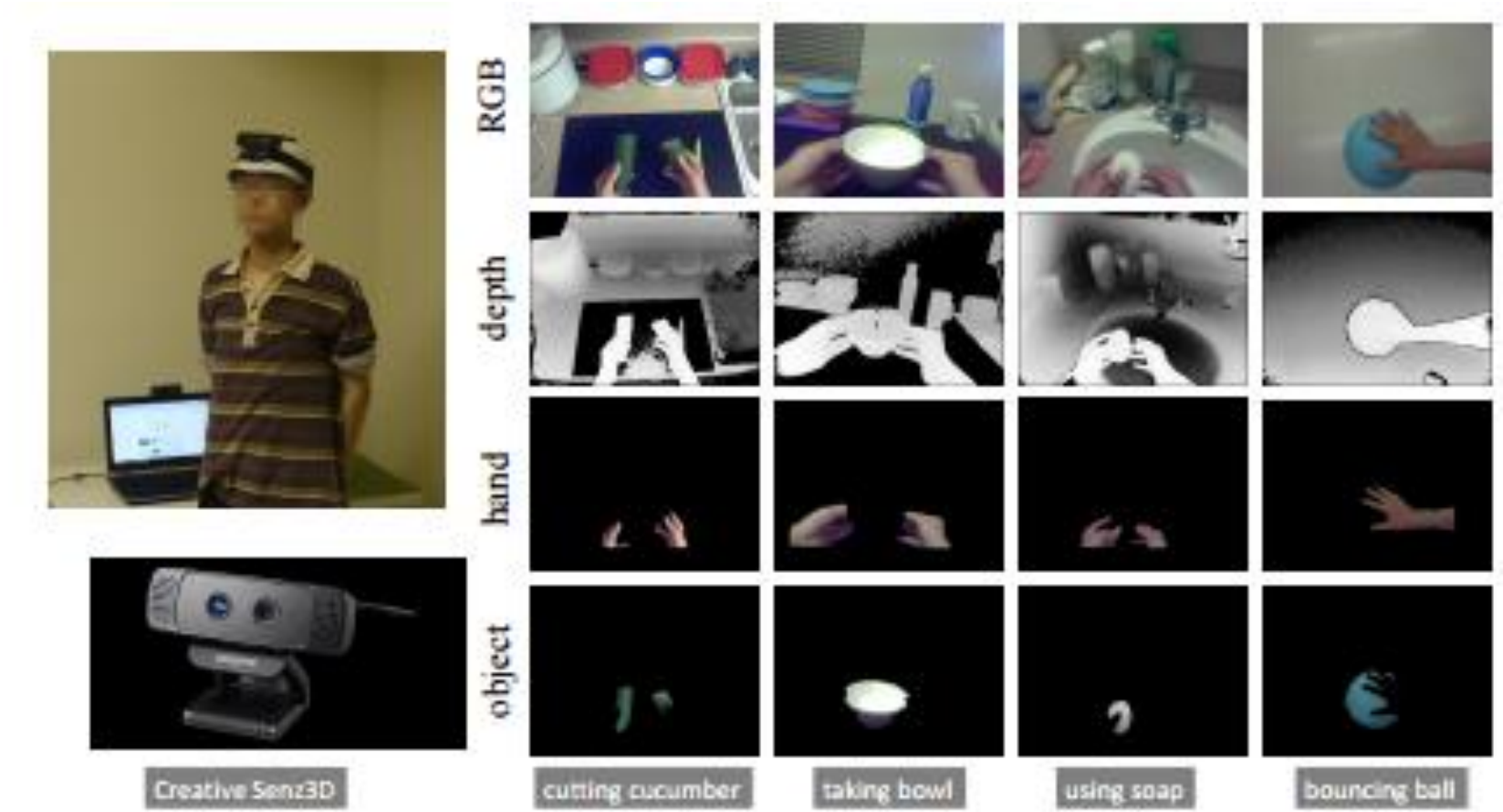
- Por ejemplo “beber de una taza” vs “beber de una botella”

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

Redes híbridas vídeo-pose-objeto

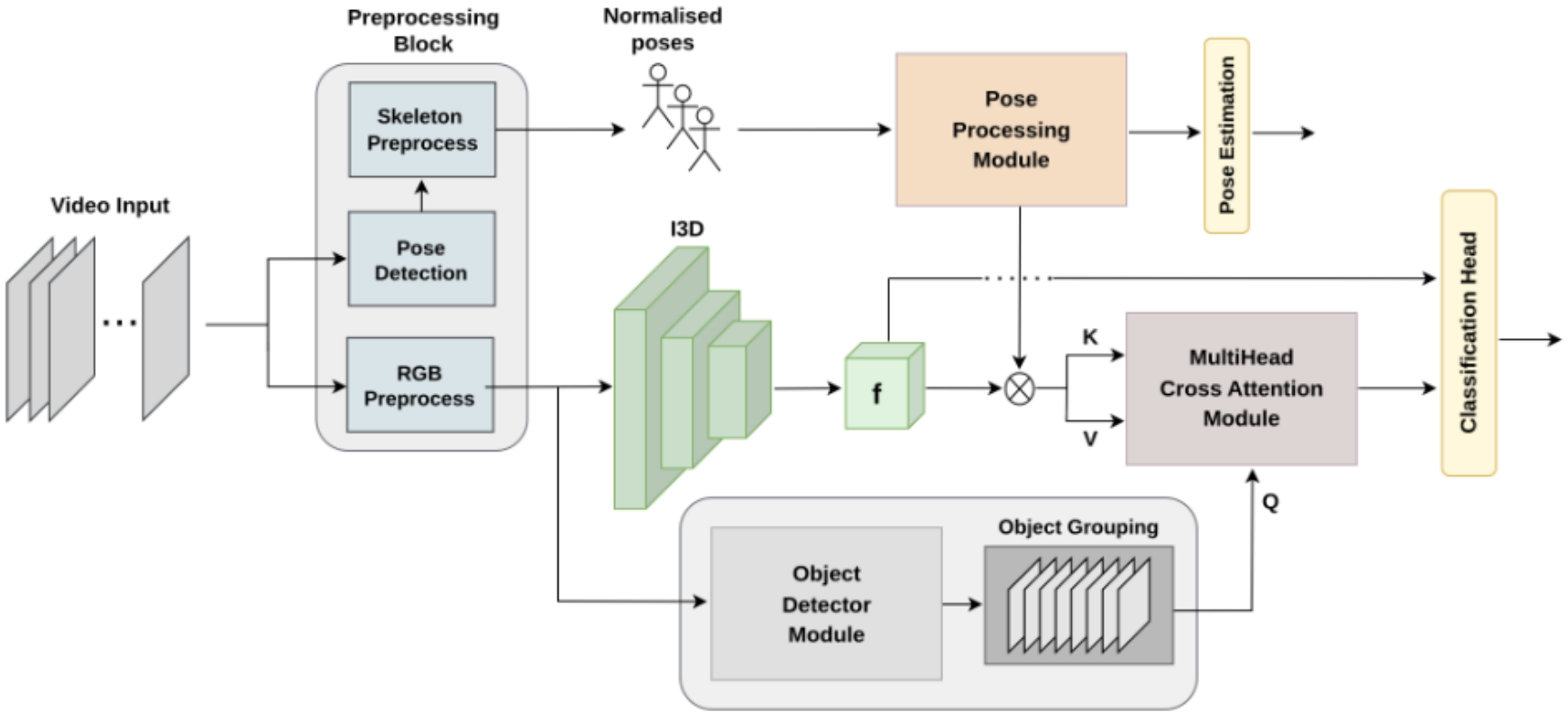
Esto ya se viene haciendo en visión egocéntrica



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

Vistas alternativas

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 oclusiones 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

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.

Visión egocéntrica

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).

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)

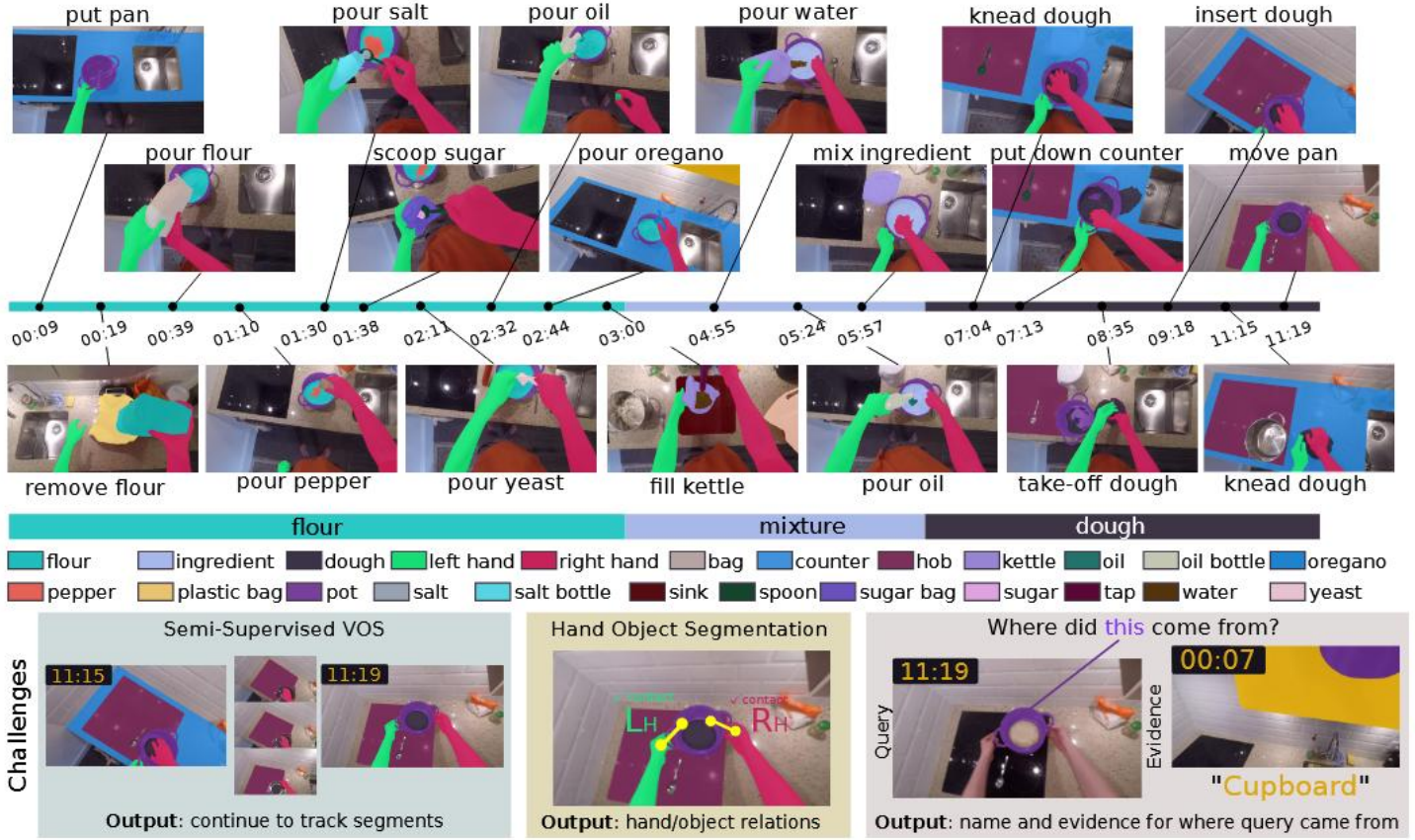
Visión egocéntrica

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

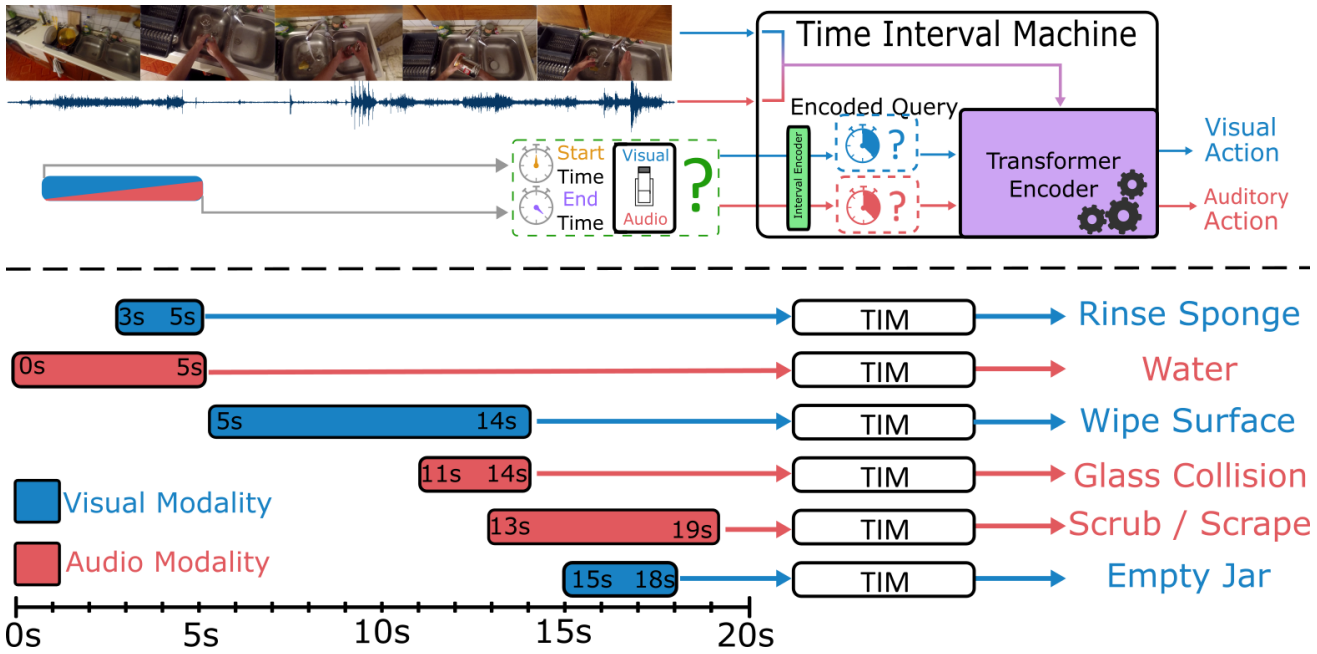


Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., ... & Damen, D. (2022). Epic-kitchens visor benchmark: Video segmentations and object relations. Advances in Neural Information Processing Systems, 35, 13745-13758.

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

AmI4AHA Ambient Intelligence for Active and Healthy Ageing

DEPARTAMENTO DE TECNOLOGÍA INFORMÁTICA Y COMPUTACIÓN

Universitat d'Alacant
Universidad de Alicante