Reconocimiento de acciones egocéntricas usando Visual Transformers

Maximiliano Giusto[1] and María Elena Buemi[1,2]

[1] Departamento de Computación, FCEyN, UBA, Argentina,
maxi.giusto@gmail.com
[2] Instituto de Investigación en Cs. de la Computación (ICC), CONICET-UBA

Abstract. El reconocimiento de acciones es importante para la comprensión de videos, más aún cuando se trata de videos egocéntricos. Existen múltiples aplicaciones para este tipo de reconocimiento, como el monitoreo automático y continuo de actividades de la vida cotidiana, resumen de videos, interpretación de interacciones sociales, entre otros. El desafío de este tipo de videos se encuentra en la velocidad de la cámara, oclusiones y fondos de escena desordenados. Una manera de abordar este problema se centra en explorar la información de la ubicación de las manos y objetos del portador del dispositivo de captura (Gopro, HoloLens). Con el fin de disminuir el costo computacional, un abordaje es utilizar datos 2D. En este estudio se propone el reconocimiento de acciones egocéntricas empleando la pose 2D de manos y objetos para la clasificación de secuencias utilizando un método basado en la combinación de YOLOv8 y Visual Transfomer sobre un subconjunto del dataset H2O.

Keywords: Reconocimiento de acciones egoncéntricas, Visual Transformer, datos 2D, keypoints

1 Introducción

El reconocimiento de acciones en videos egocéntricos es un área en desarrollo que tiene gran cantidad de aplicaciones. La comprensión de videos egocéntricos puede ayudar en control automático y continuo de las actividades de la vida cotidiana, resumen de videos, interacción en reuniones sociales, y siendo más específicos por ejemplo comprender una escena donde hay un objeto olvidado, predicción de posiciones en juegos deportivos, en cuidado de personas. Buena parte de la literatura se centra en explorar la información de la pose 3D de la mano para el reconocimiento de acciones como una manera de tener información realista, pero disponer de este tipo de datos resulta costoso por los equipos de captura y la necesidad de tener ambientes muy controlados [5]. Muchos autores utilizan datos 2D y los convierten a 3D utilizando redes de estimación de la profundidad con un uso intensivo de cálculos o, en otros casos, recurren a sensores de profundidad. Sin embargo, en el caso de convertirlos a 3D, no emplean la información de profundidad, sino que generan su estimación a partir de frames RGB, cuyo

resultado posee errores por el cálculo de la estimación. Mucha etal. [3] utiliza este tipo de datos transformados desde 3D a 2D, los autores informan que la diferencia de precisión, para los casos en los que no se utilizan sensores de profundidad, es de 37 mm entre los datos reales y los transformados. El enfoque 2D baja la complejidad computacional y resulta un buen punto para explorar. En este trabajo presentamos el reconocimiento de acciones egocéntricas basadas en el seguimiento mano-objeto mediante puntos 2D que se usan como input de un Visual Transformers (ViT) para concluir en la acción que se lleva a cabo en el video. Los ViT son una arquitectura de red neuronal basada en Transformers (hoy muy utilizados en NLP), para el procesamiento de imágenes, donde las imágenes son divididas en secuencias de patches que tendrán un tratamiento similar a las secuencias de tokens de los Transformers. Este enfoque exige menos potencia computacional que en las versiones 3D y ofrece una mejora de la privacidad frente al procesamiento de frames RGB completos, que podrían contener datos sensibles. El conjunto de datos H2O [2](del cual utilizamos un pequeña parte), permite extraer poses de manos y objetos que se transforman al espacio 2D [3]. Este conjunto de resulta oportuno de ser utilizado ya que contiene información relevante que permite el reconocimiento de las interacciones, utilizando de dos manos manipulando objetos. El reconocimiento de las acciones egocéntricas se realiza en dos etapas: en la primera se estiman los esqueletos de las manos, el bounding box y la etiqueta del objeto utilizando la arquitectura de red YOLOv8 [4] para pose sobre frames independientes; en la segunda, esta información es enviada a un Visual Transformer [1] para obtener la etiqueta de la acción que se está realizando.
Este artículo describe el conjunto de datos 3D utilizado, la metodología basada en Visual Transformer, los resultados obtenidos y los próximos pasos a seguir como conclusiones del trabajo en curso.

2 Datos de Manos y Objetos: H2O dataset

Basados en que en los videos egocéntricos la ocurrencia de una acción sucede donde se localizan mano-objeto, el conjunto de datos H2O (2 Hands and Objects)[2] es útil debido a que provee los esqueletos de las manos, los bounding-box de los objetos y la etiqueta de la acción realizada en una secuencia de frames. Además, contiene un Ground Truth (GT) de 36 etiquetas de acción que se realizan sobre 8 objetos. Los creadores del dataset proponen un método para crear un conjunto de datos unificado para el reconocimiento en 3D.

Este dataset contiene anotaciones de la pose 3D de dos manos y la pose 6D de los objetos manipulados para cada frame, junto con sus etiquetas de interacción para cada secuencia.

Los datos contenidos en este conjunto proporcionan imágenes RGB-D multivista, además de un gran número de anotaciones que incluyen etiquetas de acción, de objetos, poses 3D de ambas manos (izquierda y derecha), poses 6D de objetos, poses de cámara y nubes de puntos de la escena. En la nube de puntos de la escena se encuentran los keypoints de cada mano y el cubo contenedor del

objeto. Lo relevante de estas dos últimas anotaciones, es que si bien están en 3D, pueden ser convertidas a 2D utilizando los parámetros intrínsecos de la cámara. Las imágenes de H2O se toman en entornos interiores en los que los sujetos interactúan con ocho objetos diferentes utilizando ambas manos. El conjunto de datos incluye 571.645 fotogramas RGB-D y presenta a cuatro participantes realizando 36 clases de acciones distintas en tres entornos diferentes. Se capturaron con cinco cámaras para obtener imágenes sincronizadas RGB y de profundidad. Cuatro de estas cámaras son estáticas y colocadas de manera arbitraria en la escena y la quinta es una cámara egocéntrica montada en el frente de un casco que fue ajustada por los participantes para establecer vistas egocéntricas. Los datos se adquieren en tres escenarios (hall, oficina, cocina). Se grabaron vídeos con una resolución de 1.280 x 720 píxeles tanto para las imágenes RGB como para las de profundidad, con una frecuencia de imagen de 30 fps. Cada video corresponde a una serie de acciones que implican diversas interacciones mano-objeto, de las cuales también se dispone de las etiquetas que identifican dichas acciones.

3 Enfoque YOLO-Transformer

El enfoque presentado en este trabajo consiste en la detección de mano-objeto, a partir de puntos clave y otras informaciones adicionales. Los algoritmos Yolo que se fueron desarrollando desde 2015, se han centrado en devolver características que permiten la localización e identificación de objetos. Para la detección de mano-objeto y los keypoints, Mucha et al. [3] utilizan dos módulos: 1) YOLOv7 para obtener la localización de mano-objeto, y 2) con la información anterior, consiguen los esqueletos de las manos con una arquitectura construida sobre EfficientNetV2.

En nuestro estudio, también utilizamos dos etapas, pero para la primera (detección mano-objeto) tenemos un único módulo, que consiste en tomar el módulo para pose de YOLOv8 [4] que estima la posición de las manos en el espacio 2D usando 21 puntos clave para cada mano, más el bounding box del objeto con su respectiva etiqueta. Luego de dicho procesamiento, cada pose de la mano es representada por $Ph_t^i(x, y)$ donde $t \in \{l, r\}$ representa la mano izquierda o derecha y $i \in [1..21]$ representa los 21 puntos clave de la mano. A su vez, el objeto es representado como $Po_{bb}^i(x, y)$ donde $i \in [1..4]$ corresponde las 4 esquinas del bounding box y Po_l es la etiqueta del objeto. En resumen, como resultado del primer paso del procesamiento tenemos que cada imagen de entrada tiene como salida un frame f_n que concatena todos estos puntos y los podemos representar como:

$$f_n = Ph_l^i(x, y) \oplus Ph_r^i(x, y) \oplus Po_{bb}^i(x, y) \oplus Po_l \tag{1}$$

Por lo tanto, para la secuencia de imágenes de entrada, suponiendo que se compone de m imágenes, se crea el vector de salida V_{seq} que se describe como:

$$V_{seq} = [f_0..f_n], n \in [1..m] \tag{2}$$

Este vector de información V_{seq} va a ser la entrada de la segunda etapa, que es un modelo de secuencias inspirado en Visual Transformers [1] para determinar la acción que se está llevando a cabo. El método se describe esquemáticamente en la Fig. 1.

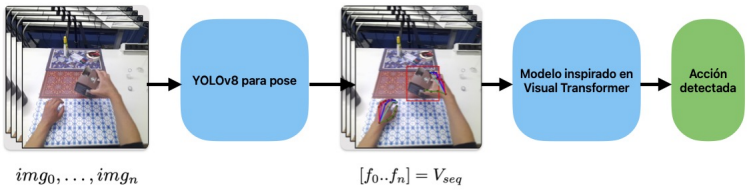

Fig. 1: Esquema que muestra los pasos realizados desde que ingresa un video hasta que se enuncia la acción reconocida

La decisión de utilizar, como alternativa al módulo combinado YOLOv7 + EfficientNetV2, una arquitectura unificada basada exclusivamente en YOLOv8-Pose se basa en tres razones principales. En primer lugar, YOLOv8-Pose combina en una sola red la detección de bounding boxes y la regresión de keypoints, lo que reduce el tiempo de inferencia significativamente, una diferencia que se vuelve importante en tareas de video en tiempo real. Pueden verse los valores en la tabla 1. En segundo lugar, al tratarse de una red única, se simplifica considerablemente la arquitectura del sistema, evitando pasos intermedios como el recorte de regiones y la sincronización entre modelos. Por último, YOLOv8-Pose ofrece resultados competitivos en métricas estándar como mAP, siendo suficiente para la tarea abordada.

Modelo	**Tarea**	**Precisión**	**Tiempo (ms)**
YOLOv7	Bounding Box (COCO)	$mAP_{50\text{-}95}$: 55.9%	5.6/17.9
EfficientNetV2-s	Clasificación (ImageNet)	Top-1: 83.9%	1.58
YOLOv8n-pose	Keypoints (COCO-Pose)	$mAP_{50\text{-}95}$: 50.4% mAP_{50}: 80.1%	1.18
YOLOv8x-pose	Keypoints (COCO-Pose)	$mAP_{50\text{-}95}$: 69.2% mAP_{50}: 90.2%	3.73/6.2

Table 1: Comparación entre modelos para detección de bounding boxes y keypoints

4 Resultados preliminares

Cada una de las dos etapas descriptas en la sección anterior requieren un preprocesamiento de los archivos de imágenes, keypoints (tanto de manos como de objetos) y etiquetas para que puedan ser entrada de las redes en cuestión. Como primera aproximación se pudo entrenar, validar y testear con YOLOv8 para pose con un subconjunto de 448 imágenes extraídas del dataset H2O, que permitió confirmar la posibilidad de procesamiento del input en entrenamiento, validación y testeo de modo correcto. Es importante observar que se trata de un número muy pequeño de imágenes, respecto del total disponible, ya que en este trabajo nos enfocamos en la factibilidad. En la Fig. 2, se muestran dos ejemplos de detección realizada con YOLOv8 sobre dos frames del mismo video y pertenecientes al dataset H2O, éstos corresponden al primer y último frame de la secuencia de acción *place cappuccino*. En ellos se muestran las detecciones: la mano izquierda (identificada como "left hand" o el número 9) con sus correspondientes keypoints; la mano derecha (identificada como "right hand" o el número 10) con sus correspondientes keypoints; el objeto, en este caso cappuccino, identificado con el número 8 o la etiqueta "cappuccino" (obstruida por las otras dos etiquetas). Finalmente, en la Fig. 3 se presentan los gráficos estándar resultantes del entrenamiento de YOLOv8. Estos incluyen las curvas de F1-score, Precisión, Recall, Precisión vs Recall (PR) y la matriz de confusión normalizada. Respecto a los valores de accuracy, los resultados aún no alcanzan lo reportado, se acercan más en validación (82,08% vs 94,26%), pero en test la diferencia es muy grande (25,05% vs 76,03%) por lo que estamos trabajando para achicar esta diferencia sin utilizar H2O, por cuestiones de infraestructura. Cabe aclarar que los resultados mostrados corresponden a una prueba preliminar, por lo que se espera que el rendimiento mejore significativamente con una mayor cantidad de muestras y fine-tuning del modelo.

5 Conclusiones y trabajo futuro

En este trabajo abordamos el problema de la detección mano-objeto utilizando un subconjunto pequeño del dataset H2O y módulos específicos de YOLOv8 para obtener los puntos con los que ViT detecta la acción que transcurre en un video. El entrenamiento se realizó con YOLOv8 para pose utilizando H2O con todas sus imágenes (y sus correspondientes nubes de puntos). La próxima tarea a abordar será achicar la brecha entre el rendimiento obtenido y lo reportado por Mucha et al., por lo que sumaremos un mayor número de imágenes y evaluaremos las precisiones.

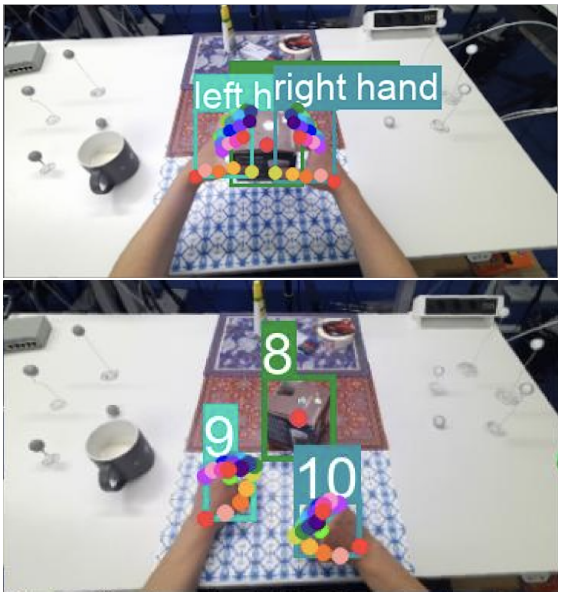

Fig. 2: Ejemplos de muestra de detección de puntos clave del primer frame(arriba), con oclusión, y el último frame(abajo) donde se ven claramente los puntos claves para ser ingresados en el Visual Transformer

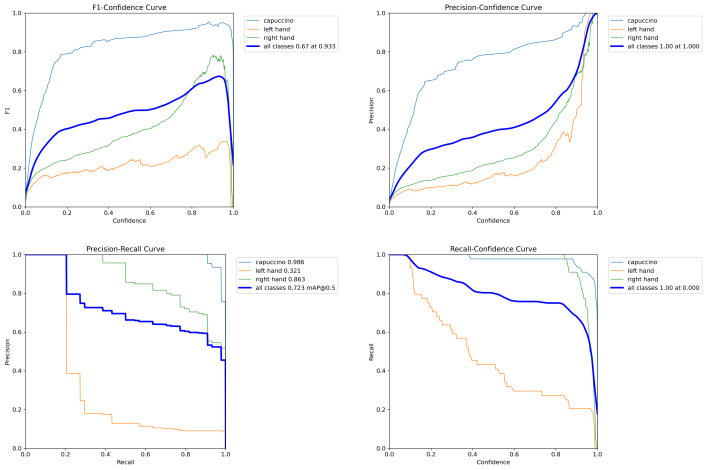

Fig. 3: Gráficos de los resultados obtenidos del entrenamiento de YOLOv8

References

[1] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929 (cit. on pp. 2, 4).

[2] Taein Kwon et al. "H2O: Two Hands Manipulating Objects for First Person Interaction Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 10138–10148 (cit. on p. 2).

[3] Wiktor Mucha and Martin Kampel. *Human Action Recognition in Egocentric Perspective Using 2D Object and Hands Pose*. 2023. arXiv: 2306.05147 [cs.CV]. URL: https://arxiv.org/abs/2306.05147 (cit. on pp. 2, 3).

[4] Joseph Redmon et al. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: 1506.02640 [cs.CV]. URL: https://arxiv.org/abs/1506.02640 (cit. on pp. 2, 3).

[5] Bugra Tekin, Federica Bogo, and Marc Pollefeys. "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on p. 1).