Generación de datos virtuales con objetos de cocina para entrenamiento de redes neuronales

Javier Fañanás-Anaya, Gonzalo López-Nicolás, Carlos Sagüés

Afiliación: Grupo de robótica, percepción y tiempo real (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: javierfa@unizar.es

Resumen

Los algoritmos de visión basados en redes neuronales requieren grandes cantidades de datos etiquetados de forma precisa. Obtener estos datos de forma manual es un proceso muy costoso. En este trabajo se propone utilizar motores gráficos para generar imágenes etiquetadas automáticamente con las que entrenar redes neuronales.

Asistente de cocina

Este trabajo está centrado en el desarrollo de un asistente de cocina que cuenta con una cámara con vista cenital de la vitrocerámica y encimera de la cocina. Este asistente, debe ser capaz de comprender las acciones del usuario y guiarlo en el cocinado de recetas. Para comprender las acciones del usuario, el primer paso es detectar, clasificar y segmentar (Segmentación de instancias) los objetos relacionados con el cocinado que se observan desde la cámara. Uno de los principales problemas de los algoritmos de visión por computador basados en aprendizaje automático es la obtención de una base de datos etiquetada de forma precisa. Generar imágenes y vídeos virtuales es una práctica que ha sido utilizada para el entrenamiento de algoritmos de visión en el problema de la conducción autónoma [1]. En este trabajo se propone adaptar este tipo de técnicas para la segmentación de instancias, tanto a partir de imágenes como vídeos, en el entorno de la cocina.

Entorno virtual

Se ha utilizado el motor gráfico Unity para crear un entorno virtual realista de una cocina. Se han implementado detalles como la iluminación exterior e interior, así como los reflejos de la encimera y vitrocerámica. El problema de visión se ha centrado en la segmentación semántica de 5 clases: Plato, cubierto, vaso, cazuela y sartén. Para ello, se cuenta con un conjunto de objetos que tienen 3 objetos por cada clase.

Con el objetivo de generar imágenes virtuales de forma automática (Fig. 1), se ha implementado un programa en Unity que realiza las siguientes acciones de forma pseudoaleatoria: Modifica la intensidad de la iluminación exterior e interior, cambia los objetos que aparecen en la escena y su posición, y modifica la posición y ángulo de la cámara.

Para obtener un etiquetado automático y preciso de las escenas generadas, se ha trabajado con los *shaders* de Unity. Al aplicar un filtro de un color a cada clase y utilizar un filtro negro para el resto de la escena, se obtiene una máscara precisa para cada uno de los objetos presentes en cada imagen, lo cual resulta ideal para la segmentación de instancias. Utilizando funciones de OpenCV en Python, se realiza el procesamiento de las máscaras para obtener los datos necesarios para el entrenamiento de algoritmos de visión basados en aprendizaje automático. Esto implica la extracción de la lista de píxeles que delimitan el contorno de cada objeto y la identificación de la clase de cada uno según su color. Estos datos se guardan en formato JSON.

Entrenamiento de la red

Con las imágenes virtuales y la información almacenada en formato JSON se puede entrenar algoritmos de visión basados en aprendizaje automático. En este trabajo se ha decidido trabajar con la red neuronal convolucional Mask R-CNN [3], especializada en segmentación de instancias. Se ha utilizado el algoritmo de entrenamiento de 3 fases propuesto en el artículo de Mask R-CNN, partiendo de los pesos de la red ya entrenados con el dataset de COCO, es decir, haciendo *Transfer Learning*. Para el entrenamiento de la red neuronal se generó un conjunto de 5500 imágenes de

resolución 1024 x 1024 píxeles, reservando 4000 imágenes para entrenamiento, 1000 imágenes para validación y 500 imágenes para test. El entrenamiento se ha realizado en una GPU NVIDIA Titan XP, que ha permitido trabajar con un tamaño de lote, o *Batch Size,* de 2 imágenes.

Validación experimental

El modelo se ha validado a partir de las 500 imágenes de test, obteniendo la matriz de confusión de la Tabla 1 y una precisión del 96% y un *recall* del 97%. Para evaluar de forma cualitativa cómo funciona este modelo en entornos reales se han utilizado imágenes de distintas distribuciones (Fig. 2), incluyendo vídeos de EPIC-KITCHENS [3].

Plato	345	0	1	1	0	2
Sartén	1	364	2	1	0	2
Cazuela	0	0	366	1	0	9
Vaso	0	0	0	213	0	19
Cubierto	0	4	0	0	325	18
FN	6	3	11	1	17	0
	Plato	**Sartén**	**Cazuela**	**Vaso**	**Cubierto**	**FP**

Tabla 1. Matriz de confusión con las 500 imágenes de test. Filas: predicciones del modelo. Columnas: objetos reales.

Los resultados positivos de los experimentos realizados demuestran el potencial que tiene el uso de motores gráficos para generar datos etiquetados y entrenar redes neuronales a partir de estos en el problema del asistente de cocina. Como trabajo futuro se propone obtener métricas a partir de un conjunto de datos de imágenes reales etiquetadas. Esto permitirá tener una evaluación cuantitativa en este contexto, y comparar las redes entrenadas a partir de datos virtuales y las entrenadas con datos reales. También, sería interesante evaluar el rendimiento de un modelo entrenado a partir de un conjunto de datos híbridos (Virtuales y reales).

Referencias

[1] GAIDON, A., et al., 2016. VirtualWorlds as Proxy for Multi-object Tracking Analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4340-4349. DOI 10.1109/CVPR.2016.470.

[2] HE, K., et al., 2017. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*. pp. 2980-2988. DOI 10.1109/ICCV.2017.322.

[3] DAMEN, D., et al., 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, vol. 130, no. 1, ISSN 1573-1405. DOI 10.1007/s11263-021-01531-2.

Agradecimientos

Trabajo financiado por el Gobierno de Aragón, grupo T45_23R, y por proyectos CPP2021-008938, PID2021-124137OB-I00 y TED2021-130224B-I00, financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea-NextGenerationEU/PRTR.

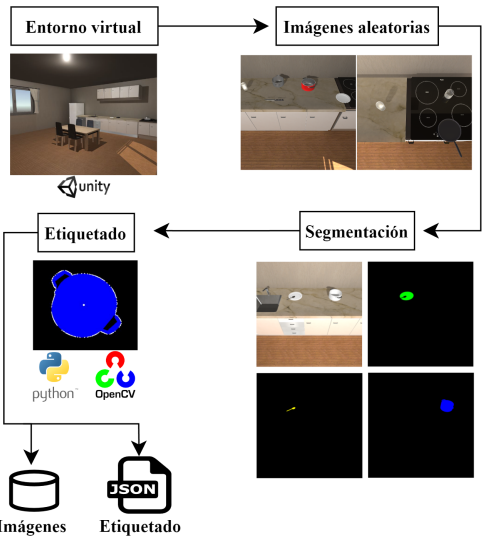

Figura 1. Resumen del proceso de generación de datos virtuales etiquetados de forma automática

Figura 2. Ejemplos de segmentación semántica en distintas distribuciones de imágenes.