vinetur

Una nueva inteligencia artificial separa troncos y sarmientos de la vid para la poda robotizada

Logra un 85,12% en sarmientos finos, con 14,99 puntos más que el mejor sistema previo

sábado 14 de noviembre de 2026



Un equipo de ingeniería agrícola ha desarrollado un sistema de inteligencia artificial capaz de separar en imágenes de viñedo los troncos y los sarmientos de la vid durante el periodo de reposo invernal, un paso previo para la poda robotizada. El trabajo,

publicado en la revista Smart Agricultural Technology, presenta una red de aprendizaje profundo llamada DTMF y la sitúa como una herramienta de visión artificial para una de las tareas más intensivas en mano de obra dentro del viñedo.

La poda de invierno es una práctica básica en viticultura porque regula el crecimiento vegetativo, la producción de uva, el vigor de la planta y la calidad final del fruto. En la mayoría de las explotaciones se realiza a mano y requiere personal con experiencia. Eso eleva el coste de la operación, introduce diferencias entre unas podas y otras y limita la eficiencia cuando la superficie es grande.

Los autores parten de un problema técnico conocido: identificar ramas finas y de bajo contraste en escenas reales del viñedo. En esas imágenes aparecen al mismo tiempo suelo, alambres de la espaldera, vegetación y fondos irregulares. Esa mezcla complica que un sistema automático separe con precisión qué parte de la imagen corresponde a la planta y qué parte pertenece al fondo.

Para resolverlo, el equipo diseñó una arquitectura de doble rama que combina varias técnicas. Por un lado, emplea un Vision Transformer para captar relaciones espaciales en la escena completa. Por otro, usa transformadas wavelet discretas y bloques Mamba, un tipo de modelo orientado al tratamiento de secuencias, para procesar información de distintas escalas y frecuencias. La idea es que una rama recoja la estructura general de la vid y la otra afine los bordes y contornos de los sarmientos, que son mucho más finos.

El modelo distribuye el trabajo en dos subredes, llamadas U1 y U2. La primera divide la imagen en fragmentos y analiza relaciones amplias entre zonas de la escena. La segunda separa componentes de baja y alta frecuencia para tratar de forma distinta las superficies más lisas del tronco y las texturas o líneas delgadas de las ramas. Después, ambas salidas se combinan en la fase final para producir una máscara semántica que distingue fondo, tronco y sarmiento.

Los investigadores también modificaron la forma de entrenar el sistema. Durante las cuatro primeras épocas usaron una función de pérdida compuesta por entropía cruzada binaria, una pérdida Dice y una supervisión híbrida entre las dos ramas del modelo. A partir de la quinta época sustituyeron la entropía cruzada por una pérdida basada en

superpíxeles, agrupaciones locales de color y textura que ayudan a que la clasificación se ajuste mejor a los límites físicos de los sarmientos sin perder estabilidad en la segmentación general.

La evaluación se hizo con un conjunto experimental de 2.150 fotografías RGB de alta resolución tomadas en invierno en dos zonas vitícolas de Xinjiang, en China. El equipo recopiló 1.250 imágenes en noviembre de 2024 en un viñedo experimental de la Universidad de Shihezi y otras 900 en diciembre de 2024 en una explotación comercial de Turpan. Todas se captaron con luz exterior natural y con una cámara digital Sony equipada con un objetivo zoom de 18 a 150 milímetros, desde ángulos y distancias variables para recoger arquitecturas de vid distintas, interferencias de la espaldera y condiciones de suelo diversas.

El conjunto se dividió en 1.505 imágenes para entrenamiento, 215 para validación y 430 para prueba final. Con recortes aleatorios, traslaciones horizontales y verticales, inyección de ruido gaussiano y rotaciones de hasta 30 grados, el volumen útil de entrenamiento aumentó hasta 4.515 imágenes. Todas las fotos fueron anotadas de forma manual a nivel de píxel con el programa Labelme. Las 430 imágenes de prueba se reservaron solo para la evaluación final y no participaron ni en el entrenamiento ni en la selección de puntos de control.

Según los resultados del estudio, la red DTMF alcanzó una precisión media por píxel del 94,24% y una intersección sobre unión media del 88,25%. La comparación con otros modelos de segmentación semántica fue amplia. DTMF superó a DeepLabV3, que obtuvo un 69,65% de mIoU; a PSPNet, con un 60,53%; a FCN, con un 72,20%; a Segformer, con un 68,12%; a U-Net, con un 72,03%; y a una versión de U-Net con base Transformer, con un 73,12%.

La diferencia fue menor en la detección del tronco, donde varios modelos ofrecieron cifras cercanas. DTMF logró un 87,68% de IoU en esa categoría, frente al 87,46% del U-Net con Transformer. Donde sí apareció una distancia amplia fue en la segmentación de los sarmientos. El nuevo sistema llegó al 85,12% de IoU en ramas, 14,99 puntos por encima del mejor resultado previo en esa clase, que correspondió a FCN con un 70,13%.

El equipo también realizó pruebas para medir cuánto aporta cada parte de la arquitectura. Cuando se aisló solo el codificador Transformer con su decodificador, el sistema se quedó en un 81,11% de precisión media por píxel y un 71,98% de mIoU. El uso aislado del codificador con wavelet y Mamba elevó esas cifras a un 81,74% y un 74,00%. La unión de ambos codificadores en un único decodificador llevó el resultado al 85,73% y al 80,65%. La configuración completa, con doble codificador y doble decodificador, alcanzó los mejores valores, con un 94,24% y un 88,25%.

Las pruebas internas de los módulos de fusión siguieron la misma línea. Añadir bloques Mamba a una base wavelet elevó el mIoU del 80,32% al 83,65%, y sumar un mecanismo de autoatención para alta frecuencia lo llevó al 86,65%. En el entrenamiento, la estrategia en dos fases con supervisión híbrida y restricciones por superpíxeles mejoró la IoU de ramas del 74,33% al 85,12% frente a un entrenamiento básico con Dice y entropía cruzada.

Los autores afirman que las inspecciones visuales muestran máscaras más continuas sobre los sarmientos principales y laterales, incluso cuando se cruzan con alambres o se confunden con texturas irregulares del suelo. También hicieron ensayos cualitativos con imágenes externas de otras plantas leñosas. Aun así, precisan que esas pruebas solo tienen valor ilustrativo y no equivalen a una validación medida de la capacidad del sistema para funcionar bien fuera del dominio en el que fue entrenado.

El trabajo no presenta una máquina que pode de forma autónoma una viña completa. Los investigadores dejan claro que la segmentación semántica es solo una parte del problema. Para llegar a un robot operativo harían falta, además, sistemas de esqueletización en tres dimensiones, algoritmos que localicen el punto exacto de corte, planificación geométrica de trayectorias y un efector físico capaz de ejecutar la poda en campo.

También enumeran varias limitaciones del estudio. La arquitectura paralela, con doble codificador y doble decodificador, aumenta el número de parámetros y la complejidad computacional, lo que puede introducir problemas de latencia en plataformas embarcadas con recursos limitados. Además, el conjunto de datos procede solo de dos regiones de Xinjiang y solo del invierno, de modo que no se ha verificado su

comportamiento con follaje de verano ni con otras configuraciones de espaldera. A eso se suma que la división del conjunto se hizo por imagen y no por planta, y que no se repitieron entrenamientos con distintas semillas para medir la variación entre ejecuciones.

La siguiente etapa, según los investigadores, pasa por reducir la carga computacional sin perder precisión. Para ello plantean varias vías: poda estructurada de parámetros, cuantización de la red y destilación de conocimiento para entrenar modelos más ligeros. También prevén incorporar pérdidas sensibles a la topología de la planta, mapas de profundidad estereoscópicos, secuencias de vídeo y técnicas de adaptación durante la inferencia para depender menos del etiquetado manual intensivo.

El avance tiene interés para el sector de las bebidas porque ataca uno de los cuellos de botella de la viticultura: una labor estacional, especializada y cara en el viñedo. Si este tipo de visión artificial logra integrarse más adelante en robots de campo y mantiene su precisión fuera del laboratorio, podría ayudar a reducir la dependencia de mano de obra en la poda y a hacer más uniforme el manejo de la vid. Ese efecto potencial sería relevante para productores de uva y para la cadena del vino, aunque el estudio se limita por ahora a la identificación de troncos y sarmientos en imagen y no a la ejecución completa de la poda.