X Congreso Ibérico de Agroingeniería
X Congresso Ibérico de Agroengenharia
Huesca, 3-6 septiembre 2019

Uso de redes neuronales convolucionales para la detección remota de frutos con cámaras RGB-D

J. Gené-Mola[1], V. Vilaplana[2], J.R. Rosell-Polo[1], J.R. Morros[2], J. Ruiz-Hidalgo[2], E. Gregorio[1]

[1] Grupo de Investigación en AgróTICa y Agricultura de Precisión, Departamento de Ingeniería Agroforestal, Universitat de Lleida (UdL) – Agrotecnio Center, Lleida, España; j.gene@eagrof.udl.cat
[2] Departamento de Teoría de la Señal y Comunicaciones, Universitat Politècnica de Catalunya, Barcelona, España.

Resumen: La detección remota de frutos será una herramienta indispensable para la gestión agronómica optimizada y sostenible de las plantaciones frutícolas del futuro, con aplicaciones en previsión de cosecha, robotización de la recolección y elaboración de mapas de producción. Este trabajo propone el uso de cámaras de profundidad RGB-D para la detección y la posterior localización 3D de los frutos. El material utilizado para la adquisición de datos consiste en una plataforma terrestre autopropulsada equipada con dos sensores Kinect v2 de Microsoft y un sistema de posicionamiento RTK-GNSS. Con este equipo se escanearon 3 filas de manzanos Fuji de una explotación comercial. El conjunto de datos adquiridos está compuesto por 110 capturas que contienen un total de 12,838 manzanas Fuji. La detección de frutos se realizó mediante los datos RGB (imágenes de color proporcionadas por el sensor). Para ello, se implementó y se entrenó la red neuronal convolucional de detección de objetos Faster R-CNN, la cual está compuesta por dos módulos: red de propuesta de regiones de interés y red de clasificación. Ambos módulos comparten las primeras capas convolucionales siguiendo el modelo VGG-16 pre-entrenado con la base de datos ImageNet. Los resultados de test muestran un porcentaje de detección del 91.4% de los frutos con un 15.9% de falsos positivos (F1-score = 0.876). La evaluación cualitativa de las detecciones muestra que los falsos positivos corresponden a zonas de la imagen que presentan un patrón muy similar a una manzana, donde, incluso a percepción del ojo humano, es difícil de determinar si hay o no manzana. Por otro lado, las manzanas no detectadas corresponden a aquellas que estaban ocultas casi en su totalidad por otros órganos vegetativos (hojas o ramas) o a manzanas cortadas por los márgenes de la imagen. De los resultados experimentales se concluye que el sensor Kinect v2 tiene un gran potencial para la detección y localización 3D de frutos. La principal limitación del sistema es que el rendimiento del sensor de profundidad se ve afectado en condiciones de alta iluminación.

Palabras clave: Cámaras de profundidad, RGB-D, Detección de frutos, Redes neuronales convolucionales, Robótica agrícola

1. Introducción

Para garantizar las necesidades alimentarias en una población mundial creciente será imprescindible incrementar la producción actual de frutos y vegetales [1]. Esto supone un reto para las comunidades agrícolas, especialmente en un contexto de mano de obra escasa y elevados costes de los insumos agrícolas. La aplicación de nuevas tecnologías al campo puede ser de gran ayuda para el manejo eficiente y sostenible de la producción frutícola, con aplicación a distintos procesos como la poda, la fertilización, el aclareo y la recolección [2], [3].

doi: 10.26754/c_agroing.2019.com.3325

Los últimos avances en los campos de la informática, robótica y visión artificial han facilitado el desarrollo de sistemas de detección remota de frutos, herramientas de gran utilidad en aplicaciones como la predicción de cosecha, la elaboración de mapas de producción y la recolección automatizada [4]–[6]. Los sensores más utilizados hasta el momento para la detección remota de frutos son las cámaras de color o RGB [7]. Sin embargo, la principal desventaja de estos sensores es que están influenciados por las condiciones de iluminación y que solamente proporcionan información 2D de la plantación. Otros sensores 2D como las cámaras térmicas, multi-espectrales e hiperespectrales también han demostrado tener potencial [7], aunque su uso es menor debido a su mayor precio y a la necesidad de personal especializado para su manejo y operación.

Los avances en fotónica han dado paso a sensores 3D, tales como los sistemas LiDAR (Light Detection and Ranging), que se han utilizado en la agricultura para obtener modelos 3D de las plantaciones [8]. Estos sensores trabajan bajo el principio de *time-of-flight* (ToF – Tiempo de vuelo), que consiste en medir distancias a partir del tiempo requerido por un pulso láser para recorrer el viaje de ida y vuelta entre el sensor y el blanco. Por otro lado, los sensores RGB-D o cámaras de profundidad proporcionan información 3D y de color, permitiendo la detección y la posterior localización 3D de los frutos. Su principio de funcionamiento puede basarse en estereoscopía [9], o en la combinación de una cámara RGB y un sensor de profundidad, ya sea de luz estructurada [10] o ToF [11].

Respecto a las técnicas de procesado, las herramientas de visión artificial más habituales utilizan características tradicionales para codificar las imágenes mediante un número limitado de descriptores y luego clasificarlas mediante algoritmos tales como K-medias (conocido en inglés como K-means), K vecinos más próximos (en inglés K-NN o K-nearest neighbours) o máquinas de soporte vectorial (SVM - Support Vector Machines) [7]. Más recientemente, el desarrollo de redes neuronales profundas ha supuesto un importante avance para la detección remota de objetos, y, por consiguiente, para la detección de frutos. La red neuronal Faster R-CNN [12] es la red más utilizada para la detección remota de frutos [13]–[15].

En este trabajo se implementó la red neuronal Faster R-CNN para la detección de manzanas Fuji mediante el uso del sensor de profundidad Kinect v2 (Microsoft, Redmond, WA, USA). Para ello, se propone utilizar los datos RGB para detectar las manzanas, con la posibilidad de inferir posteriormente la localización 3D de cada detección mediante los datos de profundidad. En la sección 2 se describe el equipamiento y la metodología utilizada para la adquisición de datos, así como la red neuronal implementada para la detección de manzanas Fuji. En la sección 3 se muestran y se discuten los resultados obtenidos, evaluando de forma cualitativa y cuantitativa la bondad del sistema de detección propuesto. Finalmente, las conclusiones que se desprenden de este trabajo se presentan en la sección 4.

2. Materiales y métodos

2.1. Adquisición y preparación de datos

Los datos utilizados en este trabajo se adquirieron en una plantación comercial de manzanas Fuji (*Malus domestica* Borkh. cv. Fuji) localizada en Agramunt (provincia de Lleida, España). Las capturas se realizaron entre los días 25 y 28 de Setiembre de 2017, tres semanas antes de la cosecha. Para ello, se utilizaron dos cámaras de profundidad instaladas en una plataforma móvil a 1 m y 3 m de altura (Figura 1) para poder capturar datos de toda la altura de los árboles. Las cámaras RGB-D utilizadas fueron dos Microsoft Kinect v2 (Microsoft, Redmond, WA, USA), las cuales incorporan una cámara de color RGB y un sensor de profundidad que trabaja bajo el principio de ToF. Ambas cámaras fueron conectadas y sincronizadas con un sensor de posicionamiento RTK-GNSS mediante un ordenador de campo que se comunica con los sensores vía un software desarrollado ad-hoc. Este software permite la adquisición automática de datos y

la georreferenciación de cada una de las capturas. Dado que el rendimiento del sensor de profundidad se ve afectado en condiciones de exposición directa al sol, las imágenes fueron capturadas en horario nocturno mediante luz artificial.

Figura 1. Equipamiento de adquisición de datos mostrando los sensores Kinect v2 montados en una plataforma autopropulsada.

Una vez adquiridos los datos, se etiquetaron manualmente las imágenes, identificando un total de 12,839 manzanas mediante la herramienta de etiquetaje *Pychet Labeller toolbox* [16]. Debido al elevado número de manzanas que aparecen por imagen (más de 100 frutos/imagen), y dado que las dimensiones de los frutos (44 ± 6 píxeles de diámetro) son relativamente pequeñas con respecto al tamaño de la imagen (1600 x 1080 píxeles), cada captura se dividió en 9 sub-imágenes de 548 x 373 píxeles (px), con un solapamiento de 20 px entre sub-imágenes (Figura 2).

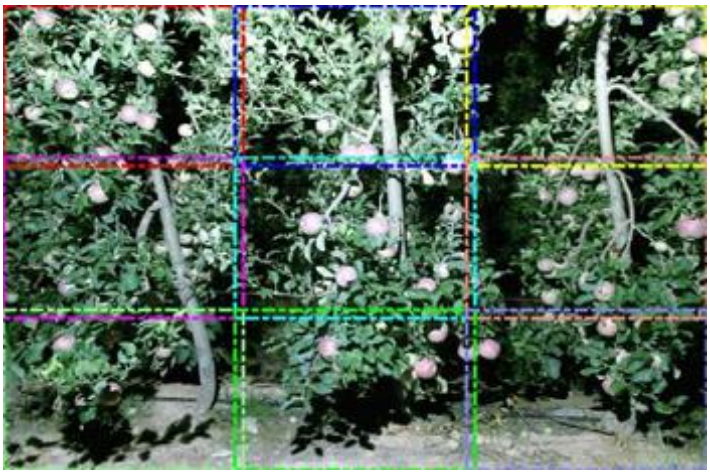

Figura 2. Subdivisión de imágenes. Cada imagen original fue dividida en 9 sub-imágenes para obtener una mejor relación entre el tamaño de las manzanas y la imagen.

2.2. Implementación de la red neuronal

En este trabajo, se utilizó la red neuronal Faster R-CNN [12] para la detección remota de frutos. Esta red fue desarrollada originalmente para detectar objetos en imágenes. La implementación original se testeó con las bases de datos PASCAL VOC [17] y COCO [18], presentando una precisión media (AP) de 78.8% y 42.7% con los datos de test de VOC 2007 y COCO, respectivamente.

Faster R-CNN está compuesta por dos módulos: (1) red de propuesta de regiones (en inglés *Region Proposal Network*, RPN), la cual pretende identificar las regiones de interés con mayor probabilidad de contener un objeto; (2) red de clasificación, que clasifica cada región propuesta por la RPN. Ambos módulos comparten las primeras capas convolucionales, aumentando así la eficiencia computacional de la red. En este trabajo, para estas primeras capas convolucionales, se utilizó el modelo VGG-16 [19] pre-entrenado con la base de datos ImageNet [20].

Para generar propuestas, la RPN evalúa distintas regiones rectangulares en cada posición de la imagen. Los diferentes tipos de regiones evaluadas (llamadas *anchors*) se caracterizan por su escala (área de la región) y la relación de aspecto (relación entre la altura y la anchura de la región). La implementación original de Faster R-CNN propone utilizar escalas de 8, 16 y 32, correspondientes a una superficie de región de 128^2, 256^2 y 512^2 pixeles, y relaciones de aspecto de 1:2, 1:1 y 2:1. Dado que las manzanas a localizar en el presente conjunto de datos son más pequeñas que los objetos localizados en la implementación original de [12], en este trabajo se utilizaron escalas de 2, 4 y 8 (superficie de región de 32^2, 64^2 y 128^2 píxeles, respectivamente), debido a que se adaptan mejor al tamaño de las manzanas objetivo. Se han mantenido las mismas relaciones de aspecto para facilitar la detección de manzanas ocluidas por hojas, ramas u otras manzanas.

3. Resultados y discusión

3.1. Entrenamiento de la red

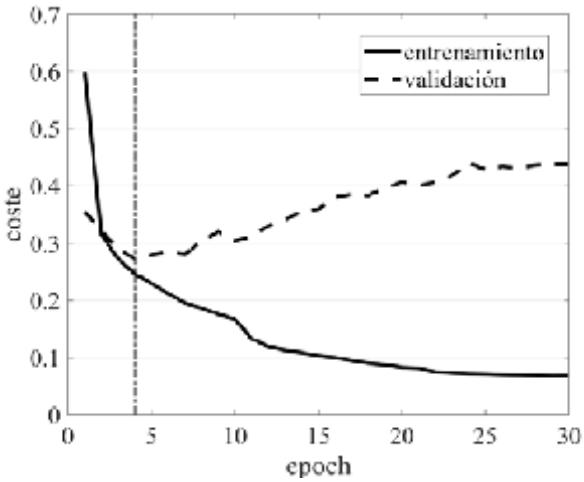

Figura 3. Función de coste en los datos de entrenamiento (línea continua) y de validación (línea discontinua) en función del número de iteraciones de entrenamiento.

La red neuronal fue entrenada utilizando la función de coste propuesta en [12]. Para actualizar los pesos durante el entrenamiento se utilizó el optimizador Adam [21] con una tasa de aprendizaje de 0.0001. El número de imágenes utilizadas para el entrenamiento, validación y test fue de 619, 155 y 193, respectivamente. La Figura 3 ilustra la evolución de la función de coste para el conjunto de datos de entrenamiento y de validación. Se observa que en la iteración (*epoch*) 4, la función de coste de validación alcanza el valor mínimo. Esto significa que a partir de esta

iteración se produce un sobre-entrenamiento de los datos que no generaliza con el conjunto de datos de validación; por este motivo, para evaluar los resultados de test (sección 3.2) se utilizaron los pesos entrenados con las 4 primeras iteraciones.

3.2. Resultados de test

En la Tabla 1 se presentan los resultados obtenidos con el conjunto de datos de test (193 imágenes). Las métricas utilizadas para evaluar estos resultados son: el porcentaje de detección (DR); porcentaje de falsos positivos (FDR); F1-score, definido como la media armónica entre los resultados de precisión (1-FDR) y porcentaje de detección (DR); precisión media (AP); y número de imágenes inferidas por segundo. Para esta evaluación se consideraron detecciones correctas (*true positives*) aquellas que presentan un solapamiento entre la detección y una manzana etiquetada superior a 0.4 (IoU>0.4).

Tabla 1. Evaluación de los resultados de detección de manzanas Fuji.

DR*	FDR*	F1-score	AP*	imágenes/s
91.4 %	15.9 %	0.876	93.9 %	17.3

*DR: porcentaje de detección // FDR: porcentaje de falsos positivos // AP: precisión media.

Los resultados muestran un porcentaje de detección (DR) del 91.4 % de los frutos con un 15.9% de falsos positivos (F1-score=0.876). El tiempo de computación medio fue de 17.3 imágenes por segundo, lo que permite su aplicación en tiempo real. Aunque es difícil comparar metodologías testeadas con distintos conjuntos de datos, los resultados demuestran una robustez similar a otros trabajos del estado del arte basados en redes neuronales como [13]–[15], los cuales reportaron valores de F1-score entre 0.838 y 0.929. Sin embargo, el uso de sensores RGB-D tiene como ventaja adicional que, aunque la detección se realiza en imágenes 2D, se puede inferir la localización 3D de cada detección.

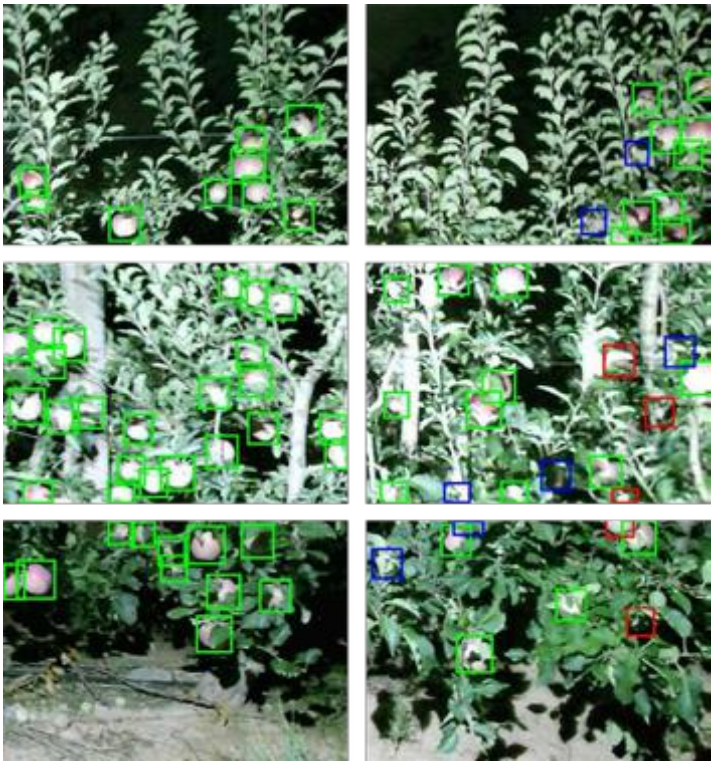

Figura 4. Detecciones obtenidas con la red neuronal Faster R-CNN. Las detecciones correctas se muestran de color verde, los falsos positivos de color rojo y los falsos negativos de color azul.

La Figura 4 ilustra ejemplos de las detecciones obtenidas utilizando la presente metodología. Las imágenes fueron seleccionadas para mostrar casos donde la red neuronal funciona con éxito (primera columna), y casos donde falla (segunda columna). Se observa que la mayoría de los falsos positivos (cuadrados rojos) corresponden a regiones de la imagen que son muy similares a una manzana o bien a manzanas que no fueron etiquetadas debido a errores humanos durante el etiquetado. Por otra parte, muchos de los falsos negativos (manzanas no detectadas, cuadrados azules) corresponden a manzanas altamente ocluidas por otros órganos vegetativos (ramas u hojas) o bien a manzanas que fueron cortadas en los bordes de la imagen.

4. Conclusiones

Este trabajo presenta un sistema de detección remota de manzanas mediante sensores RGB-D y el uso de la red neuronal convolucional Faster R-CNN. Los resultados muestran un porcentaje de detección del 91.4% de los frutos con un 15.9% de falsos positivos. De estos resultados se concluye que el sensor Kinect v2 presenta un gran potencial para la detección y localización 3D de frutos, ya que se obtienen resultados similares a otros trabajos del estado del arte, con la ventaja que con los sensores RGB-D se puede obtener la localización 3D de cada detección. Esto lo que lo hace interesante para aplicaciones como mapeo de la producción y robotización de la cosecha. La evaluación cualitativa de las detecciones muestra que los frutos no detectados corresponden a aquellos que estaban ocultos casi en su totalidad o bien a manzanas cortadas por los bordes de la imagen. Por otra parte, los falsos positivos corresponden a manzanas que no fueron etiquetadas (por errores humanos durante el etiquetaje) o bien a zonas de la imagen que presentan un patrón muy similar a una manzana y donde, incluso según la percepción del ojo humano, es difícil determinar si hay o no una manzana. La principal limitación del sistema es que el desempeño del sensor de profundidad se ve afectado en condiciones de iluminación de elevada intensidad, lo que limita su uso a días poco soleados o a trabajar en horario nocturno con luz artificial. Dado que los sensores RGB-D permiten medir distancias de las escenas capturadas, trabajos futuros podrían estudiar la capacidad de estos sensores para medir las dimensiones de los frutos de forma remota.

5. Agradecimientos

Este trabajo ha sido parcialmente financiado por la Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, el Ministerio de Economía y Competitividad de España, y the European Regional Development Fund (ERDF), mediante los proyectos 2017 SGR 646, AGL2013-48297-C2-2-R, MALEGRA TEC2016-75976-R. Se agradece al Ministerio de Educación de España por la beca pre-doctoral de J. Gené (FPU15/03355). También queremos agradecer a Nufri y Vicens Maquinària Agrícola S.A. por su apoyo durante la adquisición de datos, así como a Adrià Carbó por su colaboración en la implementación de la red neuronal.

Referencias

1. K. R. Siegel, M. K. Ali, A. Srinivasiah, R. A. Nugent, and K. M. V. Narayan, "Do we produce enough fruits and vegetables to meet global health need?," *PLoS One*, vol. 9, no. 8, 2014.
2. F. A. Auat Cheein and R. Carelli, "Agricultural robotics: Unmanned robotic service units in agricultural tasks," *IEEE Ind. Electron. Mag.*, vol. 7, no. 3, pp. 48–58, 2013.
3. S. Bargoti and J. P. Underwood, "Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards," *J. F. Robot.*, vol. 00, no. 0, pp. 1–22, 2017.
4. J. P. Underwood, C. Hung, B. Whelan, and S. Sukkarieh, "Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors," *Comput. Electron. Agric.*, vol. 130, pp. 83–96, 2016.
5. R. Linker, "Machine learning based analysis of night-time images for yield prediction in apple orchard," *Biosyst. Eng.*, 2018.

6. C. W. Bac, E. J. Van Henten, J. Hemming, and Y. Edan, "Harvesting Robots for High-value Crops: State-of-the-art Review and Challenges Ahead," *J. F. Robot.*, vol. 31, no. 6, 2014.
7. A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis, "Sensors and systems for fruit detection and localization: A review," *Comput. Electron. Agric.*, vol. 116, pp. 8–19, 2015.
8. A. Escolà *et al.*, "Mobile terrestrial laser scanner applications in precision fruticulture/horticulture and tools to extract information from canopy point clouds," *Precis. Agric.*, vol. 18, no. 1, pp. 111–132, 2017.
9. D. Font *et al.*, "A proposal for automatic fruit harvesting by combining a low cost stereovision camera and a robotic arm," *Sensors (Switzerland)*, vol. 14, no. 7, pp. 11557–11579, 2014.
10. T. T. Nguyen, K. Vandevoorde, N. Wouters, E. Kayacan, J. G. De Baerdemaeker, and W. Saeys, "Detection of red and bicoloured apples on tree with an RGB-D camera," *Biosyst. Eng.*, vol. 146, pp. 33–44, 2016.
11. E. Barnea, R. Mairon, and O. Ben-Shahar, "Colour-agnostic shape-based 3D fruit detection for crop harvesting robots," *Biosyst. Eng.*, vol. 146, pp. 57–70, 2016.
12. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
13. S. Bargoti and J. Underwood, "Deep Fruit Detection in Orchards," *2017 IEEE Int. Conf. Robot. Autom.*, pp. 3626–3633, 2017.
14. H. Gan, W. S. Lee, V. Alchanatis, R. Ehsani, and J. K. Schueller, "Immature green citrus fruit detection using color and thermal images," *Comput. Electron. Agric.*, vol. 152, no. July, pp. 117–125, 2018.
15. I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "DeepFruits: A Fruit Detection System Using Deep Neural Networks," *Sensors*, vol. 16, no. 8, p. 1222, 2016.
16. S. Bargoti, "Pychet Labeller. Available online: https://github.com/acfr/pychetlabeller." https://github.com/acfr/pychetlabeller, 2016.
17. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, 2010.
18. T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.
19. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14, 2014.
20. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
21. D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," *ICLR Int. Conf. Learn. Represent.*, 2015.