IDENTIFICACIÓN Y CONTEO DE ACEITUNAS EN IMÁGENES DIGITALES TOMADAS EN EL OLIVAR MEDIANTE MORFOLOGÍA MATEMÁTICA Y REDES NEURONALES CONVOLUCIONALES

Arturo Aquino, Juan Manuel Ponce, Borja Millan, Diego Tejada-Guzmán, José Manuel Andújar
{arturo.aquino, jmponce.real, borja.millan, diego.tejada, andujar}@diesia.uhu.es
Departamento de Ingeniería Electrónica, Sistemas Informáticos y Automática. Universidad de Huelva.
Carretera Huelva-Palos, s/n, 21810 Palos de la Frontera (España).

Resumen

La estimación precoz y precisa de la producción es un objetivo muy codiciado en la agricultura moderna. En el caso de la olivicultura, ello toma una especial relevancia debido al alto valor económico que alcanza su producción. Este artículo presenta una metodología enfocada a lograr dicho objetivo. Concretamente, se propone un algoritmo de visión artificial capaz de detectar las aceitunas visibles en una imagen digital de un árbol de olivo, tomada directamente en campo, de noche y con iluminación artificial. En primera instancia, esta imagen es preprocesada mediante técnicas de morfología matemática y filtrado estadístico para, a partir de ella, obtener un conjunto de subimágenes con alta probabilidad de contener una aceituna. Este preprocesamiento reduce el espacio potencial de búsqueda en una magnitud de 10^3. A continuación, estas subimágenes son clasificadas por una red neuronal convolucional como 'aceituna' o 'descarte'. De un total de 304.483 subimágenes, extraídas de 21 imágenes, la red clasificó correctamente el 98,23%, y arrojó un coeficiente de determinación R^2 igual a 0,9875, al enfrentar el número de aceitunas detectadas con el obtenido manualmente. Esta precisión alcanzada indica que el algoritmo desarrollado constituye un paso certero en la implementación de un futuro sistema de estimación de la producción de cultivos de olivo.

Palabras clave: Agricultura de precisión, estimación de la producción, aceituna, visión artificial, red neuronal convolucional.

1 INTRODUCCIÓN

El cultivo del olivo (*Olea europaea* L.), y su mercado asociado, es un notable motor económico para buena parte de la cuenca mediterránea. Lo es especialmente en el caso de España, la cual aportó en 2017 el 31,38% del total de aceitunas producidas mundialmente [1]; esta cifra asciende hasta el 61,95% cuando se considera también la producción de Portugal, Italia y Grecia [1]. Además, el cultivo del olivo trasciende de lo económico a lo social, ya que, por ejemplo, el aceite de oliva es la piedra angular de la dieta mediterránea, la cual ha sido declarada Patrimonio Inmaterial de la Humanidad por la UNESCO [2].

La agricultura en general, y la olivicultura en particular, se enfrenta a importantes desafíos, como son la sostenibilidad medioambiental y la productividad [3]. Estas dificultades, lejos de estar cerca de ser solventadas, se recrudecerán al ritmo exponencial de crecimiento esperado de demanda de alimentos, ligado éste a las estimaciones de desarrollo de la población mundial (según estima la ONU, desde los actuales 7.600 M de habitantes, se llegará a los 9.800 M en 2050 [4]).

En este contexto, es donde la agricultura de precisión recibe un apoyo y una atención muy relevantes en los últimos años [5]. Tradicionalmente, los cultivos han sido gestionados gracias a la capacidad de percepción de los agricultores para evaluar su desarrollo vegetativo, su necesidad de riego o nutrientes, la madurez del fruto, la cantidad de cosecha esperada, etc. En contraposición, la agricultura de precisión auspicia el uso de tecnología para adquirir información objetiva y de alta precisión espacio-temporal acerca del estado de los cultivos, para posibilitar así una óptima gestión de los mismos [5].

De entre la información deducible del olivar, la estimación temprana y precisa de la cosecha destaca en interés, ya que sería una valiosa herramienta de apoyo al sector [6]. Y es que, en efecto, una estimación precisa de la producción tendría aplicaciones prácticas en muy diversos aspectos, tales como: eficiencia en la transformación del aceite de oliva, gestión del stock, o la optimización de los recursos humanos necesarios para la recolección [7]. También tendría una importancia crucial en la regulación del precio de mercado del aceite de oliva. La volatilidad de este mercado está fuertemente ligada a las expectativas de producción, actualmente evaluadas de manera visual.

https://doi.org/10.17979/spudc.9788497497169.818

Ello ha llevado frecuentemente a unos niveles de inestabilidad de precios declaradamente excesivos [8]. La importancia de la estimación temprana de la cosecha es transversal a todos los cultivos con productos derivados de alto valor de mercado. Como ejemplo valga el cultivo de vid, para el que se pueden encontrar en la bibliografía una variedad de metodologías experimentales basadas en visión artificial enfocadas a ello [9-12]. Otros cultivos como el de la palma [13], el arroz [14], los cereales [15,16], o los frutales [17-19], han recibido también una atención relevante a este respecto. Sin embargo, para el cultivo del olivo, la bibliografía es parca en estudios y, en su mayoría, están basados en estimar la producción a través de variables indirectas. Ejemplo de ello son los trabajos referenciados en [20-22], los cuales plantean la resolución del problema a través de la generación de modelos alimentados con variables meteorológicas y de polinización. La principal debilidad de este enfoque radica en la naturaleza misma de estas variables, con un fuerte componente estocástico, lo cual compromete la reproducibilidad de los modelos. Además, el fenómeno de formación y mantenimiento del fruto es virtualmente muy complejo y resultado de la conjunción de multitud de variables. A este respecto, los trabajos referenciados no demuestran la representatividad de las variables seleccionadas con relación a las omitidas.

Este artículo presenta una metodología basada en visión artificial para la detección y conteo de las aceitunas visibles en imágenes digitales tomadas en campo. Para ello, primero se genera, a partir de las imágenes, un conjunto reducido de candidatos (subimágenes) mediante operadores de morfología matemática, reduciendo así el espacio potencial de búsqueda en una magnitud de 10^3. A continuación, estas subimágenes son preprocesadas y clasificadas por una red neuronal convolucional (CNN), la cual valida o descarta cada instancia en función de si contiene una aceituna o no. El nivel de precisión de la metodología propuesta argumenta a favor de su integración en un futuro sistema para la estimación precoz de la cosecha del olivar.

2 MATERIALES Y MÉTODOS

2.1 ADQUISICIÓN DE IMÁGENES

En septiembre de 2018 (dos meses antes de recolección), se adquirió un conjunto de 36 imágenes de la variedad Picual en un olivar de cultivo intensivo situado en Gibraleón, (37°20'09.2"N 7°02'19.8"W), provincia de Huelva (Andalucía, España); el marco de plantación de la parcela era de 5,5 × 7 m. Los individuos de estudio se seleccionaron para conseguir la máxima variabilidad disponible en cuanto a su capacidad productiva de aceitunas.

Las imágenes se tomaron de noche con la iluminación artificial proporcionada por un foco halógeno de 500 W, ya que ello favorece la simplificación de la escena en el fondo del objeto de interés. Esta decisión fue tomada gracias a la experiencia previa de los autores en trabajos similares realizados en vid [12]. La captura de las imágenes se realizó de manera manual y con trípode, usando para ello una cámara sin espejo Sony α7-II (Sony Corp., Tokyo, Japan) equipada con una óptica Zeiss estabilizada de 24/70 mm (Zeiss Gruppe, Oberkochen, Germany). Dicha cámara monta un sensor CMOS de 24 Mpx estabilizado en cinco ejes, y fue configurada con las siguientes características: apertura de f/14, tiempo de exposición de 1/200 s, distancia focal de 24 mm, resolución de 6.000 x 3.376 píxeles, profundidad de color de 24 bits, y formato de almacenamiento de imagen JPEG de mínima compresión. La distancia a la que se tomaron las imágenes osciló entre los 2 m y los 4 m dependiendo del porte del árbol para, en cada caso, obtener un encuadre apropiado de toda su superficie foliar. La figura 1 muestra un ejemplo de imagen capturada bajo los preceptos descritos.

Figura 1: imagen de olivo de la variedad Picual, sujeto de este estudio, obtenida de noche con iluminación artificial.

2.2 METODOLOGÍA DE ANÁLISIS DE IMAGEN

La metodología se basa en el uso de una red neuronal convolucional (CNN) para detectar las aceitunas presentes en una imagen de olivo como la mostrada en la figura 1. Para ello, previamente se aplica un preprocesamiento dirigido a: (1) mejorar las condiciones de partida de la imagen; (2) reducir el espacio de búsqueda mediante la obtención de un conjunto reducido de subimágenes (candidatos) con alta probabilidad de contener individualmente las aceitunas de la imagen; (3) configurar la composición y contenido de las subimágenes para optimizar el rendimiento de la CNN.

La metodología descrita en este artículo fue implementada usando la plataforma Matlab R2018b (The MathWorks, Inc., Natick, Massachusetts, USA).

2.2.1 Preprocesamiento

Sea la imagen digital *I*, análoga a la mostrada en la figura 1, generada de manera nativa de acuerdo con el espacio de color RGB, y codificada con 8 bits por canal. Esta imagen inicial es transformada al espacio de color CIE 1976 *L*a*b**, ya que ello posibilita el análisis independiente de las componentes de iluminación (canal *L**) y color (canales *a** y *b**) [23]. Así, quedan definidas las imágenes *L*, *A* y *B*, correspondientes a los canales *L**, *a** y *b**, respectivamente, que componen la imagen transformada. Para este trabajo, se descartó el uso de la imagen *A* por ofrecer un contraste de niveles de gris ínfimo en las regiones de interés.

En primer lugar, para reducir ruido de "sal y pimienta", se aplica un filtrado promedio circular, dentro de un kernel de convolución *k* de tamaño 11 × 11, a las componentes *L* y *B*:

$$L_s = L * k;\ B_s = B * k \tag{1}$$

donde el círculo de activación del filtro tiene un diámetro de 11 píxeles. El reducido tamaño de este filtro comparado con el de la imagen, 11 x 11 vs 6.000 x 3.376, profiere cierta tolerancia al establecimiento de su valor para conseguir el objetivo perseguido. A su vez, el promediado circular favorece el "limpiado" de los patrones de iluminación análogos las aceitunas.

A continuación, a partir de L_s, se genera un conjunto de semillas que posteriormente servirá para generar un conjunto de subimágenes o candidatos que permita reducir el espacio de búsqueda. Para ello, cada semilla será una agrupación píxeles vecinos (conocida formalmente como componente conexa, *CC*), que representa la ocurrencia de una región máxima local de iluminación. Gracias a la ley del coseno de Lambert [10], se conoce que una superficie convexa, como la de una aceituna, produce un intenso patrón circular de reflexión de la luz. Por ello, esta operación favorece la localización de las aceitunas en la imagen. Estas regiones máximas se obtienen mediante la aplicación de la transformada morfológica *h-maxima*, la cual encuentra aquellas regiones en la imagen que cumplen tener una elevación menor o igual que el escalar *h*. La transformada, calcula en primer lugar la reconstrucción morfológica, *R*, de la imagen L_s a partir del marcador $L_s - h$ (consultar los detalles del operador reconstrucción en [24]),

$$L_{filt} = R_{L_s}(L_s - h); h = 3 \tag{2}$$

para finalmente extraer las regiones máximas en la imagen L_{filt}, mediante el cálculo de

$$L_{RM} = L_{filt} - R_{L_{filt}}(L_{filt} - 1) \tag{3}$$

y la segmentación de la imagen resultante:

$$L_{RMbin} = \begin{cases} 255 & if\ \ L_{RM}(x,y) > 0 \\ 0 & en\ otro\ caso \end{cases} \tag{4}$$

En situaciones de aplicación análogas, se ha discutido acerca del valor óptimo para el parámetro *h* [12, 25]. Estos trabajos concluyeron que valores en torno al 3 o el 4 eran óptimos, aunque no excluyentes, ya que ligeras variaciones también proporcionaban resultados satisfactorios; para la experimentación descrita en este trabajo se estableció $h = 3$. Finalmente, el conjunto de regiones máximas de iluminación (o semillas) se compone con las componentes conexas, CC_i, existentes en la imagen binaria L_{RMbin}:

$$S_{RM} = \{CC_i \subseteq L_{RMbin}\} \tag{5}$$

La figura 2 ilustra el resultado del cálculo del conjunto de componentes conexas provenientes de regiones máximas de iluminación.

La siguiente operación consiste en obtener un conjunto de subimágenes candidatas centradas en las semillas encontradas, tanto para L_s como para B_s; estos conjuntos se denotarán como S_{L_s} y S_{B_s}, respectivamente. Para ello, para cada componente conexa $CC_i \in S_{RM}$ se extrae, en torno a su centroide, una subimagen de L_s y otra de B_s de tamaño 41 × 41. Así, se cumple que

$$S_{L_s} = \{s_{L_s}^i\}, S_{B_s} = \{s_{B_s}^j\}; 1 \le i, j \le \#(S_{RM}) \tag{6}$$

y que

$$\begin{gathered} \forall i\ 1 \le i \le \#(S_{RM}), \\ s_{L_s}^i\ "deriva\ de"\ CC_i\ \wedge s_{B_s}^i\ "deriva\ de"\ CC_i \end{gathered} \tag{7}$$

Debido a la variación en la distancia de fotografiado y al volumen de los olivos, las aceitunas impresionan con muy diversos tamaños en las distintas imágenes. Las dimensiones para las subimágenes fueron establecidas empíricamente para que contuvieran completamente a una aceituna independientemente de su tamaño. Por otro lado, en media, se generaron 14.499 subimágenes a partir de una imagen. Teniendo en cuenta que, para imágenes de resolución 6.000 x 3.376 píxeles, el espacio potencial de búsqueda está compuesto por 20.256.000 candidatos (número de píxeles que la componen), éste quedó reducido en un orden de magnitud de 10^3.

El último paso del preprocesamiento consiste en, una vez obtenidas las subimágenes de los conjuntos S_{L_s} y S_{B_s}, encontrar una combinación de ellas tal que a la postre favorezca el rendimiento de una CNN en la tarea de determinar si una subimagen contiene una aceituna o no. Es más, la red deberá ser capaz de discernir si la subimagen en evaluación está centrada

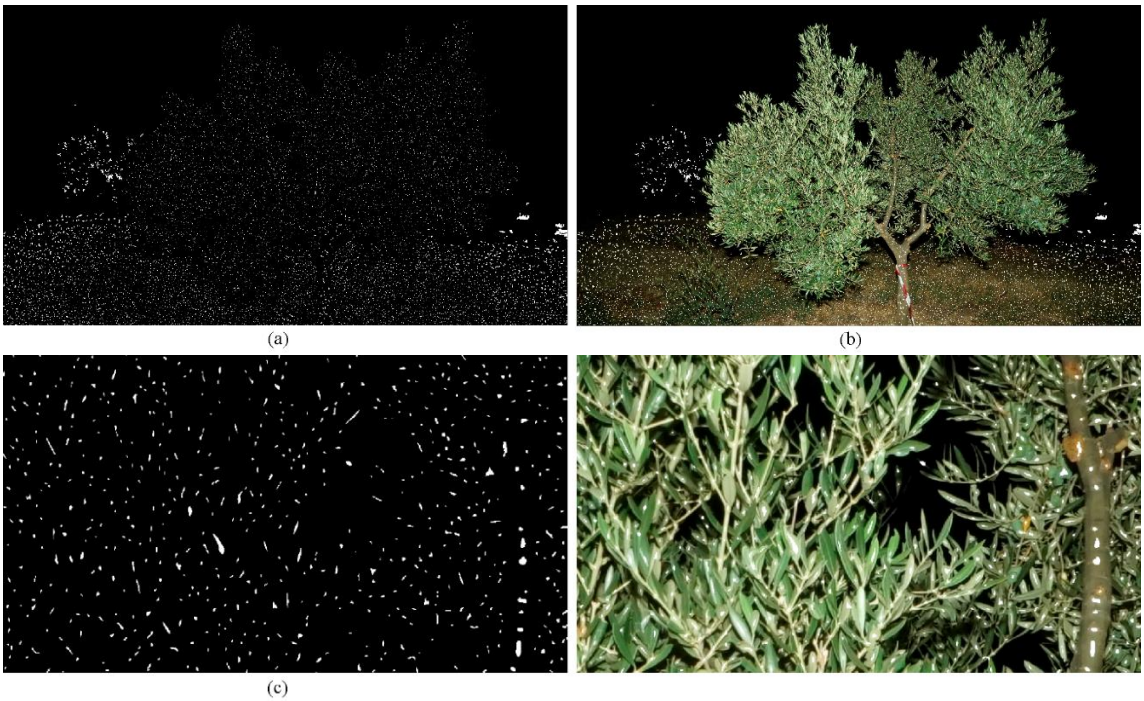

Figura 2: Resultado del procedimiento de obtención de componentes conexas de iluminación, o semillas, para la reducción del espacio de búsqueda: (a) conjunto de semillas obtenidas a partir de la imagen de la figura 1 representadas en blanco sobre fondo negro; (b) conjunto de semillas de (a) representadas sobre la imagen original; (c) sección de la imagen (a); (d) sección de la imagen (b), análoga a la (c).

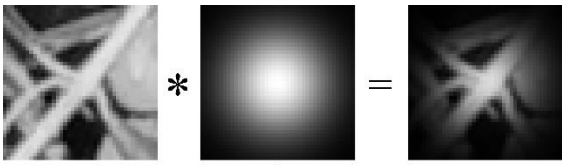

Figura 3: (a) subimagen de L_s, $s^i_{L_s}$, extraída en torno al centroide de una semilla producida por el brillo de una rama; nótese que la imagen contiene representaciones parciales de dos aceitunas; (b) ilustración de una matriz Gaussiana de tamaño 41 × 41 y $\sigma = 9{,}5$; (c) subimagen $s^i_{L'_s}$, resultado de la multiplicación elemento a elemento de (a) y (b), en la que se atenúa la presencia de las aceitunas por no estar centradas en (a).

en un brillo proveniente de una aceituna para validarla, o de cualquier otro elemento, como por ejemplo una hoja, para descartarla. En este sentido, es muy frecuente la existencia de subimágenes que contienen alguna aceituna, total o parcialmente, pero que han sido generadas a partir de un brillo de otro elemento. En este caso, la red deberá de ser capaz de discernir que la subimagen candidata está centrada en un elemento anómalo para descartarla. Para favorecer este precepto, se crea un nuevo conjunto a partir de S_{L_s}, $S_{L'_s}$, en el que sus subimágenes son multiplicadas, elemento a elemento, por una matriz Gaussiana normalizada G de tamaño 41 × 41 y $\sigma = 9{,}5$:

$$S_{L'_s} = \left\{ s^i_{L'_s} | s^j_{L'_s}(k,l) = s^j_{L_s}(k,l) \times G(k,l); \quad \forall k \forall l, 1 \le k, l \le 41, 1 \le j \le \#(S_{RM}) \right\} \tag{8}$$

En efecto, la subimagen $s^i_{L'_s}$ contiene la misma información que $s^i_{L_s}$, pero ponderada en importancia de manera decreciente desde el centro hacia el exterior de ésta (ver la figura 3).

Finalmente, la unión de los conjuntos ordenados S_{L_s}, S_{B_s} y $S_{L'_s}$ configura el espacio de búsqueda de la imagen I de la que proceden. Así, cada tripleta de elementos i-ésimos puede ser considerada una subimagen candidata de tres canales y tamaño 41 × 41, en la que $s^i_{L_s}$ aporta información de iluminación, $s^i_{L'_s}$ pondera la relevancia de dicha información, y $s^i_{B_s}$ proporciona conocimiento relativo al color (Figura 4).

2.2.2 Procedimiento de detección de aceitunas

Esta etapa de la metodología de análisis de imagen trata sobre la configuración y el entrenamiento de una CNN capaz de clasificar las subimágenes candidatas generadas en la fase de preprocesamiento, como *aceituna* (caso positivo) o *descarte* (caso negativo).

2.2.2.1 Arquitectura de la CNN

A grandes rasgos, las CNN están compuestas de dos estructuras principales conectadas. La primera,

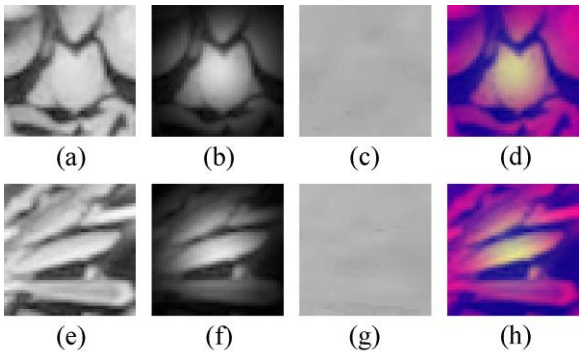

Figura 4: (a)-(d) subimágenes $s_{L_s}^i$, $s_{L_s'}^i$, $s_{L_B}^i$ y representación como imagen RGB de su combinación, respectivamente; son resultado de la semilla *i*-ésima procedente del máximo de iluminación local producido por una aceituna; (e)-(h) subimágenes $s_{L_s}^j$, $s_{L_s'}^j$, $s_{L_B}^j$ y representación como imagen RGB de su combinación, respectivamente; son resultado de la semilla *j*-ésima procedente del máximo de iluminación local producido por una hoja.

encargada de la extracción de características, está en esencia conformada por un conjunto de capas, de dimensionalidad uno o más, que implementan filtros convolucionales, y que están a menudo intercaladas con otras de normalización o reducción del muestreo. Esta primera estructura, a su salida, conecta con un perceptrón multicapa clásico, el cual se encarga de aprender las directrices de clasificación. El valor de todas las variables derivadas de este planteamiento multicapa, es ajustado de forma concurrente e iterativa mediante un algoritmo que persigue la minimización de una función de coste que expresa el error de clasificación de la red. Para una descripción más precisa sobre las CCN, consúltese [26].

Existen arquitecturas que han demostrado gran solvencia en la resolución de problemas complejos, y que han contribuido enormemente al avance del aprendizaje profundo. Entre las más relevantes, consideradas hoy prácticamente un estándar, cabe destacar: AlexNet [27], VGG-16 [28], GoogLeNet/Inception [29], ResNet [30] o Inception-ResNet [31], que hibrida las arquitecturas Inception y ResNet. Debido a la complejidad de las imágenes y las diferentes magnificaciones de los objetos que contienen, se seleccionó Inception-ResNet para el trabajo que se presenta. Por un lado, el enfoque Inception postula la aplicación de convoluciones de distintos tamaños en la misma capa, lo cual aporta flexibilidad en la detección de patrones que pueden aparecer con muy diversas dimensiones. Estos postulados impresionan favorecer el reconocimiento de aceitunas, independientemente de su tamaño. Por su lado, el paradigma ResNet aporta a la red la potencia del tratamiento de los residuos de aprendizaje. Para una capa dada, el residuo es básicamente la diferencia entre lo aprendido al inicio y al final de la misma. La arquitectura ResNet explota este residuo en su misma topología para favorecer y optimizar la convergencia a la solución óptima.

Inception-Resnet, concretamente su versión 2 que fue la usada en este trabajo, es una red de 164 capas y 55,9 M de parámetros. Ello, junto con su arquitectura híbrida anteriormente discutida, le confiere una gran potencia de razonamiento. A su entrada, la red admite imágenes de tamaño 299 × 299 × 3.

2.2.2.2 Entrenamiento y optimización de la CNN

Las CNN enumeradas anteriormente, y algunas otras populares, están disponibles a través de internet, entrenadas con millones de imágenes, para ser manipuladas o usadas con distintas plataformas de desarrollo. Aunque las versiones disponibles no hayan sido entrenadas para el caso de interés requerido, el conocimiento con el que se proporcionan es de gran valor ya que posibilita efectuar lo que se conoce como "transfer learning". Este paradigma consiste en sustituir el perceptrón de la CNN seleccionada por uno configurado y sin entrenar, conservando el bloque convolucional. De esta manera, se parte para el entrenamiento deseado de una CNN con gran capacidad adquirida de extracción de características. Este fue el enfoque empleado en esta investigación.

Previamente a poder realizar el entrenamiento por transferencia de aprendizaje de la CNN, hubo que crear un conjunto de instancias de entrenamiento. Cada instancia estuvo constituida por la tripleta de subimágenes $s_{L_s}^i$, $s_{L_s'}^i$ y $s_{L_B}^i$ (ver ejemplos en la figura 4, imágenes (a)-(c) y (e)-(g)), y fue catalogada como positiva, o *aceituna*, si éstas estaban centradas en una aceituna, o como negativa, o *descarte*, en caso contrario. Para generar dichas instancias catalogadas, en primer lugar, se seleccionaron 15 imágenes de las 36 totales (las 21 restantes se reservaron para validación externa). A continuación, se calcularon sus componentes conexas provenientes de regiones máximas de iluminación aplicando el preprocesamiento formulado en las ecuaciones (1)-(5). Estas componentes conexas se representaron sobre las imágenes originales tal y como se ilustra en la figura 2 y sobre ellas, usando un software de edición de imagen, se etiquetaron aquellas componentes correspondientes con brillos de aceitunas, independientemente del grado de oclusión que presentaran (ver la figura 5). De esta forma, se etiquetaron un total de 9.575 casos positivos, y 208.848 negativos. En torno a las componentes ya etiquetadas, tanto de los casos positivos como de los negativos, se extrajeron las subimágenes $s_{L_s}^i$, $s_{L_s'}^i$ y $s_{L_B}^i$ correspondientes siguiendo el resto del preprocesamiento especificado por las ecuaciones (6)-

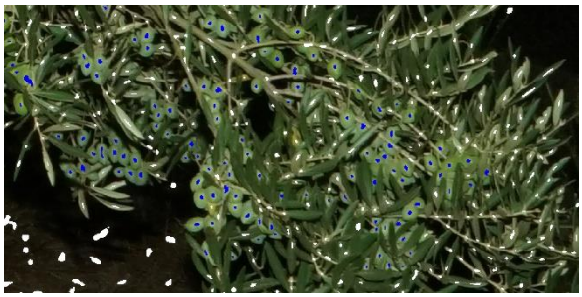

Figura 5: ilustración del etiquetado de instancias de entrenamiento. Se muestra una sección de la imagen de la figura 2-(b) en la que las componentes conexas coincidentes con brillos de aceitunas son etiquetadas en azul como casos positivos. Los casos negativos son etiquetados en color blanco. Nótese que todas las aceitunas, independientemente de su grado de oclusión, son etiquetadas como caso positivo.

(8). Finalmente, para ampliar el conjunto de instancias positivas, las subimágenes implicadas se rotaron 30º, 60º, 90º, …, 330º, para así configurar un conjunto definitivo de instancias positivas de 117.084; de entre las 208.848 instancias negativas, se seleccionó aleatoriamente el mismo número, para obtener así conjuntos balanceados. Finalmente, todas las subimágenes fueron escaladas al tamaño de entrada de la CNN, esto es, 299 × 299, usando interpolación bilineal.

El 80% de las instancias positivas y de las negativas, seleccionadas de forma aleatoria, se usó como conjunto de entrenamiento, mientras que el 20% restante fue empleado en validación para reducir las probabilidades de aparición de sobre-entrenamiento. Como algoritmo de aprendizaje se empleó el del descenso estocástico del gradiente optimizado por momentum (sgdm) [32], con una ratio de aprendizaje que varió desde 10^{-2} hasta 10^{-5} en función de la evolución del proceso de convergencia. Se configuró un tamaño para los mini-conjuntos de entrenamiento de 28 instancias, lo que resultó en 6.690 iteraciones por época. Para llegar a una solución óptima, la CNN entrenó un total de 60 épocas.

3 RESULTADOS Y DISCUSIÓN

3.1 METODOLOGÍA DE EVALUACIÓN DE RESULTADOS

Los resultados de la metodología presentada fueron evaluados utilizando métricas basadas en tablas de contingencia para clasificación binaria. Para ser factible este enfoque, en primer lugar, se generó un conjunto *gold standard* para las 21 imágenes de validación externa. Ello se realizó siguiendo un procedimiento análogo al descrito anteriormente, e ilustrado en la figura 5, para etiquetar las instancias de entrenamiento. Esto es, las imágenes se preprocesaron aplicando las operaciones (1)-(5) para, a continuación, etiquetar manualmente las componentes conexas como *aceituna* o *descarte* usando un software de edición de imagen.

Con la disponibilidad del conjunto *gold standard*, las posibles ocurrencias de clasificación que se podían producir al evaluar las imágenes de validación externa, quedaron definidas como:

- TP (verdadero positivo): instancia catalogada como *aceituna* por la CNN, que también fue etiquetada como tal en el conjunto *gold standard*.
- TN (verdadero negativo): instancia catalogada como *descarte* por la CNN, que también fue etiquetada como tal en el conjunto *gold standard*.
- FP (falso positivo): instancia catalogada como *aceituna* por la CNN, que fue etiquetada como *descarte* en el conjunto *gold standard*.
- FN (falso negativo): instancia catalogada como *descarte* por la CNN, que fue etiquetada como *aceituna* en el conjunto *gold standard*.

Con estas definiciones, las métricas usadas para evaluar el rendimiento del algoritmo de análisis de imagen fueron:

$$PR = \frac{TP}{TP + FP};$$

$$RC = \frac{TP}{TP + FN};$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}; \quad (9)$$

$$F - score = 2 \times \frac{PR \times RC}{PR + RC}$$

Donde *PR* (*precision)* es la tasa de acierto alcanzada en la inferencia realizada sobre la clase *aceituna*, *RC* (*recall*) es la fracción de instancias *aceituna* que el algoritmo logra catalogar, *ACC* (*accuracy*) evalúa la tasa de acierto global alcanzada para las dos clases, *y F-score* es la media armónica de *PR* y *RC*.

Adicionalmente, también se estudió el coeficiente de determinación R^2 resultante de enfrentar, para cada imagen, el número de aceitunas detectadas con el establecido mediante conteo manual (el establecido en el conjunto *gold standard*).

3.2 EVALUACIÓN Y DISCUSIÓN DE RESULTADOS

La tabla 1 detalla, para cada imagen del conjunto de validación externa, el número de aceitunas recogido en el conjunto *gold standard*, el número de instancias generado por el preprocesamiento, y las ocurrencias

Tabla 1: Especificación por imagen perteneciente al conjunto de validación externa, y en total, de las instancias a clasificar generadas por el preprocesamiento, el número de aceitunas reales recogidas en el conjunto *gold standard*, y las ocurrencias de clasificación.

Imagen	Aceitunas reales	Instancias	TP	TN	FP	FN
1	907	13.745	823	11.827	84	188
2	904	14.058	788	12.307	117	58
3	1.225	17.158	981	14.831	244	121
4	702	14.219	544	12.879	158	94
5	1.026	15.973	890	13.907	137	149
6	1.227	14.972	1.113	12.432	114	200
7	882	14.315	794	12.353	88	286
8	795	14.146	727	12.165	69	458
9	984	14.092	891	11.877	93	340
10	965	16.167	862	14.106	103	234
11	368	12.261	299	11.539	69	55
12	560	15.378	469	14.234	91	115
13	740	13.871	646	12.354	94	131
14	566	15.193	507	13.988	59	132
15	493	15.646	432	14.601	61	120
16	543	12.375	483	11.239	60	110
17	939	13.252	846	11.318	94	148
18	938	15.972	842	13.986	97	205
19	969	14.143	866	12.098	105	208
20	75	11.674	25	11.566	50	8
21	63	15.873	38	15.764	25	8
Total	15.871	304.483	13.866	271.371	2.012	3.368

Tabla 2: Resultados de la metodología de análisis de imagen calculados sobre el total de TP, TN, FP y FN producidos para las 21 imágenes de validación externa.

PR	RC	F-score	ACC
0,8733	0,8046	0,8375	0,9823

de clasificación producidas por la CNN; también se incluye esta información agrupando los datos de todas las imágenes como un todo. La tabla 2 muestra los resultados medidos en términos de las métricas definidas anteriormente en (9). Por un lado, y atendiendo al resultado de *Acc*, el rendimiento general del clasificador podría ser calificado de destacable, ya que clasificó correctamente el 98,23% de las instancias. Por otro lado, de acuerdo con el valor de *PR*, se produjo un 87,33% de acierto cuando la CNN clasificó una instancia como "aceituna", mientras que el *RC* indica que el 80,46% de todas las instancias "aceituna" fueron correctamente clasificadas. La conjunción de ambos aspectos, en términos de *F-score*, dio como resultado un rendimiento del 83,75%. Para contextualizar estos resultados, conviene ponderar varias circunstancias. La complejidad de las imágenes para los fines perseguidos fue notable, contribuyendo a esa dificultad: la similitud en color de las aceitunas y las hojas de olivo, la marcada variabilidad en tamaño de las aceitunas, que éstas se presentan tanto aisladas como junto a otras, y que pueden presentar importantes grados de oclusión con otros elementos como hojas, ramas u otras aceitunas. Ninguno de estos distractores fue evadido, ya que se persiguió la detección de aceitunas en cualquier circunstancia, y con cualquier apariencia y grado de oclusión, para lo que se diseñó un preprocesamiento específico que posibilitó alcanzar los resultados descritos.

La figura 6 muestra el resultado del análisis de correlación realizado al enfrentar, por imagen, el número de aceitunas reales recogidas en el conjunto *gold standard*, con las detectadas por la metodología descrita en este artículo. Como se puede apreciar, tanto en la gráfica como en el valor medido de R^2 que alcanzó el 0,9875, el grado de congruencia estadística entre las dos variables fue verdaderamente elevado. En el trabajo presentando en [12] para la estimación de la producción en viñedos, planteado con un enfoque similar, los autores encontraron un valor de R^2 de 0,9829 al enfrentar el número de granos de uva visibles en las imágenes con el determinado mediante etiquetado manual. En esa misma investigación, se encontró también una fuerte correlación entre el número de granos visibles detectados en las imágenes y el número absoluto de granos que contenían las plantas, incluyendo la fracción de ellos no visibles y, por tanto, no detectables en las imágenes. Ello resultó

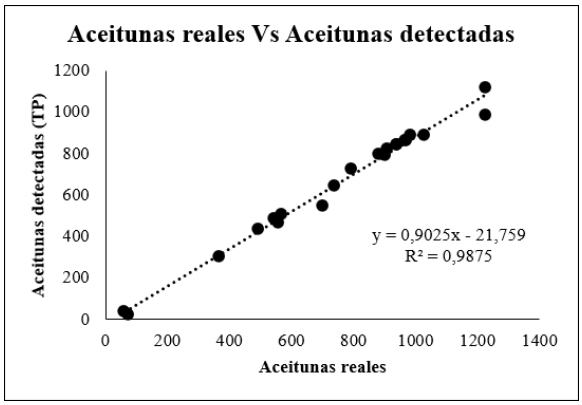

Figura 6: Resultado del estudio de correlación realizado enfrentando el número de aceitunas detectadas por el algoritmo con el número real.

en la emisión de una estimación de cosecha en gramos con un error medio del 1,3%. Por todo ello, el valor de R^2 obtenido en este trabajo argumenta a favor de que, si se encuentra una fuerte correlación entre el número de aceitunas detectadas mediante análisis de imagen y las totales contenidas en el árbol, se puede llegar a una metodología precisa de estimación temprana de la producción del olivar.

4 CONCLUSIONES

Este artículo presenta una metodología de análisis de imagen para la detección de aceitunas en imágenes de olivo, encaminada al desarrollo de un sistema de estimación de la producción del olivar. Dicha metodología implementa un preprocesamiento dirigido a hacer viable la detección de las aceitunas en las imágenes, bajo cualquier circunstancia de apariencia; mediante una red neuronal convolucional. Los resultados obtenidos muestran fuertes indicios de que el desarrollo presentado es un paso certero en la consecución del objetivo descrito. Además, es esperable que la metodología sea aplicable a otras variedades de aceituna gracias a que, en una fase temprana de maduración como la considerada en este trabajo, las características morfológicas y colorimétricas presentan una variabilidad contenida.

Como trabajo futuro, se pretende automatizar la toma de imágenes con el empleo de robots autónomos, así como mejorar las condiciones de captura con un enfoque de iluminación de mayor rendimiento. A su vez, se diseñarán experimentos de campo dirigidos a poder correlacionar la información extraída por análisis de imagen con valores reales de producción.

Agradecimientos
Esta investigación fue financiada por el Programa de Cooperación INTERREG V-A España-Portugal (POCTEP), y cofinanciada por fondos FEDER.

Los autores también quieren agradecer a la Cooperativa Virgen de la Oliva (Gibraleón, Huelva, Andalucía, España) por generosamente prestar sus olivares para el desarrollo de la experimentación de campo necesaria para culminar este trabajo.

English summary

OLIVE IDENTIFICATION AND COUNTING IN DIGITAL IMAGES TAKEN IN OLIVE ORCHARDS USING MATHEMATICAL MORPHOLOGY AND CONVOLUTIONAL NEURAL NETWORKS

Abstract

Early and accurate yield estimation is a very valued objective for modern agriculture. In the case of oliviculture, it is especially relevant due to the high economic value of its production. This paper presents a methodology aimed at achieving that end. Concretely, it comprises an artificial vision algorithm able to detect those olives that are visible in a digital image of an olive tree, captured directly in the field, at night-time and with artificial illumination. First, the image is preprocessed by means of mathematical morphology techniques and statistical filtering to, from this output, generate a subset of images with high probability of containing an olive. Thus, this preprocessing reduces the search space in a magnitude of 10^3. Next, these subimages are classified by a convolutional neural network as 'olive' or 'discarded'. From a total of 304,483 subimages, extracted from 21 images, the net correctly classified 98.23% of cases, and gave a coefficient of determination R^2 of 0.9875 when facing the number of detected olives to the real one. This achieved accuracy indicates that the found algorithm constitutes a solid step towards the implementation of a future system for early yield estimation of olive orchards.

Keywords: Precision agriculture, yield estimation, olive, artificial vision, convolutional neural network.

Referencias

[1] Food and Agriculture Organization of the United Nations (FAOSTAT). [Online]. Available: http://www.fao.org/faostat/en/#home.

[2] Saulle, R., La Torre, G., (2010) "The Mediterranean Diet, recognized by UNESCO as a cultural heritage of humanity", *Italian Journal of Public Health*, vol. 7(4), pp. 414-415.

[3] Rockström, J., Williams, J., Daily, G., Noble, A., Matthews, N., Gordon, L., Wetterstrand, H., DeClerck, F., Shah, M, Steduto, P., de Fraiture, C., Hatibu, N., Unver, O., Bird, J., Sibanda, L., Smith, J., (2010) "Sustainable intensification of agriculture for human prosperity and global sustainability", *Ambio*, vol. 46(1), pp. 4-17.

[4] Organization of United Nations (ONU). [Online]. Available: https://www.un.org/en/development/desa/news/population/2015-report.html

[5] Zarco-Tejada, P., Hubbard, N., Loudjani, P., (2014) "Precision Agriculture: An Opportunity for EU Farmers—Potential Support with the CAP 2014-2020". *Joint Research Centre (JRC) of the European Commission*. [Online]. Available: http://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL-AGRI_NT%282014%29529049

[6] Orlandi, F., Sgromo, C., Bonofiglio, T., Ruga, L., Romano, B., Fornaciari, M., (2010) "Yield modelling in a Mediterranean species utilizing cause-effect relationships between temperature forcing and biological processes". *Scientia Horticulturae*, vol. 123(3), pp. 412-417.

[7] Aguilera, F., Ruiz-Valenzuela, L. (2014) "Forecasting olive crop yields based on long-term aerobiological data series and bioclimatic conditions for the southern Iberian Peninsula". *Spanish Journal of Agricultural Research*, vol. 12(1), pp. 215-224.

[8] European Commission–Agriculture and Rural Development, (2011) "Agricultural Markets Brief–Brief 1: High commodity price and volatility: what lies behind the roller coaster ride?". [Online]. Available: http://ec.europa.eu/agriculture/analysis/tradepol/commodityprices/market-briefs/01_en.pdf.

[9] Nuske, S., Achar, S., Bates, T., Narasimhan, S., Singh, S., (2011) "Yield estimation in vineyards by visual grape detection". In *Proceedings of the International Conference on Intelligent Robots and Systems (IEEE/RSJ)*. San Fracisco, USA, Sept. 25-30, pp. 2352-2358.

[10] Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Narasimhan, S., Singh, S., (2014) "Automated visual yield estimation in vineyards". *Journal of Field Robotics*, vol. 31(5), pp. 837-860.

[11] Font, D., Tresanchez, M., Martínez, D., Moreno, J., Clotet, E., Palacín, J., (2015) "Vineyard yield estimation based on the analysis of high resolution images obtained with artificial illumination at night". *Sensors*, vol. 15(4), pp. 8284-8301.

[12] Aquino, A., Millan, B., Diago, M.P., Tardaguila, J., (2018) "Automated early yield prediction in vineyards from on-the-go image acquisition". *Computers and electronics in agriculture*, vol. 144, pp. 26-36.

[13] Balasundram, S.K., Memarian, H., Khosla, R., (2013) "Estimating oil palm yields using vegetation indices derived from Quickbird". *Life Sci. J*, vol. 10(4), pp. 851-860.

[14] Chang, K.W., Shen, Y., Chung, L.J., (2005) "Predicting rice yield using canopy reflectance measured at booting stage". *Agronomy Journal* vol. 97, pp. 872-878.

[15] Fang, H., Liang, S., Hoogenboom, G., (2011) "Integration of MODIS LAI and vegetation index products with the CSM-CERES maize model for corn yield estimation". *International Journal of Remote Sensing*, vol. 32, pp. 1039-1065.

[16] Hayes, M.J., Decker, W.L., (1996) "Using NOAA AVHRR data to estimate maize production in the United States corn belt". *International Journal of Remote Sensing*, vol. 17, pp. 3189-3200.

[17] Bargoti, S., Underwood, J., (2017) "Deep fruit detection in orchards". In *Proceedings of the International Conference on Robotics and Automation (IEEE/ICRA)*, Singapore, Jun. 29-3, pp. 3626-3633.

[18] Stein, M., Bargoti, S., Underwood, J., (2016) "Image based mango fruit detection, localisation and yield estimation using multiple view geometry". *Sensors*, vol. 16(11), pp. 1915.

[19] Bargoti, S., Underwood, J.P., (2017) "Image segmentation for fruit detection and yield estimation in apple orchards". *Journal of Field Robotics*, vol. 34(6), 1039-1060.

[20] Fornaciari, M., Orlandi, F., Romano, B., (2005) "Yield forecasting for olive trees". *Agronomy Journal*, vol. 97(6), pp. 1537-1542.

[21] Galán, C., Vázquez, L., Garcia-Mozo, H., Dominguez, E., (2004) "Forecasting olive (Olea europaea) crop yield based on pollen emission". *Field Crops Research*, vol. 86(1), pp. 43-51.

[22] Minero, F.J.G., Candau, P., Morales, J., Tomas, C., (1998) "Forecasting olive crop production based on ten consecutive years of monitoring airborne pollen in Andalusia (southern Spain)". *Agriculture, Ecosystems & Environment*, vol. 69(3), pp. 201-215.

[23] Smith, W.J., (2007) Modern Optical Engineering, 4th ed. The Design of Optical Systems. McGraw-Hill Education – Europe, USA.

[24] Soille, P., (2004) Morphological Image Analysis. Principles and Applications, 2nd ed. Springer – Verlag, Berlin, Germany.

[25] Aquino, A., Diago, M.P., Millán, B., Tardáguila, J., (2017) "A new methodology for estimating the grapevine-berry number per cluster using image analysis". *Biosystems engineering*, vol. 156, pp. 80-95.

[26] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., (1998) "Gradient-based learning applied to document recognition". *Proceedings of the IEEE,* vol. 86(11), pp. 2278-2324.

[27] Krizhevsky, A., Sutskever, I., Hinton, G.E., (2012) "Imagenet classification with deep convolutional neural networks". In *Proceedings of the 25th international conference on neural information processing systems*, Lake Tahoe, Nevada, USA, Dec. 3-6, vol. 1, pp. 1097-1105.

[28] Simonyan, K., Zisserman, A., (2014) "Very deep convolutional networks for large-scale image recognition". *Cornell University arXiv*:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556

[29] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., (2015) "Going deeper with convolutions". In *28th IEEE conference on computer vision and pattern recognition (CVPR)*, Boston, USA, pp. 1–9.

[30] He, K., Zhang, X., Ren, S., Sun, J., (2015) "Deep residual learning for image recognition". *Cornell University arXiv*:1512.03385. [Online]. Available: https://arxiv.org/abs/1512.03385

[31] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., (2017) "Inception-v4, inception-resnet and the impact of residual connections on learning". In *31st AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, California, USA, pp. 4278-4284.

[32] Qian, N, (1999) "On the momentum term in gradient descent learning algorithms". *Neural networks: the official journal of the International Neural Network Society*, 12(1), pp. 145–151.