Revista Iberoamericana de Automática e Informática Industrial

Revista Iberoamericana de Automática e Informática Industrial 15 (2018) 231-242

www.revista-riai.org

Una Revisión Sistemática de Métodos para Localizar Automáticamente Objetos en Imágenes

Deisy Chaves[a,b,*], Surajit Saikia[b], Laura Fernández-Robles[b], Enrique Alegre[b], Maria Trujillo[a]

[a] *Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Ciudad Universitaria Meléndez, Calle 13 No. 100-00, Cali, Colombia.*
[b] *Escuela de Ingenierías Industrial, Informática y Aeronáutica, Universidad de León, Campus de Vegazana, 24071, León, 24071,España.*

Resumen

Actualmente, muchas aplicaciones requieren localizar de forma precisa los objetos que aparecen en una imagen, para su posterior procesamiento. Este es el caso de la inspección visual en la industria, los sistemas de diagnóstico clínico asistido por computador, la detección de obstáculos en vehículos o en robots, entre otros. Sin embargo, diversos factores como la calidad de la imagen y la apariencia de los objetos a detectar, dificultan la localización automática. En este artículo realizamos una revisión sistemática de los principales métodos utilizados para localizar objetos, considerando desde los métodos basados en ventanas deslizantes, como el detector propuesto por Viola y Jones, hasta los métodos actuales que usan redes de aprendizaje profundo, tales como Faster-RCNN o Mask-RCNN. Para cada propuesta, describimos los detalles relevantes, considerando sus ventajas y desventajas, así como sus aplicaciones en diversas áreas. El artículo pretende proporcionar una revisión ordenada y condensada del estado del arte de estas técnicas, su utilidad y sus implementaciones a fin de facilitar su conocimiento y uso por cualquier investigador que requiera localizar objetos en imágenes digitales. Concluimos este trabajo resumiendo las ideas presentadas y discutiendo líneas de trabajo futuro.

Palabras Clave: Algoritmos de detección, Aprendizaje máquina, Procesamiento de imágenes, Reconocimiento de objetos, Reconocimiento de patrones.

A Systematic Review on Object Localisation Methods in Images

Abstract

Currently, many applications require a precise localization of the objects that appear in an image, to later process them. This is the case of visual inspection in the industry, computer-aided clinical diagnostic systems, the obstacle detection in vehicles or in robots, among others. However, several factors such as the quality of the image and the appearance of the objects to be detected make this automatic location difficult. In this article, we carry out a systematic revision of the main methods used to locate objects by considering since the methods based on sliding windows, as the detector proposed by Viola and Jones, until the current methods that use deep learning networks, such as Faster-RCNN or Mask-RCNN. For each proposal, we describe the relevant details, considering their advantages and disadvantages, as well as the main applications of these methods in various areas. This paper aims to provide a clean and condensed review of the state of the art of these techniques, their usefulness and their implementations in order to facilitate their knowledge and use by any researcher that requires locating objects in digital images. We conclude this work by summarizing the main ideas presented and discussing the future trends of these methods.

Keywords: Detection algorithms, Image processing, Machine learning, Object recognition, Pattern recognition.

1. Introducción

La localización o detección automática de objetos tiene como objetivo determinar la ubicación de los objetos de interés en una imagen, si existen (Lampert et al., 2008; Zitnick and Dollar, 2014), siendo este un problema aún abierto en visión por computador. Además, la localización previa es fundamen-

*Autor para correspondencia: deisy.chaves@correounivalle.edu.co
To cite this article: Chaves, D., Saikia, S., Fernández-Robles, L., Alegre, E., Trujillo, M. 2018. A Systematic Review on Object Localisation Methods in Images. Revista Iberoamericana de Automática e Informática Industrial 15, 231-242. https://doi.org/10.4995/riai.2018.10229

tal en el análisis automático de imágenes, cuyo objetivo puede ser la segmentación, que consiste en separar los objetos de interés del fondo, en el reconocimiento o recuperación de objetos (Saikia et al., 2017; García-Olalla et al., 2018), o en el análisis de la relación espacial entre los objetos contenidos en una imagen (Lampert et al., 2008). Actualmente, hay una gran cantidad de aplicaciones que requieren una localización precisa de los objetos, como es el caso de la necesidad que tienen los vehículos autónomos de localizar peatones (Dollár et al., 2009; Li et al., 2018; Du et al., 2017; Brazil et al., 2017; Wang et al., 2018) u obstáculos (Shah et al., 2018; Yi et al., 2018; Garnett et al., 2017; Sepúlveda et al., 2017), la localización de vehículos en sistemas de control de tráfico, se encuentren o no en imágenes áereas (Zhong et al., 2017; Ammour et al., 2017; Tang et al., 2017; Xu et al., 2017a; Lee et al., 2017), la localización de lesiones o anormalidades en tejidos que usan los sistemas de diagnóstico clínico asistido por computador (He et al., 2018; Ma et al., 2017; Jiamin et al., 2017; Sa et al., 2017; Heo et al., 2017), la detección de objetos para el control de calidad que requieren los sistemas de inspección visual (Cao et al., 2018; Chen et al., 2018; Shi et al., 2017; Ferguson et al., 2017), o la localización de obstáculos que tienen que realizar los sistemas de navegación de robots (Lee et al., 2015; Luo et al., 2017), entre otras. Sin embargo, la correcta localización de objetos es difícil debido a múltiples factores, entre los que destacan la falta de calidad de la imagen, condiciones de iluminación cambiantes, objetos con forma no rígida o los cambios en la apariencia de los objetos a localizar (Felzenszwalb et al., 2010; Dalal and Triggs, 2005). Por otra parte, existen diferentes formas de representar la ubicación de un objeto. Las más importantes son (Lampert et al., 2008): las coordenadas del punto central del objeto, el contorno, una región rectangular rodeando al objeto (Viola and Jones, 2004; Girshick et al., 2013), o una máscara procedente de su segmentación (He et al., 2017). Entre ellas, las regiones rectangulares, que se denominan rectángulos circunscritos, constituyen la representación más utilizada debido a su simplicidad para generar anotaciones y predicciones de localización (Lampert et al., 2008).

El problema de localización puede verse como un problema de clasificación en el cual se obtiene un conjunto de regiones en una imagen, que se clasifican indicando si contienen o no el objeto de interés (ver Figura 1). En la solución al problema de localización, algunos métodos emplean esta aproximación mediante el uso de una ventana de tamaño variable, ventana deslizante, que se desplaza sobre una imagen obteniendo una clasificación para cada región (Viola and Jones, 2004). Recientemente —además de algoritmos de clasificación clásicos como las máquinas de soporte vectorial (Dalal and Triggs, 2005; Felzenszwalb et al., 2010)— se están utilizando estrategias de aprendizaje profundo para clasificar regiones (Girshick et al., 2013; Girshick, 2015; Ren et al., 2015; Redmon et al., 2015; Liu et al., 2016; He et al., 2017). Otra forma de abordar el problema es utilizando métodos que optimizan la búsqueda de regiones que contienen objetos en la imagen (Lampert et al., 2008; Hosang et al., 2016).

Este artículo presenta una revisión sistemática de métodos que permiten localizar objetos relevantes, desde los métodos clásicos, basados en la búsqueda de objetos con ventana deslizante, hasta los métodos más recientes, basados en redes de aprendizaje profundo. Para cada método seleccionado, se indican sus ventajas y desventajas, así como dónde se aplican para resolver problemas actuales. En particular, se hace énfasis en cuatro aplicaciones específicas, como son los vehículos autónomos, la imagen médica, la inspección visual en la industria, y la robótica.

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se describen los principales métodos de localización de objetos. La Sección 3 contiene las librerías y conjuntos de datos utilizados para la evaluación de los métodos de localización. En la Sección 4 se presentan diversas aplicaciones en las que se utilizan los métodos descritos anteriormente. Finalmente, en la Sección 5 se bosquejan algunas conclusiones y se discuten líneas de trabajo futuro.

2. Métodos de Localización de Objetos

Hemos agrupado los métodos revisados en tres grandes categorías, en función de la estrategia que utilicen para realizar el análisis de la imagen y la propuesta de regiones, que son: (i) ventana deslizante, (ii) regiones candidatas, y (iii) aprendizaje profundo, como se muestra en la Figura 2. Hemos revisado los métodos más utilizados, porque se considera que funcionan, o funcionaban en el caso de los clásicos, mejor, prestando especial atención a los utilizados en aplicaciones industriales, tales como los vehículos autónomos, la seguridad, la vigilancia y la inspección visual, entre otras. La Tabla 1 contiene un resumen de las ventajas y las desventajas identificadas en cada uno de los métodos revisados. A continuación se presentan los métodos seleccionados, según el criterio comentado anteriormente.

2.1. Estrategias Basadas en Ventana Deslizante

La localización de objetos se realiza empleando una ventana de tamaño variable que se desplaza sobre la imagen. En cada región donde se ubica la ventana deslizante, se evalúa una función que determina si la región contiene o no un objeto de interés, como se ilustra en la Figura 2a. Generalmente, se considera que valores altos de dicha función, con respecto a un umbral, indican la presencia de un objeto. Esta función puede corresponder por ejemplo, a un puntaje (*score*) o valor de clasificación obtenido después de usar un clasificador. Sin embargo, incluso una imagen de baja resolución puede contener miles de propuestas de regiones, haciendo que la búsqueda exhaustiva de objetos sea computacionalmente costosa (Lampert et al., 2008). Por este motivo, se han propuesto diferentes estrategias para optimizar la búsqueda de objetos en la imagen.

VJ. Viola and Jones (2004) emplean una serie de clasificadores en cascada para hacer la búsqueda de regiones más eficiente y rápida. Esta propuesta se basa en centrar el análisis en las regiones con mayor probabilidad de contener objetos de interés, que corresponden en este caso a rostros, al estar orientado este método a su detección. El sistema emplea descriptores de Haar correspondientes a filtros rectangulares verticales u horizontales que codifican las características de los rostros a detectar. Los descriptores se calculan rápidamente empleando una imagen integral, donde cada píxel contiene la suma de los valores acumulados en sus intensidades hacia arriba y hacia la izquierda del píxel. La búsqueda de objetos se realiza con una ventana deslizante, iniciando con regiones de 24×24 píxeles

Figura 1: Esquema general de los pasos a seguir habitualmente para la localización automática de objetos en una imagen.

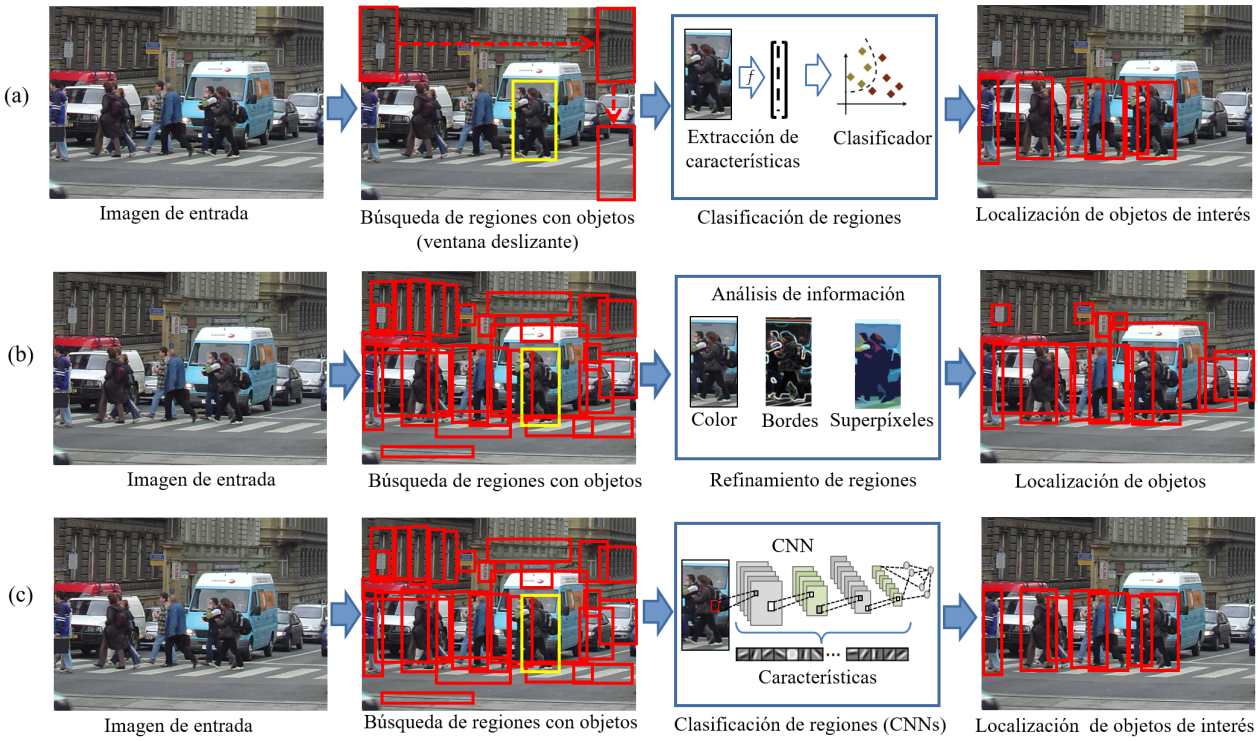


Figura 2: Localización automática de objetos en una imagen con métodos basados en: (a) ventana deslizante, (b) conjuntos de regiones candidatas, (c) aprendizaje profundo. En los casos (a) y (c) los peatones son el objeto de interés.

que incrementan el tamaño usando un factor de 1.5. Cada región es analizada usando una serie de clasificadores en cascada para determinar la presencia de objetos. Los clasificadores iniciales son simples y están construidos para descartar un elevado número de regiones sin objetos, generando un gran número de falsos positivos. Los siguientes clasificadores son más complejos y se especializan en distinguir las regiones que puedan contener objetos en el conjunto de regiones detectadas por los clasificadores iniciales. Todos los clasificadores se entrenan empleando el algoritmo AdaBoost (Freund and Schapire, 1999) a partir de un conjunto de clasificadores débiles, llamados así porque su precisión, de forma individual, durante la clasificación es baja. Finalmente, dado que el detector es invariante a pequeños cambios de escala y traslación, se detectan múltiples regiones alrededor de cada rostro, siendo necesario combinar las regiones superpuestas, en una sola. Para ello se agrupan las regiones en conjuntos disjuntos (dos regiones están en un mismo conjunto si sus coordenadas se superponen) y se promedian las coordenadas de las esquinas de las regiones por conjunto para generar la localización final de los objetos en las imágenes. En este método, rotaciones mayores a 45 grados, cambios en la iluminación y oclusión en el rostro, principalmente de los ojos y la boca, afectan a una correcta detección de los objetos.

HOG+SVM. Dalal and Triggs (2005) proponen un descriptor basado en información del gradiente para identificar peatones en condiciones de iluminación variables. Inicialmente se utiliza una función de corrección gamma para reducir los efectos de posibles variaciones en la iluminación. Posteriormente, la imagen se divide en pequeñas regiones o celdas, y se calculan histogramas acumulados de gradientes u orientaciones de bordes. La distribución de las intensidades de los gradientes, o sus orientaciones, permite caracterizar localmente los contornos, brindando información de forma y textura de los objetos. La combinación de estos histogramas corresponde a una primera representación de histogramas de orientaciones. Posteriormente, se realiza una normalización de contraste local para hacer el descriptor robusto a cambios en la iluminación y sombras. Para ello, se agrupan varias celdas en bloques y se calcula una medida de energía, que es utilizada para normalizar los histogramas de todas las celdas que lo conforman. El descriptor de bloques normalizado es conocido como Histograma de Orientaciones de Gradientes (HOG). Finalmente, los bloques, descriptores HOG, son clasificados usando una máquina de soporte vectorial (SVM) (Boser et al., 1992).

DPM. Felzenszwalb et al. (2010) desarrollan un método para localizar objetos con una gran variabilidad en su apariencia y para ello consideran el conjunto de partes que componen los objetos en su representación. En particular, se realiza una búsqueda de ventana deslizante y cada región se representa empleando una estructura jerárquica de partes o filtro raíz, que esta conformado por un conjunto de filtros de las partes que componen los objetos y modelos asociados a sus deformaciones. El descriptor HOG se emplea como filtro. Los objetos se representan como una mezcla de modelos de partes deformables y son clasificados usando SVM con variables latentes.

ESS. Lampert et al. (2008) plantean optimizar la búsque-

Tabla 1: Ventajas y desventajas de algunos métodos para la localización de objetos en imágenes.

Método	Autores	Ventajas	Desventajas
VJ	(Viola and Jones, 2004)	El descriptor en cascada reduce el tiempo de análisis de las imágenes sin afectar significativamente la precisión.	No robusto a cambios de rotación, ni oclusión.
HOG+SVM	(Dalal and Triggs, 2005)	Robusto a problemas de iluminación.	El descriptor HOG se ve afectado por los cambios en el tamaño de los objetos.
DPM	(Felzenszwalb et al., 2010)	Considera las partes que componen los objetos para su localización, permite ubicar objetos con diferentes poses.	El tiempo requerido para detectar los objetos está determinado por el número de partes consideradas para modelarlos. Modelos más complejos requieren un mayor tiempo de procesamiento.
ESS	(Lampert et al., 2008)	El uso de un algoritmo de ramificación y poda permite la identificación rápida de regiones que pueden contener objetos.	La función de clasificación debe poseer un límite superior.
Objectness	(Alexe et al., 2010)	El uso de información de color, bordes y superpíxeles en la imagen permite determinar de forma rápida un conjunto de regiones que puede contener objetos.	Es necesario procesar el conjunto de regiones obtenidas para localizar los objetos de interés, empleando por ejemplo un clasificador.
Selective Search	(Uijlings et al., 2013)	No requiere del aprendizaje de parámetros para localizar el conjunto de regiones que pueden contener objetos.	Es necesario procesar el conjunto de regiones obtenidas para localizar los objetos de interés.
Edge Boxes	(Zitnick and Dollar, 2014)	La información de bordes en la imagen permite la generación de un conjunto de regiones candidatas de forma rápida y eficiente.	Es necesario fijar dos parámetros — α y β — para la generación del conjunto de regiones candidatas. Adicionalmente, las regiones obtenidas deben ser procesadas para localizar los objetos de interés.
RCNN	(Girshick et al., 2013)	Las características obtenidas con CNNs permiten una mejor representación de los objetos que se refleja en valores altos de precisión durante la detección de los objetos.	Deben entrenarse tres modelos diferentes para la detección de los objetos. Además, RCNN tiene un alto coste computacional, dado que es necesario extraer las características de forma individual para cada una de las regiones candidatas de contener objetos. Se requieren 50 segundos, en promedio, para analizar una imagen.
Fast-RCNN	(Girshick, 2015)	El cálculo de las características de CNN se realiza en una sola iteración, logrando que la detección de objetos sea 25 veces más rápida que el método RCNN (requiere 20 segundos en promedio para analizar una imagen).	El uso de un generador de regiones candidatas externo crea un cuello de botella en el proceso de detección.
Faster-RCNN	(Ren et al., 2015)	El método de RPN permite que la detección de objetos pueda ser casi en tiempo real, aproximadamente 0.12 segundos por imagen.	A pesar de la eficiencia del algoritmo, no es lo suficientemente rápido para ser usado en aplicaciones que requieran tiempo real, como serían los vehículos autónomos.
Mask-RCNN	(He et al., 2017)	La localización de los objetos es más precisa, al realizar una segmentación de los objetos en las imágenes.	Su tiempo de ejecución es mayor al empleado por el método de Faster-RCNN, que extiende. Por lo tanto, tampoco puede ser empleado en aplicaciones que requieren tiempo real.
YOLO	(Redmon et al., 2015)	La localización de los objetos es muy eficiente, permitiendo su uso en aplicaciones en tiempo real.	El método tiene dificultades para detectar correctamente objetos pequeños.
SSD	(Liu et al., 2016)	El uso de una sola red, hace que la localización de los objetos sea más rápida que los métodos Fast-RCNN y Faster-RCNN.	La precisión de la detección de los objetos es menor en comparación con los métodos Fast-RCNN y Faster-RCNN.

da con ventanas deslizantes usando un método conocido como Búsqueda Eficiente en Subventanas (ESS, del inglés *Efficient Subwindow Search*). ESS es un algoritmo de ramificación y poda que permite maximizar eficientemente un conjunto de funciones de clasificación aplicadas sobre todos los posibles subconjuntos de imágenes en una región dada. Esta estrategia converge a un óptimo global siempre y cuando se pueda definir para la función de clasificación, una función que acote el valor máximo para un conjunto de regiones.

2.2. *Estrategias Basadas en Conjuntos de Regiones Candidatas*

A pesar de las mejoras en la búsqueda de regiones basadas en ventana deslizante, es necesario analizar un gran número de regiones, incrementadas por la necesidad de analizar múltiples escalas y considerar diferentes relaciones de aspecto en los tamaños de los objetos a localizar. Adicionalmente, utilizar clasificadores más complejos, que permiten mejorar la calidad de las detecciones, incrementa el tiempo requerido para analizar cada región de la imagen. En este escenario, surgen los métodos de propuesta de regiones candidatas, especializados en generar de forma rápida y eficiente un conjunto de regiones que tienen mayor probabilidad de contener objetos, bajo la hipótesis de que todos los objetos de interés en las imágenes comparten características comunes que los diferencian del fondo de las mismas, ver Figura 2b. Estos métodos generan un conjunto de ventanas candidatas mucho menor al conjunto generado usando ventanas deslizantes. Generalmente, se utilizan clasificadores más complejos, en comparación con los que se usan con ventanas deslizantes, mejorando los tiempos de ejecución sin afectar la precisión en la detección (Hosang et al., 2016). A continuación presentamos algunos de los principales métodos que se utilizan para generar el conjunto de regiones candidatas.

Objectness. Alexe et al. (2010) presentan este método, que es uno de los pioneros en la generación de regiones candidatas. Objectness genera una métrica que permite determinar si una región en una imagen contiene o no un objeto, el cual puede pertenecer a cualquier clase. Para ello se identifica un conjunto inicial de regiones empleando una medida de la prominencia de los objetos (*salency*) en la ventana analizada. Posteriormente, a cada ventana se le asigna un puntaje, empleando una medida para cuantificar una serie de condiciones que indican la presencia de un objeto. Dichas condiciones incluyen el análisis dentro de una región de: (i) el contraste de la información de color, (ii) la concentración de la información de bordes, (iii) el número de superpíxeles (i.e. pequeñas regiones en una imagen que poseen un color o textura similar) contenidos completamente en dicha región.

Selective Search. Uijlings et al. (2013) proponen un método donde se agrupan superpíxeles para generar el conjunto de regiones candidatas, empleando medidas de similitud para determinar qué superpíxeles combinar. Los superpíxeles del con-

junto inicial se identifican aplicando un algoritmo de segmentación basado en grafos (Felzenszwalb and Huttenlocher, 2004). En particular, para combinar los superpíxeles se consideran cuatro medidas de similitud basadas en: (i) información de color, (ii) información de textura empleando una versión rápida del descriptor SIFT, (iii) tamaño de las regiones dando prioridad a la combinación temprana de pequeñas regiones, y (iv) forma de las regiones, para determinar si una región está contenida en otra y combinarlas con el fin de evitar huecos en las regiones identificadas.

Edge Boxes. Zitnick and Dollar (2014) desarrollan un método que inicia la búsqueda de regiones candidatas siguiendo una estrategia de ventana deslizante, en la cual se evalúa una función de puntaje basada en la información de los bordes. Esta función calcula la diferencia entre el número de bordes completamente contenidos en la región analizada y el número de bordes que atraviesan el borde de dicha región. Los bordes se obtienen empleando el método de detección de bordes estructurado (Dollar and Zitnick, 2013). El tamaño de la ventana deslizante se determina usando un parámetro α, que corresponde a la *Intersección sobre la Unión* (IoU, del inglés *Intersection-over-Union*) de las regiones vecinas. Una vez realizada la búsqueda de objetos, se refinan todas las regiones identificadas para mejorar su ubicación. Primero, se ordenan las regiones usando los valores obtenidos con una función de puntaje, y luego se eliminan las regiones cuyo IoU es mayor que un umbral, β, en comparación con la región con un valor de puntaje mayor (supresión de valores no máximos).

2.3. Estrategias Basadas en Aprendizaje Profundo

Los descriptores de características visuales tradicionales como HOG y Haar, utilizados en los métodos mencionados anteriormente, tienen limitaciones para describir y representar las características de las imágenes naturales (Deng and Yu, 2014). Sin embargo, los métodos de aprendizaje profundo (*deep learning*) aprenden directamente de los datos (i.e píxeles en las imágenes) considerando diferentes niveles de abstracción (ver Figura 2c). En particular, las Redes Neuronales Convolucionales (CNN, del inglés *Convolutional Neural Network*) son un tipo de red de aprendizaje profundo especializada en el análisis de imágenes, la cual ha mostrado un buen rendimiento en problemas de reconocimiento, detección y segmentación de objetos (Krizhevsky et al., 2012; He et al., 2016; Szegedy et al., 2017). Cuantas más capas de profundidad tenga la red neuronal, mejor será la posibilidad de aprender estructuras más complejas en las imágenes. Esta abstracción jerárquica se alcanza con el apilamiento de capas de funcionamiento sencillo (Lecun et al., 2015), donde las capas iniciales ayudan a identificar características de bajo nivel, como esquinas y bordes, mientras que las capas finales contienen detalles más significativos de los datos, como un patrón específico de forma o textura conformado por características de bajo nivel (Deng, 2014). De forma similar a los sistemas de clasificación tradicional, las características se calculan usando un conjunto de filtros de convolución presentes en las capas de la red, y se refinan durante su entrenamiento, permitiendo predecir una clase para una imagen dada. El uso de estrategias basadas en aprendizaje profundo se ha incrementado desde 2012 al igual que el desarrollo de arquitecturas más sofisticadas (Simonyan and Zisserman, 2014; He et al., 2016). A continuación presentamos las estrategias basadas en aprendizaje profundo más usadas para la localización de objetos.

RCNN. Girshick et al. (2013) proponen un método de localización y detección de objetos en imágenes, usando características enriquecidas obtenidas con CNNs. El método propuesto genera inicialmente un conjunto de regiones candidatas con el método Selective Search. Posteriormente, cada región se redimensiona a un tamaño fijo para entrenar una CNN con la arquitectura Alexnet (Krizhevsky et al., 2012). Alexnet extrae las características correspondientes a cada una de las regiones candidatas las cuales se clasifican para determinar si contienen ó no un objeto de interés. La localización de las regiones que contienen objetos se refina usando un modelo de regresión lineal que ajusta las coordenadas del área rectangular identificada. Este método tiene altos valores de precisión. Sin embargo, para su ejecución se analizan al menos 2000 regiones candidatas por imagen teniendo un alto coste computacional.

Fast-RCNN. Girshick (2015) desarrollan un método para crear una versión de RCNN más rápida y eficiente. El análisis de las regiones candidatas no se hace de forma individual, en su lugar, se comparten los cálculos realizados para todo el conjunto de regiones candidatas. De esta forma, en vez de usar el clasificador para detectar objetos en cada una de las 2000 regiones candidatas consideradas por imagen, se determina la presencia de objetos en todas las regiones en una sola iteración empleando una técnica conocida como Conjunto de Regiones de Interés (*Region of Interest Pooling*). Esta técnica comparte los cálculos realizados para cada subregión en la imagen analizada, haciendo posible analizar una imagen con una sola iteración en lugar de múltiples iteraciones. Las características para cada una de las regiones se obtienen del correspondiente mapa de características.

En Fast-RCNN, la red CNN y la regresión de las regiones candidatas se entrenan conjuntamente como un solo modelo de red. Mientras que en RCNN se obtienen tres modelos que se entrenan de forma independiente (i.e. CNN para la extracción de características, SVM para determinar si las regiones contienen o no objetos de interés, y un modelo de regresión lineal para refinar la localización de las regiones que contienen objetos). Adicionalmente, en Fast-RCNN, el clasificador de regiones entrenado con el algoritmo SVM se reemplaza por una capa adicional en la red CNN que emplea una función exponencial normalizada o softmax para la clasificación. Paralelamente a la capa softmax se emplea una capa encargada de realizar la regresión lineal de las coordenadas de las regiones y generar la localización final. En resumen, Fast-RCNN se compone de una sola red y, para cada objeto identificado, permite obtener la ubicación de la región rectangular circunscrita que lo contiene.

Faster-RCNN. Los métodos RCNN y Fast-RCNN dependen de métodos externos como Selective Search para generar el conjunto de regiones candidatas inicial a analizar por imagen. Este proceso constituye un cuello de botella que afecta el tiempo requerido para la detección de los objetos. Por este motivo, Ren et al. (2015) proponen el método Faster-RCNN, con el objetivo de hacer más eficiente y rápido el proceso de generación de regiones candidatas. Para ello desarrollan un sistema de propuesta de regiones candidatas (RPN, en inglés *Region Proposal System*) a partir del mapa de características generado por

la red convolucional durante la primera iteración de la detección de objetos. El uso de este sistema evita realizar cálculos adicionales al usar métodos de generación de regiones candidatas. Para manejar las variaciones en aspecto y escala de los objetos, Faster-RCNN introduce la idea de regiones de tamaño prefijado. En cada ubicación se emplean 3 diferentes regiones de tamaño prefijado para las escalas 128×128, 256×256 y 512×512 píxeles. De forma similar, se consideran las diferencias de tamaño en el aspecto de los objetos mediante 3 relaciones de aspecto $1:1$, $2:1$ y $1:2$. Así, por cada ubicación se tienen 9 regiones en las cuales RPN predice la probabilidad de que contengan un objeto o fondo. Cada región se refina mediante regresión lineal. RPN da como resultado un conjunto de regiones candidatas y la probabilidad de que contengan un objeto de interés.

YOLO. Redmon et al. (2015) crean un método para la detección de objetos empleando CNNs en tiempo real. YOLO (del inglés *You Only Look Once*) no emplea regiones candidatas para detectar los objetos. En su lugar, utiliza una única CNN para realizar la localización de los objetos con regiones rectangulares circunscritas y su posterior clasificación. La localización de los objetos se aborda como un problema de regresión simple, donde para cada imagen de entrada, se aprenden las probabilidades de que contenga una clase con su respectiva ubicación, representada como coordenadas de una región rectangular. La imagen se divide en una cuadrícula de tamaño fijo, de $M \times M$ píxeles, y para cada cuadrícula se predicen un conjunto de regiones y la probabilidad de que contengan un objeto determinado o no. Dicho valor de probabilidad refleja la precisión de la localización de una región que contiene un objeto. Durante la localización de objetos, la CNN sólo requiere ejecutarse una vez por imagen, haciendo que este proceso se realice en tiempo real. A diferencia de los métodos mencionados anteriormente, como RCNN, YOLO analiza la imagen en una única iteración empleando información de contexto que evita la identificación de regiones con falsos positivos.

SSD. Liu et al. (2016) proponen el método conocido como SSD (del inglés *Single Shot Detector*), para la detección de objetos teniendo un balance entre la precisión de los resultados y el tiempo requerido para obtenerlos. Para ello, a partir de una imagen, se genera un mapa de características con una CNN empleando una iteración. Posteriormente, se predicen las regiones rectangulares y la probabilidad de que contengan objetos, utilizando un kernel de convolución de tamaño 3×3 píxeles sobre el mapa de características obtenido. De manera similar al método de Faster-RCNN, SSD emplea diferentes tamaños predefinidos para escalas con diferentes tamaños de aspecto. Dado que cada capa en la red convolucional realiza cálculos a diferentes escalas, la red puede predecir regiones rectangulares y detectar objetos en diferentes escalas.

Mask-RCNN. He et al. (2017) presentan un método que extiende Faster-RCNN para segmentar los objetos en las imágenes en lugar de solo identificar las regiones rectangulares que los contienen. Mask-RCNN predice la localización exacta de cada uno de los píxeles en las imágenes que corresponden a objetos empleando una máscara binaria. La máscara corresponde a una CNN totalmente conectada que a partir de mapas de características obtenidos con una CNN, genera una matriz binaria, en la que el valor 1 indica que dicho píxel pertenece a un objeto. Para mejorar la precisión, los autores emplean una estrategia de refinamiento de las posiciones de las regiones identificadas.

3. Software y Conjuntos de Datos

La Tabla 2 presenta algunas de las implementaciones de software libre más populares de los métodos de detección de regiones descritos anteriormente. Para cada implementación se indican los autores, el lenguaje de programación utilizado, así como la URL donde está disponible. Además, la Tabla 3 muestra los conjuntos de datos más utilizados en la literatura referente a este ámbito de estudio.

4. Aplicaciones

El desarrollo de sistemas automáticos que apoyen diferentes tareas es uno de los grandes retos de la computación. En este contexto están los sistemas de detección de obstáculos en vehículos autónomos tales como peatones, ciclistas u otros vehículos, los sistemas de apoyo al diagnóstico de enfermedades o los sistemas de inspección de calidad, entre otros. A continuación presentamos algunas aplicaciones que hacen uso de los métodos revisados en la Sección 2 para localizar objetos en imágenes.

4.1. Vehículos Autónomos

Se espera que en el futuro los vehículos autónomos contribuyan a incrementar la seguridad en las carreteras y así disminuir el número de accidentes de tráfico. Es posible que también ayuden a reducir las congestiones de tráfico y la contaminación en las ciudades, al mejorar la eficiencia en la conducción y el uso del combustible. Por todo ello, el desarrollo de tecnologías para vehículos autónomos es cada vez más importante en la industria automotriz. Sin embargo, estas tecnologías aún deben ser perfeccionadas. Las ciudades y autopistas son entornos muy dinámicos e impredecibles, donde actúan múltiples actores, como peatones, animales u otros vehículos. En este sentido, es necesario proveer sistemas robustos a los vehículos autónomos que permitan detectar correctamente objetos en tiempo real, para analizar su entorno y actuar en consecuencia.

4.1.1. Detección de Peatones

La detección de peatones es uno de los problemas más desafiantes e importantes en el desarrollo de vehículos autónomos, donde se deben considerar diferentes posiciones y vestimentas de los peatones, así como los problemas producidos por posibles oclusiones, iluminación y fondos variables. La detección de peatones es un problema abierto en visión por computador, con aplicaciones en vehículos autónomos, vigilancia y robótica.

Los primeros métodos desarrollados para abordar este problema fueron VJ y HOG+SVM. Desde entonces, el uso de clasificadores entrenados con el algoritmo de AdaBoost ha sido clave para la detección de peatones. En particular, ICF (Dollár et al., 2009), que se basa en VJ, es uno de los métodos clásicos más utilizados para la detección de peatones. Este detector emplea clasificadores entrenados con AdaBoost y un conjunto de características jerárquicas obtenidas para los distintos canales de la imagen (e.g. en el espacio de color YUV, el descriptor

Tabla 2: Información sobre implementaciones de software libre de métodos de localización de objetos.

Método	Autores	Lenguaje de programación	URL
VJ	Colaboradores en el desarrollo de OpenCV	Python, C++ / Librería OpenCV	https://docs.opencv.org/2.4/modules/objdetect/doc/cascade_classification.html
HOG+SVM	Colaboradores en el desarrollo de scikit-image	Python / Toolbox scikit-image	http://scikit-image.org/docs/dev/auto_examples/features_detection/plot_hog.html
DPM	Ross Girshick	Matlab	https://github.com/rbgirshick/voc-dpm
ESS	Christoph Lampert	C++	https://github.com/npinto/ESS
Objectness	Bogdan Alexe, Thomas Deselaers, Vittorio Ferrari	Matlab	http://groups.inf.ed.ac.uk/calvin/objectness
Selective Search	J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, A.W.M. Smeulders	Matlab	https://www.koen.me/research/selectivesearch
Edge Boxes	Piotr Dollar	Matlab	https://github.com/pdollar/edges
RCNN	Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik	Matlab / Framework caffe	https://github.com/rbgirshick/rcnn
Fast-RCNN	Ross Girshick	Python / Framework caffe	https://github.com/rbgirshick/fast-rcnn
Faster-RCNN	Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun	Matlab / Framework caffe	https://github.com/ShaoqingRen/faster_rcnn
Mask-RCNN	División de Investigación Facebook AI	Python / Framework caffe	https://github.com/facebookresearch/Detectron
YOLO	Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi	Librería OpenCV / CUDA	https://github.com/pjreddie/darknet/wiki/YOLO:-Real-Time-Object-Detection
SSD	Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg.	Python / Framework caffe	https://github.com/weiliu89/caffe/tree/ssd

Tabla 3: Información sobre conjuntos de datos disponibles para métodos de localización de objetos.

Conjunto de datos	Clase de Objetos	URL
CMU/MIT	Rostros	http://vasc.ri.cmu.edu/idb/images/face/frontal_images/images.tar
INRIA	Personas	http://pascal.inrialpes.fr/data/human/INRIAPerson.tar
UIUC	Autos	http://l2r.cs.uiuc.edu/~cogcomp/Data/Car/CarData.tar.gz
PASCAL VOC	20 clases de objetos de contexto general, en las categorías: personas, vehículos, animales y objetos de interiores	http://host.robots.ox.ac.uk/pascal/VOC/index.html
MS COCO	80 clases de objetos de contexto general con bordes definidos (i.e. vehículos) y 91 clases de objetos sin bordes bien definidos (i.e. césped)	http://cocodataset.org

HOG), que han sido calculadas eficientemente mediante el uso de imágenes integrales.

Recientemente, dado el éxito de las estrategias de aprendizaje profundo como RCNN para la detección de objetos en general, se han desarrollado una serie de métodos que realizan el procesamiento en dos etapas. Inicialmente, se detectan regiones candidatas, y a continuación se utilizan CNNs para la extracción de características y la detección de los peatones. Los métodos SA-FastRCNN (Li et al., 2018) y MS-CNN (Cai et al., 2016) abordan el problema de detección de peatones en diferentes escalas usando redes multi-escala que se integran, respectivamente, en los métodos Fast-RCNN y Faster-RCNN. El método RPN-BF (Zhang et al., 2016) muestra que la red para la identificación de regiones candidatas RPN de Faster-RCNN funciona bien como detector de peatones. Sin embargo, el clasificador empleado en Faster-RCNN no obtiene resultados óptimos dado que no identifica correctamente peatones en una escala pequeña. Por este motivo, RPN-BF emplea mapas de características con mayor resolución y reemplaza el clasificador por un conjunto de clasificadores potenciados, en este caso, *bosted forest*. F-DNN (Du et al., 2017) usa una modificación de Faster-RCNN, en la cual emplea como generador de regiones candidatas el método SSD y como clasificador una combinación en paralelo de múltiples clasificadores de aprendizaje profundo, entre los que se encuentran RestNet50 (He et al., 2016) y GoogleNet (Szegedy et al., 2017).

SDS-RCNN (Brazil et al., 2017) emplea RPN para generar un conjunto de regiones iniciales que se refinan empleando una red de clasificación binaria. SDS-RCNN integra una capa a la CNN que codifica los resultados de una segmentación semántica con el fin de mejorar la identificación de regiones con peatones. Finalmente, PCN (Wang et al., 2018) emplea información de las partes del cuerpo de los peatones e información de contexto para su detección. En particular, este método usa una CNN que consta de dos partes: la primera, modela las partes del cuerpo empleando un módulo de memoria a corto plazo (LSTM, del inglés *Long Short-Term Memory*) para integrar información semántica; y la segunda, considera múltiples escalas para integrar información de contexto.

En la primera fila de la Figura 3 se muestra la identificación de peatones empleando el método Mask-RCNN.

4.1.2. Detección de Obstáculos

Los obstáculos en la vía pueden provocar accidentes de tráfico y por tanto, tienen un impacto en la seguridad de los conductores, así como en el tráfico de vehículos. Es fundamental el desarrollo de sistemas que permitan, en tiempo real, la correcta detección de obstáculos, tanto estáticos como en movimiento, en las vías así como en las aceras (e.g. señales de construcción, animales, niños). Sin embargo, la presencia de sombras, entornos cambiantes y objetos en movimiento, como son los vehículos adelantando, dificultan esta tarea. Existen varias propuestas que emplean sensores especializados, como Lidar (del inglés, *LIght Detection And Ranging*), que tienen como objetivo la conducción totalmente autónoma. No obstante, debido al alto coste de estos sistemas y la mejora en la detección de objetos mediante visión computacional, en la actualidad se han desarrollado más sistemas para la detección de objetos en vehículos automáticos mediante el análisis de imágenes basados en estrategias de ventana deslizante (Shah et al., 2018; Yi et al., 2018; Sepúlveda et al., 2017) y de aprendizaje profundo (Garnett et al., 2017; Levi et al., 2015a).

Shah et al. (2018) proponen una estrategia para la detección de obstáculos en la vía que utiliza una cámara de vídeo fijada al vehículo. El sistema se basa en VJ y emplea una serie de clasificadores en cascada para detectar los obstáculos en cada fotograma del vídeo. Por otra parte, Yi et al. (2018) usan inicialmente el algoritmo MSER (Donoser and Bischof, 2006) para identificar un conjunto inicial de regiones con obstáculos

en imágenes adquiridas con una cámara monocular. Estas regiones se analizan mediante un conjunto de clasificadores en cascada para detectar la presencia o no de obstáculos. Siguiendo otro enfoque, Garnett et al. (2017) proponen un sistema basado en SSD con el fin de detectar obstáculos estáticos y en movimiento. Consideraron como obstáculos los objetos verticales con tamaño mayor al bordillo de una acera. Se modifica la red StixelNet (Levi et al., 2015b), que analiza columnas en la imagen, para considerar dos casos específicos: (i) obstáculo cercanos que se encuentra ocluidos parcialmente. (ii) regiones que no contienen ningún tipo de obstáculo. Posteriormente, se aplica SSD para detectar cuatro categorías de objetos: bicicletas, vehículos, peatones y fondo.

En la segunda y tercera fila de la Figura 3 se ilustra la identificación de diferentes obstáculos con el método Mask-RCNN.

4.1.3. Detección de Vehículos en Imágenes Áreas de Drones y Avionetas

Los vehículos aéreos no tripulados (UAV, del inglés *Unmanned Aerial Vehicles*) como los drones y avionetas, tienen un gran potencial para el transporte o la vigilancia. Los UAV son portátiles, de bajo coste y tamaño, permitiendo recolectar datos de forma rápida sobre el tráfico incluso en áreas de difícil acceso geográfico. El análisis de los datos obtenidos permiten mejorar el tráfico, así como el monitoreo de emergencia en las vías, que son fundamentales para el desarrollo de sistemas inteligentes de transporte. En particular, uno de los retos en los sistemas de monitoreo de tráfico basados en UAV, es la detección de vehículos, debido a: las condiciones de iluminación variable, los movimientos en el fondo causados por el movimiento del UAV y las distintas condiciones de tráfico con alta o baja congestión. En la literatura existen varios métodos basados en estrategias de ventana deslizante (Xu et al., 2017b, 2016) así como aprendizaje profundo (Zhong et al., 2017; Ammour et al., 2017; Tang et al., 2017; Xu et al., 2017a; Lee et al., 2017; Coifman et al., 2006).

Xu et al. (2017b) adaptan el método de VJ para identificar vehículos en imágenes aéreas obtenidas con drones del tipo quadcopter. Dado que VJ es sensible a las rotaciones de los objetos, no identifica correctamente los vehículos en las imágenes de UAV bajo rotaciones. Para conseguir invarianza a la rotación, se realiza un preprocesamiento para corregir la orientación de las imágenes, en el cual se mide la orientación de las avenidas empleando un detector de segmentos, con base en dicha orientación se alinean horizontalmente los vehículos y vías. Posteriormente, se emplea el método de VJ para detectar los vehículos. Adicionalmente, Xu et al. (2016) proponen una estrategia que combina VJ y HOG+SVM, mejorando el tiempo de análisis de vehículos en vídeos obtenidos con quadcopter. Inicialmente, se alinea cada fotograma del vídeo y a continuación se selecciona una estrategia (VJ o HOG+SVM) para detectar los vehículos. La selección se realiza en función del tiempo empleado por cada estrategia para analizar un fotograma y el número de vehículos detectados. Por otra parte, Ammour et al. (2017) inspirados por RCNN, emplean el algoritmo de mean-shift para identificar las regiones candidatas que son posteriormente descritas usando características de una red pre-entrenada con la arquitectura VGG16 (Simonyan and Zisserman, 2014). Finalmente, se emplea SVM para determinar qué regiones contienen o no vehículos. De forma similar, inspirados por Faster-RCNN, Tang et al. (2017) proponen el método HRPN, para generar regiones candidatas, en el cual se emplean mapas de características enriquecidos con el fin de mejorar la identificación de objetos pequeños. Además, como clasificador se utiliza un conjunto de clasificadores en cascada entrenados con AdaBoost. El método se evaluó con un conjunto de imágenes áreas de Munich.

4.2. Diagnóstico Clínico Asistido por Computador

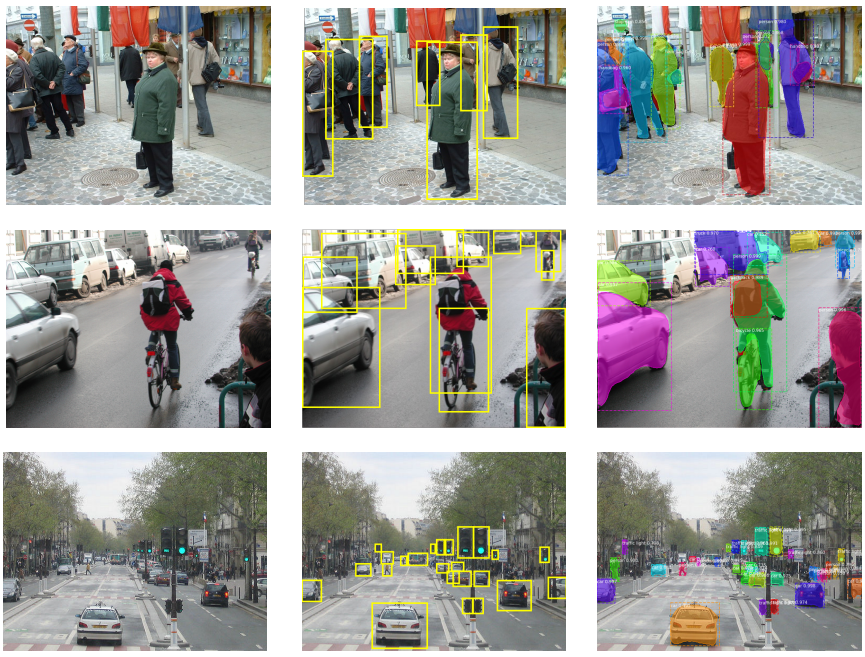
El diagnóstico asistido por computador (CAD, del inglés *Computer Aided Diagnosis*) tiene como fin dar soporte al diagnóstico realizado por un especialista. Una tarea clave en este tipo de sistemas es la detección de lesiones u objetos en las imágenes, que es una de las labores que más tiempo y dedicación requiere por parte de los médicos especialistas. La tarea consiste en la localización e identificación de pequeñas lesiones o regiones anormales en una imagen, perteneciente a un órgano o tejido, con el fin de determinar si el paciente sufre de una enfermedad ó para evaluar la efectividad de un tratamiento determinado. A diferencia de las estrategias desarrolladas para la detección de objetos en vehículos autónomos donde se analizan imágenes en entornos naturales, en los sistemas CAD es necesario considerar las características de diferentes tipos de imágenes, adquiridas en condiciones más controladas y con equipos especializados como por ejemplo: resonancia magnética, tomografía axial computarizada, ultrasonido y microscopía (patología), entre otros. Muchos de estos métodos se inspiran en estrategias de detección de objetos con ventana deslizante (Říha et al., 2013; Tek, 2013; Jung et al., 2013; Dženan et al., 2014; He et al., 2018), otros en conjuntos de regiones candidatas (Cireşan et al., 2013; Ma et al., 2017), y recientemente, en estrategias de aprendizaje profundo (Jiamin et al., 2017; Akselrod-Ballin et al., 2016; Kisilev et al., 2016; Sa et al., 2017; Yang et al., 2017; Heo et al., 2017)

Říha et al. (2013) adaptan VJ para detectar el corte transversal de la arteria carótida en imágenes de ultrasonido obtenidas en modo B simple y para la descripción de las regiones emplean las características de Haar y el coeficiente de Matthews. Además, durante el entrenamiento del detector en cascada se utiliza el algoritmo de AdaBoost con una estrategia evolutiva. Por otra parte, Sa et al. (2017) comparan el desempeño de Faster-RCNN y el método basado en ventana deslizante, HOG+SVM, en la detección de una región lumbar específica en imágenes de Rayos-X. En particular, se aplica Faster-RCNN con una CNN de arquitectura ZF (Zeiler and Fergus, 2014) para la extracción de características. Los resultados muestran que Faster-RCNN presenta una mejor precisión en la localización en comparación con HOG+SVM.

En la primera fila de la Figura 4 se ilustra la identificación de vértebras en la columna vertebral empleando el método de Edge Boxes en imágenes de resonancia magnética.

4.3. Inspección de Calidad

El control de calidad es una de las tareas más importantes en los procesos de fabricación modernos, que influye en la competitividad de las industrias debido a que cada día se desea aumentar la producción manteniendo unos estándares de calidad. Debemos tener en cuenta el alto coste de la inspección manual



(a) Imagen original (b) Localización manual (c) Localización con Mask-RCNN

Figura 3: Localización manual *vs* los resultados obtenidos con el método Mask-RCNN. (a) Imagen original correspondiente a la identificación de diferentes obstáculos en vehículos autonómos, tomadas de los conjuntos de datos INRIA y MS COCO. (b) Localización manual de objetos de interés con rectángulos amarillos. (c) Localización automática de objetos de interés usando Mask-RCNN con rectángulos y su segmentación.

donde se pueden presentar errores debido a la fatiga visual. Actualmente, se han desarrollado sistemas de visión computacional que permiten examinar automáticamente la apariencia visual de los productos en las líneas de producción. Muchos de estos sistemas usan estrategias de aprendizaje profundo (Cao et al., 2018, 2016; Park et al., 2016; Chen et al., 2018; Shi et al., 2017; Ferguson et al., 2017). Por otra parte, se han desarrollado sistemas que se enfocan en la monitorización de las condiciones de las herramientas utilizadas durante la producción, con el fin de identificar de forma temprana los fallos en el sistema (García-Ordás et al., 2017; Fernández-Robles et al., 2017a,b). Estos sistemas deben ser precisos, robustos y eficientes.

Ferguson et al. (2017) presentan una comparación de cuatro estrategias (ventana deslizante, Faster-RCNN, R-FCN (Dai et al., 2016), SSD) para la detección de defectos generados durante la fundición de piezas metálicas que afectan la calidad del producto final. Esta es una tarea difícil debido a la baja ocurrencia de los defectos y la gran variabilidad en la apariencia de los mismos. La evaluación se realizó usando un conjunto de imágenes de rayos X, GRIMA (Mery et al., 2015), que permite un análisis no invasivo de las piezas metálicas. Las CNNs para la extracción de características de los métodos Faster-RCNN, R-FCN, SSD han sido reentrenadas empleando imágenes de defectos. Los resultados obtenidos indican que Faster-RCNN con ResNet101 (He et al., 2016) es la más precisa en la localización de los defectos. Sin embargo, SSD con VGG16 requiere un menor tiempo de ejecución para analizar las imágenes.

En la segunda fila de la Figura 4 se ilustran los resultados de la identificación de placas de corte utilizadas en procesos de fresado con Edge Boxes, mientras que en los artículos originales puede verse la localización propuesta por los autores mediante otros métodos (García-Ordás et al., 2017; Fernández-Robles et al., 2017a,b).

4.4. Robótica

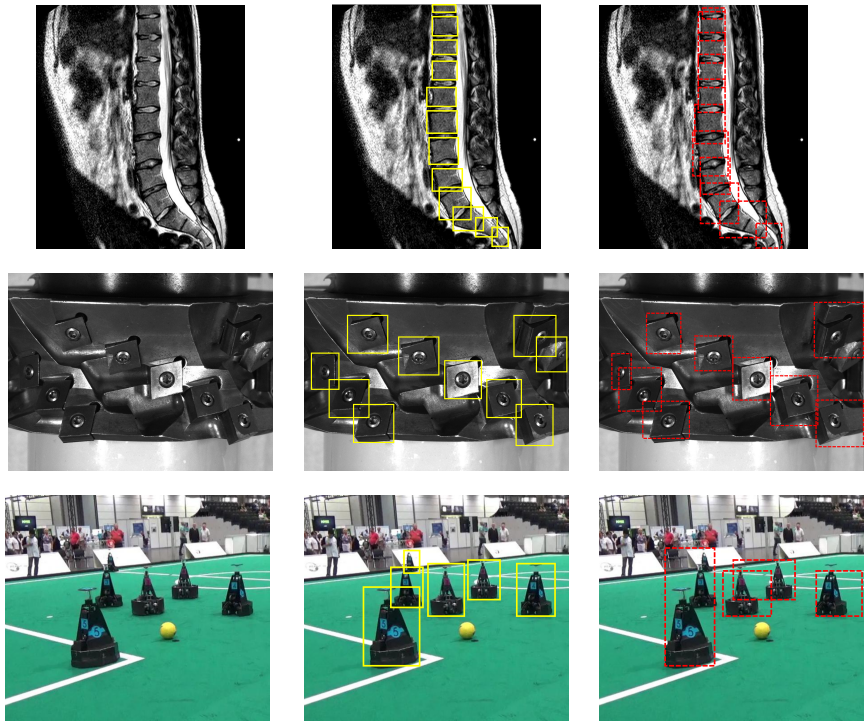
Una detección de objetos rápida y precisa determina el éxito de muchas tareas realizadas por robots, tales como: el reconocimiento, el seguimiento y el rescate en diferentes aplicaciones. En particular, Lee et al. (2015) proponen un sistema basado en VJ para la detección de obstáculos con forma de conos en un robot móvil con ruedas, para decidir la ruta a seguir al desplazarse de un punto inicial a otro. Por otra parte, Luo et al. (2017) presentan un método para la localización de otros robots de uno de los equipos participantes en el mundial de fútbol de robots. El método propuesto usa una versión reducida de YOLO para analizar imágenes en RGB y de profundidad. Los resultados obtenidos se refinan con una métrica basada en el tamaño de los robots.

En la tercera fila de la Figura 4 se presenta la identificación con Edge Boxes de robots utilizados por uno de los equipos durante el mundial de fútbol de robots.

5. Discusión, Conclusiones y Líneas Futuras

En este trabajo hemos realizado una revisión sistemática de más de 50 artículos sobre la localización de objetos de forma automática en imágenes digitales. En particular, hemos recopilado métodos más tradicionales basados en el uso de ventanas deslizantes y conjuntos de regiones candidatas junto con métodos más recientes basados en redes de aprendizaje profundo. Además de explicar brevemente las particularidades de cada método, se han mostrado las ventajas e inconvenientes que éstos presentan ante diversidad de usos.

En términos de coste computacional, las estrategias basadas en ventanas deslizantes son las más costosas ya que realizan una búsqueda exhaustiva de objetos emulando ventanas de tamaño



(a) Imagen original (b) Localización manual (c) Localización con Edge Boxes

Figura 4: Localización manual *vs* los resultados obtenidos con el método Edge Boxes. (a) Imagen original correspondiente a las aplicaciones de: detección de vértebras en la columna vertebral en la primera fila, imagen del conjunto de datos de Dženan et al. (2014); detección de plaquitas de corte en la segunda fila, imagen del conjunto de datos de Fernández-Robles et al. (2017b); y detección de robots en la tercera fila, imagen del conjunto de datos de Luo et al. (2017). (b) Localización manual de objetos de interés con rectángulos amarillos. (c) Localización semiautomática de objetos de interés usando Edge Boxes con rectángulos rojos. En cada caso se generaron 2000 regiones candidatas y se muestran las regiones más similares a las regiones identificadas manualmente.

variable que se desplazan sobre la imagen. Le siguen los métodos que proponen un conjunto de regiones candidatas, ya que reducen la búsqueda a regiones que poseen ciertas características que determinan que la probabilidad de contener un objeto sea mayor. Por último los métodos de aprendizaje profundo, si bien necesitan de redes pre-entrenadas, son muy rápidos a la hora de detectar objetos, llegando incluso a poder utilizarse en tiempo real.

En cuanto a eficiencia y aplicaciones, se ha puesto de manifiesto que los métodos basados en aprendizaje profundo son los que normalmente consiguen los mejores resultados en imágenes naturales. Asimismo, pueden conseguir altas tasas de acierto en otro tipo de imágenes mediante un ajuste de la red pre-entrenada. En cambio, los métodos basados en ventana deslizante no tienen estas limitaciones en cuanto a aplicación y llegan a conseguir buenos resultados para cualquier tipo de objeto, pero a cambio tienen un alto coste computacional. Los métodos que proponen regiones candidatas parten de la idea de que todos los objetos comparten ciertos rasgos similares, por lo que podrían estar limitados a objetos que tengan regiones características y fallar en objetos más planos y homogéneos. La pérdida de precisión debido a la generalización de los objetos se suele ver compensada con el uso de clasificadores más específicos.

Junto a la descripción de los algoritmos revisados, hemos publicado un listado de implementaciones de software libre y de conjuntos de datos para el análisis de métodos de localización. Esto permite disponer de una colección de métodos que pueden ser fácilmente evaluados en conjuntos de datos de interés para la comunidad científica y la rápida comparación de los mismos con otros métodos propuestos.

Por otra parte, se han revisado un conjunto de aplicaciones especialmente útiles en el campo de la automática e informática industrial, como son la localización de peatones y obstáculos para vehículos autónomos, la localización de objetos en imágenes aéreas tomadas desde drones y avionetas, la localización de objetos que pueden corresponder con órganos o alguna de sus partes, ayudando al diagnóstico clínico asistido por computador y la localización de objetos que permitan tanto la inspección de calidad mediante control visual como la localización de piezas y objetos en robótica. Los métodos típicamente empleados para cada aplicación han sido citados y los artículos más relevantes en la temática han sido comentados.

Creemos que la información aquí presentada, que engloba las principales técnicas, software y conjuntos de datos existentes, y sus aplicaciones a diferentes contextos, puede resultar de gran utilidad para investigadores que quieran explorar este área de conocimiento.

En cuanto a líneas futuras, aunque recientemente los métodos de detección de objetos basados en redes de aprendizaje profundo hayan conseguido grandes logros en este campo de investigación, hay aún varios retos muy interesantes que necesitan ser abordados próximamente. Consideramos que hay cuatro líneas en las que se debe realizar un esfuerzo en el futuro. La primera sería el *entrenamiento de detectores de objetos con anotación manual limitada*. La mayoría de los métodos necesitan gran cantidad de datos anotados por seres humanos para su entrenamiento. La tarea de anotar es costosa, tediosa, consu-

me mucho tiempo y además se requiere cierto conocimiento del problema para poder realizar una anotación de calidad. Dada la gran diversidad de aplicaciones y casos particulares existentes, la cantidad de datos a anotar es realmente elevada. Es por ello, que un reto futuro puede consistir en el desarrollo de métodos que no sean completamente supervisados o que no requieran de una gran cantidad de imágenes anotadas. La segunda línea correspondería a la *detección de categorías de objetos que no se han presentado como ejemplos en el entrenamiento*. Esta línea está en cierto modo relacionada con la anterior. Como es costoso disponer de un alto número de imágenes anotadas, es necesario crear métodos que aprendan de conjuntos de imágenes que contengan características comunes de objetos no entrenados. El tercer reto futuro estaría en la mejora de la *localización de objetos que pertenecen a categorías con pocos datos o datos ruidosos*. Aunque los avances hayan sido importantes, aún hay cabida a realizar mejoras en este aspecto. Y, finalmente, la *interpretación visual de alto nivel en base a la detección*. La detección puede proporcionar información relevante para interpretar las escenas en base a la posición relativa de los objetos en una imagen y así formar frases que describan el contenido de una imagen. La conexión entre el entorno visual y el entorno lingüístico es una tendencia en investigación debido a sus múltiples aplicaciones prácticas.

Agradecimientos

Este trabajo ha sido financiado parcialmente por diferentes instituciones. Deisy Chaves cuenta con una beca "Estudios de Doctorado en Colombia 2013" de COLCIENCIAS. Surajit Saikia cuenta con una beca de la Junta de Castilla y León con referencia EDU/529/2017. También queremos agradecer el apoyo de INCIBE (Instituto Nacional de Ciberseguridad) mediante la Adenda 22 al convenio con la Universidad de León. Finalmente, agradecemos a NVIDIA la donación de GPUs (GeForce GTX Titan X y K-40) que hemos utilizado en las pruebas realizadas.

Referencias

Akselrod-Ballin, A., Karlinsky, L., Alpert, S., Hasoul, S., Ben-Ari, R., Barkan, E., 2016. A region based convolutional network for tumor detection and classification in breast mammography. In: Deep Learning and Data Labeling for Medical Applications. pp. 197–205.

Alexe, B., Deselaers, T., Ferrari, V., 2010. What is an object? In: CVPR. pp. 73–80.

Ammour, N., Alhichri, H., Bazi, Y., Benjdira, B., Alajlan, N., Zuair, M., 2017. Deep learning approach for car detection in uav imagery. Remote Sens. 9 (4).
DOI: 10.3390/rs9040312

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. In: COLT. pp. 144–152.

Brazil, G., Yin, X., Liu, X., 2017. Illuminating pedestrians via simultaneous detection & segmentation. CoRR abs/1706.08564.

Cai, Z., Fan, Q., Feris, R. S., Vasconcelos, N., 2016. A unified multi-scale deep convolutional neural network for fast object detection. CoRR abs/1607.07155.

Cao, X., Gong, G., Liu, M., Qi, J., 2016. Foreign object debris detection on airfield pavement using region based convolution neural network. In: DICTA. pp. 1–6.
DOI: 10.1109/DICTA.2016.7797045

Cao, X., Wang, P., Meng, C., Bai, X., Gong, G., Liu, M., Qi, J., 2018. Region based cnn for foreign object debris detection on airfield pavement. Sensors 18 (3).
DOI: 10.3390/s18030737

Chen, J., Liu, Z., Wang, H., Núñez, A., Han, Z., 2018. Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network. IEEE T Instrum Meas 67 (2), 257–269.
DOI: 10.1109/TIM.2017.2775345

Cireşan, D. C., Giusti, A., Gambardella, L. M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI. pp. 411–418.

Coifman, B., McCord, M., Mishalani, R. G., Iswalt, M., Ji, Y., 2006. Roadway traffic monitoring from an unmanned aerial vehicle. IEE Proceedings - Intelligent Transport Systems 153 (1), 11–20.
DOI: 10.1049/ip-its:20055014

Dai, J., Li, Y., He, K., Sun, J., 2016. R-FCN: object detection via region-based fully convolutional networks. CoRR abs/1605.06409.

Dalal, N., Triggs, B., June 2005. Histograms of oriented gradients for human detection. In: CVPR. Vol. 1. pp. 886–893 vol. 1.
DOI: 10.1109/CVPR.2005.177

Deng, L., 2014. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing 3, e2.

Deng, L., Yu, D., 2014. Deep learning: Methods and applications. Foundations and Trends in Signal Processing 7 (3-4), 197–387.

Dollár, P., Tu, Z., Perona, P., Belongie, S. J., 2009. Integral channel features. In: BMVC. pp. 1–11.

Dollar, P., Zitnick, L., 2013. Structured forests for fast edge detection. In: ICCV. pp. 1841–1848.

Donoser, M., Bischof, H., 2006. Efficient maximally stable extremal region (mser) tracking. In: CVPR. pp. 553–560.
DOI: 10.1109/CVPR.2006.107

Du, X., El-Khamy, M., Lee, J., Davis, L., 2017. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In: WACV. pp. 953–961.
DOI: 10.1109/WACV.2017.111

Dženan, Z., Aleš, V., Jan, E., Daniel, H., Christopher, N., Andreas, K., 2014. Robust detection and segmentation for diagnosis of vertebral diseases using routine mr images. Computer Graphics Forum 33 (6), 190–204.
DOI: 10.1111/cgf.12343

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32 (9), 1627–1645.
DOI: 10.1109/TPAMI.2009.167

Felzenszwalb, P. F., Huttenlocher, D. P., 2004. Efficient graph-based image segmentation. IJCV 59 (2), 167–181.
DOI: 10.1023/B:VISI.0000022288.19776.77

Ferguson, M., Ak, R., Lee, Y. T. T., Law, K. H., 2017. Automatic localization of casting defects with convolutional neural networks. In: IEEE International Conference on Big Data. pp. 1726–1735.
DOI: 10.1109/BigData.2017.8258115

Fernández-Robles, L., Azzopardi, G., Alegre, E., Petkov, N., 2017a. Machine-vision-based identification of broken inserts in edge profile milling heads. Robot Comput Integr Manuf 44, 276 – 283.
DOI: https://doi.org/10.1016/j.rcim.2016.10.004

Fernández-Robles, L., Azzopardi, G., Alegre, E., Petkov, N., Castejón-Limas, M., 2017b. Identification of milling inserts in situ based on a versatile machine vision system. JMSY 45, 48 – 57.
DOI: https://doi.org/10.1016/j.jmsy.2017.08.002

Freund, Y., Schapire, R. E., 1999. A short introduction to boosting. In: IJCAI. pp. 1401–1406.

García-Ordás, M. T., Alegre, E., González-Castro, V., Alaiz-Rodríguez, R., 2017. A computer vision approach to analyze and classify tool wear level in milling processes using shape descriptors and machine learning techniques. Int J Adv Manuf Technol 90 (5), 1947–1961.
DOI: 10.1007/s00170-016-9541-0

García-Olalla, O., Alegre, E., Fernández-Robles, L., Fidalgo, E., Saikia, S., 2018. Textile retrieval based on image content from cdc and webcam cameras in indoor environments. Sensors 18 (5).
DOI: 10.3390/s18051329

Garnett, N., Silberstein, S., Oron, S., Fetaya, E., Verner, U., Ayash, A., Goldner, V., Cohen, R., Horn, K., Levi, D., 2017. Real-time category-based and general obstacle detection for autonomous driving. In: ICCVW. pp. 198–205.
DOI: 10.1109/ICCVW.2017.32

Girshick, R. B., 2015. Fast R-CNN. CoRR abs/1504.08083.

Girshick, R. B., Donahue, J., Darrell, T., Malik, J., 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR abs/1311.2524.

He, B., Xiao, D., Hu, Q., Jia, F., 2018. Automatic magnetic resonance image prostate segmentation based on adaptive feature learning probability boosting tree initialization and cnn-asm refinement. IEEE Access 6, 2005–2015.

He, K., Gkioxari, G., Dollár, P., Girshick, R. B., 2017. Mask R-CNN. CoRR abs/1703.06870.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

Heo, Y. J., Lee, D., Kang, J., Lee, K., Chung, W. K., 2017. Real-time Image Processing for Microscopy-based Label-free Imaging Flow Cytometry in a Microfluidic Chip. Scientific Reports 7 (1), 11651.
DOI: 10.1038/s41598-017-11534-0

Hosang, J., Benenson, R., Dollár, P., Schiele, B., 2016. What makes for effective detection proposals? IEEE Trans. Pattern Anal. Mach. Intell. 38 (4), 814–830.
DOI: 10.1109/TPAMI.2015.2465908

Jiamin, L., David, W., Le, L., Zhuoshi, W., Lauren, K., B., T. E., Berkman, S., A., P. N., M., S. R., 2017. Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. Medical Physics 44 (9), 4630–4642.
DOI: 10.1002/mp.12399

Jung, F., Kirschner, M., Wesarg, S., 2013. A generic approach to organ detection using 3d haar-like features. In: Bildverarbeitung für die Medizin 2013. pp. 320–325.

Kisilev, P., Sason, E., Barkan, E., Hashoul, S., 2016. Medical image description using multi-task-loss cnn. In: Deep Learning and Data Labeling for Medical Applications. pp. 121–129.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Adv Neural Inf Process Syst. pp. 1097–1105.

Lampert, C. H., Blaschko, M. B., Hofmann, T., 2008. Beyond sliding windows: Object localization by efficient subwindow search. In: CVPR. pp. 1–8.
DOI: 10.1109/CVPR.2008.4587586

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lee, C. J., Tseng, T. H., Huang, B. J., Jun-Weihsieh, Tsai, C. M., 2015. Obstacle detection and avoidance via cascade classifier for wheeled mobile robot. In: ICMLC. Vol. 1. pp. 403–407.
DOI: 10.1109/ICMLC.2015.7340955

Lee, J., Wang, J., Crandall, D., Šabanović, S., Fox, G., 2017. Real-time, cloud-based object detection for unmanned aerial vehicles. In: IRC. pp. 36–43.
DOI: 10.1109/IRC.2017.77

Levi, D., Garnett, N., Fetaya, E., September 2015a. Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In: BMVC. pp. 109.1–109.12.
DOI: 10.5244/C.29.109

Levi, D., Garnett, N., Fetaya, E., 2015b. Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In: BMVC. pp. 109.1–109.12.
DOI: 10.5244/C.29.109

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S., 2018. Scale-aware fast r-cnn for pedestrian detection. IEEE Trans Multimedia 20 (4), 985–996.
DOI: 10.1109/TMM.2017.2759508

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., 2016. Ssd: Single shot multibox detector. In: ECCV. pp. 21–37.

Luo, S., Lu, H., Xiao, J., Yu, Q., Zheng, Z., 2017. Robot detection and localization based on deep learning. In: CAC. pp. 7091–7095.

Ma, Y., Jiang, Z., Zhang, H., Xie, F., Zheng, Y., Shi, H., 2017. Proposing regions from histopathological whole slide image for retrieval using selective search. In: ISBI. pp. 156–159.
DOI: 10.1109/ISBI.2017.7950491

Mery, D., Riffo, V., Zscherpel, U., Mondragón, G., Lillo, I., Zuccar, I., Lobel, H., Carrasco, M., 2015. Gdxray: The database of x-ray images for nondestructive testing. Journal of Nondestructive Evaluation 34 (4), 42.
DOI: 10.1007/s10921-015-0315-7

Park, J.-K., Kwon, B.-K., Park, J.-H., Kang, D.-J., 2016. Machine learning-based imaging system for surface defect inspection. IJPEM-GT 3 (3), 303–310.
DOI: 10.1007/s40684-016-0039-x

Redmon, J., Divvala, S. K., Girshick, R. B., Farhadi, A., 2015. You only look once: Unified, real-time object detection. CoRR abs/1506.02640.

Ren, S., He, K., Girshick, R. B., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. CoRR abs/1506.01497.

Říha, K., Mašek, J., Burget, R., Beneš, R., Závodná, E., 2013. Novel method for localization of common carotid artery transverse section in ultrasound images using modified viola-jones detector. Ultrasound Med Biol 39 (10), 1887 – 1902.
DOI: 10.1016/j.ultrasmedbio.2013.04.013

Sa, R., Owens, W., Wiegand, R., Studin, M., Capoferri, D., Barooha, K., Greaux, A., Rattray, R., Hutton, A., Cintineo, J., Chaudhary, V., 2017. Intervertebral disc detection in x-ray images using faster r-cnn. In: EMBC. pp. 564–567.
DOI: 10.1109/EMBC.2017.8036887

Saikia, S., Fidalgo, E., Alegre, E., Fernández-Robles, L., 2017. Object detection for crime scene evidence analysis using deep learning. In: ICIAP. pp. 14–24.

Sepúlveda, G. V., Torriti, M. T., Calero, M. F., 2017. Sistema de detección de señales de tráfico para la localización de intersecciones viales y frenado anticipado. Revista Iberoamericana de Automática e Informática Industrial 14 (2), 152–162.
DOI: 10.1016/j.riai.2016.09.010

Shah, V. R., Maru, S. V., Jhaveri, R. H., 2018. An obstacle detection scheme for vehicles in an intelligent transportation system. IJCNIS 8 (10), 23–28.
DOI: 10.5815/ijcnis.2016.10.03

Shi, Y., Li, Y., Wei, X., Zhou, Y., 2017. A faster-rcnn based chemical fiber paper tube defect detection method. In: International Conference on Enterprise Systems. pp. 173–177.
DOI: 10.1109/ES.2017.35

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. pp. 4278–4284.

Tang, T., Zhou, S., Deng, Z., Zou, H., Lei, L., 2017. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors 17 (2).
DOI: 10.3390/s17020336

Tek, F., 2013. Mitosis detection using generic features and an ensemble of cascade adaboosts. J Pathol Inform 4 (1), 12.
DOI: 10.4103/2153-3539.112697

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., Smeulders, A. W. M., 2013. Selective search for object recognition. IJCV 104 (2), 154–171.

Viola, P., Jones, M. J., May 2004. Robust real-time face detection. IJCV 57 (2), 137–154.
DOI: 10.1023/B:VISI.0000013087.49260.fb

Wang, S., Cheng, J., Liu, H., Tang, M., 2018. Pcn: Part and context information for pedestrian detection with cnns. CoRR abs/1804.04483.

Xu, Y., Yu, G., Wang, Y., Ma, Y., 2017a. Car detection from low-altitude uav imagery with the faster r-cnn. JAT 2017.
DOI: https://doi.org/10.1155/2017/2823617

Xu, Y., Yu, G., Wang, Y., Wu, X., Ma, Y., 2016. A hybrid vehicle detection method based on viola-jones and hog + svm from uav images. Sensors 16 (8).
DOI: 10.3390/s16081325

Xu, Y., Yu, G., Wu, X., Wang, Y., Ma, Y., 2017b. An enhanced viola-jones vehicle detection method from unmanned aerial vehicles imagery. IEEE trans Intell Transp Syst 18 (7), 1845–1856.
DOI: 10.1109/TITS.2016.2617202

Yang, S., Fang, B., Tang, W., Wu, X., Qian, J., Yang, W., 2017. Faster r-cnn based microscopic cell detection. In: SPAC. pp. 345–350.
DOI: 10.1109/SPAC.2017.8304302

Yi, X., Song, G., Derong, T., Dong, G., Liang, S., Yuqiong, W., 2018. Fast road obstacle detection method based on maximally stable extremal regions. IJARS 15 (1), 1–10.
DOI: 10.1177/1729881418759118

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833.

Zhang, L., Lin, L., Liang, X., He, K., 2016. Is faster r-cnn doing well for pedestrian detection? In: ECCV. pp. 443–457.

Zhong, J., Lei, T., Yao, G., 2017. Robust vehicle detection in aerial images based on cascaded convolutional neural networks. Sensors 17 (12).
DOI: 10.3390/s17122720

Zitnick, L., Dollar, P., 2014. Edge boxes: Locating object proposals from edges. In: ECCV. pp. 391–405.