Actas de las XXXIX Jornadas de Automática, Badajoz, 5-7 de Septiembre de 2018
https://doi.org/10.17979/spudc.9788497497565.1022

DETECCIÓN DE EQUIPOS DE PROTECCIÓN PERSONAL MEDIANTE RED NEURONAL CONVOLUCIONAL YOLO

Manlio Massiris
Depto. de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur y Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Av. San Andrés 800, 8000, Bahía Blanca, Argentina, manlio.massiris@uns.edu.ar

Claudio Delrieux
Depto. de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur y Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Av. San Andrés 800, 8000, Bahía Blanca, Argentina, cad@uns.edu.ar

J. Álvaro Fernández
Depto. de Ingeniería Eléctrica, Electrónica y Automática, Escuela de Ingenierías Industriales, Universidad de Extremadura, Av. Elvas s/n, 06006, Badajoz, España, jalvarof@unex.es

Resumen

En un número creciente de entornos de trabajo está tornándose obligatorio el uso de equipos de protección personal, debido a que son la última barrera para detener situaciones potenciales de riesgo físico para el trabajador. Eso determina que controlar en forma periódica y fehaciente el cumplimiento de las normas de seguridad laboral sea una tarea demandante, por lo cual el monitoreo no supervisado representa una solución de alto impacto para la seguridad industrial. El presente artículo propone utilizar visión artificial como alternativa cuantitativa para monitorear la utilización de equipo de protección personal. Se entrenó la red neuronal YOLO con la intención de detectar guantes, cascos, ropa de alta visibilidad y a los trabajadores con un dataset *creado a partir de videos generados utilizando cámaras deportivas. Con el sistema entrenado, se presenta un análisis de caso* in the open *con un video grabado con cámara deportiva sujeta al casco de un trabajador metalúrgico en el sector de la construcción. Los resultados son promisorios y muestran que la estrategia planteada es adecuada para llegar a una solución implantable en ambientes de trabajo.*

Palabras Clave: Visión por computador, Equipos de protección personal (EPP), Seguridad industrial.

1 INTRODUCCIÓN

En la seguridad e higiene en el trabajo con maquinaria pesada en general, y en la industria de la construcción en particular, la ausencia o el uso inadecuado de los Equipos de Protección Personal (EPP) por parte de los trabajadores y operarios es una de las principales causas de accidentes y lesiones evitables [10]. Por dicha razón, las acciones preventivas y proactivas basadas en la comprensión del comportamiento de personas en entornos laborales es un área de investigación de creciente interés, ya que sus resultados permiten reducir el costo económico, social y ético generado por los accidentes laborales [6]. La metodología tradicional se basa en detectar las conductas de riesgo e intervenir utilizando realimentación visual o auditiva para prevenir o mitigar el peligro [16]. Esta metodología tiene varias deficiencias asociadas, y en general depende en gran medida de la intervención humana, en especial la de expertos en observaciones *in situ*, para reunir y analizar la información [16]. Además, el auto-reporte de situaciones que podrían acarrear peligros y la colaboración en general por parte de los trabajadores en la medición de utilización de EPP, se tornan dificultosos e imprecisos a lo largo del tiempo [5].

El uso de EPP es obligatorio en una creciente cantidad de contextos laborales, dado que es la última medida de seguridad que separa al trabajador de la fuente de riesgo [11]. Sin embargo, en la práctica, muchos accidentes laborales se producen por la falta de uso correcto de EPP normalizado (véase Fig. 1). El EPP regulado para una cierta actividad puede incluir elementos como guantes, gafas o calzado de seguridad, tapones para los oídos u orejeras, cascos, respiradores, o monos, chalecos y trajes de cuerpo completo, entre otros.

Recientemente, ha aumentado de forma significativa el interés por el uso de sistemas de visión artificial para prevenir riesgos laborales. Esto se debe, entre otras causas, al abaratamiento y uso masivo de cámaras digitales, el surgimiento de plataformas *open source* que facilitan el prototipado rápido de soluciones de visión artificial, y al crecimiento de las comunidades que comparten sus *datasets* y código [5]. La monitorización de la

proximidad del trabajador respecto a un objeto riesgoso [9], la evaluación de la ergonomía ocupacional [16], o la detección de utilización de casco son algunas de las aplicaciones más relevantes en la literatura reciente [3, 5, 10].

Este trabajo se centra en la monitorización de la presencia y el uso adecuado del EPP por parte de trabajadores de la construcción. Para ello, se utiliza la red neuronal convolucional (CNN) denominada "*You Only Look Once*" (YOLO) [13], cuya topología está específicamente diseñada para resolver problemas de visión artificial con un alto rendimiento.

El resto del trabajo se organiza como sigue: en la Sección 2 se revisan los antecedentes técnicos respecto a los EPP y las CNN objeto de estudio; en la Sección 3 se describe la metodología empleada en la solución al problema planteado; en la sección 4 se presentan los experimentos y los resultados de la investigación, teniendo como eje un estudio de caso, para finalmente, en la sección 5, resumir las conclusiones y plantear posibles trabajos futuros.

2 ANTECEDENTES

2.1 EQUIPOS DE PROTECCIÓN PERSONAL (EPP)

Figura 1: Ejemplo de mal uso de EPP: falta arnés de seguridad y ropa de alta visibilidad.

La norma OHSAS 18001 [12] propone una cadena preventiva jerárquica que determina que primero se debe intentar la eliminación del riesgo, sustitución del riesgo, los controles de ingeniería, la señalización, las advertencias, los controles administrativos y, por último, los equipos de protección personal. Por esto los EPP no tienen como objetivo evitar el accidente, sino que buscan reducir o minimizar las consecuencias del siniestro.

Otro estudio de [6] menciona que tanto el uso inapropiado como la falta de uso de EPP hizo que el sector de la construcción fuera el sector industrial menos seguro en el Reino Unido. Las razones principales fueron declaradas como "sin EPP", seguido de "EPP no utilizado". Se concluyó además que la deficiente supervisión del sitio fue la deficiencia clave cuando no se utilizó el EPP.

2.2 REDES NEURONALES CONVOLUCIONALES (CNN)

El aprendizaje profundo o Deep Learning (DL) permite que modelos computacionales compuestos por varias capas de procesamiento puedan aprender representaciones sobre datos con múltiples niveles de abstracción y, mediante ese concepto, descubrir representaciones precisas de forma autónoma en grandes volúmenes de datos. El DL ha logrado recientemente grandes avances en el reconocimiento de imágenes y video [3]. Un caso particular de DL son las redes convolucionales o Convolutional Neural Networks (CNN) [4, 7], las cuales constituyen actualmente el estado del arte de varios problemas de visión computacional, dado su buen desempeño problemas de reconocimiento e interpretación en imágenes y video [3]. Su capacidad para actuar adecuadamente en estos contextos está basada en características fundamentales: conexiones locales, pesos compartidos, *pooling* y el uso de una gran cantidad de capas [4]. El propósito de CNN es extraer todas las características de una imagen y luego usar dichas características para detectar o clasificar los objetos en una imagen. Los parámetros de los filtros que se pueden aprender en estas capas; se ajustarán y optimizarán junto con los componentes de clasificación para minimizar el error de clasificación total [4, 7].

La aparición de CNN ha llevado a un rápido desarrollo del campo de detección de objetos. Por esto nos permitimos mencionar los algoritmos avanzados y más recientes en el campo de detección de objetos con redes neuronales convolucionales: Faster R-CNN [15], Single Shot Multibox Detector (SSD) [8] y *You Only Look Once* (YOLO) [13], que es la adoptada en este trabajo y se describe en detalle a continuación.

2.3 LA RED YOU ONLY LOOK ONCE (YOLO)

En YOLO se toma la detección de objetos como un problema único de regresión, una única red convolucional predice simultáneamente múltiples cuadros delimitadores que enmarcan los objetos en la imagen y predice probabilidades condicionales por cada clase *p(Clase | Objeto)* para cada uno de estos cuadros delimitadores [13]. La red neuronal puede lograr una velocidad de ejecución de 45 fotogramas por segundo (fps) en computadoras de propósito general [13].

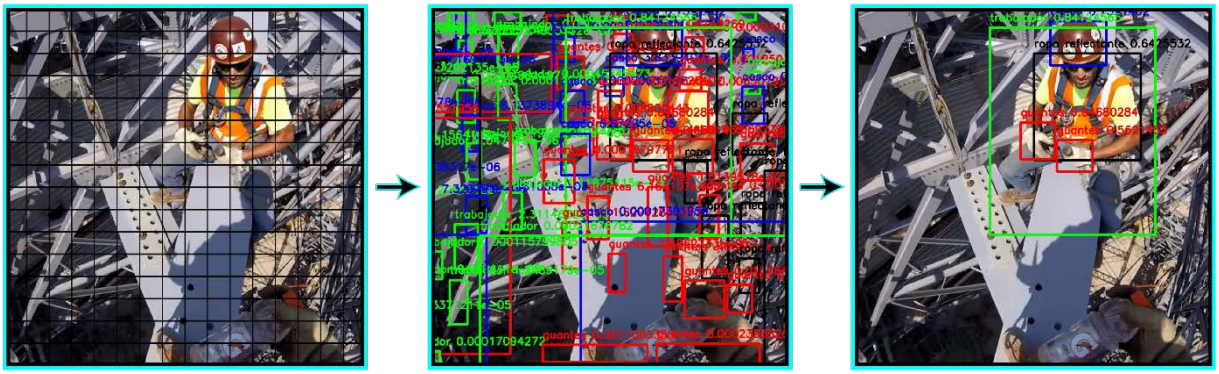

Figura 2: Proceso de YOLO [13], de izquierda a derecha: (1) división en bloques 13x13, (2) realización de predicciones, y (3) umbralización para obtener solo las detecciones más fehacientes.

YOLO trabaja globalmente sobre la imagen cuando hace predicciones, a diferencia de la técnica de ventana deslizante y las técnicas basadas en el análisis de las regiones en una imagen [13]. Por esto, codifica implícitamente la información contextual, modela el tamaño y la forma de los objetos, así como su apariencia [13] (véase Fig. 2).

Como se puede apreciar en la Fig. 3, YOLO en su primera versión tiene 24 capas convolucionales seguidas por 2 capas completamente conectadas, y en lugar de los módulos de iniciales propuestos por GoogLeNet, YOLO utiliza capas de reducción de 1×1 seguidas de capas convolucionales de 3×3 [13]. En la siguiente versión YOLO9000 [14], al agregar la normalización de lotes en todas las capas convolucionales, se obtiene una mejora de más del 2% en la precisión promedio (mAP, por sus siglas en ingles). La normalización por lote también ayuda a regularizar el modelo y se elimina el sobreajuste. YOLO9000 predice las detecciones en un mapa de características de 13×13. Si bien esto es suficiente para objetos grandes, debe beneficiarse de funciones de grano más fino para localizar objetos más pequeños [14].

3 METODOLOGÍA

El presente artículo propone la utilización de un sistema basado en la red neuronal YOLO, que nos permite mediante imágenes la detección de objetos

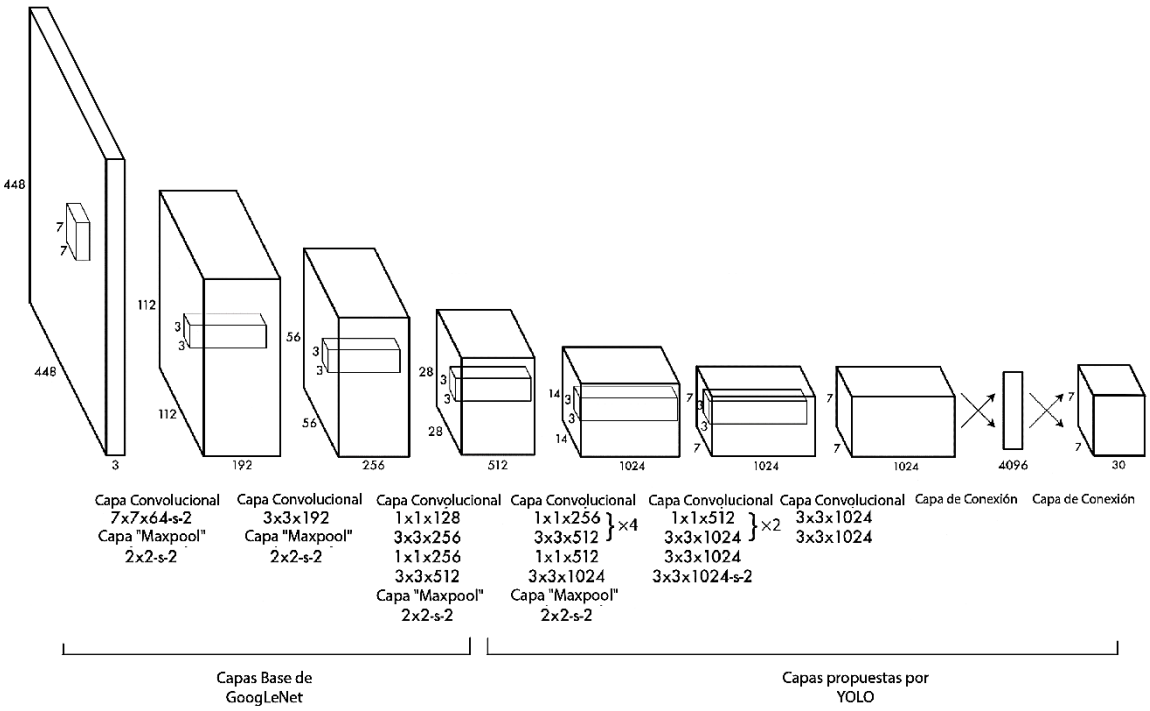

Figura 3: Arquitectura de la red neuronal YOLO [13]

con una alta efectividad, y que a diferencia de los trabajos mencionados en la revisión bibliográfica, es capaz de detectar en una imagen guantes, ropa de alta visibilidad, casco y además a los trabajadores, lo cual es novedoso. Según lo mencionado en los antecedentes, no se encontró evidencia de desarrollos que trabajen en visión por computador y puedan detectar varios equipos de protección personal en el trabajador al mismo tiempo.

3.1 DESARROLLO DEL DATASET DE TRABAJADORES METALÚRGICOS

Debido a las normativas europeas y argentinas de protección de datos, se tomaron de muestra videos obtenidos de YouTube.com, grabados con cámaras deportivas, en los cuales se aprecian a trabajadores metalúrgicos en su trabajo diario o de construcción realizando tareas relacionadas con la metalurgia, y vistiendo diferentes tipos de EPP. Para esto, se realizó una búsqueda y se seleccionaron 5 videos tomados en diferentes condiciones tales como: cámara deportiva sujeta al casco de un trabajador, cámara deportiva estática sobre una maquinaria y videos filmados con un dron que porta una cámara deportiva, los cuales se dividieron en imágenes.

Consecutivamente, se procedió a realizar el etiquetado manual de los distintos EPP en cada una de las 1354 imágenes, utilizando la herramienta YOLO_mark [14] de los mismos creadores de YOLO. Las anotaciones fueron guardadas en archivos TXT con el formato de la herramienta. Esto resultó ser una etapa engorrosa, que tomó gran cantidad de tiempo, debido a que en algunas imágenes hay gran cantidad de trabajadores, y cada trabajador debe vestir dos guantes, un casco y un chaleco de alta visibilidad (véase Fig. 4).

Posteriormente, siguiendo con las recomendaciones de YOLO, se dividió el *dataset* de forma aletatoria: el 80% de las imágenes fue destinado para el entrenamiento, el 10% para el testeo y, por último, el 10% restante para validación. Además, se generaron los directorios para cada grupo en los cuales el *dataset* fue dividido. Por último, todo el *dataset* fue empaquetado para su posterior entrenamiento en una máquina dedicada.

3.2 ENTRENAMIENTO

Para el proceso de entrenamiento, se utilizó la distribución de YOLO9000 original, a través de Darknet [14]. Para el proceso de entrenamiento se utilizó un ordenador con 16 GB de memoria, una tarjeta gráfica NVIDIA GeForce GTX 1080 TI y un procesador Intel Core i5. El proceso de entrenamiento fue relativamente sencillo, ya que, está muy bien explicado en la página de los autores.

Finalmente, después de 22000 *epochs* (períodos de entrenamiento) se compararon los resultados entre ellos y se seleccionaron los coeficientes para la red neuronal generados a los 9000 *epochs*.

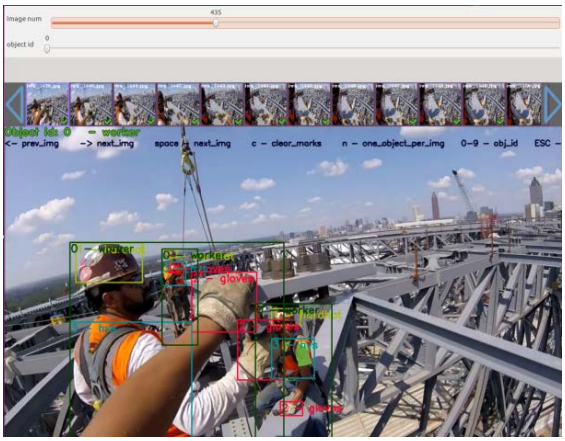

Figura 4: Ejemplo de fotograma etiquetado manualmente.

Tabla 1: Matriz de contingencia.

		Predicho por la red	
		Objeto presente	Objeto no presente
Situación real	Objeto presente	VP	FN
	Objeto no presente	FP	VN

4 RESULTADOS Y EXPERIMENTOS

Debido a la gran cantidad de herramientas y flexibilidad que ofrece Python, para la evaluación de resultados y los experimentos de este trabajo se utilizó YOLOv2 en una versión desarrollada en Keras con *backend* de Tensorflow [2] y OpenCV. El rendimiento de la red YOLO entrenada se midió en función de sus métricas de calidad y velocidad.

4.1 MÉTRICAS DE CALIDAD

Las métricas de calidad se seleccionaron en función del objetivo principal del método para detectar las clases casco, ropa de alta visibilidad, guantes y trabajador individualmente. La Tabla 1 muestra el concepto de matriz de contingencia, donde se ilustra el significado de VP (verdadero positivo), FP (falso positivo), FN (falso negativo) y VN (verdadero negativo). Específicamente, VP es el número de clases detectados correctamente, FP es la cantidad de objetos detectados que no corresponden con los etiquetados manualmente [1],

y FN es la cantidad de objetos para cada clase que la red neuronal no ha detectado.

VN es la situación que ocurre cuando en un cuadro no aparece un objeto etiquetado, y la red reporta correctamente tal situación. Este caso es el más frecuente y esperado para cada clase y, por lo tanto, en una métrica de calidad, tener en cuenta los verdaderos negativos genera un sesgo inadecuado. Por dicha razón, en este contexto de identificación se evalúa la calidad en términos de obtener la mayor tasa de VN y simultáneamente disminuir FP y FN [1].

La primera métrica utilizada usualmente en el contexto de la identificación en imágenes es la *precisión* (o valor predictivo positivo), que se define como VP / (VP + FP) y mide la fiabilidad de la detección o proporción de casos VP entre todos los casos positivos detectados por la prueba [1]. Teniendo en cuenta que VP + FP es la cantidad de objetos detectados en cada clase, la precisión indica qué porcentaje de las detecciones tiene realmente valor. La segunda métrica es la *sensibilidad*, también denominada *recall* o tasa de verdaderos positivos, definida como VP / (VP + FN), la cual básicamente determina el porcentaje de objetos existentes que fueron efectivamente detectados [1].

Un clasificador muy específico (que detecte objetos solamente cuando tiene mucha certeza) tendrá alta precisión, pero su sensibilidad será baja (y viceversa) [1]. La curva precisión-sensibilidad muestra la compensación entre precisión y sensibilidad a diferentes umbrales de detección. La alta precisión se relaciona con una baja tasa de falsos positivos, y la alta sensibilidad se relaciona con una alta tasa de falsos positivos. Los puntajes altos de ambos muestran que el clasificador devuelve resultados precisos (alta precisión), así como también devuelve la mayoría de todos los resultados positivos (alta sensibilidad) [1]. El área debajo de la curva de precisión-sensibilidad (AP, por sus siglas en inglés) es entonces un indicador de la calidad de un detector, dado que aumenta tanto cuando es alta la sensibilidad como la precisión para cada clase. Finalmente, la precisión promedio de cada clase es equivalente al AP, teniendo este valor se calcula la media. En la práctica, un valor de media de la precisión promedio (mAP, por sus siglas en inglés) más alto indica un mejor comportamiento del clasificador [1].

4.2 ANALISIS DE CASO

En este apartado se recogen los resultados de análisis de un vídeo de 14’23’’, disponible para consulta en la URL: https://youtu.be/54gwd2lcjpg

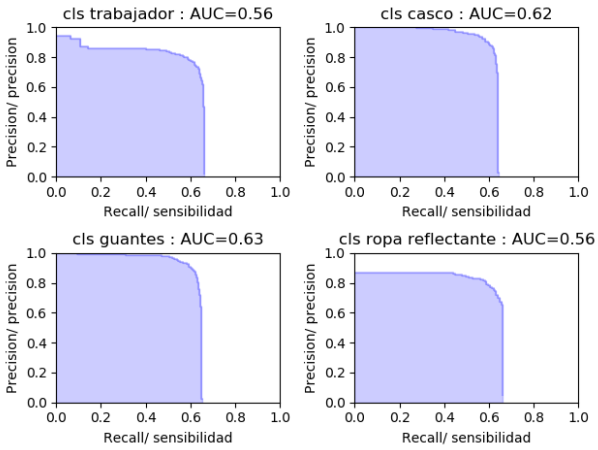

Figura 5: Curvas precisión-sensibilidad para cada una de las clases.

Este video no forma parte de los utilizados para entrenar y testear la red neuronal, pero reúne las características de los casos de aplicación esperados como resultado de esta investigación. Fue grabado con una cámara deportiva sujeta al casco de un trabajador metalúrgico en una construcción real, por lo cual es una oportunidad de medir los resultados y ahondar sobre las posibilidades de mejora. En el vídeo se etiquetan de color verde los trabajadores detectados, de color negro la ropa de alta visibilidad detectada, de color rojo los guantes detectados y de color azul los cascos detectados.

Como resultado de las curvas precisión-sensibilidad para cada una de las clases (véase Fig. 5), se eligió 0,45 como el umbral de confianza óptimo para el estudio, obteniendo un mAP general de 0,57.

4.2.1 Verdadero Positivo

En la Fig. 6 se puede apreciar la dificultad de reconocer el casco que tiene diferentes pegatinas que hacen difícil su reconocimiento por color o forma, además de, reconocer los guantes en dos diferentes posturas y al trabajador sentado de una forma poco convencional.

Figura 6: Ejemplo de verdadero positivo.

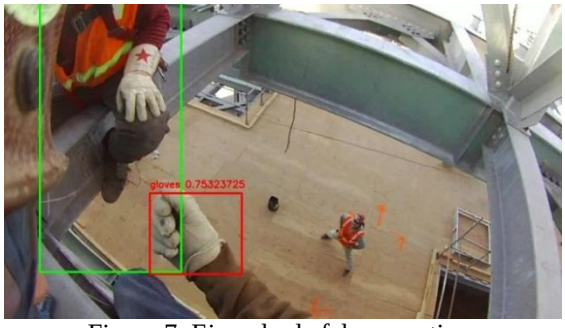

Figura 7: Ejemplo de falso negativo.

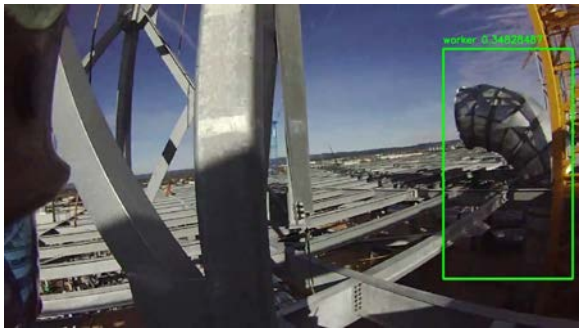

Figura 8: Ejemplo de falso positivo.

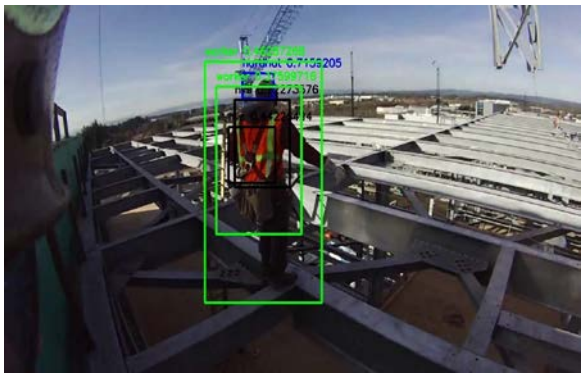

Figura 9: Ejemplo de etiquetado múltiple.

Figura 10: Detección de objeto lejano.

4.2.2 Falso negativo

En la imagen mostrada en la Fig. 7, se puede apreciar que la red neuronal es capaz de realizar la detección parcial del trabajador, sin embargo, no es capaz de detectar el guante de este ni tampoco el trabajador que se encuentra abajo a la derecha.

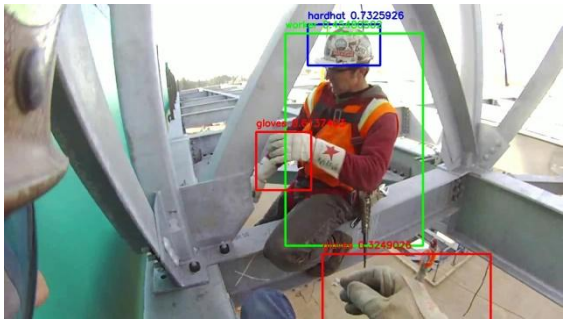

Figura 11: Ejemplo de oclusión parcial.

Figura 12: Ejemplo de oclusión parcial.

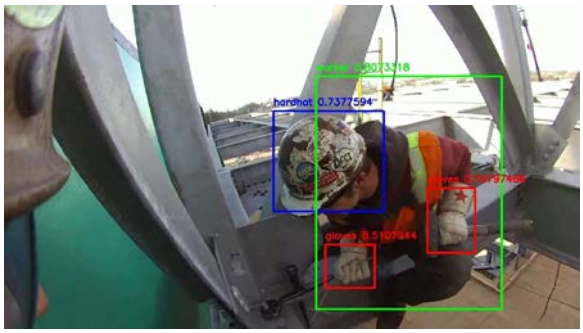

Figura 13: Ejemplo de VP en postura no habitual.

4.2.3 Falso positivo

En varios fotogramas del vídeo se identifican otros objetos como trabajador, es decir, como FP (véase ejemplo en Fig. 8).

4.2.4 Etiquetado múltiple

Como se puede apreciar en la imagen de la Fig. 9, es necesario implementar un supresor para encontrar los no máximos, de tal forma que un determinado objeto no sea reconocido dos veces.

4.2.5 Impacto del rango visual

Dado que la trayectoria de los trabajadores es de naturaleza estocástica [16], los trabajadores fueron capturados en diferentes tamaños en los videos. Si los trabajadores están cerca de la cámara, se capturan en un tamaño de píxel mayor y tienen características de imagen más ricas. Por el

contrario, los trabajadores que se encuentran lejos de la cámara se capturan en un tamaño de píxel más pequeño y tienen características de imagen de menor calidad. Este hecho se amplifica al considerar un EPP específico ya que estos son más pequeños que el trabajador.

En particular, como se puede apreciar en la Fig. 10, hay una dificultad al detectar los guantes cuando el trabajador se encuentra lejos. Esta es una de las desventajas heredadas al utilizar YOLO9000 al cual le es relativamente difícil encontrar objetos pequeños en comparación con otros métodos de detección [14], pero este problema puede ser solucionado utilizando cámaras de mayor resolución como las que están surgiendo en el mercado.

4.2.6 Impacto de las oclusiones

La red neuronal presenta un buen rendimiento a la hora de reconocer los guantes parcialmente cuando estos son cercanos a la cámara. Por ejemplo, en la imagen de la Fig. 11, la red detecta que el trabajador que porta la cámara está utilizando guantes. La red neuronal también presenta un buen desempeño al detectar trabajadores parcialmente, como se aprecia en la Fig. 12.

4.2.7 Impacto de la postura individual y de la distorsión generada por el tipo de lente.

Al ser altamente generalizable, es menos probable que YOLO falle cuando se entrene con las diferentes posturas que puede adoptar un trabajador. Como se puede apreciar en la imagen de la Fig. 13, la distorsión en el lente genera dificultad para reconocer las líneas rectas, sin embargo, la red neuronal entrenada es tiene la habilidad de superar las dificultades planteadas en este aspecto.

4.2.8 Velocidad

La velocidad de YOLOv2 entrenada se refiere al tiempo consumido por dicha red neuronal al realizar la detección de todas las clases entrenadas en el video de 14’23’’, con formato 1200x720 pixeles a 30 fps. Para el cálculo de la velocidad se utiliza la misma tarjeta gráfica GTX 1080 utilizada en el entrenamiento. Si bien es cierto que no es necesario que este tipo de análisis se realice en tiempo real, la velocidad es una métrica importante ya que uno de los objetivos del presente trabajo es minimizar el tiempo empleado por los ingenieros en seguridad para analizar este tipo de escenarios. Como resultado, se obtiene que la red neuronal emplea 22’40’’ en analizar los 14’23’’ del vídeo mencionado, logrando así una tasa superior a 19 fps sobre los 30 fps originales.

5. CONCLUSIONES Y TRABAJO FUTURO

En este artículo se ha dado el primer paso en una solución basada en visión artificial para la medición cuantitativa de la utilización de EPP, mediante una CNN tipo YOLO, y cuyo objetivo es la detección de cascos, guantes, ropa de alta visibilidad y a los propios trabajadores que los portan. Se llevó a cabo un análisis de caso y los resultados revelaron que el detector YOLO entrenado era robusto en varios escenarios y condiciones. Además, se comprobó que es relativamente eficiente en tiempo de procesado. Los resultados pueden mejorarse agregando un supresor de no máximos para así reducir los etiquetados múltiples.

Se trabajará más para mejorar la precisión del proceso de detección y eliminar las detecciones falsas entrenando y evaluando el *dataset* en la nueva versión de YOLOv3, la cual reduce un poco la velocidad de ejecución de la red neuronal, pero aumenta su tasa de aciertos. También se requieren explorar pruebas adicionales para evaluar la precisión de otros algoritmos de detección de objetos, por lo cual se debe ampliar el *dataset* tanto en entrenamiento como en nuevas pruebas. Los estudios futuros también tendrán como objetivo integrar este proceso de detección de EPP en un software completo de inspección de en seguridad industrial que pueda ser utilizado por el usuario final.

Agradecimientos

Este trabajo ha sido financiado por la Junta de Extremadura a través de los Fondos Europeos de Desarrollo Regional (FEDER), y elaborado con el apoyo de la UE bajo el programa EuLaLinks Sense Erasmus y del Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET).

English summary

DETECTION OF PERSONAL PROTECTION EQUIPMENT USING THE YOLO CONVOLUTIONAL NEURAL NETWORK

Abstract

In an increasing number of working environments, the use of personal protective equipment is becoming mandatory, since they are the last barrier to stop potential situations of physical risk for the worker. This means that periodically and reliably monitoring compliance with labor safety standards is a demanding task, which is why unsupervised monitoring represents a high impact solution for safety. This article proposes using artificial vision

as a quantitative alternative to monitor the use of personal protective equipment. The YOLO neural network was trained with the intention of detecting gloves, hard hats, high visibility suits and workers with a dataset *created from videos generated using sports cameras. With the trained system, an in-the-open case analysis is presented with a video recorded with a sports camera attached to the helmet of a metallurgical worker in a real construction site. The results are promising and show that the proposed strategy is adequate as implantable solution for these work environments.*

Keywords: Computer vision, Personal protective equipment (PPE), Industrial safety.

Referencias

[1] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2), 303–338.

[2] Experiencor (2018). YOLOv2 in Keras and Applications. Disponible en línea en: https://github.com/experiencor/keras-yolo2

[3] Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., & An, W. (2018). Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. Automation in Construction, 85, 1-9.

[4] Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Networks, 1(2), 119–130.

[5] Guo, H., Yu, Y., & Skitmore, M. (2017). Visualization technology-based construction safety management: A review. Automation in Construction, 73, 135-144.

[6] Kelm, A., Laußat, L., Meins-Becker, A., Platz, D., Khazaee, M. J., Costin, A. M., Helmus, M., & Teizer, J. (2013). Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. Automation in Construction, 36, 38–52. https://doi.org/10.1016/J.AUTCON.2013.08.009

[7] LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (1998). Efficient backprop. In Neural networks: Tricks of the trade (pp. 9–50). Springer.

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European conference on computer vision (pp. 21–37). Springer.

[9] Memarzadeh, M., Golparvar-Fard, M., & Niebles, J. C. (2013). Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. Automation in Construction, 32, 24–37. https://doi.org/10.1016/J.AUTCON.2012.12.002

[10] Mneymneh, B. E., Abbas, M., & Khoury, H. (2017). Automated Hardhat Detection for Construction Safety Applications. Procedia Engineering, 196, 895–902. https://doi.org/10.1016/J.PROENG.2017.08.022

[11] Moreno, M. D. & Fernández, J. A. (2016). An access detection and machine cycle tracking system for machine safety. International Journal of Manufacturing Technology, 87, 77-101. https://doi.org/10.1007/s00170-016-8446-2

[12] OHSAS, B. S. (2007). 18001: 2007. Occupational Health and Safety Management Systems. London.

[13] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779–788).

[14] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. ArXiv Preprint.

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R_CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91–99).

[16] Yu, Y., Guo, H., Ding, Q., Li, H., & Skitmore, M. (2017). An experimental study of real-time identification of construction workers' unsafe behaviors. Automation in Construction, 82, 193-206.