DETECCIÓN DE PARTES DEL CUERPO EN IMÁGENES MULTIMODALES DE BÚSQUEDA Y RESCATE

Alejandro González Centeno, Ricardo Vázquez-Martín, Anthony Mandow y Alfonso García-Cerezo
Universidad de Málaga, Andalucía Tech, Departamento de Ingeniería de Sistemas y Automática
amandow@uma.es

Resumen

Los sistemas de visión son fundamentales en tareas de búsqueda y rescate (SAR), principalmente en misiones cruciales como la detección de posibles víctimas en entornos de desastre. El uso de imágenes de los espectros visible (RGB) e infrarrojo térmico (TIR) para la detección de objetos son complementarias, y permiten la detección en condiciones de visibilidad limitadas. El presente trabajo analiza cómo las técnicas de aprendizaje profundo basadas en redes neuronales convolucionales (CNN) pueden aplicarse a ambas modalidades para la detección de partes del cuerpo en escenarios de catástrofe. Con este fin, se ha empleado la red YOLOv5 en ambos espectros utilizando el conjunto público de datos UMA-SAR dataset. Finalmente, se evalúan sus resultados en distintas condiciones de visibilidad.

Palabras clave: visión por computador, aprendizaje profundo, redes neuronales convolucionales, YOLOv5, imágenes térmicas, robótica para catástrofes.

1. INTRODUCCIÓN

Las tareas de búsqueda y rescate (SAR) en catástrofes y entornos hostiles demandan una respuesta rápida y eficiente para salvar vidas humanas y conllevan un cierto riesgo para los intervinientes. Por esta razón, en la robótica aplicada a operaciones SAR es importante que los vehículos no tripulados sean capaces de efectuar misiones con precisión y seguridad. Con este fin, resulta crucial que los robots dispongan de capacidades de percepción mediante sistemas de visión que puedan identificar y clasificar características del entorno y distintos objetos clave, tales como víctimas, partes del cuerpo de las mismas o personal de rescate [1, 7, 11, 12].

La robótica SAR puede beneficiarse de los desarrollos recientes en aprendizaje automático [16], y especialmente de las técnicas de aprendizaje profundo mediante de redes neuronales convolucionales o *Convolutional Neural Network* (CNN), las cuales han demostrado en los últimos años ser una herramienta altamente eficaz en aplicaciones de visión artificial como el reconocimiento de patrones y la clasificación [18]. Además, la disponibilidad de unidades de procesamiento gráfico (GPU) embarcables en el vehículo facilita el procesamiento de imágenes en tiempo real [2]. No obstante, las redes requieren un gran volumen de datos para alcanzar un nivel aceptable de aprendizaje y proporcionar resultados útiles. Las condiciones adversas en operaciones SAR dificultan la adquisición de los datos y su disponibilidad. Además, se necesitan muchos recursos para el etiquetado de conjuntos de datos, aún cuando es posible recurrir a la transferencia de conocimiento a partir de redes previamente entrenadas [5].

Asimismo, debido a las condiciones desfavorables en las que se suele trabajar en estos entornos (escasa visibilidad, polvo, humo), el reconocimiento de objetos en imágenes de catástrofes puede requerir la combinación de distintas modalidades de percepción donde las imágenes del espectro visible (RGB) se complementan con nubes de puntos lidar [4] o imágenes térmicas en el espectro infrarrojo [3, 6].

El objetivo principal de este trabajo es conseguir una red CNN capaz de detectar partes del cuerpo humano, en concreto cuatro (cabeza, torso, piernas y brazos), en condiciones de operaciones SAR. Para alcanzar este objetivo se han empleado modelos YOLOv5 [8], una arquitectura que presenta un buen equilibrio entre precisión y velocidad [15]. Estos modelos se reentrenan con un dataset específico en condiciones SAR que incluye imágenes del espectro visible y del espectro infrarrojo térmico, tomadas del *UMA-SAR dataset* [10]. Asimismo, se compara el rendimiento de dos modelos de distinto tamaño.

2. RED YOLOv5

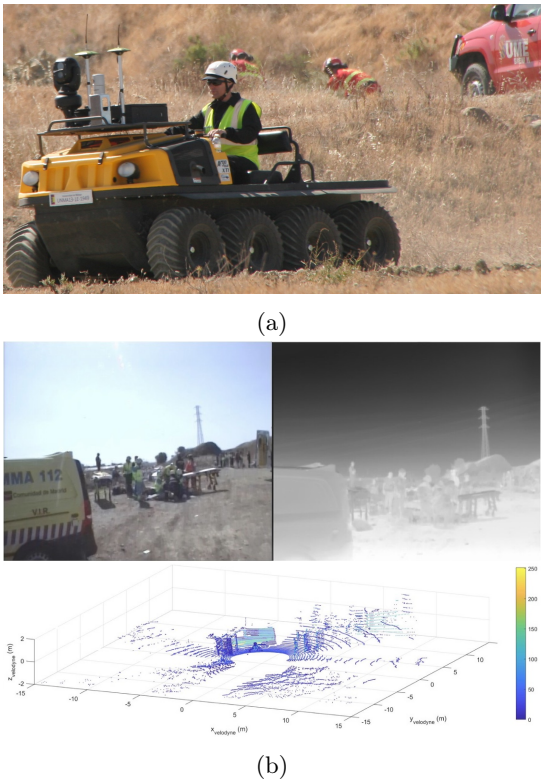
En este trabajo se emplea la arquitectura YOLOv5 [8], la cual ofrece algunas mejoras respecto a las versiones anteriores [14], principalmente en cuanto a rapidez, al mismo tiempo que mantiene un alto nivel de precisión. Además, incorpora distintos modelos preentrenados con el dataset de

ISBN 978-84-09-41095-8

COCO [9] que pueden utilizarse con técnicas de transferencia del conocimiento.

La arquitectura de YOLOv5 se compone de tres partes principales. En primer lugar, la columna vertebral o *Backbone* se encarga de realizar la convolución para extracción de características. La segunda, denominada cuello o *Neck*, se encarga de realizar la reducción del muestreo mediante la fusión de las características obtenidas en la capa anterior. Finalmente, la parte denominada cabeza o *Head*, realiza la clasificación de las características obtenidas (etiquetado, localización en la imagen, tamaño de la caja delimitadora, etc). Esta parte es la que se necesita reentrenar para las nuevas clases correspondientes a partes del cuerpo.

YOLOv5 proporciona cuatro modelos con distintos tamaños de red adecuados para distintas aplicaciones. Las redes de menor tamaño sacrifican precisión, pero resultan interesantes para aplicaciones en condiciones de tiempo real en equipos embarcados con capacidades de computación limitada. Estas versiones se denominan YOLOv5x, YOLOv5l, YOLOv5m y YOLOv5s en orden decreciente de tamaño. En este trabajo se evalúan dos de estos cuatro modelos, la versión YOLOv5x y YOLOv5m, con el objetivo de analizar la precisión de redes de distinto tamaño en aplicaciones SAR.



(a)

(b)

Figura 1: (a) Vehículo todoterreno y sistema sensorial, (b) Imágenes de cámaras RGB/TIR y nube de puntos 3D del LIDAR.

3. MODELO DE DATOS

3.1. EL DATASET UMA-SAR

Para el entrenamiento hemos utilizado nuestro conjunto de datos *UMA-SAR Dataset* [10], disponible públicamente en www.uma.es/robotics-and-mechatronics/sar-datasets.

Este *dataset* recopila información sensorial multimodal capturada desde un vehículo tripulado todoterreno durante ejercicios realistas de búsqueda y rescate celebrados en 2018 y 2019 en el Área de Experimentación en Nuevas Tecnologías para la Intervención en Emergencias de la Universidad de Málaga [17]. En la Fig. 1(a) se muestra el conjunto sensorial, formado por dos cámaras monoculares sincronizadas de luz visible (RGB) e infrarrojo térmico (TIR), un lidar tridimensional (3D) Velodyne HDL-32, así como una unidad de medición inercial (IMU) y dos receptores del sistema de posicionamiento global (GPS) con los que obtener el *ground truth*.

Nuestra misión en los ejercicios fue recoger una amplia gama de datos del dominio SAR (*Search and Rescue*), incluyendo personas, vehículos, escombros y actividad SAR en terreno no estructurado. En concreto, se recogieron cuatro secuencias de datos siguiendo rutas cerradas durante los ejercicios, con una longitud total de la trayectoria de 5,2 km y un tiempo total de 77 minutos. Además, proporcionamos tres secuencias más del lugar vacío (es decir, antes o después del ejercicio) con fines de comparación (4,9 km adicionales y 46 min). Adicionalmente, los datos se ofrecen tanto en formato legible para el ser humano como en formato de archivos *rosbag*, y se proporcionan dos herramientas de software específicas para extraer y adaptar este conjunto de datos a la preferencia de los usuarios. La Figura 1(b) muestra un ejemplo de imágenes y nubes de puntos 3D capturadas durante el ejercicio de 2019.

El dataset contiene un total de 32.260 imágenes, de las cuales la mitad pertenecen al espectro visible y la otra mitad al infrarrojo térmico. En las imágenes aparecen numerosas instancias de personas, ciudadanos de a pie y personal de emergencias. También se utilizan imágenes obtenidas durante las jornadas sobre Seguridad y Emergencias de 2020 (no incluidas aún en el dataset UMA-SAR) en situaciones de visibilidad escasa, para la validación de la red bajo estas condiciones.

ISBN 978-84-09-41095-8

3.2. DATOS Y CATEGORÍAS PARA EL ENTRENAMIENTO

Para este trabajo se han etiquetado un total de 1142 pares de imágenes (un total de 2284, considerando RGB y su correspondiente TIR) para el entrenamiento, validación y test. Además, se etiquetaron 116 imágenes para tests adicionales en distintos escenarios. A la hora de elección en el espectro infrarrojo térmico, se ha intentado que las imágenes térmicas elegidas correspondan en su mayoría con las imágenes del espectro visible. Para el etiquetado, se hace uso del programa CVAT (*Computer Vision Annotation Tool*). CVAT es una herramienta gratuita de código abierto para anotación de imágenes y vídeos basada en la web. La distribución de imágenes empleada ha sido la siguiente:

- Entrenamiento (70 % del total), para reentrenar el modelo.
- Validación (15 %), utilizadas durante el entrenamiento para ajustar los parámetros del modelo y evitar el sobreentrenamiento.
- Test (15 %), para la evaluación del modelo, posterior al entrenamiento. Además de este 15 % inicial, se han utilizado 116 imágenes adicionales con el fin de evaluar distintos escenarios.

El objetivo de este trabajo es la detección de partes del cuerpo humano. Para ello, el conjunto de clases de interés se compone de los siguientes elementos:

- *Brazo*. Esta clase puede aparecer en distintas posturas, desde doblado hasta completamente extendido. Además, son susceptibles a ser parcialmente ocluidos por otras partes del cuerpo, especialmente el tronco.
- *Pierna*. Al igual que el brazo, puede presentar mucha variabilidad en las imágenes, si bien es menos frecuente su oclusión.
- *Cabeza*. A priori, presenta una menor variabilidad en la imagen.
- *Torso*. Clase de mayor tamaño, con menor variabilidad.

3.3. AUMENTO DE DATOS

El aumento de datos consiste en ampliar el número de imágenes para el entrenamiento aplicando transformaciones a las imágenes originales y a sus etiquetas. Las transformaciones modifican las imágenes para que puedan aportar información, tal y como distintos encuadres del objeto, giros, o condiciones iluminación. En este trabajo, el aumento de datos se ha desarrollado en Python [13], para realizar varias transformaciones sobre las imágenes y sus etiquetas, haciendo uso de la librería de código abierto *Albumentations*.

Sobre las imágenes originales se realizaron cinco transformaciones:

- **Volteo horizontal**. Esta transformación se aplica sobre la totalidad de las imágenes.
- **Cambio en la saturación**. Se aplica un cambio en la saturación en un rango entre el -20 % y el 20 %.
- **Cambio en el contraste**. Cambio en el contraste en un rango entre el -20 % y el 20 %.
- **Rotación**. Rotación en un rango entre -90º y 90º.
- **Escalado**. Cambio de escala en un rango entre el 80 % y el 20 % del tamaño de la imagen.

Después de estas transformaciones, aplicadas sobre la totalidad de las imágenes del dataset utilizado, se obtuvo un dataset aproximadamente cinco veces más grande que el original.

4. ENTRENAMIENTO, EXPERIMENTACIÓN Y RESULTADOS

4.1. METODOLOGÍA

El entrenamiento se ha realizado empleando transferencia de conocimiento, que consiste en utilizar una red entrenada previamente la cual se reentrena para la clasificación según el conjunto de clases de partes del cuerpo. En este caso, se ha utilizado una red pre-entrenada en el dataset COCO, con los pesos proporcionados en el mismo repositorio. El modelo aplica su conocimiento sobre el dataset inicial para la detección de las nuevas clases entrenando la última capa (es decir, la capa de clasificación). En esta capa se define el número de clases de salida, sustituyendo las 80 clases del dataset COCO por las cuatro que resultan de interés en este trabajo. Las otras capas del modelo no se modifican ('congelamiento'del modelo) a fin de aprovechar el entrenamiento del extenso dataset de COCO. De esta forma, se necesitan menos épocas en el proceso de entrenamiento, el cual se hace más eficiente y preciso.

Los parámetros empleados durante el entrenamiento se muestran en la Tabla 1. El entrenamien-

Tabla 1: Parámetros de entrenamiento.

Parámetro	**Valor**
Num. de épocas	5000 (YOLOv5x) y 2000 (YOLOv5m)
Tamaño de lote	8
Tamaño de la imagen	704
Pesos iniciales	Proporcionados por la red entrenada del repositorio

Tabla 2: Comparación de modelos.

Modelo	**mAP@.5**	**mAP@.5:.95**
YOLOv5x RGB	0.656	0.234
YOLOv5x TIR	**0.695**	**0.291**
YOLOv5m RGB	0.638	0.229
YOLOv5m TIR	0.668	0.275

to se ha realizado en una estación de trabajo desarrollada por Nvidia (*WorkStation "Nvidia DGX Station"*) con 4 GPUs NVIDIA Tesla® V100 32 GB/GPU, y una CPU Intel Xeon E5-2698 v4 2.2 GHz 20-Core. En el entrenamiento solo se ha empleado una GPU de las cuatro disponibles en la estación de trabajo.

4.2. COMPARATIVA ENTRE YOLOv5x Y YOLOv5m

En el entrenamiento de los modelos YOLOv5x y YOLOv5m con imágenes en los espectros visible (RGB) e infrarrojo térmico (TIR) se han obtenido las precisiones medias (mAP) de las cuatro clases, mostradas en la Tabla 2. Los mejores resultados para cada modalidad se obtienen con el modelo YOLOv5x, aunque la diferencia con el modelo de mayor tamaño no es muy significativa. Por otra parte, la detección de partes del cuerpo es más efectiva en imágenes térmicas para ambas, lo cual es consistente con la precisión obtenida en trabajos anteriores al detectar personas al completo [3]. Esta ventaja del modelo térmico podría explicarse por que la temperatura característica del cuerpo humano ayuda a mejorar su detección. En cuanto al tiempo de inferencia, se han obtenido valores medios en torno a 3 ms, observándose una reducción media de 0.5 ms para el modelo de menor tamaño. En un sistema embarcado con capacidad de cómputo limitada la diferencia de tiempos puede resultar más significativa.

En la Tabla 3 se muestra la precisión media de cada clase, en el caso del modelo YOLOv5x con imágenes del infrarrojo térmico, resultados correspondientes al caso más favorable mostrado en la Tabla 2. Los mejores resultados se obtienen para las clases cabeza y torso, que son aquellas con menor variabilidad en las imágenes (ver sección 3.2). El resultado menos favorable corresponde a la clase brazo.

Tabla 3: Precisión media para cada clase con el modelo YOLOv5x con imágenes del infrarrojo térmico.

Clase	**mAP@.5**	**mAP@.5:.95**
Todas las clases	0.695	0.291
Cabeza	**0.767**	0.306
Brazo	0.609	0.251
Pierna	0.658	0.267
Torso	0.745	**0.339**

Tabla 4: Comparación entre escenarios para modelo YOLOv5x con imágenes RGB.

Escenario	**mAP@.5**	**mAP@.5:.95**
Cerca	0.617	0.288
Lejos	0.552	0.200
Día	0.693	0.334
Noche	0.0529	0.0239

Tabla 5: Comparación entre escenarios para modelo YOLOv5x con imágenes TIR.

Escenario	**mAP@.5**	**mAP@.5:.95**
Cerca	0.582	0.256
Lejos	0.588	0.234
Día	0.569	0.238
Noche	0.256	0.115

4.3. EVALUACIÓN EN DISTINTAS CONDICIONES DE VISIBILIDAD

Por otro lado, se ha evaluado el modelo YOLOv5x, que ha ofrecido mayor precisión en la detección de las clases, para imágenes de test correspondientes a distintas condiciones de visibilidad y proximidad a las cámaras. Los tests evaluados son:

- Cerca: Escenarios con buenas condiciones de visibilidad y objetos cercanos a la cámara (por debajo de 5 metros).
- Lejos: Escenarios con buenas condiciones de visibilidad y objetos lejanos (a más de 5 metros).
- Día: Escenarios con buenas condiciones de visibilidad.
- Noche: Escenarios con malas condiciones de visibilidad.

Para este análisis ha sido necesario añadir imágenes distintas (un total de 116) a las utilizadas

en el entrenamiento y en la evaluación anterior, ya que el UMA-SAR dataset no incluye imágenes nocturnas.

En las Tablas 4 y 5 se muestran los resultados cualitativos obtenidos en cada escenario, con imágenes en el espectro visible e infrarrojo térmico, respectivamente. Los valores de precisión para las imágenes diurnas y objetos cercanos, correspondientes a las condiciones de los datos de entrenamiento, resultan semejantes a los obtenidos en el apartado anterior. Sin embargo, para imágenes RGB se aprecia una reducción de la precisión para los objetos lejanos y se hace evidente la limitación del espectro visible en condiciones difíciles de iluminación.

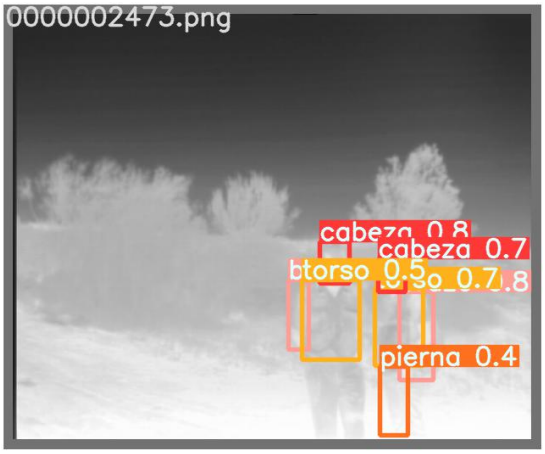
En el caso de las imágenes TIR (ver Tabla 5), el índice de precisión obtenido es menos sensible a la distancia de los objetos. Para imágenes nocturnas, la precisión mejora notablemente respecto a imágenes RGB, si bien se mantiene por debajo de la precisión diurna (más del 50 %). Este resultado puede obedecer a que las imágenes recogidas en el escenario nocturno son notablemente distintas (y más exigentes) que las del UMA-SAR dataset empleado en el entrenamiento y los otros tests. En concreto, los datos nocturnos corresponden a un conjunto de víctimas yacentes (tanto actores reales como *dummies* sin emisión de calor) en un área de escombros con oclusión de partes del cuerpo.

En cuanto a resultados cualitativos, la Figura 2 ilustra un par de imágenes de test RGB/TIR para la misma escena, adquiridas con buena visibilidad y personas cercanas a las cámaras. En este ejemplo, el modelo logra detectar todas las partes del cuerpo (cabeza, torso, brazos y piernas) a excepción de las partes ocluidas. En la imagen TIR, no se detectan todas las piernas, que no son completamente visibles debido a un menor ángulo de visión de la cámara térmica.

La Figura 3 ilustra un par de imágenes con buena visibilidad con personas lejanas. En este caso la detección de partes del cuerpo está limitada por resolución de los objetos en la imagen. Este ejemplo corrobora los resultados de las Tablas 4 y 5 en cuanto que en las imágenes TIR se logra mejor precisión. Así, en la izquierda de la Figura 3.(b) se detecta la cabeza de una persona que no es detectada en la imagen RGB correspondiente. Estos resultados indican que a largas distancias es posible mejorar la detección de personas y partes del cuerpo debido a su patrón de temperatura.

Por último, la Figura 4.(c) ofrece un ejemplo de imagen en el espectro infrarrojo térmico para el escenario de imágenes nocturnas. Se observa cómo el modelo es capaz de detectar alguna parte

(a) Imagen RGB.



(b) Imagen TIR.

Figura 2: Imágenes con objetos detectados por la red YOLOv5x. Escenario: cerca.

del cuerpo de una persona presente en la escena, aunque confunde brazo con cabeza. No obstante, no se detectan partes del cuerpo parcialmente ocluido de la persona en la parte derecha de la imagen. En la Figura 4.(a) se muestra la misma escena en una imagen capturada de día, mientras que en la Figura 4.(b) se muestra la imagen RGB de la escena nocturna.

En resumen, los resultados con imágenes a color son superiores en condiciones de cercanía a la cámara y en buenas condiciones de visibilidad, mientras que en condiciones de visibilidad extrema, como la lejanía a las cámaras y la baja visibilidad nocturna las imágenes del infrarrojo térmico ofrecen un resultado superior.

5. CONCLUSIONES

En este trabajo se ha evaluado el rendimiento de la red YOLOv5 para la detección automática de partes del cuerpo en imágenes de los espectros visible (RGB) e infrarrojo térmico (TIR) en situa-

ISBN 978-84-09-41095-8

(a) Imagen RGB.

(b) Imagen TIR.

Figura 3: Imágenes con objetos detectados por la red YOLOv5x. Escenario: lejos

ciones de búsqueda y rescate (SAR). Con este fin se han etiquetado 2400 imágenes RGB y TIR pertenecientes al UMA-SAR dataset ([10]), etiquetas correspondientes a partes del cuerpo humano. En concreto cuatro categorías: cabeza, torso, piernas y brazos. Se ha comparado el rendimiento de dos tamaños de modelos YOLOv5, se ha seleccionado el que ofrece un mejor rendimiento precisión/velocidad de inferencia y se han validado en cuatro escenarios clasificados por proximidad de personas a las cámaras e imágenes diurnas y nocturnas. La complementariedad de ambos espectros debe combinarse para conseguir mejorar los resultados de precisión en la detección.

Finalmente, como trabajo futuro es importante aumentar el número de imágenes en el dataset en diversas situaciones, que incluyan víctimas yacientes e imágenes en condiciones de visibilidad extrema (noche, polvo, túnel, etc). Por otro lado, la implantación de este trabajo en un vehículo terrestre no tripulado para su uso en misiones SAR en tiempo real es un desarrollo importante para

(a) Imagen diurna.

(b) Imagen nocturna RGB.

(c) Imagen nocturna TIR.

Figura 4: Imágenes con objetos detectados por la red YOLOv5x en escenario nocturno.

avanzar en la detección de víctimas.

Agradecimientos

Este trabajo ha recibido financiación del proyecto nacional RTI2018-093421-B-I00 y de la Universidad de Málaga (Andalucía Tech).

English summary

HUMAN PARTS DETECTION IN SEARCH AND RESCUE MISSIONS

Abstract

Vision systems are essential in search and rescue (SAR) tasks, mainly in crucial missions such as the detection of potential victims in disaster environments. The use of images from the visible (RGB) and thermal infrared (TIR) spectra for object detection are complementary, and allow detection in limited visibility conditions. To this end, this paper analyses how deep learning techniques based on convolutional neural networks (CNN) can be applied to both modalities. The YOLOv5 network is employed to optimise the detection of human body parts in both spectra in disaster situations using the public UMA-SAR dataset. Finally, their results are evaluated under different visibility conditions.

Keywords: Computer vision, deep learning, convolutional neural networks, YOLOv5, thermal infrared imaging

Referencias

[1] Al-Kaff, A., Gómez-Silva, M. J., Moreno, F. M., de la Escalera, A. and Armingol, J. M.: 2019, An appearance-based tracking algorithm for aerial search and rescue purposes, *Sensors* **19**(3).

[2] Barba-Guaman, L., Naranjo, J. E., Ortiz, A. and Gonzalez, J. G. P.: 2021, Object detection in rural roads through SSD and YOLO framework, *Adv Intell Syst Comp* **1365 AIST**, 176–185.

[3] Bañuls, A., Mandow, A., Vázquez-Martín, R., Morales, J. and García-Cerezo, A.: 2020, Object detection from thermal infrared and visible light cameras in search and rescue scenes, *IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 380–386.

[4] Beltrán, J., Guindel, C., Moreno, F. M., Cruzado, D., García, F. and De La Escalera, A.: 2018, BirdNet: A 3D object detection framework from lidar information, *IEEE Conf Intel Transp Sy*, pp. 3517–3523.

[5] Cruz, E., Rangel, J. C., Gomez-Donoso, F. and Cazorla, M.: 2020, How to add new knowledge to already trained deep learning models applied to semantic localization, *Applied Intelligence* **50**(1), 14 – 28.

[6] Cruz Ulloa, C., Prieto Sánchez, G., Barrientos, A. and Del Cerro, J.: 2021, Autonomous thermal vision robotic system for victims recognition in search and rescue missions, *Sensors* **21**(21).

[7] Dubé, R., Cramariuc, A., Dugas, D., Sommer, H., Dymczyk, M., Nieto, J., Siegwart, R. and Cadena, C.: 2020, Segmap: Segment-based mapping and localization using data-driven descriptors, *The International Journal of Robotics Research* **39**(2-3), 339–355.

[8] Jocher, G., Stoken, A., Borovec, J. and others: 2021 (Consultado en mayo de 2022), ultralytics/yolov5: v5.0 - yolov5-p6 1280 models, aws, supervise.ly and youtube integrations.
URL: *https://zenodo.org/record/4679653.YnBjS8lO_I*

[9] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: 2014, Microsoft COCO: Common objects in context, *European Conference on Computer Vision*, Springer International Publishing, pp. 740–755.

[10] Morales, J., Vázquez-Martín, R., Mandow, A., Morilla-Cabello, D. and García-Cerezo, A.: 2021, The UMA-SAR dataset: Multimodal data collection from a ground vehicle during outdoor disaster response training exercises, *The International Journal of Robotics Research* **40**(6-7), 835–847.

[11] Murphy, R. R., Duncan, B. A., Collins, T., Kendrick, J., Lohman, P., Palmer, T. and Sanborn, F.: 2016, Use of a small unmanned aerial system for the sr-530 mudslide incident near oso, washington, *Journal of field Robotics* **33**(4), 476–488.

[12] Petříček, T., Šalanskỳ, V., Zimmermann, K. and Svoboda, T.: 2019, Simultaneous exploration and segmentation for search and rescue, *Journal of Field Robotics* **36**(4), 696–709.

[13] Raschka, S.: 2015, *Python machine learning*, Packt publishing ltd.

[14] Redmon, J., Divvala, S. K., Girshick, R. B. and Farhadi, A.: 2015, You only look once: Unified, real-time object detection, *CoRR* **abs/1506.02640**.
URL: *http://arxiv.org/abs/1506.02640*

[15] Sabater, A., Montesano, L. and Murillo, A. C.: 2020, Robust and efficient post-processing for video object detection, *IEEE International Conference on Intelligent Robots and Systems*, pp. 10536 –10542.

[16] Sampedro, C., Rodriguez-Ramos, A., Bavle, H., Carrio, A., de la Puente, P. and Campoy, P.: 2019, A fully-autonomous aerial robot for

search and rescue applications in indoor environments using learning-based techniques, *Journal of Intelligent and Robotic Systems: Theory and Applications* **95**(2), 601 – 627.

[17] Universidad de Málaga: 2021 (Consultado en julio de 2021), Área de experimentación en nuevas tecnologías para la intervención en emergencias (LAENTIEC).
URL: *www.uma.es/LAENTIEC*

[18] Zhao, Z.-Q., Zheng, P., Xu, S.-t. and Wu, X.: 2019, Object detection with deep learning: A review, *IEEE transactions on neural networks and learning systems* **30**(11), 3212–3232.