Artículo Científico / Scientific Paper

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología
https://doi.org/10.17163/ings.n27.2022.08
pISSN: 1390-650X / eISSN: 1390-860X

Detección de peatones en el día y en la noche usando YOLO-v5

Pedestrian detection at daytime and nighttime conditions based on YOLO-v5

Bryan Montenegro[1,3], Marco Flores-Calero[2,3,*]

Recibido: 13-05-2021, Recibido tras revisión: 12-08-2021, Aceptado: 13-09-2021, Publicado: 01-01-2022

Resumen

En este artículo se presenta un nuevo algoritmo basado en aprendizaje profundo para la detección de peatones en el día y en la noche, denominada multiespectral, enfocado en aplicaciones de seguridad vehicular. La propuesta se basa en YOLO-v5, y consiste en la construcción de dos subredes que se enfocan en trabajar sobre las imágenes en color (RGB) y térmicas (IR), respectivamente. Luego se fusiona la información, a través, de una subred de fusión que integra las redes RGB e IR, para llegar a un detector de peatones. Los experimentos, destinados a verificar la calidad de la propuesta, fueron desarrollados usando distintas bases de datos públicas de peatones destinadas a su detección en el día y en la noche. Los principales resultados en función de la métrica mAP, estableciendo un IoU en 0.5 son 96.6 % sobre la base de datos INRIA, 89.2 % sobre CVC09, 90.5 % en LSIFIR, 56 % sobre FLIR-ADAS, 79.8 % para CVC14, 72.3 % sobre Nightowls y KAIST un 53.3 %.

Palabras clave: infrarrojo, color, multiespectral, peatones, aprendizaje profundo, YOLO-v5

Abstract

This paper presents new algorithm based on deep learning for daytime and nighttime pedestrian detection, named multispectral, focused on vehicular safety applications. The proposal is based on YOLO-v5, and consists of the construction of two subnetworks that focus on working with color (RGB) and thermal (IR) images, respectively. Then the information is merged, through a merging subnetwork that integrates RGB and IR networks to obtain a pedestrian detector. Experiments aimed at verifying the quality of the proposal were conducted using several public pedestrian databases for detecting pedestrians at daytime and nighttime. The main results according to the mAP metric, setting an IoU of 0.5 were: 96.6 % on the INRIA database, 89.2 % on CVC09, 90.5 % on LSIFIR, 56 % on FLIR-ADAS, 79.8 % on CVC14, 72.3 % on Nightowls and 53.3 % on KAIST.

Keywords: Infrared, color, multispectral, pedestrian, deep learning, YOLO-v5

[1]Ingeniería en Electrónica, Automatización y Control, Universidad de las Fuerzas Armadas ESPE, Av. Gral. Rumiñahui s/n, PBX 171-5-231B, Sangolquí (Pichincha), Ecuador.
[2]Departamento de Eléctrica, Electrónica y Telecomunicaciones, Universidad de las Fuerzas Armadas ESPE, Av. Gral. Rumiñahui s/n, PBX 171-5-231B, Sangolquí (Pichincha), Ecuador.
[3,*]Departamento de Sistemas Inteligentes I&H Tech. Autor para correspondencia ✉: mjflores@espe.edu.ec.

Forma sugerida de citación: Montenegro, B. y Flores-Calero, M. "Detección de peatones en el día y en la noche usando YOLO-v5," *Ingenius, Revista de Ciencia y Tecnología*, N.° 27, pp. 85-95, 2022. DOI: https://doi.org/10.17163/ings.n27.2022.08.

1. Introducción

En la actualidad, los accidentes de tráfico son un problema de salud pública a nivel mundial, porque ocasionan un alto número de víctimas y lesionados, costos de tratamientos médicos, rehabilitación, alteraciones psicológicas, seguros personales y materiales, consumen recursos que podrían destinarse a otros campos de la salud [1], donde los peatones están expuestos a un alto porcentaje de accidentalidad, llegando hasta el 22 % de los casos [2]. Muchos de estos infortunios pueden ser evitados, debido a que son generados por la acción riesgosa, negligente o irresponsable de los conductores y/o los mismos peatones [3]. En el caso del Ecuador, los atropellamientos representan más del 10 % de las defunciones por accidentes de tráfico.

En este escenario, los sistemas de detección de peatones (SDP) son uno de los componentes tecnológicos más importantes para evitar posibles situaciones de peligro y reducir los atropellamientos. Por lo tanto, la detección de peatones es un tema de investigación activo y desafiante, debido a los retos que se deben superar al trabajar en ambientes no controlados y con sensores limitados en la percepción de la escena vial.

En el caso de las condiciones atmosféricas, el exceso de sol, las lluvias, la niebla o la neblina cambian las condiciones de iluminación, y para peor, la noche magnifican estos factores de riesgo debido a la falta de luz natural [4–6]. Respecto a los peatones, estos usan diferentes tipos de ropa, en colores variados, cambian la postura del cuerpo y pueden estar en cualquier posición de la escena vial. En lo que tiene que ver con la información captada por la cámara, en general, es incompleta debido al reducido campo de visión del sensor, la distancia que separa al peatón de la cámara, disminuye la resolución de la imagen capturada. El movimiento y la vibración del vehículo generan distorsión de la imagen. Además, la geometría de la carretera incide directamente en la calidad de la información captada por la cámara [5], [7].

Afortunadamente, hoy por hoy existen bases de datos públicas, especializadas en la detección de peatones, en el día, en la noche, en conjunto o por separado, en el contexto de vehículos inteligentes y autónomos, que pueden ser usadas para la parte experimental [8–10].

Así, el principal objetivo de este trabajo es instrumentar una nueva arquitectura de aprendizaje profundo (*DL, Deep Learning*) basada en YOLO-v5 [4], [11–15], para obtener un sistema de vanguardia y especializado en la detección de peatones en la noche y/o en el día, usando información visual en el rango de la luz visible y en el infrarrojo, que genere resultados comparables a los existentes en el estado de la cuestión.

El contenido de este documento está organizado de la siguiente manera: la sección 2 presenta el estado de la cuestión en el campo de los SDP usando técnicas DL. A continuación, el apartado 3 describe la arquitectura del sistema de detección basado en una nueva arquitectura basada en YOLO-v5 para la clasificación/detección de peatones en la noche y/o en el día. El siguiente apartado exhibe los resultados de la evaluación experimental, desarrollada sobre varias bases de datos públicas destinada a la implementación de SDP; en el día y la noche. Finalmente, la última parte está dedicada a las conclusiones y los trabajos futuros.

1.1. Estado de la cuestión

Actualmente, las arquitecturas DL están siendo ampliamente usadas en la construcción de SDP, cuyo objetivo es la detección de peatones en escenarios reales de conducción [4], [6], [12], [15, 16]. Para este fin se han usado cámaras en el rango de la luz visible (imágenes RGB) y en el infrarrojo, lejano o cercano, (imágenes IR) para captar la información visual en el día y en la noche, en conjunto o por separado.

Así, Kim *et al.* [17] usaron CNN sobre imágenes nocturnas capturadas con una cámara del espectro visible. Los experimentos se han desarrollado sobre las bases de datos KAIST [18] y CVC-14 [10].

Ding *et al.* [18] pusieron en funcionamiento una arquitectura CNN basada en dos subredes R-FCN, una red para imágenes en color y otra para térmicas. Las subredes de gran medida, térmica y en color, se fusionan en la mitad de la arquitectura, de manera similar para las subredes de pequeña medida. Obteniendo detecciones por separado para peatones de grande y pequeña escala, al final de la red se utiliza el algoritmo de NMS (no máxima supresión) para fusionar los resultados de las dos subredes y obtener una detección robusta. Con la fusión de los dos canales se disminuye la tasa de error versus FPPI del 40 % a 34 %, que se obtiene con los canales por separado. Además, el porcentaje de pérdidas con R-FCN es del 69 %, mientras que con Faster-RCNN es del 51 %.

Kïng *et al.* [5] han instalado una red RPN para detectar personas en el espectro visible y en el infrarrojo; luego para fusionar la información han utilizado la técnica Boosted Decision Tree obteniendo una tasa de error del 29.83 % sobre la base de datos KAIST [19].

Kim *et al.* [16] combinaron RPN y Boosted Forest para la detección de peatones sobre las bases de datos Caltech [20], INRIA [21], ETHZ y KITTI [22]; para mejorar el entrenamiento utilizaron técnicas de boostrap para llegar a una tasa de error del 9.6 %; el algoritmo tiene un tiempo de procesamiento de 0.6 segundos por fotograma. Además, comprobaron que Faster R-CNN no funciona adecuadamente, debido a que los mapas de características no presentan la suficiente información para detectar peatones a larga distancia, lo que resulta una desventaja a ser resuelta.

Zhang *et al.* [15] desarrollaron una arquitectura Faster R-CNN en el espectro visible e infrarrojo. Los resultados experimentales se desarrollaron sobre la base de datos Caltech [20], y en situaciones nocturnas sobre una base propia, obteniendo una tasa de error del 19 % y del 24 %, respectivamente, con un tiempo de procesamiento de 103 milisegundos (9.7 fps) sobre imágenes de 640 × 480 píxeles.

Liu *et al.* [4] emplearon una arquitectura Faster-RCNN para la detección de peatones en los espectros visible e infrarrojo, con una tasa de error del 37 % sobre la base de datos KAIST [19].

Song *et al.* [11] propusieron una red híbrida basada en Yolo-v3 llamada MSFFN (*multispectral feature fusion network*), la cual está compuesta por una estructura DarkNet-53 y dos subredes MFEV y MFEI, para imágenes en color e infrarrojo, respectivamente. Los mapas de características de MFEV se dividen en tres escalas de (13 × 13), (26 × 26) y (52 × 52), de manera análoga para MFEI, para ser fusionados en la parte final de la arquitectura. MSFFN alcanza un mAP de 85.4 %, con respecto al 84.9 % de Faster-RCNN sobre KAIST [19], otro aspecto sobresaliente es los 56 fps de MSFFN, contra los 28 fps de Faster-RCNN.

Cao *et al.* [8] expusieron mejoras en los parámetros para la detección en YOLO-v3, modificando el tamaño de cuadrícula a (10 × 10), aplicando Soft-NMS en lugar de NMS, con un umbral de 0.2 de superposición y, finalmente, agregando un nuevo mapa de características de (104 × 104). Los experimentos los realizaron sobre INRIA [21], obteniendo una precisión del 93.74 % y un recall de 88.14 %, con una velocidad de procesamiento 9.6 milisegundos por fotograma.

Yu *et al.* [23] modificaron Faster R-CNN, concatenando tres niveles diferentes de VGG16 con las ROI, luego se los normaliza, escala y dimensiona. Con estos cambios se obtienen un *miss-rate* (MR) de 10.31 % sobre la base de datos INRIA [21].

Zhou *et al.* [24] propusieron un sistema para mejorar el rendimiento en la detección con oclusión con su red MSFMN (*Mutual-Supervised Feature Modulation Network*), compuesto por dos ramas supervisadas por anotaciones de cuerpo completo y de partes visibles, que genera ejemplos de entrenamiento mejor enfocados. Además, se calcula la semejanza en la pérdida entre las cajas de cuerpo completo y las partes visibles, permitiendo aprender características más robustas, principalmente para peatones ocluidos. La fusión se realiza al final multiplicando los dos puntajes de clasificación. Los experimentos los desarrollaron sobre la base de datos CityPersons [24] obteniendo un 38.45 % para una fuerte oclusión.

Por otra parte, Tesema *et al.* [25] pusieron en marcha una arquitectura híbrida que recibe el nombre de HCD (SDS-RPN), con un Log-average Miss Rate de 8.62 % sobre Caltech [20]. Por otra parte, Kyrkou [26] presentó el sistema YOLOPED que se basa en la arquitectura DenseNet. En lugar de FPN, cada resolución se redimensiona al tamaño del mapa de características más profundo en la columna, permitiendo combinarlos mediante una concatenación, la cual es usada en la detección de cabecera. Finalmente, se implanta una nueva función de pérdida, combinando las características de YoloV2 [27], SSD [28] y lapNet [29].

Evaluando en PETS2009, se obtiene una precisión del 85.7 %, miss rate del 12 %, con un procesamiento de 33.3 fps. Wolpert *et al.* [12] han propuesto combinar imágenes RGB y térmicas, usando Faster R-CNN sin cajas de anclaje, adaptando la arquitectura CSP-Net [12] para fusionar las imágenes IR en el final de la arquitectura, alcanzando un promedio de MS 7.40 % sobre KAIST [19]. Zhou *et al.* [30] han presentado la red MBNet (*Modality Balance Network*), basada en SSD con un módulo DMAF (*Differential Modality Aware Fusion*), el cual fusiona y complementa la información entre las características RGB y térmicas.

La detección IAFA (*Illumination Aware Feature Alignment*) maneja el equilibrio entre las dos modalidades en la detección, el desempeño alcanza un miss rate de 21.1 % y 8.13 % sobre CVC-14 [10] y KAIST [19] respectivamente. Wang [31] utiliza una arquitectura llamada CSP, compuesta por una parte de extracción de características basada en Resnet-101 y una etapa de detección, la cual a su vez es usada para predecir el centro, escalar y offset. En el cual usan Batch Normalization (BN), para acelerar el proceso de entrenamiento y mejorar el desempeño de las CNN. Otra técnica más reciente es Switch Normalizacion (SN), la cual emplea un promedio ponderado de la media y la varianza estadística de la normalización por bloques.

Para el modelo CSP se comprobó que usando BN se obtiene 11.29 % MR (miss rate), mientras que, SN obtiene un 10.91 % MR en la base de datos CityPersons. Escalar adecuadamente las imágenes ayudan a disminuir la carga computacional y a eliminar ruido, CSP con SN y una entrada de (1024 × 2048), se tiene un 11.41 % MR, mientras que, con una entrada de (640 × 1280), se tiene un MR de 10.80 %. Shopovska *et al.* [32] presentaron una arquitectura similar a las redes generativas adversarias (GAN). Esta red tiene dos entradas, una RGB y una térmica, dando como salida una imagen que mantiene los peatones con buena visibilidad, mientras que, con la información obtenida de las imágenes térmicas realza el color de los peatones con mala visibilidad.

Esta imagen es utilizada de entrada para red Faster RCNN VGG16, con lo cual se obtiene un 52.07 % MR y 43.25 % MR, para las imágenes del día y la noche, respectivamente en la base de datos KAIST [19], en CVC-14 [10] se obtienen un 69.14 % MR y 63.52 % MR, para imágenes en el día y en la noche, respectivamente.

2. Materiales y métodos

La Figura 1 muestra el esquema general del sistema multiespectral propuesto para la detección de peatones. El sistema toma la información visual, proveniente de las imágenes en color o térmicas, para alimentar dos subredes, denominadas RBG e IR, respectivamente. Luego, la red de fusión concatena las salidas para localizar peatones en el día y en la noche, de manera conjunta o por separado. Las subredes están compuestas por una arquitectura basada en YOLO-v5 (You Only Look Once) [11], [26], [33–35].

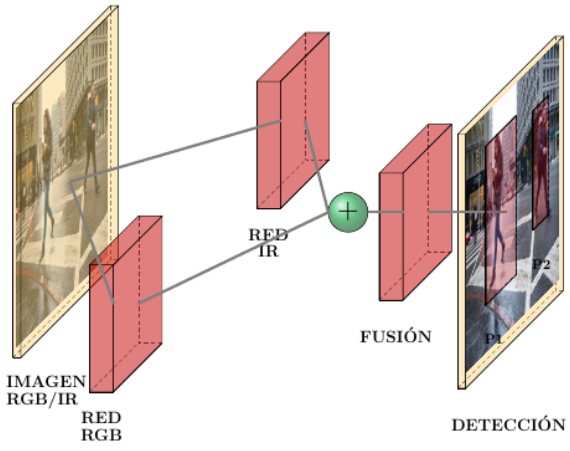

Figura 1. Esquema general del sistema multiespectral para la detección de peatones sobre imágenes en color y térmicas, basado en YOLO-v5

2.1. Descripción de la arquitectura YOLO-v5

YOLO es un acrónimo de «You Only Look Once» [11], [27], [33–35]. Es un modelo muy popular y de alto rendimiento en el campo de detección de objetos, es considerado como la tecnología de punta en detecciones en tiempo real (FPS). YOLO-v5 es la quinta generación de los detectores de una sola etapa [36]. YOLO-v5 está implementado en Pytorch. La Tabla 1 muestra la composición de las capas personalizadas que describen la arquitectura, en función de las capas base de Pytorch.

En la Tabla 1, la sigla SF es un acrónimo para *Scale Factor*, por otra parte, el símbolo #s representa parámetros variables que se manejan de acuerdo con los valores establecidos en la columna de parámetros de la Tabla 2, estos definen principalmente el tamaño del Kernel, Stride, Padding y Canales.

Finalmente, el símbolo – representa que no recibe ningún parámetro.

La Figura 2 muestra la arquitectura YOLO-v5, que constituyen las subredes IR y RGB; con las capas mencionadas en la Tabla 1.

Tabla 1. Composición de las capas personalizadas implementadas en YOLO-v5 [36]

Nombre	**Composición**	**Parámetros**		
		Kernel	**Stride**	**Canales**
Conv	conv2d	#	#	#
	BatchNom2d	-	-	-
	Hardwish	-	-	-
Focus	Conv	3 × 3	1	32
	concat	-	-	-
BottleNeckCSP	Conv	3 × 3	1	#
	Conv	3 × 3	1	#
	Conv	3 × 3	1	#
	conv2d	3 × 3	1	#
	conv2d	3 × 3	1	#
	concat	-	-	-
	BatchNom2d	-	-	-
	LeakyRelu	-	-	-
	Conv	3 × 3	1	#
SPP	Conv	3 × 3	1	512
	-	**Kernel**	**Stride**	**Padding**
	Maxpool2d	5 × 5	1	2
	Maxpool2d	9 × 9	1	4
	Maxpool2d	13 × 13	1	6
	concat	-	-	-
	Conv	3 × 3	1	512
Upsample	nn.Upsample	**Size**	**SF**	**Mode**
		none	2	nearest

2.2. Arquitectura propuesta

La arquitectura propuesta se enfoca en crear un sistema capaz de fusionar dos subredes que trabajan con imágenes RGB e IR, respectivamente. La red de fusión concatena las capas 17 y 40 (peatones pequeños), y las capas 20 y 43 (peatones grandes), descritos en la Tabla 2, para localizar peatones en el día y en la noche, de manera conjunta o por separado.

La Tabla 2 muestra las capas específicas que componen cada una de las subredes; cada capa cuenta con un identificador (id), el cual se utiliza en procedencia para identificar a qué capas están conectadas. La procedencia –1 indica que es una conexión a la capa anterior; el número indica la cantidad de veces que se repite la capa, por último, en parámetros se indica los argumentos que recibe cada capa.

Las capas que contienen los mapas de características de las redes RGB e IR están concatenadas para la fusión de la información, a través, de una capa BottleneckCSP. Esta información combinada es la que se envía a la capa de detección para la generación de los cuadros delimitadores y la predicción de clase.

Tabla 2. Distribución y conexiones de subredes que conforman la arquitectura del sistema para la detección de peatones en el día y la noche, basado YOLO-v5 [36]

Red	Id	Procedencia	Número	Módulo	Parámetros
RGB	0	–1	1	Focus	[32,3]
	1	–1	1	Conv	[64,3,2]
	2	–1	3	BottleneckCSP	[64]
	3	–1	1	Conv	[128,3,2]
	4	–1	9	BottleneckCSP	[128]
	5	–1	1	Conv	[256,3,2]
	6	–1	9	BottleneckCSP	[256]
	7	–1	1	Conv	[512,3,2]
	8	–1	1	SPP	[512,[5,9,13]]
	9	–1	3	BottleneckCSP	[512,False]
	10	–1	1	Conv	[1]
	11	–1	1	Upsample	[256,False]
	12	[–1,6]	1	concat	[1]
	13	–1	3	BottleneckCSP	[256,False]
	14	–1	1	Conv	[128,1,1]
	15	–1	1	Upsample	[None,2,Nearest]
	16	[–1,4]	1	concat	[1]
	17	–1	3	BottleneckCSP	[128,False]
	18	–1	1	Conv	[128,3,2]
	19	[–1,14]	1	concat	[1]
	20	–1	3	BottleneckCSP	[256,False]
	21	–1	1	Conv	[256,3,2]
	22	[–1,10]	1	concat	[1]
	23	–1	3	BottleneckCSP	[512,False]
IR	24	0	1	Conv	[64,3,2]
	25	–1	3	BottleneckCSP	[64]
	26	–1	1	Conv	[128,3,2]
	27	–1	9	BottleneckCSP	[128]
	28	–1	1	Conv	[256,3,2]
	29	–1	9	BottleneckCSP	[256]
	30	–1	1	Conv	[512,3,2]
	31	–1	1	SPP	[512,[5,9,13]]
	32	–1	3	BottleneckCSP	[512,False]
	33	–1	1	Conv	[1]
	34	–1	1	Upsample	[256,False]
	35	[–1,29]	1	concat	[1]
	36	–1	3	BottleneckCSP	[256,False]
	37	–1	1	Conv	[128,1,1]
	38	–1	1	Upsample	[None,2,Nearest]
	39	[–1,27]	1	concat	[1]
	40	–1	3	BottleneckCSP	[128,False]
	41	–1	1	Conv	[128,3,2]
	42	[–1,37]	1	concat	[1]
	43	–1	3	BottleneckCSP	[256,False]
	44	–1	1	Conv	[256,3,2]
	45	[–1,33]	1	concat	[1]
	46	–1	3	BottleneckCSP	[512,False]
Fusión	47	[17,40]	1	concat	[1]
	48	–1	3	BottleneckCSP	[128,False]
	49	[20,43]	1	concat	[1]
	50	–1	3	BottleneckCSP	[256,False]
	51	[23,46]	1	concat	[1]
	52	–1	3	BottleneckCSP	[512,False]
Detect	53	[48,50,52]	3	Detect	[1, anchors]

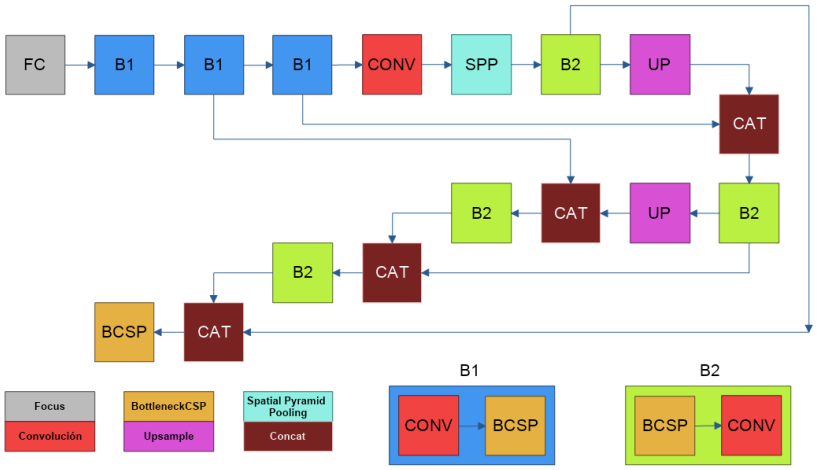

Figura 2. Representación gráfica de la arquitectura YOLO-v5

3. Resultados y discusión

Para llegar al modelo propuesto se han desarrollado múltiples experimentos, usando bases de datos de referencia en el estado de la cuestión y, las métricas estándar de evaluación destinados a la detección de objetos.

3.1. Descripción de las bases de datos

Las bases de datos públicas de peatones, en los espectros visible e infrarrojo, son INRIA [21], CVC 09 [9], CVC-14 [10], LSI *Far Infrared Pedestrian Dataset* (LSI-FIR) [37], FLIR-ADAS [38], Nightowls [39] y KAIST [19].

Estas bases de datos fueron elegidas porque están especializadas en aplicaciones vehiculares durante el día y la noche, e incluyen la etiquetación de la región verdadera, B_{gt}, donde se localizan efectivamente los peatones.

- **INRIA** [21]. La base de datos pública INRIA es una de las más utilizadas en detección de peatones. Cuenta con un conjunto de imágenes divididas en «*train*» y «*test*»; la carpeta «train» contiene 614 imágenes para el entrenamiento mientras que, la carpeta «test» incluye 288 imágenes para test. En la Tabla 3 se muestra el contenido.

Tabla 3. Contenido de la base de datos INRIA

	Detección
Entrenamiento	614(614)[a]
Prueba	288(288)

.[a] El valor entre paréntesis representa el número de fotogramas que contienen peatones.

- **CVC-09** [9]. Estas son las bases de datos más utilizadas para la detección de peatones en la noche y en el día, respectivamente. En este caso se la usó para el entrenamiento, y posteriormente para la validación. En la Tabla 4 se describen los conjuntos de entrenamiento y de prueba. Esta base de datos viene etiquetada con los peatones presentes en la escena, B_{gt}.

Tabla 4. Contenido de la base de datos CVC-09 durante la noche

	Positivos	**Negativos**
Entrenamiento	2200	1002
Prueba	2284	-

- **LSIFIR** [37]. Es otra base de datos importante para el desarrollo de algoritmos de detección de peatones en la noche. En la Tabla 5 se describen los conjuntos de entrenamiento y de prueba, con sus respectivos tamaños. En este caso al igual que CVC09 se la usó para el entrenamiento, validación y prueba de la propuesta.

Tabla 5. Contenido de la base de datos LSI FIR

	Clasificación	**Detección**
Entrenamiento	43 391(10 209)[a]	2936(3225)
Prueba	22 051(5945)	5788(3279)

.[a] El valor entre paréntesis representa el número de fotogramas que contienen peatones.

- **FLIR-ADAS** [38]. Esta base cuenta con imágenes térmicas para el desarrollo sistemas de conducción autónoma. El objetivo de estas imágenes es ayudar al desarrollo de sistemas más seguros, que, combinados con imágenes en color, información de sensores LIDAR, se pueda crear un sistema robusto para la detección de peatones. Con unas 8862 imágenes para el entrenamiento y 5838 para el *test*, ver Tabla 6.

Tabla 6. Contenido de la base de datos FLIR-ADAS

	Detección
Entrenamiento	8862(5838)[a]
Prueba	1366(1206)

.[a] El valor entre paréntesis representa el número de fotogramas que contienen peatones.

- **CVC-14** [10]. Está compuesta por dos secuencias de imágenes térmicas tomadas durante el día y la noche. Con más de 6000 imágenes para el entrenamiento y 700 para validación.

- **Nightowls** [39]. Se enfoca en la detección de peatones en la noche. Las imágenes son capturadas con una cámara estándar, con una resolución de 1024 × 640. Las secuencias fueron capturadas en tres países, bajo todas las condiciones climáticas y en todas las estaciones, para obtener una mayor variabilidad de escenas.

- **KAIST** [19]. Base de datos multiespectral que cuenta con un conjunto de imágenes de 640 × 480, tomadas por dos cámaras una térmica y otra en color con una frecuencia de 20 Hz. Tomadas durante el día y la noche para considerar distintas condiciones de iluminación. Existen la misma cantidad de imágenes térmicas y en color con un total de 100 368 imágenes para entrenamiento y 90 280 para el test, ver Tabla 7.

Tabla 7. Contenido de la base de datos KAIST

	Detección	
	Color	**Térmica**
Entrenamiento	50 184(#)[a]	50 184(#)
Prueba	45 140(#)	45 184(#)

.[a] El valor entre paréntesis representa el número de fotogramas que contienen peatones.

3.2. Métricas de evaluación

Para la evaluación se seguirán los siguientes protocolos:

- Curva P-R (*Precision-Recall*). La precisión (Pres) es la fracción de casos relevantes entre los casos recuperados. El recall (*Rec*) es la fracción de casos relevantes que se han recuperado sobre la cantidad total de casos relevantes. Las ecuaciones para estos casos son las siguientes:

$$Pres = \frac{TP}{TP + FP} \tag{1}$$

$$Rec = \frac{TP}{TP + FN} \tag{2}$$

- AP (*Average Precision*). Este índice fue propuesto para el desafío VOC2007 [40] para evaluar el desempeño de detectores, y está relacionado con el área bajo la curva, de la curva P-R, de una clase. El mAP que es un promedio de los AP de todas las clases.

Para estimar las métricas se necesita un índice que permita identificar una correcta predicción, en este caso es IoU (*Intersection-over-Union*). IoU determina la relación entre las regiones que corresponden a los verdaderos positivos (TP) y falsos positivos (FP), mediante (3).

$$IoU = \frac{Area(B_{det} \cap B_{gt}}{Area(B_{det} \cup B_{gt}} \tag{3}$$

Donde B_{gt} es la ROI verdadera y B_{det} es la ROI detectada. En este caso, se tiene un TP si el valor de IoU mayor a 0.5, caso contrario es un FP. Con estos valores se puede evaluar las ecuaciones (1) y (2).

3.3. Detalles de la implementación

La arquitectura propuesta cuenta con cuatro partes principales que son las subredes IR y RGB, el bloque de fusión de características y el bloque de detección. El entrenamiento de la arquitectura contara con una etapa de entrenamiento de ajuste fuerte y una etapa de entrenamiento de ajuste fino. Para el entrenamiento de ajuste fuerte se utiliza el algoritmo de optimización SGD (*stochastic gradient descent*) y una tasa de aprendizaje (LR, *Learning rate*) de 0.01, esta técnica evita quedarse estancado en un mínimo relativo de la función de optimización, y se fijan 100 épocas para el entrenamiento de la arquitectura completa con las imágenes RGB. A continuación, se congelan los pesos correspondientes a la subred RGB, para fijar 100 épocas de entrenamiento a la arquitectura con las imágenes IR.

Finalmente, para concluir con la etapa de ajuste fuerte congelaremos los pesos correspondientes a las

subredes IR y RGB, entrenaremos 50 épocas las capas de fusión con las imágenes IR y RGB combinadas en partes iguales para evitar que las capas de fusión sean segadas por las características de las imágenes IR o RGB.

En la etapa de ajuste fino se modifica LR a 0.0001, se congelan todos los pesos de la arquitectura excepto los correspondientes a la subred RGB, luego se entrenan 50 épocas con las imágenes RGB, consecuentemente se congelan todos los pesos excepto los de la subred IR y entrenamos 50 épocas con imágenes IR. Como último paso se congelan todos los pesos excepto los de la capa de fusión y se realiza un entrenamiento de 25 épocas con las imágenes IR y RGB en partes iguales.

Al momento, este procedimiento fue aplicado a cada una de las bases de datos anotadas en este trabajo.

3.4. Resultados

En la Tabla 8 se presenta el desempeño del método de detección, al ser evaluado con varias métricas sobre las bases de datos elegidas.

En todos los casos, el tiempo de procesamiento fue de 29.8 milisegundos.

A continuación, en la Figura 3 se exhiben los gráficos de las curvas P-R de la arquitectura propuesta sobre cada una de las bases de datos elegidas. A partir de la Tabla 8 y la Figura 3 se puede concluir que el mejor desempeño se realiza sobre INRIA [21], seguido de CVC09 [9] y LSIFIR [37].

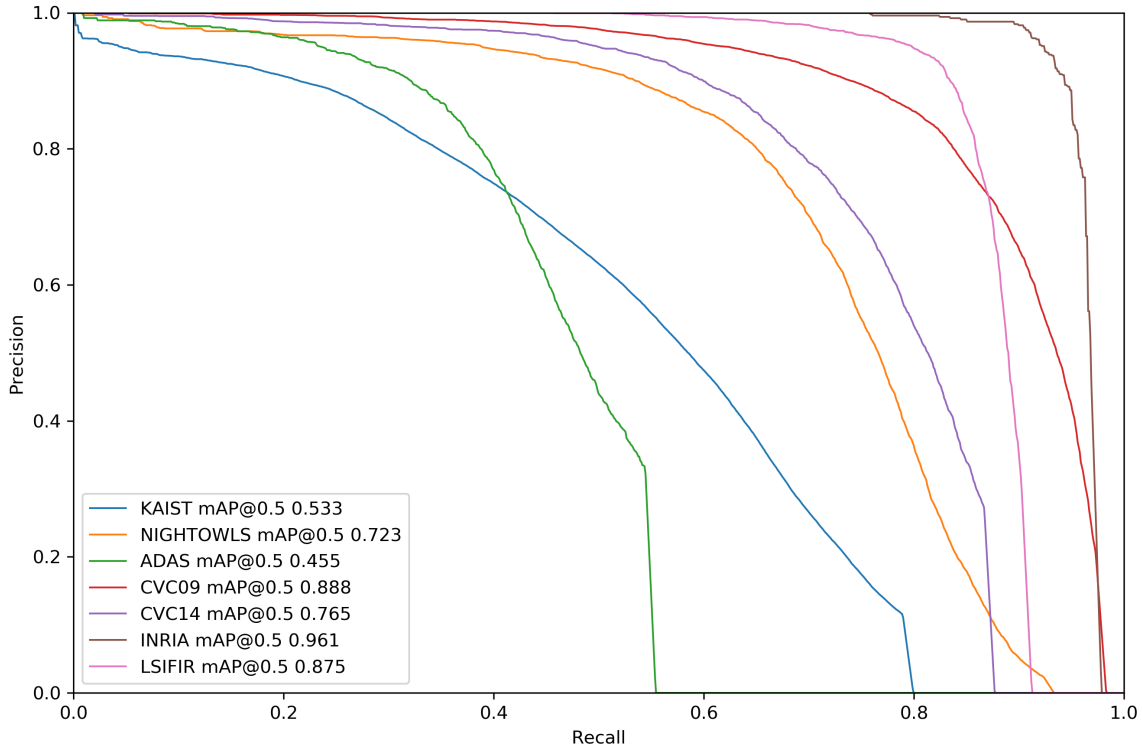

Figura 3. Gráficas de las curvas P-R sobre las distintas bases de datos de peatones

Tabla 8. Evaluación de la arquitectura Yolo-v5 [40], sobre varias bases de datos públicas en el espectro visible e infrarrojo. LAMS es un acrónimo para Log Average Miss Rate

	INRIA	**CVC09**	**LSIFIR**	**FLIR-ADAS**	**CVC14**	**Nightowls**	**KAIST**
mAP@50	96.6	89.2	90.5	56	79.8	72.3	53.3
Precisión	69.8	67.4	89.2	72.1	86.4	80.7	52.5
Recall	90	89	83.4	40.1	61.6	64.6	53.7
LAMS	6	20	17	69	36	36	67

4. Conclusiones

En este trabajo se ha presentado un sistema para la detección de peatones en el día y en la noche usando modernas técnicas de procesamiento de imágenes y aprendizaje profundo, donde desarrolló una nueva arquitectura DL basada en YOLO-v5, con DenseNet, para la detección de peatones en el día y en la noche usando imágenes en el espectro visible y en el infrarrojo lejano, cuyo mAP es de 96.6 % para el caso INRIA, 89.2 % sobre CVC09, 90.5 % en LSIFIR, 56 % sobre FLIR-ADAS, 79.8 % para CVC14, 72.3 % sobre Nightowls y 53.3 % para KAIST.

Como trabajo futuro se plantea perfeccionar la arquitectura propuesta y probarla sobre las bases de datos más relevantes en este campo del conocimiento.

Agradecimientos

El equipo de computación, GPU, ha sido financiado por la empresa I&H Tech.

Además, deseamos dar las gracias a los investigadores que han publicado sus bases de datos y arquitecturas de aprendizaje profundo, para el uso de la comunidad científica; sin ellos no hubiese sido posible este trabajo.

Finalmente, deseamos expresar nuestros agradecimientos a los revisores anónimos que colaboran.

Referencias

[1] WHO. (2018) Road traffic injuries. World Health Organization. [Online]. Available: https://bit.ly/3pmr9Rc

[2] ANT. (2015) Estadísticas de siniestros de tránsito octubre 2015. Agencia Nacional de Tránsito del Ecuador. [Online]. Available: https://bit.ly/3aUIWGv

[3] ——. (2017) Estadísticas de siniestros de tránsito agosto 2017. Agencia Nacional de Tránsito del Ecuador. [Online]. Available: https://bit.ly/3aUIWGv

[4] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," 2016. [Online]. Available: https://bit.ly/2Z3BLJu

[5] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 243–250. [Online]. Available: https://doi.org/10.1109/CVPRW.2017.36

[6] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019. [Online]. Available: https://doi.org/10.1016/j.inffus.2018.11.017

[7] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018. [Online]. Available: https://doi.org/10.1109/TMM.2017.2759508

[8] J. Cao, C. Song, S. Peng, S. Song, X. Zhang, Y. Shao, and F. Xiao, "Pedestrian detection algorithm for intelligent vehicles in complex scenarios," *Sensors*, vol. 20, no. 13, p. 3646, 2020. [Online]. Available: https://doi.org/10.3390/s20133646

[9] Caltech. (2016) Caltech pedestrian detection benchmark. [Online]. Available: https://bit.ly/3aXuZb4

[10] Pascal. (2016) Inria person dataset. [Online]. Available: https://bit.ly/30APbxi

[11] X. Song, S. Gao, and C. Chen, "A multispectral feature fusion network for robust pedestrian detection," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 73–85, 2021. [Online]. Available: https://doi.org/10.1016/j.aej.2020.05.035

[12] A. Wolpert, M. Teutsch, M. S. Sarfraz, and R. Stiefelhagen, "Anchor-free small-scale multispectral pedestrian detection," 2020. [Online]. Available: https://bit.ly/3G8k5gL

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: https://bit.ly/3B167d1

[14] C. Ertler, H. Possegger, M. Opitz, and H. Bischof, "Pedestrian detection in RGB-D images from an elevated viewpoint," in *Proceedings of the 22nd Computer Vision Winter Workshop*, W. Kropatsch, I. Janusch, and N. Artner, Eds. Austria: TU Wien, Pattern Recongition and Image Processing Group, 2017. [Online]. Available: https://bit.ly/3AYTI9w

[15] X. Zhang, G. Chen, K. Saruta, and Y. Terata, "Deep convolutional neural networks for all-day pedestrian detection," in *Information Science and Applications 2017*, K. Kim and N. Joukov, Eds. Singapore: Springer Singapore, 2017, pp. 171–178. [Online]. Available: https://doi.org/10.1007/978-981-10-4154-9_21

[16] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 443–457. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_28

[17] J. H. Kim, H. G. Hong, and K. R. Park, "Convolutional neural network-based human detection in nighttime images using visible light camera sensors," *Sensors*, vol. 17, no. 5, 2017. [Online]. Available: https://doi.org/10.3390/s17051065

[18] L. Ding, Y. Wang, R. Laganiere, D. Huang, and S. Fu, "Convolutional neural networks for multispectral pedestrian detection," *Signal Processing: Image Communication*, vol. 82, p. 115764, 2020. [Online]. Available: https://doi.org/10.1016/j.image.2019.115764

[19] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1037–1045. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298706

[20] Caltech. (2012) Caltech pedestrian detection benchmark. [Online]. Available: https://bit.ly/3pkn93o

[21] Pascal. (2012) INRIA person dataset. [Online]. Available: https://bit.ly/3IAO6Hw

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. [Online]. Available: https://bit.ly/3n6oBnq

[23] X. Yu, Y. Si, and L. Li, "Pedestrian detection based on improved faster rcnn algorithm," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, 2019, pp. 346–351. [Online]. Available: https://doi.org/10.1109/ICCChina.2019.8855960

[24] Y. He, C. Zhu, and X.-C. Yin, "Mutual-supervised feature modulation network for occluded pedestrian detection," 2020. [Online]. Available: https://bit.ly/3C14eyn

[25] F. B. Tesema, H. Wu, M. Chen, J. Lin, W. Zhu, and K. Huang, "Hybrid channel based pedestrian detection," *Neurocomputing*, vol. 389, pp. 1–8, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.12.110

[26] C. Kyrkou, "Yolopeds: efficient real time single shot pedestrian detection for smart camera applications," *IET Computer Vision*, vol. 14, no. 7, pp. 417–425, Oct 2020. [Online]. Available: http://dx.doi.org/10.1049/iet-cvi.2019.0897

[27] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. [Online]. Available: https://bit.ly/3nuyCv1

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_2

[29] F. Chabot, Q.-C. Pham, and M. Chaouch, "Lapnet : Automatic balanced loss and optimal assignment for real-time dense object detection," 2020. [Online]. Available: https://bit.ly/3FYZDPo

[30] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," 2020. [Online]. Available: https://bit.ly/2Z6qKaV

[31] W. Wang, "Detection of panoramic vision pedestrian based on deep learning," *Image and Vision Computing*, vol. 103, p. 103986, 2020. [Online]. Available: https://doi.org/10.1016/j.imavis.2020.10398

[32] I. Shopovska, L. Jovanov, and W. Philips, "Deep visible and thermal image fusion for enhanced pedestrian visibility," *Sensors*, vol. 19, no. 17, 2019. [Online]. Available: https://doi.org/10.3390/s19173727

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. [Online]. Available: https://bit.ly/3aWg3tO

[34] D. Heo, E. Lee, and B. Chul Ko, "Pedestrian detection at night using deep neural networks and saliency maps," *Journal of Imaging Science and Technology*, vol. 61, no. 6, pp. 604 031–604 039, 2017. [Online]. Available: https://doi.org/10.2352/J.ImagingSci.Technol.2017.61.6.060403

[35] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://bit.ly/30Lg81v

[36] G. Jocher, A. Stoken, J. Borovec, NanoCode012, A. Chaurasia, TaoXie, L. Changyu, V. Abhiram, Laughing, tkianai, yxNONG, A. Hogan,

lorenzomammana, AlexWang1900, J. Hajek, L. Diaconu, Marc, Y. Kwon, oleg, wanghaoyang0106, Y. Defretin, A. Lohia, ml5ah, B. Milanko, B. Fineran, D. Khromov, D. Yiwei, Doug, Durgesh, and F. Ingham, "ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4679653

[37] D. Olmeda, C. Premebida, U. Nunes, J. M. Armingol, and A. de la Escalera, "Pedestrian detection in far infrared images," *Integrated Computer-Aided Engineering*, vol. 20, no. 4, pp. 347–360, 2013. [Online]. Available: http://dx.doi.org/10.3233/ICA-130441

[38] Teledyne Flir. (2021) Free flir thermal dataset for algorithm training. Teledyne FLIR LLC All rights reserved. [Online]. Available: https://bit.ly/2Xxe3F4

[39] NightOwls. (2021) About nightowls. NightOwls Datasets. [Online]. Available: https://bit.ly/3pof6m9

[40] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4