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

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

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

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

Received: 13-05-2021, Received after review: 12-08-2021, Accepted: 13-09-2021, Published: 01-01-2022

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

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

[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. Corresponding author ✉: mjflores@espe.edu.ec.

Suggested citation: Montenegro, B. and Flores-Calero, M. "Pedestrian detection at daytime and nighttime conditions based on 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. Introduction

At present, car accidents are a public health problem worldwide, since they cause a high number of victims and injured, medical treatment costs, rehabilitation, psychological disorders, personal and property insurance, they consume resources that might be aimed at other health areas [1], where pedestrians are exposed to a high accident rate, reaching up to 22 % of the cases [2]. Many of these misfortunes may be prevented, because they are generated by the risky, negligent and irresponsible action of drivers and/or the pedestrians themselves [3]. In the case of Ecuador, run overs represent more than 10 % of the deaths due to car accidents.

In this scenario, pedestrian detection systems (PDS) are one of the most important technological components to prevent possible dangerous situations and reduce run overs. Therefore, pedestrian detection is an active and interesting research topic, due to the challenges that must be overcome when working in uncontrolled environments and with limited sensors in the perception of the road scene.

In the case of atmospheric conditions, excessive sun, rains, or mist change lighting conditions, and to make matters worse, the night magnifies these risk factors due to the absence of natural light [4–6]. With respect to pedestrians, they use different types of clothes, in different colors, change the body posture and may be at any position of the road scene. Regarding the information captured by the camera, it is generally incomplete due to the reduced field of vision of the sensor, the distance that separates the pedestrian from the camera reduces the resolution of the captured image. The movement and vibration of the vehicle generate distortion of the image. In addition, the geometry of the road has direct influence on the quality of the information captured by the camera [5], [7].

Fortunately, at present there are public databases specialized in pedestrian detection, at daytime or nighttime, together or separate, in the context of intelligent and autonomous vehicles, which may be used for the experimental part [8–10].

Thus, the main objective of this work is the implementation of a new deep learning (DL) architecture based on YOLO-v5 [4], [11–15], to obtain a cutting-edge system specialized in pedestrian detection at nighttime and/or daytime, using visual information in the range of visible and infrared light, that generates results comparable to the existing ones in the state of the art.

The content of this document is organized as follows: section 2 presents the state of the art in the field of the PDS using DL techniques. Afterwards, section 3 describes the architecture of the detection system based on YOLO-v5 for classification/detection of pedestrians at nighttime and/or daytime. The following section shows the results of the experimental evaluation, conducted on various public databases aimed at the implementation of PDS, at daytime or nighttime. Finally, the last part is devoted to the conclusions and future works.

1.1. State of the art

At present, DL architectures are being widely used in the construction of PDS, whose objective is the detection of pedestrians in real driving scenarios [4], [6], [12], [15], [16]. For this purpose, cameras in the range of visible light (RGB images) and infrared (IR images) have been used, to capture visual information at daytime or nighttime, far or close, together or separate.

Thus, Kim *et al.* [17] used CNNs on night images captured with a visible spectrum camera. The experiments were conducted on the KAIST [18] and CVC-14 [10] databases.

Ding *et al.* [19] put into operation a CNN architecture based on two R-FCN subnetworks, one for color images and another for thermal images. Large subnetworks, thermal and color, were merged in the middle of the architecture; it was done similarly for small subnetworks. To obtain separate detections for pedestrians of large and small scale, the NMS (non-maximum suppression) algorithm is used at the end of the network to merge the results of the two subnetworks and obtain a robust detection. By merging the two channels, the error rate versus FPPI is reduced from 40 %, obtained with separate channels, to 34 %. In addition, the percentage of losses with R-FCN is 69 %, whereas with Faster-RCNN is 51 %.

Köing *et al.* [5] have installed an RPN network for detecting persons in the visible and infrared spectra; then, they have used the Boosted Decision Tree technique to merge the information, obtaining an error rate of 29.83 % on the KAIST database [18].

Zhang *et al.* [16] combined RPN and Boosted Forest for detecting pedestrians on the Caltech [20], INRIA [21], ETHZ and KITTI [22] databases; they used bootstrap techniques to improve the training, reaching an error rate of 9.6 %; the algorithm has a processing time of 0.6 seconds per frame. In addition, they proved that Faster R-CNN does not work properly, because the feature maps do not have enough information to detect pedestrians at a great distance, which results in a drawback to be overcome.

Zhang *et al.* [15] developed a Faster R-CNN architecture in the visible and infrared spectra. The experimental results were obtained on the Caltech database *20*, and in nighttime situations on an own database, obtaining an error rate of 19 % and 24 %, respectively, with a processing time of 103 milliseconds (9.7 fps) on 640 × 480 pixels images.

Liu *et al.* [4] used a Faster-RCNN architecture for detecting pedestrians in the visible and infrared spectra, with an error rate of 37 % on the KAIST database [18].

Song *et al.* [11] proposed a hybrid network based on Yolo-v3 called MSFFN (multispectral feature fusion network), which consists of a DarkNet-53 structure and two subnetworks, MFEV and MFEI for color and infrared images, respectively. The feature maps of MFEV are divided in three scales of (13 × 13), (26 × 26) and (52 × 52), and analogously for MFEI, and then merged in the final part of the architecture. MSFFN achieves a mAP of 85.4 % compared to the 84.9 % of Faster-RCNN on KAIST [18], another remarkable aspect is the 56 fps of MSFFN, compared to the 28 fps of Faster-RCNN.

Cao *et al.* [8] presented improvements in the parameters for the detection in YOLO-v3, modifying the size of the grid to (10 × 10), applying Soft-NMS instead of NMS, with a superposition threshold of 0.2 and, finally, adding a new feature map of (104 × 104). The experiments were conducted on INRIA [21], obtaining a precision of 93.74 % and a recall of 88.14 %, with a processing speed of 9.6 milliseconds per frame.

Yu *et al.* [23] modified the Faster R-CNN, concatenating three different levels of VGG16 with the ROIs, which is then normalized, scaled and dimensioned. A miss-rate (MR) of 10.31 % was obtained on the INRIA database [21] with these changes.

Zhou *et al.* [24] proposed a system to improve the performance in the detection with occlusion with their MSFMN (Mutual-Supervised Feature Modulation Network), constituted by two branches supervised by annotations of entire body and visible parts, that generates training examples which are better focused. In addition, it is calculated the similarity in the losses between the boxes corresponding to entire body and visible parts, enabling learning more robust features, mainly for occluded pedestrians. The merge is performed at the end, multiplying the two classification scores. The experiments were conducted on the CityPersons database [24], obtaining a 38.45 % for a strong occlusion.

On the other hand, Tesema *et al.* [25] put into operation a hybrid architecture that receives the name of HCD (SDS-RPN), with a Log-average Miss Rate of 8.62 % on Caltech [20]. On the other hand, Kyrkou [26] presented the YOLOPED system which is based on the DenseNet architecture. Instead of FPN, each resolution is resized to the size of the deepest feature map in the column, enabling to combine them through a concatenation which is used in header detection. At last, a new loss function is implemented, combining the features of YoloV2 [27], SSD [28] and lapNet [29].

An evaluation in PETS2009 yielded a precision of 85.7 %, a miss rate of 12 %, with a processing of 33.3 fps. Wolpert *et al.* [12] have proposed to combine RGB and thermal images, using Faster R-CNN without anchor boxes, adapting the CSPNet [12] architecture to merge the IR images at the end of the architecture, reaching an MS average of 7.40 % on KAIST [18]. Zhou *et al.* [30] have presented the MBNet (Modality Balance Network), based on SSD with a DMAF (Differential Modality Aware Fusion) module, which merges and complements the information between the RGB and thermal features. The IAFA (Illumination Aware Feature Alignment) detection handles the equilibrium between the two detection modalities, and the performance achieves miss rates of 21.1 % and 8.13 % on CVC-14 [10] and KAIST [18], respectively.

Wang [31] uses an architecture called CSP, constituted by a feature extraction part based on Resnet-101 and a detection stage, which in turn is used to predict the center, scalar and offset. They use Batch Normalization (BN) to accelerate the training process and improve the performance of the CNNs. A most recent technique is Switch Normalization (SN), which uses a weighted average of the statistical mean and variance of the normalization by blocks. It was proved that using BN for the CSP model an MR (miss rate) of 11.29 % was obtained, whereas SN yields a MR of 10.91 % on the CityPersons database.

Appropriate scaling of images helps to reduce the computational load and to eliminate noise, using CSP with SN and an input of (1024 × 2048) yields an MR of 11.41 %, whereas an MR of 10.80 % is achieved with an input of (640 × 1280). Shopovska *et al.* [32] presented an architecture similar to the generative adversarial networks (GAN). This network has two inputs, an RGB and a thermal, giving as output an image that maintains the pedestrians with good visibility, whereas the information obtained from the thermal images enhances the color of pedestrians with bad visibility. This image is used as the input to a Faster RCNN VGG16 network, yielding MRs of 52.07 % and 43.25 %, for daytime and nighttime images, respectively, in the KAIST database [18], and MRs of 69.14 % and 63.52 % for daytime and nighttime images, respectively, in the CVC-14 database [10].

2. Materials and methods

Figure 1 shows the general scheme of the proposed multispectral system for pedestrian detection. The system takes visual information coming from color or thermal images, to feed two subnetworks, named RGB and IR, respectively. Then, the merging network concatenates the outputs to locate pedestrians at daytime or nighttime, jointly or separately. The subnetworks are constituted by an architecture based on YOLO-v5 (You Only Look Once) [11], [26], [33–35].

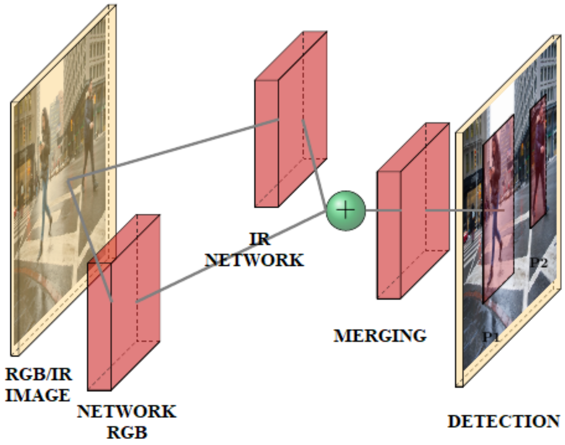

Figure 1. General scheme of the multispectral system based on YOLO-v5, for pedestrian detection on color and thermal images

2.1. Description of the YOLO-v5 architecture

YOLO is an acronym for «You Only Look Once» [11], [27], [33–35]. It is a very popular model with high-performance in the field of object detection, being considered a cutting-edge technology in real-time detection (FPS). YOLO-v5 is the fifth generation of one-stage detectors [36]. Yolo-v5 is implemented in Pytorch. Table 1 shows the composition of the customized layers that describe the architecture, according to the base layers of Pytorch.

In Table 1, SF is an acronym for Scale Factor; on the other hand, the symbol #s represents variable parameters which are handled according to the values established in the column of parameters in Table 2, which mainly define the size of the Kernel Stride, Padding and Channels.

Finally, the symbol – represents that it receives no parameter.

Figure 2 shows the YOLO-v5 architecture, constituted by subnetworks IR and RGB, with the layers mentioned in Table 1.

Table 1. Composition of the customized layers implemented in YOLO-v5 [36]

Name	**Composition**	**Parameters**		
		Kernel	**Stride**	**Channels**
Conv	conv2d	#	#	#
	BatchNom2d	-	-	-
	Hardwish	-	-	-
Focus	Conv	3 x 3	1	32
	concat	-	-	-
BottleNeckCSP	Conv	3 x 3	1	#
	Conv	3 x 3	1	#
	Conv	3 x 3	1	#
	conv2d	3 x 3	1	#
	conv2d	3 x 3	1	#
	concat	-	-	-
	BatchNom2d	-	-	-
	LeakyRelu	-	-	-
	Conv	3 x 3	1	#
SPP	Conv	3 x 3	1	512
	-	**Kernel**	**Stride**	**Padding**
	Maxpool2d	5 x 5	1	2
	Maxpool2d	9 x 9	1	4
	Maxpool2d	13 x 13	1	6
	concat	-	-	-
	Conv	3 x 3	1	512
Upsample	nn.Upsample	**Size**	**SF**	**Mode**
		none	2	nearest

2.2. Proposed architecture

The proposed architecture is focused on creating a system capable of merging two subnetworks that work with RGB and IR images, respectively. The merging network concatenates layers 17 and 40 (small pedestrians), and layers 20 and 43 (large pedestrians), described in Table 2, to locate pedestrians at daytime or nighttime, jointly or separately.

Table 2 shows the specific layers that constitute each of the subnetworks; each layer has an identifier (id), which is used in origin to identify the layers to which they are connected. The origin –1 indicates that it is a connection to the previous layer; the number indicates the number of times that the layer is repeated, and finally, the arguments received by each layer are indicated in parameters.

The layers that contain the feature maps of the RGB and IR networks are concatenated, to merge the information through a BottleneckCSP layer. This combined information is sent to the detection layer to generate bounding boxes and the class prediction.

Table 2. Distribution and connections of the subnetworks that constitute the architecture of the system based on YOLO-v5 [36], for detecting pedestrians at daytime and nighttime

Network	Id	Origin	Number	Module	Paremeters
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]

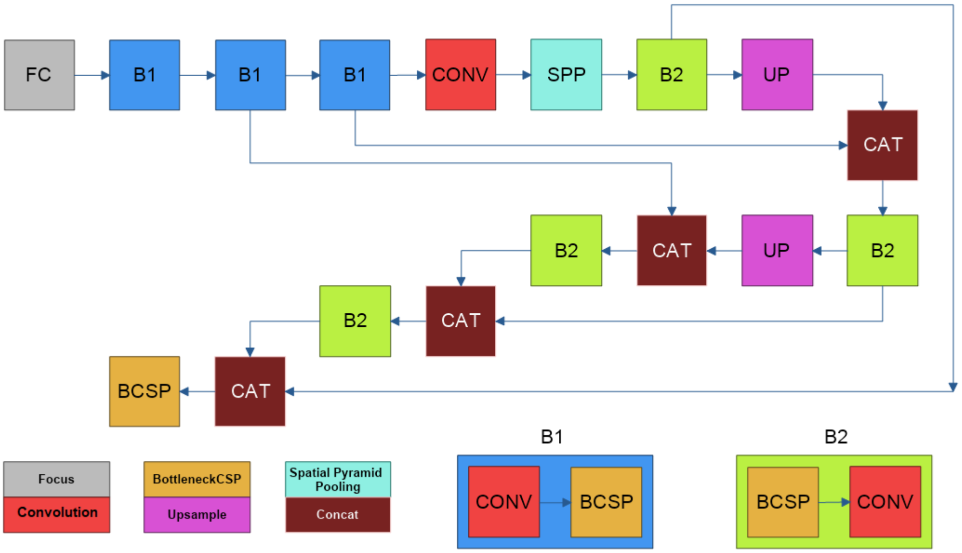

Figure 2. Graphical representation of the YOLO-v5 architecture

3. Results and discussion

Multiple experiments have been conducted to get the proposed model, using reference databases in the state of the art, and standard evaluation metrics used in object detection.

3.1. Description of the databases

The public pedestrian databases in the visible and infrared spectra are INRIA [21], CVC 09 [9], CVC-14 [10], LSI Far Infrared Pedestrian Dataset (LSIFIR) [37], FLIR-ADAS [38], Nightowls [39] and KAIST [18].

These databases were selected because they are specialized in daytime and nighttime vehicle applications, and include labeling of the true region, B_{gt}, where pedestrians are effectively located.

- **INRIA** [21]. The INRIA public database is one of the most widely used in pedestrian detection. It has a set of images divided in «train» and «test»; the «train» folder contains 614 images for training, whereas the «test» folder includes 288 images for testing. Table 3 shows the content.

Table 3. Content of the INRIA database

	Detection
Train	614(614)^a
Test	288(288)

.^{a} The value in parenthesis represents the number of frames that contain pedestrians.

- **CVC-09** [9]. These are the databases most widely used for detecting pedestrians at nighttime and daytime, respectively. In this case it was used for training, and afterwards for validation. Table 4 describes the train and test sets. This database is labeled with the pedestrians present in the scene, B_{gt}.

Table 4. Content of the CVC-09 database at nighttime

	Positive	Negative
Train	2200	1002
Test	2284	-

- **LSIFIR** [37]. It is another important database for developing algorithms for pedestrian detection at nighttime. Table 5 describes the train and test sets, with their corresponding sizes. As it was the case for the CVC-09, this database was used for the training, validation and testing of the proposal.

Table 5. Content of the LSI FIR database

	Classification	Detection
Train	43 391(10 209)^a	2936(3225)
Test	22 051(5945)	5788(3279)

.^{a} The value in parenthesis represents the number of frames that contain pedestrians.

- **FLIR-ADAS** [38]. This database includes thermal images for developing autonomous driving systems. The objective of these images is to help in the development of safer systems, which combined with color images and information from LIDAR sensors, may enable creating a robust system for pedestrian detection. It has 8862 images for training and 5838 for testing, see Table 6.

Table 6. Content of the FLIR-ADAS database

	Detection
Train	8862(5838)[a]
Test	1366(1206)

.[a] The value in parenthesis represents the number of frames that contain pedestrians.

- **CVC-14** [10]. It is constituted by two sequences of thermal images taken at daytime and nighttime. It includes more than 6000 images for training and 700 for validation.

- **Nightowls** [39]. It is focused on the detection of pedestrians at nighttime. The images were captured using a standard camera, with a 1024 × 640 resolution. The sequences were captured in three countries, under all weather conditions and at all seasons, to obtain a greater variability in the scenes.

- **KAIST** [18]. Multispectral database that contains a set of 640 × 480 images, taken by two cameras, one thermal and one color with a frequency of 20 Hz. They were taken at daytime and nighttime to consider different lighting conditions. The number of thermal and color images is the same, for a total of 100,368 images for training and 90,280 for testing, see Table 7.

Table 7. Content of the KAIST database

	Detection	
	Color	**Thermal**
Trains	50 184(#)[a]	50 184(#)
Test	45 140(#)	45 184(#)

.[a] The value in parenthesis represents the number of frames that contain pedestrians.

3.2. Evaluation metrics

The following protocols will be followed for the evaluation:

- P-R Curve (Precision-Recall). Precision (Prec) is the ratio between relevant cases and cases recovered. Recall (Rec) is the ratio between relevant cases that have been recovered and total of relevant cases. The equations for these cases are the following:

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

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

- AP (Average Precision). This index was proposed for the VOC2007 challenge [40] to evaluate the performance of detectors, and is related to the area under the P-R curve of one class. The mAP is an average of the APs for all classes.

To estimate the metrics, it is required an index that enables identifying a correct prediction, which in this case is the IoU (Intersection-over-Union). IoU determines the ratio between the regions that correspond to true positives (TP) and false positives (FP), by means of (3).

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

Where B_{gt} is the true ROI and B_{det} is the detected ROI. In this case, a TP occurs for an IoU greater than 0.5; otherwise, it is an FP. Equations (1) and (2) may be evaluated with these values.

3.3. Implementation details

The proposed architecture is constituted by four main parts, which are the IR and RGB subnetworks, the feature merging block and the detection block. The training of the architecture will consist of a training stage of strong adjustment, and a training stage of fine adjustment. The SGD (stochastic gradient descent) optimization algorithm with a learning rate (LR) of 0.01 is used for the training of strong adjustment, fixing 100 epochs for training the whole architecture with the RGB images; the SGD technique prevents being stuck in a relative minimum of the objective function. Then, the weights corresponding to the RGB subnetwork are frozen, and 100 epochs are fixed for training the architecture with the IR images.

At last, to conclude the strong adjustment stage, the weights corresponding to the IR and RGB subnetworks will be frozen, and the merge layers will be trained for 50 epochs with the IR and RGB images combined in equal parts, to prevent the merge layers from being biased by the features of the IR or RGB images.

In the fine adjustment stage the LR is modified to 0.0001, all weights of the architecture are frozen except the ones corresponding to the RGB subnetwork and the training is performed for 50 epochs with the RGB images; then, all the weights are frozen except the ones of the IR subnetwork, and the training is carried out for 50 epochs with the IR images. In the last step all the weights are frozen except the ones of the merge layer, and a training is performed for 25 epochs with the IR and RGB images in equal parts.

At this time, this procedure was applied to each of the databases listed in this work

3.4. Results

Table 8 presents the performance of the detection method, when it is evaluated with various metrics on the selected databases.

In all cases, the processing time was 29.8 milliseconds.

Figure 3 displays plots of the P-R curves for the proposed architecture on each of the selected databases. It may be concluded from Table 8 and Figure 3 that the best performance was obtained on INRIA [21], followed by CVC09 [9] and LSIFIR [37].

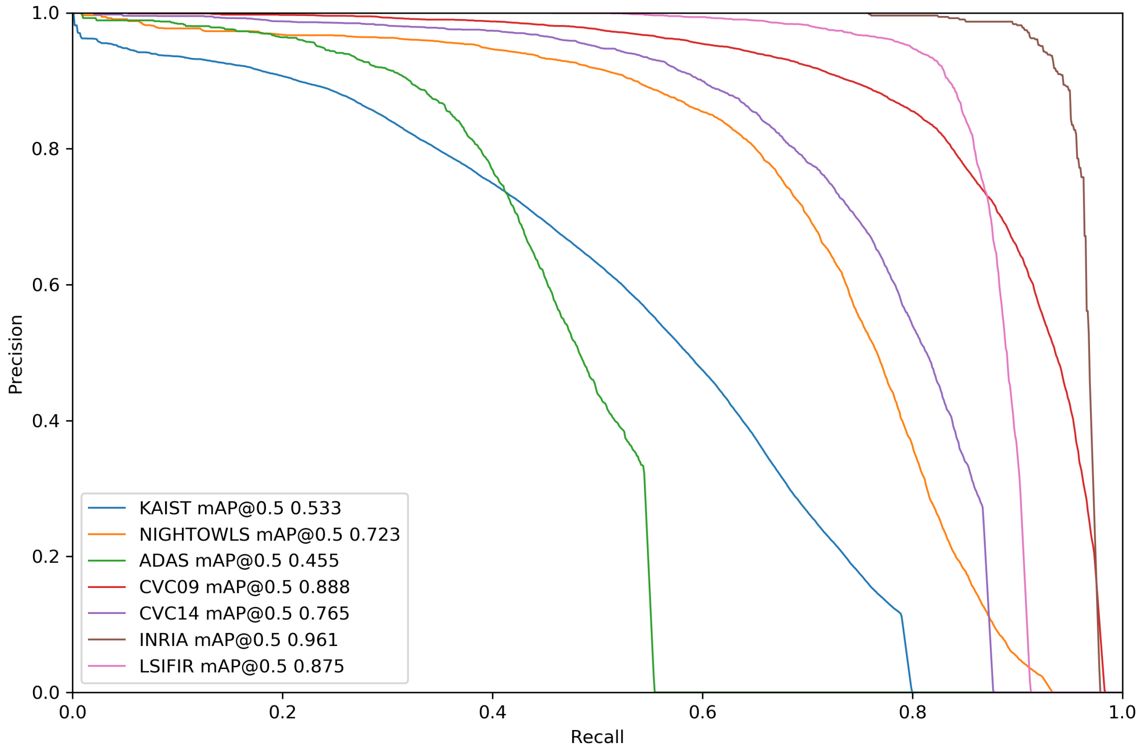

Figure 3. Plots of the P-R curves for the different pedestrian databases

Table 8. Evaluation of the Yolo-v5 [36] architecture on various public databases on the visible and infrared spectra. LAMS is an acronym for 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. Conclusions

This work has presented a system for detecting pedestrians at daytime and nighttime using modern image processing techniques and deep learning, where a new DL architecture based on YOLO-v5 was developed with DenseNet, for detecting pedestrians at daytime and nighttime using images in the visible and far infrared spectra, whose mAP is 96.6 % for the case of INRIA, 89.2 % on CVC09, 90.5 % on LSIFIR, 56 % on FLIR-ADAS, 79.8 % for CVC14, 72.3 % on Nightowls and 53.3 % for KAIST.

Future work will be aimed at improving the proposed architecture and testing it on the most relevant databases in this field of knowledge.

Acknowledgements

The computing equipment, GPU, has been funded by the I&H Tech company.

In addition, we wish to thank the researchers that have published their databases and deep learning architectures, for the use of the scientific community; this work would not have been possible without them.

Finally, we wish to express our acknowledgements to the anonymous reviewers for their collaboration.

References

[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] 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

[19] 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

[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