R B V
Badajoz 2024

ACTAS DEL SIMPOSIO DE ROBÓTICA, BIOINGENIERÍA Y VISIÓN POR COMPUTADOR

Universidad de Extremadura.
Escuela de Ingenierías industriales
Badajoz, 29 a 31 de mayo de 2024

UNIVERSIDAD DE EXTREMADURA

CEA

SIMPOSIO DE ROBÓTICA, BIOINGENIERÍA Y VISIÓN POR COMPUTADOR

Actas

Badajoz, 29-31 de Mayo de 2024

UNIVERSIDAD DE EXTREMADURA

Cáceres, 2024

Actas de Simposio de Robótica, Bioingeniería y Visión por Computador
Badajoz, 29-31 de Mayo de 2024

Editores: Santiago Salamanca Miño
Emiliano Pérez Hernádez
Patricia Arroyo Muñoz
Antonio J. Calderón Godoy
Isaías González Pérez
Jesús Lozano Rogado
Pilar Merchán García
José Ignacio Suárez Marcelo
Inés Tejado Balsera
Blas M. Vinagre Jara

1ª edición, 2024

Edita:
Universidad de Extremadura. Servicio de Publicaciones
Plaza de Caldereros, 2. 10003 Cáceres (España)
Tel. 927 257 041; Fax 927 257 046
publicac@unex.es
http://publicauex.unex.es/

E- ISBN.: 978-84-9127- 262-5 (edición digital)

Imagen de la portada generada con DALL-E de OpenAI

Comité Organizador

Patricia Arroyo Muñoz
Antonio J. Calderón Godoy
Isaías González Pérez
Jesús Lozano Rogado
Pilar Merchán García
Emiliano Pérez Hernádez
Santiago Salamanca Miño
José Ignacio Suárez Marcelo
Inés Tejado Balsera
Blas M. Vinagre Jara

Comité Científico

Óscar Reinoso García	(UMH)	Robótica
Eduardo Rocon de Lima	(CSIC)	Bioingeniería
Luis Payá Castelló	(UMH)	Visión por Computador

Prefacio

Este simposio representará un punto de encuentro único para la presentación y discusión de los trabajos más recientes de los grupos temáticos de Robótica, Bioingeniería y Visión por Computador del Comité Español de Automática (CEA). Investigadores, académicos y profesionales convergerán en este espacio propicio para el intercambio de conocimientos y la exploración de colaboraciones futuras.

El escenario elegido para este evento es la Escuela de Ingenierías Industriales de la Universidad de Extremadura. Con Badajoz como telón de fondo, esta ciudad impregnada de historia y cultura, los participantes podrán disfrutar no solo de la riqueza científica-tecnológica del evento, sino también de la belleza y hospitalidad que ofrece la región de Extremadura.

El programa abarca una amplia gama de temas de los 3 grupos, sesiones paralelas, charlas plenarias, mesas redondas, presentaciones de empresas y un reconfortante programa social, proporcionando un ambiente propicio para el networking y el establecimiento de conexiones duraderas entre los participantes.

Los trabajos aceptados por los revisores de los distintos grupos temáticos han sido 25 de robótica, 16 de bioingeniería y 5 de visión por computador. Todos estos artículos son los que se encuentran recogidos en estas actas publicadas por el Servicio de Publicaciones de la UEx.

Óscar Reinoso García
Coordinador Grupo Robótica CEA

Eduardo Rocon de Lima
Coordinador Grupo Bioingeniería CEA

Luis Payá Castelló
Coordinador Grupo Visión por Computador CEA

ÍNDICE

Simposio de Robótica, Bioingeniería y Visión por Computador 2024

Sesión: Visión por computador

Fusión estática de imágenes del espectro visible y térmico para una mejor detección de personas mediante Redes Neuronales Convolucionales: un análisis del rendimiento

Heredia-Aguado, E.[a,*], Valiente, D., Gil, A., Máximo, M., Paya, L.

[a] *Instituto de Investigación en Ingeniería de Elche (I3E). Universidad Miguel Hernández de Elche. Avda. de la Universidad s/n, 03202 Elche (Alicante), Spain*

To cite this article: Heredia-Aguado, E., Valiente, D., Gil, A., Máximo, M., Paya, L. 2024. Fusión estática de imágenes del espectro visible y térmico para una mejor detección de personas mediante Redes Neuronales Convolucionales: un análisis del rendimiento.
Simposio de Robótica, Bioingeniería y Visión por Computador, 2024, 1-5. https://orcid.org/0009-0001-7717-1428

Resumen

Este artículo presenta diferentes métodos de fusión estática para imágenes compuestas de dos espectros, imagen del espectro visible RGB e imagen térmica (infrarrojo lejano) para ser empleadas en tareas de detección mediante redes neuronales convolucionales. Aunque las tareas de detección mediante imagen están muy desarrolladas, siguen estando fuertemente limitadas por las condiciones de iluminación, tanto del *dataset* como de las condiciones de trabajo. Esta limitación hace que estas técnicas no consigan un grado de fiabilidad suficiente para ser aplicadas a gran escala. Las imágenes térmicas añaden información que, de por sí, es invariante a las condiciones de iluminación, complementando las imágenes del espectro visible allí donde son menos robustas. En este trabajo se analiza el rendimiento de diferentes técnicas que, al margen del *dataset* empleado, permiten fusionar los cuatro canales de información (RGBT) para aprovechar la potencia de detección de algoritmos empleados en imágenes RGB. Enfocados en operaciones de búsqueda y salvamento, seguridad o vigilancia, se pretenden detectar personas de manera robusta. Se aprovechará la potencia de YOLOv8 haciendo uso de un *dataset* de imágenes multiespectro: Kaist.

Palabras clave: imagen térmica, detección de personas, fusion imagen multiespectral, aprendizaje profundo, visión por computador

Abstract

This paper presents a review of different image fusion methods for the combination of visible spectrum images with thermal spectrum (far-infrared) images, aimed to enhance people detection by means of Convolutional Neural Networks (CNNs). While image detection with RGB images is a well-developed area, it still heavily relies on and is greatly limited by lighting conditions. This limitation poses a significant challenge for image detection to play a larger role in everyday technology, where illumination cannot always be controlled. Far-infrared images (which are invariant to light conditions), can serve as a valuable complement to RGB images in environments where illumination cannot be controlled and a robust detection of objects is needed. In this work, various fusion techniques are presented so that the information fused can be used in more advanced detection algorithms. Focussing on the field of search and rescue operations, security of vigilance the work focuses on pedestrian detection, taking advantage of the power of YOLOv8 algorithm. All the fusion techniques are tested making use of a multispectral *dataset*, Kaist, with the aim of addressing these limitations and improving detection performance.

Keywords: thermal images, person detection, multiespectral image fusion, deep learning, computer vision

[*] Autor para correspondencia: e.heredia@umh.es

1. Introducción

El campo de la detección de objetos en base a imágenes ha evolucionado mucho en los últimos años, dando resultados muy avanzados. Esta área del procesamiento de imagen, centrada en combinar la clasificación de objetos junto con la localización de la misma en una imagen, se ha focalizado mayormente en imágenes del espectro visible. No es raro que sea así, pues la mayoría de cámaras y aplicaciones funcionan en dicho espectro. Pero en según qué situaciones limitar el espectro de imagen a un rango tan pequeño, el visible, puede ser contraproducente dejando información relevante fuera de la imagen. Además, muchas de las técnicas dependen directamente de un control preciso de la iluminación, condición que no se cumple en todos los casos de uso. Tareas como son las de búsqueda y rescate (SAR *Search And Rescue*), vigilancia o seguridad limitan la influencia que puede tenerse sobre la iluminación del entorno. Son tareas que a su vez requieren una gran madurez en el sistema empleado y una robustez a cambios ambientales que la imagen visible no es capaz de ofrecer.

En la última década han destacado dos familias de detectores basados en Redes Neuronales Convolucionales (CNNs, *Convolutional Neural Networks*). La primera, basada en detección en dos etapas, está representada por la familia de detectores RCNN Girshick et al. (2013) y sus derivados Fast-RCNN (Girshick, 2015) y Faster-RCNN (Ren et al., 2015). Por otro lado, otros detectores combinan la detección y clasificación en una sola etapa como las redes YOLO (Redmon et al., 2015), así como todas las versiones que de ella han derivado. En los últimos años, las arquitecturas basadas en *transoformers* (Vaswani et al., 2017), inicialmente desarrolladas para el ámbito del procesamiento del lenguaje, se han abierto camino en el mundo del procesamiento de imagen, destacando la red DETR (Carion et al., 2020).

Figura 1: El poder de combinar imágenes visuales y térmicas (Kalita et al., 2020)

Los últimos avances siguen mejorando centrándose sobre imágenes en el espectro visible, algunas cuestiones de interés surgen: ¿Qué pasa cuando las condiciones de iluminación no pueden controlarse? ¿Qué pasa cuando hay oclusiones que impiden detectar la presencia de ciertos objetos? Merecen especial atención las oclusiones que, en algunos casos, sólo se dan en cierto espectro de la luz, tal y como puede verse claramente ejemplificado en la Figura 1. En la misma puede verse claramente como los arbustos impiden identificar a la persona que hay detrás, claramente visible en el espectro térmico. Aplicaciones como las mencionadas, de SAR, seguridad o vigilancia no pueden permitirse el precio de no detectar individuos presentes, así como excusarse en una falta de iluminación.

Un espectro de la luz especialmente interesante, dada la situación planteada, es el espectro térmico. Las cámaras térmicas son capaces de procesar luz infrarroja, que es sabido es emitida por los diferentes cuerpos del entorno de manera proporcional a la temperatura de los mismos. El espectro infrarrojo es un espectro bastante amplio, dependiendo del rango de temperaturas que se quiera medir se centrará la adquisición de imagen en una zona concreta. En este caso se centrará en el infrarrojo lejano o LWIR (*Longwave Infrared*), que abarca desde los 8µm hasta 14µm. En función de la sensibilidad del sensor de la cámara, y variando la misma, este espectro permite aproximar temperaturas del rango de -20ºC hasta los 1000ºC. Una configuración típica permitiría medir desde los -20ºC hasta los 120ºC, que es el rango óptimo para detectar personas, por ejemplo desaparecidas en una montaña, y diferenciarlas del entorno circundante.

Hay algunos trabajos centrados en emplear las imágenes térmicas por sí solas, como en el caso de Ivašić-Kos et al. (2019) o Kalita et al. (2020), evaluados con la red YOLOv3 (parte de los detectores previamente mencionados. Puede verse una descripción completa expuesta por Redmon and Farhadi (2018), muy relevante e ilustrativa para entender la familia YOLO). Aunque las imágenes térmicas pueden aportar gran cantidad de información relevante, siendo esta invariante a condiciones de iluminación (una de las limitaciones previstas en el caso de imágenes visuales), aún presenta sus propias limitaciones. Por ejemplo algunos objetos transparentes en el espectro visible se verán opacos, con un tono proporcional a la temperatura del mismo, en la imagen térmica. También son susceptibles a cambios ambientales: no hace la misma temperatura en exteriores en invierno que en verano, ni las personas a detectar llevan la misma cantidad de ropa. Estos son parte de los retos que se plantean frente a una tarea de detección.

Teniendo cada tipo de imagen sus ventajas y desventajas, la mejor solución posible podría pasar por combinarlas para suplir con una los problemas de la otra. Aunque hay diferentes enfoques al problema de fusión de imágenes, este trabajo se va a centrar en aquellos que aprovechan el desarrollo de algoritmos de detección tradicionales fusionando ambas imágenes en una nueva imagen que se usará como entrada para afinar el algoritmo dado. Estos algoritmos de fusión para comprimir ambas imágenes pueden ir desde los más sencillos como puede ser una media entre la información de ciertos canales (Vandersteegen et al., 2018) a soluciones más complejas como *superpixel segmentation* (Mao et al., 2021) o el uso de *wavelets* (Chipman et al., 1995) o (Su and Jung, 2018). Otros enfoques se plantean sustituir las arquitecturas clásicas de detección para integrar la etapa de fusión en la propia red profunda, integrándola en la etapa de extracción de características (Xiang et al., 2022) o en la etapa de clasificación al final de la red.

Es bastante tentador, cuando se buscan nuevas soluciones a un problema tal y como el que se ha planteado, optar por soluciones especialmente complejas, sumergiéndose en las últimas técnicas del estado del arte e involucrando arquitecturas cada vez más complejas. Pero antes de dar ese paso creemos que es importante establecer una línea base, sobre la que evaluar la mejoría y su coste en términos de complejidad, tiempo de cómputo o gasto de recursos. En consecuencia, este trabajo plantea: ¿hasta donde pueden llevarnos los métodos más simples de fusión de imagen?

Establecida la pregunta de investigación, se continúa con Sección 2 explicando diferentes algoritmos de fusión de imagen. La Sección 3 cubre los aspectos relacionados con la metodología del estudio, así como la descripción del algoritmo y el *dataset* empleados. También se resaltan algunas limitaciones que serán relevantes en la Sección 4, donde se comentan los resultados obtenidos para cada una de las condiciones y algoritmos de fusión. Por último, la Sección 5 incluye algunas conclusiones que se han considerado relevantes, así como unas pinceladas de lo que serán futuros desarrollos.

2. Fusión de imagen: térmica y visible

El objetivo de la fusión de imágenes que se plantea en este trabajo es el de comprimir la información de dos imágenes en una sola. Las imágenes visibles incluyen tres canales, RGB, mientras que las imágenes térmicas incluyen un solo canal de intensidad; de esta manera, la imagen resultante debe incluir información de los cuatro canales compactada en solo tres.

La mayoría de desarrollos en torno a la detección de imagen, tal y como se ha introducido previamente, están optimizados para emplear imágenes del espectro visible, típicamente con tres canales de entrada. Es por eso que se ha optado por este enfoque en la fusión puesto que la idea es aprovechar la potencia de dichos enfoques adaptándolos ligeramente al caso que nos ocupa.

Los métodos que se plantean en este artículo se les ha denominado métodos estáticos de fusión, pues son métodos predefinidos que se aplican por igual a todas las imágenes sea cual sea el *dataset*. Nótese que esta serie de métodos requieren que la escena en ambas imágenes, visible y térmica, sea exactamente la misma.

2.1. Fusión RGBT

Tal y como proponen en Vandersteegen et al. (2018), y para evitar la menor pérdida de información posible, ambas imágenes se comprimen haciendo la media aritmética entre sus canales. Los tres canales de la imagen del espectro visual, por separado, con el canal de intensidad de la imagen térmica. Puede observarse la expresión matemática en (1).

$$\begin{aligned} ch_1 &= (R + T)/2 \\ ch_2 &= (G + T)/2 \\ ch_3 &= (B + T)/2 \end{aligned} \tag{1}$$

La imagen resultante se conforma combinando de vuelta los tres canales ch_1, ch_2 y ch_3 como si se tratasen de los canales RGB originales.

2.2. Fusión HSVT

Para comprimir ambas imágenes, en este caso tomando la visual en el espacio de color HSV, se da mayor importancia a la información de color. De esta forma se mantienen los canales de tono (canal H) y saturación (canal S) de la imagen, combinando la intensidad (canal V) con la imagen térmica. Ambos canales se suman y se reescalan para entrar de nuevo en una representación de 8 bits, típica en imágenes. De esta manera se comprime la información de intensidades haciendo uso de todo el rango de valores del último canal de la imagen. Las ecuaciones a continuación describen la transformación realizada:

$$\begin{aligned} ch_1 &= H \\ ch_2 &= S \\ ch_3 &= 255 * (V + T)/max(V + T) \end{aligned} \tag{2}$$

Nuevamente, la imagen fusionada es el resultado de combinar de vuelta los canales ch_1, ch_2 y ch_3 en una sola imagen como si de los canales HSV originales se tratasen.

2.3. Fusión VTHS

En función de la iluminación puede darse el caso en que el color no sea especialmente relevante, y la mayoría de la información útil esté contenida en el canal de intensidad. En este caso, la fusión descrita previamente estaría ocupando mayor espacio para la información más escasa, comprimiendo en un solo canal la más representativa. Es importante remarcar que las diferencias pueden deberse tanto al *dataset* como a la importancia del color en los objetos que se pretende detectar.

Asumiendo una mayor importancia de la intensidad frente al color, se reservan dos canales de la imagen resultante exclusivamente para el canal de intensidad de la imagen visible y el canal de intensidad de la imagen térmica. El tercer canal se recompone con la información más significativa de los canales de tono y saturación. Nótese que contienen información diferente y no sería conveniente realizar una media. La compresión se hace concatenando los 4 bits más significativos de cada uno de los canales en un solo canal, tal y como describe (3).

$$\begin{aligned} ch_1 &= V \\ ch_2 &= T \\ ch_3 &= (H \gg 4)\&((S \gg 4) \ll 4) \end{aligned} \tag{3}$$

La imagen resultante surge de recomponer los tres canales de vuelta.

2.4. Fusión VT

Llevando al extremo el caso anterior, y asumiendo que la información de color no esté aportando beneficio a la detección, podrían obtenerse buenos resultados dando aun mayor importancia a los canales de intensidad. Nótese, nuevamente, que las diferencias entre métodos tendrán una gran dependencia sobre los datos de partida.

La fusión en este caso mantendrá los canales V y T, como en el caso anterior, y rellenará el tercer canal con la media de ambos. La entrada que se pretende tener es de tres canales y es posible que cierta información recurrente pueda aprovecharse dentro de la red profunda. La ecuación (4) describe la fusión planteada.

$$\begin{aligned} ch_1 &= V \\ ch_2 &= T \\ ch_3 &= (V + T)/2 \end{aligned} \tag{4}$$

La combinación de los tres canales daría como resultado la imagen fusionada.

3. Metodología

3.1. Algoritmo de detección: YOLOv8

Presentado en la Sección 1, YOLO es uno de los algoritmos estrella en cuanto a tareas de detección. El trabajo comenzó con las últimas actualizaciones de YOLOv5, desarrollado en Python (Pytorch) por la empresa Ultralitics (Jocher, 2020), finalmente se actualizó a la versión v8, con la que se han realizado las pruebas, también haciendo uso del paquete Pytorch (Jocher et al., 2023). YOLOv8 incluye, además de la detección, capacidades para funcionar como clasificador, segmentar diferentes clases de objetos en imágenes y otra serie de funcionalidades. En estas pruebas solo se ha evaluado funcionando como detector.

En este caso, YOLOv8 toma como entrada una imagen de 3 canales, y habiendo sido entrenado para ello, es capaz de detectar un conjunto de objetos así como su posición sobre la imagen. Bajo este contexto se ha reentrenado la red para detectar una sola clase, la clase 'peatón'. Nótese que en el texto se usará indistintamente la palabra persona o peatón.

En las tareas de detección, para medir el rendimiento del algoritmo, se hace uso de dos métricas combinadas: la probabilidad sobre la posición de la *bounding box*, y la probabilidad de una clase. Para cuantificar la exactitud de la posición estimada por la red se usa la primera, *box confidence*, que combina dos valores: *objectness score*, ¿contiene la caja un objeto? así como la intersección sobre la unión (IoU, *Intersection over Union*. Para cuantificar la exactitud de la clasificación realizada se toma una probabilidad condicionada de que sea una clase habiéndose detectado un objeto sobre esta. Se ha trabajado con los valores por defecto de YOLO, que pueden resultar algo restrictivos.

3.2. dataset: Kaist

Para evaluar la fusión de imágenes se ha recurrido al *dataset* Kaist (Hwang et al., 2015). Este *dataset* contiene 95k pares de imágenes térmicas-visibles (640x480, 20Hz). Ha sido etiquetado manualmente incluyendo las clases 'peatón', 'ciclista' y 'grupo de gente'. Todas las imágenes están calibradas de manera que la escena vista en cada par de imágenes es completamente coincidente, requisito necesario para la fusión planteada.

El *dataset* no incluye suficientes instancias de las clases 'ciclista' ni 'grupo de personas' para el tipo de entrenamiento que se va a hacer, así que se han eliminado. Se usarán las clases 'peatón' y 'ciclista', interpretadas en ambos casos como 'persona'.

El propio *dataset* incluye una propuesta de división de las imágenes en subsets tanto para entrenamiento como para validación/test. Estos subsets de imágenes han sido equilibrados en cuanto al número de imágenes e instancias de objetos que se encuentran para cada caso. La Tabla 1 muestra un resumen de cómo se han utilizado los sets propuestos por Kaist para el entrenamiento posterior, revisando el número de imágenes en cada caso, el número de fondos y el número de instancias de persona. Los subsets de *train* se han empleado para entrenar la red y los de *test* para validación y test.

Tabla 1: Resumen de las características de los sets empleados en el entrenamiento y validación del modelo

Nombre del set	Imágenes	Fondos	Instancias
test-day-01	29178	15191	34492
train-day-02	16694	10803	12521
test-night-01	15962	10253	11999
train-night-02	8392	4817	8671

Se observa que el número de imágenes entre condiciones, día y noche no está equilibrado, habiendo muchas más imágenes del caso diurno. Para evitar posible sesgo debido a este desbalanceo, no habiendo demasiadas imágenes del caso nocturno, se ha optado por entrenar dos modelos diferenciando, por separado, el modelo diurno del modelo nocturno.

El *dataset* de Kaist, tal y como detallan los autores del mismo, está enfocado a la conducción autónoma, concretamente en el ámbito de la detección de personas. El *dataset* por tanto incluye una serie de imágenes consecutivas en entornos urbanos tomándolas desde un coche (perspectiva diferente a la que podría tener un UAV, un dron o un robot terrestre). Estas características suman una serie de limitaciones a la hora de traspasar la información al caso de uso que nos ocupa.

3.3. Entrenamiento del modelo

Aunque YOLO incluye modelos preentrenados con el *dataset* COCO (Lin et al., 2014), incluyendo información que puede ser muy relevante a la hora de entrenar una red, presenta una gran limitación: requiere utilizar exactamente la misma arquitectura sobre la que cargar los pesos preentrenados. Como a futuro se plantean modificaciones de arquitectura, se ha optado por entrenar de cero la red, fijando Kaist como la única fuente de datos, pudiendo así comparar el rendimiento sobre una base común. De esta manera, diferentes arquitecturas y algoritmos de fusión serán evaluados con una base común, sin incluir la mejora del *dataset* externo.

Evidentemente este marco nos permite hacer comparaciones justas entre los algoritmos empleados, pero nos impide comparar con modelos o algoritmos externos que hagan uso de otras fuentes de información o condiciones de ejecución. Para poder hacer una comparación más justa sería necesario adaptar otros *dataset*s con información térmica y visual que permitan aumentar el tamaño de datos en la etapa de entrenamiento.

Pese a la limitada cantidad de imágenes con las que contamos, podemos obtener conclusiones prácticas sobre la dirección de avance de la investigación, dejando el trabajo más costoso (reentrenar con más datos) para aquellos casos más prometedores. Es importante recordar que las conclusiones extraídas con un *dataset* a escala no tienen por qué ser directamente trasladables a un *dataset* más grande; habría que hacer pruebas con varios modelos para poder confirmarlo con seguridad.

3.4. Evaluación y métricas

Algunas métricas típicas han sido empleadas para evaluar el rendimiento de la red. Se definen las mismas a continuación para poder presentar el posterior análisis de resultados en la Sección 4:

- *Precision* (P): permite evaluar la capacidad del modelo de evitar equivocarse, de detectar falsos positivos.

- *recall* (R): permite medir la capacidad del sistema para detectar todas las instancias de una clase dado un nivel de confianza para la detección.
- *Intersection over Union* (IoU): permite cuantificar hasta qué punto el algoritmo es capaz de encontrar adecuadamente (posición y tamaño) las instancias de cada clase. Para ello compara la intersección de ambas *bounding boxes*, la predicción y la etiqueta, sobre la unión de las mismas.
- *Average Precision* (AP): se calcula integrando el área bajo la curva de Precision-*recall* para una clase dada. Permite resumir en un solo valor el rendimiento del modelo.
- *Mean Average Precision* (mAP): cuantifica el rendimiento general del modelo haciendo la media de los AP para diferentes clases. En el caso de uso que se plantea se evaluará solo una clase, 'persona', aun así se empleará esta métrica por ser la comúnmente utilizada. Esta métrica se calcula para diferentes IoU: mAP50 para un límite de IoU de 0.5 (detecciones más sencillas); mAP50-95 acumula la media de AP calculada para diferentes mínimos de IoU que van desde 0.5 hasta 0.95, siendo una métrica mucho más restrictiva.

4. Resultados

La Tabla 2 contiene un resumen de los resultados comentados a continuación. Además de los resultados obtenidos en base a los cuatro métodos de fusión presentados, se han añadido tanto los resultados con imágenes del espectro visible y térmico, por separado, a modo de referencia para la condición diurna y nocturna respectivamente.

Tal y como se plantea el escenario de trabajo, parece lógico dar mayor prioridad a un mejor *recall* (detectar todas las instancias) sin descuidar la *precision* (evitar falsos positivos). Puede verse una comparación de la curva de *precision-recall* para las condiciones diurnas y nocturnas en la Figura 2(a) y la Figura 2(b) respectivamente.

En el espectro visual, y con las condiciones del *dataset* presentado, no hay método de fusión que supere a las imágenes del espectro visible. Tiene un mejor mAP para el límite de 0.5 de IoU, mejor *recall* y el menor tiempo de duración del entrenamiento, algo que también se debe considerar. Cuando la restricción de IoU es mayor, el valor de mAP50-95 puede ser superado por otros métodos (VT y VTHS). VT y VTHS proporcionan mejor resultados en la búsqueda de personas, del *bounding box* al priorizar la imagen térmica, que en este *dataset* resalta especialmente a los peatones. Aunque el método RGBT ofrece una mayor *precision* se sigue prefiriendo un mayor *recall*.

Respecto a la condición nocturna, de entre la imagen visual y LWIR, es la LWIR la que contiene mayor información y la que se empleará como referencia. En este caso, los mejores resultados, nuevamente en cuanto a mAP y *recall*, se obtienen con el método de fusión VTHS. Aunque inspeccionando las imágenes puede verse cómo el color no es un componente muy relevante en el *dataset* (además de dar resultados muy pobres solo con la imagen visible), la información que proporciona sigue siendo mejor que la eliminación completa del mismo como en la fusión VT. Nuevamente, en cuanto a precisión, VT es ligeramente mejor al reducir los falsos positivos incidiendo más en la imagen térmica, que permite distinguir mejor a los peatones (se ven resaltados con respecto al entorno en las imágenes de este *dataset*). Aun cuando se restringe el IoU sigue siendo superior el desempeño del método VTHS. En ambos casos el tiempo de entrenamiento es razonable, siendo muy superior para métodos con peor desempeño.

El equilibrio entre la *precision* y *recall* puede evaluarse en las curvas de la Figuras 2(a) y 2(b). En el caso diurno puede comprobarse cómo ambos, RGBT y visible, dan buenos resultados. A su vez se valida que el método VTHS llega a proporcionar una mejor *precision* a costa de un muy bajo *recall*. En el caso nocturno se observa mayor diferencia; el método VTHS estaría proporcionando resultados significativamente mejores que el VT en cuanto *recall* a igual *precision*.

5. Conclusiones

Es generalmente aceptado que tener más cantidad de datos puede generar mejores resultados incluso que algoritmos más refinados. Es por ello que trabajos futuros deberían enfocarse en mejorar los datos para el caso de uso que se plantea. El *dataset* debería complementarse con más datos (día, noche y diferentes peatones, no solo imágenes contiguas), y también con información más variada (oclusiones que apliquen a una de las imágenes pero no a la otra, como la mostrada en la Figura 1, más cambios en iluminación y condiciones climatológicas). Además de ampliar la información se pretende ajustar la división de datos entre entrenamiento y validación para ajustarla a las necesidades del caso planteado, en vez de las propuestas por los autores de Kaist (Hwang et al., 2015).

Estos cambios en los datos deberían tener un claro impacto en los resultados obtenidos. El rendimiento solo con imágenes de espectro visual empeorará claramente una vez se incluyan este tipo de datos dando mejores resultados y mayor importancia a la fusión en ambas condiciones, diurna y nocturna.

Aunque la decisión de eliminar COCO de la ecuación nos permite comparar de manera justa diferentes enfoques de la fusión, ninguna conclusión debería tomarse como definitiva sin añadir información extra. No debe obviarse la posibilidad de incluir otro tipo de *dataset*s que puedan complementar la información sobre la que se va a trabajar.

Por último, para poder comparar los métodos en otros rankings sería preciso evaluar también métricas específicas como las propuestas por Hwang et al. (2015) como el *miss rate* o el *Log Average Miss Rate*.

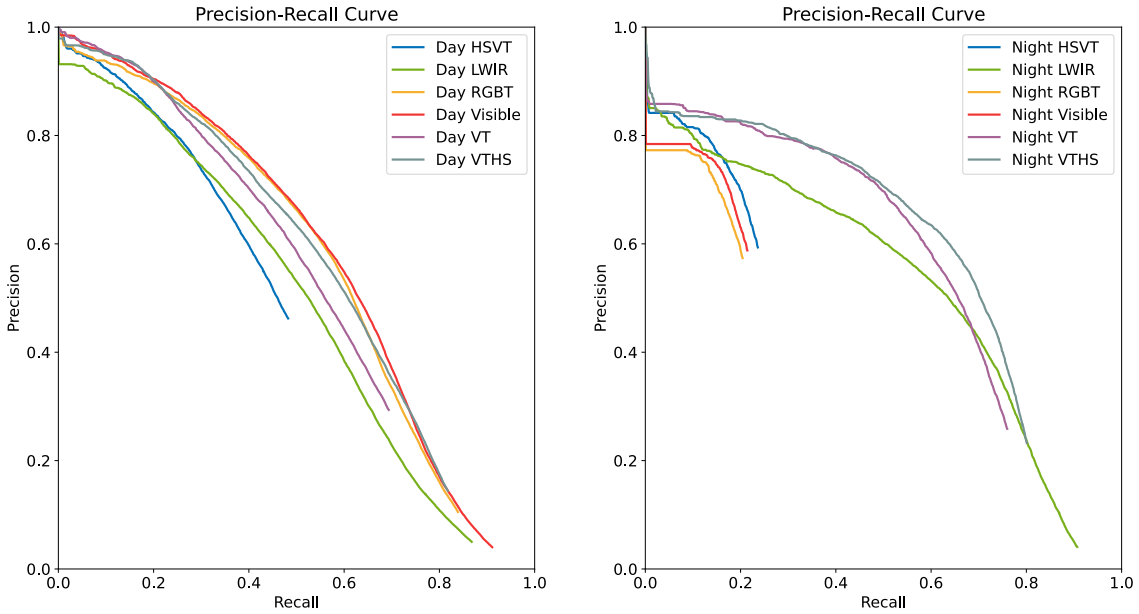
Mejorando la información es un paso clave para obtener unas conclusiones más robustas en cuanto a la fusión de imágenes que nos ocupa.

Agradecimientos

Esta publicación forma parte del proyecto PROMETEO/2021/075 financiado por la Generalitat Valenciana y al proyecto TED2021-130901B-I00, financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea "NextGenerationEU"/PRTR.

Tabla 2: Resultados de entrenamiento y validación de diferentes test separados por condición diurna/nocturna para cada algoritmo de fusión de imagen

	P	R	mAP(50)	mAP(50-95)	Best epoch
HSVT day	0.4612	0.4832	0.4935	0.2186	59
VT day	0.5969	0.4915	0.5466	0.2378	20
VTHS day	0.5922	0.5372	0.5657	0.2345	9
RGBT day	0.6106	0.5466	0.5664	0.2284	6
Visible day	**0.6041**	**0.5560**	**0.5762**	**0.2292**	**5**
RGBT night	0.5712	0.2048	0.3783	0.1533	69
HSVT night	0.5913	0.2364	0.4104	0.1568	69
LWIR night	**0.5383**	**0.5916**	**0.5201**	**0.1718**	**3**
VT night	0.6268	0.5652	0.5671	0.1892	20
VTHS night	0.6133	0.6251	0.5894	0.2012	20



(a) Precision-*recall* condición diurna

(b) Precision-*recall* condición nocturna

Figura 2: Resultados conjuntos de los test de entrenamiento

Referencias

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. CoRR abs/2005.12872.

Chipman, L., Orr, T., Graham, L., 1995. Wavelets and image fusion. In: Proceedings., International Conference on Image Processing. Vol. 3. pp. 248–251 vol.3.

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

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

Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I. S., 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1037–1045.

Ivašić-Kos, M., Krišto, M., Pobar, M., 2019. Human detection in thermal imaging using yolo. In: Proceedings of the 2019 5th International Conference on Computer and Technology Applications. ICCTA '19. Association for Computing Machinery, New York, NY, USA, p. 20–24.

Jocher, G., May 2020. YOLOv5 by Ultralytics.
URL: `https://github.com/ultralytics/yolov5`

Jocher, G., Chaurasia, A., Qiu, J., Jan. 2023. YOLOv8 by Ultralytics.
URL: `https://github.com/ultralytics/ultralytics`

Kalita, R., Talukdar, A. K., Kumar Sarma, K., 2020. Real-time human detection with thermal camera feed using yolov3. In: 2020 IEEE 17th India Council International Conference (INDICON). pp. 1–5.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: common objects in context. CoRR abs/1405.0312.

Mao, S., Duan, J., Zhang, Z., Zhang, Z., 2021. Visible and infrared image fusion via superpixel segmentation and salient region detection. In: 2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC). pp. 643–648.

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

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. CoRR abs/1804.02767.

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

Su, H., Jung, C., 2018. Multi-spectral fusion and denoising of rgb and nir images using multi-scale wavelet analysis. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 1779–1784.

Vandersteegen, M., Van Beeck, K., Goedemé, T., 2018. Real-time multispectral pedestrian detection with a single-pass deep neural network. In: Campilho, A., Karray, F., ter Haar Romeny, B. (Eds.), Image Analysis and Recognition. Springer International Publishing, Cham, pp. 419–426.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. CoRR abs/1706.03762.

Xiang, J., Gou, S., Li, R., Zheng, Z., 2022. Rgb-thermal based pedestrian detection with single-modal augmentation and roi pooling multiscale fusion. In: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 3532–3535.