SEGMENTACIÓN AUTOMÁTICA DE MOVIMIENTO EN PROYECCIONES COMPRIMIDAS DE SECUENCIAS DE VÍDEOS MULTIESPECTRALES MEDIANTE APRENDIZAJE PROFUNDO.

LISETH VERÓNICA LUCENA LUNA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2021

SEGMENTACIÓN AUTOMÁTICA DE MOVIMIENTO EN PROYECCIONES COMPRIMIDAS DE SECUENCIAS DE VÍDEOS MULTIESPECTRALES MEDIANTE APRENDIZAJE PROFUNDO

LISETH VERÓNICA LUCENA LUNA

Tesis presentada en cumplimiento de los requisitos para el grado de Ingeniera de Sistemas.

Directora:
Claudia Victoria Correa Pugliese
Doctora en Ingeniería Eléctrica y Computación.

Codirector:
Henry Arguello Fuentes
Doctor en Ingeniería Eléctrica y Computación

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2021

AGRADECIMIENTOS

El autor expresa sus agradecimientos:

A Dios, por ser mi guía, por darme fuerza y sabiduría para cumplir mis metas.

A mi madre, Eugenia Lucena, por su amor, esfuerzo y valentía. Gracias por su apoyo incondicional, por impulsarme a cumplir mis sueños, por ser la principal patrocinadora de este logro y un pilar fundamental en mi vida.

A mi hermana, Alejandra Lucena, por brindándome su apoyo en todo momento, por acompañarme en las etapas más importantes, siendo mi mayor motivación.

A mis amigos y compañeros de estudio, en especial a Karina, Andrea, Natalia, Jessica y Esteban por acompañarme en este proceso.

A mi directora de tesis, Claudia Correa, por su dedicación, paciencia y compromiso. Gracias por todas sus enseñanzas.

Al grupo de investigación HDSP, por brindarme su ayuda en el desarrollo de este proyecto.

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

[1] PINILLA, Samuel, *et al.* "Salient Motion Detection for Spectral Video on the Compressive Domain". En: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2019, págs. 106-110.

LISTA DE TABLAS

pág.

---

[2] PINILLA, Samuel, *et al.* "Salient Motion Detection for Spectral Video on the Compressive Domain". En: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2019, págs. 106-110.

[3] PINILLA, Samuel, *et al.* "Salient Motion Detection for Spectral Video on the Compressive Domain". En: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2019, págs. 106-110.

[4] BOYD, Stephen , *et al.* "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". En: *Foundations and Trends in Machine Learning* 3 (2011), págs. 1-122.

LISTA DE ANEXOS

pág.

RESUMEN

TÍTULO: SEGMENTACIÓN AUTOMÁTICA DE MOVIMIENTO EN PROYECCIONES COMPRIMIDAS DE SECUENCIAS DE VÍDEOS MULTIESPECTRALES MEDIANTE APRENDIZAJE PROFUNDO. [*]

AUTOR: LISETH VERÓNICA LUCENA LUNA [**]

PALABRAS CLAVE: MUESTREO COMPRESIVO, VÍDEO ESPECTRAL, REDES NEURONALES CONVOLUCIONALES, SEGMENTACIÓN DEL MOVIMIENTO

DESCRIPCIÓN: El vídeo espectral ha surgido como una herramienta científica no invasiva para analizar el comportamiento de escenas dinámicas en alta resolución espectral, siendo de gran interés en diversas áreas del conocimiento enfocadas a la detección de anomalías, clasificación de materiales, y seguimiento de objetos. Dada su importancia, diversas técnicas se han desarrollado para obtener la información espectral de la escena en formato comprimido, a fin de solventar los problemas de adquisición, almacenamiento y procesamiento. Estas técnicas consisten en sistemas ópticos que captan y codifican la información tridimensional de la escena (espacio-espectral) en un conjunto de proyecciones bidimensionales, a altas velocidades de fotograma. A partir de estas medidas comprimidas, se puede recuperar el vídeo espectral subyacente mediante algoritmos de reconstrucción computacional, para realizar las diversas tareas de procesamiento como la detección, identificación y seguimiento de objetos en movimiento. Sin embargo, el principal desafío de los enfoques basados en la reconstrucción es el alto costo computacional, que incrementa significativamente con el número de fotogramas. En este trabajo se presenta una estrategia para segmentar objetos en movimiento en el dominio de compresión. El método propuesto incorpora redes neuronales convolucionales para modelar características espacio-temporales que permiten detectar las regiones de movimiento prominente sin necesidad de reconstruir el conjunto de datos. Los experimentos realizados muestran la calidad de la segmentación del enfoque propuesto respecto a un método del estado del arte, que también realiza esta tarea de inferencia en el dominio de compresión, reportando en promedio una mejora del 29 % en C-CASSI y 24 % en CASSI en términos de *F-measure*.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Claudia V. Correa, Ph.D. Codirector: Henry Arguello, Ph.D.

ABSTRACT

TITLE: AUTOMATIC MOTION SEGMENTATION IN COMPRESSED PROJECTIONS OF MULTISPECTRAL VIDEO SEQUENCES USING DEEP LEARNING. *

AUTHOR: LISETH VERÓNICA LUCENA LUNA. **

KEYWORDS: COMPRESSIVE SAMPLING, SPECTRAL VIDEO, CONVOLUTIONAL NEURAL NETWORKS, MOTION SEGMENTATION.

DESCRIPTION: Spectral video has emerged as a non-invasive scientific tool to analyze the behavior of dynamic scenes in high spectral resolution, being of great interest in several areas of knowledge focused on anomaly detection, material classification, and object tracking. In view of its importance, numerous techniques have been developed to obtain the spectral information of the scene in compressed format, in order to solve the issues related to acquisition, storage and processing. These techniques consist of optical systems that capture and encode the three-dimensional (spatio-spectral) scene information into a set of two-dimensional projections at high frame rates. From these compressed measurements, the underlying spectral video can be recovered using computational reconstruction algorithms to later perform various processing tasks such as detection, identification and tracking of multiple moving objects. However, the main challenge of reconstruction-based approaches is the high computational cost, which increases significantly with the number of frames. Therefore, this paper presents a strategy for segmenting moving objects in the compressive domain. The proposed method incorporates convolutional neural networks (CNN) to model spatio-temporal features to detect regions of prominent motion without involving data reconstruction. The experiments show that the proposed method outperforms a state-of-the-art approach that also detects motion on the compressive domain, showing an average improvement of 29 % in C-CASSI and 24 % in CASSI in terms of F-measure.

---

* Undergraduate project

** Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Claudia V. Correa, Ph.D. Co-advisor: Henry Arguello, Ph.D.

INTRODUCCIÓN

Una imagen espectral (IE) es un conjunto de datos que contiene información espacial de una escena en un amplio rango del espectro electromagnético. Específicamente, este conjunto se compone de varias imágenes monocromáticas adquiridas en diferentes longitudes de onda. La información que proporciona este tipo de imagen es valiosa para diversas aplicaciones enfocadas al diagnóstico médico [1], monitoreo del medio ambiente [2], inspección de productos agrícolas [3], entre otras. Una secuencia de estas imágenes captadas en diferentes instantes de tiempo conforma un vídeo espectral. Este tipo de dato es gran interés en la comunidad científica, dado que permite analizar en cortos periodos de tiempo la variación espectral de la escena bajo estudio [4]. En los últimos años, la teoría de muestreo compresivo se ha implementado en los sistemas de captura de imagen espectral, con el objetivo de reducir los costos de adquisición, almacenamiento y procesamiento [5]. Estos sistemas ópticos se han adaptado para adquirir y simultáneamente comprimir la información espacio-espectral de escenas dinámicas en un conjunto de proyecciones

---

[1] LU, Guolan y FEI, Baowei. "Medical hyperspectral imaging: a review". En: *Journal of Biomedical Optics* 19.1 (2014), págs. 1-24.

[2] SHAW, Gary y BURKE, Hsiao-hua. "Spectral Imaging for Remote Sensing". En: *Lincoln Laboratory Journal* 14 (2003), págs. 3-28.

[3] WANG, Yujie, *et al.* "Discrimination of nitrogen fertilizer levels of tea plant (Camellia sinensis) based on hyperspectral imaging". En: *Journal of the Science of Food and Agriculture* 98 (2018), págs. 4659-4664.

[4] LÓPEZ, Kareth León; GALVIS, Laura y FUENTES, Henry Arguello. "Temporal Colored Coded Aperture Design in Compressive Spectral Video Sensing". En: *IEEE Transactions on Image Processing* 28.1 (2019), págs. 253-264.

[5] CAO, Xun, *et al.* "Computational Snapshot Multispectral Cameras: Toward dynamic capture of the spectral world". En: *IEEE Signal Processing Magazine* 33.5 (2016), págs. 95-108.

bidimensionales (2D) [6]. En estos sistemas se asume que la señal de interés tiene una representación escasa en un determinado dominio de transformación, es decir, posee pocas entradas no nulas que contribuyen a la misma, de manera que puede ser aproximada por una pequeña cantidad de proyecciones aleatorias[7]. A partir de este conjunto de proyecciones 2D, se puede recuperar el vídeo espectral mediante un algoritmo computacional, para realizar las diversas tareas de procesamiento. Por ejemplo, en los sistemas de vigilancia, los vídeos espectrales se han utilizado para la detección, identificación y seguimiento de objetos [8]. Sin embargo, el proceso de reconstrucción conlleva un alto costo computacional, que incrementa con el número de fotogramas, así como con la resolución espacial y espectral. En vista de ello, varios trabajos centrados en el análisis de IE [9] [10] [11] han explorado la alternativa de realizar tareas de inferencia, como la clasificación de imágenes y detección de objetos, directamente en las proyecciones comprimidas. Estos enfoques

---

[6] LÓPEZ, Kareth León; GALVIS, Laura y FUENTES, Henry Arguello. "Spatio-spectro-temporal coded aperture design for multiresolution compressive spectral video sensing". En: *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017, págs. 728-732.

[7] PUGLIESE, Claudia; CARREÑO, Diana y ARGUELLO, Henry. "Sparse representations of dynamic scenes for compressive spectral video sensing". En: *DYNA* 83 (2016), págs. 42-51.

[8] LIU, Rongrong; RUICHEK, Yassine y EL BAGDOURI, Mohammed. "Multispectral Dynamic Codebook and Fusion Strategy for Moving Objects Detection". En: *Image and Signal Processing*. 2020, págs. 35-43.

[9] HINOJOSA, Carlos; RAMIREZ, Juan Marcos y ARGUELLO, Henry. "Spectral-Spatial Classification from Multi-Sensor Compressive Measurements Using Superpixels". En: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, págs. 3143-3147.

[10] RAMIREZ, Juan Marcos y ARGUELLO, Henry. "Spectral Image Classification From Multi-Sensor Compressive Measurements". En: *IEEE Transactions on Geoscience and Remote Sensing* 58.1 (2020), págs. 626-636.

[11] VARGAS, Hector; FONSECA, Yesid y ARGUELLO, Henry. "Object Detection on Compressive Measurements using Correlation Filters and Sparse Representation". En: *2018 26th European Signal Processing Conference (EUSIPCO)*. 2018, págs. 1960-1964.

han demostrado que es posible realizar con precisión dichas tareas en un espacio de menor dimensión.

Por otra parte, en el campo de visión por computadora, la detección de movimiento desempeña un papel importante, ya que es la base de muchas aplicaciones enfocadas al reconocimiento de acciones, estimación de pose, control del tráfico, entre otras. Múltiples trabajos han sido desarrollados alrededor de esta temática, en donde se involucran diferentes técnicas para detectar cambios sobresalientes en secuencias de vídeo en escala de grises y RGB. En general, estos enfoques incluyen el flujo óptico, mezcla de distribuciones gaussianas y análisis de bajo rango para modelar el fondo de escenarios dinámicos y obtener los objetos de interés. Recientemente, los modelos de aprendizaje profundo han logrado un alto rendimiento en diversas aplicaciones destinadas al análisis de imagen y vídeo. Varios enfoques basados en redes neuronales convolucionales (RNC) han demostrado que este tipo de red neuronal puede modelar adecuadamente el movimiento en secuencias de vídeo [12] [13]. En este sentido, se han logrado avances significativos en tareas de etiquetado a nivel de píxel, tal como la segmentación de objetos en movimiento. No obstante, esta tarea de inferencia en secuencias de vídeo espectral, dentro del marco de muestreo compresivo ha sido poco explorada en la literatura. Por ejemplo, los autores del método [14] realizan la detección de movimiento directamente en el dominio de compresión, mediante técnicas tradicionales de procesamiento de señales como

---

[12] YASIN, Hashim y HAYAT, Saqib. "DeepSegment: Segmentation of motion capture data using deep convolutional neural network". En: *Image and Vision Computing* 109 (2021), pág. 104147.

[13] JAIN, Suyog; XIONG, Bo y GRAUMAN, Kristen. "FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos". En: (ene. de 2017).

[14] PINILLA, Samuel, *et al.* "Salient Motion Detection for Spectral Video on the Compressive Domain". En: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2019, págs. 106-110.

diferencia temporal y filtrado espacial (operaciones morfológicas).

En este trabajo de investigación se introduce un enfoque basado en RNC para segmentar automáticamente objetos en movimiento en el dominio de compresión de secuencias de vídeo espectral, mediante el modelamiento de características espacio-temporales. En general, el método propuesto toma como entrada un conjunto de proyecciones comprimidas y proporciona un conjunto de máscaras binarias que indican, a nivel de píxel, si un objeto se ha movido en la escena. El proceso de detección de movimiento directamente sobre las proyecciones comprimidas evita la ejecución de algoritmos de reconstrucción, reduciendo considerablemente el tiempo de cómputo. Los resultados obtenidos demuestran que el enfoque propuesto obtiene mayor precisión en la detección de movimiento respecto al método tradicional mencionado previamente [14], mostrando en promedio una mejora del 29 % y 24 % en C-CASSI y CASSI, respectivamente en términos de *F-measure*. Además, el método propuesto obtiene resultados similares a los de un enfoque que trabaja sobre los datos reconstruidos [8].

1. OBJETIVOS

1.1. OBJETIVO GENERAL

- Desarrollar un algoritmo computacional basado en aprendizaje profundo para segmentar objetos en movimiento en proyecciones comprimidas de secuencias de vídeos multiespectrales.

1.2. OBJETIVOS ESPECÍFICOS

- Establecer las técnicas de aprendizaje profundo que permitan determinar características de apariencia y patrones de movimiento en secuencias de imágenes.
- Conformar el conjunto de datos de secuencias de vídeo espectral, necesario para llevar a cabo el proyecto.
- Diseñar un algoritmo computacional basado en las técnicas de aprendizaje profundo seleccionadas, para realizar clasificación a nivel de píxel en el dominio de compresión de secuencias de vídeos multiespectrales.
- Implementar el método computacional para proporcionar la segmentación de movimiento a nivel de píxel en el dominio de compresión.
- Evaluar y comparar el desempeño del algoritmo propuesto en la investigación respecto a técnicas del estado del arte.

2. MARCO DE REFERENCIA

2.1. MUESTREO COMPRESIVO

El muestreo compresivo (CS, por sus siglas en inglés) es un protocolo de adquisición que permite muestrear y simultáneamente comprimir señales, sin perder información esencial. Esta teoría establece que una señal puede ser recuperada con alta probabilidad a partir de un pequeño conjunto de mediciones lineales no adaptivas, reduciendo sustancialmente la cantidad de muestras exigidas por el teorema tradicional de Shannon-Nyquist [15]. Esto es posible gracias a dos principios fundamentales: la escasez, que caracteriza la señal de interés y la incoherencia, que establece la estructura o modalidad de muestreo [16]. En los últimos años, CS ha sido adoptado por diversos sistemas para adquirir eficientemente imágenes espectrales, aumentando la velocidad de sensado y disminuyendo notoriamente los costos de adquisición, almacenamiento y procesamiento.

2.1.1. ESCASEZ Una señal se considera escasa si la mayor parte de su información está concentrada en un pequeño conjunto de sus elementos. Pocas señales tienen esta propiedad por naturaleza, en tal caso, es posible encontrar una representación concisa (escasa) en un determinado dominio de transformación, donde el número de elementos no nulos es menor que la longitud total de la señal [16]. Usualmente se emplean bases de representación ortonormales (como Wavelet, Coseno, Fourier, entre otras) para encontrar los coeficientes significativos de la señal

---

[15] STANKOVIC, Ljubisa, *et al.* "A Tutorial on Sparse Signal Reconstruction and Its Applications in Signal Processing". En: *Circuits, Systems, and Signal Processing* 38 (mar. de 2019).

[16] CANDES, E. J. y WAKIN, M. "An Introduction To Compressive Sampling". En: *IEEE Signal Processing Magazine* 25 (2008), págs. 21-30.

de interés, y así poder descartar sin mucha pérdida perceptual los coeficientes despreciables [15], con el fin de obtener una representación comprimida de dicha señal.

Considere $\mathcal{F} \in \mathbb{R}^{N\times N\times L\times D}$ un vídeo espectral, con $N \times N$ píxeles espaciales, L bandas espectrales y D fotogramas. Su forma vectorial $\boldsymbol{f} \in \mathbb{R}^{N^2LD}$ puede ser representada en la base $\boldsymbol{\Psi}$ como:

$$\boldsymbol{f} = \boldsymbol{\Psi}\boldsymbol{\theta}, \tag{1}$$

donde $\boldsymbol{\theta} \in \mathbb{R}^{N^2LD}$ es un vector de coeficientes de transformación que tiene a lo sumo k valores no nulos. Con base en esto, se dice que la señal $\boldsymbol{f}$ es $k - escasa$ si puede ser aproximada por una combinación lineal de términos de la base vectorial $\boldsymbol{\Psi}$, donde $k \ll N^2LD$.

2.1.2. INCOHERENCIA Para obtener una adecuada compresión de la señal de interés, es necesario analizar el grado de correlación entre los elementos de la base de representación ($\boldsymbol{\Psi}$) y la base de muestreo ($\boldsymbol{\Phi}$). Formalmente, la coherencia entre dos bases ortonormales $\boldsymbol{\Phi} \in \mathbb{R}^{V\times N^2LD}$ y $\boldsymbol{\Psi} \in \mathbb{R}^{N^2LD\times N^2LD}$ con $V \ll N^2LD$, se define como el máximo valor absoluto del producto interno entre los elementos de las dos bases [17]:

$$\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) = \sqrt{N^2LD} * \max_{1\leq i,j\leq N^2LD} |\langle \Psi_i, \Phi_j \rangle|. \tag{2}$$

La coherencia cuantifica la mayor correlación entre cualquier par de elementos de $\boldsymbol{\Psi}$ y $\boldsymbol{\Phi}$. Si estas bases tienen elementos correlacionados el valor de la coherencia será grande, de lo contrario será pequeño. En el muestreo compresivo se busca que este

---

[17] CANDÈS, E. y ROMBERG, J. "Sparsity and incoherence in compressive sampling". En: *Inverse Problems* 23 (2007), págs. 969-985.

par de bases sea lo menos coherente posible [16], ya que entre menor sea su valor de coherencia, menor será el número de medidas necesarias para captar y recuperar con precisión la señal de interés. Usualmente se utilizan matrices aleatorias para realizar el muestreo de la señal, dado que tienen un alto grado de incoherencia con cualquier base de representación $\mathbf{\Psi}$.

2.2. ADQUISICIÓN COMPRESIVA DE SECUENCIAS DE VÍDEO ESPECTRAL

El sistema de adquisición de imágenes espectrales basado en aperturas codificadas (CASSI, por su sigla en inglés) es una de las arquitecturas ópticas de muestreo compresivo más conocidas, que capta y condensa información espacio-espectral en un conjunto de proyecciones 2D [5]. Diversos autores han adaptado esta arquitectura para muestrear escenas dinámicas a determinadas tasas de fotogramas. En este esquema de adquisición, la luz incidente es modulada espacialmente por una apertura codificada, y dispersada espectralmente por un prisma. Posterior a este paso, la energía codificada y dispersada se integra en el detector de imágenes denominado matriz de plano focal (FPA, por su sigla en inglés). En concreto, una apertura codificada es una arreglo matricial de valores binarios, en el cual los elementos iguales a 0 impiden el paso de radiación electromagnética. Una variante más versátil de este sistema es la arquitectura CASSI de color (C-CASSI) [4], donde se utiliza un arreglo de filtros ópticos de color en lugar de la apertura codificada binaria, que no sólo bloquea o deja pasar la luz a través de un conjunto de puntos espaciales, sino que también realiza un proceso de codificación en diferentes regiones del espectro electromagnético. En este trabajo, se adoptan ambos esquemas de adquisición para obtener información espacial, espectral y temporal en un formato comprimido. Matemáticamente, el proceso de adquisición compresiva de vídeo espectral con CASSI y C-CASSI se modela como se describe a continuación:

Considere $\mathcal{F} \in \mathbb{R}^{N \times N \times L \times D}$ una escena espectral dinámica, con $N \times N$ píxeles espaciales, L bandas espectrales y D cuadros temporales (fotogramas). La proyección CASSI en el FPA para un tiempo d puede modelarse como [7]:

$$(G_d)_{i,j} = \sum_{l=0}^{L-1} (\mathcal{F}_d)_{i,j-l,l} \, (\mathcal{T}_d)_{i,j-l} + \omega, \tag{3}$$

donde $(\mathcal{F}_d)_{i,j,l}$ representa las entradas discretas del d-ésimo fotograma, con i,j como índices espaciales, y l como índice de bandas espectrales; $(\mathcal{T}_d)_{i,j} \in \{0, 1\}$ son las entradas del código de apertura, que varían a lo largo al tiempo; y ω representa el ruido del sistema. Note que el desplazamiento en el eje j indica el efecto de dispersión inducido por el prisma. De manera similar, las proyecciones para el sistema C-CASSI vienen dadas por:

$$(G_d)_{i,j} = \sum_{l=0}^{L-1} (\mathcal{F}_d)_{i,j-l,l} \, (\mathcal{T}_d)_{i,j-l,l} + \omega, \tag{4}$$

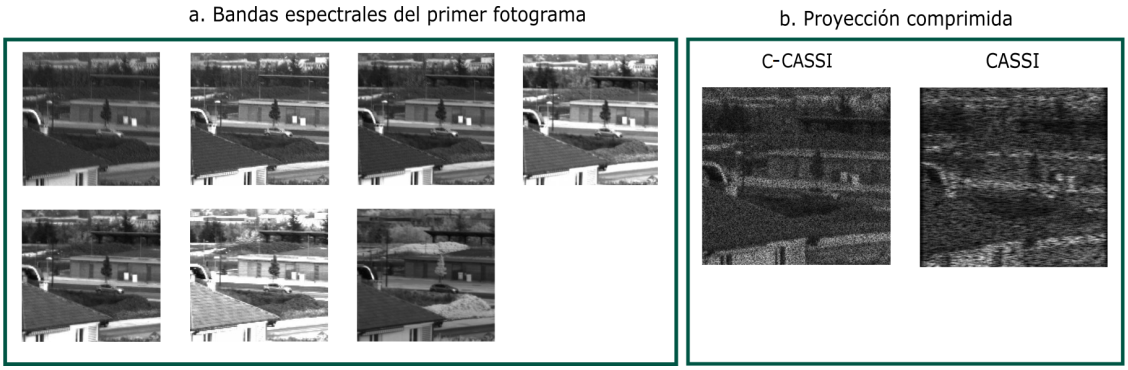
donde $(\mathcal{T}_d)_{i,j,l}$ contiene las respuestas espectrales de la apertura codificada de color. Teniendo en cuenta la adquisición de proyecciones de todos los cuadros temporales, el modelo discreto puede reescribirse en forma matricial como:

$$\boldsymbol{g} = \mathbf{H}\boldsymbol{f} + \boldsymbol{\omega}, \tag{5}$$

donde $\boldsymbol{f} \in \mathbb{R}^{N^2LD}$ representa la vectorización por columnas del vídeo espectral, $\boldsymbol{g} \in \mathbb{R}^{V}$ es la representación vectorial de las medidas comprimidas de todos los fotogramas G, con $V \ll N^2LD$; y $\mathbf{H}$ denota la matriz de proyección del sistema, que modela los fenómenos físicos de codificación y dispersión. Observe que, en general, la estructura de $\mathbf{H}$ para CASSI y C-CASSI es la misma, pero sus entradas difieren en función de la codificación binaria o de color empleada. En la figura 1 se presenta un ejemplo de un vídeo espectral de vigilancia. Específicamente, en la fig. 1(a) se muestra el primer fotograma del vídeo en diferentes longitudes de onda y en la fig.

1(b) se ilustra la proyección comprimida tanto en el sistema CASSI como C-CASSI.

Figura 1. Fotograma de un vídeo espectral y su correspondiente proyección comprimida



2.3. FLUJO ÓPTICO

El flujo óptico es un concepto que proviene de la psicofísica, fue introducido en la década de 1950 por el psicólogo James Gibson en su teoría sobre la percepción directa del movimiento [18]. Desde entonces, se ha propuesto un gran número de técnicas para cuantificar el movimiento en secuencias de imágenes, donde el objetivo es calcular un campo vectorial que modele los desplazamientos de la escena a nivel de píxel. Entre los métodos disponibles en la literatura se destacan las técnicas basadas en RNC y los enfoques variacionales, que son una extensión del algoritmo clásico de Horn-Schunck [19]. En general, los métodos variacionales estiman el flujo óptico a través de un problema de minimización de energía, donde se integra un término de fidelidad de datos y un término de suavidad. En el término de fidelidad de datos se asume que alguna propiedad en la imagen es invariante en el tiempo, mien-

---

[18] GIBSON, James J. y CARMICHAEL, Leonard. *The Perception of the Visual World*. Houghton Mifflin, 1950.

[19] HORN, Berthold y SCHUNCK, Brian. “Determining Optical Flow”. En: *Artificial Intelligence* 17 (ago. de 1981), págs. 185-203.

tras que en el término de suavidad se regulan los desplazamientos para garantizar la coherencia espacial [20]. Estos enfoques incorporan una estrategia multi-escala para detectar grandes desplazamientos, donde se genera una pirámide de múltiples resoluciones por cada imagen de entrada, de manera que los desplazamientos se van reduciendo en cada nivel de la jerarquía. Esto facilita el cómputo de las derivadas que conforman la función objetivo y permite refinar de forma iterativa la estimación del flujo óptico. Aunque la precisión de estos enfoques ha mejorado notablemente, el tiempo de ejecución es significativamente alto[21].

Recientemente, varios trabajos han optado por el uso de RNCs para la estimación del flujo óptico [22]. Estos enfoques realizan un entrenamiento de extremo a extremo, es decir, la RNC realiza tanto la extracción de características como el proceso de emparejamiento entre las imágenes de entrada, donde el objetivo es encontrar correspondencias por píxel entre ellas para estimar el campo de movimiento. La ventaja de estos enfoques radica en la precisión de estimación y en la eficacia del tiempo de ejecución.

2.4. SEGMENTACIÓN DE OBJETOS MEDIANTE APRENDIZAJE PROFUNDO

En los últimos años se ha incrementado el uso de redes neuronales convolucionales para resolver distintos desafíos relacionados con el análisis de imagen y vídeo. Se

---

[20] TU, Zhigang, *et al.* “A survey of variational and CNN-based optical flow techniques”. En: *Signal Processing: Image Communication* 72 (2019), págs. 9-24.

[21] SUN, Deqing, *et al.* “PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume”. En: (sep. de 2017).

[22] HUR, Junhwa y ROTH, Stefan. “Optical Flow Estimation in the Deep Learning Age”. En: *Modelling Human Motion: From Human Perception to Robot Design*. Ed. por NOCETI, Nicoletta; SCIUTTI, Alessandra y REA, Francesco. Springer International Publishing, 2020, págs. 119-140.

ha logrado avances significativos en tareas de segmentación, clasificación y detección de objetos. Actualmente, los métodos de mejor rendimiento en segmentación semántica [23] [24] [25] tienen en común una arquitectura profunda (ResNet, VGG, GoogleNet) que ha sido entrenada previamente en el conjunto de datos ImageNet. Estos trabajos reemplazan las capas totalmente conectadas de dichas arquitecturas por capas convolucionales, para poder generar una máscara de segmentación con la misma resolución de la imagen de entrada. Las arquitecturas convencionales de clasificación aplican operaciones de agrupamiento en cada bloque convolucional para reducir gradualmente la dimensión espacial del dato de entrada y así aumentar el campo receptivo, ya que se adquiere características en distintas jerarquías. Esto se conoce como muestreo descendente, el cual permite identificar qué objetos están presentes en la escena. Sin embargo, en este proceso se pierde información sobre la ubicación de dichos objetos, por tal razón es necesario emplear operaciones de interpolación que permitan escalar el último mapa de características a la resolución de la imagen de entrada, con el fin de generar predicciones densas a nivel de píxel. Estos cambios se manejan en distintas proporciones, por ejemplo, en el trabajo de J. Long et al. [23] se emplea una capa de convolución transpuesta, también denominada deconvolución, al final de la arquitectura para realizar el muestreo ascendente. Aunque se obtienen buenos resultados con esta red totalmente convolucional, se pierde un poco la información del contexto debido a las múltiples operaciones de submuestreo. En [24] se presenta una mejora a esta arquitectura, se incrementan las

---

[23] SHELHAMER, Evan; LONG, Jonathon y DARRELL, Trevor. “Fully Convolutional Networks for Semantic Segmentation”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (mayo de 2016), págs. 1-1.

[24] RONNEBERGER, O.; FISCHER, P. y BROX, T. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. En: *MICCAI*. 2015.

[25] CHEN, Liang-Chieh, *et al.* “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. En: (feb. de 2018).

capas de convolución transpuesta y se conectan los bloques de muestreo descendente y ascendente en distintos niveles para transferir características de bajo y alto nivel, mejorando notoriamente la precisión de segmentación. Esta arquitectura de red simétrica de tipo codificador-decodificador se conoce actualmente como U-net, la cual fue diseñada para trabajar con pocos datos de entrenamiento. Por otra parte, C. Chen et al. [25] desarrolla una red que emplea una arquitectura de tipo codificador-decodificador con una arquitectura de agrupación espacial en pirámide, la cual permite capturar información del contexto en diferentes escalas. Este trabajo emplea convoluciones dilatadas para controlar el aumento de parámetros y la complejidad computacional. Esta red denominada DeepLab logra segmentar objetos genéricos con gran precisión. Diversos proyectos [26] [27] [28] han adoptado estas arquitecturas para trabajar sobre secuencias de vídeo, con el propósito de segmentar objetos en movimiento en entornos dinámicos, así como realizar seguimiento a objetos específicos, generando la ubicación exacta de estos en cada cuadro del vídeo. Tokmakov et al. [26] emplea una arquitectura de estructura similar a la red U-net para aprender patrones de movimiento característicos, provenientes del flujo óptico de cuadros consecutivos de un conjunto de vídeos sintéticos. Asimismo, emplea campos aleatorios condicionales (CRF, por sus siglas en inglés) para refinar las máscaras binarias generadas por la red. En un trabajo posterior Tokmakov et al [27] integra un flujo con información de apariencia y un módulo de memoria visual basado en unidades recurrentes convolucionales (GRU) para propagar información espacial de los objetos

---

[26] TOKMAKOV, P.; KARTEEK, Alahari y SCHMID, C. “Learning Motion Patterns in Videos”. En: *CVPR*. 2017.

[27] TOKMAKOV, P.; KARTEEK, Alahari y SCHMID, C. “Learning Video Object Segmentation with Visual Memory”. En: *ICCV*. 2017.

[28] JAIN, Suyog; XIONG, Bo y GRAUMAN, Kristen. “FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos”. En: (ene. de 2017).

en el dominio temporal. De igual forma, S. Jain et al [28] propone una red totalmente convolucional de dos flujos, donde cada flujo codifica señales genéricas de apariencia y movimiento. Estas señales individuales se fusionan en la red para producir con precisión máscaras binarias que distinguen si un objeto se mueve o no en cada fotograma. Sin embargo, los métodos de alto rendimiento previamente mencionados han sido desarrollados para trabajar sobre secuencias de vídeos RGB. En el estado del arte no se han propuesto enfoques de aprendizaje profundo para detectar y segmentar objetos en movimiento en vídeos espectrales directamente en el dominio de compresión.

3. DETECCIÓN DE MOVIMIENTO EN MEDIDAS COMPRIMIDAS DE SECUENCIAS DE VÍDEO ESPECTRAL

Este trabajo introduce una estrategia computacional para detectar y segmentar regiones de movimiento prominente en secuencias de vídeo multiespectral directamente en el dominio de compresión. El enfoque propuesto puede discriminar entre el movimiento del entorno y del objeto de interés, considerando la información del espectro condensada en las proyecciones comprimidas junto con la información temporal. Para ello, se adoptan técnicas del estado del arte que permiten abordar el problema en cuestión. En primera instancia, se realiza una estimación del flujo óptico, el cual proporciona información sobre el movimiento aparente de los elementos de una escena, donde es posible determinar la magnitud y dirección de cada desplazamiento. En segunda instancia, se establece una arquitectura de red neuronal totalmente convolucional, para efectuar el aprendizaje de patrones característicos de apariencia y movimiento, donde es fundamental realizar un equilibrio de clases para evitar el sesgo del modelo. Las características de apariencia agrupan toda información semántica como texturas, bordes, formas y relaciones complejas entre las mismas que permiten identificar objetos en diferentes contextos, mientras que las características de movimiento representan patrones, que permiten determinar si ciertos cambios en las posiciones espaciales de los objetos son relevantes en relación con su entorno. A continuación, se describe en detalle las estrategias del enfoque propuesto.

3.1. ESTIMACIÓN DEL MOVIMIENTO

El flujo óptico es una herramienta que proporciona información sobre la disposición espacial de los objetos de interés en escenarios dinámicos, por lo cual es amplia-

mente utilizado en tareas de reconocimiento de acciones, detección, segmentación y seguimiento de objetos en secuencias de vídeo [29]. En esta sección se estudia el comportamiento de dos algoritmos de flujo óptico sobre las medidas comprimidas, donde se analiza la precisión de estimación del movimiento, así como el tiempo de ejecución. Los algoritmos explorados son DeepFlow [30] y LiteFlowNet [31]. DeepFlow, es un método variacional que incorpora una técnica para encontrar correspondencias densas entre los datos de entrada mediante descriptores de características. Esta técnica consiste en fragmentar las imágenes en pequeños segmentos, llamados parches, para encontrar similitudes en diferentes escalas por medio de una arquitectura piramidal, la cual integra capas de convolución y de agrupación máxima, cuya estructura se asemeja a una RNC. Sin embargo, en este enfoque no se considera el proceso de aprendizaje para el ajuste de los parámetros. LiteFlowNet, por su parte, es un método que introduce una arquitectura RNC compuesta por dos subredes: un codificador que genera dos pirámides de características de alta dimensión por cada par de imágenes de entrada, y un decodificador que desplaza (deforma) los mapas de activación de la segunda imagen hacia la primera, para hallar correspondencias en el espacio de características a fin de estimar el flujo óptico en cada nivel de la jerarquía. Además, en este método se incorpora un módulo de regularización que permite mejorar progresivamente la precisión de estimación. Específicamente, en este módulo se crean filtros de convolución local que se adaptan a las característi-

---

[29] GUPTA, Arpan y BALAN, M. Sakthi. “Action Recognition from Optical Flow Visualizations”. En: *Proceedings of 2nd International Conference on Computer Vision & Image Processing*. Springer Singapore, 2018, págs. 397-408.

[30] WEINZAEPFEL, Philippe, *et al.* “DeepFlow: Large Displacement Optical Flow with Deep Matching”. En: *2013 IEEE International Conference on Computer Vision* (2013), págs. 1385-1392.

[31] HUI, Tak-Wai; TANG, Xiaoou y LOY, Chen Change. “LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation”. En: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), págs. 8981-8989.

cas del codificador y a mapas de probabilidad de oclusión.

Aunque estos algoritmos han sido desarrollados para trabajar directamente sobre secuencias en escala de grises y RGB, se pueden adoptar para obtener un campo vectorial $\mathcal{O} = (\mathcal{O}_1, \mathcal{O}_2)$ por cada par de proyecciones comprimidas consecutivas $\{g_d,\ g_{d+1}\}$, el cual describe el movimiento horizontal y vertical en cada posición espacial (x, y). A partir de las componentes de dicho campo vectorial, se puede calcular la magnitud y dirección de cada desplazamiento, respectivamente como:

$$\|\mathcal{O}\|_{x,y} = \sqrt{\mathcal{O}_1(x,y)^2 + \mathcal{O}_2(x,y)^2} \tag{6}$$

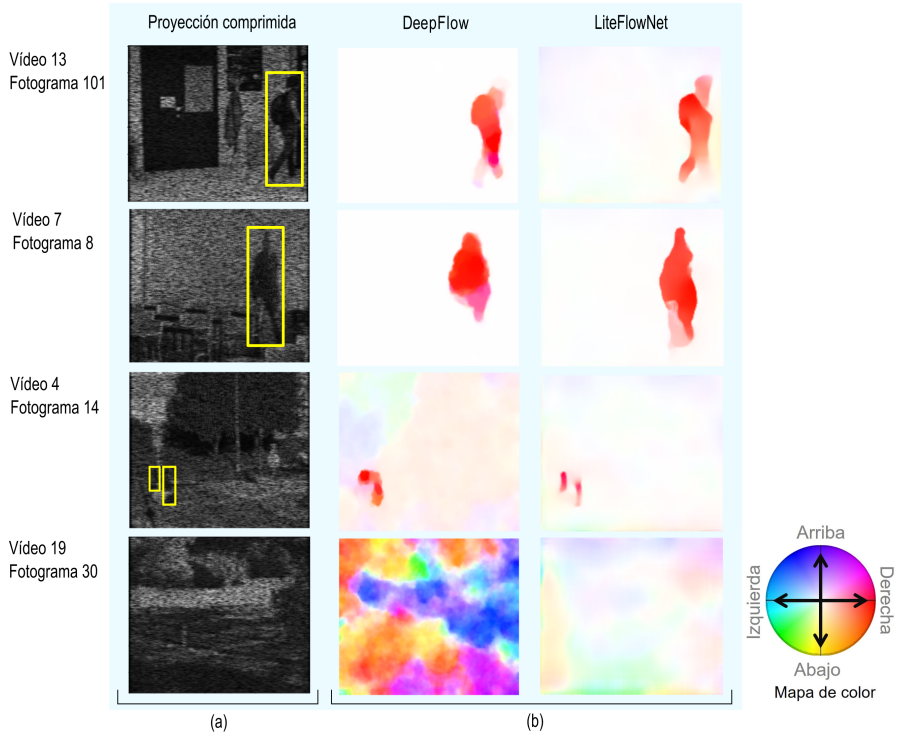
$$\theta_{x,y} = \tan^{-1} \frac{\mathcal{O}_1(x,y)}{\mathcal{O}_2(x,y)} \tag{7}$$

Para visualizar el flujo óptico, se utiliza la técnica propuesta en [32], donde se asigna un tono de color a cada dirección y la saturación se asocia con la magnitud del movimiento. La figura 2 muestra los resultados cualitativos de los algoritmos para 4 mediciones comprimidas de vídeo multiespectral. Cada proyección representa un escenario con un desafío específico: variabilidad en la apariencia del objeto (primera y segunda fila), oclusiones (tercera fila) y cambios de iluminación en ausencia de objetos en movimiento (cuarta fila). En esta figura, se puede observar que el método DeepFlow es sensible al ruido, siendo afectado por los cambios de iluminación de la escena. En cuanto al tiempo de ejecución, el método LiteFlowNet es $5{,}4$ veces más rápido. Por consiguiente, se adopta esta arquitectura de 5 millones de parámetros para estimar los vectores de desplazamientos. En el enfoque propuesto, el flujo óptico calculado a partir de cada par de proyecciones comprimidas, correspondientes a

---

[32] BAKER, S., *et al.* "A Database and Evaluation Methodology for Optical Flow". En: *International Journal of Computer Vision* 92 (2007), págs. 1-31.

Figura 2. Representación del flujo óptico en un mapa de color. (a) Proyeccción comprimida. (b) Estimación del movimiento con los métodos DeepFlow y LiteFlowNet

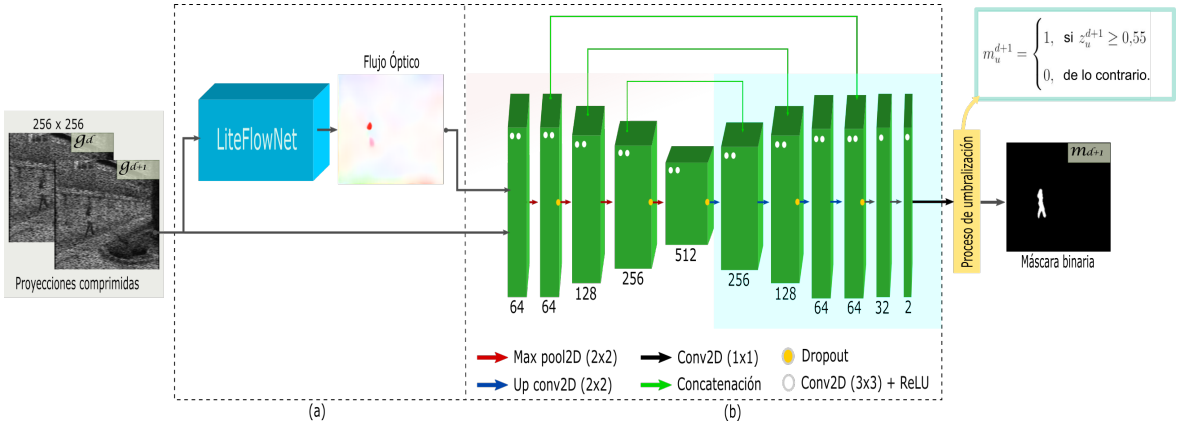


fotogramas consecutivos, se incorpora como una tercera entrada a una arquitectura de RNC 2D, donde se realizará la caracterización del movimiento y se descartarán los valores atípicos inducidos por el entorno. La figura 3 ilustra el esquema general del método propuesto anteriormente descrito; a continuación se describe en detalle la etapa que comprende el aprendizaje del movimiento en las proyecciones comprimidas.

3.2. APRENDIZAJE DE CARACTERÍSTICAS DE APARIENCIA Y MOVIMIENTO

En los últimos años, las arquitecturas convolucionales 3D se han utilizado para abordar tareas de análisis de vídeo, dado que permiten modelar el movimiento mediante representaciones de características espacio-temporales. Sin embargo, estas arqui-

Figura 3. Esquema ilustrativo del método propuesto. (a) Estimación del flujo óptico. (b) Aprendizaje de características espacio-temporales para la segmentación de objetos en movimiento en el dominio de compresión.



tecturas trabajan con un tamaño de entrada fijo, es decir, los vídeos del conjunto de datos deben tener el mismo número de fotogramas, no pueden variar en su resolución temporal[33]. Además, estas arquitecturas son costosas desde el punto de vista computacional, ya que los núcleos volumétricos de cada capa de convolución incrementan el número de parámetros[34] [35]. En este trabajo se presenta una método de aprendizaje profundo que emplea convoluciones 2D para capturar la información espacio-temporal de cuadros adyacentes en el dominio de compresión. En este sentido, se propone una extensión de la arquitectura U-net para llevar a cabo la segmentación del movimiento. Esta tarea se aborda como un problema de seg-

---

[33] KÖPÜKLÜ, Okan; HERZOG, Fabian y RIGOLL, G. "Comparative Analysis of CNN-based Spatio-temporal Reasoning in Videos". En: *ICPR Workshops*. 2020.

[34] TOUDJEU, Ignace Tchangou y TAPAMO, Jules-Raymond. "A 2D Convolutional Neural Network Approach for Human Action Recognition". En: *2019 IEEE AFRICON*. 2019, págs. 1-5.

[35] BURNEY, Atika y SYED, Tahir Q. "Crowd Video Classification Using Convolutional Neural Networks". En: *2016 International Conference on Frontiers of Information Technology (FIT)*. 2016, págs. 247-251.

mentación binaria, es decir, a cada píxel de la imagen se le asigna una de las dos etiquetas (objeto en movimiento o fondo). La arquitectura de red totalmente convolucional se ha utilizado con éxito en diversas aplicaciones, como la segmentación semántica, la estimación del flujo óptico y la superresolución [24] [36]. Este logro se atribuye a su capacidad de aprender características en diferentes niveles de abstracción con pocos datos de entrenamiento. La Figura 3(b) ilustra la topología de la arquitectura propuesta, cuyas entradas son pares de proyecciones comprimidas, y su correspondiente campo de flujo óptico$\{g_d,\ g_{d+1},\ \mathcal{O}_{d+1}\}$. Esta arquitectura aprovecha la correlación espacio-temporal de los datos para obtener mapas de probabilidad de movimiento de tamaño $N \times N \times 1$. En concreto, la red propuesta se compone de dos partes principales. La primera parte es la codificación, en la cual la resolución espacial de los datos se reduce gradualmente (en un factor de 2) para obtener características en múltiples jerarquías, proporcionando información semántica del contexto. Esta etapa consta de cinco bloques de muestreo descendente, cada uno de los cuales contiene dos capas de convolución con filtros de tamaño 3×3, y una capa de agrupación máxima de tamaño 2×2. La segunda parte es la decodificación, donde se utilizan operaciones de interpolación para escalar el mapa de características a la resolución original de los datos. De manera similar al método de Tokmakov et al. [26] y Rahmon et al [37], en este proceso se incluyen conexiones de salto entre el codificador y el decodificador para transferir representaciones de bajo y alto nivel, lo cual mejora significativamente la consistencia de la segmentación y solventa el problema del desvanecimiento del gradiente. La etapa de decodificación consta de

---

[36] HU, Xiaodan, *et al.* "RUNet: A Robust UNet Architecture for Image Super-Resolution". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

[37] RAHMON, Gani, *et al.* "Motion U-Net: Multi-cue Encoder-Decoder Network for Motion Segmentation". En: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, págs. 8125-8132.

seis bloques de muestreo ascendente, cada uno de los cuales contiene una capa de convolución transpuesta de tamaño 2×2, y entre una y dos capas de convolución con filtros de tamaño 3×3. La unidad lineal rectificada *(ReLU)* se aplica como función de activación en todas las capas de convolución, excepto en la última, donde se emplea una función sigmoide para producir un mapa de predicción a nivel de píxel. Finalmente, se realiza un proceso de umbralización para generar la máscara binaria de movimiento m_{d+1} del fotograma g_{d+1}. Con el objetivo de mejorar la precisión del enfoque propuesto, se incorporan capas de normalización de lotes y de abandono *(Dropout)* en diferentes bloques de la arquitectura, lo cual estabiliza el proceso de aprendizaje, reduce el sobre-ajuste y asegura la generalización del modelo. Cabe destacar que las múltiples capas de la red son necesarias para aprender eficazmente los patrones de movimiento y apariencia, así como para descartar los valores atípicos inducidos por la complejidad y diversidad de las escenas a exteriores.

4. DISEÑO EXPERIMENTAL

4.1. CONJUNTOS DE DATOS

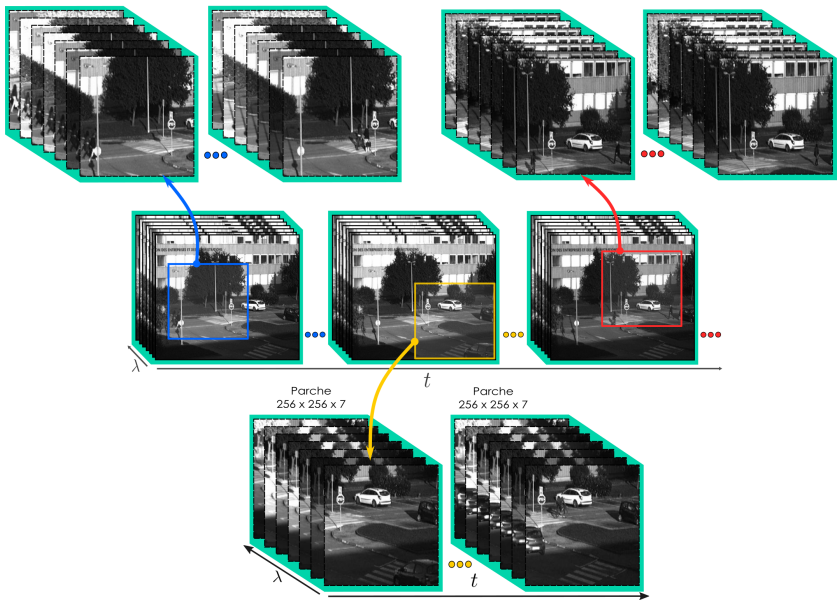
El método propuesto fue evaluado en dos conjuntos de datos que proporcionan vídeos de escenarios desafiantes con cambios de iluminación, fondo dinámico, efectos de camuflaje, oclusiones y objetos con movimiento intermitente. Estos conjuntos de datos contienen vídeos espectrales reales y estimados, donde cada fotograma cuenta con una máscara de referencia, la cual ha sido segmentada manualmente. En las siguientes subsecciones se consignan los detalles de ambas bases de datos.

4.1.1. FluxData FD-1665 dataset Este conjunto de datos ofrecido por el laboratorio de investigación ImVia [38], está constituido por un total de 5 vídeos multiespectrales, que contienen entre 250 y 2300 fotogramas. Cada fotograma cuenta con una resolución espacial de 658×492 píxeles y 7 canales espectrales, de los cuales 6 pertenecen al espectro visible y el restante, al infrarrojo cercano (NIR, por su sigla en inglés). Con el propósito de incrementar el número de muestras de este conjunto de datos, se decide fragmentar en 3 regiones los vídeos grabados en exteriores. Estos vídeos se caracterizan por los diversos eventos que se pueden presentar a lo largo de la secuencia. Específicamente, de cada región se seleccionó una secuencia de parches volumétricos de longitud arbitraria; cada parche tiene una dimensión de $256 \times 256 \times 7$. Como se ilustra en la Figura 4, las secuencias seleccionadas son las más representativas, ya que contienen más de un objeto en movimiento. Es necesario aclarar que los parches de cada región se muestrearon consecutivamente en el eje

[38] BENEZETH, Yannick; SIDIBÉ, Désiré y THOMAS, Jean Baptiste. “Background subtraction with multispectral video sequences”. En: (jun. de 2014).

temporal.

Figura 4. Ilustración del proceso de extracción de parches. Cada cuadrado de color indica una región espacial diferente, de la cual se extrae una secuencia representativa de parches volumétricos. Cada parche cuenta con una dimensión de $256 \times 256 \times 7$. Las secuencias seleccionadas contienen múltiples objetos en movimiento.



4.1.2. LASIESTA Database Es una colección de más de 20 vídeos RGB que agrupan diversos desafíos para evaluar las técnicas de detección de objetos en movimiento. Cada vídeo tiene entre 250 y 1400 fotogramas, los cuales poseen una resolución de 352×258 píxeles espaciales [39]. De esta base de datos se seleccionaron 13 vídeos para construir un conjunto de secuencias multiespectrales de 7 canales

[39] CUEVAS, Carlos; YÁÑEZ, Eva María y GARCÍA, Narciso. "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA". En: *Computer Vision and Image Understanding* 152 (2016), págs. 103-117.

por medio de la red neuronal HSCNN+ [40]. Esta red ganadora del desafío *NTIRE 2018 Spectral Reconstruction Challenge*, permite estimar bandas espectrales en 31 longitudes de onda para cada fotograma RGB.

4.2. PROYECCIONES COMPRIMIDAS

Todos los fotogramas se redimensionaron a 256×256 píxeles en resolución espacial para unificar ambos conjuntos de datos. Las medidas comprimidas se adquirieron simulando los modelos de los sistemas CASSI y C-CASSI descritos en la Ecuación (3) y Ecuación (4), respectivamente. Para ello, se utilizaron aperturas codificadas con un nivel de transmitancia del 50 % (la proporción de información que pasa a través de la apertura); estos patrones de codificación se generaron aleatoriamente a partir de una distribución Bernoulli. De esta manera, se conforma un total de 24 mediciones comprimidas de vídeo multiespectral (MCVM) que suman 2770 fotogramas. El 75 % de estas MCVM se utilizaron para entrenar el modelo convolucional propuesto y 25 % para validar su desempeño en diferentes escenarios. Específicamente, para construir el conjunto de entrenamiento, se seleccionaron aleatoriamente 10 y 8 MCVM del primer y segundo conjunto de datos, respectivamente. Las MCVM restantes se utilizaron para conformar el conjunto de pruebas, el cual permite evaluar el método propuesto en cuanto a la capacidad de segmentar objetos en movimiento en escenarios desconocidos. Este procedimiento se realizó tanto para las proyecciones CASSI como C-CASSI.

---

[40] SHI, Zhan , *et al.* "HSCNN+: Advanced CNN-Based Hyperspectral Recovery from RGB Images". En: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, págs. 1052-10528.

4.3. EQUILIBRIO DE CLASES

Mediante un análisis experimental, se identificó que solo el 4,81 % de los píxeles del conjunto de datos de entrenamiento representaba un objeto en movimiento. Esto es un claro ejemplo de un alto desequilibrio de clases, lo cual afecta el rendimiento de la RNC, ya que la desproporción de observaciones entre categorías sesga al modelo. Para abordar este problema, se adopta una técnica proveniente de la regresión logística para ponderar las clases en la función de pérdida; a fin de reducir tanto los falsos negativos como los falsos positivos. El objetivo es penalizar más al modelo por clasificar erróneamente las muestras de la clase minoritaria. Como se describe en la ecuación 8, el peso w_p es inversamente proporcional a la frecuencia de la clase p,

$$w_p = \frac{n}{km_p}, \qquad p \in \{0, 1\}, \tag{8}$$

donde $k = 2$ representa el total de clases, n el total de muestras y m_p el número de observaciones en la clase p. En este trabajo, se utiliza la función pérdida entropía cruzada binaria, la cual se minimiza durante el proceso de entrenamiento para encontrar los parámetros óptimos θ de la red, que permiten extraer características espacio-temporales de los datos para realizar la clasificación a escala de píxel. La función de pérdida se define como

$$\mathcal{L}_d(z_d, y_d) = \frac{-1}{N^2} \sum_{u=1}^{N^2} [\ w_1 y_u^d \log(z_u^d)\ +\ w_0 (1 - y_u^d) \log(1 - z_u^d)\ ], \tag{9}$$

con $z_d = \mathcal{N}_\theta(g_{d-1},\ g_d,\ \mathcal{O}_d)$, donde $\mathcal{L}_d(.)$ denota el valor de pérdida correspondiente al d^{th} ejemplo; y_u^d y z_u^d representa la etiqueta de referencia y la salida de la red en la ubicación del píxel u del ejemplo d, respectivamente; w_1 es el costo de clasificar erróneamente una muestra de la clase de movimiento como fondo (falso negativo) y w_0 es el costo de clasificar incorrectamente una muestra de la clase de fondo como

movimiento (falso positivo).

4.4. AUMENTO DE DATOS

En este trabajo se implementa una técnica de aumento de datos para generar variaciones de las medidas comprimidas, dado que los datos obtenidos no son suficientes para garantizar la generalización del modelo de clasificación. Esta técnica consiste en aplicar transformaciones geométricas a las proyecciones comprimidas, que se utilizan en la fase de entrenamiento para ajustar el modelo. Para este fin, se consideran rotaciones entre $10°$ y $30°$, desplazamientos horizontales y verticales en un rango de $[-38, 38]$ píxeles. Con esta técnica se obtienen muestras ligeramente modificadas que no se alejan de la realidad y por ende no afecta el aprendizaje del algoritmo. Cabe aclarar que dichas transformaciones también se aplicaron a las máscaras de referencia y al flujo óptico. Esta técnica nace bajo la premisa de que a mayor número de muestras, mejor será la precisión de estimación del modelo [41].

4.5. MÉTRICAS DE VALIDACIÓN

El rendimiento del método propuesto se evalúa mediante tres métricas que permiten cuantificar la calidad de la segmentación en las proyecciones comprimidas: Intersección sobre unión (*IoU*), medida F (*F-score*) y porcentaje de clasificaciones erróneas (*PWC*). Para calcular estas métricas se deben comparar las predicciones del modelo con las máscaras de referencia, donde se asignan los siguientes valores: Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN). VP representa las muestras de la clase 'movimiento' que

---

[41] CUBUK, Ekin D., *et al.* "AutoAugment: Learning Augmentation Strategies From Data". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

se clasificaron correctamente, FP indica el número de muestras de la clase 'fondo' que el modelo etiquetó como 'movimiento', VN corresponde al total de muestras de la clase 'fondo' predichas correctamente y FN representa las muestras de la clase 'movimiento' etiquetadas como 'fondo'. Teniendo en cuenta estos valores, las métricas mencionadas previamente se describen a continuación:

- **Intersección sobre Unión:** También conocida como índice Jaccard, es esencialmente una métrica de superposición que permite cuantificar el grado de correspondencia entre las máscaras de referencias y las predicciones del modelo a nivel de píxel. Este medida se expresa como:

$$IoU = \frac{VP}{VP + FP + VN} \tag{10}$$

- **Medida-F:** Se utiliza para evaluar los sistemas de clasificación binaria, ya que permite agrupar en una sola medida la precisión (P) y la exhaustividad (E) del modelo, dando a cada propiedad la misma ponderación. Esta medida se calcula a partir de:

$$F = 2\frac{E * P}{E + P}, \tag{11}$$

  donde $\mathrm{E} = \frac{\mathrm{VP}}{\mathrm{VP+FN}}$ y $\mathrm{P} = \frac{\mathrm{VP}}{\mathrm{VP+FP}}$

- **PWC:** Mide la proporción de predicciones incorrectas con respecto al total de clasificaciones realizadas. Este porcentaje de error se define como:

$$PWC = \frac{FN + FP}{VP + FN + FP + VN} \tag{12}$$

4.6. CONFIGURACIÓN DE LA ARQUITECTURA CONVOLUCIONAL 2D

Para el entrenamiento del modelo convolucional se utilizó el optimizador *Adam*, el cual es una versión eficiente del método de gradiente descendente, donde cada pa-

rámetro tiene su propia tasa de aprendizaje adaptativo. Experimentalmente, las mejores configuraciones de los hiperparámetros fueron un tamaño de muestra (*batch-size*) de 5 y una tasa de aprendizaje de 0,001. La arquitectura se entrenó durante 45 épocas para cada sistema de adquisición.

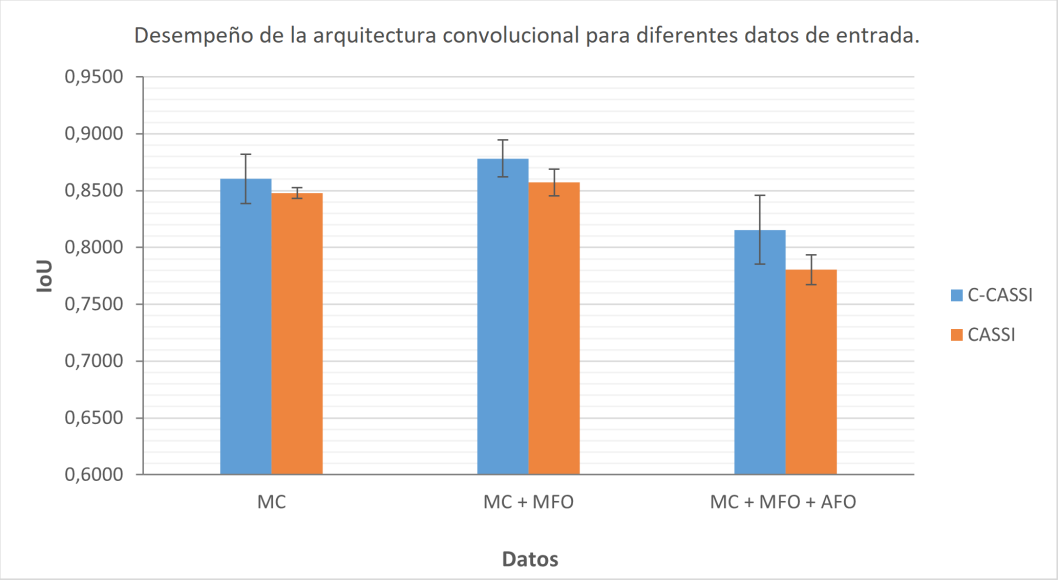
5. EVALUACIÓN Y RESULTADOS

5.1. MAGNITUD Y DIRECCIÓN DEL CAMPO DE FLUJO ÓPTICO

El objetivo de esta sección es analizar la contribución del ángulo y la magnitud del flujo óptico en la detección de objetos en movimiento en el dominio de compresión. Para este fin, se realizaron 3 experimentos con la arquitectura convolucional 2D propuesta en este trabajo. En cada experimento se aplicó una técnica de validación cruzada denominada *K-fold cross-validation*, donde $K = 4$ representa el número de iteraciones. En este caso, las MCVM se agruparon en 4 subconjuntos de igual tamaño, a fin de reservar en cada iteración un subconjunto para la prueba y el resto para el entrenamiento. La figura 5 resume los resultados obtenidos en cada experimento al suministrar diferentes datos de entrada: proyecciones comprimidas, magnitud y dirección del desplazamiento. El primer experimento se diseñó con el propósito de evaluar el comportamiento del modelo convolucional en los datos comprimidos, por lo cual solo se utilizaron los subconjuntos de MCVM, es decir, no se agregó información del flujo óptico. En general, es posible notar que la arquitectura convolucional logra capturar e integrar adecuadamente características espacio-temporales de cuadros adyacentes, alcanzando en promedio un IoU de 84,80 % $\pm$ 0,47 % (desviación estándar) y 86,03 % $\pm$ 2,17 % (desviación estándar) en las medidas de los sistemas de adquisición CASSI y C-CASSI, respectivamente. Al suministrar las proyecciones comprimidas en pares independientes, es decir, al considerar solo el cambio de dos fotogramas consecutivos, se garantiza un apropiado modelamiento de los patrones de movimiento, debido a que se reduce la captura de ruido presente en las secuencias. De esta manera se evitan las distorsiones temporales, especialmente en aquellas secuencias grabadas en exteriores. En el segundo experimento se incluyó la magnitud del flujo óptico para complementar la información de las CSVM.

Como era de esperar, el dato adicional permite reforzar la tarea de clasificación, siendo favorable para ambos sistemas de adquisición, reportando en promedio una ganancia del 1,08 % en las proyecciones CASSI y 2,09 % en C-CASSI. En el tercer experimento se agregó la dirección del flujo óptico, donde se obtuvo una puntuación media de 78,05 % $\pm$ 1,32 % y 81,54 % $\pm$ 3,03 % en las medidas CASSI y C-CASSI, respectivamente. En este caso, el rendimiento del modelo es afectado por el ruido que introduce este dato adicional, el cual puede estar asociado a los fuertes cambios de iluminación en las secuencias. Con base en estos resultados, se puede afirmar que el ángulo del flujo óptico no aporta significativamente al problema en cuestión; a diferencia de otras aplicaciones donde es un punto clave para identificar gestos y acciones en secuencias de vídeo, e incluso para realizar seguimiento a objetos específicos.

Figura 5. Resultados de la validación cruzada *4-fold*. Se puede apreciar que el modelo logra una adecuada clasificación de movimiento con los datos MC + MFO. La línea vertical en cada barra representa la desviación estándar. *MC: Medidas comprimidas, MFO: Magnitud del flujo óptico, AFO: Ángulo del flujo óptico



5.2. DESEMPEÑO DEL ENFOQUE PROPUESTO RESPECTO A MÉTODOS DEL ESTADO DEL ARTE

Una vez identificadas las ventajas de utilizar la magnitud del flujo óptico como complemento de las CSVM, se procede a analizar la calidad de la segmentación del enfoque propuesto en las secuencias del conjunto de prueba. Asimismo, se realiza una comparación con un método del estado del arte que trabaja en el dominio de compresión [14]. Los autores de este método utilizan operaciones morfológicas junto con técnicas de umbralización para segmentar la diferencia temporal de proyecciones adyacentes. Es importante aclarar que en los experimentos se utilizó el código fuente original del artículo en mención. Las tablas 1 y 2 presentan los resultados promedio de cada enfoque para las métricas de precisión estudiadas, en las medidas CASSI y C-CASSI, respectivamente. Aunque la técnica tradicional requiere de menos tiempo de ejecución para generar las máscaras binarias, el método propuesto obtiene mejores puntuaciones en cada secuencia. Por otra parte, las figuras 6, 7, 8, 9, 10 y 11 muestran los resultados cualitativos de cada método con sus respectivos valores de IoU. En estas figuras, se puede apreciar la consistencia de la segmentación de la estrategia propuesta con respecto a las máscaras de referencia. En contraste, la técnica tradicional presenta ciertas limitaciones debido a la configuración (tamaño y forma) del elemento estructurante y al valor de intensidad del proceso de umbralización. Como se observa en las figuras, el método tradicional captura ruido del entorno y en muy pocos casos segmenta completamente al objeto de interés.

Adicionalmente, se lleva a cabo una comparación con un método que utiliza la información del espectro para detectar objetos en movimiento [8]. En este método se propone una técnica de sustracción de fondo que mide la similitud espectral entre los píxeles del fotograma actual y un modelo de referencia; en dicha técnica los

umbrales se establecen y actualizan automáticamente. Los autores de este método reportan para el vídeo 1 un *F-measure* de 0.9307. Al revisar la tabla 2, se puede observar que la estrategia propuesta obtiene un resultado comparable utilizando sólo las proyecciones comprimidas del vídeo espectral, los valores no difieren en gran medida. La figura 7 da cuenta de ello, donde claramente se puede apreciar que las máscaras binarias de estos dos enfoques son muy similares. Dado que realizar la detección de movimiento en el vídeo espectral implica la reconstrucción de cada fotograma, en este trabajo se utiliza el algoritmo iterativo ADMM[42] para estimar el tiempo de reconstrucción de cada secuencia del conjunto de prueba. En la tabla 3, se puede observar que el costo computacional para recuperar un vídeo espectral a partir de medidas comprimidas es significativamente alto. Con la estrategia propuesta se evita este costo adicional, ya que el proceso de reconstrucción no es necesario.

Tabla 1. Resultados promedio del método propuesto y del enfoque tradicional [44] en las proyecciones CASSI. En los datos del método propuesto se presenta el tiempo de respuesta de la red U-net (izquierda) y LiteFlowNet (derecha).

Secuencias de prueba	Total de fotogramas	IoU		F-measure		PWC		Tiempo de CPU (s)	
		Método propuesto	Técnica tradicional	Método propuesto	Técnica tradicional	Método propuesto	Técnica tradicional	Método propuesto	Técnica tradicional
Vídeo 4	55	**0,8615**	0,7007	**0,8497**	0,6582	**0,4784**	1,5783	34,10 + 48,61	**3,09**
Vídeo 7	82	**0,8421**	0,7741	**0,8236**	0,7499	**0,7206**	1,2172	50,10 + 74,84	**4,50**
Vídeo 10	16	**0,8726**	0,7330	**0,8959**	0,7724	**2,2218**	5,3927	10,30 + 12,26	**1,07**
Vídeo 13	111	**0,8597**	0,6540	**0,8480**	0,5679	**2,1949**	5,1525	68,40 + 111,29	**5,53**
Vídeo 19	109	**0,8604**	0,6655	**0,8487**	0,5649	**0,8098**	1,7486	67,40 + 103,45	**4,87**
Vídeo 21	152	**0,8471**	0,7813	**0,8108**	0,7694	**0,7130**	0,9034	93,00 + 148,43	**9,91**

[42] BOYD, Stephen , *et al.* “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. En: *Foundations and Trends in Machine Learning* 3 (2011), págs. 1-122.

Tabla 2. Resultados promedio del método propuesto y del enfoque tradicional [45] en las proyecciones C-CASSI. En los datos del método propuesto se presenta el tiempo de respuesta de la red U-net (izquierda) y LiteFlowNet (derecha).

Secuencias de prueba	Total de fotogramas	IoU		F-measure		PWC		Tiempo de CPU (s)	
		Método propuesto	Técnica tradicional	Método propuesto	Técnica tradicional	Método propuesto	Técnica tradicional	Método propuesto	Técnica tradicional
Video 1	254	**0,9315**	0,84153	**0,9113**	0,79533	**1,0274**	2,3964	111,00 + 235,60	**4,43**
Video 4	55	**0,8582**	0,67613	**0,8514**	0,60802	**0,5158**	1,9782	24,00 + 42,62	**1,97**
Video 8	88	**0,8733**	0,74585	**0,8616**	0,73351	**1,1729**	2,2613	38,60 + 78,74	**1,89**
Video 14	83	**0,9067**	0,74065	**0,8987**	0,67576	**0,2631**	0,7720	36,20 + 69,33	**1,42**
Video 15	85	**0,8919**	0,7087	**0,8884**	0,57524	**1,2930**	4,2812	37,20 + 69,28	**1,77**
Video 22	62	**0,8010**	0,7034	**0,7748**	0,62799	**0,3392**	0,5384	27,20 + 50,34	**1,36**

Tabla 3. Tiempos de reconstrucción en CPU utilizando el algoritmo iterativo ADMM [46]

Secuencias	Tiempo de reconstrucción (s)
Vídeo 1	2125,5
Vídeo 4	370,47
Vídeo 7	360,80
Vídeo 8	428,04
Vídeo 10	152,13
Vídeo 13	699,30
Vídeo 14	557,76
Vídeo 15	493,85
Vídeo 19	773,90
Vídeo 21	931,76
Vídeo 22	560,10

5.3. DESEMPEÑO RESPECTO A LA TASA DE ADQUISICIÓN DE DATOS

En esta sección se analiza el comportamiento del método propuesto en cuanto a la capacidad de detectar movimiento en proyecciones de secuencias de vídeo con un mayor número de bandas espectrales. Esto debido a que la tasa de datos adquiri-

dos en los sistemas como CASSI y C-CASSI está definido por la proporción entre el número de proyecciones captadas (V) y el número total de variables del conjunto de datos original (N^2LD). En este sentido, a medida que se aumenta el número de bandas espectrales, se obtiene una menor tasa de adquisición de datos. En otras palabras, este valor indica el porcentaje de datos utilizados por los algoritmos que trabajan con las medidas comprimidas. En este trabajo, todos los experimentos contemplan una adquisición del sistema compresivo por cada fotograma (aproximadamente N^2 valores para CASSI o C-CASSI), en consecuencia, aumentar el valor de L implica una menor tasa de adquisición de datos, que se traduce en mayor complejidad para resolver cualquier problema de inferencia directamente sobre las proyecciones comprimidas. El objetivo de este experimento consiste en verificar el comportamiento del algoritmo desarrollado en la detección del movimiento cuando las proyecciones utilizadas como entrada involucran una mayor cantidad de bandas espectrales, es decir cuando se varía la tasa de adquisición de datos. Para este fin, se utilizó el modelo convolucional previamente entrenado en las medidas del conjunto de datos de 7 bandas de los experimentos anteriores. Es importante aclarar que el modelo se utilizó únicamente en la fase de inferencia, y no se realizó ningún proceso de entrenamiento en función del número de bandas.

El experimento se realizó sobre las proyecciones comprimidas de 4 vídeos, cuya resolución espectral se incrementó gradualmente, con el objetivo de analizar el desempeño en diferentes tasas de adquisición de datos, como se indica en la Tabla 4. La figura 13 presenta los resultados en términos de IoU para los sistemas CASSI y C-CASSI. La precisión del modelo disminuye a medida que aumenta el número de bandas, debido a la complejidad que implica agrupar mayor cantidad de información en el mismo número de medidas. Sin embargo, se puede apreciar que el algoritmo desarrollado detectó el movimiento en vídeos con alrededor del doble de bandas

espectrales que los vídeos usados para el entrenamiento, con una disminución en la precisión de a lo sumo 5.6 % en CCASSI y 17.6 % en CASSI. Por otra parte, las figuras 14 y 15 ilustran las máscaras binarias en 4 tasas de adquisición de datos; en dichas figuras se puede apreciar la similitud de los resultados y la consistencia de la segmentación respecto a las máscaras de referencia. Estos resultados permiten concluir que el algoritmo propuesto está en capacidad de detectar el movimiento en medidas comprimidas de videos espectrales con un mayor número de bandas sin necesidad de reentrenamiento, proporcionando niveles de pérdida de precisión aceptables (menores del 18 %) tanto para CASSI como para C-CASSI.

Tabla 4. Tasa de adquisición de datos según el número de bandas espectrales.

L	7	10	13	16
Tasa de adquisición	14 %	10 %	7.6 %	6.25 %

Figura 6. Visualización de las máscaras binarias generadas por cada método en las proyecciones C-CASSI. En la parte superior e inferior se presentan los resultados obtenidos para los vídeos 1 y 4, respectivamente. En (a) se muestran las máscaras de 4 fotogramas con sus respectivos valores de IoU. En (b) se ilustran las entradas del modelo convolucional.

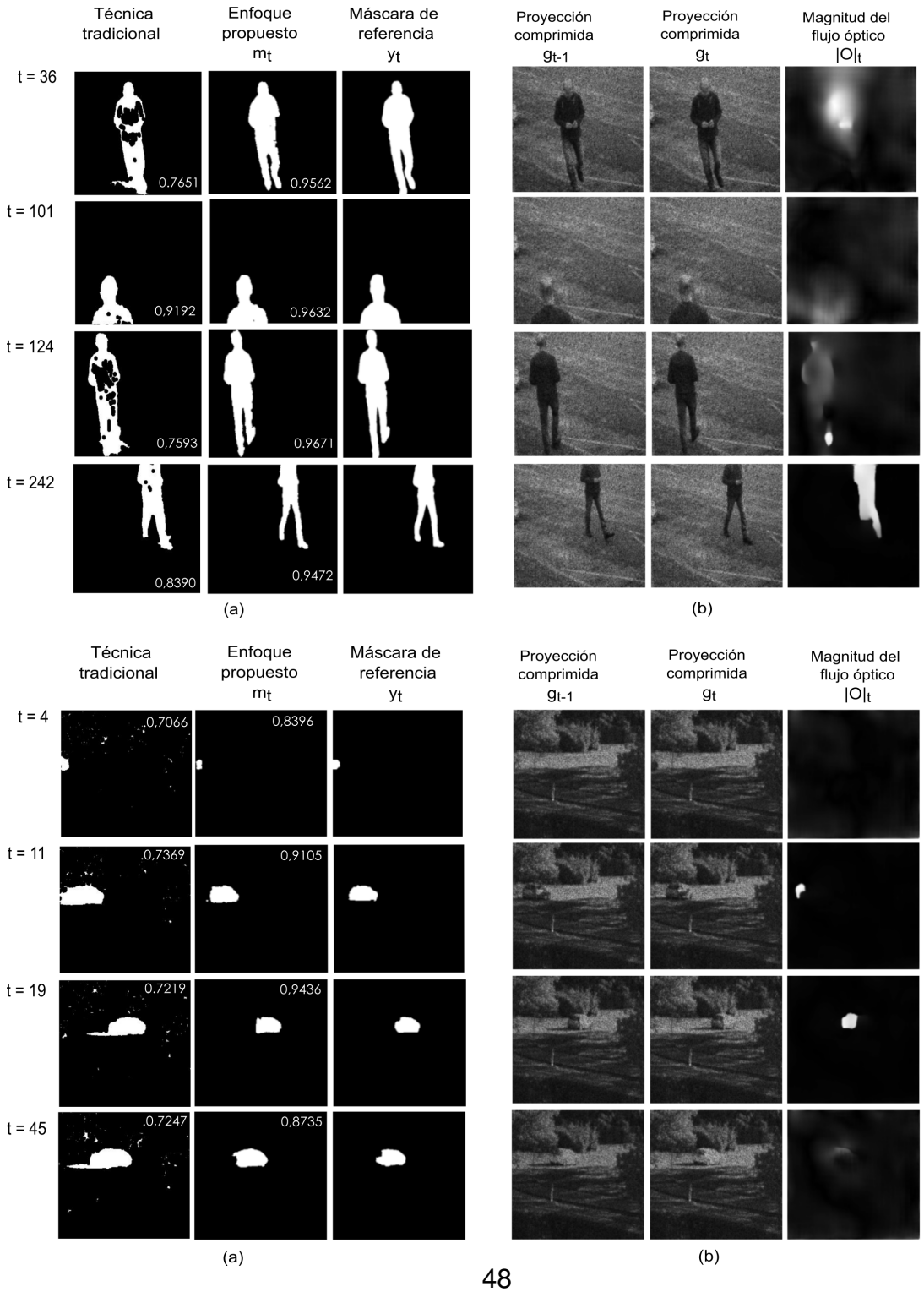


Figura 7. Visualización de las máscaras binarias generadas por cada método en las proyecciones C-CASSI. En la parte superior e inferior se presentan los resultados obtenidos para los vídeos 8 y 14, respectivamente. En (a) se muestran las máscaras de 4 fotogramas con sus respectivos valores de IoU. En (b) se ilustran las entradas del modelo convolucional.

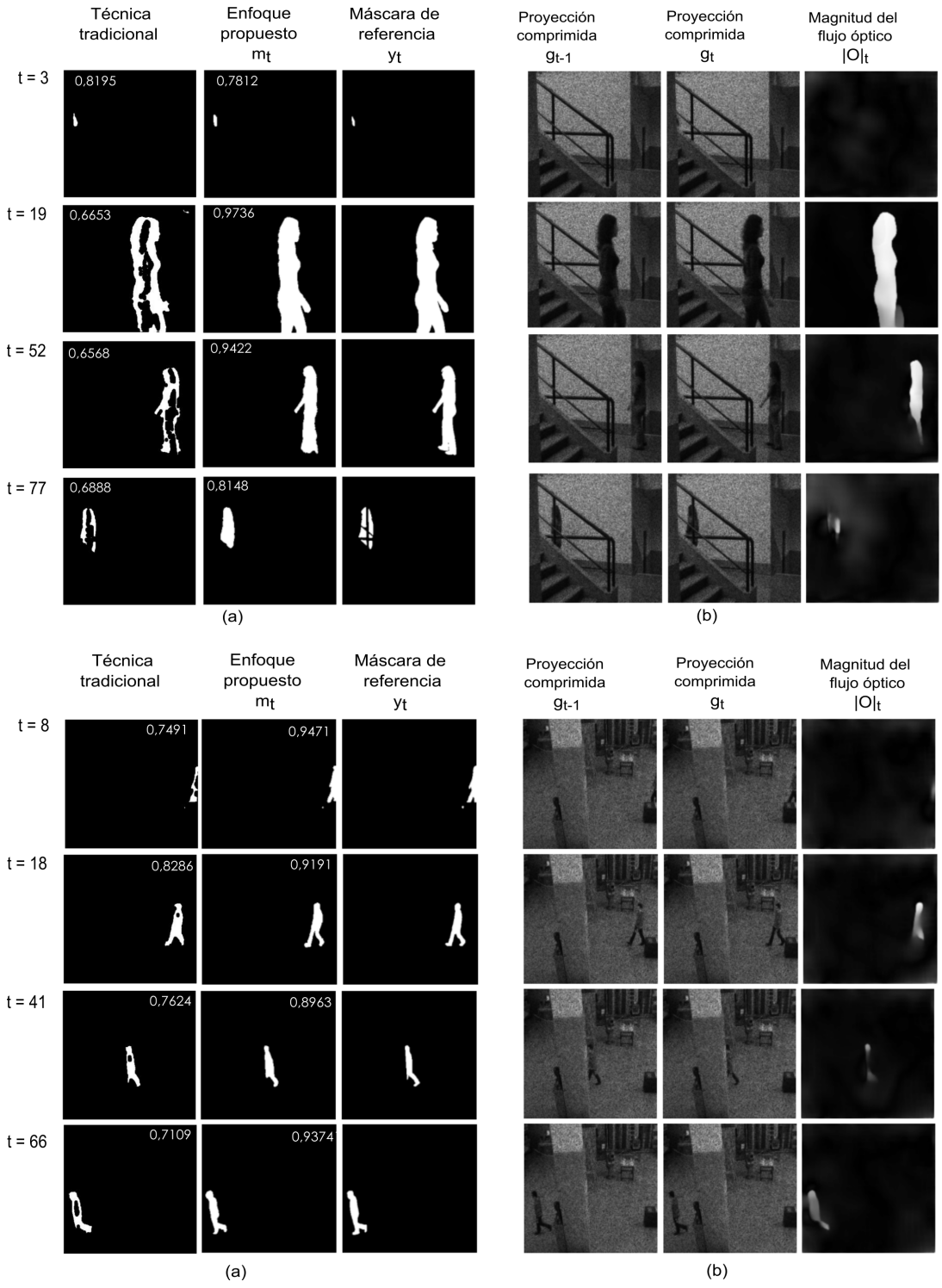


Figura 8. Visualización de las máscaras binarias generadas por cada método en las proyecciones C-CASSI. En la parte superior e inferior se presentan los resultados obtenidos para los vídeos 15 y 22, respectivamente. En (a) se muestran las máscaras de 4 fotogramas con sus respectivos valores de IoU. En (b) se ilustran las entradas del modelo convolucional.

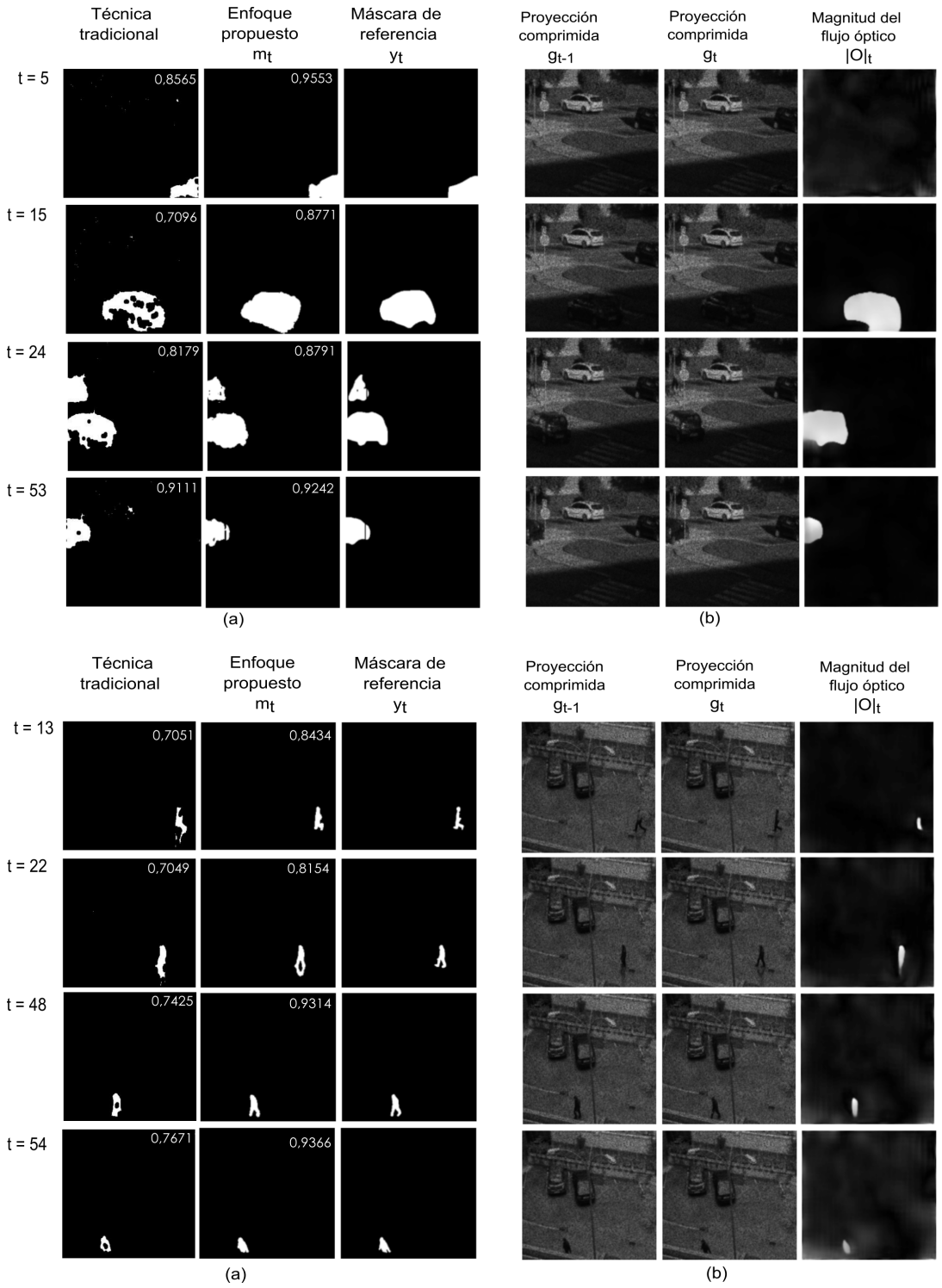


Figura 9. Visualización de las máscaras binarias generadas por cada método en las proyecciones CASSI. En la parte superior e inferior se presentan los resultados obtenidos para los vídeos 7 y 10, respectivamente. En (a) se muestran las máscaras de 4 fotogramas con sus respectivos valores de IoU. En (b) se ilustran las entradas del modelo convolucional.

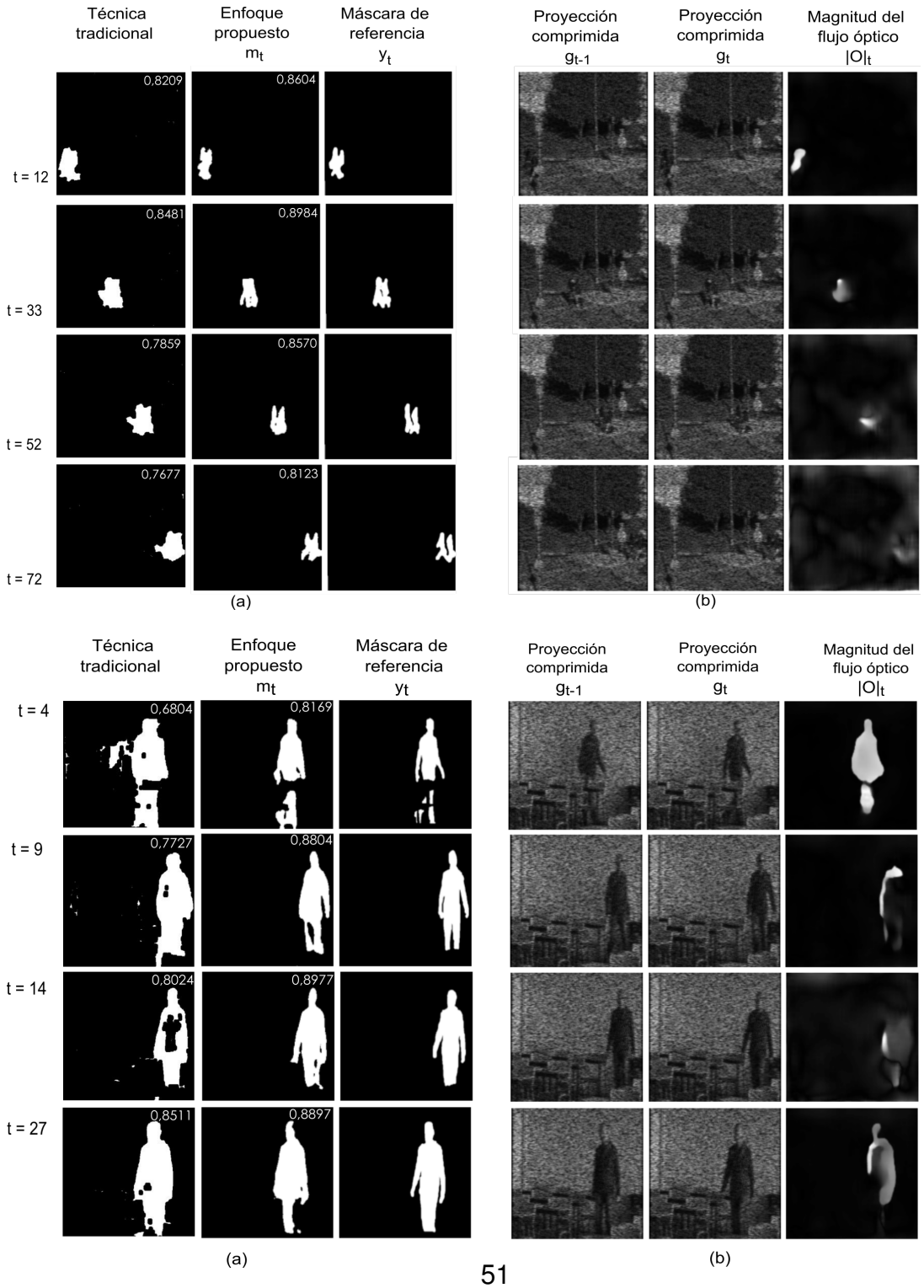


Figura 10. Visualización de las máscaras binarias generadas por cada método en las proyecciones CASSI. En la parte superior e inferior se presentan los resultados obtenidos para los vídeos 19 y 13, respectivamente. En (a) se muestran las máscaras de 4 fotogramas con sus respectivos valores de IoU. En (b) se ilustran las entradas del modelo convolucional.

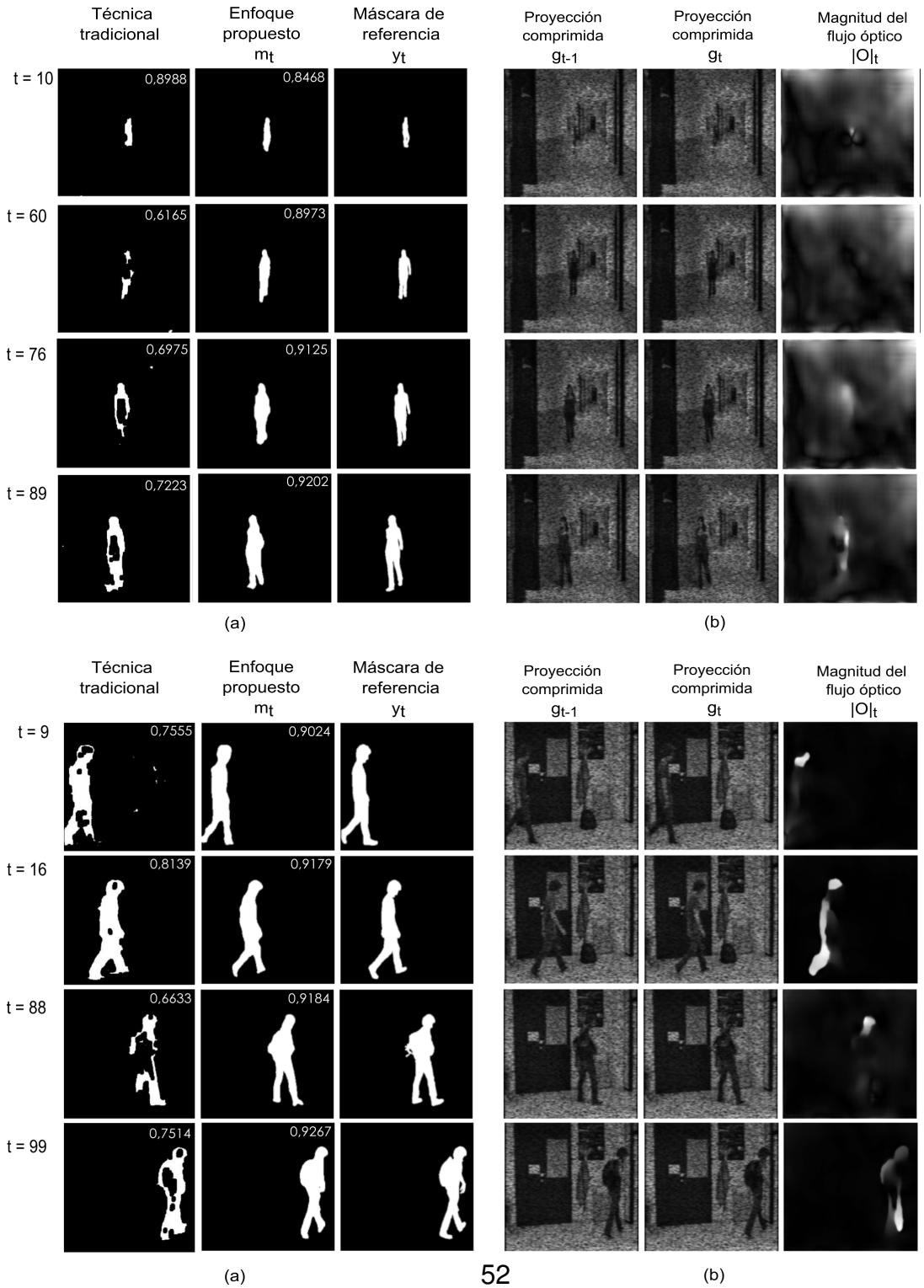


Figura 11. Visualización de las máscaras binarias generadas por cada método en las proyecciones CASSI. En la parte superior e inferior se presentan los resultados obtenidos para los vídeos 4 y 21, respectivamente. En (a) se muestran las máscaras de 4 fotogramas con sus respectivos valores de IoU. En (b) se ilustran las entradas del modelo convolucional.

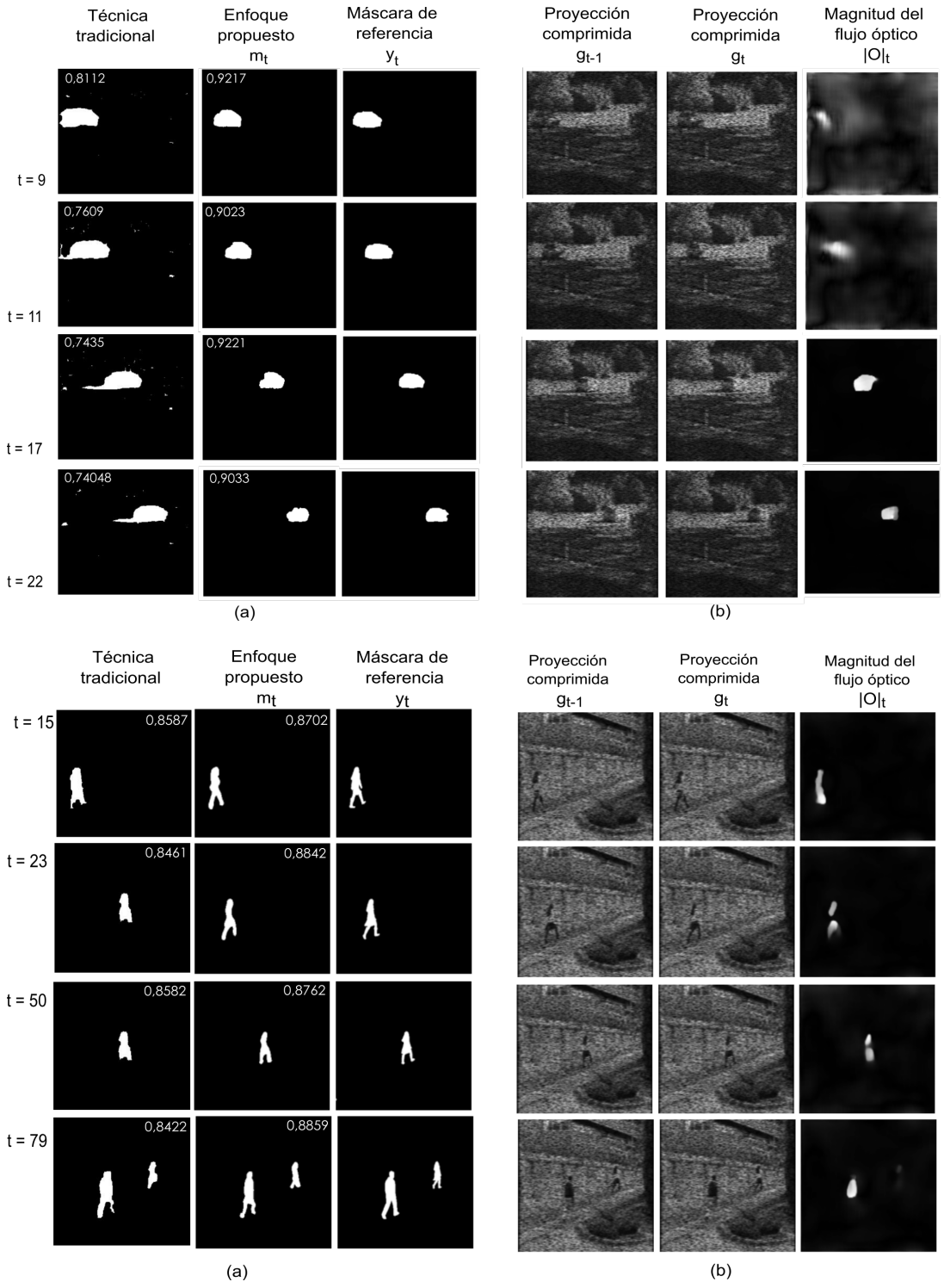


Figura 12. Comparación de resultados de la metodología propuesta respecto a la técnica de sustracción de fondo en datos multidimensionales[43]

Figura 13. Comportamiento del método propuesto para CASSI y C-CASSI en función del número de bandas de los vídeos originales.

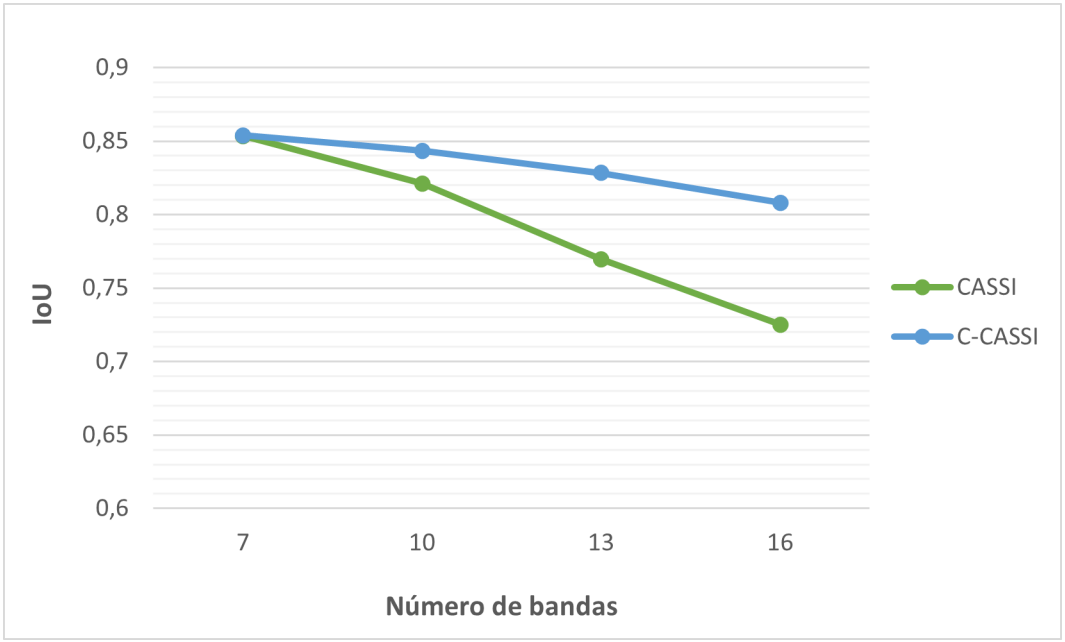


Figura 14. Resultados cualitativos del método propuesto en las proyecciones C-CASSI para 4 tasas de adquisición de datos.

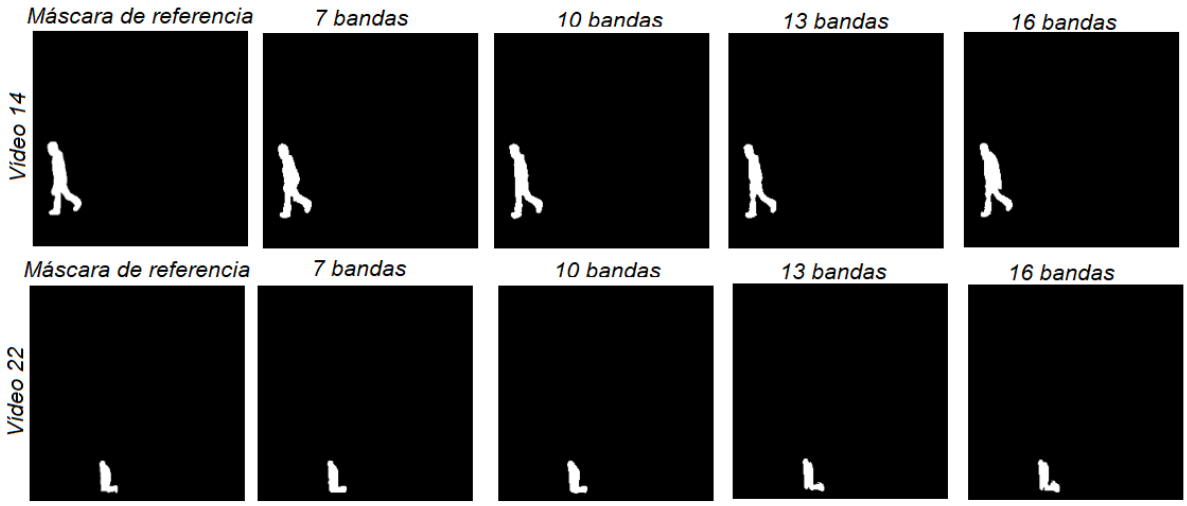
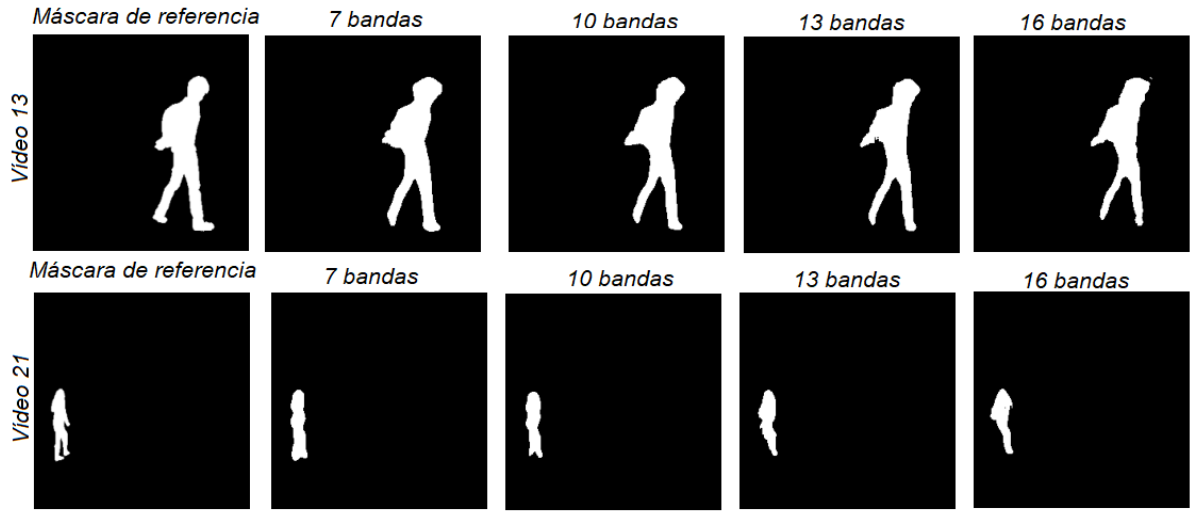


Figura 15. Resultados cualitativos del método propuesto en las proyecciones CASSI para 4 tasas de adquisición de datos.



6. CONCLUSIONES

En este trabajo de investigación se desarrolló un método para segmentar objetos en movimiento en mediciones comprimidas de vídeo multiespectral. La estrategia propuesta incorpora una arquitectura de red totalmente convolucional para modelar características espacio-temporales, que permiten realizar la discriminación del movimiento en diferentes escenarios, evitando el costo computacional de los algoritmos de reconstrucción. Las capas de convolución 2D de dicha arquitectura logran capturar e integrar adecuadamente patrones de apariencia y movimiento de proyecciones adyacentes. Con la finalidad de garantizar la generalización del modelo convolucional, se implementaron tres estrategias de aumento de datos: extracción de parches, transformaciones geométricas y estimación de bandas espectrales mediante la red HSCNN+. Además, se adoptó una técnica de ponderación para penalizar al modelo en la función de perdida, a fin de reducir las clasificaciones erróneas en la clase minoritaria. En el desarrollo de este proyecto se pudo evidenciar que la magnitud del flujo óptico contribuye en gran medida a la detección de movimiento en el dominio de compresión. El método propuesto fue evaluado mediante diferentes métricas que permitieron cuantificar el desempeño en cuanto a la capacidad de segmentar objetos en escenarios desconocidos. Los experimentos realizados demuestran que el enfoque propuesto proporciona los mejores resultados respecto al método del estado del arte, que utiliza técnicas tradicionales de procesamiento de señales como operaciones morfológicas, filtrado espacial y diferencia temporal. Sin embargo, requiere más tiempo de CPU para generar las máscaras binarias que dicho método. En general, se obtiene en promedio una ganancia de 29 % en C-CASSI y 24 % en CASSI en términos de *F-measure*. Adicionalmente, el método propuesto proporciona resultados comparables a los de un enfoque que requiere del paso de reconstrucción para llevar acabo esta tarea de inferencia. Por otra parte, se resalta

la capacidad de generalización del modelo convolucional para realizar la detección de movimiento en proyecciones con más multiplexación de bandas espectrales, sin necesidad de reentrenamiento, proporcionando pérdida de precisión menor de 18 % tanto para CASSI como para C-CASSI.

BIBLIOGRAFÍA

BAKER, S., *et al.* “A Database and Evaluation Methodology for Optical Flow”. En: *International Journal of Computer Vision* 92 (2007), págs. 1-31 (vid. pág. 29).

BENEZETH, Yannick; SIDIBÉ, Désiré y THOMAS, Jean Baptiste. “Background subtraction with multispectral video sequences”. En: (jun. de 2014) (vid. pág. 34).

BOYD, Stephen, *et al.* “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. En: *Foundations and Trends in Machine Learning* 3 (2011), págs. 1-122 (vid. págs. 44, 45).

BURNEY, Atika y SYED, Tahir Q. “Crowd Video Classification Using Convolutional Neural Networks”. En: *2016 International Conference on Frontiers of Information Technology (FIT)*. 2016, págs. 247-251 (vid. pág. 31).

CANDÈS, E. y ROMBERG, J. “Sparsity and incoherence in compressive sampling”. En: *Inverse Problems* 23 (2007), págs. 969-985 (vid. pág. 19).

CANDES, E. J. y WAKIN, M. “An Introduction To Compressive Sampling”. En: *IEEE Signal Processing Magazine* 25 (2008), págs. 21-30 (vid. págs. 18, 20).

CAO, Xun, *et al.* “Computational Snapshot Multispectral Cameras: Toward dynamic capture of the spectral world”. En: *IEEE Signal Processing Magazine* 33.5 (2016), págs. 95-108 (vid. págs. 13, 20).

CHEN, Liang-Chieh, *et al.* “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. En: (feb. de 2018) (vid. págs. 24, 25).

CUBUK, Ekin D., *et al.* “AutoAugment: Learning Augmentation Strategies From Data”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (vid. pág. 38).

CUEVAS, Carlos; YÁÑEZ, Eva María y GARCÍA, Narciso. “Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA”. En: *Computer Vision and Image Understanding* 152 (2016), págs. 103-117 (vid. pág. 35).

GIBSON, James J. y CARMICHAEL, Leonard. *The Perception of the Visual World*. Houghton Mifflin, 1950 (vid. pág. 22).

GUPTA, Arpan y BALAN, M. Sakthi. “Action Recognition from Optical Flow Visualizations”. En: *Proceedings of 2nd International Conference on Computer Vision & Image Processing*. Springer Singapore, 2018, págs. 397-408 (vid. pág. 28).

HINOJOSA, Carlos; RAMIREZ, Juan Marcos y ARGUELLO, Henry. “Spectral-Spatial Classification from Multi-Sensor Compressive Measurements Using Superpixels”. En: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, págs. 3143-3147 (vid. pág. 14).

HORN, Berthold y SCHUNCK, Brian. “Determining Optical Flow”. En: *Artificial Intelligence* 17 (ago. de 1981), págs. 185-203 (vid. pág. 22).

HU, Xiaodan, *et al.* “RUNet: A Robust UNet Architecture for Image Super-Resolution”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019 (vid. pág. 32).

HUI, Tak-Wai; TANG, Xiaoou y LOY, Chen Change. “LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation”. En: *2018 IEEE/CVF Confe-*

rence on Computer Vision and Pattern Recognition (2018), págs. 8981-8989 (vid. pág. 28).

HUR, Junhwa y ROTH, Stefan. “Optical Flow Estimation in the Deep Learning Age”. En: *Modelling Human Motion: From Human Perception to Robot Design*. Ed. por Nicoletta NOCETI; Alessandra SCIUTTI y Francesco REA. Springer International Publishing, 2020, págs. 119-140 (vid. pág. 23).

JAIN, Suyog; XIONG, Bo y GRAUMAN, Kristen. “FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos”. En: (ene. de 2017) (vid. pág. 15).

— “FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos”. En: (ene. de 2017) (vid. págs. 25, 26).

KÖPÜKLÜ, Okan; HERZOG, Fabian y RIGOLL, G. “Comparative Analysis of CNN-based Spatiotemporal Reasoning in Videos”. En: *ICPR Workshops*. 2020 (vid. pág. 31).

LIU, Rongrong; RUICHEK, Yassine y EL BAGDOURI, Mohammed. “Multispectral Dynamic Codebook and Fusion Strategy for Moving Objects Detection”. En: *Image and Signal Processing*. 2020, págs. 35-43 (vid. págs. 14, 16, 43).

LU, Guolan y FEI, Baowei. “Medical hyperspectral imaging: a review”. En: *Journal of Biomedical Optics* 19.1 (2014), págs. 1-24 (vid. pág. 13).

LÓPEZ, Kareth León; GALVIS, Laura y FUENTES, Henry Arguello. “Spatio-spectro-temporal coded aperture design for multiresolution compressive spectral video sensing”. En: *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017, págs. 728-732 (vid. pág. 14).

LÓPEZ, Kareth León; GALVIS, Laura y FUENTES, Henry Arguello. “Temporal Colored Coded Aperture Design in Compressive Spectral Video Sensing”. En: *IEEE Transactions on Image Processing* 28.1 (2019), págs. 253-264 (vid. págs. 13, 20).

PINILLA, Samuel, *et al.* “Salient Motion Detection for Spectral Video on the Compressive Domain”. En: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2019, págs. 106-110 (vid. págs. 15, 16, 43-45, 54).

PUGLIESE, Claudia; CARREÑO, Diana y ARGUELLO, Henry. “Sparse representations of dynamic scenes for compressive spectral video sensing”. En: *DYNA* 83 (2016), págs. 42-51 (vid. págs. 14, 21).

RAHMON, Gani, *et al.* “Motion U-Net: Multi-cue Encoder-Decoder Network for Motion Segmentation”. En: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, págs. 8125-8132 (vid. pág. 32).

RAMIREZ, Juan Marcos y ARGUELLO, Henry. “Spectral Image Classification From Multi-Sensor Compressive Measurements”. En: *IEEE Transactions on Geoscience and Remote Sensing* 58.1 (2020), págs. 626-636 (vid. pág. 14).

RONNEBERGER, O.; FISCHER, P. y BROX, T. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. En: *MICCAI*. 2015 (vid. págs. 24, 32).

SHAW, Gary y BURKE, Hsiao-hua. “Spectral Imaging for Remote Sensing”. En: *Lincoln Laboratory Journal* 14 (2003), págs. 3-28 (vid. pág. 13).

SHELHAMER, Evan; LONG, Jonathon y DARRELL, Trevor. “Fully Convolutional Networks for Semantic Segmentation”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (mayo de 2016), págs. 1-1 (vid. pág. 24).

SHI, Zhan , *et al.* “HSCNN+: Advanced CNN-Based Hyperspectral Recovery from RGB Images”. En: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, págs. 1052-10528 (vid. pág. 36).

STANKOVIC, Ljubisa, *et al.* “A Tutorial on Sparse Signal Reconstruction and Its Applications in Signal Processing”. En: *Circuits, Systems, and Signal Processing* 38 (mar. de 2019) (vid. págs. 18, 19).

SUN, Deqing, *et al.* “PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume”. En: (sep. de 2017) (vid. pág. 23).

TOKMAKOV, P.; KARTEEK, Alahari y SCHMID, C. “Learning Motion Patterns in Videos”. En: *CVPR*. 2017 (vid. págs. 25, 32).

— “Learning Video Object Segmentation with Visual Memory”. En: *ICCV*. 2017 (vid. pág. 25).

TOUDJEU, Ignace Tchangou y TAPAMO, Jules-Raymond. “A 2D Convolutional Neural Network Approach for Human Action Recognition”. En: *2019 IEEE AFRICON*. 2019, págs. 1-5 (vid. pág. 31).

TU, Zhigang, *et al.* “A survey of variational and CNN-based optical flow techniques”. En: *Signal Processing: Image Communication* 72 (2019), págs. 9-24 (vid. pág. 23).

VARGAS, Hector; FONSECA, Yesid y ARGUELLO, Henry. “Object Detection on Compressive Measurements using Correlation Filters and Sparse Representation”. En: *2018 26th European Signal Processing Conference (EUSIPCO)*. 2018, págs. 1960-1964 (vid. pág. 14).

WANG, Yujie, *et al.* “Discrimination of nitrogen fertilizer levels of tea plant (Camellia sinensis) based on hyperspectral imaging”. En: *Journal of the Science of Food and Agriculture* 98 (2018), págs. 4659-4664 (vid. pág. 13).

WEINZAEPFEL, Philippe, *et al.* “DeepFlow: Large Displacement Optical Flow with Deep Matching”. En: *2013 IEEE International Conference on Computer Vision* (2013), págs. 1385-1392 (vid. pág. 28).

YASIN, Hashim y HAYAT, Saqib. “DeepSegment: Segmentation of motion capture data using deep convolutional neural network”. En: *Image and Vision Computing* 109 (2021), pág. 104147 (vid. pág. 15).

ANEXOS

Anexo A. Productos Académicos

- Lucena, V., Correa, C., & Arguello, H. (2021, September). Automatic Motion Segmentation of Spectral Videos in the Compressed Domain using a Fully Convolutional Network. In 2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA).