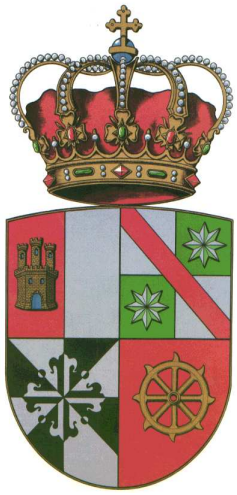
UNIVERSIDAD DE CASTILLA-LA MANCHA

Escuela Superior de Ingeniería Informática
Departamento de Sistemas Informáticos



Ph.D. Dissertation

ROBUST HUMAN DETECTION THROUGH FUSION OF COLOR AND INFRARED VIDEO

A dissertation submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

Juan Serrano Cuerda

Advisors

Prof. Antonio Fernández-Caballero
Dr. María Teresa López Bonal

Agradecimientos

El desarrollo de este trabajo no hubiera sido posible sin el apoyo de una serie de personas e instituciones. Espero que estas líneas sirvan para expresar mi gratitud hacia ellos, por mucho que cueste condensarla en solo una pequeña porción de texto.

En primer lugar, a mis directores Antonio y Maite, por ayudarme en mis primeros pasos en el mundo de la investigación. Por guiarme aportando consejos e ideas reorientándome cuando perdía de vista el camino a seguir. Por su paciencia y por haberme brindado el honor de trabajar con ellos.

A mis padres por haberme dado su cariño y amor durante todos este largo camino. Por aportar siempre fuerza en los malos momentos y alegría en los buenos, así como su ánimo incondicional.

A mis compañeros del Instituto de Investigación en Informática de Albacete, por ofrecer su apoyo, amistad incondicional y confianza cuando yo mismo dudaba. Sin su colaboración no habría sido posible esta tesis y, aunque muchos ya no sigan allí por desgracia, su influencia sigue presente. Estas líneas por mucho que intente, se quedarán cortas para expresar mi gratitud hacia ellos.

A todos los amigos con los que sé que puedo contar a diario, haciéndome reír y brindándome consuelo en los malos momentos y disfrutando y celebrando a mi lado los buenos. Por constituir conmigo una gran familia con la que sé que puedo contar a diario al igual que ellos saben que siempre estaré ahí cuando lo necesiten. Porque todos estos años espero que parezcan poco en relación al brillante futuro que estoy convencido que nos espera, convencido de que, aunque la vida nos pueda distanciar físicamente en algún momento, siempre estaremos juntos.

A todos los profesores que han contribuido a mi formación en estos años y sin cuyas enseñanzas este trabajo no sería posible.

A los Ministerios de Economía y Competitividad, de Industria, Energía y Turismo y la Junta de Comunidades de Castilla-La Mancha, por los proyectos de investigación que han permitido el desarrollo de este trabajo.

Summary

Nowadays, human detection systems are a key challenge in the field of Computer Vision. Many people detection systems are based on the use of color cameras. Yet, these cameras have problems when the scene is poorly illuminated or when there are sudden lighting changes. This is why, the use of thermal-infrared cameras seems to be an interesting alternative. Indeed, these cameras show a good performance in cold environments, but they offer many troubles in warm scenarios. In these conditions humans' temperature is similar to the thermal readings of the remaining elements in the scene. This fact makes it hard to distinguish humans from the environment.

This work aims at the development and implementation of a robust human detection system based on fusing the information provided after segmenting infrared and color spectra videos. The final system has been developed based on the *INT*^{3}*-Horus* framework. This framework was recently designed and implemented in collaboration with other members and PhD students of the *n&aIS* research group.

The framework starts with an initial acquisition level which grabs frames from both cameras and synchronizes their output. The features from the captured images are also used to assign a confidence level to each spectrum. This level is a key to perform information fusion between the infrared and visible spectra. The next level is the segmentation of infrared and visible videos. A series of different human detection algorithms have been implemented at this level, both in infrared and visible spectrum. In the second case color information is used. The different approaches are compared between each other with the objective of choosing the most suitable human detection algorithm for each spectrum. The results of the selected algorithms are used at the fusion level.

Now, the fusion level analyzes the results achieved by the human detection approaches in each spectrum. Fusion is based on a rule-based system which uses the confidence level assigned to each spectrum. The location of each human detected by each segmentation approach is also used. In accordance with the rules, the decision taken can be (1) to add humans to the final result, (2) to better adjust the dimensions and the amount of detected humans by analyzing the results gotten by the other spectrum, or (3) to ignore the current detection if the confidence level assigned to the spectrum is not above a certain value. Finally, a tracking algorithm decides if the humans previously detected remain in the scene, although they have not been detected un the current frame.

An outdoor environment has been chosen to evaluate the system. Different sequences were recorded in diverse atmospherical and illumination conditions. The sequences also offer a varying complexity, where a different number of humans are present in diverse situations. These situations can be as simple as a single human walking on the scene, and as complex as multiple people walking in a group. The performed tests show a significant improvement between the results achieved by human detection algorithms focused on a single spectrum and those offered after information fusion. This improvement does not only occur in adverse conditions for each spectrum, but also in situations where both spectra collaborate through reinforcing their results. Thus, it is confirmed that the system has a stable performance regardless of the monitored environment's conditions.

Resumen

Hoy en día los sistemas de detección de humanos constituyen un desafío en boga dentro del campo de la visión artificial. Muchos de estos sistemas se encuentran basados en el uso de cámaras en color, aunque pueden presentar problemas ante condiciones adversas de iluminación, ya sea debidas a la oscuridad o a cambios súbitos de luz. Una alternativa interesante la constituye el uso de cámaras térmico-infrarrojas, las cuales funcionan bien en entornos fríos, pero son problemáticas en ambientes cálidos, ya que la temperatura de los humanos es la misma que la del resto de elementos presentes en la escena y es difícil distinguir a los primeros del resto del entorno.

Esta tesis persigue el diseño e implementación de un sistema robusto de detección de personas basado en la fusión de la información proporcionada por la segmentación de humanos en los espectros infrarrojo y color. El sistema se ha elaborado a partir de la arquitectura *INT*^{3}*-Horus*, diseñada e implementada en colaboración con otros miembros y doctorandos del grupo de investigación *n&aIS*.

Parte esta arquitectura de un nivel inicial de adquisición que se encarga de capturar las imágenes de ambas cámaras y sincronizarlas. Además, las características de las imágenes capturadas sirven para asignar un nivel de confianza a cada espectro. Este nivel de confianza es decisivo a la hora de realizar la fusión entre ambos espectros. El siguiente nivel lo constituye la segmentación de los vídeos infrarrojo y visible. En esta etapa se han implementado una serie de algoritmos de detección de humanos, tanto en el espectro infrarrojo como en el visible, usando en este último la información del color, con el fin de poder comparar sus resultados en distintas situaciones. Los diversos algoritmos son comparados entre sí con el fin de elegir únicamente una técnica de detección de humanos en cada espectro. Los resultados de estos algoritmos son los que se tienen en cuenta en el nivel de fusión.

El nivel de fusión se encarga de analizar los resultados obtenidos en las detecciones de humanos en cada espectro. La fusión se basa en un sistema de reglas que utiliza la confianza asignada a cada espectro. También se tiene en cuenta la localización de los humanos detectados por cada algoritmo de segmentación. La decisión tomada a partir de estas reglas puede ser (1) incorporar el humano al resultado final, (2) ajustar mejor las dimensiones y el número de personas detectadas a partir del análisis de los resultados del otro espectro, o (3) ignorar la detección actual por no tener asignada un nivel de confianza suficiente el espectro analizado. Los humanos incorporados al resultado final del sistema por estas reglas son comparados con los humanos detectados anteriormente en la escena. Para ello se utiliza un algoritmo de identificación que establece correspondencias entre los humanos actuales y aquellos detectados en la escena en la iteración anterior. Finalmente, un algoritmo de seguimiento decide si todavía permanecen en la escena aquellos humanos que se encontraban anteriormente en el escenario y no han sido localizados en el fotograma actual.

Con el fin de evaluar el sistema, se ha escogido un entorno de exteriores, en el cual se han grabado una serie de secuencias en diversas condiciones atmosféricas y de iluminación. Además, las secuencias presentan una complejidad variable, donde pueden aparecer distintas cantidades de humanos en diferentes situaciones, desde las más simples, como un humano caminando sólo, a las más complejas, como personas caminando en grupo. Las pruebas realizadas demuestran una mejora significativa en-

tre los resultados alcanzados por los algoritmos de detección en cada espectro en solitario y aquellos logrados gracias a la fusión. Esta mejora no se da únicamente en situaciones adversas para uno de los dos espectros, sino en aquéllas en las que los dos espectros colaboran reforzando mutuamente sus resultados. Con ello, se confirma que el sistema presenta un funcionamiento estable independientemente de las condiciones en que se encuentre el entorno monitorizado en cada momento.

Índice general

Índice de figuras

Índice de tablas

Lista de acrónimos

Término	Significado
B_{IR}	Candidato a humano detectado en en el espectro visible
BF	Humano detectado inicialmente por la fusión
B_V	Candidato a humano detectado en en el espectro visible
C_{IR}	Nivel de confianza en el espectro infrarrojo
C_V	Nivel de confianza en el espectro visible
HT	Humano detectado en la escena al final del algoritmo de fusión
L_{BF}	Lista de humanos detectados inicialmente por el algoritmo de fusión
L_{ID}	Lista de humanos identificados actualmente en la escena
$L_{HT}(t)$	Lista final de humanos en la escena en el instante t
ROI	Región de interés
$ROI_{B_{IR}}$	Región de interés relativa a una mancha B_{IR}
ROI_{B_V}	Región de interés relativa a una mancha B_V

Chapter 1

Introduction

This chapter explains the motivations and main objectives of the proposed PhD dissertation. After introducing the background of this dissertation, the context in which this research takes place is described. Then, the main motivations of this work are explained. Next, the main objective of this work is described. The general objective is divided into a series of sub-objectives. Finally, the outline of this PhD dissertation is detailed.

1.1. Background

Detecting people is a key technology for many applications, especially in the video surveillance domain. At the same time it is one of the most challenging problems in computer vision and remains a scientific challenge for realistic and challenging scenes.

People surveillance is one of the hottest topics of the last decade in computer vision and pattern recognition research; it covers all aspects of computer engineering and computer science, models and algorithms, software and hardware architecture, real-time data processing and management, machine learning and knowledge-based reasoning, to detect in the space-time dimensions people living in the real world starting from tsunami of visual data, acquired by networks of static and moving cameras, recognizing their presence also in cluttered and crowded environment, extracting information about their aspect, motion, action and interaction and eventually behavior (Rita Cucchiara, 2010).

Indeed, visual processing of people, including detection, tracking, recognition, and behavior interpretation, is a key component of intelligent video surveillance systems. A number of surveillance applications require the detection and tracking of people to ensure security and safety (Martínez-Tomás et al., 2013). That is, many video surveillance systems require the ability to determine if an image region contains people. This is none but a specific case of object classification in which there are only two object classes: person and non-person. Object classification in general is difficult and people detection is even harder. Also, video-surveillance systems must run at video-rate and thus require a trade-off between precision and computing time. Moreover, any people detection method

highly depends on segmentation, which remains a primitive problem.

In computer vision, multisensor image fusion is the process of combining relevant information from two or more images into a single image (or subimage). The resulting image (or subimage) will be more informative than any of the input images. Most common advantages of image fusion are improvement of reliability ((by redundant information) and improvement of capability (by complementary information). Image fusion is closely linked to the multisensory interpretation of human activities, where the very human detection is key. More generally, multisensor interpretation combines multiple information sources. These are presented by several sensors to generate a more accurate and solid interpretation of the environment (Pavón et al., 2007). Behaviors' and situations' interpretation is currently based in the use of a series of sensors that present a high ratio of false alarms.

The availability of new kinds of sensors in monitoring tasks allows dealing with new challenges in the field of multisensor data fusion (e.g. (Navarro et al., 2013)). Nowadays it is possible to build environment models and to diagnose situations from the analysis of undefined sequences of sensory information provided by several kinds of sensors. Multisensor data fusion - known as data fusion from several identical sensors or, complementarily, as data fusion from different kinds of sensors - is, from the efficiency perspective, a key task in advanced monitoring. Thus, the research and development in the multisensor field has experimented a fast growth. Multisensor systems include vision, audio, heat (thermal), appearance (volumetric), or vibration sensors, among others

Multisensory monitoring systems require of three important components (Zhu and Huang, 2007):

1. sensors to capture the information to be processed;
2. data fusion algorithms to process information captured by the sensors;
3. architectures to model and build real-time systems.

The goal of this kind of multisensor interpretation of situations and behaviors (particularly, human ones) is:

1. to monitor the environment, by using all features from the different kinds of sensors, and, moreover, in a cooperative/collaborative way among the different system's sensors,
2. to implement and optimize algorithms for the segmentation of the elements of interest, and also the labeling and tracking of objects to grant the system real-time processing ability,
3. to diagnose behaviors (spatio-temporal relationships among different elements of interest) starting from each sensor information, as well as situations obtained by means of the system sensors fusion, and,
4. to act in an intelligent manner on the scenario according to the previous diagnosis.

This dissertation addresses precisely robust human detection in video surveillance by fusing color and infrared video as part of a larger system of interpretation of situations and human behavior in complex environments (Fernández-Caballero et al., 2013; Gascueña et al., 2013).

1.2. Context and Motivation

This thesis is framed in the context of the research carried out during the last four years within the research group *natural and artificial Interaction Systems* (n&aIS) belonging to the *Laboratory of User Interaction and Software Engineering* (LoUISE) at the *Universidad de Castilla-La Mancha*. For over 20 years the (n&aIS) group is performing a series of research lines related to the computer vision in the *Instituto de Investigación en Informática de Albacete* (i3A). Lately, the group is developing algorithms and methodologies dedicated to robust people detection for video surveillance applications.

Part of this thesis has been funded by the Spanish national R&D project called “Multisensor data fusion in complex and dynamic environments: Bio-inspired methods, intelligent agent-based architectures, and multimodal and augmented interfaces” (TIN2010-20845-C03-01), part of the coordinated project “INT3 - Multisensor INTerpretation of behaviors and situations for an INTelligent INTervention in complex and dynamic environments”. The coordinated project, continuation of projects TIN2007-67586, TIN2004-07661 and TEC2008- 02077/TEC, deepens in the task of environment understanding, with the remarkable newness of the introduction of multisensor data fusion. The primary objective is to support the human operator in objective and consistent decision making in complex and dynamic environments. The project includes the detection and interpretation of predefined events capable of sending a warning of anomalous behavior or situation.

We can say that today human detection continues to be found especially in vogue within the field of computer vision. This is due to several reasons, among which we can cite the rise of surveillance applications. In some cases it is necessary to detect intrusions in security perimeters (Iwata et al., 2006; Kuno et al., 1996). It is also necessary to distinguish humans in the case of vehicle navigation systems (Armingol et al., 2007), especially for night driving (Xu et al., 2005). Another reason why people detection is becoming increasingly important is the increasing popularity of robots that need to perform object recognition and obstacle avoidance in order to avoid them in their navigation, and even alert a vigilant in case of an intrusion, being even necessary to perceive people in the scenario for the purpose of interaction (Gascueña and Fernández-Caballero, 2011; Nourbakhsh et al., 1999) or simply for secure navigation (Fernández-Caballero et al., 2010).

A widespread approach for detecting people is the use of color information (Capellades et al., 2003; Rodriguez and Shah, 2007; Schwartz et al., 2009). These are usually problematic when facing changes in lighting in a scene or visibility problems therein. To guard against these failures, you can find an alternative in the use of the infrared spectrum (Sun and Park, 2001; Kumar et al., 2006; Li et al., 2010). Even so, there remains some problems to solve at the time of performing an accurate segmentation of humans, as for instance the appearance of halos that decrease accuracy when defining the outlines of the silhouettes or problems that arise in situations where the temperature of objects and

Figure 1.1: Common problems of thermal cameras. (a) Difficulty to distinguish the human in certain circumstances (b) Appearance of halos

persons present in the scene are quite homogeneous. Examples related to this problems are shown in Figure 1.1a, where a human standing in the left side of the lateral area of the building can not be easily distinguished, whereas in Figure 1.1b the halo that surrounds humans prevents from isolating them correctly by using the contrast with the background.

In order to combine the strengths of both approaches while minimizing the flaws inherent to each one, in this thesis we have chosen to make a fusion between human segmentation in the thermal-infrared and the color spectrum, in order to achieve more robust detection of humans. It is evident that the idea of using color and infrared video fusion arises because both the visible and the infrared spectrum have different strengths and weaknesses when performing human detection. The advantages and disadvantages are also complementary in both spectra. On the one hand, although the information obtained from an infrared camera is useful for detecting humans in nocturnal environments, it presents several problems in other environments (Goubet et al., 2006). This is the case for hot or thermally homogeneous environments. On the other hand, the color yields good results when conducting human detection in well lit environments, but it is problematic in dark environments or in areas of the scene that present shadows or have low visibility in general. Therefore, the most recommended approach when performing a human detection is to use the support of learning or fusion techniques (Goubet et al., 2006; Kumar et al., 2006; Wang and Terman, 1997). This thesis focuses on the latter.

There is no doubt that segmentation and tracking are two essential concepts when addressing human detection. Image segmentation consists in partitioning an image into its constituent parts or objects (Gonzalez and Woods, 2007). On the other hand, image tracking can be defined as the process of consistently locating a desired feature in each frame of an input sequence (Sim, 1998). Obviously, this is a process dependent on the results achieved by the segmentation. An example of tracking is shown in Figure 1.2.

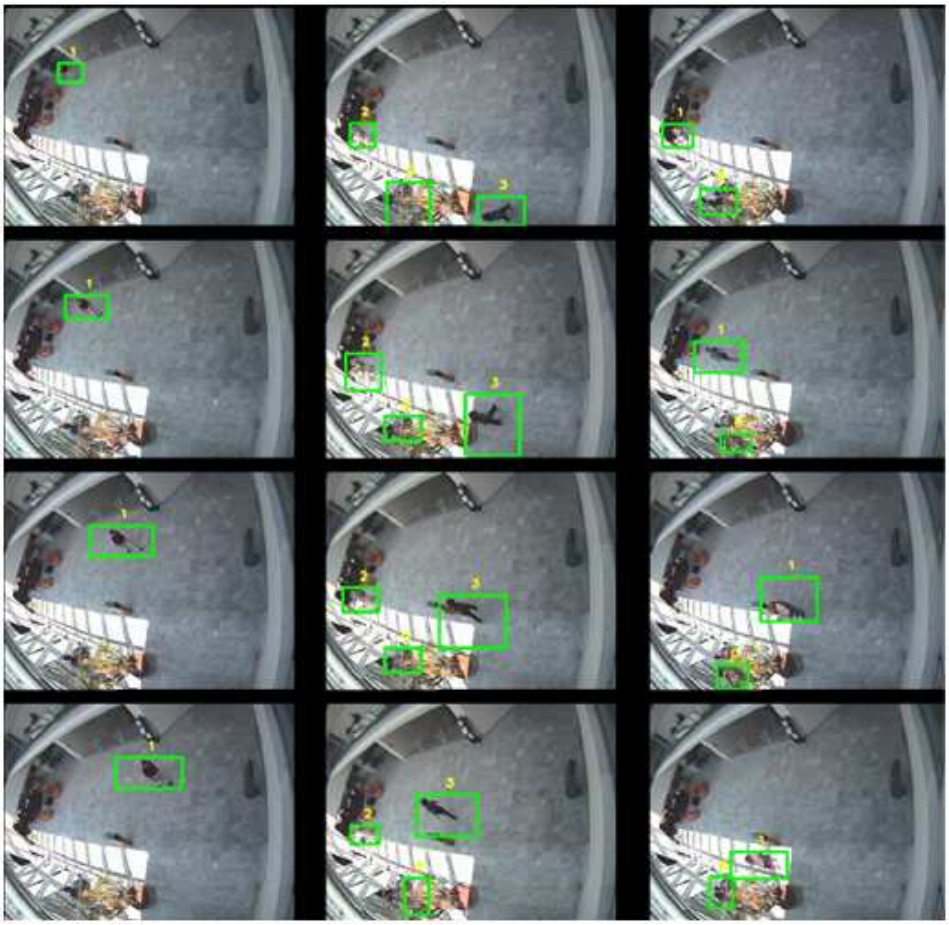


Figure 1.2: An example of tracking through several image frames.

1.3. Objectives

The major objective addressed in this thesis is the “design and implementation of a robust people detection system based on fusing the information provided after human segmentation in infrared and color spectra”. This objective has been divided, in turn, in the following sub-objectives:

1. Study of algorithms for motion-based object segmentation and tracking widely used in the literature, specially focusing on those centered on human detection, facing their future evaluation depending on the monitored scene.

2. Study of current image fusion techniques, specially focusing on those based on video captured taken by infrared and color cameras, and then proposal of a rule-based fusion mechanism obtained from experimentation in color and infrared spectra under different environmental conditions.

3. Proposal of a method for validating the results of human detection, and analysis of the results obtained in a real environment in order to validate the correct operation of the developed system.

1.4. Dissertation Outline

This PhD dissertation is structured as follows:

- Chapter 1 provides is a brief introduction explaining the main concepts to be covered, the problems to be addressed and the main objectives to be achieved. First, it presents an overview of the problem at hand, that is robust people detection through multispectra image fusion. After explaining the motivation for addressing the problem at hand, the chapter details the objectives to be achieved and finally arises the structure of the thesis.

- Chapter 2 is a study of the major phases covering a video-based human detection system, specifically segmentation, information fusion and tracking. For each of these areas an overview is provided. Then each stage is decomposed into various substeps, by studying and analyzing the latest trends emerged in each of the substeps. Then, we pass to particularize the study in each phase to the people detection problem. Finally, for each of these phases a classification of the various techniques studied in order to select those that best suit the objectives of the thesis are also performed.

- Chapter 3 introduces a comprehensive proposal to carry out a robust human detection system. Initially, we describe the work environment. Then, the development levels of the chosen framework are detailed, conducting a general description of the purpose of each along with their inputs and outputs. Finally, the evaluation metrics that will be used to validate the results achieved in each of the chosen levels are explained.

- In chapter 4 a detailed description of the elaborated processing levels takes place. We start from the image acquisition level, in addition explaining the synchronization mechanism chosen. Then the human segmentation algorithms developed both for infrared and color images captured in the previous level are described. Subsequently, the process of fusing the results obtained by the segmentation in each spectrum, and the tracking carried out using as a basis the said fusion, are explained in detail.

- Chapter 5 provides an assessment of the developed people detection system. First, a description is made of the chosen environment, also describing each of the selected sequences. Subsequently, all the people segmentation algorithms made in each of the spectra are compared with each other, in order to choose in each spectrum the most appropriate to the environment in which the tests are taking place. Next, we analyze the results obtained in each sequence after applying the developed fusion and tracking algorithm using as input the results of the chosen segmentation algorithms. These results are compared with those of the proper segmentation

algorithms to see if there is an improvement after the fusion process, thereby validating the correct operation of the system developed.

- Chapter 6 embodies the conclusions reached during the development of the thesis, focusing on the results achieved and the knowledge obtained and provided. Also some ideas are added about possible work that could start from this thesis. Finally, the publications arisen from the development of this thesis and its related research projects are detailed.
- The literature used during the preparation of the thesis is also included in this memory.

Capítulo 2

Detección de humanos en vídeo

En el presente capítulo se presenta una visión general del estado del arte en los niveles de un sistema de visión artificial que cubre la presente tesis, es decir, segmentación de humanos en imágenes en infrarrojo y color, fusión de los resultados obtenidos en cada segmentación en un espacio común, y seguimiento de los humanos detectados sobre dicho espacio, correspondiéndose cada paso con una sección del capítulo. Se concluye finalmente con un breve resumen de las conclusiones generales extraídas.

2.1. Segmentación de humanos en vídeo

El primer paso del estudio necesario para realizar el sistema que se desarrolla en esta tesis se corresponde con la segmentación de humanos en color e infrarrojo. La presente sección está centrada en dicho aspecto de la visión artificial, comenzando con una visión general de los principales conceptos asociados a la segmentación y que, por tanto, entran más en juego a la hora de explicar los diversos algoritmos encontrados en la literatura. Posteriormente, se particulariza sobre la segmentación en color e infrarrojo, para finalmente plasmar las principales conclusiones extraídas de la lectura de los diversos artículos que describen los algoritmos estudiados centrándose finalmente en el campo particular de la detección de humanos.

2.1.1. Definición de segmentación

Tal y como se ha enunciado en el Capítulo 1, la segmentación de imágenes consiste en particionar una imagen en las partes u objetos que la constituyen (Gonzalez and Woods, 2007). El objetivo es la localización de áreas significativas de la imagen, tales como imperfecciones en una herramienta, áreas urbanas en el caso de que se esté trabajando sobre un mapa, intrusos en el caso de un sistema de vigilancia de un entorno, etc

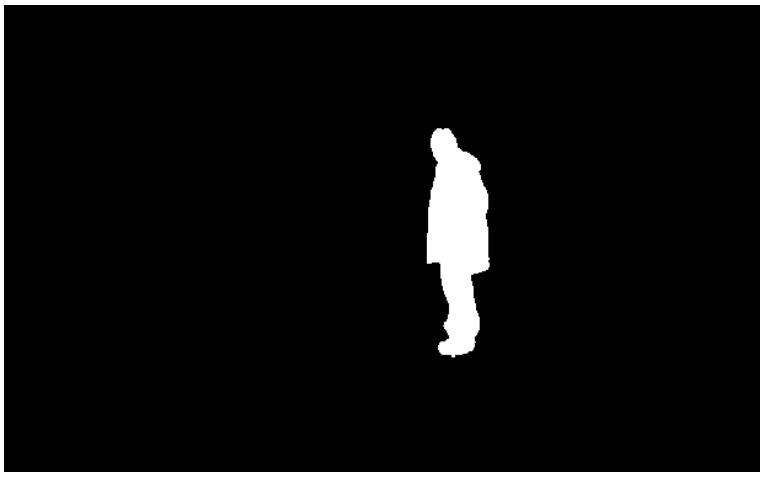


Figura 2.1: Imagen dividida en dos regiones: Humano y fondo

.

Con el fin de explicar más claramente el concepto de segmentación, se incide en la segmentación orientada a regiones, en la que la imagen se divide en n regiones. Tomando R como la representación de la región completa de la imagen que contiene a todas las demás (tal y como se puede ver en la ecuación (2.1)), se puede entender la segmentación orientada a regiones de la siguiente forma:

$$\bigcup_{i=1}^{n} R_i = R \tag{2.1}$$

donde cada R_i es una región $i = 1, 2, .., n$ en la que cada uno de sus puntos se encuentra conectado con otro. Por su parte, estas regiones son disjuntas entre sí, tal y como se aprecia en la ecuación (2.2).

$$R_i \bigcap R_j = \emptyset, \forall i, j, |i, j \in (1, .., n), i \neq j \tag{2.2}$$

Sea $P(R_i)$ una determinada propiedad que deben satisfacer los puntos pertenecientes a una región (por ejemplo, un determinado nivel de gris) tenemos:

$$P(R_i) = CIERTO, \forall i = 1, 2, .., n. \tag{2.3}$$

$$P(R_i \bigcup R_j) = FALSO, \forall i \neq j \tag{2.4}$$

donde la ecuación (2.3) nos indica que todos los puntos pertenecientes a una región cumplen una determinada propiedad mientras que la ecuación (2.4) establece que dicha propiedad no es válida entre puntos pertenecientes a distintas regiones. Por ejemplo, en la imagen de la Figura 2.1 podemos ver como la imagen original se ha dividido en dos regiones, de forma que el humano aparece en blanco y el fondo en negro.

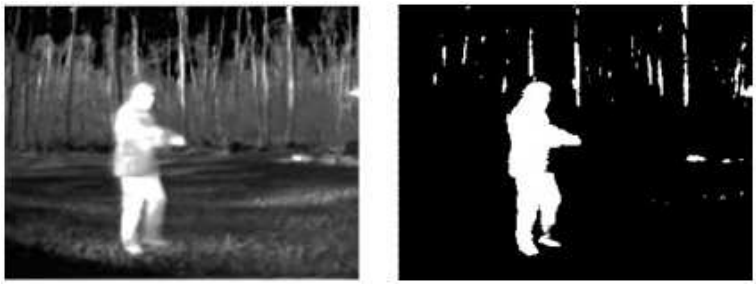


Figura 2.2: Imagen térmica infrarroja y binarización de la misma.

2.1.1.1. Conceptos básicos sobre segmentación

Generalmente los algoritmos de segmentación se basan en el estudio de dos propiedades de las imágenes: la similitud y la discontinuidad (Gonzalez and Woods, 2007).

La *similitud* se basa en la suposición de que los píxeles pertenecientes a un mismo objeto deben presentar una apariencia homogénea. Sin embargo, en la realidad es difícil que todos los píxeles dentro de un mismo objeto presenten propiedades uniformes, ya que influyen factores tales como la iluminación, el ruido de la imagen, los reflejos que pueda proyectar el objeto en caso de encontrarse sobre una superficie especular, etc. A pesar de estos factores que la dificultan, en base a la propiedad de la similitud se pueden utilizar técnicas de umbralización, las cuales destacan las zonas de la imagen que cumplen una determinada propiedad (normalmente las que se encuentran dentro de un determinado rango de brillo), descartando el resto. Un ejemplo de imagen umbralizada lo podemos encontrar en la Figura 2.2, donde observamos que el proceso de binarización obtenido en la imagen de la derecha no es siempre suficiente para realizar una segmentación válida de la imagen de la izquierda, sino que en muchas ocasiones se hace necesario eliminar posteriormente los elementos conectados de la imagen que no sean de interés. Existen numerosas formas de determinar el umbral óptimo a aplicar a una imagen o secuencia de imágenes en base a los objetos de interés buscados, si bien un método muy utilizado es el de Otsu (1979). Este algoritmo divide los píxeles de la imagen en dos clases, de las cuales una contendrá los objetos de interés y la otra el resto (lo que denominaremos *fondo* en el resto de esta memoria). El objetivo es encontrar el umbral óptimo para separar estas dos clases, de manera que la varianza dentro de cada clase sea mínima, es decir, que los píxeles dentro de cada clase sean tan similares entre sí como sea posible. Si bien el método es fácilmente extensible hasta la obtención de tres o más clases de regiones, a partir de una cantidad superior a tres el autor considera que las expresiones para obtener los umbrales se vuelven demasiado complejas, repercutiendo en la pérdida de capacidad de distinción. Alternativamente, a la hora de escoger umbrales se han planteado diversas técnicas, tales como la de Sun and Park (2001), que usa técnicas de lógica difusa a la hora de establecer un umbral óptimo, combinando sus resultados con los de la detección de bordes mediante el algoritmo de Canny (1986), que se explicará a continuación ya que hace uso de la siguiente propiedad.

Por su parte, la *discontinuidad* se basa en buscar cambios de contraste en la imagen. Estas zonas

de contraste pueden corresponderse a los límites de los objetos de interés, ayudando así a poder delimitarlos y distinguirlos respecto del fondo. Existen numerosos algoritmos de detección de bordes basados en discontinuidades que permiten trazar los contornos de los objetos que forman la imagen. En estos métodos un píxel suele ser considerado como borde cuando presenta niveles de gris diferentes a los de sus vecinos. Los algoritmos basados en esta propiedad suelen utilizarse para refinar los resultados de métodos previamente aplicados, con el fin de delimitar mejor los objetos de interés. Uno de los principales algoritmos de detección de bordes es el de Canny (1986), que se basa en la definición de un conjunto de objetivos para computación de los puntos pertenecientes a los bordes. Estos objetivos consisten en una serie de criterios de detección y localización de dichos puntos. Estos criterios se basan en una buena detección (debe delimitar el máximo número de bordes posibles), una buena localización (los bordes deben ser marcados lo más cerca posible a la localización real de los mismos) y una única respuesta para cada borde (un borde en la imagen sólo debe marcarse una vez y el ruido no debe influir, dando lugar a la detección de falsos bordes).

A la hora de realizar una detección de bordes, la técnica de detección del *gradiente* es una de las más utilizadas. Cuando se trabaja en una sola dimensión, un borde está asociado a un máximo local en la primera derivada. El gradiente es una medida de las variaciones en los valores de una función, pudiendo considerarse una imagen como un vector de muestras de una función de intensidad continua (Jain et al., 1995). Análogamente, los cambios significativos en los niveles de gris de una imagen pueden detectarse usando una aproximación discreta al gradiente, que es el equivalente bidimensional de la primera derivada, definiéndose como el vector indicado en la ecuación (2.5).

$$G[f(x,y))] = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{2.5}$$

Para calcular el gradiente se suelen aplicar el operador de Sobel o el de Prewitt, los cuales realizan aproximaciones al gradiente en una vecindad 3×3. Las máscaras de convolución de Sobel pueden verse en las ecuaciones (2.6) y (2.7), mientras que las de Prewitt se pueden observar en las ecuaciones (2.8) y (2.9) respectivamente.

$$s_x = \begin{array}{|c|c|c|} \hline -1 & 0 & 1 \\ \hline -2 & 0 & 2 \\ \hline -1 & 0 & 1 \\ \hline \end{array} \tag{2.6}$$

$$s_y = \begin{array}{|c|c|c|} \hline 1 & 2 & 1 \\ \hline 0 & 0 & 0 \\ \hline -1 & -2 & -1 \\ \hline \end{array} \tag{2.7}$$

$$p_x = \begin{array}{|c|c|c|} \hline -1 & 0 & 1 \\ \hline -1 & 0 & 1 \\ \hline -1 & 0 & 1 \\ \hline \end{array} \tag{2.8}$$

$$p_y = \begin{array}{|r|r|r|} \hline 1 & 1 & 1 \\ \hline 0 & 0 & 0 \\ \hline -1 & -1 & -1 \\ \hline \end{array} \tag{2.9}$$

2.1.2. Fases en la segmentación

Antes de explicar los diferentes algoritmos de segmentación, debemos delimitar claramente las fases en las que se puede dividir este proceso. Son la extracción de características, el filtrado de objetos de interés y la clasificación.

En una *extracción de características* inicial, se extraen los posibles puntos de interés o regiones de la imagen que pueden dar una primera visión aproximada de los objetos que la componen. Una vez que tenemos esta visión preliminar es necesario refinar los resultados, para lo cual se realiza un *filtrado de objetos de interés* en base a diversos criterios, como pueden ser su composición, su distribución interna, etc. Finamente, aquellos objetos que no han sido descartados por este proceso son sometidos a un proceso de *clasificación*, que se encarga de establecer a qué categoría pertenece cada objeto de interés, pudiendo descartarlos en caso de que no se correspondan al tipo de objeto que se pretende detectar en la aplicación en la que se está trabajando.

A continuación, profundizaremos más a fondo en cada una de estas fases.

2.1.2.1. Extracción de características

El primer paso en un algoritmo de segmentación consiste en la extracción inicial de características de la imagen. Por característica podemos entender cualquier rasgo distintivo que nos permita realizar una aproximación a los puntos de interés. Según Gonzalez and Woods (2007), las principales características que se pueden extraer de una imagen son, de menor a mayor complejidad, puntos, líneas y bordes, a los que podemos añadir un nivel superior consistente en regiones.

Dependiendo de la metodología utilizada para la segmentación, podemos encuadrar la mayoría de los diversos algoritmos de segmentación en técnicas basadas en detección de movimiento, propuestas estadísticas y probabilísticas, y técnicas basadas en crecimiento de regiones y en grafos, si bien existen otras aproximaciones a la segmentación que se alejan de esta clasificación o combinan varias técnicas de las citadas.

Detección de movimiento

Las *técnicas basadas en detección de movimiento* se fundamentan en la noción de que las diferencias entre imágenes consecutivas se deben normalmente a movimientos llevados a cabo por objetos de interés, lo que significa que, encontrando dónde se ha producido el movimiento, se podrá encontrar al objeto que lo ha realizado. Entre otras formas, el movimiento se puede caracterizar mediante el uso del flujo óptico o de la diferencia de imágenes. Los primeros trabajos de detección de movimiento

estaban basados en diferencia de fotogramas consecutivos (Jain and Nagel, 1979), siendo el algoritmo más sencillo y directo la simple umbralización de la imagen diferencia. El procedimiento estándar a seguir pasa por restar individualmente los píxeles de cada imagen y binarizar posteriormente el resultado, representando como valor máximo en la imagen resultado aquellos que en la imagen obtenida de la resta sobrepasan un determinado umbral T, tal y como se puede apreciar en la ecuación (2.10) para los píxeles de las imágenes i y j consecutivas en el tiempo, donde (x,y) son las coordenadas de cada píxel y t_i y t_j los instantes actual y anterior respectivamente.

$$d_{ij}(x,y) = \begin{cases} \text{MAX si } |f(x,y,t_i) - f(x,y,t_j)| > T, \\ \text{MIN, en otro caso} \end{cases} \tag{2.10}$$

La elección del umbral T necesario depende en gran medida de la secuencia, su ruido y su movimiento, no habiendo ningún motivo para que este valor sea constante en toda la imagen, ya que objetos y tipos de movimiento distintos suelen presentar valores diferentes de intensidad. Se han desarrollado muchos métodos para decidir si un píxel se ha movido o no, pudiendo realizarse la decisión de forma independiente para cada píxel (Konrad, 2000) o bien para pequeños bloques de píxeles (Fablet et al., 1999). Con detecciones de píxel independientes, los mapas de detección resultantes están normalmente *corruptos* a causa de huecos en la máscara de objetos en movimiento (es decir, en el conjunto de píxeles que se han detectado como objetos en movimiento en la imagen) y falsos positivos (que se definen como detecciones de objetos de interés que realmente no se encuentran en esas coordenadas de la imagen) debidos al ruido. Estos errores pueden ser atenuados usando restricciones de regularización e información adicional.

Los métodos basados en la diferencia de fotogramas consecutivos son muy sensibles al ruido y a cambios de iluminación (por ejemplo, (Fernández-Caballero et al., 2001, 2003; López et al., 2006)), por lo que la resta de fondo constituye una solución apropiada cuando hay que procesar un gran número de fotogramas donde aparecen pocos cambios entre fotogramas consecutivos. Esta técnica, ampliamente utilizada en la literatura, tal y como veremos a lo largo del presente capítulo, consiste en elaborar un modelo del fondo de la escena, entendiendo como fondo la imagen captada por la cámara sin humanos presentes en dicha imagen. Una vez modelado el fondo, se aplica la resta de imágenes descrita entre el último fotograma capturado y el fondo previamente modelado, de forma que los objetos que no estuvieran originalmente presentes en la escena aparecerán como objetos de interés. Esto presenta un problema, y es que el fondo debe ser actualizado periódicamente, ya que puede suceder que el fondo de la escena varíe debido a objetos depositados en la misma (los cuales no interesa captar si se están buscando humanos) o simplemente cambios de iluminación debido al paso del tiempo. Los métodos de modelado de fondo se pueden clasificar a su vez como predictivos y no predictivos. Los no predictivos construyen una función de probabilidad de densidad para la intensidad por cada píxel individual, pudiendo ser representada la distribución estadística de un píxel mediante una distribución gaussiana (Kanade et al., 1998; Cavallaro and Ebrahimi, 2001; Huwer and Niemann, 2000; Xiao et al., 2013). De esta forma se determinan los píxeles de interés como aquellos para los que el valor de intensidad se aleja de la media del modelo de fondo, con lo que un agrupado de

píxeles permite la detección de objetos de interés. Por otro lado, los métodos predictivos usan un modelo dinámico para predecir la intensidad de un píxel a partir de las observaciones previas. Debido a que esta técnica requiere un escenario con características sujetas a pocas variaciones, normalmente se usa en aplicaciones de vigilancia en las que la cámara se encuentra siempre en una posición fija. Incluso realizando las mejoras nombradas, los métodos de resta de fondo y umbralización son un paso preliminar para la detección de objetos en movimiento haciéndose necesario un procesado posterior para obtener de forma más exacta las regiones donde se encuentran dichos objetos.

Otra alternativa muy utilizada en la literatura la constituyen los métodos basados en el flujo óptico. Éste consiste en obtener la distribución de las velocidades aparentes del movimiento de los patrones de brillo entre dos imágenes (Gibson, 1950; Horn and Schunck, 1981; Lucas and Kanade, 1981; Tomasi and Kanade, 1992). Su idea principal se basa en que cuando el observador se encuentra en movimiento todos los objetos parecen aproximarse a él o alejarse de él desde su punto de vista. Por tanto, para distinguir los objetos que se encuentran realmente en movimiento, bastará con discriminar aquellos que presentan características de movimiento distintas a las mayoritarias en la escena. Estas técnicas son recomendables para la segmentación desde cámaras en movimiento, tal y como puede suceder si se desea obtener los peatones presentes en imágenes tomadas desde un vehículo en tránsito. Por ello, en la literatura podemos encontrar algoritmos como el de Talukder and Matthies (2004), donde se usa esta metodología para encontrar objetos en movimiento desde un vehículo en marcha, o el de Klappstein et al. (2009), en el que se describe la detección de objetos mediante flujo óptico tanto en imágenes monoculares como estereoscópicas.

Propuestas estadísticas y probabilísticas

Las *propuestas estadísticas y probabilísticas* añaden a la segmentación conceptos propios de minería de datos y de estadística, incorporando una capa superior que sirve para ir actualizando los resultados. Para realizar este proceso, se basan en la coherencia temporal de las secuencias de imágenes, recabando información estadística. Estos algoritmos utilizan en algunas ocasiones conceptos de inteligencia artificial agregando una capa de aprendizaje en estos casos.

Entre los métodos estadísticos, basándose en las características de los datos de la imagen, podemos citar Kass et al. (1988), donde se propuso el concepto de *serpiente*, que consiste en un modelo de contorno activo. Este modelo se basa en una curva definida en la que se intenta minimizar la energía asociada a ella. Esta energía se define como la suma de las fuerzas externas (las cuales son mínimas cuando la serpiente se encuentra en los límites de un objeto, y cuyo enfoque más común consiste en alcanzar valores mínimos cuando el gradiente en torno al objeto alcanza su valor más alto) e internas (que se basan en criterios tales como la longitud de la serpiente o su curvatura). Esto significa que las serpientes se van deformando progresivamente de acuerdo al contorno más cercano, localizándolo de forma precisa, aunque el comportamiento de las serpientes depende de la definición de la función que se escoja para definir tanto las fuerzas externas como las internas. Estos métodos han tenido éxito realizando tareas como la detección de bordes, la detección de esquinas, el seguimiento de movimiento y el *matching* estéreo.

Crecimiento de regiones

Las *propuestas basadas en crecimiento de regiones* se basan en la idea de que la segmentación se basa en dividir una imagen en un conjunto de regiones, constituyendo una alternativa al uso de umbrales. En base a ello surgen las técnicas de crecimiento de regiones, donde de forma iterativa a través de sucesivas pasadas los píxeles pertenecientes a un mismo objeto se van uniendo en base a características comunes, especialmente en cuanto a la intensidad de su nivel de gris.

Por ejemplo, en Beucher (1991) se utiliza la transformada watershed ("cuencas de agua") para la segmentación de la imagen. Esta transformada se construye implementando un proceso de inundación en una imagen en niveles de gris. Considerando dicha imagen como una superficie topográfica y definiendo las cuencas y las líneas divisorias de agua de la misma, se imagina que se perfora cada mínimo de la superficie topográfica y que se sumerge esta superficie en un lago con una velocidad vertical constante, de forma que el agua que entra por los agujeros inunda la superficie. Durante esta inundación, dos o más caudales viniendo de distintos mínimos se pueden fusionar. Para evitar esta fusión, se construye un dique en los puntos de la superficie donde puede ocurrir. Al final del proceso, los diques sobresalen por encima del agua solamente. Estos diques definen las líneas de agua de la imagen, separando las diversas cuencas, cada una de las cuales contiene un único mínimo. Para aplicar este algoritmo a la segmentación de imágenes, se utiliza la imagen de gradiente, que se representa topográficamente de forma similar a un volcán. En la nueva imagen, cada punto de la original se transforma en un mínimo dentro de una región rodeado de una cadena cerrada de montañas, de forma similar a una cuenca. La altitud variante de la cadena de montañas expresa la variación de contraste a lo largo del contorno del punto original. Se realiza una aplicación jerárquica de este algoritmo en la que, tras calcular las líneas de agua de la imagen del gradiente y etiquetar cada cuenca con el valor mínimo de gris de la imagen original en esa región, se fijan los criterios para extraer los objetos deseados de la imagen. En la Figura 2.3 podemos observar cómo los lagos (correspondientes a los puntos de la imagen de bajo contraste) se van inundando hasta alcanzar un dique que impedirá que se fusionen. Si, por ejemplo, eliminamos el dique 2, podemos ver como los lagos 1 y 2 se fusionarían, lo que en una imagen significaría que se agruparían en un único objeto de interés. Otra aplicación de la técnica de cuencas de agua la podemos encontrar en Navon et al. (2005), donde también se divide la imagen en regiones, siendo dichas regiones homogéneas en esta ocasión, obtenidas mediante umbrales locales.

Por otra parte, en Kim and Kim (2003) se presenta un método basado en una aplicación multiresolución de la transformada wavelet ("cordillera"), que descompone la imagen en diversas subimágenes de diferentes niveles de detalle que explicaremos más tarde en esta tesis, y en el algoritmo watershed. El proceso consiste en realizar, en primer lugar, una representación piramidal, creando imágenes de diversas resoluciones mediante el uso de la transformada wavelet. A continuación, se segmenta la imagen de menor resolución obtenida mediante el algoritmo watershed, para a continuación unificar la región segmentada en las diversas imágenes con mayor resolución. Finalmente, la imagen de baja resolución se proyecta en la imagen original utilizando la transformada inversa.

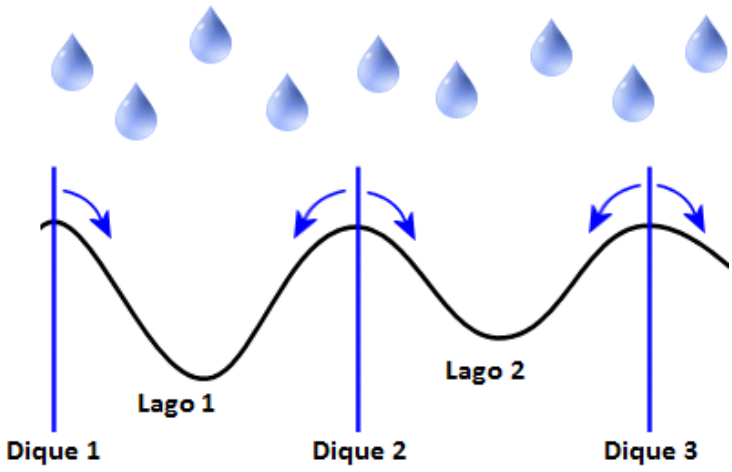


Figura 2.3: Esquema del algoritmo watershed, extraído de Image Metrology (2013)

Propuestas basadas en grafos

Las *propuestas basadas en grafos* se basan en la teoría de grafos, utilizando técnicas de este campo para interpretar la imagen como un grafo estableciendo sus arcos y nodos en base a diversos criterios y utilizando técnicas basadas en grafos tales como encontrar el grafo de recubrimiento mínimo. Muchas de estas técnicas (Shi and Malik, 1997; Birchfield et al., 2007; Bugeau and Pérez, 2008; Kohli et al., 2008; Li and Yang, 2008) utilizan el concepto de *corte*, basado en que a la hora de dividir un grafo en dos grafos disjuntos, se hace necesario eliminar las aristas que conectan ambos grafos. Se define el corte como el peso de las aristas que han debido ser eliminadas, siendo el bi-particionamiento óptimo de un grafo aquél que minimiza el valor del corte. Entre los ejemplos que hemos citado de este enfoque podemos destacar el de Shi and Malik (1997), cuya propuesta se centra en resolver problemas de percepción de grupos. En vez de basarse en características locales y sus consistencias en los datos de la imagen, se intenta extraer la impresión global de la imagen, tratando la segmentación de la misma como un problema de particionamiento de grafos. Una aproximación reciente presenta como novedad el uso de un algoritmo de seguimiento rápido de objetos con una segmentación basada en corte de grafos en un marco de trabajo jerárquico (Hu et al., 2013). También encontramos un algoritmo basado en corte aplicado en sucesivos frames (Collomosse and Wang, 2012).

Otras propuestas

Entre *otras propuestas*, podemos destacar el algoritmo SIFT (transformación de características invariante a la escala, en inglés Scale Invariant Feature Transform, SIFT), que fue propuesto por Lowe (1999) y está considerado como uno de los mejores algoritmos para extraer puntos característicos. Estos puntos son muy distintivos, en el sentido de que una sola característica puede ser correspondida muy probablemente con otra que esté dentro de una base de datos amplia con características de muchas imágenes. Los principales pasos del algoritmo son los siguientes (Xiong and Zhang, 2010):

1. *Detección de extremos espaciales y de escala*: La primera etapa busca extremos en todas las escalas y localizaciones de la imagen. Se implementa de forma eficiente usando una función de diferencia de gaussianos para identificar puntos potenciales de interés que sean invariantes a la escala y a la orientación.

2. *Localización de puntos clave*: Se ajusta un modelo detallado para determinar la localización y escala de cada zona candidata. Los puntos clave se seleccionan en base a las medidas de su estabilidad. Aquellos con bajo contraste o con grandes diferencias entre sus autovalores máximo y mínimo son eliminados.

3. *Asignación de la orientación*: A cada punto clave se le asigna una o más orientaciones de acuerdo a las direcciones del gradiente local de la imagen. Todas las operaciones futuras se llevarán a cabo sobre los datos de la imagen transformados en base a la localización, escala y orientación asignados a cada característica, proporcionando invarianza respecto a estas transformaciones.

4. *Descriptor de puntos clave*: Los gradientes de imagen locales son medidos en la escala seleccionada en la región en torno a cada punto clave. Estos puntos son transformados a una representación que permite niveles significativos de distorsión de forma local y cambios en iluminación.

Basada en SIFT, otra tendencia muy utilizada consiste en el uso de histogramas de gradientes orientados (en inglés, Histograms of Oriented Gradients, HOGs). Su idea básica se fundamenta en que la apariencia y la forma locales de un objeto pueden caracterizarse en la mayoría de los casos mediante una distribución de los gradientes de intensidad locales o las direcciones de los bordes (Dalal and Triggs, 2005), sin ser necesario un conocimiento previo de las posiciones de los mismos. En la práctica, esto se implementa dividiendo la ventana de la imagen en pequeñas regiones espaciales ("celdas"), de forma que cada celda acumule un histograma local unidimensional de las direcciones del gradiente o las orientaciones de los bordes sobre los píxeles que pertenecen a la misma. La combinación de los diferentes histogramas es lo que acaba formando la representación. Para lograr una mayor invarianza a la iluminación, sombras, etc., también se suele normalizar previamente las respuestas locales antes de usarlas. Se puede lograr mediante la acumulación de la energía local del histograma sobre regiones espaciales mayores ("bloques") para posteriormente usar los resultados obtenidos para normalizar todas las celdas del bloque. A estos bloques descriptores normalizados es a lo que se denomina histograma de gradientes orientados, que se suele combinar con un clasificador a nivel de ventana basado en una máquina de vector soporte convencional, que explicaremos posteriormente en esta tesis. Igualmente parcialmente basado en SIFT, SURF (Speeded Up Robust Features) es un detector local de características, presentado por primera vez por Bay et al. (2008). Muy recientemente se ha propuesto un marco de trabajo para la detección en escenas basado en entropía y SURF (Baber et al., 2013).

2.1.2.2. Filtrado de objetos de interés

Una vez que tenemos una representación inicial de los rasgos que caracterizan a los objetos de interés deseados, se hace necesario realizar un filtrado de dichos objetos con el fin de obtener una aproximación más fina de los objetos en la escena que pueden ser considerados de interés para el sistema. En base a las características de estos objetos, podemos distinguir una serie de representaciones, que clasificaremos utilizando una adaptación de las que establece Yilmaz et al. (2006).

En las representaciones que utilizan las *densidades de probabilidad de la apariencia del objeto*, se utiliza una función de densidad de probabilidad (FDP), que se define como una función matemática que caracteriza el comportamiento probable de un conjunto de variables. Esta función $f(x)$ especifica la posibilidad relativa de que una variable aleatoria continua X tome un valor cercano a x, y se define como la probabilidad de que la variable aleatoria continua tome un valor dentro del intervalo $[x, x+dx]$ siendo dx un número infinitesimalmente pequeño. La mayoría de las funciones de densidad de probabilidad requieren uno o más parámetros para especificarlas totalmente. Las estimaciones de densidad de probabilidad de la apariencia del objeto pueden ser paramétricas, como, por ejemplo, gaussianas (Zhu and Yuille, 1996), o no paramétricas (Elgammal et al., 2002), entre las que podemos citar los histogramas (Comaniciu et al., 2003). Las características de las densidades de probabilidad de la apariencia del objeto pueden ser calculadas a partir de las regiones de la imagen especificadas por los modelos de forma, pudiendo comprender una región interior de una elipse o contorno.

Dentro de las paramétricas, una propuesta muy utilizada consiste en realizar una *mezcla de gaussianos* (en inglés, Mixture of Gaussians, MoG). En Friedman and Russell (1997) se explica este principio describiendo una de sus primeras aplicaciones. A la hora de segmentar el tráfico, se establecen diversas categorías: vehículo, sombra y carretera. Cada píxel se modela para cada una de estas categorías usando una distribución gaussiana, de forma que la composición de las tres puede dar una idea de la probabilidad de que el píxel pertenezca a cada una de estas categorías.

Otra aproximación consiste en usar *plantillas*, las cuales se forman usando formas geométricas simples o siluetas. Una ventaja de una plantilla es que lleva tanto información espacial como de apariencia. Como inconveniente podemos citar que las plantillas únicamente codifican la apariencia del objeto generada a partir de una sola vista. Por tanto, son exclusivamente apropiadas para el seguimiento de objetos cuyas poses no varían considerablemente durante el transcurso del seguimiento. Como ejemplo de estos métodos, podemos citar el algoritmo propuesto en Nanda and Davis (2002), que realiza un análisis basado en probabilidades. Para ello, se usan plantillas probabilísticas con el fin de capturar las distintas variaciones que puede presentar una forma humana, siendo estas plantillas especialmente útiles en los casos en que el contraste sea bajo y falten partes del cuerpo.

2.1.2.3. Clasificación de objetos

Una vez que se han extraído las principales características de la escena y se ha realizado un primer filtrado de los objetos de interés presentes en la misma, falta clasificar los resultados obtenidos. Dicha clasificación, que constituye el último paso en un algoritmo de segmentación, puede ser realizada de múltiples formas. Entre las diversas aproximaciones a la clasificación de objetos, destacaremos las técnicas basadas en aprendizaje, en las que Yu et al. (2011) establecen una división entre métodos basados en aprendizaje supervisado y basados en aprendizaje no supervisado.

Si se puede especificar un pequeño conjunto de píxeles de entrenamiento con los contornos trazados previamente de forma aproximada, la segmentación basada en aprendizaje supervisado constituye una buena elección, mientras que si el usuario desea poder realizar la segmentación de imágenes automáticamente sin intervención humana en ningún momento, es más recomendable usar los métodos de aprendizaje no supervisado, ya que estos últimos no requieren un entrenamiento inicial. A continuación se describen con mayor profundidad ambas aproximaciones.

Aprendizaje supervisado

Los algoritmos basados en *aprendizaje supervisado* seleccionan en primer lugar un pequeño conjunto de píxeles en diferentes regiones para que sirvan como conocimiento previo a la hora de entrenar un clasificador. El resto de píxeles se considera como el conjunto de pruebas, siendo particionados en varias regiones significativas una vez que el clasificador ha sido entrenado. Como ejemplo de esta metodología podemos citar las máquinas de soporte vectorial (en inglés, Support Vector Machines, SVM) (Cortes and Vapnik, 1995). El objetivo de las SVM es encontrar un hiperplano que maximice el margen geométrico y minimice el error de clasificación dando un problema de dos clases linealmente separables. El margen entre ellas se mide por las distancias entre los vectores de soporte y el hiperplano. Por ejemplo, en la Figura 2.4 se pueden ver varios hiperplanos que establecen la clasificación entre dos clases (los puntos rojos y los verdes) pudiéndose apreciar que el hiperplano óptimo es el (d), que se corresponde con la máxima distancia posible a las dos clases a las que divide.

En Cao et al. (2009) se propone una estrategia de reconocimiento de movimiento que representa cada vídeo como un conjunto de imágenes filtradas, cada una de las cuales corresponde a un fotograma. Usando un clasificador de imágenes filtradas basado en SVMs, se clasifica un vídeo aplicando votación por mayoría sobre las etiquetas predichas de sus imágenes filtradas, identificando el tipo más probable de acción en cada momento. Además se define una confianza de clasificación y el umbral asociado en ambos casos, lo que permite identificar la existencia de un tipo desconocido de movimiento cuando aparece alguno no clasificado anteriormente. Junto con la estrategia de reconocimiento propuesta, esto posibilita construir un sistema de reconocimiento que no solo puede hacer clasificaciones en tiempo real, sino que además es capaz de aprender nuevos tipos de movimiento y reconocerlos en el futuro.

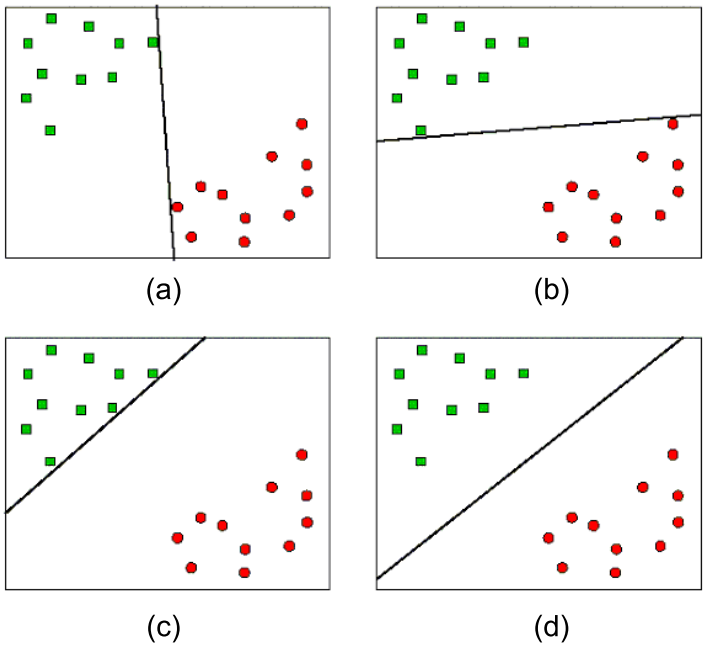


Figura 2.4: Ejemplo de diversos hiperplanos cuya misión es poder clasificar categorías en una SVM. El hiperplano (d) es el óptimo al proporcionar la mayor distancia entre las dos clases. Imagen extraída de Improved Outcome Software (2013)

Otra tendencia muy utilizada es el uso de *Adaptative Boosting* (Adaboost) (Freund and Schapire, 1995), consistente en hallar un clasificador muy preciso mediante la combinación de muchos clasificadores base, cada uno de los cuales puede ser solo moderadamente preciso. En la fase de entrenamiento del algoritmo Adaboost, el primer paso es construir una distribución inicial de pesos sobre el conjunto de entrenamiento. Entonces el mecanismo de estímulos selecciona un clasificador base que dé el mínimo error, donde el error es proporcional a los pesos de los datos mal clasificados. A continuación, se aumentan los pesos asociados con los datos mal clasificados por el clasificador base seleccionado. Así, el algoritmo favorece la selección de otro clasificador que rinda mejor con los datos mal clasificados en la siguiente iteración. En el contexto de la detección de objetos, los clasificadores débiles pueden ser operadores simples tales como un conjunto de umbrales, aplicados a las características del objeto extraídas de la imagen. Esta técnica presenta numerosas aplicaciones, como detección de caras (Kage et al., 2007; Vincenzo and Lisa, 2007) o la clasificación de objetos en tráfico (Zhang et al., 2007).

Aprendizaje no supervisado

Por otra parte, los métodos basados en *aprendizaje no supervisado* estiman en primer lugar el número óptimo de regiones (K) en la imagen mediante un índice de validez de región. Este índice incluye índices internos (tales como el índice de Dunn (Dunn, 1974), o el de Davies-Bouldin (Davies and Bouldin, 1979)) y criterios basados en información, como por ejemplo el criterio de información de Akaike (en inglés, Akaike Information Criterion, AIC) (Akaike, 1974), el cuál compara modelos estadísticos entre sí a partir de un conjunto de datos, con el fin de establecer qué modelo se ajusta

mejor a la información proporcionada. Por su parte, el criterio de longitud de descripción mínima (en inglés, Minimum Description Length, MDL) (Rissanen, 1978) busca la descripción mínima de un conjunto de datos que permita estimar en tiempo real tanto los parámetros estructurales como los del sistema. Finalmente, podemos citar el criterio de información bayesiana (en inglés, Bayesian Information Criterion, BIC) (Schwarz, 1978), que busca encontrar una solución bayesiana a la hora de seleccionar un modelo de entre varios de distintas dimensiones. A continuación, se aplican diferentes algoritmos de clustering, como las k-medias (Lloyd, 1982), que intenta particionar n observaciones en k clusters, con cada observación perteneciendo a la media más cercana. Así mismo, la maximización de expectativas (en inglés, Expectation Maximization, EM) (Dempster et al., 1977), consiste en encontrar estimaciones de máxima probabilidad de parámetros en modelos estadísticos. Otra alternativa para realizar el clustering se basa en la elaboración de mapas auto-organizativos (en inglés, Self Organizing Map, SOM) (Kohonen et al., 2001), consistentes en un tipo de red neuronal entrenada para producir una representación discretizada de las muestras de entrenamiento, con el fin de poder particionar la imagen en k regiones. Los métodos de aprendizaje no supervisados se usan para realizar búsquedas semánticas de imágenes, anotaciones automáticas y segmentaciones multibanda. Podemos encontrar una aplicación de estos algoritmos en Caetano and Barone (2001), donde se usa EM para realizar un detector de piel usando una normalización del espacio de color RGB.

2.1.3. Detección de humanos

La presente tesis se centra en el campo particular de la detección de humanos a partir de la segmentación de imágenes. Concretamente la idea consiste en realizar una detección inicial de humanos tanto en el espectro infrarrojo como en el color. Como aproximación inicial a las principales ideas en detección de humanos, podemos citar la clasificación establecida por (Moeslund et al., 2006) según la forma de la detección, la cual se puede realizar basada en el movimiento producido en la escena (la cual ya hemos explicado antes), en la apariencia de los humanos, en su forma o en los datos de profundidad de la escena. A continuación, se procederá a explicar en qué consiste cada uno de estos enfoques.

Segmentación basada en la apariencia

La segmentación basada en la *apariencia* del humano se fundamenta en la idea de que las apariencias de éste y el fondo son diferentes, al igual que las de los individuos entre sí. Estos métodos se basan en construir un modelo de apariencia de cada humano. A partir de éste se pueden adoptar dos alternativas. En la primera, se construyen modelos de apariencia de los objetos segmentados en la imagen actual y se comparan con los modelos previamente elaborados. Otra metodología consiste en segmentar directamente los píxeles en la imagen actual que pertenezcan a alguno de los modelos previamente definidos. Algunos de estos métodos son independientes en el contexto temporal, lo que significa que se aplican al modelo de apariencia general de un humano permaneciendo este modelo invariable, en oposición a los métodos donde el modelo de apariencia del humano se actualiza basándose en imágenes previas de la secuencia actual.

Los métodos libres de contexto temporal se utilizan para detectar humanos en imágenes estáticas (Mohan et al., 2001), donde se detectan individualmente las diversas partes del cuerpo y el aprendizaje se realiza a lo largo de diversas etapas, para detectarlos entrando en una escena (Okuma et al., 2004), o para indexar imágenes en bases de datos (Burak Ozer and Wolf, 2002). En primer lugar se detectan las regiones en movimiento, agrupando los vectores de movimiento de forma automática mediante el uso de las proporciones de las partes del cuerpo humano, utilizando posteriormente modelos para saber si cada región obtenida pertenece al mismo. Estas propuestas se centran mayoritariamente en el empleo de grandes cantidades de datos para el aprendizaje de clasificadores. Por ejemplo, en Okuma et al. (2004) se utilizan 6000 imágenes para entrenar un clasificador basado en Adaboost. Otros ejemplos utilizan métodos basados en bloques, tal y como podemos encontrar en Utsumi and Tetsutani (2002), donde la imagen se divide en una serie de bloques, calculándose para cada bloque la media y la matriz de covarianza de las intensidades de sus píxeles. En base a estas medidas estadísticas, se construye una matriz de distancia donde cada entrada representa la distancia generalizada de Mahalanobis entre las medidas de dos bloques. Gracias a esto, la detección se basa en el hecho de que en las imágenes que no contienen humanos las distancias entre bloques cercanos dentro de la imagen son mayores que en aquellas que contienen humanos. Estos métodos tienen en común que el humano es detectado como una caja (normalmente la caja que lo rodea), con lo que el fondo contenido dentro del rectángulo tendrá un efecto en los resultados. Además, al representar a los humanos como una única entidad en lugar de una serie de subentidades, las oclusiones tienen un gran impacto en estos métodos, a los que también afectan los cambios repentinos de iluminación, pues los modelos han sido elaborados de forma general y no se adaptan a los cambios en la escena.

El contexto temporal hace referencia a aquellos métodos donde se utiliza un modelo aprendido que se actualiza gracias a las imágenes anteriores para detectar píxeles de interés o para clasificar píxeles de los humanos sobre los que se realiza un seguimiento. Estos métodos pueden trabajar tanto a nivel de píxel como a nivel de región. A nivel de píxel se calcula la probabilidad para cada píxel perteneciente a un modelo de humano. El nivel de región se considera cuando una región en la imagen, como un rectángulo, se compara con el modelo de apariencia del humano que se ha predicho en el fotograma actual. Por ejemplo, la probabilidad de que una región en una imagen se corresponda con un modelo de humano en particular. Los modelos de apariencia basados en color se han popularizado recientemente, permitiendo mejores algoritmos de seguimiento en exteriores con oclusiones parciales. Esto ha llevado a la necesidad de modelos capaces de representar las diferencias entre individuos incluso ante oclusiones parciales. Entre los trabajos que usan esta aproximación citaremos el de Jepson et al. (2003), donde se propone la actualización de cada píxel del modelo de apariencia mediante una combinación ponderada de un modelo poco cambiante, un modelo muy cambiante y un modelo ruidoso. Los pesos en la imagen actual se actualizan apoyándose en los diferentes modelos.

Segmentación basada en la forma

La *forma* de los humanos suele ser muy diferente de la forma de otros objetos de la escena. Por eso, la detección de humanos basada en su forma puede generar buenos resultados. En contraposición con los modelos basados en apariencia, las formas de los individuos son frecuentemente muy similares.

Debido a este hecho, los métodos basados en forma se aplican para realizar un seguimiento basado en correspondencias sencillas. Estos métodos se dedican a la detección de humanos y a su seguimiento en entornos no controlados. Recientemente, los avances en resta de fondo permiten obtener mejores siluetas que describen la forma de los humanos en la secuencia de imágenes. Además, han aparecido nuevos métodos para la representación y segmentación de humanos en imágenes estáticas. Como ocurría con los métodos basados en apariencia, se dividen los métodos basados en forma en aquellos que no utilizan un contexto temporal y aquellos que sí lo utilizan.

Entre los métodos libres de contexto temporal podemos destacar el trabajo Leibe et al. (2005), donde se aprenden y almacenan en una serie de plantillas los contornos de humanos caminando. Cada una de estas plantillas se compara con los bordes detectados en las imágenes de entrada combinando los resultados con con la probabilidad de que una persona esté presente, la cual se mide comparando pequeñas zonas de las apariciones de humanos con su distribución de ocurrencia. En el caso de que el contexto temporal sea tenido en cuenta, los métodos basados en forma pueden aplicarse al seguimiento de individuos a lo largo del tiempo. Un ejemplo que podemos destacar se encuentra en Mikolajczyk et al. (2004), donde se usa un método estadístico para detectar humanos usando estructuras geométricas comunes a las de los aspectos que presentan las figuras humanas, entendiéndose un humano como un conjunto flexible de varias partes más simples y utilizando después un clasificador Adaboost entrenado previamente. Finalmente, destacaremos la propuesta de Pedersoli et al. (2011), donde se realiza una detección de humanos basada en una cascada multiresolución de HOGs aplicadas sucesivamente. Estos histogramas pueden reducir el coste computacional de la búsqueda necesaria para la detección sin afectar a la precisión del algoritmo. El método descrito se basa en una cascada de detectores de ventana deslizante, cada uno de los cuales es una máquina de vector soporte lineal compuesta de características extraídas a partir de HOGs a distintas resoluciones, desde la menos detallada en el primer nivel hasta la última con los detalles más finos.

Por su parte, entre los métodos que usan contexto temporal, destacaremos las metodologías descritas en Fan et al. (2008) y Siddiqi et al. (2011). En la primera se describe un mecanismo de detección de humanos que usa SVM. En primer lugar, se utilizan un vector de cuadrícula y otro que parte del centro de forma radial para representar las características de objetos en movimiento extraídos a partir de una resta de fondo. Los datos de muestra se obtienen combinando por igual humanos y no humanos, formando la entrada de entrenamiento a la SVM. También en Siddiqi et al. (2011) se utiliza el contexto temporal adaptando un modelo de contorno activo de forma que sea robusto a los cambios de iluminación y ropa y utilizando flujo óptico para situar el contorno inicial en el fotograma actual.

2.1.3.1. Detección de humanos usando la información del infrarrojo

Comparadas con las imágenes de luz visible, las imágenes térmicas infrarrojas tienen características únicas (Li et al., 2010). Generalmente, la intensidad de los objetos se determina principalmente por su temperatura y calor irradiado y es independiente de las condiciones de luz actual, de forma que un sistema de detección en este espectro se puede aplicar igualmente tanto en día como en noche. Por tanto, la idea más intuitiva para realizar la detección de humanos en el espectro térmico-infrarrojo es

partir de la base de que los humanos se aprecian más calientes que el resto de los objetos de la escena.

Sin embargo, esto no siempre es así, tal y como se describe en Goubet et al. (2006), donde se explica que, si bien esto se cumple especialmente en invierno y durante la noche pudiéndose detectar humanos mediante la aplicación de un umbral; existe el problema de que, aunque se puede realizar una segmentación eficiente en verano usando resta de fondo, es más difícil realizar una clasificación basada en la forma o en las características del cuerpo humano. Esto se debe a que las propiedades de los objetos en la escena (emisividad, reflectividad y transmisibilidad) y la longitud de onda afectan a la intensidad de las imágenes infrarrojas. Los objetos no humanos y los fondos, como animales, coches, postes de la luz y edificios, producen áreas brillantes en dichas imágenes, especialmente en las tardes de verano. Esto hace imposible detectar humanos basándose únicamente en su brillo. En segundo término, debido a las limitaciones en las tecnologías de las cámaras, la mayoría de la imágenes infrarrojas tienen una resolución espacial baja y menos sensibilidad que las imágenes visibles, lo que lleva a menudo a una baja calidad de la imagen, como desenfocado, bajo contraste del objetivo con el fondo, y una gran cantidad de ruido. Se puede decir que muchas de las técnicas encontradas en infrarrojo combinan propiedades de apariencia y forma, ya que detectan los humanos inicialmente en base a la primera (su apariencia suele ser más brillante que la del resto de objetos en la escena) y se filtran y clasifican de acuerdo a la segunda.

Una tendencia muy utilizada en la literatura es el uso de HOGs, combinados con un clasificador basado en una SVM convencional. Un ejemplo de aplicación de esta técnica lo podemos encontrar en O' Malley et al. (2010), donde se segmentan los humanos usando crecimiento de regiones con semillas de alta intensidad. A continuación, se usan características extraídas del histograma de gradientes orientados de las regiones candidatas obtenidas, siendo estas características utilizadas por un clasificador SVM. En Chang et al. (2011) también se usan HOGs combinados con algunas de las técnicas que se han explicado en esta sección. En primer lugar, se realiza una resta de fondo, sobre la que se aplica el algoritmo de Otsu y una serie de operaciones morfológicas de apertura y cierre. En la segunda fase, se filtran las regiones sin humanos mediante restricciones de tamaño y forma. El paso final consiste en mejorar el reconocimiento usando características extraídas mediante HOG, tales como la magnitud y la orientación del gradiente, analizando las estadísticas de cada región y comprobando que se corresponden con humanos mediante el uso de un conjunto de clasificadores Adaboost entrenado previamente.

Como ejemplo de aproximación de estas técnicas, podemos destacar el trabajo de Davis and Sharma (2004), donde se realiza en primer lugar una resta de fondo para identificar objetos de interés locales. Entonces se combina la información de los objetos y el fondo para elaborar un mapa de contornos representando para cada píxel la confianza de que pertenezca a los bordes de una persona. Finalmente, se realiza una búsqueda de caminos para completar segmentos rotos. También podemos destacar la propuesta de Fang et al. (2004), donde se presenta una segmentación basada en histogramas en la que se realizan ajustes tanto en horizontal como en vertical en las regiones candidatas a humanos, estableciéndose dos casos diferentes para el verano y el invierno. En el invierno se usan los valores de niveles de gris para delimitar los humanos, mientras que en el verano se buscan los cambios

de intensidad en los límites de los humanos. En Li et al. (2010) los candidatos a humanos se detectan originalmente usando umbralización, gracias a la propiedad que enunciamos previamente sobre la intensidad mayor de los humanos en una imagen infrarroja. Los candidatos obtenidos se descomponen en diversas capas mediante wavelet, extrayéndose sus características a partir de las subbandas de alta frecuencia. Finalmente, las regiones de humanos se clasifican de acuerdo a una SVM. También se utilizan SVMs en Xu et al. (2005), donde se detectan puntos cálidos, se estima el tamaño de los peatones, y tras separar esas regiones como candidatos a humanos, se estima si son o no humanos mediante una SVM. Más recientemente, en Wang et al. (2012a), se construye el fondo de la escena utilizando una mezcla de gaussianos y realizando un cierre vertical parcial sobre los candidatos a humanos extraídos para corregir las distorsiones que aparecen debidas a la ropa en el espectro infrarrojo. Para diferenciarlos de otros objetos, se realiza una ponderación basada en que la escala de los seres humanos se encuentra siempre dentro de un intervalo determinado. Una vez que se han obtenido los modelos de características, se aplica un clasificador SVM para clasificar los candidatos en humanos y no-humanos.

2.1.3.2. Detección de humanos usando la información del color

El color aparente de un objeto se ve influenciado principalmente por dos factores físicos como son la distribución de energía espectral de la fuente de luz y las propiedades reflexivas de la superficie del objeto (Yilmaz et al., 2006). Si bien el RGB (Red, Green, Blue, en español, rojo, verde, azul) es el espacio más comúnmente utilizado para la adquisición de imágenes, no es un espacio perceptivamente uniforme ya que las diferencias entre los colores en el espacio no se corresponden a las percibidas por los humanos. Por contra, $L, u, v*$ y $L, a, b*$ son espacios de color perceptivamente uniformes, mientras que HSV (Hue, Saturation, Value, en español, tono, saturación y valor) es un espacio aproximadamente uniforme (entendiéndose como espacio de color uniforme aquel en el que las diferencias entre los colores representados en la escala del espacio de color se corresponden con la diferencia visual entre los colores representados). No obstante, estos espacios son sensibles al ruido. Finalmente, si bien no centrado específicamente en la detección de humanos sino de objetos en general, podemos citar métodos de segmentación tales como el propuesto en el artículo de Carmona et al. (2008), en el que el fondo se adapta dinámicamente mediante el filtrado de elementos tales como sombras, reflejos, etc., que aparecen normalmente en la resta de fondo.

Frecuentemente, la representación de un humano completo mediante un único modelo de color es demasiado restrictiva, incluso si el modelo contiene diferentes modos. Por eso, recientemente se está incluyendo información espacial. Por ejemplo, utilizando un correlograma, que es una matriz de co-ocurrencia que expresa la probabilidad de que píxeles de dos colores diferentes se encuentren a una cierta distancia uno de otro (Huang et al., 1999). Por ejemplo, en Capellades et al. (2003) se utiliza una combinación de correlograma e histograma tras detectar previamente los humanos mediante una resta de fondo.

En Rodriguez and Shah (2007) se describe una propuesta para detectar y segmentar humanos en secuencias con aglomeraciones. En primer lugar, dado un conjunto de vídeos de entrenamiento,

se realiza una resta de fondo y se extraen los contornos de cada fotograma mediante la detección de bordes en las manchas (es decir, regiones conexas) de interés correspondientes a humanos en la escena. Estos contornos se integran en vectores de forma que a su vez se agrupan en clusters de posturas usando el algoritmo de las k-medias. Estos contornos se muestrean usando el contexto de sus formas, buscando instancias del descriptor de formas aprendido, y realizando votaciones para las localizaciones humanas y sus posturas respectivas en el fotograma. Otra propuesta se puede encontrar en Schwartz et al. (2009), donde se añade información de textura y color a las características más usadas extraídas mediante HOGs, produciendo un espacio multidimensional enorme (más de 170000 características). Al ser este espacio inabarcable por clasificadores tradicionales, se emplea un método llamado análisis de mínimos parciales (en inglés, Partial Least Square analysis, PLS), que conserva la información significativa que puede servir para distinguir los humanos del fondo, proyectando los datos en un subespacio dimensional mucho más pequeño (20 dimensiones). La idea básica de PLS es construir nuevas variables a partir de combinaciones lineales de las variables originales. La información extraída es entonces utilizada por un clasificador genérico previamente entrenado, tal y como pueden ser las SVMs.

Finalmente, entre las propuestas más actuales que se pueden encontrar en la literatura, destacamos Muhammad Anwer et al. (2011), donde nuevamente se utilizan HOGs para la detección inicial de humanos y SVMs para su clasificación. Sin embargo, la aportación más interesante del artículo se encuentra en la descripción del espacio de colores oponentes (OPP, del inglés OPPonent colors) como una alternativa biológicamente inspirada para detección de humanos, utilizando HOG y afirmando que presenta mejores resultados que la segmentación equivalente en RGB. El espacio OPP se basa en que los fotorreceptores de color de la retina (los conos) son sensibles a las longitudes de onda largas (conos L), media (conos M) y cortas (conos S). La teoría de los procesos oponentes postula que la información de color amarillo-azul y rojo-verde se representa mediante dos canales paralelos en el sistema visual que combinan las señales de los conos de forma diferente. En la etapa inicial de los colores rojo-verde se oponen señales de los conos L y M, mientras que el los colores amarillo-azul las señales de los conos S se oponen a una señal combinada de los conos L y M. Se puede concluir que L, M y S pertenecen a una primera capa de la retina mientras que la luminancia y los colores oponentes pertenecen a una segunda capa de la misma, formando la base de la entrada cromática al córtex visual primario. Por tanto, se puede establecer un espacio de color a partir del espacio RGB consistente en rojo-verde, amarillo-azul y luminancia.

2.1.4. Resumen y conclusiones

Durante la presente sección se han estudiado las principales propuestas encontradas en el campo de la segmentación de imágenes. Se ha comenzado con un estudio preliminar de los principales conceptos básicos que entran en juego a la hora de elaborar un algoritmo de segmentación para posteriormente incidir en las fases que suele presentar un algoritmo de este tipo, presentando dentro de cada fase las principales tendencias que se pueden encontrar en la literatura. Finalmente, se ha incidido especialmente en el tema de la detección de humanos, principal campo de estudio de esta tesis. En

la tabla 2.1 se pueden apreciar los principales algoritmos estudiados.

En general, se aprecia que una de las aproximaciones a la extracción de características más utilizadas es la resta de fondo, ya que constituye un método fácil de implementar y no muy caro computacionalmente. Sin embargo, también se ha explicado que este método puede ser problemático en entornos muy complejos y con numerosos cambios. Aun así, su facilidad de uso y poca complejidad hace que muchos algoritmos partan de esta base, para centrarse más extensamente en el uso de clasificadores y considerando la primera etapa como algo casi trivial. También se ha podido comprobar el auge en los últimos años de los histogramas de gradientes orientados, los cuales tampoco son difíciles de implementar y arrojan buenos resultados en imágenes estáticas sin necesidad de recurrir a la información que proporciona el movimiento.

En cuanto al filtrado de objetos de interés, se ha observado que muchos algoritmos omiten este paso, centrándose en la primera y en la última fase descritas. Sin embargo, este paso es muy necesario si no se quiere proporcionar una gran cantidad de datos a la fase de clasificación, ya que afectaría al coste computacional del sistema, aparte de que puede resultar problemático en el caso de clasificadores con aprendizaje no supervisado.

Finalmente, en la fase de clasificación se ha comprobado el auge de las máquinas de soporte vectorial, las cuales son muchas veces utilizadas en conjunción con histogramas de gradientes orientados, demostrando muchos autores que constituye una buena alternativa a la hora de enfrentarse a la segmentación de humanos. Así, por ejemplo en Dalal and Triggs (2005) se alcanzan tasas de detección del $89\,\%$, y una tasa de falsos positivos del $5{,}8\,\%$ en Schwartz et al. (2009). Sin embargo, también se ha observado que el uso de clasificadores Adaboost puede resultar una alternativa a estimar a la hora de obtener buenos resultados.

2.1.4.1. Detección de humanos

La presente tesis se ha enfocado en la detección de humanos, y por ello se ha prestado especial atención a los algoritmos dedicados a esta tarea. Los algoritmos estudiados se muestran en la tabla 2.2, donde se han dividido las diversas técnicas de detección de humanos basadas en apariencia y basadas en forma.

Se ha observado que, debido a las características que presentan los humanos, muchos de estos métodos utilizan la información del movimiento, principalmente la resta de fondo, si bien también hay bastantes aproximaciones que utilizan la información del flujo óptico para detectar humanos, ya sea por estar enfocados desde vehículos en marcha (Talukder and Matthies, 2004) o desde robots (Nakada et al., 2008). De forma similar a como se ha apuntado en la parte de segmentación, también los histogramas de gradientes orientados comienzan a tomar auge en estas aplicaciones, nuevamente utilizados con SVMs.

En la detección de humanos en infrarrojo, se ha confirmado que una aproximación muy común consiste en utilizar la propiedad de que los humanos emiten más radiación que el resto de objetos en la

Tabla 2.1: Resumen de los algoritmos de segmentación estudiados

Fase	Método	Algoritmo	Ejemplos
Extracción de características	Detección de movimiento	Resta de fondo	(Kanade et al., 1998), (Huwer and Niemann, 2000), (Cavallaro and Ebrahimi, 2001), (Capellades et al., 2003), (Davis and Sharma, 2004), (Rodriguez and Shah, 2007) (Carmona et al., 2008), (Wang et al., 2012a), (Xiao et al., 2013)
		Diferencia de imágenes	(Jain and Nagel, 1979),(Fablet et al., 1999), (Konrad, 2000), (Fernández-Caballero et al., 2003), (López et al., 2006)
		Flujo óptico	(Talukder and Matthies, 2004), (Nakada et al., 2008), (Klappstein et al., 2009)
		Otros	(Burak Ozer and Wolf, 2002)
	Propuestas estadísticas y probabilísticas	Geometría	(Mikolajczyk et al., 2004)
		Contornos activos	(Siddiqi et al., 2011)
	Crecimiento de regiones	Watershed	(Kim and Kim, 2003), (Navon et al., 2005)
	Grafos	Corte	(Shi and Malik, 1997), (Birchfield et al., 2007), (Bugeau and Pérez, 2008), (Kohli et al., 2008), (Li and Yang, 2008), (Hu et al., 2013), (Collomosse and Wang, 2012)
	Otras propuestas	Histogramas de gradientes orientados	(Dalal and Triggs, 2005), (Schwartz et al., 2009), (O' Malley et al., 2010), (Muhammad Anwer et al., 2011), (Chang et al., 2011), (Pedersoli et al., 2011)
		Otros	(Lowe, 1999), (Mohan et al., 2001), (Utsumi and Tetsutani, 2002), (Xu et al., 2005), (Li et al., 2010), (Baber et al., 2013)
Filtrado de objetos de interés	Densidades de probabilidad	Paramétricos	(Zhu and Yuille, 1996), (Wang et al., 2012a)
		No paramétricos	(Elgammal et al., 2002), (Comaniciu et al., 2003),
	Plantillas	Probabilísticas	(Nanda and Davis, 2002), (Leibe et al., 2005)
Clasificación	Aprendizaje supervisado	SVM	(Xu et al., 2005), (Fan et al., 2008), (Cao et al., 2009), (Li et al., 2010), (O' Malley et al., 2010), (Muhammad Anwer et al., 2011), (Wang et al., 2012a)
		Adaboost	(Kage et al., 2007), (Vincenzo and Lisa, 2007), (Chang et al., 2011)
		Otros	(Utsumi and Tetsutani, 2002), (Schwartz et al., 2009),
	Aprendizaje no supervisado	EM	(Caetano and Barone, 2001)
		K-Medias	(Rodriguez and Shah, 2007)
		Otros	(Mohan et al., 2001)

escena, si bien esta información tiene que ser refinada o apoyada mediante clasificadores o algoritmos de resta de fondo con el fin de reducir el número de falsas detecciones. Una tendencia especialmente interesante consiste en utilizar el contraste que la radiación mencionada produce con el resto de la escena para realizar una resta de fondo inicial, si bien entonces vuelven a aparecer los problemas que hemos mencionado con respecto a la resta de fondo.

Se ha podido apreciar en la literatura que es difícil encontrar algoritmos centrados en detección de humanos que utilicen la información del color (ver (Hernández-Vela et al., 2012) como contraejemplo), ya que muchos se centran en detección de objetos en general o pasan directamente a un nivel superior de seguimiento retroalimentando a la segmentación mediante dicho nivel. Creemos que utilizar de forma más exhaustiva la información de la forma y la apariencia de los humanos puede resultar especialmente útil y es un campo poco explorado, ya que el color permite combinar información de la forma y apariencia de los humanos por igual.

Tabla 2.2: Técnicas de segmentación para detección de humanos

Espectro	**Técnica**	**Ejemplos**
Niveles de gris	Apariencia	(Utsumi and Tetsutani, 2002), (Jepson et al., 2003)
	Forma	(Mikolajczyk et al., 2004), (Leibe et al., 2005), (Pedersoli et al., 2011), (Siddiqi et al., 2011)
Infrarrojo	Apariencia	(Fang et al., 2004), (Xu et al., 2005), (Li et al., 2010),
	Forma	(Davis and Sharma, 2004), (O' Malley et al., 2010), (Fernández-Caballero et al., 2011a), (Chang et al., 2011), (Wang et al., 2012b), (Sokolova et al., 2013)
Color	Apariencia	(Capellades et al., 2003), (Schwartz et al., 2009), (Hernández-Vela et al., 2012)
	Forma	(Rodriguez and Shah, 2007), (Muhammad Anwer et al., 2011)

2.2. Fusión de imágenes

El principal campo de estudio de esta tesis es la fusión de imágenes de vídeo, que tiene como objetivo mejorar los resultados obtenidos de la anterior fase (segmentación). Este campo se enmarca dentro del ámbito de la fusión multisensorial, la cual comenzaremos definiendo y explicando brevemente para posteriormente estudiar las fases necesarias para llevar a cabo una fusión de imágenes, incidiendo además en las diversas aproximaciones a estas etapas que se pueden encontrar en la literatura. Finalmente, se llevará a cabo un breve resumen de las principales conclusiones extraídas del estudio de los diversos algoritmos de fusión explicados.

2.2.1. Definición de fusión

Se puede definir la fusión de datos como la teoría, las técnicas y las herramientas utilizadas para combinar datos de sensores o derivados de ellos en un formato de representación común. De acuerdo a Mitchell (2007) el concepto general de fusión es análogo a la forma en la que los humanos y los animales utilizan sus sentidos, experiencias y capacidad de razonar para aumentar sus posibilidades de supervivencia. Así, la fusión multisensorial intenta determinar el mejor procedimiento para combinar todas las entradas de información. El uso de modelos probabilísticos en fusión aporta la ventaja de manejar la incertidumbre subyacente a las relaciones entre sensores y fuentes de información. Como ejemplo se puede citar la metodología bayesiana que permite formular los problemas de fusión como problemas matemáticos manejando la incertidumbre del problema.

Con el empleo de más de un sensor realizando fusión multisensorial se consigue aumentar la calidad de la información obtenida de varios modos, pudiendo considerarse que la fusión de datos mejora las prestaciones de un sistema en las siguientes vías tal y como se explica en Bellot et al. (2002):

- Representación: La información posee un nivel de abstracción o una granularidad mayor que la suma de los datos de entrada.
- Certeza: La probabilidad de que la información sea correcta tras la fusión sensorial es mayor que la probabilidad a priori de los datos antes de la misma.
- Precisión: La desviación estándar de los datos tras la fusión es menor que la obtenida directamente de las fuentes. Si los datos de entrada son ruidosos o erróneos, la fusión intenta reducir o eliminar su efecto.
- Completitud: Al añadir nueva información al conocimiento que se posee de un entorno se genera una vista más completa del mismo.

Por tanto, se puede concluir que la principal motivación a la hora de realizar fusión multisensorial es el conseguir una mejora sustancial en la calidad de la información proveniente de los sensores.

Dentro de la fusión multisensorial, nos centramos a partir de ahora en el campo de la fusión de imágenes. El objetivo de este campo en particular es integrar información complementaria de datos multisensoriales de forma que las nuevas imágenes obtenidas sean más adecuadas para la percepción visual humana, así como para las tareas de visión artificial tales como la segmentación, la extracción de características y el reconocimiento de objetos (Li et al., 1995).

2.2.2. Fases en la fusión de imágenes

Para realizar una fusión adecuada de los datos de varias fuentes de adquisición es necesario convertir sus lecturas a un formato común, lo que permite compatibilizar la información de las mismas facilitando la fusión. Así, se puede decir que a la hora de realizar una fusión de imágenes es necesario que se enmarquen dentro de un sistema común de coordenadas (alineación espacial), correspondan a un mismo eje temporal (alineación temporal) y se sitúen dentro de una escala de valores común, encontrándose las imágenes normalizadas dentro de dicha escala (normalización de valores). A continuación, describiremos en mayor profundidad estas operaciones, divididas en dos grandes etapas: alineación y normalización de valores.

En la alineación de las imágenes, éstas se convierten al mismo rango temporal y espacial, con el fin de poder inscribir las imágenes que se desea fusionar dentro de un mismo eje de coordenadas y de un mismo instante de tiempo. Por tanto, este primer paso puede ser dividido en alineación espacial y temporal. En la alineación espacial, se transforman las posiciones espaciales, (x, y), a un sistema de

coordenadas común, mientras que la temporal implica la transformación del tiempo local, t, en un eje de tiempo común.

El segundo gran paso consiste en la normalización de los valores de las fuentes de adquisición. Para realizar esta última operación, tanto los valores locales de las fuentes como su incertidumbre deben normalizarse en una escala común. Este proceso se puede realizar atendiendo únicamente a la información de cada imagen a nivel de píxel, en base a las regiones obtenidas de cada imagen o utilizando la información de los descriptores obtenidos sobre dichas regiones a nivel simbólico.

2.2.2.1. Alineación de las imágenes

Tal y como hemos dicho, el primer paso a la hora de realizar una fusión de imágenes es buscar un marco común para las mismas, tanto a nivel espacial como temporal. Este paso es fundamental para que se pueda realizar correctamente la fusión de imágenes, y un fallo en el mismo será crítico en el proceso, ya que, si las imágenes no se alinean debidamente la información obtenida posteriormente será errónea.

Alineación espacial

Como ya se ha indicado, la alineación espacial consiste en la conversión de las posiciones espaciales locales en un sistema de coordenadas común. Esto supone la primera etapa para la formación de un sistema común de representación. En esta sección nos limitaremos a sensores de imágenes bidimensionales (x, y), siendo denominado en este caso el proceso de alineamiento espacial como *registro de la imagen*. Denotemos dos imágenes de entrada en escala de grises como I_1 e I_2. Estas imágenes serán, respectivamente, la imagen de *referencia* y la imagen de *test*. El registro de la imagen consiste en encontrar una transformación T que mapee coordenadas de I_1 en I_2. Así, dado un píxel (x, y) de I_1, la posición correspondiente en I_2 será $(x', y') = T(x, y)$. En la literatura podemos encontrar diferentes enfoques para realizar el registro de la imagen. En base a la forma de adquisición podemos clasificar dichos enfoques en análisis multivista, multimodal, multitemporal y registros de modelo a escena (Zitová and Flusser, 2003).

Análisis multivista

En el *análisis multivista* se toman imágenes de la misma escena desde distintos puntos de vista con el objetivo de conseguir una representación tridimensional o bidimensional más exacta de la escena. Para ello será necesario obtener múltiples funciones que transformen cada imagen de entrada al espacio de la imagen. Dentro de este campo toma especial importancia el concepto de disparidad, el cual se define como la diferencia entre las imágenes del mismo objeto proyectado sobre distintas cámaras. El grado de disparidad entre dos o más imágenes depende del ángulo de convergencia formado por las cámaras, es decir, el ángulo formado por las lentes de cada cámara que convergen en un objeto, estando dicho ángulo relacionado con la distancia entre un objeto y las lentes. En Jain and Ross (2002) podemos encontrar una aplicación de este análisis con el objetivo de realizar un sistema de verificación de huellas dactilares donde se construye una huella mosaico a partir de huellas parciales,

mientras que más recientemente, el algoritmo propuesto en Starck and Hilton (2008) usa parametrización local de formas para permitir reconstruir posiciones de los vértices de los humanos con el fin de ajustar su superficie en múltiples niveles de detalle dentro de un modelo jerárquico. Basándose en dicho modelo se usa el conocimiento previo que se posee de la escena para poder establecer el campo de visión común entre las diferentes vistas y regularizar la reconstrucción de las formas de los objetos cuando los datos originales son ambiguos o presentan ruido, utilizando en el proceso información de la silueta, de la vista estéreo y restricciones definidas previamente por el usuario. En Devarajan et al. (2008) se describe un método descentralizado para obtener el grafo de visión para una red de cámaras distribuidas, donde cada arista une dos cámaras que cubren una parte lo suficientemente amplia del mismo entorno. En base a esto se presenta un algoritmo distribuido en el que cada cámara realiza una optimización local sobre sus parámetros y los puntos de la escena cubiertos por sus vecinos en el grafo para obtener una estimación inicial de la calibración, la cual se refinará a partir de la estimación de las diversas cámaras mediante un algoritmo de inferencia distribuido de propagación de creencias basado en redes bayesianas. También en Thompson and Wettergreen (2005) podemos encontrar una aplicación al campo de la robótica de este tipo de análisis, en la que las vistas de diversos robots se fusionan usando EM para ajustarse a un modelo de mundo común.

Análisis multitemporal

En el *análisis multitemporal* se adquieren imágenes de la misma escena en momentos distintos, normalmente a intervalos regulares y muchas veces bajo condiciones distintas. El objetivo es encontrar y evaluar los cambios aparecidos en la escena entre capturas consecutivas. Entre sus aplicaciones podemos incluir seguimiento de movimiento y detección automática de cambios producidos en el entorno (como puede ser un objeto abandonado en un aeropuerto) para aplicaciones de seguridad. En Petrovic and Xydeas (2004) podemos encontrar una aplicación de este análisis donde se combinan imágenes tomadas por satélite en diversas fechas con el fin de poder observar los cambios producidos en una zona a través del tiempo.

Análisis multimodal

La presente tesis hará hincapié en el *análisis multimodal*, en el que imágenes de la misma escena son capturadas por sensores distintos, siendo el objetivo integrar la información obtenida de diferentes fuentes para lograr una representación más compleja y detallada. Entre otros campos, estas técnicas se usan en aplicaciones de imágenes médicas y en fusión de información de cámaras con características distintas, donde las imágenes de color e infrarrojo se combinan para mejorar la resolución espectral e independizarse de las condiciones lumínicas de la escena. Una forma muy sencilla de realizar esta alineación es el uso de un tablero de ajedrez iluminado por focos halógenos de intensidad alta (Krotosky and Trivedi, 2006), de manera que, al absorber las casillas negras toda la radiación éstas aparecerán más brillantes en el espectro infrarrojo. Una vez que se tiene el tablero de ajedrez enfocado en ambos espectros, la solución consiste en encontrar los bordes y vértices de la cuadrícula y estimar las transformaciones oportunas. Como ejemplo de estas técnicas, podemos citar Leykin and Hammoud (2010), donde se realiza una resta de fondo probabilística de forma que se combinan características de los dos espectros, pudiéndose eliminar las sombras gracias a la combinación de información del color

y el infrarrojo.

Una corriente muy utilizada a la hora de hacer el registro multimodal consiste en el uso de información mutua, la cual es una medida de la dependencia estadística entre dos conjuntos de datos. La información mutua entre dos variables aleatorias X e Y se define como podemos ver en la ecuación (2.11), donde $H(X) = -\sum X \in \chi log(P(X))$ representa la entropía de la variable aleatoria X, χ el conjunto de valores que puede tomar dicha variable y $P(X)$ su distribución de probabilidad. Este concepto fue propuesto en Cover et al. (1991) y desarrollado para visión artificial en Collignon et al. (1995); Maes et al. (1997); Viola and Wells III (1997), siendo el objetivo de muchos algoritmos maximizar esta medida, tal y como se muestra en Krotosky and Trivedi (2006) o en Chen et al. (2003).

$$MI(X,Y) = H(Y) - H(Y|X) = H(X) + H(Y) + H(X,Y) \tag{2.11}$$

Registro de escena a modelo

Finalmente, podemos citar las aplicaciones que realizan un *registro de escena a modelo*, el cual puede ser una representación sintética o escaneada de la escena, como por ejemplo mapas o modelos digitales de elevación en SIG (Sistema de Información Geográfica). El objetivo es localizar la imagen capturada en el modelo y/o comparar la imagen con el mismo. Este proceso también es relativamente frecuente en el campo de la medicina, donde podemos encontrar ejemplos como Chouteau et al. (2007), donde se intenta determinar la variación en la cinemática de rodillas a las que se ha realizado una artroplastia a partir de radiografías antiguas donde los parámetros de la proyección de rayos X son desconocidos, pero se sabe que se encuentran dentro de unos determinados intervalos.

Debido a la gran diversidad de imágenes que pueden ser registradas y a los diversos tipos de degradaciones que pueden aparecer, es imposible diseñar un método universal que pueda resolver todas las tareas necesarias para un registro. Cada método debe tomar en cuenta no solo el tipo de deformación asumido entre las imágenes, sino que además intervienen factores tales como las deformaciones radiométricas, el ruido, la precisión mínima deseada para el registro y las características de los datos que puede manejar la aplicación. Sin embargo, tal y como se enuncia en Zitová and Flusser (2003), la mayoría de los métodos de registro realizan una *detección de características* inicial en ambas imágenes, realizando posteriormente una *unificación* de las mismas. En base a esta unificación se lleva a cabo una *estimación del modelo de transformación* que hay que aplicar para poder mapear dichas imágenes sobre un sistema de coordenadas común, para finalmente llevar a cabo dicha transformación mediante un *remuestreo y transformación de la imagen* de test sobre la de referencia. A continuación, explicaremos con mayor detalle estos cuatro pasos.

Durante la *detección de características* se detectan los puntos o zonas distintivas (regiones cerradas, bordes, contornos, esquinas, etc.) manual o automáticamente. De cara a un nivel superior, estas características se pueden representar mediante los puntos que las definen (tales como el centro de gravedad), los cuales son conocidos como puntos de control en la literatura. Sin embargo, este paso puede presentar diversos problemas. En primer lugar, hay que decidir qué tipo de características es apropiado para el objetivo deseado. Dichas características deben ser fáciles de detectar de forma efi-

ciente tanto en la imagen de test como en la de referencia. Los métodos de detección deben además ser precisos y no ser sensibles a la degradación de la imagen.

El siguiente paso consiste en la *unificación de características*, durante la que se realiza una correspondencia entre las características detectadas en la imagen de test y en la de referencia. Con este fin, se pueden usar diversos descriptores de características y medidas de similaridad, junto con relaciones espaciales. En esta etapa pueden surgir problemas debidos a una detección errónea de características. Otro punto problemático importante consiste en que los descriptores deben ser invariantes ante degradaciones de la imagen, así como tener el suficiente poder discriminativo para distinguir entre las distintas características elegidas y ser lo bastante estables como para no verse influidos por pequeñas variaciones inesperadas y ruido.

A continuación, se realiza una *estimación del modelo de transformación*, en la que se estiman el tipo y parámetros de las funciones de mapeado, las cuales alinean la imagen de test con la de referencia. Los parámetros de estas funciones se calculan utilizando la correspondencia de características realizada en el paso anterior. El modelo elegido debe ser lo bastante flexible como para poder afrontar todas las posibles degradaciones que pueden aparecer, siendo también importante elegir el error de aproximación máximo que se considerará como aceptable. Además, se debe escoger qué diferencias de las imágenes pueden obviarse y cuáles no, siendo éste un problema de especial dificultad. El mapeo que se debe realizar debe ser planificado cuidadosamente, pues el problema abarca desde la eliminación de la distorsión óptica producida por las lentes o la perspectiva, hasta incluso la alineación de dos imágenes. Así, debe alcanzarse un compromiso entre el tiempo de procesado y la exactitud de la transformación. En algunos casos, las imágenes experimentan deformaciones locales, por lo que no será posible describir la transformación de dos imágenes utilizando una transformación sencilla. En estas situaciones se utilizan transformaciones “no-rígidas”, compuestas por la suma de una transformación global a bajo nivel y otra local en la que los parámetros cambian en cada píxel. Se puede encontrar otro ejemplo en Perperidis et al. (2005), donde se utiliza la transformación de imágenes de resonancia magnética tomadas en diferentes momentos del ciclo cardíaco para el diagnóstico de enfermedades cardiovasculares.

Finalmente, se realiza un *remuestreo y transformación de la imagen*, en el que las coordenadas de la imagen de test se transforman al ámbito de la imagen de referencia usando las funciones de mapeo obtenidas en el paso anterior. En la mayoría de los casos es suficiente buscar el vecino más cercano o interpolación bilineal. En este punto es además crucial el encontrar una técnica que defina el nivel de gris para los píxeles de la imagen de referencia que, una vez transformados, no se encuentran en la imagen de test. Para la interpolación de la imagen, es preciso reconstruir una imagen continua bidimensional $I(x, y)$ a partir de valores de píxeles discretos. Por eso, la amplitud de una posición (x, y) se estima mediante los valores de los vecinos.

En la Figura 2.5 podemos apreciar un ejemplo de las fases para el registro que hemos nombrado, en este caso aplicadas en un caso de análisis multimodal en los espectros de color e infrarrojo. En (a) podemos ver la detección de características (en este caso esquinas) en ambos espectros. A continuación, en (b) se puede apreciar la correspondencia de características, donde se ha asignado a cada

esquina un identificador numérico. En (c) se puede ver la estimación del modelo de transformación a realizar, pudiéndose apreciar el cambio de escala y rotación que se debe realizar sobre la imagen infrarroja. Finalmente, en (d) podemos ver la transformación aplicada de forma que ambas imágenes se enmarcan dentro del mismo eje de coordenadas.

Alineación temporal

El objetivo del alineamiento temporal es definir una transformación $T(t)$ que mapea el tiempo local t del fotograma capturado de una cámara en un eje de tiempo común t' (Sakoe and Chiba, 1978). Así, el alineamiento temporal es uno de los procesos básicos a la hora de crear un formato de representación común y juega un papel crítico en aplicaciones de fusión, especialmente si se opera en tiempo real. Si denotamos $I(x, y, t)$ como el fotograma capturado por una determinada cámara, entonces hablaremos de que la transformación $T(t)$ es una función de la posición (x, y) y el tiempo t. La deformación dinámica del tiempo (Dynamic Time Warping o DTW) , según Rabiner and Juang (1993); Ratanamahatana (2005), es una técnica general para realizar el alineamiento temporal entre dos secuencias de tiempo, P y Q. DTW intenta encontrar el alineamiento óptimo entre dos series de datos P y Q de tal manera que se minimice la suma de las distancias locales $d(i, j)$ entre los pares de observaciones alineadas (P_i, Q_i). Esta distancia $d(i, j)$ se define comúnmente como el cuadrado de la distancia Euclídea (ver 2.12), aunque se pueden usar otras fórmulas para calcular la distancia local. Este procedimiento deforma los ejes de tiempo t y t' de modo que las observaciones de los sensores se correspondan en un mismo eje de tiempo. En general, los ejes t y t' se alinean de modo no lineal. El esquema de esta técnica lo podemos ver en la Figura 2.6. Una aplicación de esta técnica puede ser el reconocimiento de personas a partir de su forma de caminar, como se ve en Kale et al. (2003). La propuesta del artículo consiste en usar modelos, y al no ser recomendable el hacer correspondencias fotograma a fotograma entre la captura actual y el modelo (ya que una misma persona puede cambiar su velocidad y estilo de andar dentro de una misma secuencia), se decide permitir una región de búsqueda para cada instante de tiempo durante la evaluación.

$$d(i, j) = (P_i - Q_i)^2 \tag{2.12}$$

2.2.2.2. Normalización de los valores

Existen numerosos enfoques a la hora de abordar este paso final que se puede decir que constituye el núcleo la fusión de imágenes. Una vez realizado el registro previo de las imágenes que queremos fusionar, podemos tomar diversas alternativas, que se pueden realizar a nivel de píxel, de región o simbólico (Luo and Kay, 1992). Es importante aclarar que cuando se establecen estos niveles de fusión también se suele incluir la fusión a nivel de señal, definida como la combinación de un grupo de sensores con el objetivo de producir una señal única de mayor calidad y fiabilidad (Li et al., 1995). Como anotación, podemos destacar que en la literatura muchas publicaciones están centradas únicamente en este paso, obviando la alineación previa y dando por hecho que se ha realizado previamente, siendo denominado este paso muchas veces como *fusión* propiamente dicha, denominación que también

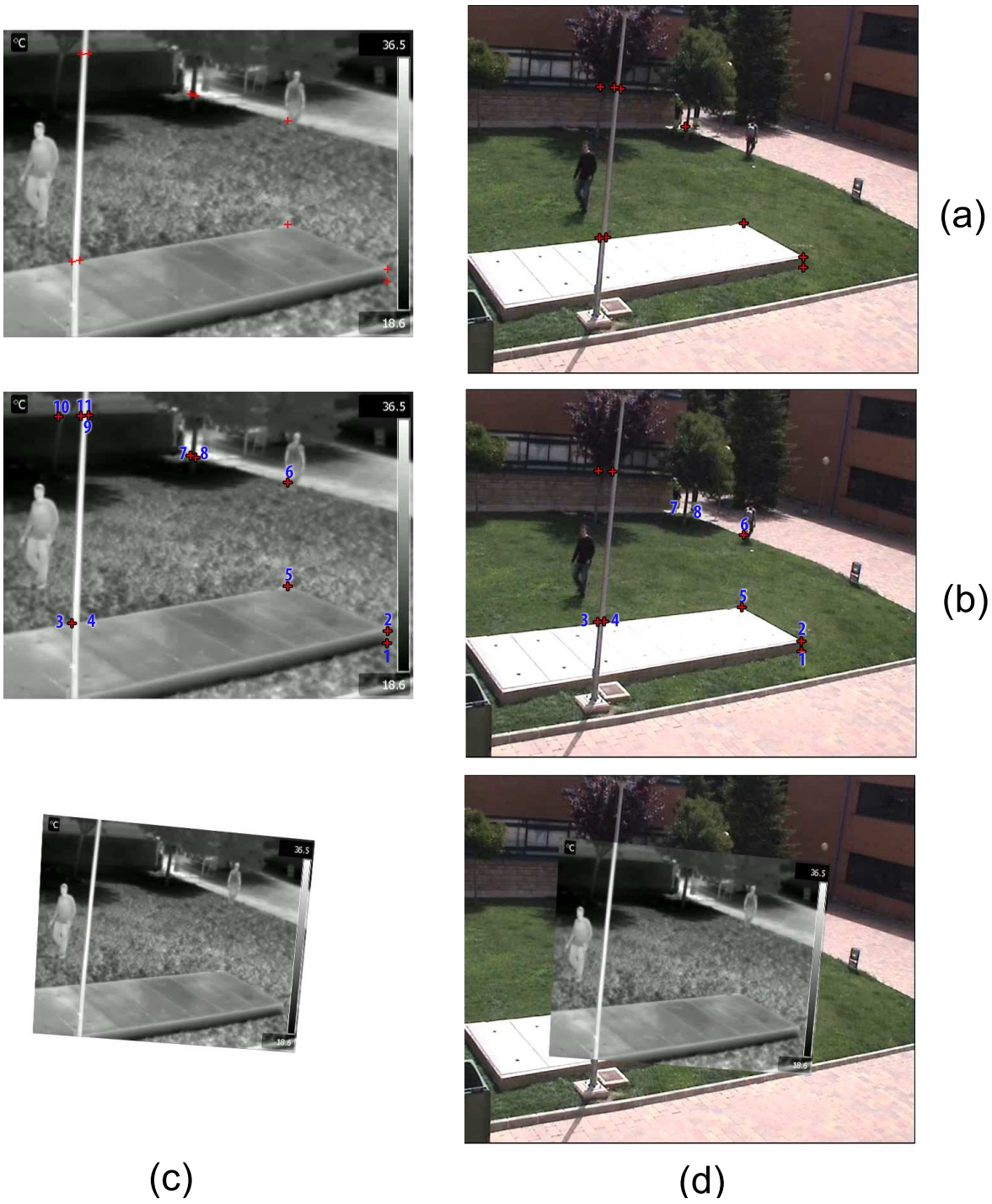


Figura 2.5: Ejemplo del registro por análisis multimodal partir de dos imágenes. La fila superior (a) muestra las características detectadas (en este caso esquinas). En la fila intermedia (b) se aprecia la correspondencia calculada de las características. Finalmente, en (c) se ve la transformación estimada, mientras que (d) muestra la transformación aplicada.

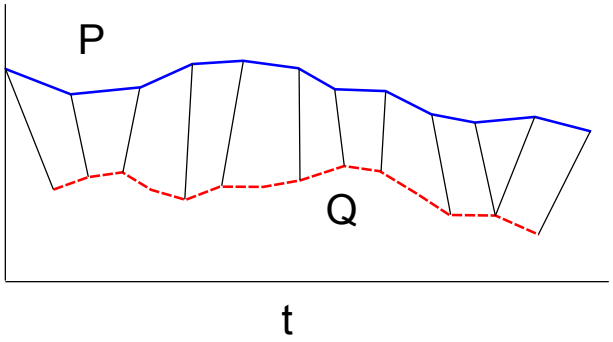


Figura 2.6: Resultado del algoritmo DTW sobre dos series de tiempo P y Q según Mitchell (2007).

emplearemos ya que es la más comúnmente utilizada en la literatura.

Algoritmos a nivel de píxel

Los algoritmos a nivel de píxel trabajan en el dominio espacial o en el de las diversas transformadas que explicaremos a continuación. Aunque la fusión a nivel de píxel es una operación local, los algoritmos del dominio de las transformadas (las cuales explicaremos a continuación en mayor profundidad) crean la imagen fusionada a nivel global. Cambiando un solo coeficiente en la imagen fusionada todos los valores del dominio espacial cambiarán (o al menos una vecindad completa). Como resultado, se pueden crear artefactos indeseados en algunas áreas de la imagen durante el proceso de acentuar las propiedades en otras zonas.

Podemos descomponer los objetivos de las técnicas de fusión a nivel de píxel dentro de los siguientes requisitos (Rockinger, 1996):

- (1) No se debe descartar ninguna información relevante ni destacada de las imágenes de entrada.
- (2) No se deben introducir artefactos ni inconsistencias que puedan distraer o confundir a un observador humano o a un sistema de procesado de imagen.
- (3) El sistema debe ser fiable, robusto y con tolerancia al ruido y a pequeños problemas de registro de la imagen.

La aproximación más simple a este tipo de técnicas consiste en hacer el promedio píxel a píxel de las imágenes que se desea fusionar. Sin embargo, esto conlleva efectos colaterales negativos tales como una reducción del contraste (Li and Yang, 2008).

Técnicas de descomposición

Una aproximación muy común en estos algoritmos consiste en descomponer las imágenes de entrada en subimágenes o subbandas de frecuencia. A continuación pasaremos a describir brevemente las principales técnicas utilizadas para la descomposición.

Las *técnicas de análisis multiresolución (MRA)* descomponen la imagen de entrada I en una secuencia de imágenes $I_l, ...I_L$,$l \in \{1, 2, \ldots, L\}$, capturando cada una la información de I en una escala

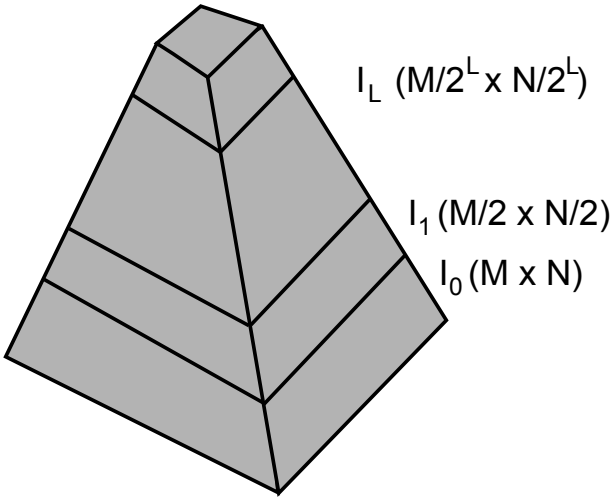


Figura 2.7: Representación piramidal del MRA. La base de la pirámide es la imagen de entrada $I_{M\times N}$. Según se asciende en la pirámide, las imágenes derivadas contienen una mayor resolución espacial. En el nivel l-ésimo, la imagen tendrá las dimensiones $M/2^L \times N/2^L$ Mitchell (2010).

y orientación distinta. Gráficamente, tal y como muestra la Figura 2.7, las L imágenes pueden verse como una pirámide, al fondo de la cual encontramos la imagen I_0, idéntica a la imagen de entrada I. Los niveles superiores de dicha pirámide se construyen sucesivamente mediante la aplicación de filtros paso bajo (como suavizados gaussianos) y sub-muestreo a la imagen I_{l-1}. Eligiendo los filtros de paso de baja frecuencia apropiados es posible seleccionar el cambio de resolución entre las imágenes I_l.

La idea básica es realizar una descomposición en multiresolución en cada imagen fuente, integrando posteriormente todas estas descomposiciones para formar una representación compuesta, y finalmente reconstruir la imagen fusionada realizando una transformada inversa de multiresolución (Li and Yang, 2008). Como ejemplo de estos algoritmos, podemos citar Petrovic and Xydeas (2004), donde la fusión se realiza en un mapa de gradientes multiresolución. En cada resolución, las imágenes de entrada se representan como mapas de gradientes, combinándose para producir nuevos mapas de gradiente fusionados. Las señales de estos mapas son procesadas, usando filtros de gradiente para producir una representación piramidal multiresolución. La imagen fusionada de salida se obtiene al aplicar en dicha pirámide un proceso de reconstrucción análogo al de la transformada wavelet tradicional.

Un caso especial de MRA es la *descomposición por transformada wavelet discreta (DWT)* (Haar, 1910), donde los filtros se diseñan para que las sucesivas capas de la pirámide solo incluyan detalles que no se encuentran en los niveles precedentes. Se utiliza así unos filtros paso-bajo y paso-alto en cascada combinados con una operación de sub-muestreo. Si consideramos la descomposición de una señal unidimensional x, el proceso para aplicar DWT a x puede verse como una serie de filtros donde en cada nivel de descomposición la señal x_l se divide en una componente de alta frecuencia y_{l+1} y otra de baja frecuencia x_{l+1}. Esta componente de baja frecuencia es sucesivamente descompuesta hasta

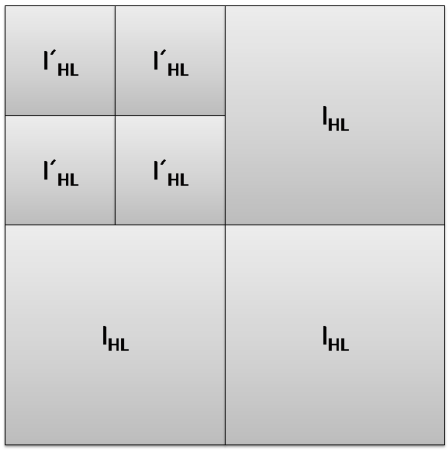


Figura 2.8: Transformación DWT bidimensional. La imagen $I(M,N)$ se descompone en tres imágenes detalle $I_{LH}(M/2,N/2)$, $I_{HL}(M/2,N/2)$ y $I_{HH}(M/2,N/2)$ y una de aproximación $I_{LL}(M/2,N/2)$. Esta se descompone a su vez en tres imágenes detalle $I'_{LH}(M/4,N/4)$, $I'_{HL}(M/4,N/4)$ y $I'_{HH}(M/4,N/4)$ y una de aproximación $I'_{LL}(M/4,N/4)$. Mitchell (2010).

alcanzar la resolución deseada. Este proceso es fácilmente ampliable a imágenes bidimensionales, sustituyendo la señal por una imagen y dividiendo la imagen en diversas subimágenes, tal y como se puede ver en la Figura 2.8. Un artículo reciente presenta un método basado en wavelet para la fusión de imágenes mutifocales registradas espacialmente (Saeedi and Faez, 2013).

Otro algoritmo de descomposición, también utilizado frecuentemente en la literatura, es la transformada NSCT (en inglés, Nonsubsampled Contourlet Transform, transformada de contornos no submuestreados) (da Cunha et al., 2006), que consiste en un conjunto de filtros que divide el plano de frecuencias bidimensional en diversas subbandas, de forma que por un lado una serie de filtros de frecuencia paso-baja posibilitan que la transformada puede dividirse en una pirámide que mantiene la multiescalabilidad del algoritmo y, por otro, una estructura de filtros direccionales proporciona direccionalidad. Recientemente, en Wang et al. (2013) se ha presentado un algoritmo de fusión de imágenes basado en FRFT (en inglés, Fractional Fourier Transform).

Como ejemplo general de algoritmo de fusión a nivel de píxel podemos citar a Jang and Ra (2008), donde se utiliza el espacio HLS para realizar fusión entre imágenes en el espectro infrarrojo e imágenes en niveles de gris. En primer lugar, se fusionan los valores de intensidad, usando una estructura wavelet basada en el gradiente mediante un mapa y filtros de gradiente. Posteriormente, se asigna la componente de tono representando la imagen del espectro visible con colores entre el amarillo y el cian, mientras que la imagen del espectro infrarrojo se representa utilizando colores entre el rojo y el azul en componente de tonalidad. Esto permite que en el espectro visible los objetos brillantes se representen en amarillo y los oscuros en cian, mientras que en la imagen en infrarrojo los objetos cálidos se representan en rojo y los fríos en azul. Finalmente, la componente de saturación se calcula en base a la tonalidad asignando un color más puro (un mayor valor de la componente de saturación) a la información más fácil de percibir, mientras que la menos destacable se representará con un menor

valor de esta componente. Para ello, se asume que la imagen en infrarrojo representa la información característica de la imagen, mientras que la imagen en color representa la información general. El problema es que en el cálculo de esta componente se utilizan una serie de constantes determinadas únicamente de forma experimental. Otra propuesta se puede encontrar en Pszczólkowski and Soto (2007), donde se pretende detectar humanos en entornos de interior usando un robot. Tras usar un sistema de visión en estéreo que obtiene una imagen de disparidad, se segmentan a partir de la misma los objetos de interés usando un algoritmo de crecimiento de regiones. Finalmente, un clasificador probabilístico basado en mezcla de gaussianos proporciona información para decidir si una región de piel determinada corresponde a un humano. Sin embargo, este método se encuentra limitado a encontrar únicamente gente de pie y mirando de cara al robot.

Sin embargo, las técnicas de descomposición presentan una serie de problemas (Gemma and Piella, 2003). Por ejemplo, el realizar un muestreo causa un deterioro en la calidad de la imagen fusionada al introducir efectos de bloque mayores que los que se habrían obtenido sin el muestreo. También, muchas veces se quiere que haya invarianza respecto a desplazamiento y rotación, ya que el resultado de la fusión no debe depender de la localización ni la orientación de los objetos en las imágenes de entrada. Por ello, estas técnicas se usan muy a menudo en algoritmos a nivel de regiones o características, con el fin de analizar la forma y tamaño de los objetos de interés.

Aunque las técnicas de fusión a nivel de píxel son las más utilizadas en la literatura ver, por ejemplo, (Petrovic and Xydeas, 2004; Meytlis and Sirovich, 2007; Correa et al., 2008; Li et al., 2013)), es fácil ver que estas metodologías presentan un problema de gran importancia, el cual reside en que son algoritmos que no son fácilmente aplicables a sistemas de fusión multimodal, lo que en el campo de fusión de imágenes significa sistemas con cámaras de diferentes resoluciones y características. Ya que estas técnicas trabajan a nivel de píxel, requieren previamente que las cámaras estén calibradas y que dicha calibración sea de gran exactitud, ya que en caso contrario presentarán diversos problemas al no utilizar información de las imágenes de entrada que se podría obtener a niveles superiores. Otro problema importante es que, al realizarse a nivel tan bajo, son sensibles al ruido, ya que valores de píxeles distorsionados dentro de una de las fuentes se propagarán a la imagen completa (Li and Yang, 2008).

Algoritmos a nivel de región

Los algoritmos a nivel de características o regiones segmentan las imágenes en regiones para posteriormente fusionar dichas regiones usando diversas propiedades. Como principal inconveniente de estos algoritmos, podemos citar que asumen que las correspondencias entre características de las imágenes a fusionar se conocen previamente (Ardeshir Goshtasby and Nikolov, 2007). Las principales ventajas de estos algoritmos consisten en que son menos sensibles al ruido a nivel de señal y, además, pueden usar reglas de fusión semántica más complejas basadas en las características totales de la imagen en vez de en píxeles individuales o en grupos arbitrarios de los mismos. En Lewis et al. (2007) podemos encontrar una descripción más extensa de estas ventajas, tal y como veremos a continuación:

- *Reglas inteligentes de fusión*: Las reglas de fusión están basadas en combinar los grupos de píxeles que forman las regiones de una imagen. Por tanto, se pueden implementar una mayor cantidad de pruebas para elegir las regiones que constituirán una imagen fusionada, en base a diversas propiedades de las regiones que la componen.

- *Capacidad de resaltar las características de interés*: Las regiones con ciertas propiedades pueden ser acentuadas o atenuadas en la imagen fusionada resultante de acuerdo a las características de dichas regiones.

- *Sensibilidad reducida al ruido*: El proceso de regiones semánticas en lugar de píxeles individuales o regiones arbitrarias puede ayudar a evitar algunos de los problemas que sufren los métodos de fusión a nivel de píxel tales como sensibilidad al ruido, falta de nitidez o pequeños fallos de registro.

- *Capacidad de ayudar al registro y fusión de vídeo*: La información de características extraída de las imágenes de entrada puede usarse para ayudar al registro de las imágenes. Los métodos de fusión de vídeo basados en características pueden usar la estimación de movimiento para realizar un seguimiento de las características fusionadas, permitiendo que la mayoría de los fotogramas puedan ser predichos rápidamente a partir de algunos fotogramas ya fusionados.

Llegados a este punto, es también importante destacar que la segmentación cobra especial importancia debido a que se realiza como paso previo a la fusión. Por ello, este primer paso debe cumplir las siguientes propiedades:

- Todas las características de interés buscadas deben segmentarse como regiones apropiadamente separadas. Si una característica se pierde, puede no estar incluida en la imagen fusionada. Si una característica se parte en más de una región, cada una será tratada por separado, introduciendo ruido en la imagen fusionada.

- El número de regiones obtenidas debe ser tan pequeño como sea posible, ya que el tiempo requerido para calcular la imagen fusionada aumenta al incrementar el número de regiones, al procesarse éstas por separado.

En Davis and Sharma (2007) se describe un esquema estándar de un algoritmo de fusión de información en color y en infrarrojo a nivel de características, mostrado en la Figura 2.9. Podemos ver cómo la segmentación inicial se realiza individualmente en ambos espectros, fusionando la información una vez que las regiones han sido obtenidas. En este algoritmo, el registro inicial de la imagen se realiza mediante análisis multimodal, seleccionando manualmente una serie de puntos de interés en cada imagen y calculando a partir de ellos una matriz de homografía. Tras realizar una segmentación inicial en la que se extraen los contornos de cada espectro mediante resta de fondo, se elabora para cada uno de estos espectros un mapa de los contornos destacados para posteriormente fusionar ambos mapas. Esta fusión se realiza de acuerdo al gradiente de los contornos obtenidos, de forma que los

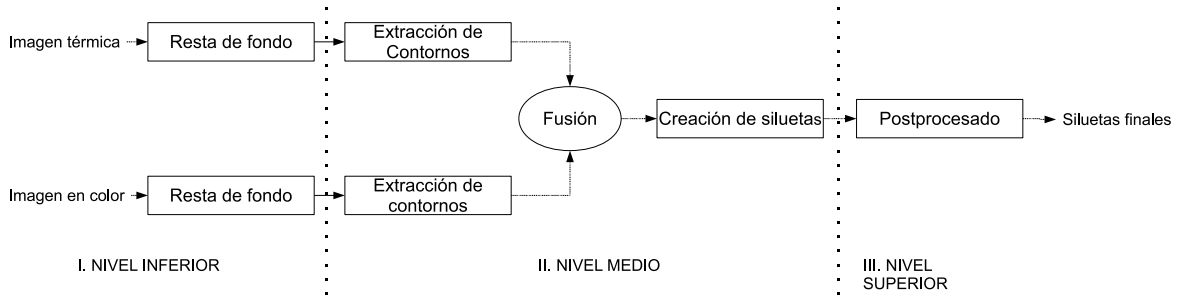


Figura 2.9: Esquema de fusión a nivel de características, correspondiente al algoritmo citado en Davis and Sharma (2007).

contornos excesivamente gruesos se descartan, mientras que en caso de que se obtengan contornos equivalentes en ambos espectros dentro de la misma región de la imagen se utilizará el que tenga mayor gradiente. Esto arroja como resultado aquellos contornos más destacados tanto en sus dominios individuales como en los dos simultáneamente.

Muchos de estos métodos utilizan las técnicas de descomposición que hemos visto en los algoritmos que trabajan a nivel de píxel, tales como MRA o especialmente DWT (Li et al., 1995; Lewis et al., 2007; Nikolov et al., 2000). En estos métodos, las imágenes sin registrar se transforman en primer lugar usando un método de análisis multiresolución. A continuación, las regiones que representan características de interés de la imagen se extraen a partir de los coeficientes de la transformada escogida mediante un método de segmentación, fusionando posteriormente las regiones de acuerdo a sus características. El problema de estos métodos reside, sin embargo, en que las imágenes que se han fusionado de esta forma pueden perder información presente en las imágenes originales debido a la implementación de la transformada inversa (Li and Yang, 2008). Por ejemplo, de forma reciente podemos destacar Wang and Li (2011), donde, tras una segmentación inicial en los espectros infrarrojo y color utilizando el fotograma actual y los tres anteriores, se realiza una descomposición de las regiones de interés obtenidas mediante el algoritmo NSCT, basado en pirámides de descomposición. Las subbandas de menor frecuencia de cada espectro obtenidas serán fusionadas obteniendo la media entre la subbanda correspondiente infrarroja y la de color, mientras que en las de alta frecuencia se tendrá también en cuenta el contraste, dividiendo dichas subbandas en regiones y eligiendo para cada región el valor de la región del espectro con mayor contraste correspondiente. Al final de este proceso, tendremos una nueva composición de subbandas de frecuencia correspondientes a la imagen fusionada, la cual se puede reconstruir aplicando a la inversa el algoritmo NSCT.

Dentro del campo específico de la detección de humanos, existe un enfoque basado en los datos de profundidad. Estas metodologías se fundamentan en la idea de que los humanos destacan dentro de un entorno tridimensional y pueden basarse directamente en los datos tridimensionales estimados para la escena (Hayashi et al., 2004; Haritaoglu et al., 2002; Yang et al., 2004) o indirectamente, para lo que se realiza una combinación de diferentes vistas después de que se hayan extraído las características principales para cada una (Iwase and Saito, 2004; Mittal and Davis, 2003; Yang et al., 2003). Estos métodos surgen debido al problema previamente comentado de que la resta de fondo

puede ser demasiado sensible a cambios de iluminación. Se puede adoptar un enfoque basado en profundidad, modelando el fondo como un modelo de profundidad y comparando dicho modelo con los datos de profundidad de cada fotograma, con el fin de obtener los objetos de interés presentes en el mismo. Sin embargo, los algoritmos de este tipo no suelen funcionar en tiempo real ya que funcionan en estéreo, por lo que se suele utilizar hardware especial. En el trabajo expuesto en Ivanov et al. (2000) se intenta afrontar este problema, evitando el uso de un mapa de profundidad. En su lugar, se aprende el mapeado entre los píxeles entre las dos cámaras, lo que permite una comparación entre los píxeles asociados mediante dicho mapeo entre las dos cámaras. La detección se realiza entonces basándose en la asunción de que el color y la intensidad son similares entre dos píxeles únicamente si representan el fondo. En Zhao and Thorpe (2000) se utiliza la información de profundidad para extraer las siluetas de los individuos en las imágenes. Gracias al entrenamiento de una red neuronal para reconocer humanos de pie, es posible verificar si las siluetas extraídas pertenecen a humanos o no. Para mejorar la robustez del método, se utilizan los gradientes del contorno de la silueta para representar la forma de los humanos. En otra propuesta (Nakada et al., 2008) se utiliza flujo óptico para detectar peatones desde un robot móvil equipado con cámara estéreo. Por su parte, en Suard et al. (2005) se presenta un método de detección de peatones usando visión estéreo y comparación de grafos. Las imágenes se segmentan aplicando el método de corte óptimo de grafo en una sola imagen y computando la disparidad a partir de una pareja de imágenes, comparando los grafos obtenidos. La fase de reconocimiento final se realiza aplicando una máquina de soporte vectorial.

Como ejemplo complementario de estos algoritmos, podemos citar a Li et al. (2013), donde la idea principal reside en que cuando se tienen imágenes de dos cámaras, cada una está enfocada sobre unos objetos distintos. Obteniendo la información de enfoque de cada imagen mediante información de altas frecuencias, se escoge posteriormente la región de cada imagen que mejor está enfocada. Otro ejemplo de estos esquemas de fusión lo podemos encontrar en Zribi (2010), aunque en esta ocasión utiliza la técnica estadística de bootstrap (Efron, 1979), en la que se intenta obtener una idea de la distribución global de la imagen a partir de una muestra aleatoria de la misma. Dicha distribución se obtiene reemplazando los valores desconocidos con los de la distribución observada dentro de los valores de muestra. Tras remuestrear las imágenes en base a esta técnica, se realiza una segmentación inicial mediante un algoritmo de EM no paramétrico basado en bootstrap, para posteriormente fusionar las regiones obtenidas de acuerdo también a dicho algoritmo. Un trabajo reciente basado en bootstrap es el de Hsiao and Leou (2013).

Otra técnica (Iwata et al., 2008) propone una red híbrida de cámaras. El sistema contiene una cámara panorámica y cámaras Pan-Tilt-Zoom (PTZ) para tomar imágenes de rango amplio e imágenes de rostros a una resolución lo suficientemente alta para tareas de identificación. Los métodos de detección robusta de humanos utilizados incluyen un método robusto de resta de fondo, segmentación del color de piel y un método de seguimiento de cámaras. Primero, el sistema detecta personas a partir de una imagen panorámica, y entonces las imágenes detalladas de caras se obtienen con las cámaras PTZ. Las cámaras PTZ pueden seguir las caras usando las cuatro características dimensionales, las cuales corresponden al resultado de aplicar el operador de Prewitt en horizontal, vertical y las dos diagonales obteniendo los gradientes de intensidad en cada una de estas direcciones, y la correspondencia

de relajación, consistente en identificar la posición de la cara optimizando las relaciones espaciales de los puntos que la componen. De forma similar, en Cielniak and Duckett (2004) los humanos son detectados inicialmente en el espectro infrarrojo usando su información de localización obtenida para segmentar entonces la región correspondiente en la imagen en color. En primer lugar, los individuos se detectan basándose en su información térmica, estableciéndose un intervalo de temperaturas en las que se sitúan los humanos. Entonces, cada región de interés que contiene a una persona se divide a su vez en tres subregiones (cabeza, torso y piernas) cuyas características térmicas y de color sirven como entrada a un sistema de reconocimiento de patrones. Para cada sub-región, se almacenan su media y su desviación típica así como su tono, saturación e intensidad (los componentes del espacio de color HSV) conformando un total de 24 características. En otra propuesta (Han and Bhanu, 2007) se realiza una resta de fondo tanto en el espectro infrarrojo como en el visible, extrayendo las siluetas de cada uno de ellos. A continuación, se emplea un algoritmo genético con el fin de encontrar correspondencias entre las siluetas preliminares del color y del infrarrojo. Finalmente, se usan estrategias probabilísticas para obtener mejores resultados a la hora de extraer la silueta del cuerpo.

En Sharma and Davis (2009) se realiza una fusión entre infrarrojo y color en la que, dependiendo de la situación, una cámara de estos espectros se usará como sensor primario y otra como secundario. Para realizar la fusión, se define en primer lugar una representación de características basada en fragmentos de contornos que captura implícitamente la forma de los distintos objetos. Tras proponer un método que genera una distribución de probabilidad a partir de una única pareja de imágenes, se computa la distribución de probabilidad condicional de acuerdo a la afinidad de contornos detectados en ambos espectros basándose en la probabilidad de que los objetos tengan formas regulares y límites continuos. En ese momento, se computa la información mutua entre las características extraídas de ambos sensores, para finalmente obtener un subconjunto de características del sensor secundario que tienen su información mutua más alta con los contornos de objetos proporcionados por el primario. El resultado final se obtiene al solapar los contornos seleccionados de ambos dominios, siendo estos contornos completados y rellenados para crear siluetas.

Por ejemplo, podemos citar el trabajo expuesto en Bertozzi et al. (2007), donde se realiza una fusión entre diversas técnicas de detección de humanos en infrarrojo, utilizando, además, técnicas de estéreo. Así, en primer lugar, se realiza una detección de las zonas cálidas en la imagen. Posteriormente, se obtiene una imagen de disparidad entre la cámara izquierda y la derecha con el fin de identificar los obstáculos, utilizándose, además, la combinación de ambas cámaras con el fin de obtener las distancias entre los humanos delimitados y sus tamaños, así como para poder delimitar mejor sus posiciones. De forma similar, también en Suard et al. (2006) se combinan HOGs con una SVM para distinguir humanos en imágenes estereoscópicas en infrarrojo, utilizando la información estereoscópica para establecer la posición del humano.

Finalmente, se hará especial hincapié en el algoritmo explicado en Kumar et al. (2006) con el fin de ilustrar el uso de técnicas heurísticas a la hora de realizar fusión de imágenes. En dicho artículo, se combinan lógica difusa y filtros de Kalman para modelar el movimiento de los objetos detectados en cada espectro. Para supervisar cada sensor se utiliza un sistema de inferencia difuso para controlar

cada canal y asignar pesos adecuados a la estimación filtrada de cada sensor. El sistema de inferencia se basa en la fiabilidad de los sensores, la cual se estima mediante dos parámetros de entrada: el Ratio de Apariencia (RA), y la Confianza (C). El parámetro RA permite apreciar la fuerza de la segmentación realizada por cada sensor en la instancia actual, mientras que el valor de C muestra la consistencia temporal del sensor a la hora de mantener una buena detección de un objeto en particular, siendo la confianza para un objeto detectado en una sola mancha mayor que la asignada si el objeto se detecta en partes fragmentadas. Basándose en los valores de C y RA, el motor asigna un peso en el intervalo $[0, 1]$ a cada una de las salidas de los filtros de Kalman. Este valor refleja la fiabilidad de la medida del sensor y actúa como un peso que le indica al defuzzificador (el sistema que transformará los datos de lógica difusa en cuantitativos) que unificará estos datos el valor de confianza con el que debe tomar la salida de cada filtro de Kalman. Esta confianza servirá para realizar finalmente una estimación apropiada del movimiento del objeto a monitorizar.

Para calcular RA, se denominará como D la imagen resultado obtenida de la diferencia en valor absoluto entre el fotograma actual y un fotograma de referencia y t como el umbral usado para binarizar D. También se tomará como B_j la mancha j extraída del sensor, con lo que tenemos la ecuación (2.13), donde $|B_j|$ es el número de píxeles de la mancha B_j. El valor de RA es proporcional a la fuerza de las manchas segmentadas por cada sensor, siendo un valor bajo de RA el indicador que la intensidad del píxel en la región de la mancha apenas ha superado el umbral, por lo que el valor de AR se puede comparar para determinar qué sensor aporta mayor cantidad de información.

$$RA(B_j) = \frac{\sum_{x,y \varepsilon B_j} D(x, y)}{|B_j| \times T} \tag{2.13}$$

Para poder calcular la confianza, debemos primero determinar una serie de propiedades, específicamente el solape entre dos manchas y su parecido. El solape entre dos manchas a y b se define como muestran las ecuaciones (2.14) y (2.15), correspondientes al solape máximo $Omax$ y mínimo $Omin$ respectivamente, donde $A(i)$ es el área de la caja que contiene a la mancha i, e $IA(a, b)$ es el área de intersección entre las dos áreas de las cajas que recubren las manchas.

$$Omax(a, b) = \max(IA(a, b)/A(a), \frac{IA(a, b)}{A(b)} \tag{2.14}$$

$$Omin(a, b) = \min(IA(a, b)/A(a), \frac{IA(a, b)}{A(b)} \tag{2.15}$$

El parecido entre las dos manchas se denomina como $R(a, b)$ y se calcula como el grado de correspondencia entre dos manchas usando el solape mínimo *Omin* y un factor de similaridad, tal y como muestra la ecuación (2.16).

$$R(a, b) = Omin(a, b) \times [1 - \frac{[|A(a) - A(b)|]}{\max(A(a), A(b))}] \tag{2.16}$$

Por tanto, podemos calcular la confianza C de una mancha a como indica la ecuación (2.17), donde a es la nueva mancha, b es la mancha anterior, y n el número de manchas anteriores que se correspondían con la actual. Tal y como se puede ver en la ecuación, la confianza en la correspondencia entre $t-1$ y t aumenta si la mancha se ha seguido durante un largo periodo de tiempo y el parecido entre sus instancias anteriores es grande, siendo 1 el valor mínimo, correspondiente a cuando la mancha aparece por primera vez en la escena.

$$C(a) = (\sum_{b=0}^{n} R(a,b) \times C(b)) + 1 \tag{2.17}$$

Las reglas para combinar los sensores se basan en dos consideraciones heurísticas. En primer lugar, si C y RA son grandes para un objeto extraído de un sensor, esto implica que el sensor es muy fiable, mientras que si esos valores se acercan al mínimo, su salida no debe ser tenida en cuenta. Por tanto, usando la regla compuesta de inferencia suma - producto, el motor calcula el peso, lo que le indicará finalmente al defuzzificador con qué nivel de confianza debe tomar la salida de cada sensor. En base a estos pesos para cada sensor, finalmente el defuzzificador realiza una estimación final del movimiento que ha llevado a cabo el objeto.

Algoritmos a nivel simbólico

Los algoritmos a nivel simbólico combinan descripciones de la imagen, por ejemplo, en forma de grafos relacionales. Esta fusión suele combinar descriptores de datos extraídos de diversos sensores, y requiere un alto conocimiento del dominio. En la literatura existen pocos artículos sobre fusión de imágenes realizada a este nivel, si bien podemos, por ejemplo, citar la metodología descrita en Brunn et al. (1996), donde se propone una fusión a alto nivel con el objetivo de realizar reconstrucción tridimensional de edificios a partir de imágenes aéreas. La idea básica consiste en encontrar una descripción simbólica S_O del objeto a partir de N segmentaciones S_i de la imagen $i = 1, 2, ..., N$. Los autores aducen como principal razón para trabajar en el nivel simbólico el poder incluir todos los tipos de conocimiento de dominio específico en todos los niveles del análisis. El proceso consiste en realizar, en primer lugar, una extracción de características (puntos, líneas y regiones), y agrupar dichas características en un grafo. De dicho grafo se derivan estructuras locales, que se denominan vértices, alas y celdas (correspondientes a puntos, líneas y regiones, respectivamente), siendo los primeros y los últimos los más apropiados para establecer correspondencias. La reconstrucción consistirá en que un vértice contenga un punto tridimensional, dos o tres bordes vecinos y las caras entre dichos bordes, usando el conocimiento específico del dominio (lo que caracteriza este método como fusión a alto nivel), imponiendo que los vértices deban unir tres celdas. Finalmente, se realiza una reconstrucción de celdas para unir la información encontrada hasta el momento, realizando hipótesis sobre las regiones superficiales contenidas en los diversos planos y usando la retroproyección de las imágenes sobre las superficies precisas y la información radiométrica de las imágenes para encontrar elementos consistentes de superficie sobre los planos predichos.

Es importante destacar que estos niveles de fusión pueden combinarse de forma que un nivel

fusione los datos de los niveles anteriores con sensores nuevos, tal y como podemos apreciar en la Figura 2.10. En esta figura se muestra un sistema estándar de vigilancia cuyo objetivo es la detección de humanos. En el nivel más bajo, el sistema reconoce que se han detectado objetos mediante ondas de dos radares. Los datos de posición arrojados por las ondas se fusionan componiendo una imagen inicial que muestra de forma aproximada las posiciones de los objetos detectados. Esa imagen es comparada con la arrojada por una cámara en color, realizando una fusión a nivel de píxel con el fin de poder contrastar las posiciones de los objetos detectados con las que detecta la cámara. Una vez que se ha realizado esta fusión, se extraen las principales características de los objetos detectados, y se comparan con las características de los humanos detectados mediante segmentación de las imágenes adquiridas por una cámara en infrarrojo. En base a esta fusión a nivel de características, se extrae una probabilidad de que en la posición comprobada efectivamente haya un humano ($0{,}8$). Esta probabilidad se compara con los datos de un micrófono omnidireccional que detecta ruido en la escena, y que mediante un algoritmo de detección de humanos a nivel de audio calcula la probabilidad de que si hay ruido en la escena, ese ruido se deba a presencia humana, arrojando una probabilidad de $0{,}2$ sobre 1. Estas probabilidades se unifican en una fusión a nivel simbólico, la cual arroja la probabilidad definitiva de que en la escena haya un humano en la posición predicha por el radar y ampliada por los diversos mecanismos de adquisición de datos.

2.2.3. Resumen y conclusiones

En esta sección se ha deseado dar una visión general del estado del arte actual en fusión de imágenes, comenzando con una perspectiva general de las características de un sistema de fusión multisensorial. Posteriormente, se ha particularizado en las fases que constituyen una metodología de fusión de imágenes, comenzando con una fase inicial en la que se realiza un alineado espacial y temporal de las mismas. Esta primera fase es crítica, ya que un error aparecido en esta fase se propagará por todo el sistema, afectando a todo el funcionamiento del mismo. En cuanto a la alineación espacial, se ha visto que es un campo de estudio muy amplio con una gran cantidad de aproximaciones posibles, debido a la gran cantidad de tipos de cámaras, entornos y objetivos con los que se puede realizar la fusión. Pueden encontrarse incluso estudios centrados en esta fase como Zitová and Flusser (2003), pudiéndose comprobar que los pasos para la alineación temporal de imágenes descritos en dicho estudio son aplicables a todos los algoritmos de alineación estudiados. También se ha realizado una breve descripción de las principales tendencias a la hora de realizar una alineación temporal de las imágenes a fusionar, si bien el uso de estas técnicas no es el principal campo de estudio de esta tesis.

Una vez que las imágenes se encuentran en la misma escala y ejes espacial y temporal, se procede a normalizar los datos de las mismas, ya sea a nivel de píxel, de región o simbólico. Esta segunda fase es el núcleo de la mayoría de algoritmos de fusión de imágenes encontrados, los cuales muchas veces dan por supuesto que la alineación se ha realizado antes de proceder a la metodología descrita.

La fusión de imágenes a nivel de píxel es la más encontrada en la literatura, existiendo múltiples aproximaciones. Si bien la más simple de ellas consiste en elaborar un promedio de las imágenes a fusionar, las más extendidas abogan por realizar una descomposición a nivel de frecuencias o resolu-

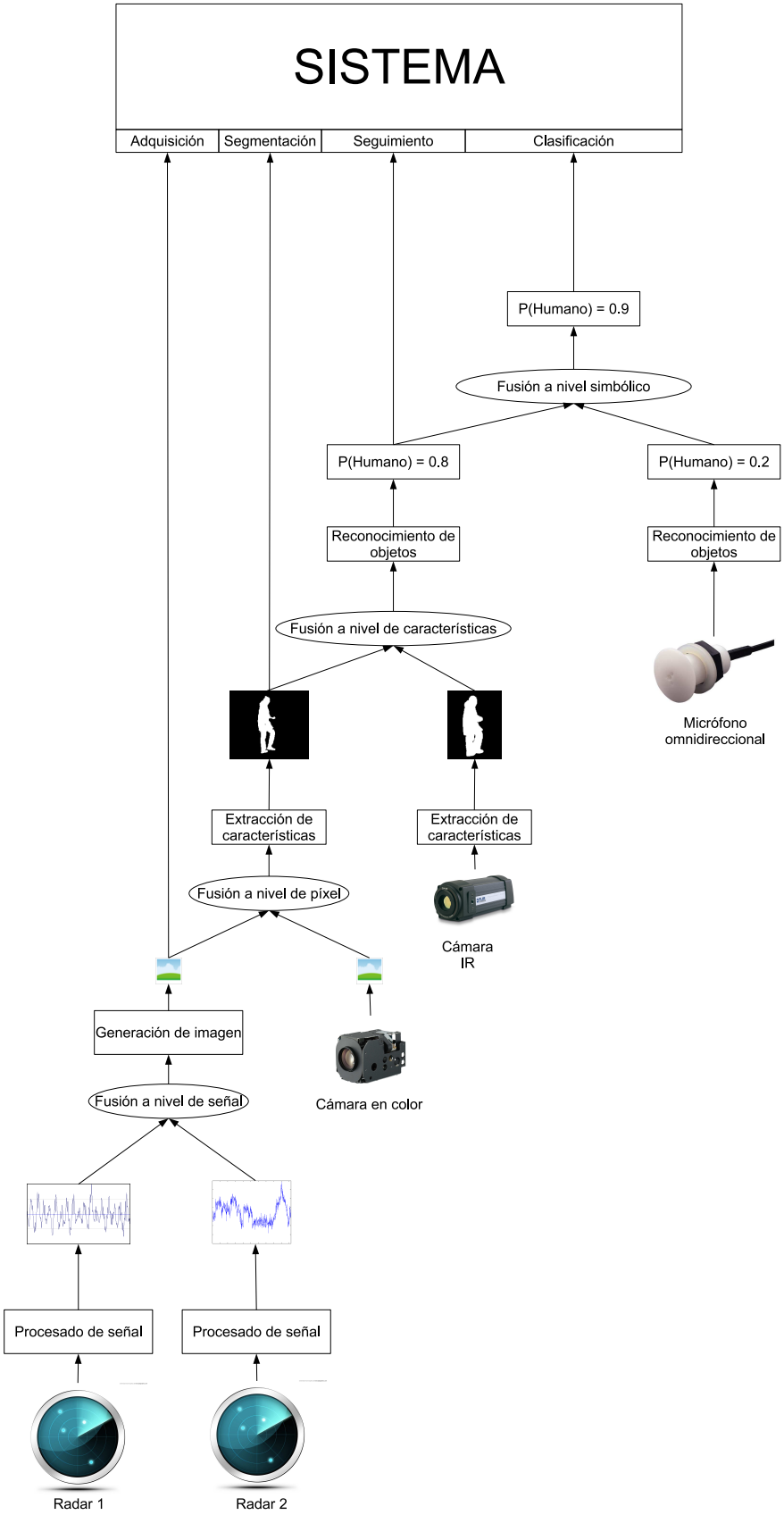


Figura 2.10: Esquema de un sistema que combina todos los niveles de fusión, adaptado de Processing and Communication Research Lab (2008)

ción de cada imagen en diversas capas, componiéndose la imagen fusionada en cada nivel con la capa más adecuada de una de las imágenes descompuestas, de forma que el resultado es una combinación de diferentes capas de las diversas imágenes descompuestas.

En la fusión a nivel de regiones, que utiliza datos a más alto nivel de las diversas imágenes a fusionar, se ha visto que una de las tendencias más utilizadas también consiste en realizar una descomposición de las imágenes a fusionar para posteriormente recomponer también por capas, si bien en este caso muchas veces no se descompone toda la imagen, sino únicamente las regiones de interés obtenidas por una segmentación realizada previamente.

También se ha observado que este nivel de segmentación es utilizado en muchas ocasiones para fusión de imágenes multimodales (encontrándose muchos casos de fusión de imágenes multiespectrales a este nivel), debido a que sus requisitos para realizar la alineación de imágenes son menos exigentes y se puede utilizar la información de características de las regiones que nos proporciona cada espectro. Así mismo, se ha podido encontrar poca literatura sobre fusión de imágenes a alto nivel. Esto es debido a que este nivel requiere un alto uso de la información a nivel del dominio en el que se desea realizar la fusión, mientras que la mayoría de los algoritmos estudiados buscan el poder aplicar las técnicas de fusión dentro del mayor número de escenarios posible.

Finalmente, en la tabla 2.3 se puede ver un resumen de los algoritmos de fusión estudiados en cada paso descrito de la fusión de imágenes.

2.3. Seguimiento de humanos en vídeo

La fase final del sistema que se propondrá en la presente tesis consistirá en un algoritmo de seguimiento de humanos basado en los resultados obtenidos de la fusión de imágenes previa. En la presente sección se realiza inicialmente una breve definición y primera aproximación a las problemáticas que se pueden encontrar en este campo, para posteriormente incidir en los diversos enfoques que se pueden utilizar para el seguimiento, tanto a nivel de representación como de correspondencia entre los humanos encontrados en el fotograma actual y aquellos detectados previamente en la secuencia. Finalmente, se llevará a cabo un resumen final con el fin de extraer las tendencias más importantes que se pueden encontrar en la literatura referidas a este campo.

2.3.1. Definición de seguimiento

En su forma más simple, el seguimiento se puede definir como el problema de realizar la estimación de la trayectoria seguida por un objeto en el plano de la imagen conforme se mueve sobre una escena (Yilmaz et al., 2006). En otras palabras, un sistema de seguimiento asigna etiquetas consistentes a los objetos rastreados en diferentes fotogramas de un vídeo. En Moeslund et al. (2006) podemos encontrar otra definición de la funcionalidad de estos sistemas, explicándose que su objetivo es, dado el estado de N personas en los fotogramas previos, hallar el estado de las mismas personas en el fo-

Tabla 2.3: Resumen de los algoritmos de fusión estudiados para cada etapa de la misma

Etapa	**Nivel**	**Tipo**	**Artículos**
Alineación	Espacial	Multitemporal	(Petrovic and Xydeas, 2004), (Perperidis et al., 2005)
		Multivista	(Jain and Ross, 2002), (Thompson and Wettergreen, 2005), (Starck and Hilton, 2008)
		Multimodal	(Chen et al., 2003), (Krotosky and Trivedi, 2006), (Sharma and Davis, 2009), (Leykin and Hammoud, 2010),
		Registro de escena a modelo	(Chouteau et al., 2007)
		Mixto	(Harmouche et al., 2010)
	Temporal	DTW	(Rabiner and Juang, 1993), (Kale et al., 2003), (Ratanamahatana, 2005)
Normalización	Nivel de píxel	Descomposición	(Petrovic and Xydeas, 2004), (Meytlis and Sirovich, 2007), (Correa et al., 2008), (Li et al., 2013)
		Otras	(Pszczólkowski and Soto, 2007), (Jang and Ra, 2008)
	Nivel de región	Descomposición	(Li et al., 1995), (Nikolov et al., 2000),(Lewis et al., 2007), (Wang and Li, 2011), (Wang et al., 2013), (Saeedi and Faez, 2013)
		Profundidad	(Ivanov et al., 2000), (Zhao and Thorpe, 2000), (Haritaoglu et al., 2002), (Mittal and Davis, 2003), (Yang et al., 2003), (Hayashi et al., 2004), (Iwase and Saito, 2004), (Yang et al., 2004), (Suard et al., 2005), (Nakada et al., 2008)
		Otras	(Cielniak and Duckett, 2004), (Kumar et al., 2006), (Suard et al., 2006), (Bertozzi et al., 2007), (Davis and Sharma, 2007), (Han and Bhanu, 2007), (Iwata et al., 2008), (Sharma and Davis, 2009), (Zribi, 2010), (Li et al., 2013), (Hsiao and Leou, 2013)
	Nivel simbólico	Grafo	(Brunn et al., 1996)

tograma actual. En este caso, normalmente el estado es la posición en la imagen de una persona, pero puede contener otros atributos, tales como el color (Hu et al., 2004), la forma (Krüger et al., 2005), etc. El autor establece dos pasos a la hora de realizar seguimiento de humanos:

1. Segmentación de figuras: En este ámbito se puede definir como el proceso de separar los objetos

de interés (humanos) del resto de la imagen (fondo). Los métodos de segmentación de figuras se aplican a menudo como primer paso en muchos sistemas, resultando un proceso crucial.

2. Correspondencia temporal: Proceso de asociar los humanos detectados en el fotograma actual con aquellos de los fotogramas previos, proporcionando trayectorias en el espacio a través del tiempo, esto es, dado el estado de N personas en los fotogramas previos y los actuales fotogramas de entrada, establecer los estados de las mismas personas en los fotogramas actuales.

El seguimiento de objetos o personas puede ser complejo debido, entre otros factores, a la pérdida de información causada por proyección del mundo tridimensional a una imagen bidimensional, los movimientos complejos de los objetos (problema especialmente acentuado en el caso de los humanos, cuyos comportamientos no tiene por qué seguir un patrón preestablecido) y la naturaleza no rígida o articulada de los objetos. También encontramos problemas derivados de la propia escena, tales como las oclusiones parciales y completas que se pueden producir con los elementos presentes en la escena o entre los propios objetos y los cambios de iluminación que puede haber en la escena, junto con el ruido que puede haber en la secuencia de imágenes que se está capturando. A todo esto se suma el agravante de que una aplicación de vigilancia debe ser capaz de funcionar en tiempo real, lo que obliga a buscar técnicas computacionalmente eficientes para realizar el seguimiento de los objetos.

En la literatura se pueden encontrar numerosas propuestas para el seguimiento de objetos. Estas difieren primariamente unas de otras basándose en las formas en que afrontan las siguientes cuestiones:

- ¿Qué representación del objeto es adecuada para el seguimiento?
- ¿Qué características de la imagen deben usarse?
- ¿Cómo deberían modelarse el movimiento, apariencia y forma de los objetos?

Las respuestas a estas preguntas dependen del contexto y entorno en que el seguimiento se realiza y el uso final para el que se busca la información del seguimiento. A la hora de realizar el presente estudio, nos centraremos en los diversos mecanismos utilizados tanto a la hora de realizar la *representación* de los objetos como de establecer la *correspondencia temporal* entre sus estados predichos y los hallados en el último fotograma procesado.

2.3.1.1. Representación

Antes de comenzar el proceso de seguimiento, se debe construir un modelo de cada individuo, centrándose muchos métodos en hacer este proceso de forma automática. En un escenario de seguimiento, un objeto puede definirse como cualquier elemento que sea de interés para un análisis posterior. Entre las representaciones de la forma de los objetos más comúnmente utilizadas podemos destacar:

- *Puntos*: En este enfoque, el objeto se representa mediante un punto, siendo el más comúnmente elegido el centroide (en la Figura 2.11 (a))(Veenman et al., 2001). Otras alternativas usan un conjunto de puntos (en la Figura 2.11 (b)), como por ejemplo aquellos correspondientes a bordes de objetos presentes en la imagen (Serby et al., 2004). En general, esta representación es adecuada para el seguimiento de objetos que ocupan pequeñas regiones en una imagen. Como ejemplo, podemos citar Weng et al. (2006), donde el usuario selecciona en primer lugar un objeto en movimiento, extrayéndose su color dominante y localizándose el conjunto de puntos que pertenecen inicialmente al objeto mediante una resta de fotogramas en un radio de 50 píxeles respecto al punto delimitado por el usuario.

- *Formas geométricas primitivas*: La forma del objeto es representada mediante un rectángulo, elipse (en la Figura 2.11 (c), (d) (Comaniciu et al., 2003)), etc. El movimiento del objeto para esa representación es normalmente modelado por traslación o transformación homográfica. Aunque las formas geométricas primitivas son más apropiadas para representar objetos simples rígidos, también se usan para representar objetos no rígidos. Por ejemplo, en Zhou et al. (2009) se explica un algoritmo para seguimiento de objetos en escenas complejas, consistente en adaptar de forma óptima la elipse que resalta los objetos de interés. Técnicamente, se pretende reducir los residuos entre la distribución de probabilidad estimada y la esperada, lo que repercute en que forma de la elipse se puede adaptar en la etapa de seguimiento. Por otra parte, en Atsushi et al. (2002) se modela la pose de los humanos en el fotograma anterior por medio de una elipse, mientras que en el fotograma actual se predicen nueve posibles poses del humano.

- *Silueta y contorno del objeto*: La representación del contorno define los límites de un objeto (como en la Figura 2.11 (g), (h)). La región dentro del contorno es conocida como la silueta del objeto (ver Figura 2.11 (i)). Las representaciones de la silueta y el contorno son adecuadas para el seguimiento de formas complejas no rígidas (Yilmaz et al., 2004). Por ejemplo, en Li et al. (2009) se propone un método para detección y seguimiento rápido de humanos basado en las características de la forma de la cabeza y los hombros de las personas, ya que este conjunto presenta una forma similar a la letra griega omega (Ω). En Haritaoglu et al. (2000); Hu et al. (2006); Yang et al. (2004) se buscan los candidatos a cabezas mediante el análisis de los límites de la silueta y la proyección vertical del histograma de la silueta. Asimismo, en Zhao and Nevatia (2003) encontramos una propuesta similar con la diferencia de que se aplican métodos basados en bordes para encontrar el conjunto cabeza-hombros dentro de las siluetas.

- *Modelos articulados de forma*: Los objetos articulados se componen de partes del cuerpo que se mantienen juntas mediante conexiones (ver Figura 2.11 (e)). Como ejemplo de estas técnicas, podemos citar Benezeth et al. (2008) , donde el seguimiento se realiza buscando intersecciones de las componentes conexas entre el fotograma actual y el anterior.

- *Modelos de esqueleto*: Este modelo se usa normalmente como representación de forma para reconocer objetos (Ali and Aggarwal, 2001), pudiendo ser usada para modelar tanto objetos articulados como rígidos (ver Figura 2.11 (f)).

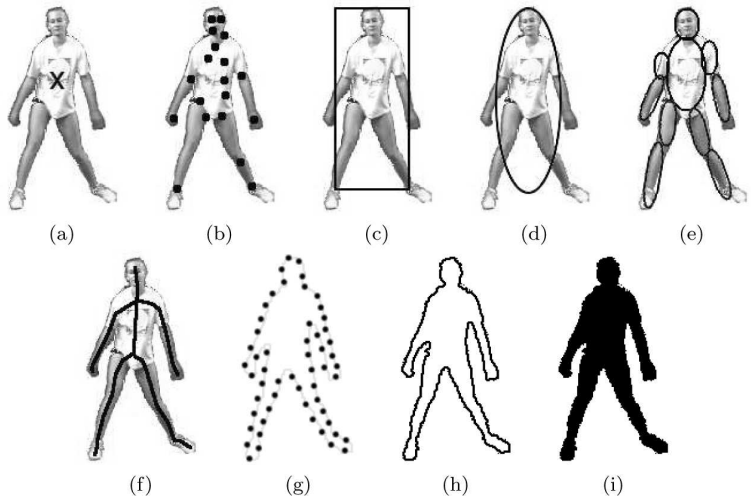


Figura 2.11: Representaciones de objetos, extraída de Moeslund et al. (2006). (a) Centroide (b) Múltiples puntos (c) Contorno rectangular (d) Contorno elíptico (e) Formas basadas en múltiples partes (f) Esqueleto del objeto (g)Puntos de control en el contorno del objeto (h) Contorno completo del objeto (i)Silueta del objeto.

Al igual que en otras clasificaciones que hemos visto, estas representaciones no son rígidas, sino que se pueden combinar entre sí con el fin de obtener mejores resultados. Por ejemplo, en Gouet-Brunet and Lameyre (2008) se presenta una aproximación para segmentación y reconocimiento de objetos en escenas muy recargadas, basado en características visuales heterogéneas. Se propone el uso conjunto de características complementarias de naturalezas diferentes. Por una parte, un conjunto de descriptores locales basados en puntos de interés, por otra, un descriptor global basado en una serpiente, proporcionando una descripción a alto nivel de la forma del objeto.

2.3.1.2. Correspondencia temporal

Una vez que ha comenzado el seguimiento y tenemos una representación, el problema reside en encontrar las correspondencias temporales entre los estados predichos y los medidos (Moeslund et al., 2006). Entre las diversas aproximaciones a la hora de encarar este proceso, podemos citar de forma general el uso de matrices de correspondencia, las cuales tienen en un eje los objetos predichos y aquellos medidos en el otro. Para cada elemento de la matriz se calcula la distancia entre la posición obtenida del objeto y su posición predicha. A partir del análisis de las columnas y filas se pueden predecir situaciones tales como la aparición de un nuevo objeto, la pérdida de uno existente, la correspondencia entre dos objetos y situaciones de disolución y reunión de grupos, siendo estos dos últimos casos especialmente complejos. Este enfoque puede estar sujeto a diversas optimizaciones.

A continuación, estudiaremos de forma específica los diferentes métodos de establecer correspondencias temporales en base a la representación de los objetos elegida.

Figura 2.12: Ejemplo de seguimiento basado en puntos, extraído de EMSI (2007)

Correspondencia basada en puntos

El seguimiento se puede formular como la correspondencia entre fotogramas de objetos detectados representados mediante puntos. La correspondencia por puntos es un problema complicado (Yilmaz et al., 2006), especialmente cuando se producen oclusiones, falsos positivos y entradas y salidas de objetos. En general, los métodos de correspondencia entre puntos se pueden dividir en dos categorías: métodos estadísticos y deterministas. Dentro de la clasificación de esquemas de representación que se ha hecho en la anterior sección, estos métodos usan las representaciones basadas en puntos (ver figuras 2.11a y 2.11b). En la Figura 2.12 podemos ver un ejemplo de un algoritmo de seguimiento mediante correspondencia basada en puntos, en este caso tomando como punto característico el centroide de un avión.

Los métodos deterministas definen un coste para asociar cada objeto en el fotograma $t-1$ a un único objeto en el fotograma t usando un conjunto de restricciones de movimiento. La minimización del coste de correspondencia se formula como un problema de optimización combinatoria, definiéndose normalmente el coste de la correspondencia entre fotogramas mediante una combinación de criterios de proximidad entre las localizaciones del objeto, la velocidad del mismo, su tipo de movimiento, etc. Entre estos métodos, podemos citar Fuentes and Velastin (2006) donde se caracteriza al objeto mediante su centroide y se utiliza la matriz de correspondencia explicada previamente para poder establecer situaciones de entrada de humanos en la escena, salida, creación de grupos, etc. Por su parte, en Shafique and Shah (2005) se pretende utilizar la coherencia temporal de la velocidad y posición de los humanos, formulando la correspondencia como un problema de teoría de grafos, representando cada detección del objeto en un fotograma como un nodo de un grafo con el objetivo de encontrar el mejor camino único.

Por su parte, los métodos estadísticos intentan mitigar los problemas del ruido producido por los sensores de vídeo o los cambios súbitos que pueden experimentar los movimientos de los objetos. Para ello, estos métodos tienen en cuenta las medidas e incertidumbre del modelo durante la estimación del estado del objeto, usando el enfoque del espacio de estados para modelar las propiedades de los objetos tales como posición, velocidad y aceleración. Estas propiedades normalmente consisten en la posición del objeto en la imagen, la cual se obtiene mediante un mecanismo de detección. Entre los métodos estadísticos para la estimación del estado de un único objeto podemos destacar los filtros de Kalman, los filtros de partículas y los métodos de Montecarlo.

Los *métodos de Montecarlo* (Metropolis and Ulam, 1949) describen las distribuciones de probabilidad como conjuntos de muestras ponderadas en el espacio de estados. Estas muestras se utilizan para simular inferencia probabilística a través de la regla de Bayes, la cual se aplica cuando se desea calcular la probabilidad condicional de un evento anterior en base a otro que ocurrió posteriormente. Los métodos de Montecarlo aproximan las distribuciones de probabilidad de las secuencias utilizando grandes conjuntos de muestras, llamados partículas, las cuales se propagan en el tiempo utilizando mecanismos de muestreo por importancia y de remuestreo. Estos métodos son adecuados para problemas donde los modelos de transiciones y de observación son altamente no lineales, debido a que los métodos basados en muestreo pueden representar densidades de probabilidad muy generales. En particular, las funciones de densidad multimodales o de múltiples hipótesis se ajustan bien a las técnicas de Montecarlo, cubriendo además estos métodos el espacio entre los métodos de fusión de datos paramétricos y los basados en rejillas. Sin embargo, los métodos de Montecarlo no se ajustan a problemas donde el espacio de los estados es altamente dimensional.

Otra aproximación probabilística son los *modelos ocultos de Markov* (en inglés Hidden Markov Models, HMM), capaces de modelar la evolución temporal, combinándolos con otros modelos probabilísticos como las redes bayesianas (en inglés, Bayesian Networks, BN). Su estructura representa un grafo dirigido que se extiende para adaptarse a la evolución temporal de las trayectorias. Así, cada HMM llevará asociada una estructura de grafo donde una variable representará el estado oculto, y otra representará la observación (vista en un instante de tiempo). Estos instantes de tiempo (nodos) representan la evolución del objeto descrito por el modelo a lo largo del tiempo. En cada instante existen arcos internos que indican la dependencia de la variable observada.

Los *filtros de Kalman* (Kalman, 1960) se basan en un estimador lineal recursivo que calcula sucesivamente una estimación para un estado continuo en base a observaciones periódicas del estado actual. Esta metodología emplea un modelo estadístico explícito que modela la evolución del parámetro de interés a través del tiempo, mientras que otro modelo se encarga de la relación del parámetro con las observaciones. Un filtro de Kalman se ejecuta recursivamente en dos etapas: predicción y actualización. La etapa de predicción estima el estado actual en base al anterior, siendo este estado predicho conocido como estimación a priori al no incluir aún la información de la observación actual. En la fase de actualización, la predicción a priori se combina con la información observada para refinar la estimación del estado, conociéndose el resultado obtenido estimación a posteriori.

Una limitación de los filtros de Kalman es que asumen que las variables de estado se ajustan a una distribución normal (gaussiana). Por tanto, estos filtros ofrecerán estimaciones defectuosas a la hora de abordar variables que no siguen una distribución gaussiana. Esta limitación puede superarse mediante el uso de *filtros de partículas* (Gordon et al., 1993). En los filtros de partículas la densidad condicional de estados $p(X_t \mid Z_t)$ en un instante t se representa mediante un conjunto de muestras $\{s_t^{(n)} : n = 1, ..., N\}$ (partículas) con pesos $\Pi_t^{(n)}$ (probabilidad de muestreado). Estos pesos definen la importancia de una muestra, es decir, su frecuencia de observación (Isard and Blake, 1998). Para reducir la complejidad computacional, por cada tupla $(s^{(n)}, \pi^{(n)})$ se almacena un peso acumulado $c^{(n)}$, donde $C^{(N)} = 1$. Las nuevas muestras en el instante t se extraen de $S_{t-1} = \{(s_{t-1}^{(n)}, \Pi_{t-1}^{(n)}, c_{t-1}^{(n)}) : n = 1, ..., N\}$ en el instante previo $t - 1$ de acuerdo a los diferentes esquemas de muestreo (MacKay, 1998). Sin embargo, estas aproximaciones asumen una sola medida en cada instante de tiempo, esto es, únicamente se estima el estado de un solo objeto. El seguimiento de múltiples objetos requiere una solución conjunta de asociación de datos y problemas de estimación de datos. De forma reciente, en Fu and Han (2012) se propone un método que utiliza una estimación ponderada del centroide de un objeto de acuerdo a la probabilidad de que los píxeles de la región que lo delimita pertenezcan realmente al objeto, utilizando posteriormente filtros de Kalman para predecir la próxima posición del objeto y resta de fondo para detectar los objetos en movimiento en la imagen.

Cuando se realiza un seguimiento de múltiples objetos usando filtros de Kalman o de partículas, es necesario asociar de forma determinista la medida más probable para un objeto en particular al estado de este objeto, es decir, el problema de correspondencia se debe resolver antes de poder aplicar estos filtros. El método más simple consiste en usar el vecino más cercano, aunque siempre puede darse una correspondencia incorrecta si los objetos están próximos entre sí.

Una alternativa la constituye el filtro de probabilidad conjunta de asociación de datos (en inglés Joint Probability Density Association Filter, JPDAF), propuesto en Fortmann et al. (1983), que asocia a cada objeto todas las posibles medidas obtenidas en la escena teniendo en cuenta la distribución estadística de los errores obtenidos hasta el momento y asumiendo que una medida puede estar asociada a más de un objeto. Por su parte, el seguimiento de múltiples hipótesis (en inglés, Multiple Hypothesis Tracking, MHT), ideado por Reid (1979), considera que se pueden mejorar los resultados del seguimiento si la decisión de la correspondencia no se toma hasta que no se hayan examinado varios fotogramas. El algoritmo MHT mantiene varias hipótesis de correspondencia para cada objeto en cada fotograma. La trayectoria final del objeto será el conjunto de correspondencias más probables durante el periodo de tiempo de la observación. El algoritmo tiene la habilidad de crear nuevas trayectorias para objetos que entran al campo de visión y finalizarlas para los objetos que salen del mismo, pudiendo además encargarse de oclusiones, esto es, prolongar las trayectorias incluso si falta alguna de las medidas de un objeto.

Se puede decir que los algoritmos de seguimiento basados en puntos son apropiados para seguir a objetos muy pequeños que pueden ser representados por un solo punto, necesitándose más puntos para seguir a objetos de mayor tamaño. En el contexto de seguir a objetos usando múltiples puntos, el agrupado automático de puntos que pertenecen al mismo objeto es un problema importante, debido a

la necesidad de distinguir entre múltiples objetos, así como entre los mismos y el fondo. En Benezeth et al. (2010) se soluciona este problema construyendo para cada mancha obtenida en la segmentación una serie de puntos de interés y forzando que cada punto de interés pertenezca a un píxel perteneciente a un objeto de interés obtenido en la segmentación imponiendo además que tenga una distancia mínima con otros puntos pertenecientes al mismo objeto, con el fin de tener una distribución espacial homogénea de los mismos. Posteriormente, se elabora una matriz de correspondencias similar a los enfoques vistos previamente.

Correspondencia basada en semillas

El seguimiento mediante semillas se realiza normalmente para calcular el movimiento de un objeto de un fotograma al siguiente, siendo el objeto representado mediante una región primitiva. Con la representación basada en semillas, el movimiento calculado define implícitamente la región del objeto así como su orientación en el siguiente fotograma, ya que para cada punto del objeto en el fotograma actual se puede determinar su localización en el siguiente fotograma de acuerdo al modelo de movimiento estimado. Dependiendo del contexto en que se usen estos métodos, solo una de estas propiedades puede ser especialmente importante. Por ejemplo, en caso de analizar el comportamiento de un objeto basándose en su trayectoria, solo el movimiento es adecuado. Sin embargo, para identificarlo, también es importante la región en que está contenido. Los algoritmos de este tipo se diferencian entre sí en la representación de la apariencia usada, el número de objetos a seguir y el método usado para estimar el movimiento del objeto, pudiéndose dividir en modelos basados en apariencia y plantillas, y en modelos de apariencia multivista (Yilmaz et al., 2006). Dentro de los esquemas de representación que hemos descrito anteriormente, estos métodos usan las representaciones basadas en regiones geométricas primitivas (ver figuras 2.11c y 2.11d).

Para seguir a un solo objeto, el enfoque más común consiste en correspondencia de plantillas, el cual es un método de fuerza bruta que busca en la imagen una región similar a la plantilla del objeto definida en el fotograma anterior, calculando la posición de la plantilla en la imagen actual mediante una medida de similaridad. Normalmente, para formar las plantillas, se usan desde características de la intensidad de la imagen o el color hasta los gradientes de la imagen. Una limitación de la correspondencia mediante plantillas es el alto coste computacional debido a la búsqueda mediante fuerza bruta. Para limitar el coste computacional, normalmente se limita la búsqueda del objeto a la vecindad de su posición previa.

En lugar de plantillas se pueden usar representaciones alternativas como histogramas de color o modelos mixtos, calculados mediante el uso de la apariencia de píxeles en el interior de las regiones rectangulares o elipsoidales. Un problema de estas aproximaciones es que requieren que alguna parte del objeto se encuentre dentro de la forma elegida, cuya localización esté definida por la posición anterior del objeto. Para eliminar este requisito, se pueden usar filtros de Kalman (como se hace en O' Malley et al. (2010) o en Xu and Puig (2005)) o partículas (como ocurre en Frintrop et al. (2010)) para predecir la posición del objeto en el siguiente fotograma. Dado el estado del objeto definido en términos de velocidad y aceleración del centroide del objeto, estos filtros estiman la posición del centroide del objeto de forma que aumente la probabilidad de observar parte del objeto dentro de la región.

Una de las limitaciones del uso de formas primitivas geométricas para la representación de los objetos es que algunas partes pueden quedarse fuera de la forma previamente definida mientras que partes del fondo se encuentren dentro. Estos fenómenos pueden apreciarse tanto en objetos rígidos (al cambiar la pose del objeto) como no rígidos (cuando el movimiento local da lugar a cambios en la apariencia del objeto). En estos casos, el movimiento del objeto estimado al maximizar la similaridad del objeto puede no ser correcto. Para evitar estos problemas, se puede forzar que la región se encuentre dentro del objeto en vez de encapsular su forma completa, siendo otra alternativa modelar la apariencia del objeto mediante funciones de probabilidad conjunta de color o textura y asignando pesos a los píxeles dentro de la forma primitiva de acuerdo a la probabilidad condicional del color/textura observado. Un ejemplo del uso de estas características lo podemos encontrar en Schiele (2006), donde se lleva a cabo un seguimiento combinando características de forma, movimiento y color, usando una variación del algoritmo de Viterbi (el cual busca encontrar la secuencia de estados ocultos más probable a partir de una secuencia de observaciones) para predecir qué regiones de la imagen actual corresponden a los objetos ya observados. Por su parte, en Zhou and Hoang (2005) se usan el histograma de color, la dirección, la velocidad y el número de píxeles y tamaño del modelo humano para describir a los humanos, asumiéndose que cada humano siempre se mueve en una dirección y velocidad similares. Si la persona no se mueve durante un periodo de tiempo, se comprueba si la detección era correcta, aprendiéndose en caso negativo las circunstancias de la falsa alarma y ajustando el fondo.

En Kelly et al. (2009) podemos encontrar una propuesta basada en grafos bipartitos ponderados donde se utilizan características del histograma para realizar la correspondencia. Como uso innovador de estos métodos, podemos destacar Wang et al. (2012b), donde se añaden características de forma (normalmente usadas en el siguiente enfoque que veremos, correspondencia basada en siluetas) a este tipo de algoritmos. Para ello, se muestrean una serie de puntos de forma uniforme a lo largo de la elipse que recubre al objeto. De esta forma, se obtiene un vector de radios con las distancias entre cada punto y el centroide del objeto. Una vez que tenemos este vector, se puede utilizar como característica la similaridad entre la forma aprendida para el modelo actual y la del objeto detectado. En la Figura 2.13 se puede ver un ejemplo de seguimiento basado en plantillas aplicado a seguimiento de caras, donde, una vez establecida la plantilla (en la esquina inferior izquierda de cada fotograma), se buscan zonas en la imagen similares a ésta.

En el seguimiento mediante modelos de apariencia multivista debe tenerse en cuenta que las vistas de un objeto pueden cambiar durante el proceso de seguimiento, por lo que el modelo de apariencia de un objeto puede dejar de ser válido y perderse su trayectoria. Para solucionar este problema, se pueden aprender varias vistas de un mismo objeto offline y usarlas para el seguimiento. Por ejemplo, en Fleuret et al. (2008) se propone un algoritmo para seguir a múltiples personas en múltiples vistas de cámara, usando un mapa probabilístico de las localizaciones de los diferentes individuos, unido a un algoritmo de programación dinámica que sigue a cada persona de forma aislada, usándose tanto un modelo de apariencia como uno de movimiento para describir los objetos que se están siguiendo.

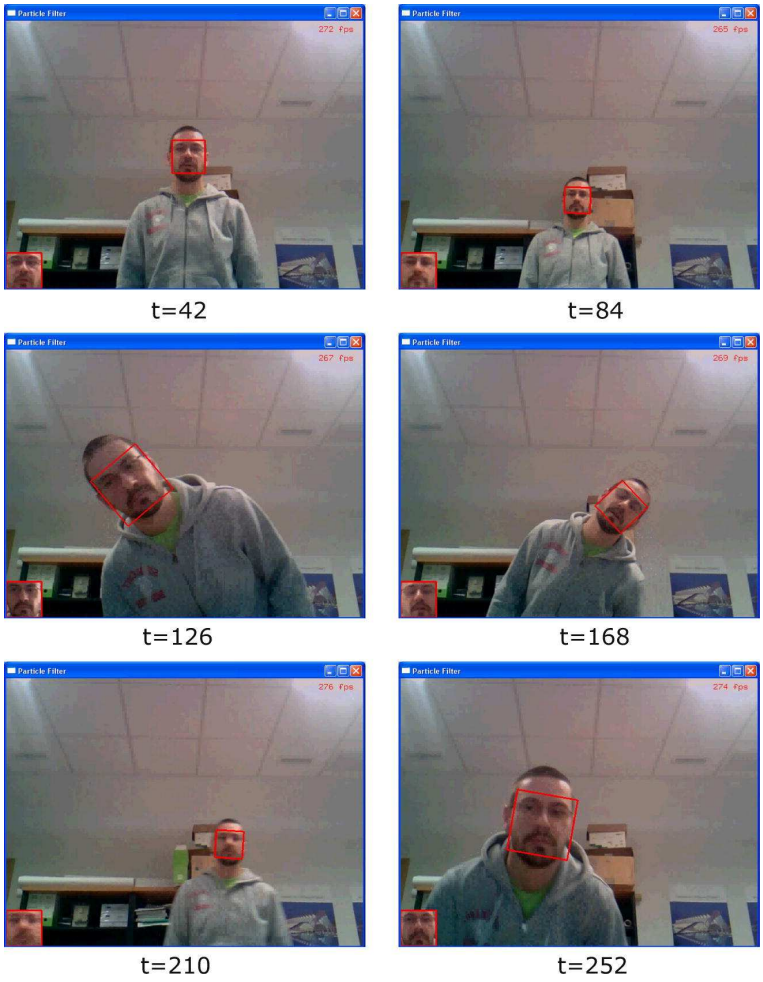


Figura 2.13: Ejemplo de seguimiento basado en plantillas extraído de Cabido et al. (2012). La plantilla utilizada se puede ver en la esquina inferior izquierda.

Correspondencia basada en siluetas

Los objetos pueden poseer formas complejas, como pueden ser las manos, la cabeza y los hombros, que no pueden ser bien descritas mediante formas geométricas simples. Los métodos basados en siluetas proporcionan una descripción de la forma de estos objetos. El objetivo de un algoritmo de seguimiento de objetos basado en siluetas consiste en encontrar la región del objeto en cada fotograma por medio del objeto generado usando los fotogramas anteriores. Este modelo puede ser en forma de un histograma de color, los bordes del objetos o su contorno. Las siluetas se pueden representar mediante diversas formas, si bien la más común consiste en una función indicadora binaria, que marca la región del objeto mediante unos y las regiones que no pertenecen al objeto como ceros. Se pueden dividir los algoritmos de seguimiento basados en siluetas en dos categorías, de acuerdo a si se basan en correspondencias de formas o de contornos. Los métodos basados en correspondencia de formas buscan la silueta del objeto en el fotograma actual, mientras que los basados en contorno parten de un contorno inicial para hallar su nueva posición en el fotograma actual, usando los modelos de espacio

de estados o realizando una minimización directa de alguna función de energía. Dentro de las representaciones que hemos establecido previamente, podemos encuadrar estos métodos dentro de los que utilizan la silueta y contorno del objeto, los basados en forma y los que utilizan modelos de esqueleto (ver figuras 2.11e, 2.11f, 2.11g, 2.11h, y 2.11i).

Las representaciones elegidas por los métodos de seguimiento de objetos basados en siluetas pueden usar modelos de movimiento (similares a los algoritmos basados en puntos), modelos de apariencia (similares a los métodos a nivel de semilla), modelos de formas o una combinación de los anteriores. La apariencia de los objetos se modela normalmente mediante funciones de densidad paramétricas o no paramétricas tales como mezclas de gaussianos o histogramas. Las formas de los objetos se pueden modelar en forma de subespacios de contornos generados en base a un conjunto de posibles contornos de objetos obtenidos a partir de diversas poses de los mismos (Blake and Isard, 1998). Además, la forma de un objeto se puede modelar implícitamente mediante una función de conjuntos de niveles en la que las posiciones de las rejillas se asignen a la distancia generada a partir de diversas funciones de conjuntos de niveles correspondientes a posturas diferentes del objeto (Yilmaz et al., 2004).

La correspondencia de formas puede llevarse a cabo de manera similar al seguimiento basado en plantillas, donde se buscaba en el fotograma actual la silueta del objeto de acuerdo a su modelo asociado. La búsqueda se realiza mediante el cálculo de la similaridad del objeto con el modelo generado a partir de la silueta del objeto predicha de acuerdo al fotograma anterior. En este enfoque, se asume que la silueta únicamente se traslada de un fotograma al siguiente, por lo que no se puede llevar a cabo seguimiento de objetos no rígidos. El modelo del objeto, que normalmente se encuentra en forma de un mapa de bordes, se puede reiniciar para poder manejar cambios de apariencia dentro de cada fotograma en el que se encuentra el objeto. Esta actualización es necesaria para afrontar problemas de seguimiento relacionados con el punto de vista y cambios de las condiciones de iluminación, así como movimiento de objetos no rígidos.

Un enfoque para hacer correspondencia de siluetas es encontrar formas que se correspondan detectadas en dos fotogramas consecutivos. La forma de establecer esta correspondencia es similar a la que se utilizaba en la correspondencia de puntos explicada previamente, si bien la principal diferencia en ambas consiste en la representación de los objetos y en los modelos de objetos usados. La detección inicial de siluetas normalmente se realiza mediante resta de fondo, realizándose la correspondencia una vez que se han extraído las siluetas de los objetos mediante cálculo de la distancia entre los modelos de los objetos asociados a cada silueta. Estos modelos normalmente consisten en funciones de densidad (usando histogramas de color o bordes), límites de siluetas (las cuales pueden consistir en contornos abiertos o cerrados de objetos), los bordes de los objetos o una combinación de los anteriores. Por ejemplo, en Haritaoglu et al. (2000) se realiza una correlación binaria de bordes entre los bordes de las siluetas en el último fotograma y los bordes obtenidos en la imagen actual. Por su parte, en Davis et al. (2000) se utiliza un modelo de distribución de puntos (del inglés Point Distribution Model o PDM) para representar la silueta de los humanos. Las configuraciones más probables para los bordes en el fotograma anterior son utilizadas para predecir la posición en el fotograma actual

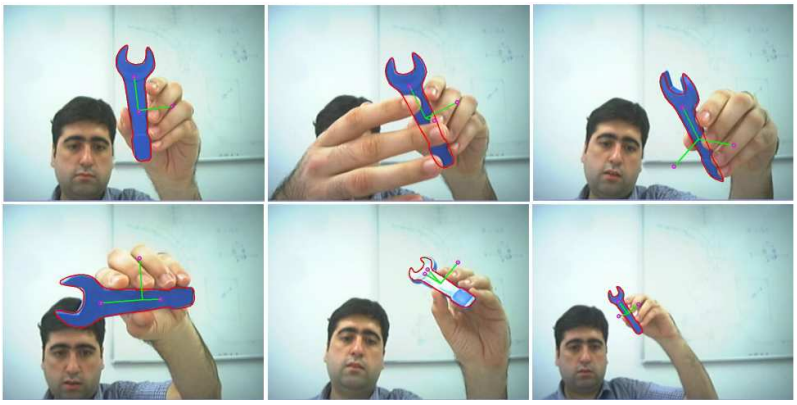


Figura 2.14: Ejemplo de seguimiento basado en contornos, extraído de (Panin et al., 2006).

utilizando un filtro de partículas. Una propuesta similar puede encontrarse en Koschan et al. (2003) donde se aplica un modelo de forma activa para encontrar coincidencias en el fotograma actual. También en Li et al. (2010); Yeh et al. (2010); Wang et al. (2012b) podemos encontrar algoritmos de seguimiento en infrarrojo basado en siluetas utilizando la información de los bordes y de la intensidad para aplicarla a un filtro de partículas. En Krüger et al. (2005) la silueta extraída se correlaciona con una jerarquía de siluetas de personas caminando aprendida a priori. La traslación, escalado y tipo de las siluetas se estiman en paralelo mediante un framework de seguimiento bayesiano en tiempo real.

Por otro lado, los métodos de seguimiento de *contornos*, a diferencia de los métodos de correspondencia de formas, parten de un contorno inicial en el fotograma anterior para hallar la nueva posición del objeto en el fotograma actual. En los métodos basados en contornos, la silueta se puede representar explícita o implícitamente. La representación explícita define los límites de la silueta mediante un conjunto de puntos de control, tal y como podemos ver en Wu and Nevatia (2005), donde se detectan cuatro partes del cuerpo diferentes: cuerpo completo, cabeza y hombros, torso, y piernas. Para cada parte se entrena un detector que utiliza un clasificador de boosting junto con “edgelets” (pequeñas cadenas de píxeles conectados del borde) los cuales se cuantifican en diferentes orientaciones. El problema de estos métodos es que, al igual que los basados en núcleos también requieren que alguna parte del objeto en el fotograma actual se solape con la región que contiene al objeto en el fotograma anterior. Este seguimiento se puede realizar usando modelos de espacio de estados para modelar la forma del contorno y su movimiento, o bien minimizando la energía del contorno mediante técnicas directas de minimización tales como descenso de gradiente. En la Figura 2.14 podemos ver un ejemplo de este tipo de seguimiento.

En los modelos de espacio de estados, se define el estado del objeto en base a la forma y los parámetros de movimiento del contorno. El estado se actualiza en cada instante de tiempo de forma que se maximice la probabilidad del contorno a posteriori, dependiente del estado anterior y la probabilidad actual, definida normalmente mediante la distancia entre el contorno y los bordes observados. Por su parte, en los métodos de minimización de energía del contorno, dicha energía se define en base a la información temporal ya sea usando el gradiente temporal (flujo óptico) o estadísticas de apariencia

generadas a partir del propio objeto, las cuales requieren la inicialización del contorno en el fotograma actual a partir de su posición anterior.

2.3.2. Propuestas de seguimiento de objetos y humanos

Dentro de la visión artificial, se pueden distinguir dos aproximaciones principales para el seguimiento de personas: aquellas basadas en modelos y aquellas basadas en características.

Aproximaciones basadas en modelos

En las aproximaciones *basadas en modelos*, un modelo del objeto se aprende por adelantado, normalmente a partir de un conjunto amplio de imágenes de entrenamiento que muestran el objeto desde diferentes puntos de vista y en distintas poses (Rohr, 1994). Ya que aprender un modelo de humano es difícil a causa de la dimensionalidad del cuerpo humano y de la variabilidad del movimiento humano, normalmente estos métodos no operan en tiempo real y se basan en fondo estático e uniforme. Usando conjuntos de características específicas de los objetos cuidadosamente elegidas, se pueden lograr detecciones muy fiables aplicables directamente como observaciones en un algoritmo de seguimiento. Por ejemplo, en Atsushi et al. (2002) se modela la pose de los humanos en el fotograma anterior por medio de una elipse, mientras que en el fotograma actual se predicen nueve posibles poses del humano. En situaciones de oclusión parcial, los métodos basados en formas descritos anteriormente suelen fallar dada la falta de información global de las formas. Aún así, existen mejoras que incluyen detección de humanos basada únicamente en algunas partes de la silueta. Cuando los humanos aparecen agrupados, el problema de las oclusiones se ve aumentado y la única información fiable de la forma es la obtenida a partir de la cabeza o el conjunto cabeza-hombros. Por ejemplo, el seguimiento del popular *Kinect* se encuentra basado en modelos, con una base de entrenamiento de quinientos mil fotogramas (Shotton et al., 2013). Aunque la mayoría de los trabajos se limitan a vistas frontales o traseras, encontramos algunas extensiones capaces de tratar vistas laterales (Wu and Nevatia, 2006).

Por otra parte, en Song et al. (2000) se presenta un método para detectar y etiquetar movimientos humanos en secuencias de imágenes. El método usa como entrada la posición y velocidad de las características más destacadas de la imagen, computadas por el algoritmo de seguimiento de características mostrado en Tomasi and Kanade (1991), sin requerirse segmentación previa. El método se basa en el modelado del movimiento humano con una aproximación de la densidad de probabilidad conjunta de la posición y el movimiento de las características asociadas al cuerpo humano, realizándose la detección mediante la suma de las probabilidades de todos los etiquetados posibles. A continuación, la localización se realiza hallando el subconjunto de las características detectadas que es más probable que esté asociado a un cuerpo humano. El entrenamiento del modelo a asociar se realiza con un conjunto de entrenamiento etiquetado a mano. También en Ghaeminia et al. (2010) se realiza un modelado del movimiento humano a partir de la función de densidad de probabilidad de aceleración y velocidad del humano a seguir.

Aproximaciones basadas en características

Por su parte, en los métodos *basados en características* no se aprende un modelo, sino que se sigue a un objeto basándose en características simples como características de color o esquinas.

En muchos sistemas de seguimiento, el color de un humano se representa mediante un histograma de color (Hu et al., 2004; McKenna et al., 2000; Okuma et al., 2004; Xu and Puig, 2005), o bien mediante una mezcla de gaussianos (Kang et al., 2005; Khan and Shah, 2000; Roth et al., 2005; Yang et al., 2005). Por ejemplo, en Pérez et al. (2002) se presenta una propuesta probabilística basada en la búsqueda determinista de una ventana cuya ventana de color se ajuste con un modelo de referencia de histograma de color. Esto se realiza en un entorno probabilístico, en el que se utiliza una técnica de seguimiento de Montecarlo. El uso de un filtro de partículas permite aplicar el algoritmo en las situaciones en que los colores se acumulan en el fondo, así como la oclusión completa de las entidades seguidas durante unos pocos fotogramas. El problema de esta propuesta es que la forma de la región a seguir debe ser fijada a priori, si bien dicha forma no se limita únicamente a elipses o planos. Dentro de estas técnicas destacaremos Frintrop et al. (2010), donde se presenta un método cuyo núcleo consiste en un descriptor basado en componentes que captura la estructura y apariencia de un objetivo de forma flexible. También podemos citar el algoritmo de seguimiento en infrarrojo expuesto en Li and Gong (2010), en el cual se usan las características de las regiones que contienen al cuerpo de los humanos. El método construye la representación del histograma de las ROIs en un modelo de proyección espacial de intensidad y distancia, con el fin de atenuar el problema que presenta tener información insuficiente cuando únicamente se considera la intensidad. Basándose en la idea de que los cuerpos de los peatones presentan formas similares, se usa esta característica para mejor el rendimiento del seguimiento. Tras elaborar un histograma de distancias (entre cada objeto y el centro de la imagen) y otro de intensidades, ambos se fusionan. Finalmente se utiliza un filtro de partículas para predecir el siguiente estado de cada objeto actualizando posteriormente las medidas estimadas. Más recientemente tenemos una propuesta basada en filtros de partículas para el seguimiento de humanos (Jharna and Kiran, 2013).

Las representaciones basadas en mezcla de gaussianos se comparan normalmente utilizando la distancia de Mahalanobis, la cual puede evaluarse utilizando un gaussiano (Kang et al., 2005) y asumiendo la independencia entre los canales de color (Cucchiara et al., 2004). Otra alternativa es la utilización de la media (Yang et al., 2005). Otro ejemplo lo podemos encontrar en Kelly et al. (2009), donde se pretende realizar detección y seguimiento de humanos en escenas con aglomeraciones de gente. La detección de peatones se realiza mediante un proceso de agrupamiento en 3D en un entorno de crecimiento de regiones, usando para caracterizarlos información de sus trayectorias previas y de sus características de color.

2.3.3. Resumen y conclusiones

En la presente sección se ha realizado un estudio general de las últimas tendencias en el campo del seguimiento de objetos. En primer lugar, se ha realizado una breve introducción aportando una visión general de la problemática que surge a la hora de elaborar un algoritmo de seguimiento. Posteriormente, se han descrito los principales esquemas que se pueden encontrar en la literatura para la representación de objetos. También se ha hecho especial hincapié en las diversas alternativas a la hora de establecer correspondencias entre los objetos detectados en el fotograma actual y los ya detectados previamente por el sistema, siguiendo la clasificación establecida por Yilmaz et al. (2006). Finalmente, se ha aportado una visión general del problema particular del seguimiento de humanos, con el fin de completar una perspectiva general del problema del seguimiento aplicado al principal campo de estudio de la tesis.

En cuanto a los sistemas de representación de objetos, tras relacionarlos con los sistemas de correspondencia temporal, se ha visto que existen tres grandes grupos: puntos, semillas y siluetas. Si bien la forma más simple es elaborar una matriz de correspondencias entre los diversos objetos, se ha podido apreciar que la forma de representación más comúnmente utilizada en la literatura es la basada en semillas, particularmente en plantillas. Esto se debe a que proporcionan una representación más compleja de los objetos que los basados en puntos, que muchas veces proporcionan una visión insuficiente. Además, esta representación, aunque sea computacionalmente compleja, es muy flexible pudiendo utilizar regiones de mayor o menor complejidad según el objetivo requerido. Otra tendencia muy utilizada es la correspondencia basada en siluetas, ya que la segmentación inicial es relativamente simple de obtener (muchas veces basta con una resta de fondo) y no es computacionalmente cara, amén de que los métodos basados en formas permiten caracterizar mucho mejor a objetos complejos compuestos de formas simples, como pueden ser humanos. En la tabla 2.4 podemos ver la clasificación de los algoritmos descritos en base a la correspondencia temporal que utilizan.

2.3.3.1. Seguimiento de humanos

Se ha realizado una particularización final en el seguimiento de humanos, el cual se ha visto que se puede basar en modelos y en características. Si bien no se aprecian tendencias tan marcadas como en el caso general de los objetos, es cierto que se puede apreciar que los métodos basándose en características son más utilizados que los basados en modelos, debido a que no requieren un aprendizaje previo a diferencia de las primeras. Dentro de las técnicas basadas en características, se ha apreciado una gran variedad de técnicas muy diferentes entre sí, cada una basada en diversas características de las formas de los humanos. En la tabla 2.5 se puede apreciar la clasificación descrita de los algoritmos de seguimiento de humanos estudiados.

Tabla 2.4: Clasificación de los algoritmos de seguimiento estudiados según el método de correspondencia temporal utilizado

Representación	**Metodología**	**Ejemplos**
Puntos	Deterministas	(Shafique and Shah, 2005), (Fuentes and Velastin, 2006)
	Probabilísticos	(Veenman et al., 2001), (Serby et al., 2004), (Weng et al., 2006), (Benezeth et al., 2010), (Fu and Han, 2012)
Semillas	Plantillas	(McKenna et al., 2000), (Song et al., 2000), (Atsushi et al., 2002) , (Pérez et al., 2002), (Comaniciu et al., 2003), (Cucchiara et al., 2004), (Okuma et al., 2004), (Xu and Puig, 2005), (Roth et al., 2005), (Yang et al., 2005), (Zhou and Hoang, 2005), (Fuentes and Velastin, 2006), (Schiele, 2006), (Kelly et al., 2009), (Zhou et al., 2009), (Frintrop et al., 2010) (Ghaeminia et al., 2010), (O' Malley et al., 2010), (Wang et al., 2012b), (Jharna and Kiran, 2013)
	Multivista	(Khan and Shah, 2000), (Kang et al., 2005), (Fleuret et al., 2008)
Siluetas	Formas	(Blake and Isard, 1998) (Ali and Aggarwal, 2001), (Koschan et al., 2003), (Hu et al., 2004), (Yilmaz et al., 2004) (Krüger et al., 2005), (Wu and Nevatia, 2006), (Benezeth et al., 2008), (Li et al., 2009), (Li et al., 2010), (Li and Gong, 2010) (Yeh et al., 2010), (Wang et al., 2012b)
	Contornos	(Davis et al., 2000), (Haritaoglu et al., 2000), (Zhao and Nevatia, 2003), (Yang et al., 2004), (Wu and Nevatia, 2005), (Hu et al., 2006),(Shotton et al., 2013)

2.4. Resumen y conclusiones generales

En el presente capítulo se ha realizado un estudio general de las fases que compondrán el sistema que se realizará durante la presente tesis, es decir, detección de humanos, fusión de las regiones de interés que contienen humanos detectadas en diversas segmentaciones y seguimiento de las regiones obtenidas.

En primer lugar, se ha realizado una aproximación inicial a la segmentación. Posteriormente, se han estudiado las diversas fases que componen un algoritmo de segmentación, aportando ejemplos de cada una y clasificando los diversos algoritmos según su aproximación a estas fases. A continuación, se ha incidido especialmente en la detección de humanos en color y en infrarrojo con el fin de tener una perspectiva general de como llevan a cabo este proceso los diversos autores que podemos encontrar en la literatura. Finalmente, se ha llevado a cabo una clasificación de los diversos algoritmos estudiados, observándose que si bien la resta de fondo es un procedimiento ampliamente utilizado, ésta requiere normalmente el uso de clasificadores que fortalezcan el resultado, lo cual repercute en el rendimiento de los algoritmos. En cuanto a la detección de humanos, se ha llegado a la conclusión de que resulta especialmente interesante utilizar las características térmicas de los humanos en el espectro infrarrojo, mientras que en el color se ha observado que existen pocos algoritmos que utilicen esta información

Tabla 2.5: Clasificación de los algoritmos de seguimiento de humanos estudiados

Representación	**Metodología**	**Ejemplos**
Modelos	Entrenamiento	Atsushi et al. (2002), Hu et al. (2004), Wu and Nevatia (2006), Shotton et al. (2013)
	Probabilidad	Haritaoglu et al. (2000), Song et al. (2000), Fleuret et al. (2008), Ghaeminia et al. (2010),
Características	Histogramas	McKenna et al. (2000), Yang et al. (2004), Hu et al. (2006), Pérez et al. (2002), Hu et al. (2004), Okuma et al. (2004), Yang et al. (2004), Xu and Puig (2005), Zhou and Hoang (2005), Hu et al. (2006), Li and Gong (2010)
	Mezcla de gaussianos	Khan and Shah (2000), Cucchiara et al. (2004), (Kang et al., 2005), Roth et al. (2005), Yang et al. (2005), Kelly et al. (2009)
	Otros	Davis et al. (2000), Zhao and Nevatia (2003), Krüger et al. (2005),Shafique and Shah (2005), Fuentes and Velastin (2006), Schiele (2006), Li et al. (2009) Frintrop et al. (2010), O' Malley et al. (2010)

a la hora de caracterizar a los diversos humanos.

Análogamente, se ha llevado a cabo un proceso similar con la fusión de imágenes, núcleo principal de la presente tesis. Tras caracterizar este proceso y plasmar su utilidad, se han explicado las fases necesarias para llevarlo a cabo, realizándose una clasificación de los diversos algoritmos estudiados en base a su aproximación a estas fases. Finalmente, se ha llevado a cabo un resumen final de los diversos métodos estudiados, obteniéndose como principal conclusión que la fusión a nivel de regiones es especialmente interesante, ya que proporciona mayor independencia de las características de las diversas cámaras y presenta mayor invariabilidad frente a problemas puntuales de ruido. En cuanto a las aplicaciones de estas técnicas, se ha podido observar que su uso puede ser especialmente recomendable en el campo de la vigilancia debido a que proporcionan información para reforzar las detecciones realizadas por diversas cámaras.

Finalmente, se ha realizado un estudio de las diversas técnicas de seguimiento que se pueden encontrar en la literatura. Tras describir las principales problemáticas que surgen a la hora de realizar un algoritmo de seguimiento, se han estudiado las diversas técnicas que existen a la hora de representar los objetos y realizar la correspondencia de los mismos entre fotogramas, incidiéndose nuevamente de forma especial en el seguimiento de humanos. Nuevamente se ha concluido este estudio con una clasificación de los algoritmos estudiados, obteniéndose como conclusión que para el seguimiento de humanos las técnicas basadas en semillas son especialmente interesantes, ya que los humanos son objetos no rígidos que presentan mayor variabilidad que la requerida por las técnicas basadas en puntos. En cuanto al seguimiento específico de humanos, se ha concluido que los métodos basados

en características pueden ser especialmente apropiados para los objetivos de la tesis al no requerir entrenamiento previo, lo que les proporciona mayor versatilidad.

Capítulo 3

Descripción del sistema de detección de humanos

El presente capítulo aborda la descripción general de los pasos seguidos a la hora de diseñar, implementar, probar y validar la propuesta de un sistema robusto de detección de humanos mediante fusión de vídeo en color e infrarrojo. En primer lugar, se describe el marco de trabajo INT^3-Horus, una arquitectura multisensorial inteligente para el desarrollo sencillo de tareas de monitorización e interpretación de actividades. A continuación, se sitúa la propuesta dentro de este marco general y nos adentramos en los niveles seleccionados del entorno para abordar el problema específico de detección de humanos mediante visión artificial. Para ello, partiendo de un nivel de adquisición de imágenes con cámaras en color e infrarrojo, se ha procedido a la elaboración e implementación de diversos algoritmos en de segmentación de humanos en ambos espectros, con el fin de elegir el más apropiado para nuestros objetivos. Una vez que se ha completado el nivel de segmentación, el siguiente objetivo consiste en combinar los resultados obtenidos en un nivel de fusión. Este paso constituye una parte significativa de la tesis, y es sin duda el más complejo, ya que requiere investigar formas de combinar los resultados de ambos espectros, basándose en las características de cada uno de ellos, buscando fortalecer las segmentaciones iniciales obtenidas. Finalmente, se procede a la identificación y seguimiento de humanos, utilizando la información obtenida de la fusión, con el fin de poder identificar a los individuos presentes en la escena y obtener información de sus trayectorias, amén de poder predecir hacia donde se dirige cada humano detectado y previamente identificado. Se descubre que este último paso es más efectivo si contribuye, a su vez, al proceso de fusión, por lo que se realiza una realimentación del nivel de seguimiento hacia el nivel de fusión.

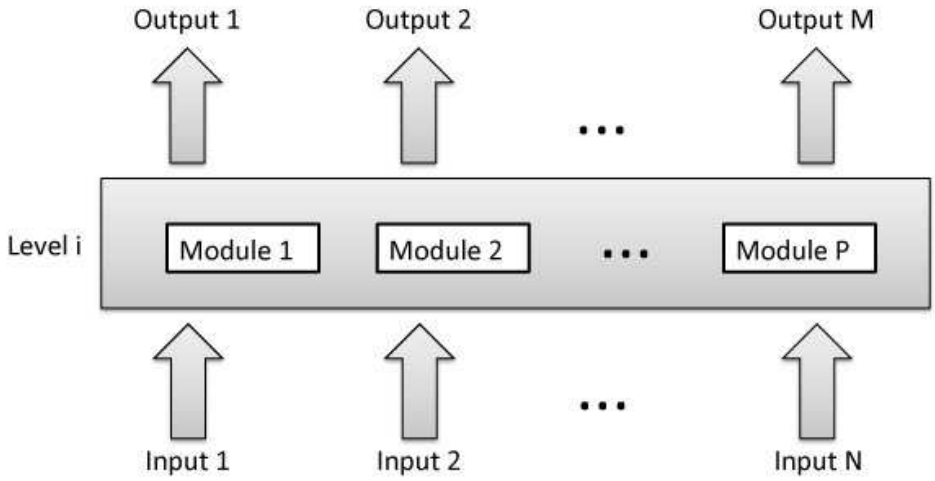


Figure 3.1: Example of level with inputs, outputs and operation modules. Despite inputs are common for the whole level, it is not mandatory for each level to manage them. The same happens with outputs; they reflect the format of the level outputs, which varies depending on the implemented modules.

3.1. Description of the Underlying Framework

INT^3-Horus is a multisensor framework to carry out monitoring and activity interpretation (e.g. (Fernández-Caballero et al., 2013; Castillo et al., 2013)). The framework establishes a set of levels with some clearly defined input/output interfaces to provide a hierarchy to the processing. The levels consist of a set of modules that incorporate the algorithms dedicated to processing at each level. If thinking of several sensors that provide input information, at the lowest level (the acquisition level) several modules, each one responsible for the acquisition of a type of sensor, are located. For each level, the framework provides a set of inputs and outputs to be met by the modules (see Figure 3.1). The inputs and outputs are independent from each other and from the number of modules. Thus, a module is not required to implement all inputs and outputs on its level, but it may implement the subset that best fits its needs. A higher level task is in charge of selecting those modules at different levels that are compatible with each other to create a monitoring system based on the framework.

Thus, INT^3-Horus allows the coexistence of modules that are in charge of information of different nature within a single level, although they conceptually work at the same processing level. For instance, at the level of acquisition there are some modules that capture information from cameras, while others are prepared to capture data from wireless sensor networks. Although both sources of information seem incompatible a priori, upper levels house algorithms to merge and operate with them, regardless of the data capture algorithms. Following the scheme described in Figure 3.1, the levels of the framework establish a hierarchy from the level of sensor information acquisition to the level of activity analysis, by connecting the inputs of the immediately upper level through the outputs of the lower level (see Figure 3.2).

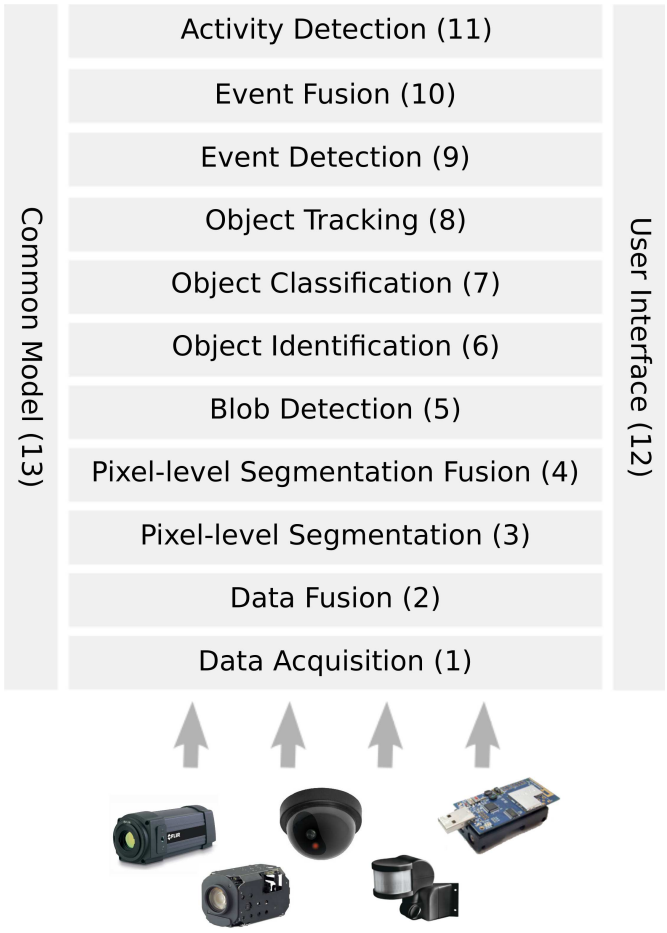


Figure 3.2: Framework levels.

The INT[3]-Horus framework uses OpenCV libraries and Qt programming environment with their own libraries. The Qt environment is a programming environment similar to Microsoft Visual Studio that allows cross-platform programming, i.e. applications that can run simultaneously on Windows, Mac and UNIX are programmed. Currently, even extensions have been developed that enable the deployment of applications for Android [1]. Qt was originally produced by the development division for Nokia environments, while the license was subsequently sold to Digia, the current developer. In this thesis we used the QtCreator 2.4.1 version and 4.7.4 version of the libraries.

For its part, the OpenCV libraries are a computer vision libraries created under BSD license (i.e. open source). These libraries are developed to maximize the power of Intel processors. Apart from providing more than 500 vision functions, their greatest advantage is that there is a large Internet development community, resulting in an increasing power and evolution of the library functionalities. An example of added functionality can be found in the CvBlobsLib[2] libraries (also used to implement this thesis) which allow the extraction and filtering of a list of blobs from a binary image. In this thesis

[1] http://qt-project.org/wiki/Qt5ForAndroid
[2] http://opencv.willowgarage.com/wiki/cvBlobsLib

we used OpenCV 2.4.0 version.

Next, the most significant features of INT^3-Horus framework are described.

3.1.1. INT^3-Horus Is Multisensory

The framework is designed to work with different sources of information. The sources are mainly based on vision sensors as they are the most widely used for monitoring tasks. Nevertheless, other sensor technologies are introduced to provide the framework of greater power and flexibility. These technologies are mostly sensors used in commercial surveillance, that is, volumetric sensors, presence detection sensors, contact sensors for doors and windows, etc. The framework also includes the ability to access information from wireless sensor networks (WSNs), which allows rapid deployment of sensors in the area to be monitored, regardless of its characteristics (indoor, outdoor, and so on).

3.1.2. INT^3-Horus Includes Information Fusion

As the framework operates with different data sources, INT^3-Horus includes information fusion algorithms in its design. For this reason, the JDL data fusion model has been used as basis, adapting some of its levels from military to civilian. The JDL model was developed by the *Joint Directors of Laboratories Data Fusion Group*, a committee of the United States Department of Defense (DoD). The architecture proposed by the JDL model is considered the de facto standard for implementing a surveillance system. The proposed model illustrates the main functions, relevant information and databases, as well as the interconnection required to perform data fusion. The JDL fusion model also provides a definition of the concept of data fusion (DoD, 1991) that was later refined as a "multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from single and multiple sources" (DoD, 1994).

After studying the different fusion levels proposed by the JDL model, Figure 3.2 shows a schematic view of our proposal for the union of traditional surveillance processing levels with the JDL fusion levels. Following this approach, processing starts at the level of acquisition (1), where several sources provide different types of information (e.g. color cameras, thermal cameras and volumetric sensors). After that, the proposal establishes an initial fusion level (2) in charge of the refinement of the sources of information, using the characteristics of some information sources to reduce the defects of others (JDL level 0). Once the input information is enhanced by means of fusion, there is a level of segmentation of the objects (3). This level is designed to locate the blobs that contain objects of interest in the scene. When working with multi-sensory systems, multiple segmentation algorithms may run in parallel, each one extracting the blobs of the information provided by a different sensor type, or by the fusion of different types of sensors. Therefore, once the segmentation level has increased the abstraction of information, a new level of fusion is proposed (4), which unites the blobs detected by the different segmentation algorithms. This level coincides with level 1 proposed by the JDL standard. Again, the robustness of the information provided by the previous level is improved

and an enhanced blob detection (5) can be performed.

Now, processing continues with object identification (6), where the information coming from the blobs of the previous levels is transformed into object-level information (that endures over time). Term object refers to real world objects detected by the framework defined by a series of attributes such as real world position, trajectory, or class, among others. The subsequent levels complete the object definition by adding some required parameters. Continuing with the object information processing, classification (7) and trajectory analysis (8) are also performed that also allows trajectory prediction. The information generated by this tracking level (8) is passed on to the next level, that is, events detection (9). Here the object level information is transformed into semantic information about the behavior of the objects in the scene. After the events are detected, there is a new level of fusion (10) that matches level 2 of the JDL model, where the actions identified for the objects are fused to search for more complex behaviors involving multiple objects. Finally, the activity interpretation (11) is performed to find more complex behaviors involving several objects and sensors in the scenario. After all this, other aspects such as user interfaces (12), execution control and information management (13) are also considered.

3.1.3. INT^3-Horus Is Based on the MVC Paradigm

This proposal extends the traditional model-view–controller (MVC) architecture (Reenskaug, 1979) in order to provide a greater flexibility when incorporating the functionalities of monitoring systems and to allow existing algorithms to be incorporated into the framework without involving a major change in design. To do this, the business logic is detached from the model, generating a new execution block and allowing the controller to invoke its functionality. This new block is called “algorithm”. Its functionality is given by the algorithms of each level of a monitoring system and other features necessary to implement a system that works with remote processing nodes. Thus, the block “algorithm” includes features starting from traditional processing algorithms and business logic to communications between remote modules or databases, passing through the acquisition of sensor data, essential in monitoring systems. It may seem a priori that this new block handles all the processing. Nevertheless, instead of a single block there will be a set of them, each linked to a module composing the levels of the framework. In addition, in this extension to MVC, the functionality of the local model is to store the application data and to provide primitives to manage them (see Figure 3.3). Each component in INT^3-Horus possesses this internal structure, being the local model in charge of storing and managing local data. Global data is stored and managed by the *Common Model*.

All system elements are included in one of three modules (model, view and controller) in the traditional MVC paradigm. In our extension, the components depicted in Figure 3.3 are considered as a single execution module. Each module is one of the levels of the INT^3-Horus framework and is responsible for providing the functionality by performing the associated processing, instantiating the required parameters in a *Common Model* and showing the result of the execution as well as the parameters to allow adjusting its settings in its interface.

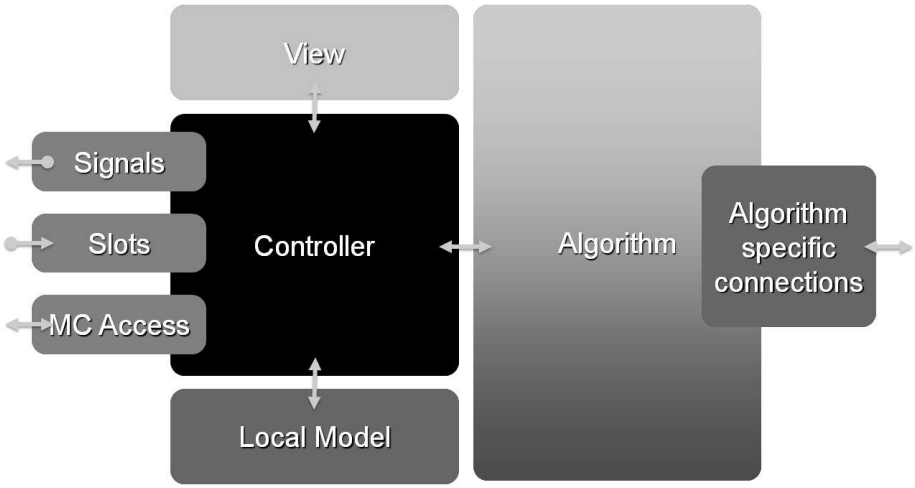


Figure 3.3: Extension to the traditional MVC.

3.1.4. INT^3-Horus Is a Hybrid Framework

It is necessary to describe the execution model before describing the levels that make up INT^3-Horus in detail. It is defined as a hybrid distributed system, where a number of remote nodes are responsible for a part of the processing (the lower levels) and a central node is responsible for collecting information from those nodes to merge and make the processing at higher levels. The central node also houses the central controller of the complete framework, as well as the common view that encompasses the views of each of the framework levels (see Figure 3.4).

Figure 3.5 shows a representation of the different hardware blocks that make up such a system. All remote nodes have the extended MVC structure, but they only perform a part of the complete framework processing. In the central node the major components of the extended MVC and their relationships are distinguished, although the remote nodes also have a *Common Model* whose functions are detailed next.

3.1.5. INT^3-Horus Provides a *Common Model*

Within the INT^3-Horus monitoring and activity interpretation framework, there is a fundamental part, namely the *Common Model*. The *Common Model* has two distinct functionalities: to host and manage both the data model and the controller of the execution of the different framework levels. These features correspond to the model and the controller of the traditional MVC paradigm and convert the *Common Model* in the core of the framework. Levels and modules can vary (or fail), but the data model and the controller have to ensure the data consistency and proper implementation of the components that form the framework, as well as to control possible errors in the levels and try to correct or minimize them as far as possible.

The *Common Model* houses the data structures that support the exchange of information between

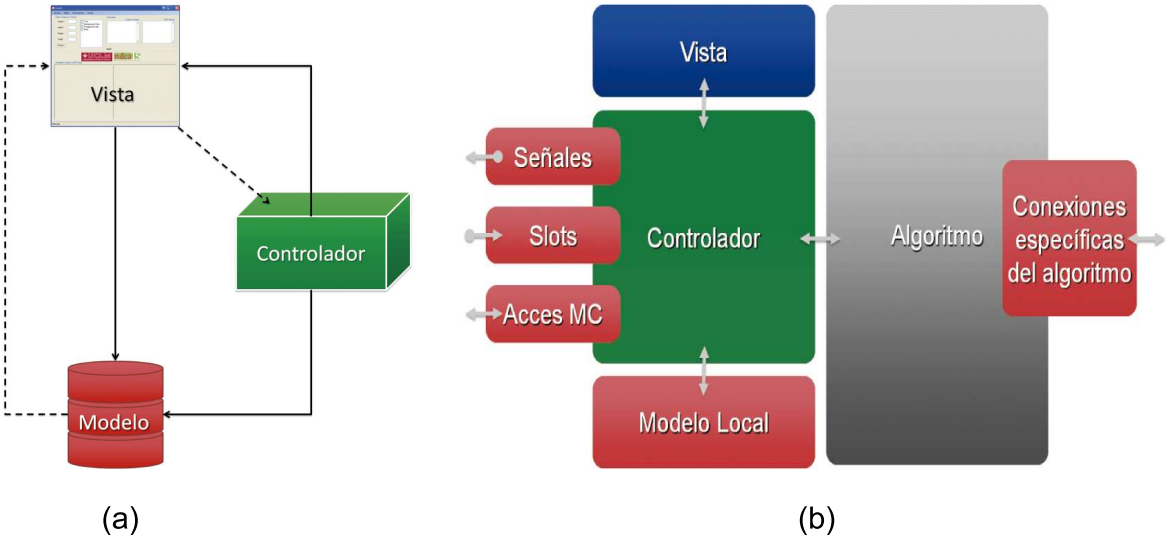


Figure 3.4: Definition of the *Common Model*.

levels of the framework. In this sense, primitives are provided for information management. On the one hand, there are primitives that add information to the *Common Model* from the levels (outputs). Similarly, in the data model there exist functions to retrieve the information from the levels that require it (inputs). On the other hand, the *Common Model* provides functions for managing the data model, allowing to update the information contained and its deletion (if no longer needed) within the framework execution flow. The process occurs through two pathways: the direct elimination of the information with the consequent release of memory, or the storage of information in hardware, freeing it from main memory. All these operations require a high degree of consistency within the data model. For this reason, the *Common Model* performs the necessary checks to ensure that the information inserted into the data model is correct and does not cause inconsistencies or errors when the modules require it for execution.

Furthermore, the *Common Model* has been designed as the "orchestra conductor" in INT^3-Horus. It is responsible for launching the execution of the modules contained in the framework levels and for receiving signals indicating the execution completion. Thus, different execution models, such as sequential or pipeline, allowing the execution of several levels in parallel, each handling relative information in different times, are permitted. In addition, the *Common Model* is responsible for monitoring the correct performance of the modules, taking action in case of malfunction (for example by restarting the modules). There is a signaling mechanism to warn the modules of the restart of one of them, so they can perform the actions necessary for not altering their operation.

3.1.6. The INT^3-Horus Processing Levels

The present subsection describes the functionality of the levels of the INT^3-Horus framework. Of course, the proposed levels are just a guideline to create the framework, but it is possible to include

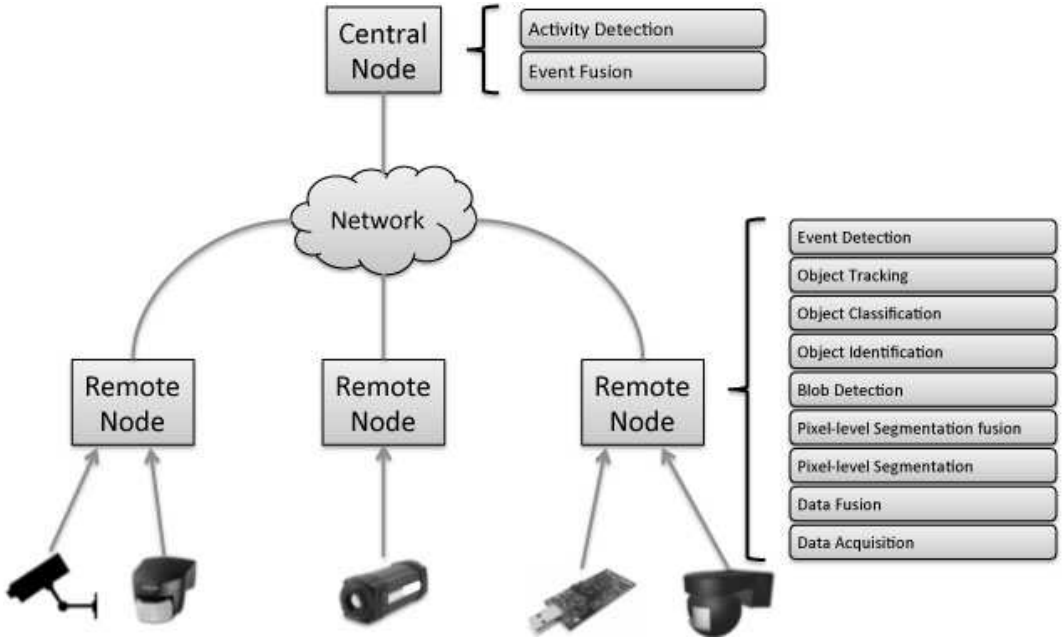


Figure 3.5: Hybrid execution model. Levels per node.

new levels according to the application requirements.

Data Acquisition: This level directly interacts with the digital analog devices, measuring from the physical world and adapting these measures to be usable by the system. The measures are data from the sensors as well as data from other information sources (disk, database, and so on). The data acquisition level also performs information preprocessing.

Sensor Fusion: This level is in charge of fusing the sensor data to improve the information quality (more complete and accurate). Fusion algorithms may also operate with different spectrum images and are capable of introducing knowledge on the domain.

Pixel-level Segmentation The third framework level is dedicated to isolate the objects of interest contained in the input images. This level may hold a wide range of methods, from the simplest one (a binarization applied to infrared (IR) images (Fernández-Caballero et al., 2011a), (Fernández-Caballero et al., 2010)) to other more complex approaches yielding better results (Moreno-Garcia et al., 2010). On the other hand, this level also filters the spots corresponding to the objects of interest with the aim of eliminating possible noise.

Pixel-level Segmentation Fusion: This level fuses images obtained in localization and filtering stage as there might be several localization and filtering approaches running in the framework (e.g.

one devoted to color images and another to IR images). Thus, this level seeks for the most benefic features from the input images.

Blob Detection: The blob detection level filters isolated spots misdetected in the previous levels. Besides, the blob detection level is in charge of extracting information associated to the spots to allow a more efficient analysis of the objects. This information is application-dependent.

Object Identification: This level operates with objects instead of blobs. This enhances the information abstraction, mapping object coordinates into the real world instead of simply operating with image coordinates.

Object Classification: This level is specially important to perform a good activity analysis because it provides knowledge about "what" the object is. Also, object classification may provide information about the objects' orientation.

Object Tracking: This level calculates the trajectories followed by the moving objects within the scenario, independently of the particular sensor that detected them. It also makes predictions about future positions of the objects on the basis of the previously detected trajectories. This level uses the information from the common model referring to the map, the sensors situation and its coverage range.

Event Detection: The event detection level generates semantic information related to the behavior of the objects in the scenario. These events happen in a short period of time and involve few objects (usually one). Some examples are events such as running, walking or falling, which can be detected with just one or at most a few input images. This is the last level of the framework held within remote nodes (see Figure 3.5). The next levels are implemented in the central node together with the common model and the central controller and view.

Event Fusion: In a multisensor monitoring and interpretation system, where several sensors monitor a common scenario, the events generated from different sources usually do not match. This is why the event fusion level is necessary to unify the information arriving from the different sensory data generated in the previous level.

Activity Detection: This final level of the architecture is in charge of the analysis and interpretation of activities already associated to temporal features. After event fusion, the current level has a better knowledge of what is happening in the scenario according to the detected events. Hence, the activities detected at this level can be translated into actions along the scenario, providing a higher abstraction level.

3.1.7. The INT^3-Horus Formal Ontology Model

If defining an ontology O in terms of an algebraic system (Gruber, 1995), we have the following three attributes:

$$O = (C, R, \Omega) \tag{3.1}$$

where C is a set of concepts, R a set of relations between the concepts, and Ω is a set of rules. Equation (3.1) proposes the ontology for a domain of interest to be described by offering proper meanings to C, R and Ω. As Protégé (2013) is used as the ontology editing software, the following specialization is obtained:

$$DO =< Individuals, Classes, Properties, Values, Restrictions, AxiomaticRules >$$

Individuals are entities representing the instances of the framework levels as well as their associated input and output parameters. *Classes* are interpreted as "sets containing individuals", and are organized in a taxonomy in accordance with the hierarchical superclass-subclass relations. *Properties* are binary relations on individuals, which enables asserting facts about classes and individuals. They are functional, inverse functional, symmetric or transitive. The properties are used in restrictions and in axioms. *Values* contains the values that can be assigned to individuals. *Restrictions* state the permitted and extreme ranges. Generally speaking, *Restrictions* impose constraints on the properties of the classes. *AxiomaticRules* use restrictions, boolean algebra and some other concepts such as general classes to create properties and class axioms.

The components of the algebraic system (see equation 3.1) are as follows. Let us take a look at the classes with respect to our domain of interest. In agreement with formula 3.1, component C for the intelligent monitoring and activity interpretation domain can be rewritten as:

$$C = \langle Level, DataType \rangle \tag{3.2}$$

Here, the first class, $Level$, is a set of processing levels, $< L_i >$ ($i = 0, 1, \ldots, n$, n is a number of levels or layers), each one in charge of a specific set of tasks. The second class, $DataType$, $< D_j, k >$ ($j = 0, 1, \ldots, m$, $k = 0, 1, \ldots, k$, where m is a number of possible data types of the first level, and k is the possible number of their correspondent subclasses), represents data type structures that constitute inputs and outputs at each level and are used at each processing level $< L_i >$.

The second component of equation 3.1, R, stands for relations between the classes defined in C.

Component R is determined as:

$$R = \langle InputData, OutputData \rangle \tag{3.3}$$

Finally, the component Ω represents a set of rules, which determine the system functionality for a given domain. In other words, component Ω is a set of goals and corresponding actions for an application. The component Ω is determined as:

$$\Omega = \langle Goal, Class, DataType \rangle \tag{3.4}$$

where $Goal$ is a set of tree-based goals $G_{i,j}$, $i = 0, 1, \ldots, n$, $j = 0, 1, \ldots, m$, being j the number of levels in the goals tree and i the number of goals within a level. $DataType$ takes values *hasInput* or *hasOutput*.

3.1.7.1. Description of the Level Class

The ontology for the levels hierarchy, as created in Protégé, is represented in Figure 3.6. The functionality of each level is described, although no algorithm is associated to keep the framework as generic as possible. Of course, the proposed levels are just a guideline to create the framework, but it is possible to include new levels according to the application requirements.

3.1.7.2. Description of the DataType Class

In a similar manner, the DataType class is represented in Figure 3.7. The class has a hierarchical structure as it includes data type structures, corresponding to the needs of many current monitoring and activity interpretation approaches. Finally, we have identified eight data structures on the first level; and two of these classes have been subdivided into subclasses. The detailed description of these classes is provided next, ordered by the abstraction information level.

Sensor Data: Raw data coming from the different sensor technologies is considered in the specification of the modules. In this sense, despite this parameter does not correspond to a real data structure by itself, it must be taken into account as it corresponds to the general input to the framework.

Image: Images managed by INT^3-Horus correspond to a data structure provided by OpenCV (2013), namely *IplImage*. This data structure is common to the first levels of the framework, starting from data acquisition, which transform the raw data coming from cameras to this image format. As *IplImage* is a standard format, it is able to hold different kinds of image technologies, such as color, infrared and thermal images.

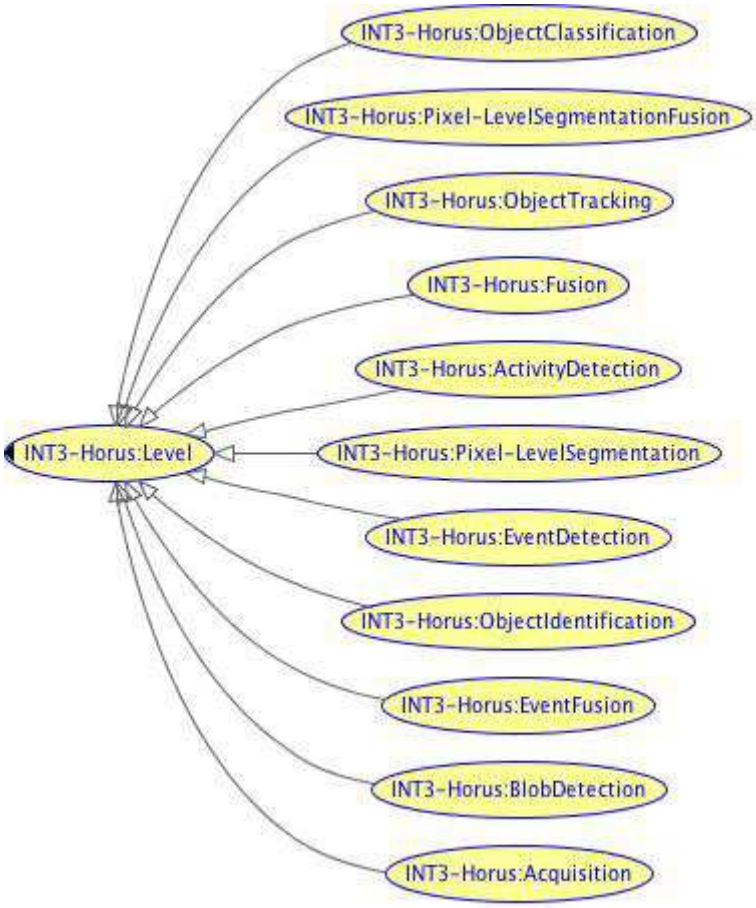


Figure 3.6: Representation of the Level Class in Protégé.

Matrix: Another way of representing sensor data is through matrixes. This parameter is used to manage data coming from range sensors, such as time-of-flight cameras. These sensors provide three-dimensional information of the scene, which makes a matrix data structure suitable for storage. On the other hand, wireless sensor networks (WSNs) are also proposed in the framework as information sources. Thus, arrays of information coming from different nodes are stored in matrixes.

Array: This parameter holds information about classic surveillance sensors. This is, boolean sensors such as volumetric or movement sensors which only provide triggered events that hold no associated value.

Blob: Image processing extracts information regarding some spots or regions. These regions and some parameters calculated (or even inferred) are stored in data structures named blobs.

Object: Growing in abstraction and temporal dependency, blobs captured from different sources are related to each other to form objects. Objects possess features to distinguish them from other entities

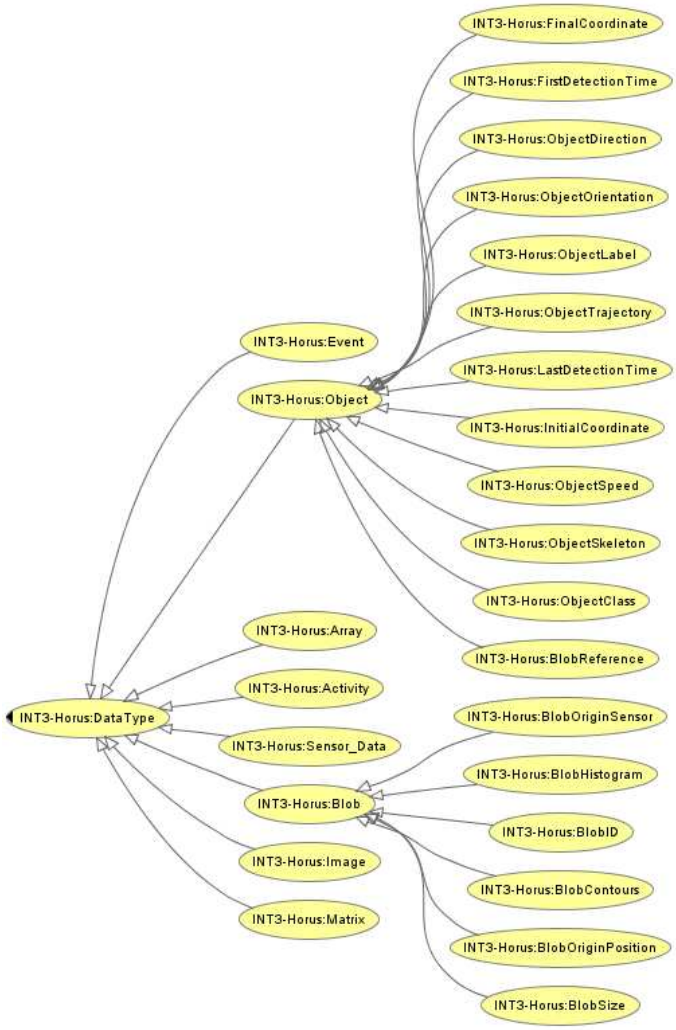


Figure 3.7: Representation of the DataType Class (including Blob and Object subclasses) in Protégé.

in the scene, tracking them along time and remaining in memory (even after leaving the scenario). This is a very powerful feature that improves the event detection process.

Event: Events are included to manage simple behaviors that occur in a short period of time (even instantly) and involves just a few objects. The events are used as primitives for higher-level behavior interpretation.

Activity: Global behaviors involving several objects (and sensors) are stored as activities. Both, event and activity, are implemented in a manner close to natural human language. This implies the use of text strings to manage the information and XML-like formats to encapsulate it.

Additionally, we have come to conclusion that the most complex classes: Blobs and Objects, have to be extended in subclasses. In the current version of INT^3-Horus, these data types contain a great amount of attributes whose number is still growing because the framework is continuously evolving. Therefore, formalization through ontology enables a flexible and well-structured knowledge representation, allowing for expansion as new features are incorporated into the framework.

3.1.7.2.1. Description of the Blob subclasses The attributes that define a blob are based on our previous works in this area (Fernández-Caballero et al., 2011a; Moreno-Garcia et al., 2010). At every step of the design, the dynamic nature of INT^3-Horus has been faced to stay open for future expansions. Therefore, the blob features that compose the Blob Subclass are easy to modify or add for a given data model. It is enough to add new attributes in the corresponding subclass and to implement the methods that operate them (or extend the existing ones). Next, the subclasses of the Blob Class are provided and described.

Contours: This is a list of points (image pixels) containing the contour of each blob. The list is implemented as a class template *QList* containing elements of type *QVector3D*. As indicated by their name, these elements allow the storage of three-dimensional points, which can be used to work with blobs obtained from range sensors, among others.

ID: This subclass has been defined to offer a unique identification to each blob. Although the utility of the *id* is restricted to the identification level, it is very useful in the higher levels when matching blobs detected over time to obtain objects. There is a possibility that the same object is detected by various sensors, where every sensor produces a separate blob with its own identifier. In this case, the incorporation of an identifier is useful to make a proper fusion of blobs and a "reconstruction" of objects.

Origin and Size: Another important parameter of a blob is the coordinates of the "bounding box" that contains it. This is why two attributes have been included to establish the origin (initial) and the final position of the box (i.e. the two opposite corners of a rectangle). Traditionally, the starting and ending points are set as the lower left and upper right corners of the rectangle, respectively.

Origin Sensor: The Blob Class must also contain the information associated with the sensor that obtained the blob. Thus, the attribute *originSensor* is included. This attribute is a data structure containing an identifier of the sensor and its associated sensor type.

Histogram: The possibility to store the histogram associated with image blobs, expanding their information, is offered by this subclass. This information is used both at the blob and the higher levels, e.g. when merging blobs detected by multiple cameras. *OpenCV* offers both a data structure dedicated to this task and a set of functions to work with in order to store the histogram.

3.1.7.2.2. Description of the Object subclasses The Object subclasses have been implemented to manage the information associated with real-world objects. These objects are obtained from one or more blobs that are captured by one or more sensors. Although the current implementation of the objects of INT^3-Horus is fully functional, the model remains open to collect the needs of new algorithms. However, the currently defined set of attributes for objects has proven very powerful for the management of objects in two-dimensional images and the real three-dimensional world. Next, a list of the subclasses of the Object Class are described.

Blob Reference: As previously mentioned, objects are formed from the information of one or more blobs. Therefore, an attribute is defined to reference objects of *Blob* type related to the blobs that from them. The blobs are stored in the list of blobs for each sensor and for each timestamp. For this reason, the *refBlob* structure is provided with a parameter to identify the sensor which detected a given blob, and another parameter to indicate the position of the blob within the list of blobs for the given sensor. Finally, there is an attribute to indicate the timestamp of the blob detection.

Label: This subclass identifies the objects and stores their labels. This parameter is critical as it used in the management methods of the Object Class to access the objects.

Trajectory: This subclass stores: the trajectory of each object based on the points where it was detected in previous timestamps; its future position based on the given trajectory; and, the current speed and direction of the object. The attribute is defined as a hash table whose index is the acquisition timestamp for the positions comprising the trajectory, while the table values are formed by three-dimensional points.

Direction and Speed: The direction and velocity of the object are encoded as a three-dimensional coordinate, where the direction is given by the vector starting from position {0,0,0} to the current objects's position, and the speed is the magnitude of the vector.

Orientation: This subclass stands for object orientation. The most important use of this subclass is activity recognition.

First Detection Time and Last Detection Time: The first attribute defines the time when an object has been detected, this is, the first timestamp when the object was detected. The second one indicates the timestamp associated with the last detection of the object. This may not correspond to the current moment of execution, since an object may not have been detected for some time, remaining in memory to facilitate the operation of some levels such as Object Tracking.

Skeleton: This subclass contains a description of the skeleton of the object. It has been included to facilitate the interpretation of events and activities at higher levels. This attribute depends on the final applications developed in the framework, as the skeleton of a person receives a different treatment to the skeleton that can be obtained to characterize an object like a car. A skeleton is considered as a graph with a set of nodes (the start and end points) and a series of arcs (lines which join these positions). Thus, the repeated positions indicate the joints of the skeleton segments.

Initial Coordinate and Final Coordinate: Unlike blobs, objects have a projection into the real world. Therefore, two subclasses have been added that provide the start and the end points of the "bounding box" which links the object to the real world. For instance, such a "bounding box" is obtained from a range sensor which provides three-dimensional information, or from a multi-camera system where it is possible to get information about the objects' dimensions. Moreover, if the information is two-dimensional, the missing depth dimension is approximated through the objects' models. The attributes used in this case are two-dimensional points: *realWorldOrigin* and *realWorldSize*. Although the last parameter seems confusing, it is actually very useful for optimizing the calculations, since it is defined as the opposite corner to the origin of the bounding box.

Class: This subclass is used to indicate a class of object. For example, in case of traffic monitoring, it is possible to distinguish among several vehicle types (cars, trucks, motorcycles, and so on).

3.1.7.3. Relations

Levels and DataType classes are linked through a set of relations *hasInput* and *hasOutput* that administrate the types of data available at each level. Relations *isInputOf* and *isOutputOf* are inferred due to the symmetric property of the previous relations. Table 3.1 shows inputs and outputs in their respective levels, which are sorted following a growing order of information abstraction. Notice that the information flows from the output of a lower level to the input of the following one.

To illustrate the previous relations in depth, two examples are provided. On the first hand, Figure 3.8 shows the relation *hasInput* with its symmetric relation *isInputOf* for a specific level. In the example, an instance is added to the Sensor Fusion level, namely *MultisensorFusion*. As shown in Table 3.1, the Sensor Fusion level permits three inputs. The Image data type stores an *ImageList* acquired from a visual sensor. The Array data type stores the readings from several sensors parameter through *ArrayList*. Analogously, the Matrix data type is instantiated in a *MatrixList*. The figure not only shows the data type class hierarchy but also the relations between each data type instance with the level instance.

The second relation, *hasOutput*, is shown in Figure 3.9. The Activity Detection level is used as an example to describe the interaction with the data types. In this case, the level in charge of activity interpretation has only one output (see Table 3.1). An individual named *ActivityList* is created to store and manage a list of activities. The figure shows its relations with an Activity Detection instantiation,

Table 3.1: Levels and DataType relations

Level	**hasInput**	**hasOutput**
Data Acquisition	Sensor Data	Image, Matrix, Array
Sensor Fusion	Image, Matrix, Array	Image
Pixel-level Segmentation	Image	Image
Pixel-level Segmentation Fusion	Image	Image
Blob Detection	Image	Blob
Object Identification	Blob	Object
Object Classification	Object	Object
Object Tracking	Object	Object
Event Detection	Object	Event
Event Fusion	Event	Event
Activity Detection	Event	Activity

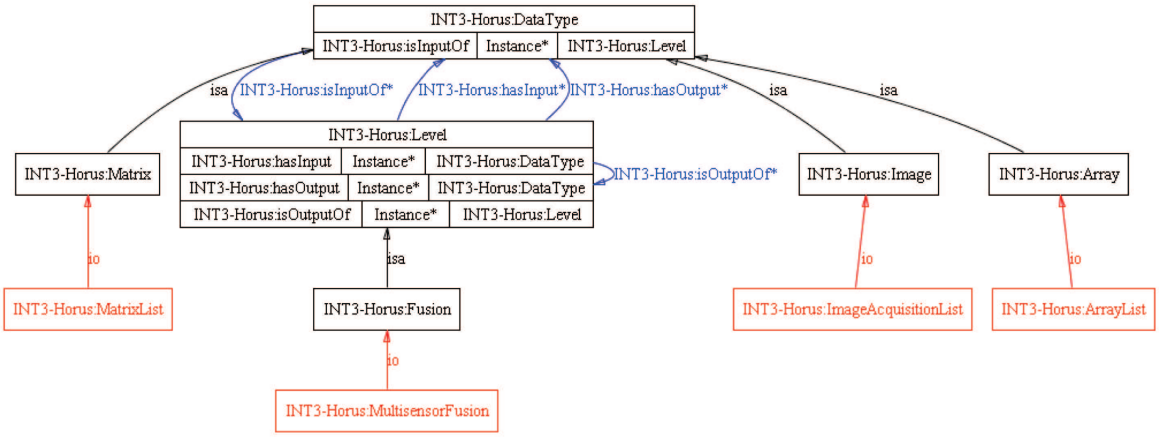


Figure 3.8: *hasInput* relation for a *MultisensorFusion* instance.

that is, *FuzzyActivityDetection*, together with the class hierarchy.

3.2. Niveles seleccionados para la detección robusta de humanos

Como ya se ha comentado al principio del capítulo, la presente propuesta de detección robusta de humanos se ha desarrollado a partir del marco de trabajo general INT^3-Horus (Fernández-Caballero et al., 2013; Castillo, 2012). Este marco está concebido, precisamente, como un entorno de desarrollo para cualquier tipo de sistema que desempeñe tareas de monitorización e interpretación de actividades. Este objetivo resulta complejo dado el gran número de escenarios y actividades que pueden ser plasmados (Kieran and Yan, 2010; Sokolova et al., 2012)), pero ya se ha visto su gran utilidad en algunas aplicaciones previas (por ejemplo, (Gascueña and Fernández-Caballero, 2011; Gascueña et al., 2011, 2012)).

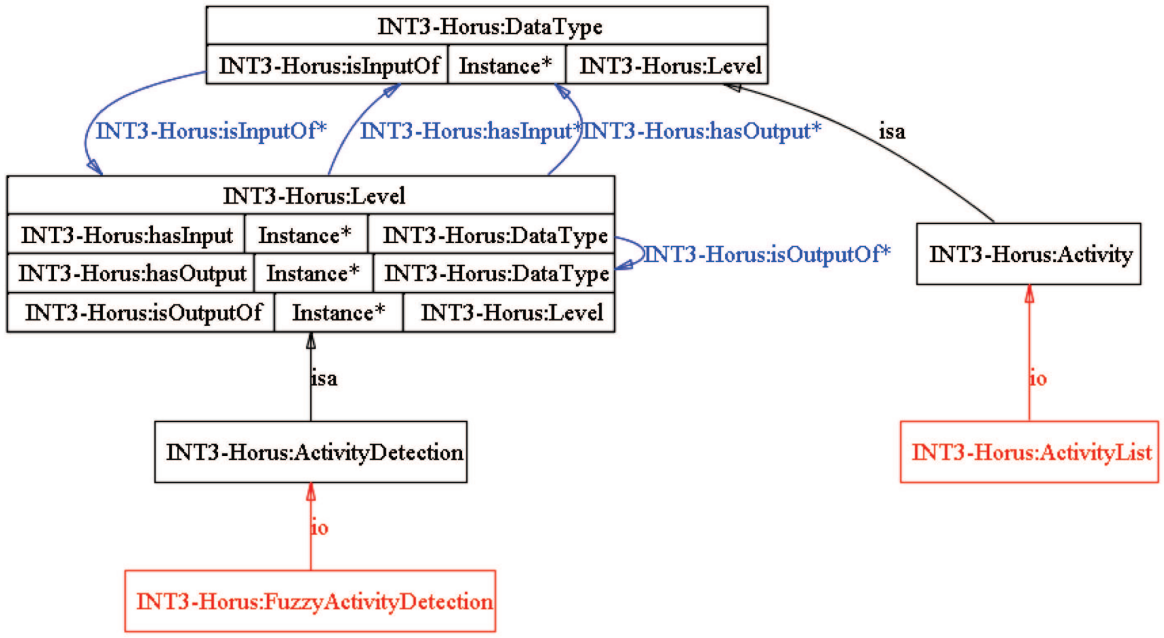


Figure 3.9: *hasOutput* relation for a *FuzzyActivityDetection* instance.

La mayor de las ventajas del marco de trabajo INT^3-Horus es que, aunque se proponen una serie de niveles con el fin de cubrir todos los pasos de un sistema genérico multisensorial de interpretación de actividades (véase, por ejemplo, (Pavón et al., 2007; Rivas et al., 2011; Fernández-Caballero et al., 2012)), la filosofía de este entorno permite que se pueda adaptar un conjunto flexible de niveles a un determinado sistema final, tal y como se muestra en Carneiro et al. (2012) y Costa et al. (2012). Por otra parte, el framework permite integrar fácilmente el código, proporcionando a los usuarios plantillas de cada módulo para introducir su código en las mismas. Estas plantillas ya tienen las conexiones necesarias para acceder al resto de componentes de INT^3-Horus conteniendo, aparte del modelo de datos y la interfaz de usuario, el controlador para lanzar la ejecución de cada módulo.

Así pues, de los niveles previamente descritos, se han seleccionado los niveles que se ajustan a las necesidades que se plantean en la actual tesis, a saber *Captura*, *Segmentación*, *Fusión*, *Identificación* y *Seguimiento*. La estructura de estos niveles, junto con sus entradas y salidas, puede observarse en la Figura 3.10.

A continuación se describe con mayor detalle cada uno de los niveles escogidos.

3.2.1. Nivel de *Captura*

El primer nivel consiste en capturar simultáneamente (y síncronamente) vídeos con una cámara en infrarrojo y otra en color.

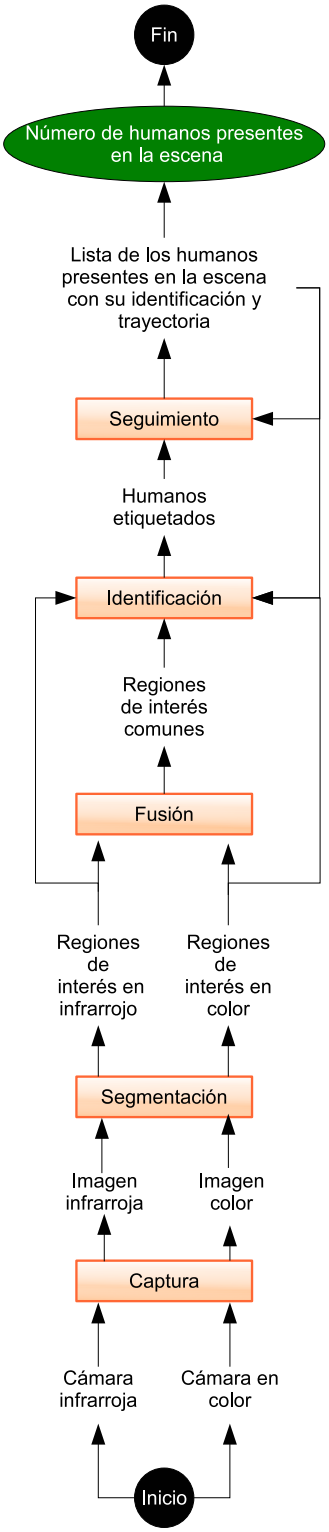


Figura 3.10: Niveles de INT^3-Horus usados para la detección robusta de humanos.

Figura 3.11: Montaje realizado para la captura simultánea en los espectros infrarrojo y color. (a) Vista trasera. (b) Vista delantera.

3.2.1.1. Montaje para la *Captura*

Las dos cámaras están colocadas en paralelo y enfocan hacia un punto de un mismo escenario, con el objetivo de obtener dos vistas similares de una misma escena. En la Figura 3.11a y Figura 3.11b podemos observar una vista trasera y delantera, respectivamente, del montaje realizado.

La captura simultánea y sincronizada de ambas cámaras es posible gracias a la utilización de un codificador de vídeo con capacidad de capturar la imagen por las dos cámaras en un mismo instante. El codificador Axis Q4704 usado posee cuatro entradas de vídeo a las que se pueden conectar hasta cuatro cámaras (ver Figura 3.12). Los fotogramas capturados por las cámaras conectadas al codificador son separados en canales (uno por cada entrada de vídeo) y añadidos a un buffer que emite el vídeo *streaming* en formato *mjpg* y al que se puede conectar una aplicación externa para captar los fotogramas del canal correspondiente. A su vez, estos canales están protegidos por un sistema de seguridad que impide que presencias no deseadas puedan acceder a la información que están grabando las cámaras. Conectando este codificador y el ordenador receptor a un router pueden capturarse directamente los fotogramas desde una red interna habilitada especialmente para ello, lo que evita la instalación obligatoria de una tarjeta capturadora de vídeo en el ordenador desde el que se captura. Al presentar una interfaz independiente de las cámaras (vídeo emitido por red) como salida, se hace posible la abstracción del protocolo usado por cada una. Gracias a este hardware, se consigue también grabar con un retardo mínimo un fotograma desde que es capturado por la cámara hasta que es recibido por el ordenador.

3.2.1.2. Captura en el espectro infrarrojo

Todas las pruebas se han realizado con la cámara infrarroja *FLIR A-320*, que provee una sensibilidad entre -40° y 70° y que captura a una resolución de 320×240 píxeles con una velocidad de 5 fotogramas por segundo. Se ha decidido activar el rango dinámico de temperaturas en esta cámara,

Figura 3.12: Codificador Axis Q4704 utilizado para la captura simultánea en los espectros infrarrojo y color.

(a) (b)

Figura 3.13: Cambio automático de la escala de niveles de gris de la cámara en infrarrojo. (a) Escena sin humano. (b) Escena con humano.

es decir, que el píxel más claro siempre será aquél que tenga la temperatura máxima detectada en la escena. Del mismo modo, más oscuro será el que presente la mínima temperatura presente. Esta decisión se ha tomado para poder distinguir siempre a los humanos con la mayor claridad posible, independientemente de las condiciones térmicas (temperatura ambiente) de la escena. Sin embargo, como resultado los humanos no siempre presentarán el mismo nivel de gris en la escena, ya que el nivel medio de gris se reajusta automáticamente cuando se produce un cambio de temperatura media en el escenario monitorizado.

Tal y como se puede apreciar en las dos imágenes de un mismo entorno mostradas en la Figura 3.13, el nivel de gris del césped (que cubre gran parte de la imagen) cambia de una imagen a otra para cubrir con 256 niveles de gris todo el intervalo de temperaturas presente en la escena. En la Figura 3.13b el punto más cálido de la escena se corresponde ahora con la cabeza del humano. Ahora, la temperatura máxima en el escenario es $7{,}6°$ mayor que antes de que el humano entrara (es decir, la temperatura media ha variado respecto de la Figura 3.13a). De esta forma, la temperatura más baja aparece siempre de color negro en la imagen capturada, mientras que los puntos con mayor calor se muestran en color blanco.

3.2.1.3. Captura en el espectro visible

Por otra parte, la captura en color se realiza con una cámara *SONY FCB-EX780bp* que captura a 384×288 píxeles con una velocidad ajustada a 5 fotogramas por segundo (con el fin de sincronizar los fotogramas capturados con aquellos adquiridos por la cámara en infrarrojo).

3.2.2. Nivel de *Segmentación*

El siguiente nivel de la presente propuesta consiste en la segmentación de humanos en color e infrarrojo sobre las imágenes capturadas en el nivel de procesamiento anterior. Este nivel de *Segmentación* tiene como principal objetivo el detectar posibles candidatos a humanos en ambos espectros. El nivel incluye una serie de algoritmos de segmentación que han sido implementados para sacar partido a las características propias de cada uno de los espectros, así como para poder hacer frente a diferentes entornos y escenarios monitorizados. De este modo la propuesta permite elegir (empíricamente) cuáles de los algoritmos de segmentación resultan más adecuados para el entorno particular en el que se requiera monitorizar humanos para su detección robusta. Cabe señalar que en este nivel de *Segmentación* se eliminan falsos positivos que puedan aparecer, imponiendo una serie de restricciones de forma que caracterizan a las personas (véase, su tamaño, su posición y su proporción entre altura y anchura).

Los candidatos a humanos detectados en este nivel serán posteriormente refinados en el nivel de *Fusión*, donde se usará un sistema basado en reglas dependiente del espectro para detectar de un modo más robusto los humanos.

3.2.2.1. *Segmentación* en el espectro infrarrojo

A la hora de segmentar humanos, la principal ventaja de usar el espectro infrarrojo es que las personas suelen aparecer con mayor luminosidad dentro de la imagen. Esto es cierto sobre todo cuando la temperatura ambiente no es demasiado elevada y en escenarios con poca iluminación en una secuencia capturada de noche. Sin embargo, tal y como mencionamos en el capítulo 1, en Goubet et al. (2006) se explica que el infrarrojo también presenta múltiples inconvenientes. Especial mención se merece la aparición de los llamados “halos”. Los halos rodean a los objetos que tienen un gran contraste con el fondo, de forma que los puntos brillantes en un fondo oscuro estarán rodeados por una región más oscura que dicho fondo y viceversa. Estos halos que se muestran en las imágenes se deben a que la radiación secundaria calienta los sensores sobre una superficie mayor que la imagen actual del objeto. Otro inconveniente, especialmente importante, lo constituye el hecho de que, cuando la temperatura ambiental es muy elevada (como es el caso en verano), los humanos aparecen más fríos que el entorno o incluso a la misma temperatura, dificultando su distinción incluso por el ojo humano, tal y como se puede apreciar en la Figura 3.14a.

Muchos algoritmos intentan minimizar el impacto de este problema mediante la aplicación de una resta de fondo. No obstante, se ha descartado esta opción debido a que la cámara infrarroja suele

(a) (b)

Figura 3.14: Imagen capturada a temperatura alta. (a) Espectro infrarrojo. (b) Espectro visible.

cambiar su escala para adaptarla al rango de temperaturas en la escena. Dados estos cambios, se ha preferido realizar una detección de humanos en infrarrojo que no dependa del fondo (excesivamente cambiante en entornos de vigilancia reales) y solucionar los inconvenientes debidos a los reajustes de escala mencionados, realizando una fusión posterior de regiones de interés (manchas detectadas) del infrarrojo con regiones de interés del color. La idea se desarrolla en su totalidad en el nivel de *Fusión*.

Tal como se ha comentado anteriormente, para la segmentación en el espectro infrarrojo (al igual que en el espectro de color) se han estudiado (e implementado) varias aproximaciones (ver (Sokolova et al., 2013; Fernández-Caballero et al., 2011a)). La primera utiliza únicamente la información de la intensidad del calor del último de los fotogramas capturados (es decir, el fotograma actual), sacando partido a la suposición de que los humanos presentan manchas más calientes que el resto de los objetos presentes en la imagen. Sobre este método, se fue añadiendo posteriormente información del movimiento de los objetos presentes en la escena y del movimiento inherente a la cámara que está capturando las imágenes. De nuevo, recuérdese que la intención es la de establecer cuál es la propuesta más apropiada para un entorno monitorizado concreto.

3.2.2.2. *Segmentación* en el espectro visible

Si bien en la literatura existen numerosos algoritmos de segmentación de objetos en vídeo color (explicados en el capítulo anterior), se ha podido apreciar que solamente unos pocos de estos algoritmos se centran específicamente en la detección de humanos. Además, en numerosas ocasiones la clasificación de los objetos detectados se hace con la ayuda en un nivel posterior de seguimiento de objetos en la escena.

También aquí se han probaron diversas aproximaciones. Esta vez se ha trabajado con el historial de movimiento de los objetos presentes en la escena y su comparación con una imagen de referencia o fondo (ver (Fernández-Caballero et al., 2011b, 2008)). En base a los resultados que se obtengan en

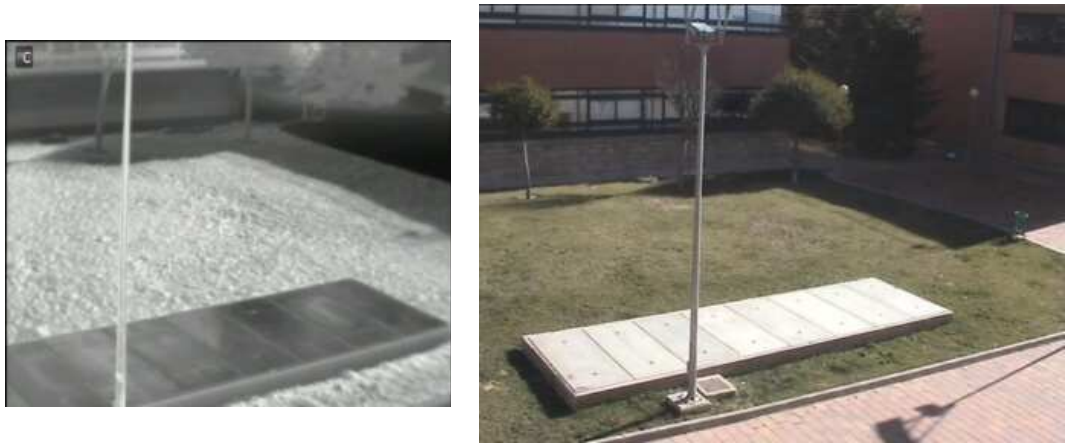


Figura 3.15: Profundidad de campo de las cámaras usadas. (a) Imagen capturada con una cámara térmica infrarroja. (b) Imagen tomada con una cámara en color.

cada tipo de escenario, se podrá elegir la propuesta más apropiada.

3.2.3. Nivel de *Fusión*

El siguiente nivel a realizar es el de la *Fusión* de los candidatos a humanos segmentados en el nivel anterior para los vídeos color y infrarrojo. Este nivel constituye una de las aportaciones más importantes y novedosas de este trabajo. Es de especial interés este nivel de *Fusión*, ya que ambos espectros presentan diversas limitaciones y puntos fuertes que deben ser tenidos en cuenta a la hora de elaborar un buen algoritmo de fusión.

Tras estudiar las diversas alternativas encontradas en la literatura a la hora de realizar algoritmos de fusión, se ha optado por realizar una *fusión a nivel de regiones*. El motivo es que la implementación de este tipo de fusión presenta ventajas muy interesantes a la hora de trabajar con cámaras de diversos espectros. Una de las ventajas es que se posibilita el uso de reglas inteligentes de fusión de acuerdo a las características propias de cada espectro. Por otra parte, se consigue mitigar problemas de nitidez que surgen en las situaciones (ya mencionadas) de oscuridad o mucho calor.

3.2.3.1. Calibración para la *Fusión*

En la Figura 3.15 se muestra cómo las dos cámaras usadas presentan una profundidad de campo distinta. De hecho, en la Figura 3.15a se aprecia que la cámara FLIR presenta una menor profundidad de campo abarcando una superficie de la escena significativamente menor que la cámara en color (ver Figura 3.15b).

Es indispensable que el algoritmo de *Fusión* trabaje sobre ambas imágenes utilizando las mismas

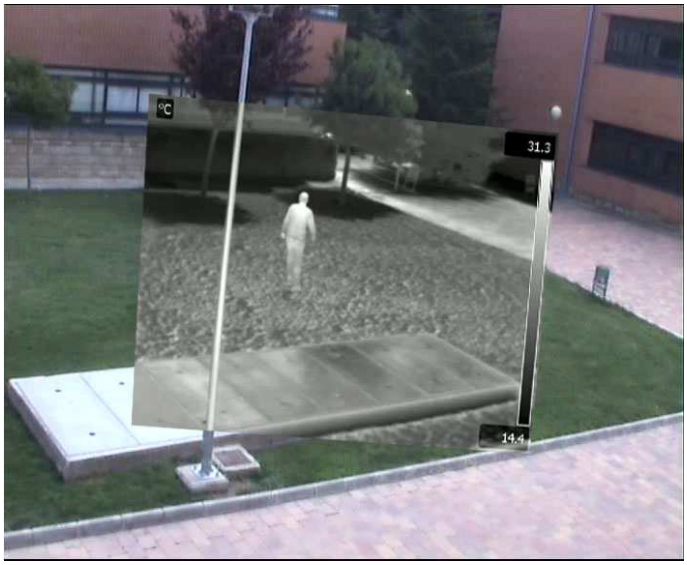


Figura 3.16: Resultado de la calibración de las imágenes en infrarrojo y color.

coordenadas con el fin de poder fusionar posteriormente la información obtenida de ambas. Además, el objetivo es abarcar la mayor extensión de escenario posible aunque solo sea cubierto en parte por una de las dos cámaras únicamente. El motivo es que la zona no común proporcione información de apoyo a la fusión cuando un humano pasa de la zona no cubierta por ambos espectros a la zona cubierta por ambas cámaras. Se hace pues necesario realizar, como primer paso de la *Fusión*, una calibración para tener un campo de visión común de las dos imágenes

Con este objetivo, se toma como referencia la imagen en color, y se aplican las transformaciones geométricas necesarias a la imagen en infrarrojo. El resultado aplicado a un escenario concreto se muestra en la Figura 3.16.

3.2.3.2. Descripción general de la *Fusión*

Una vez realizadas las transformaciones geométicas descritas, comienza el procesamiento propiamente dicho de la *Fusión*. Se trata de obtener los humanos presentes en la escena a partir de las regiones de interés (ROI) asociadas a los candidatos a humanos obtenidos en el nivel anterior de *Segmentación*. La fusión se basa en una serie de reglas en las que se tienen en cuenta varios casos bien diferenciados.

En primer lugar, se procede a fusionar en la zona común a ambos espectros los candidatos a humanos descritos por su ROI. A partir de las ROIs obtenidas en ambos espectros (infrarrojo y color), en primer lugar, se va a utilizar la información de las áreas comunes que se han segmentado previamente. Se trata, pues, de obtener inicialmente un conjunto de ROIs similares entre sí y que se han detectado en ambos espectros. Por ejemplo, en la Figura 3.17 se puede observar que el humano ha sido detectado en los dos espectros. Como era previsble, la segmentación no es perfecta en ninguno de los dos

(a) (b) (c)

Figura 3.17: Fusión de regiones de interés en infrarrojo y color. (a) ROI en infrarrojo. (b) ROI en color. (c) ROI obtenida tras la fusión.

espectros. Véase que en el infrarrojo no se han detectado los pies (ver Figura 3.17a), mientras que en el color se ha perdido parte de la cabeza y se ha tomado como región perteneciente al humano gran parte de la zona de césped situada por debajo de los pies (ver Figura 3.17b). Estas diferencias pueden unificarse en una zona común estableciendo con el infrarrojo la parte superior de la cabeza y formando una nueva región de interés, tal y como se puede apreciar finalmente en la Figura 3.17c.

Pero, además de los candidatos a humanos obtenidos al mismo tiempo en ambos espectros, se tienen también en cuenta aquellos obtenidos únicamente en el vídeo infrarrojo o en el vídeo en color, según un valor de confianza otorgado a cada espectro en base a las características fundamentales de las imágenes capturadas. Como se verá más adelante, en el capítulo 4, las características principales son la iluminación media y la desviación típica de la imagen en el espectro infrarrojo, y la intensidad media en el espectro visible.

Puede darse el caso de que el infrarrojo funcione mejor que el color y no solo en situaciones oscuras, sino en zonas con mucha sombra como se puede ver en la Figura 3.18a. En este caso, si las características de la imagen en infrarrojo aportan la confianza suficiente y se ha segmentado algún candidato a humano (ver Figura 3.18b), se tendrá en cuenta la detección del infrarrojo tal y como se aprecia en la Figura 3.18c).

Análogamente, podemos encontrarnos con el caso opuesto. En la Figura 3.19b se aprecia que el humano presente en la escena es prácticamente imposible de distinguir del fondo en el espectro infrarrojo, mientras que la escena en color permite distinguirlo perfectamente, como se muestra en la Figura 3.19a. En este caso, la característica fundamental asociada al espectro visible conlleva la asignación de una confianza alta a la segmentación. Por lo tanto, el resultado final será el mostrado en la Figura 3.19c.

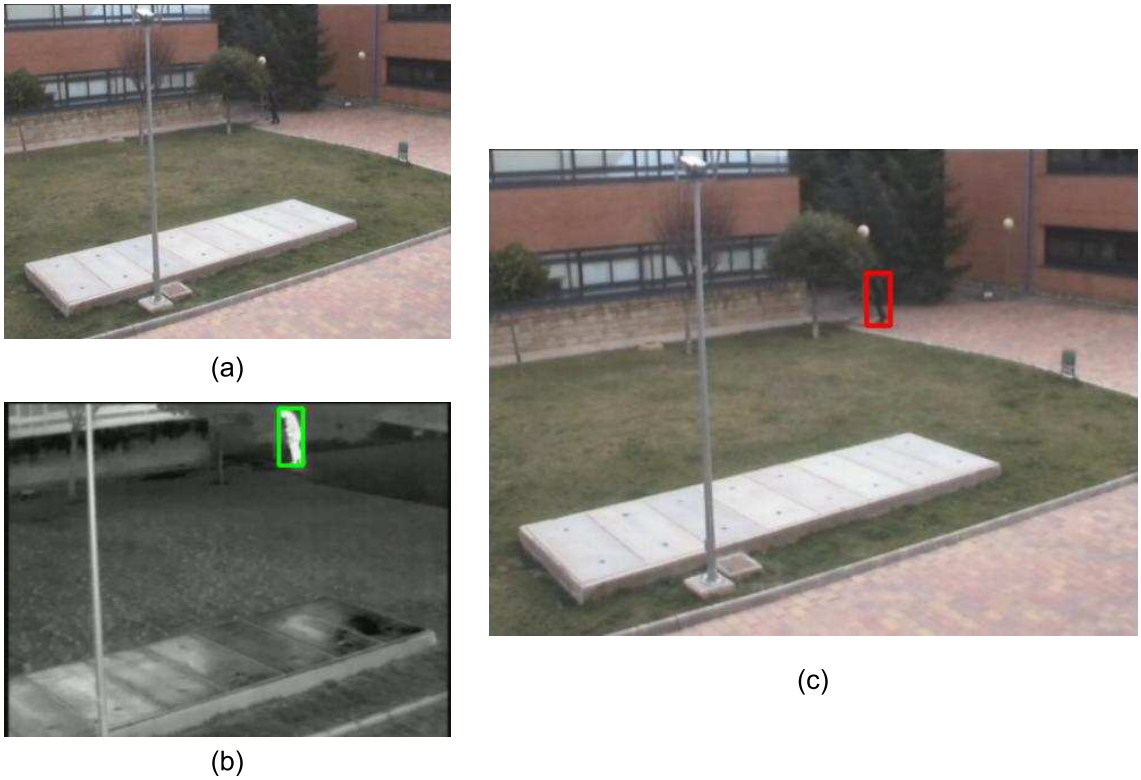


Figura 3.18: Mejora de la detección de humanos tras una mala segmentación en el espectro visible. (a) Segmentación en el espectro en color. (b) Segmentación en el espectro infrarrojo. (c) Resultado tras la fusión.

3.2.4. Nivel de *Identificación*

El siguiente nivel consiste en elaborar una lista con los humanos detectados en la secuencia y sus características. Para identificar a los humanos se utilizará un identificador numérico, denominado *etiqueta*. Así, el primer paso consiste en mirar si la última posición de cada humano detectado en el fotograma actual se encuentra cercana a la de alguno de los humanos ya etiquetados en anteriores fotogramas. En caso afirmativo, se le identificará asignándole la etiqueta correspondiente y se procederán a actualizar sus propiedades. En caso contrario, se procederá a crear una nueva etiqueta para el nuevo humano detectado. Se puede apreciar un ejemplo de etiquetado de una secuencia en la Figura 3.20, en la que las dimensiones de las regiones que contienen a los humanos se adaptan progresivamente a las características de los humanos en base a la posición que estos ocupan en la escena. Es decir, se tiene en todo momento en cuenta la geometría tridimensional de la escena.

3.2.5. Nivel de *Seguimiento*

Finalmente, tenemos el nivel de *Seguimiento*. En base a la lista de las últimas posiciones de todos los humanos, se pueden calcular sus trayectorias y sus velocidades. Para ello, se procede a realizar un

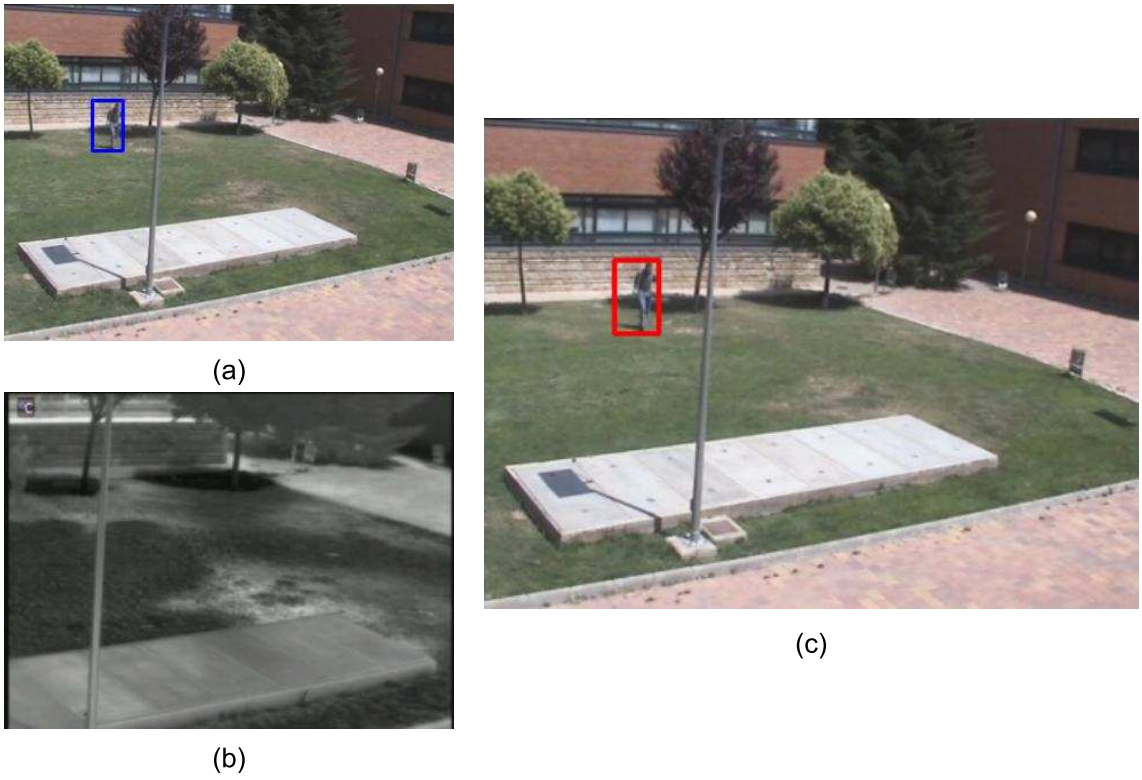


Figura 3.19: Mejora de la detección de humanos tras una mala segmentación en el espectro infrarrojo. (a) Segmentación en el espectro visible. (b) Segmentación en el espectro infrarrojo. (c) Resultado tras la fusión.

análisis de trayectorias sobre la lista de los humanos presentes en la escena, tanto si han sido detectados en el último fotograma analizado como si no. El análisis de trayectorias puede predecir la posición de cada humano en el siguiente fotograma, y, asumiendo que un humano no puede desaparecer de en medio de una escena, se mejora sensiblemente la segmentación, por lo que se puede concluir que este análisis proporciona apoyo al nivel de *Segmentación* paliando posibles falsos positivos puntuales que hayan podido aparecer.

Es también en este nivel donde se estima si un humano ha abandonado o no la escena. A la hora de realizar esta estimación, es necesario valorar la probabilidad de que un humano esté presente en la escena. En esta valoración influyen diversos factores tales como su última posición conocida o la confianza establecida por la fusión a la hora de realizar la detección del humano. Así, se estima que un humano tiene más posibilidades de encontrarse presente en la escena si fue confirmado por ambos espectros a la hora de realizar la fusión de regiones de interés que si fue detectado únicamente por uno de ellos. También se valora el número de fotogramas que el humano lleva presente en la escena. De este modo se eliminan los falsos positivos que pueden aparecer debidos a cambios de iluminación puntuales en la escena. También se tiene en cuenta la posición del humano, ya que, si un humano se encontraba cerca de los límites de la escena, es muy posible que la haya abandonado. Sin embargo, pueden aparecer excepciones a esta norma que también se hace necesario valorar, como el hecho de

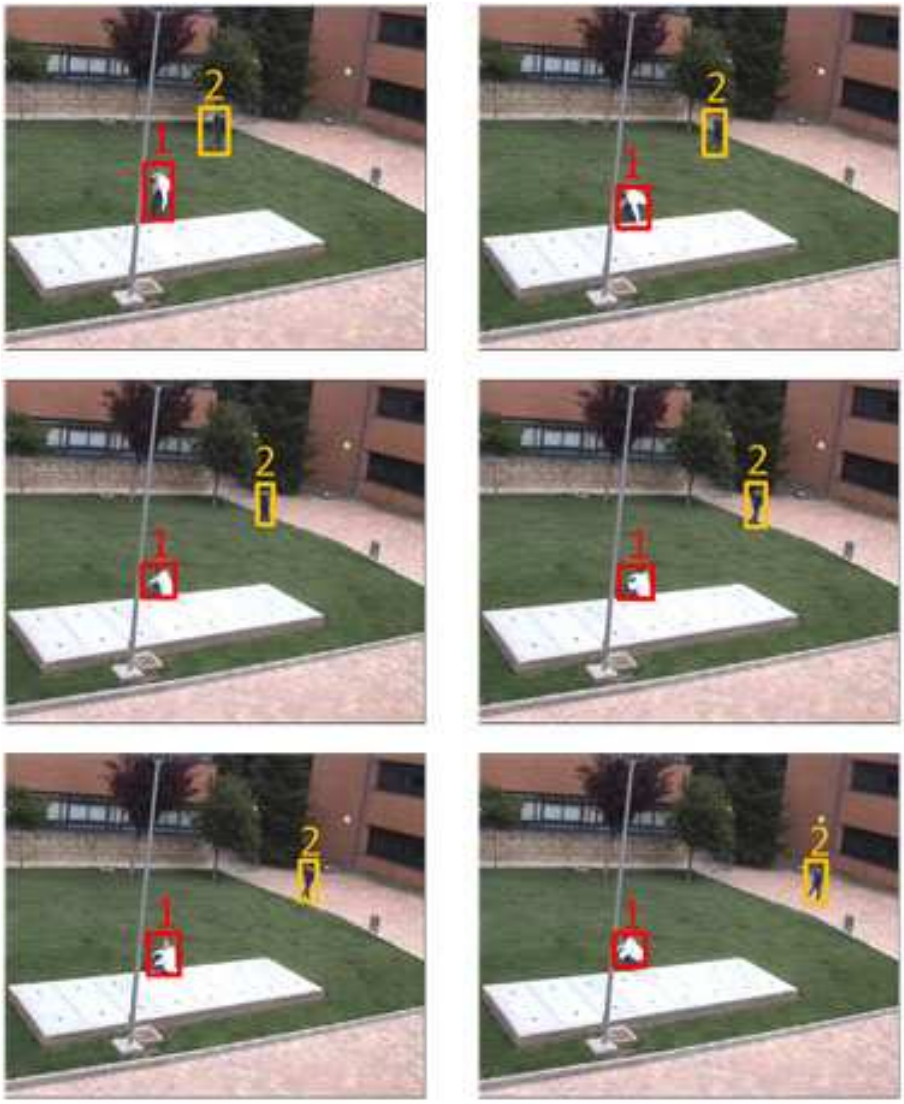


Figura 3.20: Ejemplo de identificación en una secuencia de fotogramas.

que un humano pueda haber desaparecido de la escena aun encontrándose lejos de los límites. Puede darse, en efecto, que haya pasado a formar parte de un grupo y ahora mismo se encuentre ocluido por otras personas. Se puede apreciar que es un caso opuesto a la incorporación de un nuevo humano (antes mencionada).

Atendiendo a su función, puede afirmarse que este nivel y el anterior están estrechamente relacionados, pudiéndose realimentar entre ellos. Se puede apreciar un ejemplo de análisis de trayectorias sobre una secuencia en la Figura 3.21.

Figura 3.21: Ejemplo de análisis de trayectorias (dibujado sobre un fotograma), teniendo en cuenta las últimas posiciones de los humanos en la secuencia.

3.3. Evaluación de la robustez del sistema

Los algoritmos de visión artificial no necesitan proporcionar solamente la respuesta sino también la fiabilidad de la respuesta en diversas condiciones, a fin de que los algoritmos puedan ser útiles en su uso. Sin embargo, una gran parte de la investigación en visión por computador se centra en el desarrollo de soluciones con un esfuerzo limitado en caracterizar, de un modo sistemático, la fiabilidad de la solución bajo diferentes condiciones. Es especialmente importante identificar aquellas condiciones en las que el rendimiento del algoritmo se puede optimizar y las condiciones bajo las cuales el algoritmo puede fallar. Este objetivo se logra mediante la evaluación del rendimiento.

Nuestro enfoque utiliza una evaluación empírica del rendimiento, es decir, estudiaremos la precisión del sistema robusto de detección de personas propuesto mediante datos de imágenes reales y compararemos la salida de los algoritmos con algunos datos *groundtruth*. La ventaja de este enfoque es su realismo. Las principales dificultades con este enfoque incluyen la dificultad en la obtención de datos *groundtruth*, y la suficiencia y la representación de los datos de las imágenes seleccionadas.

Por ello, de cara a evaluar la mejora del sistema propuesto de detección de humanos, se hace necesario evaluar tanto los algoritmos de segmentación implementados como el algoritmo final de fusión resultante, utilizando una serie de métricas estadísticas usadas extensamente por la comunidad de visión artificial, tales como la precisión, la sensibilidad y el F-score. Estas estadísticas se calculan tal y como muestran las ecuaciones (3.5), (3.6) y (3.7), respectivamente.

$$precision = \frac{VP}{VP + FP} \tag{3.5}$$

$$sensibilidad = \frac{VP}{TP + FN} \tag{3.6}$$

$$F - score = \frac{2 \times precision \times sensibilidad}{precision + sensibilidad} \tag{3.7}$$

donde VP (verdaderos positivos) es el número de detecciones que el sistema ha realizado de forma correcta en la secuencia, FP (falsos positivos) es la cantidad de detecciones erróneas que ha realizado el sistema, contando humanos que realmente no se encuentran en la escena, y FN (falsos negativos) es el número de humanos presentes en la escena que no son detectados por el sistema.

La precisión indica el porcentaje de verdaderos positivos respecto al total de detecciones (VP sumados a FP), es decir, la probabilidad de que una detección del sistema sea correcta y realmente pertenezca a un humano. Por otra parte, la sensibilidad muestra la probabilidad de que sea detectado un humano que está realmente presente en la escena. Una sensibilidad del $100\,\%$ significará que todos los humanos en la escena fueron detectados, mientras que una precisión del $100\,\%$ indica que, cuando el sistema realiza una detección, es siempre cierto que dicha detección pertenece a un humano. Finalmente, el F-score proporciona una visión en conjunto del comportamiento del sistema, considerando tanto la precisión como la sensibilidad. El F-score puede interpretarse como una media ponderada de la precisión y la sensibilidad, donde un sistema ideal tendrá un valor de 1 y le será asignado un valor de 0 a un sistema que nunca funcione correctamente.

Obviamente, a la hora de realizar la evaluación del sistema, es necesario especificar qué se entiende por verdadero positivo y qué no, así como definir qué tipo de datos se tomarán como referencia a la hora de evaluar los algoritmos, es decir como *groundtruth*. Debido a que se ha trabajado con más de setenta mil imágenes capturadas a propósito para la presente tesis (es decir, no existen en internet *groundtruths* ya realizados para las distintas secuencias), el trabajar a nivel de píxel para evaluar la exactitud de las detecciones hubiera supuesto un trabajo enorme que se escapa del ámbito del trabajo actual. Por ello, se ha optado por etiquetar cada fotograma con el número de humanos presentes en el mismo, así como una referencia de la zona en la que se encuentra el humano. Este número (y la zona) se comparará con el número de humanos detectados (y las zonas de detección) por cada algoritmo con el fin de poder validar su correcto funcionamiento.

3.4. Conclusiones

En este capítulo se han detallado los pasos generales que se han seguido durante la presente tesis. En primer lugar se ha enmarcado la propuesta de detección robusta de humanos en el marco de trabajo INT^3-Horus. Posteriormente han sido desglosados los niveles que se han escogido de dicho marco de trabajo, correspondientes a las etapas que lleva a cabo el desarrollo implementado en la tesis, explicando en profundidad el objetivo de cada una de las mismas. Tras una etapa inicial de *Captura* de imágenes simultánea en ambos espectros, se procede a obtener (por separado) los candidatos a huma-

nos presentes en cada uno de ellos en el nivel de *Segmentación*. En este nivel se han implementado diversos algoritmos, utilizando tanto la información del color como la del espectro infrarrojo, con el fin de elegir los más apropiados para los objetivos y el entorno de la tesis.

Las segmentaciones obtenidas de ambos espectros (realizadas con los algoritmos escogidos) se unifican en un nivel de *Fusión*, teniendo ya una única lista de posibles humanos. Sobre estos candidatos se realiza un procesamiento de *Identificación* y *Seguimiento*, con el fin de situar a los humanos detectados en el contexto de la secuencia y poder establecer su historial de trayectorias, paliando así posibles fallos de los niveles anteriores. Este proceso permite, además, estimar las futuras posiciones de los individuos detectados en la escena en el caso de aparición de problemas ocasionales (como las oclusiones), realimentando, a su vez, a los dos niveles anteriores, refinando sus resultados. El resultado final obtenido es el número de humanos de humanos presentes en la escena.

A la hora de realizar la evaluación de los diversos algoritmos implementados (tanto de segmentación como de fusión) se utilizarán una serie de métricas estándar empleadas a la hora de valorar un sistema de visión artificial, utilizando como datos a la hora de valorar estos algoritmos el número de humanos detectados en la escena.

Capítulo 4

Detailed Description of the Processing Levels

The proposed human detection system based on the fusion of color and infrared video is described in detail in this chapter. First, a general overview of the system is provided, giving a brief explanation of how the system's levels work and the way they are connected with each other. Next, the various levels are introduced. In first place the *Video Acquisition* level is described, focusing on its synchronization mechanisms and on the calibration system that is later used to perform the fusion between infrared and visible spectra. Then, the different algorithms used in the *Segmentation* level are described in depth, detailing the human detection approaches employed for the infrared and visible spectra. Finally, *People Fusion and Tracking* is disclosed as a unified level, explaining how the detections of the segmentation in the different spectra are combined with the purpose of improving the human detection performance from the different algorithms. In this sense the robustness of the system is confirmed. The results are reinforced by a tracking algorithm to improve the fusion's outcomes. The final output of the system shows the number of humans detected as well as their location within the monitored scene.

4.1. General System Overview

The designed system is composed of a series of processing levels. Each level uses inputs from the previous levels and/or generates outputs to the later ones. Each one of them is also composed of a series of stages that communicate among each other and sometimes work in parallel. In other occasions the results from other stages are necessary to generate the outputs from a level. A visual representation of the different system levels and how they intercommunicate is shown in Figure 4.1.

The *Video Acquisition* level initially performs *Infrared and Color Video Grabbing*, capturing two simultaneous frames from an infrared and color camera. This process is described in subsection 4.2.1. Once the frames have been synchronized, they are sent to the *Segmentation* level. The image captured

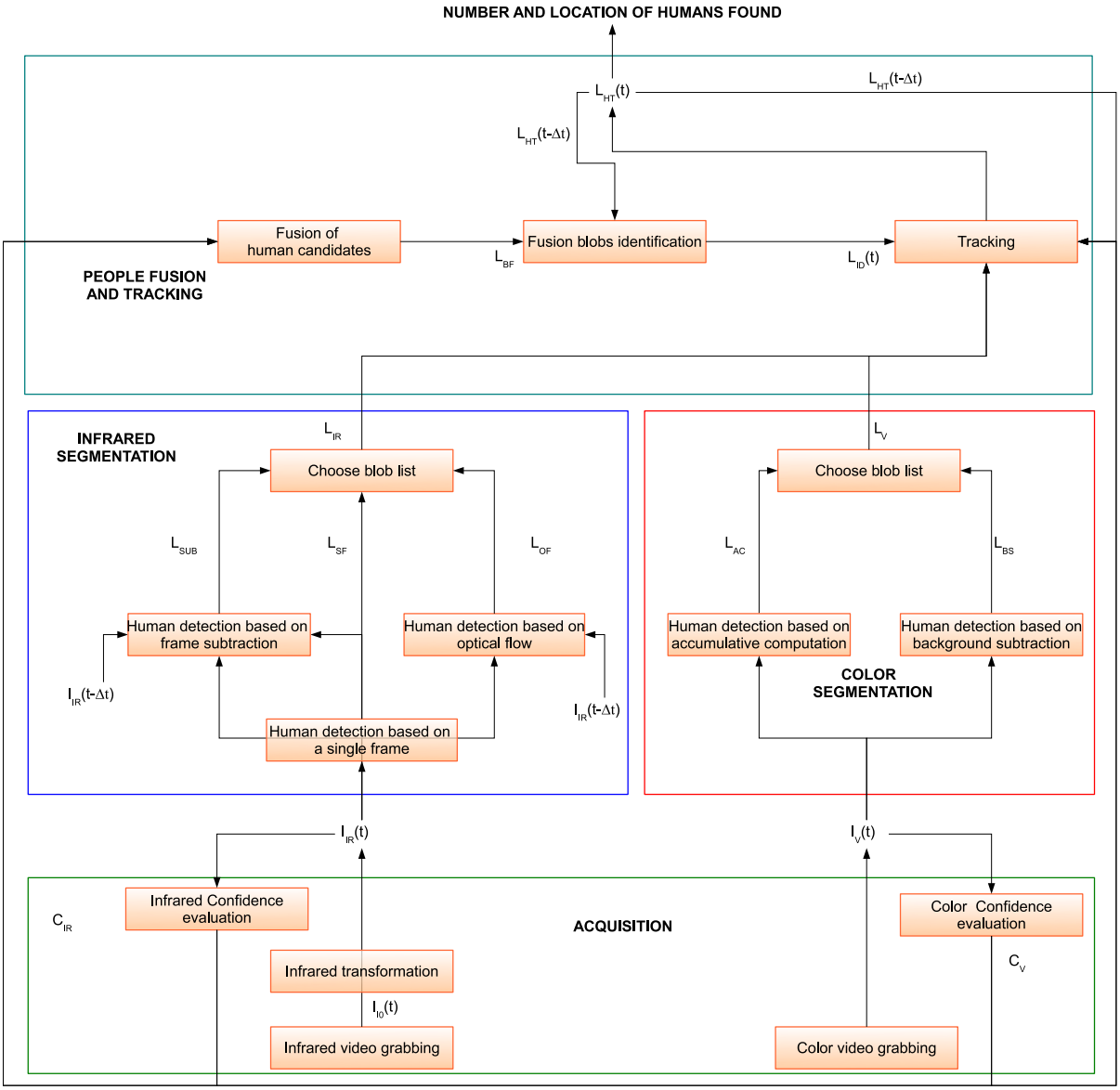


Figura 4.1: Levels used by the system.

by the infrared camera (I_{I0}) is transformed into its equivalent coordinates in the image on the color camera (by a process named as *Infrared transformation*, as seen in subsection 4.2.2) and is used as input to the infrared segmentation algorithms. The frame grabbed by the color camera (I_V) is received by the color segmentation algorithms. Confidence degrees C_{IR} and C_V are set for infrared and visible spectrum, respectively. C_{IR} is based on the contrast of the image grabbed by the infrared camera and C_V depends on the average illumination of the frame acquired by the color camera. These values take a special importance in the higher levels of the system, as it will be seen later on.

The *Segmentation* level works in parallel with infrared and color segmentation algorithms. Moreover, several segmentation approaches are tested for each spectrum. *Segmentation based on a single frame*, *Segmentation based on frame subtraction* and *Segmentation based on optical flow calculation* (described in sections 4.3.1.1, 4.3.1.2 and 4.3.1.3 respectively) are used in infrared. In addition, *segmentation based on accumulative computation* (explained in subsubsection 4.3.2.1) and *segmentation based on background subtraction* (described in subsubsection 4.3.2.2) are tested in the visible spectrum. Notice that while the first approach only uses information from a single frame, the other two algorithms add motion information to improve the segmentation results. In all segmentation methods, a list of blobs containing human candidates is obtained. The blob lists generated by human detection based on a single frame (L_{SF}) are compared to those obtained by human detection based on frame subtraction (L_{SUB}) and human detection based on optical flow (L_{OF}). On the other hand, blob lists resulting from human detection based on accumulative computation (L_{AC}) and human detection based on background subtraction (L_{BS}) are also compared in the visible spectrum. The final results of all these comparisons are precisely blob lists L_{IR} and L_V obtained in the infrared and visible spectra, respectively.

Now, the blob lists L_{IR} and L_V are the inputs to the fusion and tracking levels, composed of three stages intercommunicated between each other. The first stage, named as *Fusion of human candidates* (described in subsection 4.4.1), creates a new list L_{BF} with those blobs considered to be humans with a greater probability. The *Fusion Blobs Identification* stage (explained in subsection 4.4.2) uses L_{BF} and the human list from the previous iteration, $L_{HT}(t-\Delta t)$, as well as confidence degrees C_{IR} and C_V from the infrared and visible spectrum, respectively, to establish an initial list L_{ID} of the humans in the scene and their associated labels. Now, $L_{HT}(t-\Delta t)$ is compared with L_{ID} in the final *Tracking* (described in subsection 4.4.3) to check if any human previously detected in the scene has not been located in the current frame. For each human in this situation, a new test is applied in L_{IR} and L_V since a human previously detected in the scene should not be discarded due to a low confidence degree in the spectrum where he/she has been detected now. If this situation happens, the human is newly enlisted into L_{HT}, associating his/her location with his/her coordinates in the spectrum where located. If the human was not located in any spectrum, his/her coordinates, as well as the current confidence on the spectrum where he/she was previously detected, are examined to establish whether he/she has left the scene or is simply occluded. In this case, the potential new coordinates must be calculated according to his/her previous location in the scene. Finally, the number of humans enlisted into $L_{HT}(t)$ as well as their location are used as the output of the system.

4.2. Video Acquisition

General overviews of the video acquisition systems for infrared and color cameras are shown in Figure 4.2a and Figure 4.2b, respectively. The system operation starts with the acquisition of frames $I_{I0}(t)$ and $I_V(t)$ from an infrared camera and a color camera, respectively. A synchronization method is used to ensure that these frames belong to a same time instant t. Also a common coordinates

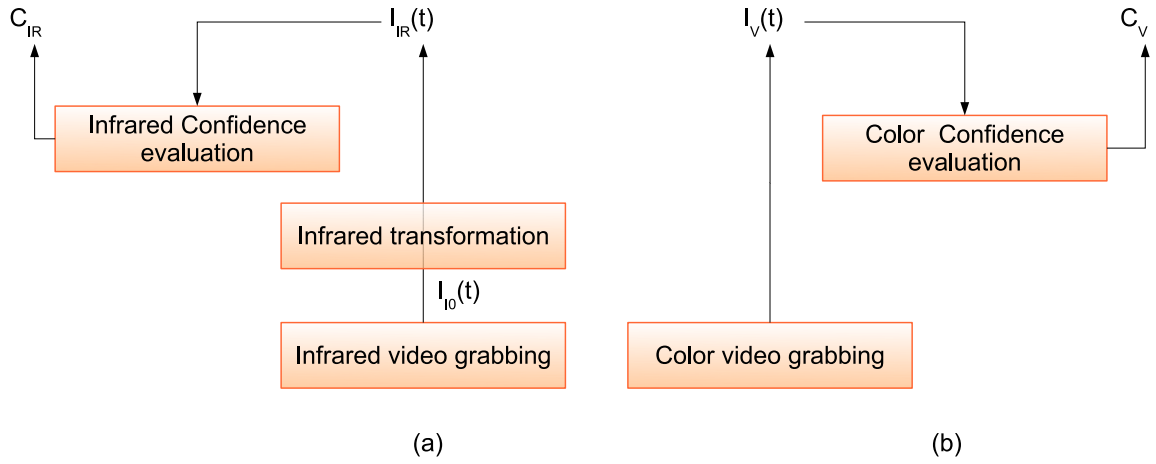


Figura 4.2: Overview of the infrared and color video acquisition system. (a) Infrared video acquisition system. (b) Color video acquisition system.

system is needed to allow later information fusion from both spectra. Thus, the infrared image $I_{I0}(t)$ coordinates will be transformed to coordinates on $I_V(t)$ by creating a transformed image $I_{IR}(t)$. A confidence degree for each spectrum will also be set using the properties of initial frames $I_V(t)$ and $I_{IR}(t)$. The confidence values (C_V for the visible spectrum and C_{IR} for the infrared spectrum) will allow the use of heuristic fusion techniques in later levels of the system.

4.2.1. Infrared and Color Video Grabbing

The process used to acquire and to synchronize frames from the infrared and color video cameras is shown in Figure 4.3. The first stage of the acquisition system starts with the capture of two frames in the infrared and visible spectra. It was explained in Chapter 3 that an *Axis Q4704* encoder is used to simultaneously grab the frames acquired by the two cameras at the same instant. The frames captured by the cameras connected to the encoder are separated into channels (with each channel set to a different video input). However, the encoder was found to have an internal buffer for the frames captured by each channel. This buffer only removes a frame when it is requested from its channel, sending the oldest frame available from that channel buffer. This feature originally led to synchronization failures since sometimes frames can be "missed" due to internal camera readjustments or small network traffic problems. When a frame is missed, it is grabbed as an empty frame from its origin channel while the other channel will have its valid frame in the buffer as usual. If that frame is not grabbed from its buffer, it will remain there causing synchronization problems. These failures are especially critical when working in an image fusion system where the frames from the different cameras are required to be as simultaneous as possible. The adopted solution is to ensure that valid frames have been captured from both cameras before sending the images to the system's upper levels. To do so, after the two frames have been captured from their channels, they are checked to be not empty. If one of them (infrared or color) is empty, the two frames are grabbed again from their buffer, whether they were empty images

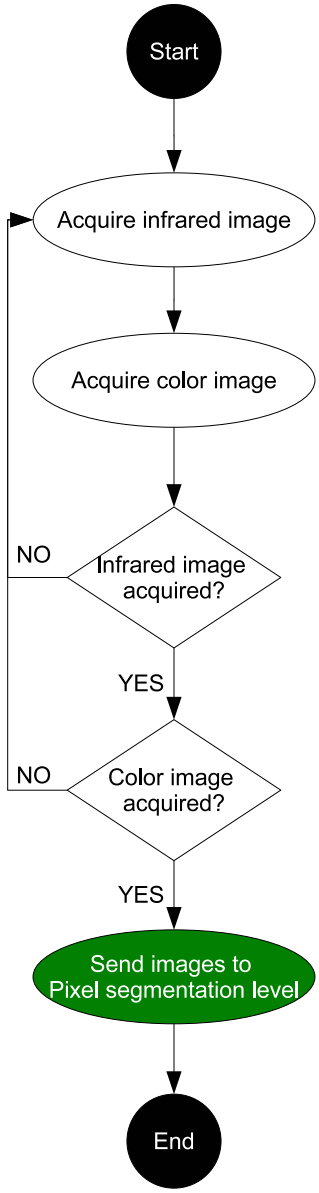


Figura 4.3: Video image acquisition in infrared and visible spectra.

or not. Thanks to this condition, we ensure that the latest images available from both buffers are sent towards the upper levels. After this first stage, image $I_{I0}(t)$ is grabbed from the infrared camera and image $I_V(t)$ is captured from the color camera.

4.2.2. Infrared Transformation

The next step allows the conversion between coordinates of $I_{I0}(t)$ and $I_V(t)$. Through this conversion a spatial alignment is performed and both spectra will operate in common coordinates, which is required to be able to compare detections in later stages by the segmentation algorithms used in

each spectrum. An homogenous equation system is used with this objective in mind.

First, a set of N pixel pairs $[((x_{i_s}, y_{i_s}), (x'_i, y'_i)]$ is set where (x_{i_s}, y_{i_s}) are the coordinates of a pixel in I_{I0} and (x'_i, y'_i) are the analogue coordinates of the same point in color image I_V, being $i \in 1, 2, ..., N$. The objective of the algorithm is, given equation (4.1) (which transforms coordinates $(x_{i_s}, y_{i_s}$ from $I_{I0})$ into coordinates (x_{i_f}, x_{i_f}) of I_V), to find the coefficients h_{jk} (with $\forall j, k \in [1, 3]$) that minimize the quadratic error E, as shown in equation (4.2). This is usual when dealing with space coordinates, since every point in the two-dimensional space can be treated as three-dimensional.

$$\begin{bmatrix} x_{i_s} \\ y_{is} \\ 1 \end{bmatrix} \times \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = \begin{bmatrix} x_{i_f} \\ y_{i_f} \\ 1 \end{bmatrix} \tag{4.1}$$

$$E = \sum_i (x'i - \frac{h_{11}x_{i_s} + h_{12}y_{i_s} + h_{1_3}}{h_{31}x_{i_s} + h_{32}y_{i_s} + h_{3_3}})^2 + (y'i - \frac{h_{21}x_{i_s} + h_{22}y_{i_s} + h_{2_3}}{h_{31}x_{i_s} + h_{32}y_{i_s} + h_{3_3}})^2 \tag{4.2}$$

When applying equation (4.1), the coordinates of every point in I_{I0} are transformed into its equivalent coordinates in I_V. The transformed image I_{IR} is generated and the system now works with a coordinate system which is independent of the origin spectrum of every human candidate being detected in later levels.

4.2.3. Infrared and Color Confidence Degree Review

Finally a confidence degree is assigned to each spectrum. These confidence degrees will be used to choose which segmented blobs will be chosen and which blobs will be discarded in the later fusion process. The confidence degree of each spectrum is based on the features of the images captured by the corresponding acquisition camera and is updated in each frame acquired. So, if the scene changes its temperature, illumination, etc., the fusion algorithm will be able to adapt itself to the new conditions. A scheme of how the confidence levels are divided and assigned is shown in Figure 4.4.

4.2.3.1. Color Confidence Assignation

The average or mean gray level value of color image $I_V(t)$ is the base for establishing confidence in the color spectrum. In order to improve the invariability from the camera settings, the color image $I_V(t)$ is transformed into a gray level image I_G. A low average gray level value of I_G means that the image is captured under poor lighting conditions, making any object (including humans) hard to distinguish from the rest of the scene, as shown in Figure 4.5a. The infrared equivalent of this frame is shown in Figure 4.6b. On the other hand, a very high gray level mean denotes that it is snowing or the environment is under fog conditions (as depicted in Figure 4.5c). Even, there can be a lighting source blinding the color camera as it is directly pointing to it. So, the visible spectrum segmentation will again be unable to distinguish anything on the scene. Thus, an intermediate gray level mean

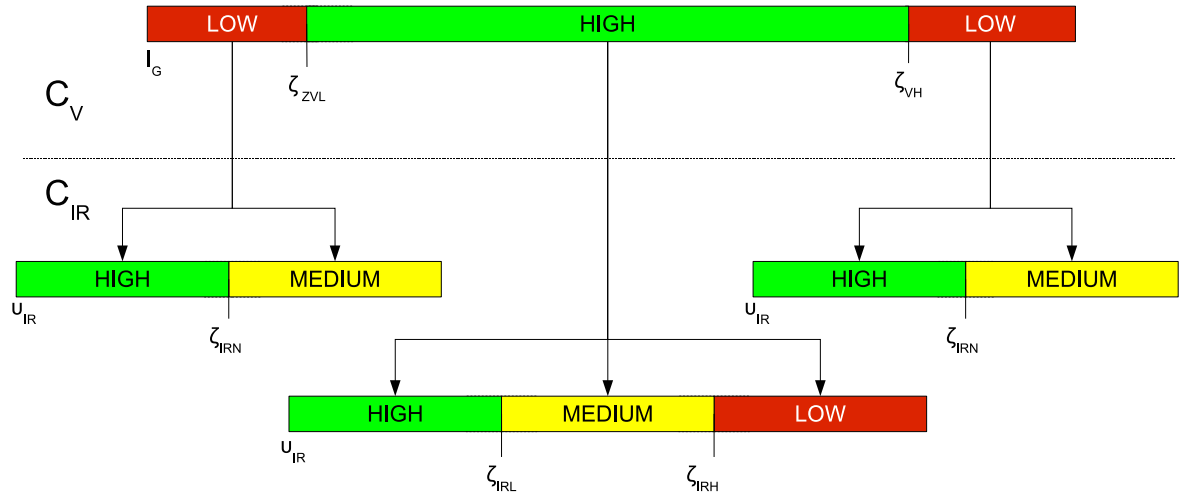


Figura 4.4: Establishment of the confidence levels

Figura 4.5: Different confidence values for the visible spectrum. (a) Color image with low confidence level in night conditions. (b) Color image with high confidence level. (c) Color image with low confidence level in fog conditions.

indicates that the lighting conditions in the scene are adequate and humans are easily distinguished. An example of this situation is depicted in Figure 4.5b. Equation (4.3) shows how the reliability of the visible spectrum is established (denoted as C_V), with thresholds ζ_{VL} and ζ_{VH} fixed experimentally, since the gray level values of the elements in the scene determine the conditions where the visual spectrum is trustful and also where it is not reliable.

$$C_V = \begin{cases} \text{HIGH}, & \text{if } \zeta_{VL} < \overline{I_{VG}} < \zeta_{VH} \\ \\ \text{LOW}, & \text{otherwise} \end{cases} \tag{4.3}$$

4.2.3.2. Infrared Confidence Assignation

The illumination mean and the standard deviation of the image provide very useful information about its contrast in the infrared spectrum. Since our infrared algorithms are based on distinguishing the humans because they usually appear warmer than the background, this information will have the key to establish their reliability on environmental conditions where this is satisfied, i.e. sequences where the human temperature is higher than the environment where he/she are placed. In other cases, new approaches must be examined, but this process is beyond the scope of this PhD thesis.

Let us define the contrast υ_{IR} as the coefficient between the average gray level value $\overline{I_{IR}}$ of infrared image I_{IR} and its standard deviation $\sigma_{I_{IR}}$, just as shown in equation (4.4). An image with a high gray level mean and a low standard deviation denotes that a great amount of pixels have similar values, making humans hard to distinguish from the background. An example of an initial frame with these features is shown in Figure 4.6a (with the equivalent color frame seen in Figure 4.5b). On the other hand, an image with a great standard deviation and a low mean value has a small number of pixels with high gray level values and the rest of them with low values. The high value pixels usually correspond to humans. An example of a frame with this situation is shown in Figure 4.6c. The intermediate case where humans are distinguished from the background not as clearly as in the previous situation can be appreciated in Figure 4.6b. Thus, the equation for establishing the reliability of the frames in the infrared spectrum (which will be denoted as C_{IR}) is shown in equation (4.5), where thresholds ζ_{IRH} and ζ_{IRL} are experimentally established since they are dependent of the particular heat distribution of the test scenario.

$$\upsilon IR = \frac{\overline{I_{IR}}}{\sigma_{I_{IR}}} \tag{4.4}$$

$$C_{IR} = \begin{cases} \text{HIGH}, & \text{if } (C_V = HIGH \text{ AND } \upsilon_{IR} < \zeta_{IRL}) \\ \text{MEDIUM}, & \text{if } (C_V = HIGH \text{ AND } \zeta_{IRL} < \upsilon_{IR} < \zeta_{IRH}) \\ \text{LOW}, & \text{if } (C_V = HIGH \text{ AND } \upsilon_{IR} > \zeta_{IRH}) \end{cases} \tag{4.5}$$

Different conditions for the infrared confidence are also imposed for high and low visibility sequences. Since segmentation in the visible spectrum is almost unable to distinguish humans when the confidence is set to LOW (except when they are in a zone directly illumined by a lamppost), the infrared spectrum will always be forced with a confidence level above LOW. This is true because in those conditions it will always work better than the color segmentation, since the temperature is lower than at daytime, and humans can partially be distinguished in those conditions. However, the infrared confidence level will not only be restricted to $HIGH$, due to the importance of the contrast υ_{IR} already mentioned. An example of this situation is seen in Figure 4.5a for the color camera and Figure 4.6b for the infrared camera, where it can be appreciated that the human is still hard to distinguish, but

Figura 4.6: Different confidence values for the infrared spectrum. (a) Infrared image with low confidence level. (b) Infrared image with medium confidence level. (c) Infrared image with high confidence level.

easier than at the color camera. Thus, the infrared confidence at night will be restricted to $MEDIUM$ and $HIGH$, with a new threshold ζ_{IRN} used to separate both values as shown in equation (4.6).

$$C_{IR} = \begin{cases} \text{HIGH}, & \text{if } (C_V = LOW \text{ AND } v_{IR} < \zeta_{IRN})) \\ \text{MEDIUM}, & \text{if } (C_V = LOW \text{ AND } v_{IR} > \zeta_{IRN}) \end{cases} \tag{4.6}$$

4.3. People Segmentation

The main features of each spectrum, exploiting their properties, are used to develop the most robust human segmentation algorithms. Thus, the thermal difference between the humans and their environment is a cue used in the infrared spectrum and the information provided by the color in the scene is exploited in the visible spectrum. Three different approaches are implemented for the infrared spectrum to study the influence of motion in the scene for human detection, while two algorithms are developed in the visible spectrum to analyze whether motion history or background information is more relevant combined with color information in the task of human segmentation.

4.3.1. People Segmentation in Infrared

It has been previously explained in Chapters 1 and 3 of this dissertation that the infrared spectrum has many interesting features which can be exploited for robust human detection. Two of these properties are clearly important: (1) the independence of lighting conditions of the scene, and specially, (2) the fact that humans tend to be clearly highlighted respect to the background of the picture. The human heads also appear hotter than the rest of the body covered with clothing. This is why an initial human detection system is developed using these properties based on a single frame (the current infra-

red frame $I_{IR}(t)$). After the initial tests, new features are subsequently added to this initial approach to exploit motion information in the scene between the current frame $I_{IR}(t)$ and the previous frame $I_{IR}(t - \Delta t)$ where frame subtraction is used to extract the moving elements in the scene which could have not been located based on the information of a single frame and adding the humans detected to those detected on the human detection based on a single frame. Finally, an alternate proposal is tested using optical flow to people previously detected on a single frame.

A visual representation of these approaches is provided in Figure 4.7. Notice that human detection based on frame subtraction directly adds its information to the results of the human detection based on a single frame, while human detection based on optical flow directly affects the outcome of previous stages of the approach based on a single frame. This is why optical flow is considered an independent algorithm. The results of the three approaches are compared in Chapter 5 to assess which algorithm fits better to the test scenario where the overall system is evaluated.

Once the experimental environment is established, all these approaches are compared to establish the algorithm which suits better for the particular needs of the selected test scenario. Each one of these approaches has an initial stage where a list of human candidates (possible humans) is extracted. The candidates (enlisted through their representing blobs) are analyzed and refined in later stages in order to separate possible groups of humans or to remove false positives (non-humans) which could not be filtered in the previous stages. The results of each algorithm are finally obtained in a new list of blobs. Both lists are later compared to obtain a final list L_{IR} of humans in the infrared spectrum.

4.3.1.1. Human Detection in Infrared Based on a Single Frame

A human detection system based on a single frame has initially been developed in the infrared spectrum using the properties already mentioned (Fernández-Caballero et al., 2011a). A complete scheme describing this algorithm is offered in Figure 4.8. First, a set of human candidates are extracted from the scene, based on their thermal properties. A series of adjustments are performed on the initial candidates to obtain a better characterization of their size and location. A series of restrictions on size and shape are finally applied over the adjusted candidates to eliminate potential false positives that may have appeared in the algorithm. Each one of the stages are now explained in more detail.

Detection of Human Candidate Blobs

The algorithm starts with the analysis of the input image, $I_{IR}(t)$, captured at time t by the acquisition system, as shown in Figure 4.9a. A threshold θ_c is used to perform a binarization for the aim of isolating the human candidate spots. This threshold obtains the image areas containing moderate heat blobs, thus belonging to human candidates. So, warmer zones of the image where humans could be present are isolated. The threshold is calculated as shown in equation (4.7), where $\sigma_{I_{IR}}$ is the standard deviation of image I_{IR}. Also, a base threshold γ which must be experimentally fixed based on the features of each scenario is used. This base threshold is added to the product of the standard deviation $\sigma_{I_{IR}}$ and an augmentation percentage factor ϕ experimentally fixed according to the features of the scenario.

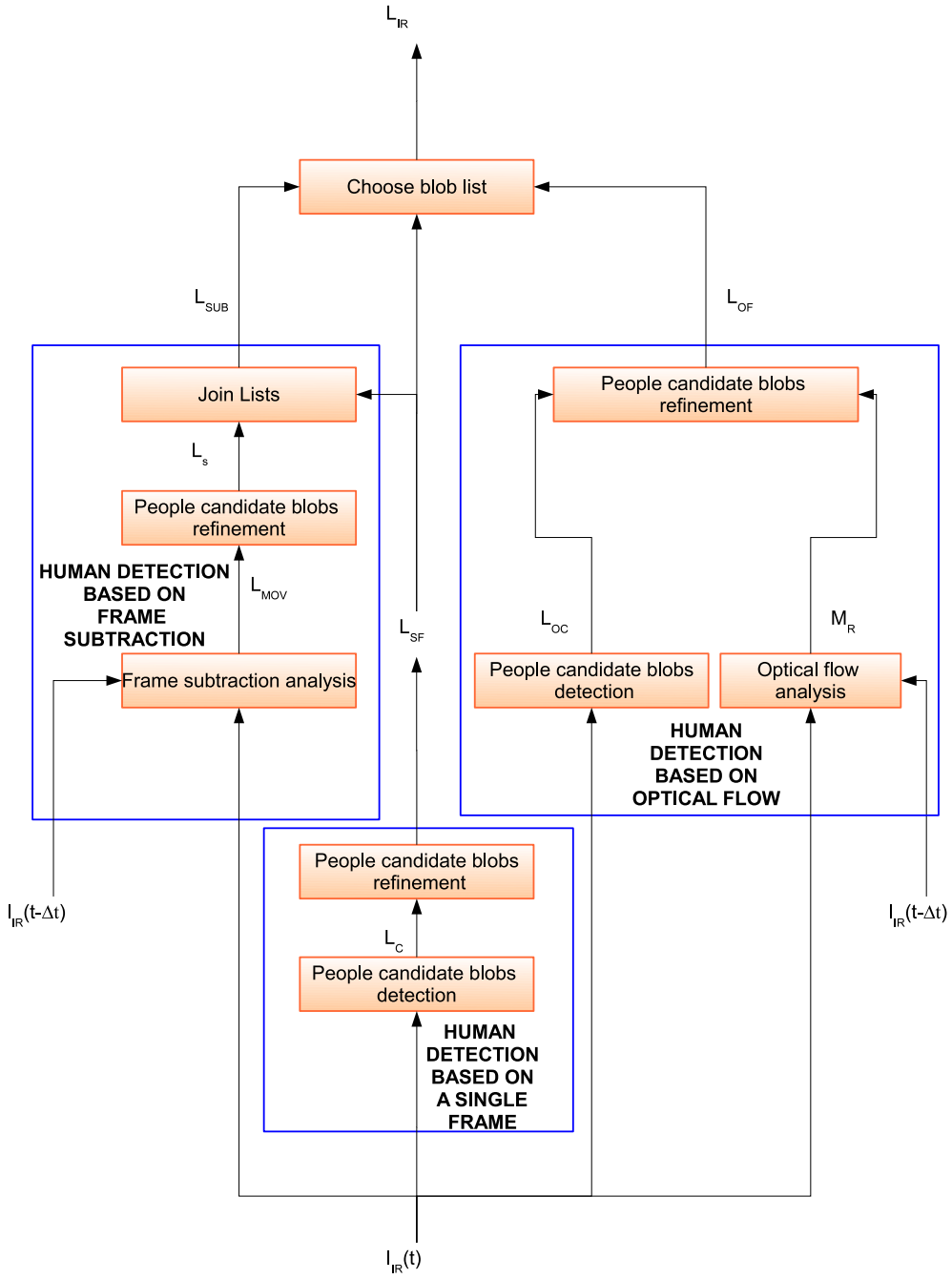


Figura 4.7: Overview of the infrared segmentation system.

Now, image I_{IR} is binarized using the obtained threshold θ_c. Pixels above the threshold are set as maximum value $max = 255$ (the maximum pixel intensity for a gray level 8 bits image) and pixels below are set as minimum value $min = 0$. This process is shown in equation (4.8) while the results

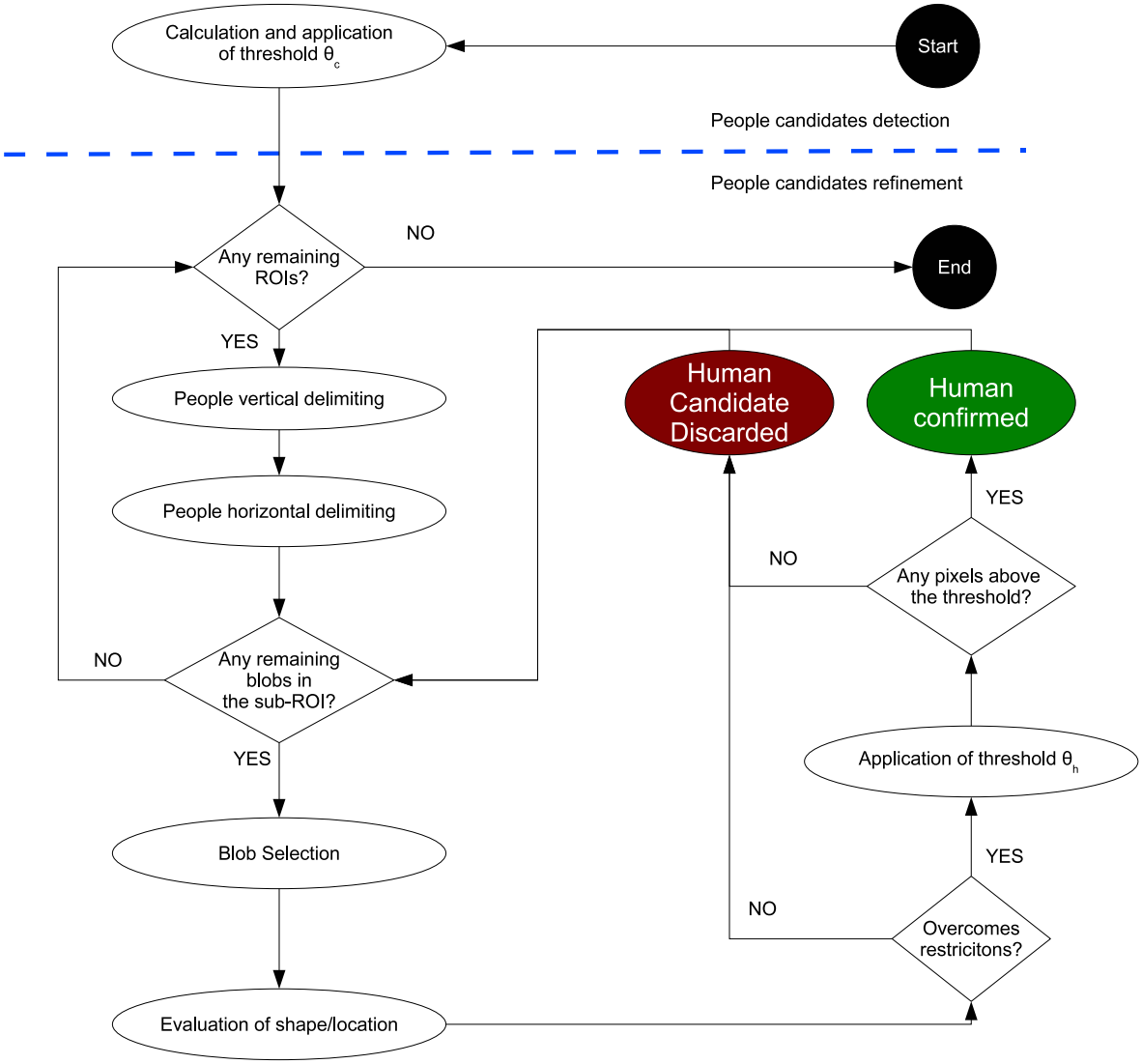


Figura 4.8: Algorithm of human detection based on a single infrared frame.

can be seen in Figure 4.9b.

$$\theta_c = \gamma + \phi \times \sigma_{I_{IR}} \tag{4.7}$$

$$I_b(x,y) = \begin{cases} \text{min, if } I_{IR}(x,y) \leq \theta_c \\ \text{max, otherwise} \end{cases} \tag{4.8}$$

Next, the algorithm performs morphological opening and closing operations to eliminate isolated pixels and to unite areas split during the binarization, obtaining image I_c. These operations are shown in equations 4.9 and 4.10, respectively, and require structuring elements which, in both cases, are 3×3

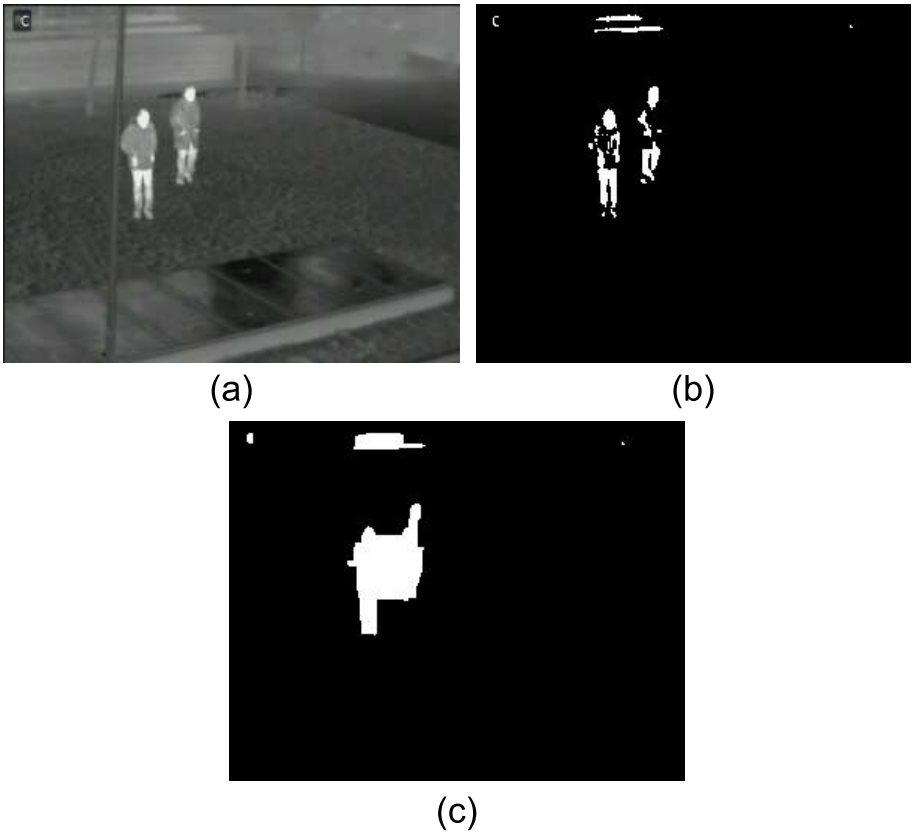


Figura 4.9: Detection of human candidate blobs in the infrared spectrum. (a) Infrared input frame. (b) Thresholded frame. (c) Frame after morphological operations.

square matrices centered at position $(1, 1)$. These operations join small artifacts which could be part of the shapes, as shown in Figure 4.9c.

$$I_o = I_b \circ \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix} \tag{4.9}$$

$$I_c = I_o \bullet \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix} \tag{4.10}$$

Once the binarized image has been obtained, the blobs contained in the image are extracted. A minimum area, A_{min}, – function of the image size – is established for a blob to be considered to contain humans. A_{I_c} is the area of image I_c, as shown in equation (4.11). This value is experimentally fixed as $\frac{1}{400^{th}}$ of the area of I_c. As a result, the list of blobs, L_C, containing people candidates in form of blobs $b_\Lambda[(x_{start}, y_{start}), (x_{end}, y_{end})]$, is generated. Λ stands for the number (index) of people candidate blob in image $I_c(x, y)$ and (x_{start}, y_{start}) and (x_{end}, y_{end}) are the upper left and lower right

coordinates, respectively, of the minimum rectangle containing the blob. As an example, consider the resulting list of blobs related to Figure 4.9 and offered in Table 4.1, where only a single blob is initially detected.

$$A_{min} = \frac{A_{I_c}}{400} \tag{4.11}$$

Tabla 4.1: Example of people candidate blobs list.

κ	x_{start}	y_{start}	x_{end}	y_{end}	area
1	84	50	482	142	2778

Refinement of Human Candidate Blobs

In this part, the algorithm works with the list of blobs L_C present in image I_c. This list was obtained at the end of the previous section. At this point, there is a need to validate the content of each blob to find out if it contains one single human candidate or more than one. Therefore, each detected blob is individually processed.

Let us define a region of interest (ROI) as the minimum rectangle containing one blob of list L_B (obtained from $I(x, y)$). A ROI may be defined as $R_\kappa = R_\kappa(i, j)$, when associated to blob $b_\kappa[(x_{start}, y_{start}), (x_{end}, y_{end})]$. Notice that $i \in [1..max_i = x_{end} - x_{start} + 1]$ and $j \in [1..max_j = y_{end} - y_{start} + 1]$.

Vertical Delimiting of Humans

The next step consists in scanning R_κ by columns, adding the gray level value corresponding to each pixel in that column, as shown in equation (4.12). This way, a histogram $H_\kappa[i]$, showing which zones of the current ROI own greater heat concentrations, is obtained. A double purpose is pursued when computing the histogram. In first place, we want to increase the certainty of the presence and situation of human heads. Secondly, as a ROI may contain several persons that are close enough to each other, the histogram helps separating human groups (if any) into single humans. This method, when looking for maxima and minima within the histogram allows differentiating among the people present in the particular ROI.

$$H_\kappa[i] = \sum_{j=1}^{max_j} R_\kappa(i, j) \forall i \in [1, 2, ..., max_i] \tag{4.12}$$

Now the histogram, $H_\kappa[i]$, is scanned to separate grouped humans, if any. Local maxima and local minima are searched in the histogram to establish the different heat sources (see Figure 4.10a) with this purpose. To assess whether a histogram column contains a local maximum or minimum, a new threshold $\theta_{v_{min}}$ is fixed. Experimentally we went to the conclusion that the local maximum threshold

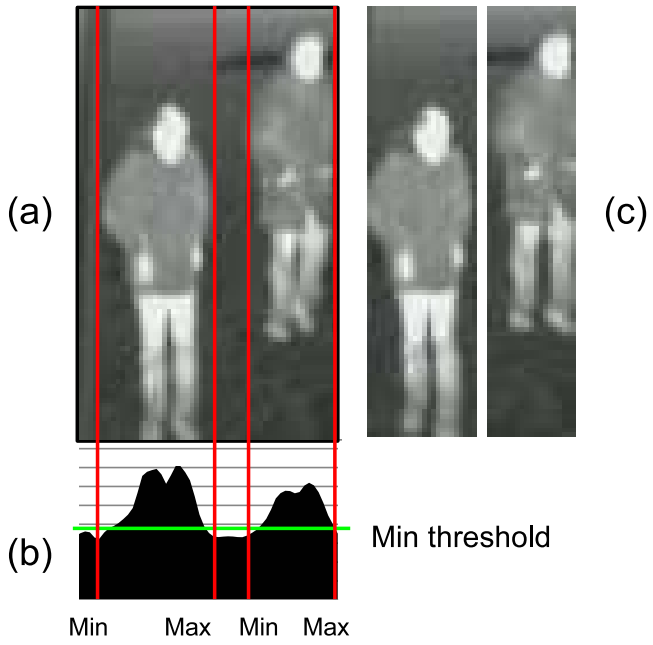


Figura 4.10: Vertical delimiting of humans in the infrared spectrum. (a) Input ROI containing a group of humans. (b) Histogram. (c) Column adjustment to obtain two human candidates.

must be set as shown in equation (4.13).

$$\theta_{v_{min}} = 0{,}6 \times \overline{R_\kappa} \times max_j \tag{4.13}$$

where $\theta_{v_{min}}$ indicates those regions of the ROI where the sum of the heat sources are really low. We are looking for columns where the $60\,\%$ of their pixels are below the mean gray value of R_κ, since those regions are supposed to belong to gaps between two humans. Figure 4.10b shows the histogram for input ROI of Figure 4.10a. Figure 4.10c shows the two humans as separated by the algorithm into sub-ROIs, $sR_{\kappa,\alpha}$.

Horizontal Delimiting of Humans

All humans contained in a sub-ROI, $sR_{\kappa,\alpha}$, and obtained in the previous section still possess the same height, namely the height of the original ROI. Now, we want to fit the height of each sub-ROI to the real height of the human contained. For this purpose row adjustment is performed for each new sub-ROI, $sR_{\kappa,\alpha}$, generated after the previous column adjustment, by applying a new threshold, θ_h. The calculation is done separately on each sub-ROI to avoid the influence of the rest of image pixels on the result. This threshold uses the value of the sub-ROI mean gray level, $\theta_h = \overline{sR_{\kappa,\alpha}}$. Thus, sub-ROI $sR_{\kappa,\alpha}$ is binarized in order to delimit its upper and lower limits, obtaining $sR_{\beta,\kappa,\alpha}$, as shown in equation (4.14) similarly to equation (4.8). The result of the threshold application over the input

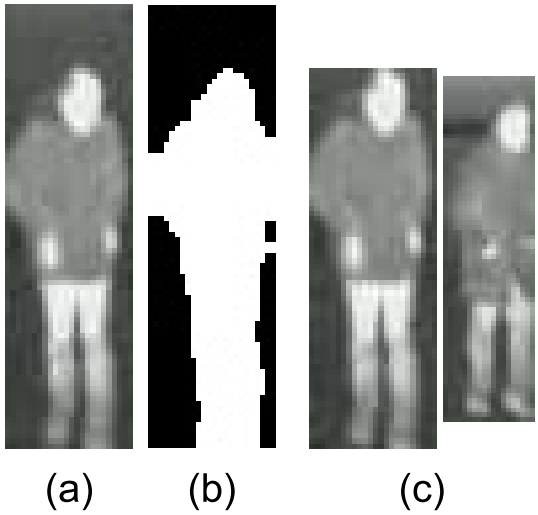


Figura 4.11: Horizontal delimiting of humans in the infrared spectrum. (a) Input sub-ROI. (b) Thresholded sub-ROI. (c) Row adjustment to obtain two human candidates.

sub-ROI shown in Figure 4.11a can be seen in Figure 4.11b.

$$sR_{\beta,\kappa,\alpha}(x,y) = \begin{cases} \text{min, if } sR_{\kappa,\alpha} \leq \theta_r \\ \text{max, otherwise} \end{cases} \tag{4.14}$$

After this, a closing operation (shown in equation (4.15)) is performed to unite spots isolated in the binarization, getting $sR_{\varsigma,\Lambda,\alpha}$ (see Figure 4.11b). Next, $sR_{\varsigma,\Lambda,\alpha}$ is scanned, searching pixels with values superior to min. The upper and lower rows of the human are equal to the first and last rows, respectively, containing pixels with a value set to max. The final result may be observed in Figure 4.11c.

$$sR_{\varsigma,\kappa,\alpha} = sR_{\beta,\kappa,\alpha} \bullet \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix} \tag{4.15}$$

Confirmation of Humans

Now a final stage is needed for each sub-ROI, $sR_{\varsigma,\kappa,\alpha}$, to confirm if the human candidate contained in it is actually a human. A flow diagram for the human confirmation process is offered in Figure 4.12.

At this point, it is interesting to remind that every sub-ROI is defined by its coordinates (x_{start}, y_{start}) and (x_{end}, y_{end}). In first place, let us define the basic parameters needed for human confirmation for every sub-ROI, $sR_{\varsigma,\kappa,\alpha}$, in equations (4.16), (4.17), (4.18) and (4.19).

$$h_{sR} = y_{end} - y_{start} \tag{4.16}$$

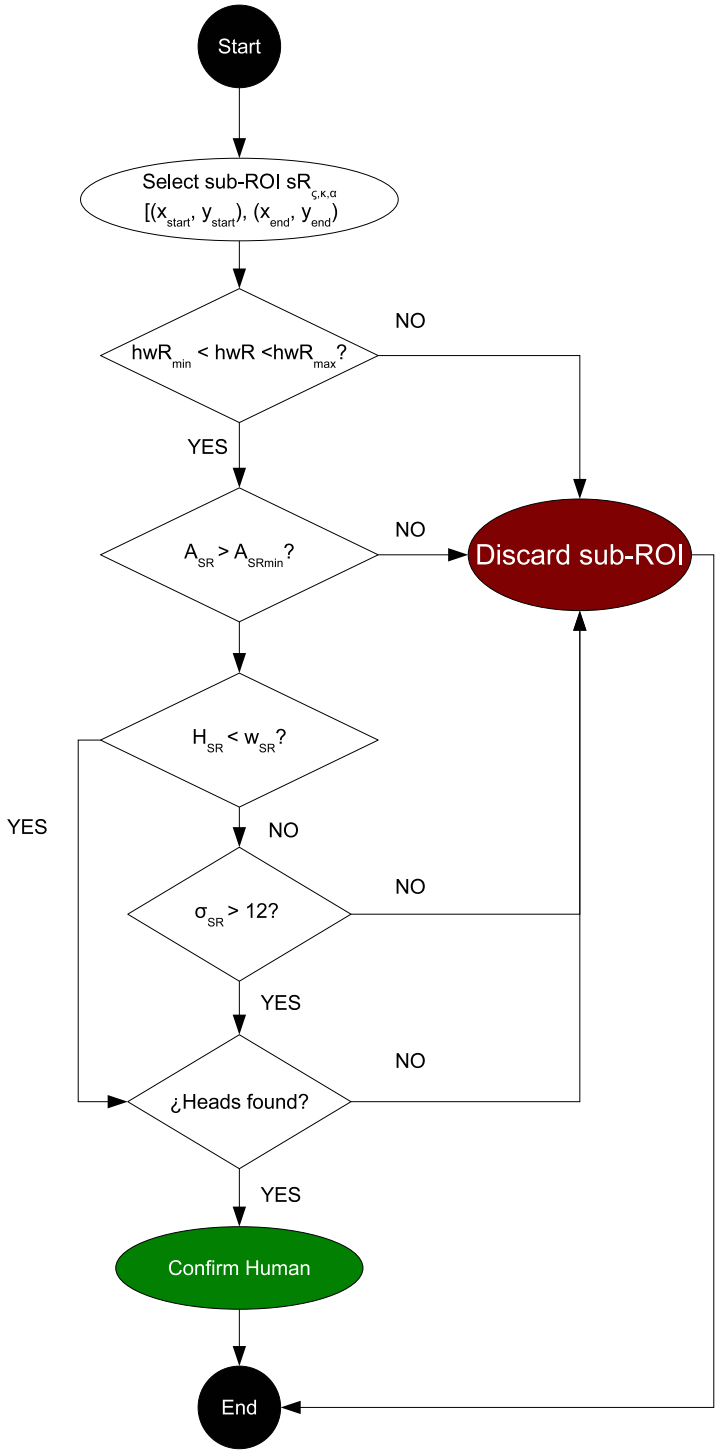


Figura 4.12: Human confirmation algorithm.

$$w_{sR} = x_{end} - x_{start} \tag{4.17}$$

$$A_{sR} = h_{sR} \times w_{sR} \tag{4.18}$$

$$hwR = \frac{h_{sR}}{w_{sR}} \tag{4.19}$$

Some of the incandescent spots in the image (such as light bulbs or big heat sources in general) can still be confused under certain circumstances with humans due to their heat properties, so another important step consists in verifying if one of these spots is being scanned instead of a human. First, the human candidate's shape is checked. The first check consists in testing its height/width ratio (see equation (4.19)). The restrictions applied are not specifically fixed since they will be relaxed when the scene has low contrast, as it will be examined in Chapter 5.

The human candidate's area A_sR is also required to be above a minimum area A_{sRmin} experimentally fixed according to features such as the camera height or the extension of the scenario. Area A_sR is calculated as shown in equation (4.18) where w_{sR} and h_{sR} are the width and height of the sub-ROI and are obtained as shown in equations (4.16) and (4.17) respectively .

Next, if the human candidate's width w_{sR} is longer than its height h_{sR}, its standard deviation (σ_{sR}) is checked. This is due to the fact that incandescent spots such as lamps or fuses have a low standard deviation since their heat distribution is uniform, while humans, as it has been previously said, have different heat concentrations in their body parts, such as the head being warmer than the rest of the body. We have determined experimentally that σ_{sR} must be greater than 12 to be a human candidate.

The final check scans if the human candidate has zones warmer than the hard threshold θ_h calculated in equation (4.20), similarly to equation (4.7), although the standard deviation of image I^n is replaced by the sub-ROI $sR_{\varsigma,\kappa,\alpha}$ standard deviation in order to use only the features of the human candidate and Γ has been experimentally fixed to approximately a $60\,\%$ of the maximum value of a 256 gray level image, i.e., 150. . The final zones obtained indicate the presence of human heads.

$$\theta_h = \Gamma + \sigma_{sR_{\varsigma,\kappa,\alpha}} \tag{4.20}$$

Finally, the blobs associated to the split ROIs that have satisfied these criteria are enlisted into the final list of humans, L_{SF}.

4.3.1.2. Human Detection in Infrared Based on Frame Subtraction

We have previously explained that certain environmental conditions affect negatively the visual contrast in the infrared spectrum. For example, humans are very hard to find in warm environments where the scene temperature is similar to the people temperature. An example of this situation can be appreciated in Figure 4.14, where the human has been manually highlighted because it cannot be

Tabla 4.2: Example of detected people as the final result of the human detection in the infrared spectrum.

κ	x_{start}	y_{start}	x_{end}	y_{end}	area
1	339	286	395	354	3808
2	396	270	481	458	15980
3	298	289	338	354	2600

easily seen in this single frame. Yet, if we use the motion information in the scene, we can find the humans in it since they do not tend to be static during long periods of time. Therefore, an extension for the human detection based on a single frame is developed using the motion information in the scene. The scheme of the algorithm with this new extension is shown in Figure 4.13. The new stages are highlighted with red lines and bold text.

While the list L_{SF} of humans obtained from detection based on a single frame is used, information from two new stages is added to the previous list. A new phase, called frame subtraction analysis, is introduced in this extension in order to take advantage of the motion information in the scene. The results L_{MOV} from this new stage are later refined into a new list L_S which will be joined with the list L_{SF} in order to reduce the number of false negatives in the scene.

Frame Subtraction Analysis

In this new phase, the previous image $I_{IR}(t-\Delta t)$, and the current one, $I_{IR}(t)$ are used. An image subtraction and thresholding is performed on these frames as shown in equation (4.21), where θ_{sub} is experimentally fixed as a $16\,\%$ of the maximum value of a 256 gray levels image. This binarized image is combined with I_c by an "AND" operation, obtaining binary image I_{sc}. This way, false positives due to small illumination changes are discarded, by ensuring that the zones with motion have also warm heat concentrations similar to humans. Now, ROIs with area superior to A_{min} (calculated as shown in equation (4.11)) and with a percentage of pixels set to MAX greater than a rate threshold ψ (experimentally fixed at a $5\,\%$ of the area of the ROI) are extracted from I_{sc} in the list of blobs L_{MOV}.

$$I_s(x,y) = \begin{cases} \text{max, if } |I_{IR}(x,y,t) - I_{IR}(x,y,t-\Delta t)| > \theta_{sub} \\ \text{min, otherwise} \end{cases} \tag{4.21}$$

Refinement of Human Candidate Blobs

ROIs obtained from the blobs in L_{MOV} are vertically and horizontally delimited in the same way as the ROIs are refined in the human detection based on a single frame. The humans are also confirmed exactly the same way. The main difference is that human candidates are here enlisted in a list L_s. This list is finally checked along with L_c (obtained from human detection based on a single frame) to remove redundancies encountered in both lists. This way, humans that can only be found through motion information are added to the initial algorithm. These humans are enlisted into the final

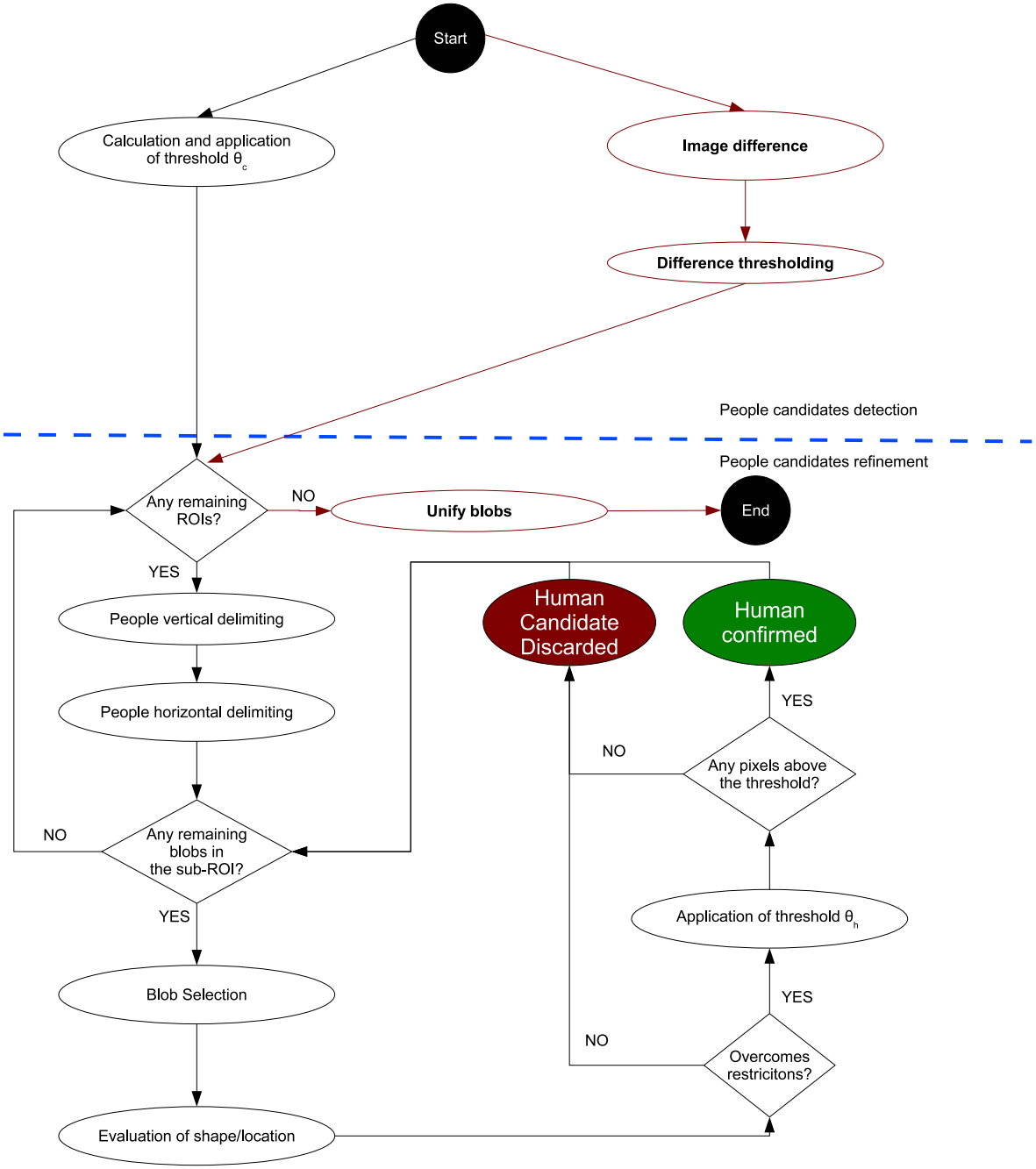


Figura 4.13: Algorithm of human detection in the infrared spectrum based on frame subtraction.

list L_{SUB}.

4.3.1.3. Human Detection in Infrared Based on Optical Flow Calculation

An extension of the algorithm for human detection based on a single frame is developed with the objective of using the motion information that can be extracted from a scene acquired from a moving

Figura 4.14: Example of a human hard to be detected in the infrared spectrum.

camera. This information can be especially useful when capturing images from a surveillance robot or a moving vehicle that must detect pedestrians with the aim of warning the driver about them. Image subtraction is not used in this new approach, since there are differences between every pixel in the scene due to the camera motion. Optical flow has been selected as it discards the scene movement due to the proper vehicle motion. A simple subtraction-based approach would indicate that everything is in movement, making impossible to really differentiate moving objets in the completely moving scene. Thus, as the majority motion is the scene movement, optical flow discards it to only focus in other different direction movements (Lucas and Kanade, 1981). Although this algorithm was originally designed for a not static location, it was tested to assess another way of the influence of motion in the scene. A scheme depicting this algorithm is shown in Figure 4.15. Again, the new stages are highlighted with red lines and bold text.

Detection of Human Candidate Blobs

This initial phase is performed exactly in the same way described in the algorithm for human detection based on a single frame with the obtained blows enlisted into a list L_{oc}. Indeed, notice that those ROIs without motion detected in the next stage run through the same stages as in the original algorithm.

Optical Flow Analysis

This phase uses two image frames, the previous image, $I_{IR}(t - \Delta t)$, and the current one, $I_{IR}(t)$ (see Figure 4.16a and Figure 4.16b). In first place, the current and the previous frames are multiplied to enhance the contrast, such that the dark values become darker and the bright values become brighter (see Figure 4.16c and Figure 4.16d). This way, the calculation of the optical flow is facilitated.

The dynamic analysis requires the calculation of the moments corresponding to each pixel (x, y) movement on the input images $I_{IR}(x, y, t - \Delta t)$ and $I_{IR}(x, y, t)$. The optical flow calculation results into two gray level images, where each pixel reflects the angular moment detected, storing the movements in X and Y axes. Firstly, the algorithm performs the speed calculation of the optical flow.

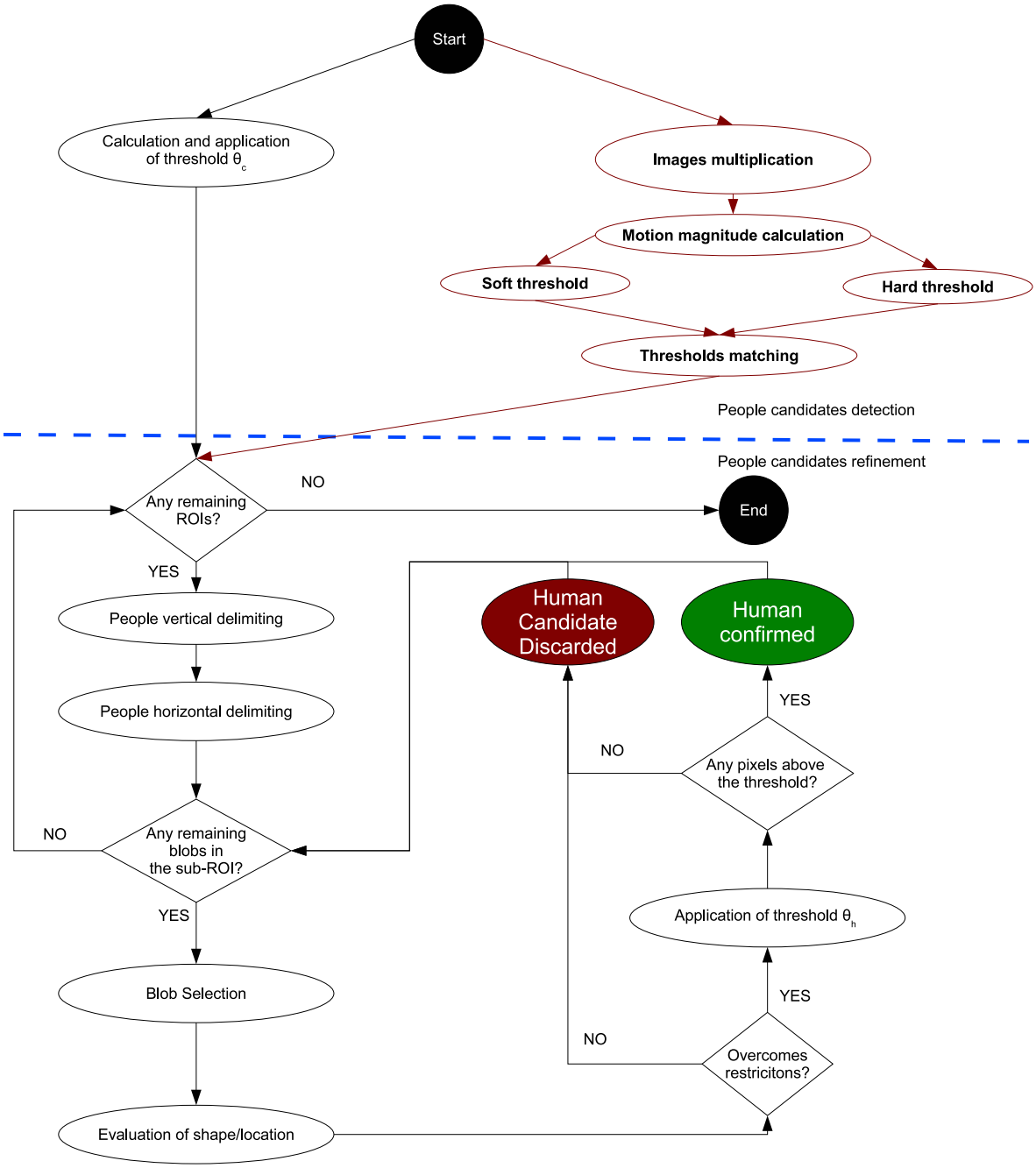


Figura 4.15: Algorithm of human detection in the infrared spectrum based on optical flow calculation.

The selected optical flow approach is Lucas-Kanade without pyramids algorithm. This algorithm is fast and offers an excellent success vs. speed ratio. The calculated speeds, as a result of the optical flow, are turned into angles, $\alpha(x, y, t)$, and magnitudes, $m(x, y, t)$. Figure 4.17a shows the magnitudes (moments), that is to say, the amount of movement at each pixel (x, y) between $I(x, y, t - \Delta t)$ and $I(x, y, t)$, in form of a moments image, $M(t)$. Similarly, Figure 4.17b shows the direction of the movement (angles). The results clearly indicate that angles are less important than moments. Indeed,

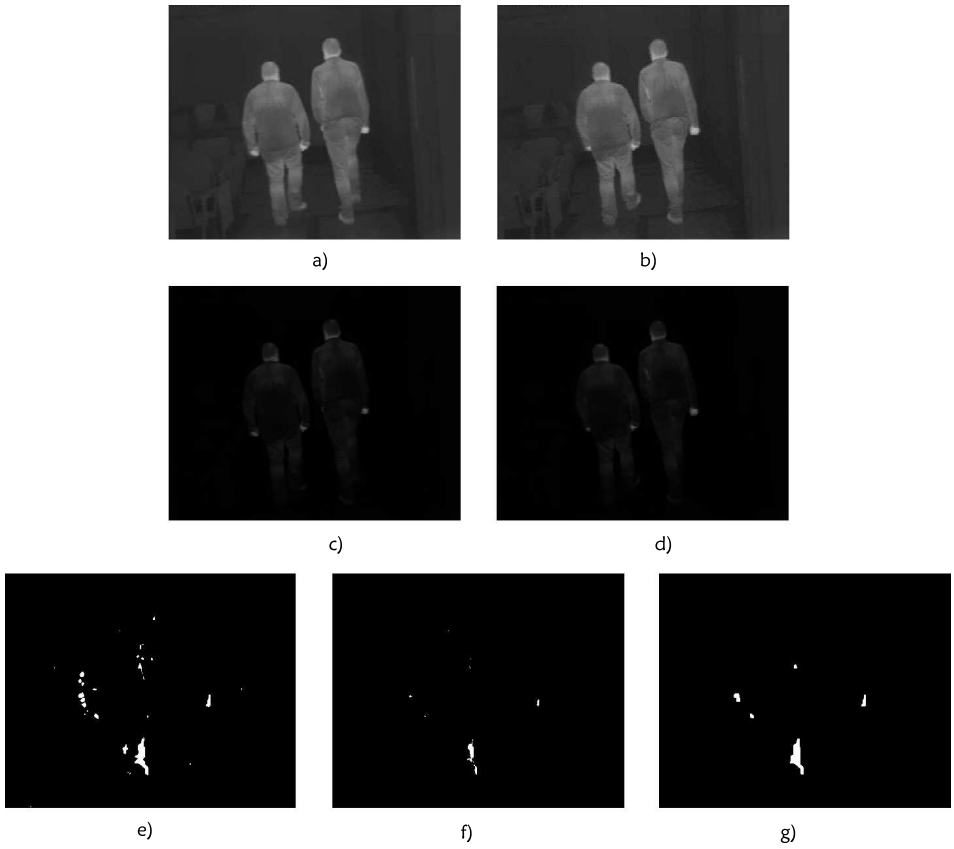


Figura 4.16: Images obtained in the human detection based on optical flow. (a) Previous frame. (b) Current frame. (c) Multiplied previous frame. (d) Multiplied current frame. (e) Soft thresholded moments. (f) Hard thresholded moments. (g) Matched thresholds.

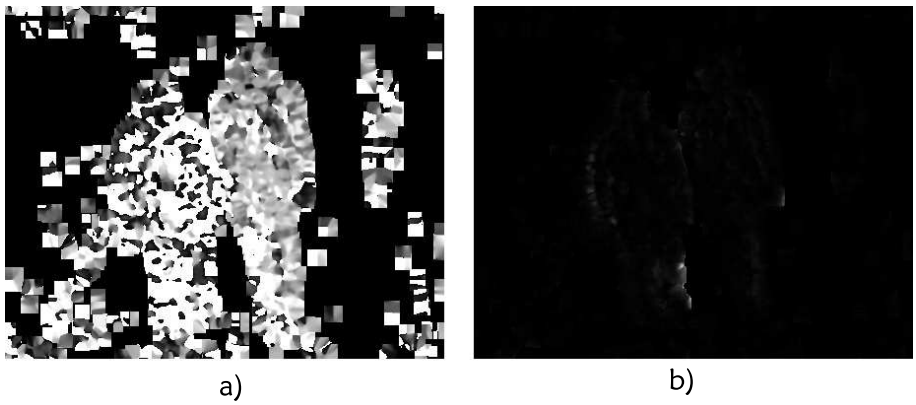


Figura 4.17: Optical flow calculation. (a) Moments. (b) Angles.

on the one hand, non-rigid objects'movements go into very different directions, and, on the other side, angles with low moments may be caused by image noise.

To efficiently use the moments image $M(t)$, its histogram, as shown in Figure 4.18 has been

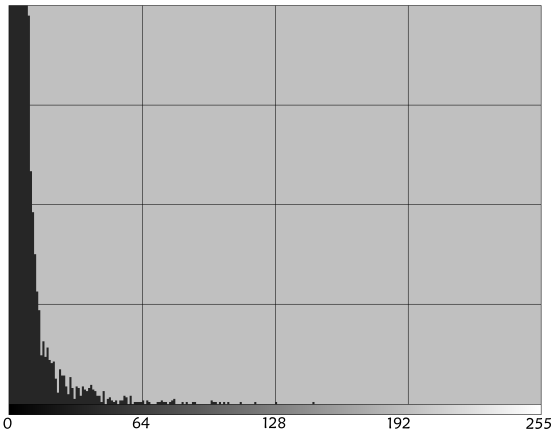


Figura 4.18: Histogram of the optical flow moments.

studied for many cases. As you may observe, most values are in the $[0, 64]$ interval, but very close to 0. Indeed, the average value is close to 1 in these moments images. Therefore, two thresholds, a moments soft threshold $\mu_s = 10$ and a moments hard threshold $\mu_h = 25$, are used to delimit the blobs of possible (candidate) humans. The aim of the soft threshold, μ_s, is to obtain the most representative values, whereas the hard threshold, μ_h, is used to refine a better matching between zones that show an elevated movement and zones with less movement but connected to the previous ones. Thus, the zones where movement has been detected are extended, and the zones with reduced movements are eliminated.

Therefore, firstly, the moments soft threshold μ_s is applied to the moments image $M(t)$ to obtain image $M_s(t)$, as shown in Figure 4.16e. The related formula is shown in equation (4.22) for each pixel (x, y) of the image.

$$M_s(x, y, t) = \begin{cases} min, & \text{if } M(x, y, t) \leq \mu_s \\ max, & \text{otherwise} \end{cases} \tag{4.22}$$

Afterwards, an opening filter is applied to erase isolated pixels, getting M_o (see equation (4.23)). In this case, disconnected areas can arise, as parts of the image may have gone in different directions.

$$M_o(x, y, t) = M_s(x, y, t) \circ \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix} \tag{4.23}$$

After this, the moments hard threshold, $\mu_h = 25$, is applied to M in order to obtain image M_h

(see Figure 4.16f and equation (4.24)).

$$M_h(r,c,t) = \begin{cases} min, & \text{if } M(x,y,t) \leq \mu_h \\ max, & \text{otherwise} \end{cases} \tag{4.24}$$

Now, the list of blobs L_o present in M_o is compared to the list of blobs L_h found in M_h. The aim is to verify if each blob detected with the hard threshold is contained in a spot detected with the soft threshold. The spots that do not meet this condition are discarded. Finally, the common blobs are depicted as white areas over a black background in a resulting binary image, called refined moments image M_r, and shown in Figure 4.16g, only contains the blobs that have met the previous condition. This image is used during the people candidate blobs refinement phase to improve the certainty about the human presence.

Refinement of Human Candidate Blobs

The first two stages of this phase (the people vertical and horizontal delimiting for every sub-ROI) are performed in exactly the same way as in the people detection in a single frame. However, a major change occurs at the end of the people horizontal delimiting stage. At this point, the equivalent region of the sub-ROI $sR_{\varsigma,\kappa,\alpha}$ is scanned in the image M_R obtained at the end of the image motion analysis stage. If any pixels are found which value is set to max, the hard threshold is not applied in the human confirmation stage, since a warm human candidate has been detected whose movement is different to the majority motion in the scene, so a major trust in that candidate is assumed. If there were no pixels found with value set to max, the human candidate is treated the same way as it was in the human detection based on a single frame algorithm. The humans found are finally enlisted into the list L_{OF}.

4.3.2. People Segmentation in Color

Since humans in the visible spectrum cannot be distinguished using only the frame color information, two different approaches are developed using the motion of the elements in the scene. Both approaches use the current visible frame $I_V(t)$ to update their information. An overview of the color segmentation system is offered in Figure 4.19.

Our first proposal uses the motion history of the elements in the scene (this denomination was introduced in Bobick and Davis (1996)). Now, we use the accumulative computation approach adapted from Delgado et al. (2010), separating $I_V(t)$ into three different color channels which are processed in parallel and later joined for the final processing. The humans detected are placed into a list L_{AC}. On the other hand, the second proposal compares the current frame $I_V(t)$ with an adaptive background regularly updated during the monitored sequence, where human candidates are filtered and extracted after performing an image noise removal from the background subtraction image I_B. The humans detected are enlisted into a list L_{BS}.

In an analogue case to people segmentation in the infrared spectrum, the results from both approaches will be compared in Chapter 5 to establish which segmentation suits better to the chosen

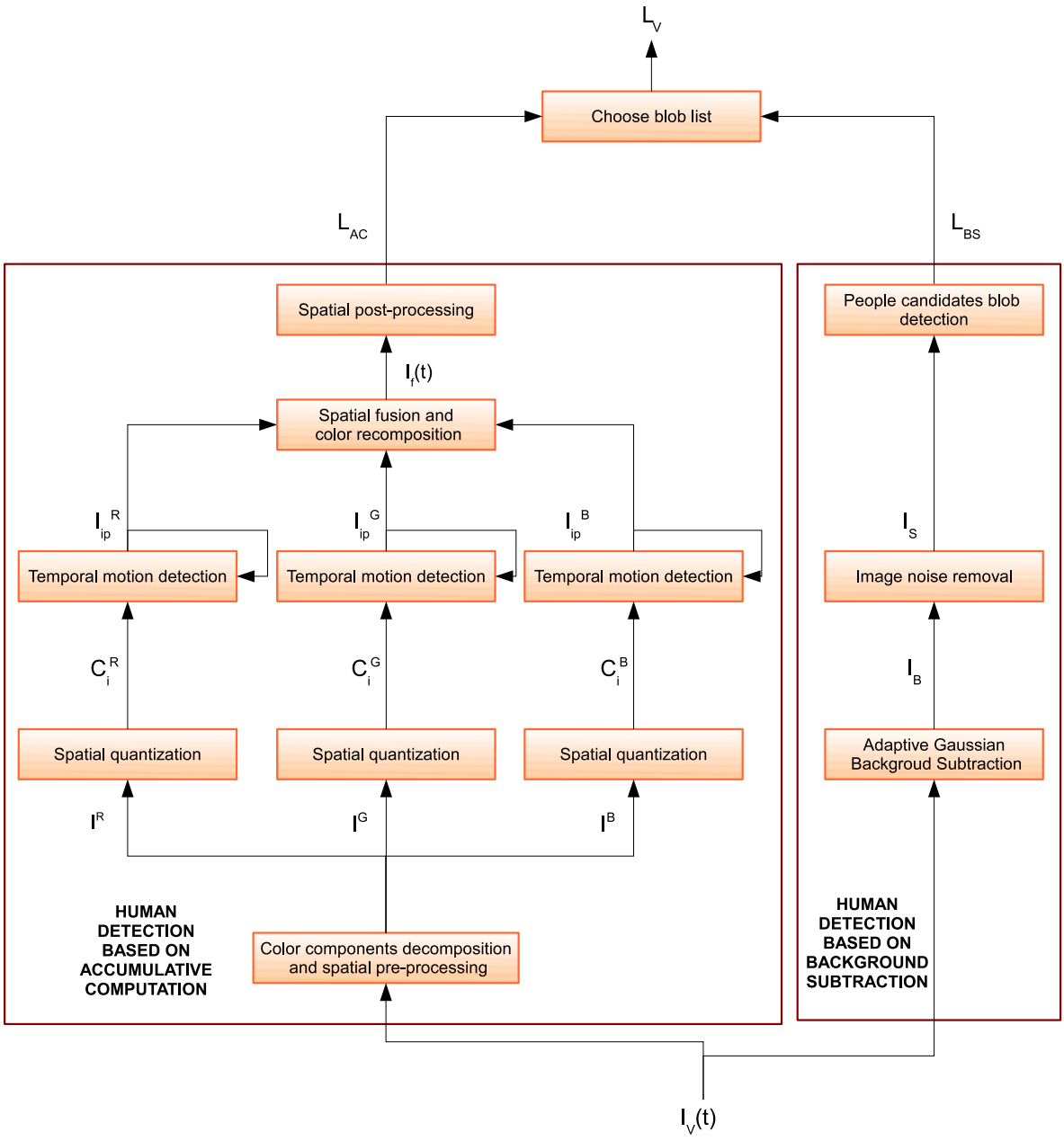


Figura 4.19: Overview of the color segmentation system.

experimental environment. This blob list will be named L_V and used as an input for the later fusion stage.

4.3.2.1. Human Detection in Color Based on Accumulative Computation

An approach for human detection in color video using accumulative computation is implemented and tested. This algorithm, introduced in Serrano-Cuerda et al. (2012), starts separating the input

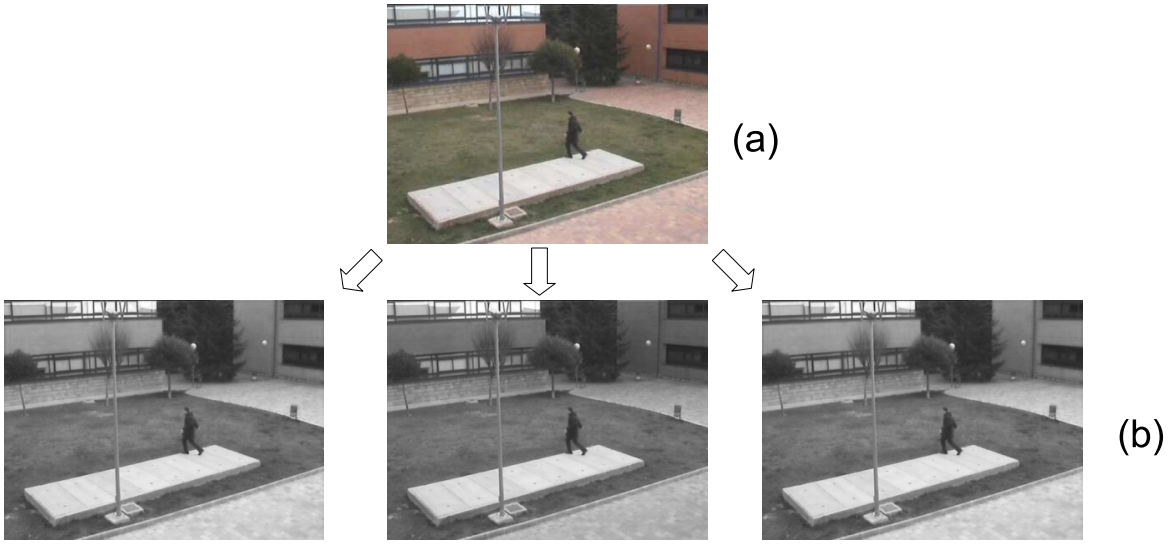


Figura 4.20: Spatial pre-processing of a color image. (a) Original input frame. (b) Separation into bands r, g and b.

image I_V into its three color components. The motion history for each color component is separately built. These three motion histories are later joined with the objective of obtaining a general overview of the scene movement avoiding false positives due to lighting changes.

Color Components Decomposition and Spatial Pre-Processing

The first step of accumulative computation in color video is to separate each RGB input image $I_V(t)$ into three images corresponding to the r, g and b components, as shown in Figure 4.20. At the end of this first stage three gray level images I^r, I^g and I^b are gotten.

Spatial Quantization

The next step covers the need to segment each component κ of the color input image $I_V(t)$, that is to say $I_i^{\gamma}(t)$ (where γ is the color band of the image, i.e. r, g or b), into a preset group of bands (N), as shown in Figure 4.21b. In the RGB color space, we have equation (4.25) for each one of the color components:

$$I_i^{r/g/b}(x,y;t) = \begin{cases} 1, \text{if } I^{r/g/b}(x,y;t) \in [\frac{256}{N} \cdot i, \frac{256}{N} \cdot (i+1) - 1] \\ 0, \text{otherwise} \end{cases} \tag{4.25}$$

where $i \in [0, 1, ..., N-1]$ is the current band and $C^{r/g/b}(x,y,t)$ is the r, g or b component, respectively, of the input color image at time instant t. The value 256 is used in the equation since we are working with images of 256 gray level values.

Temporal Motion Detection

Now, a charge or discharge is performed due to motion detection. The objective of this stage is to

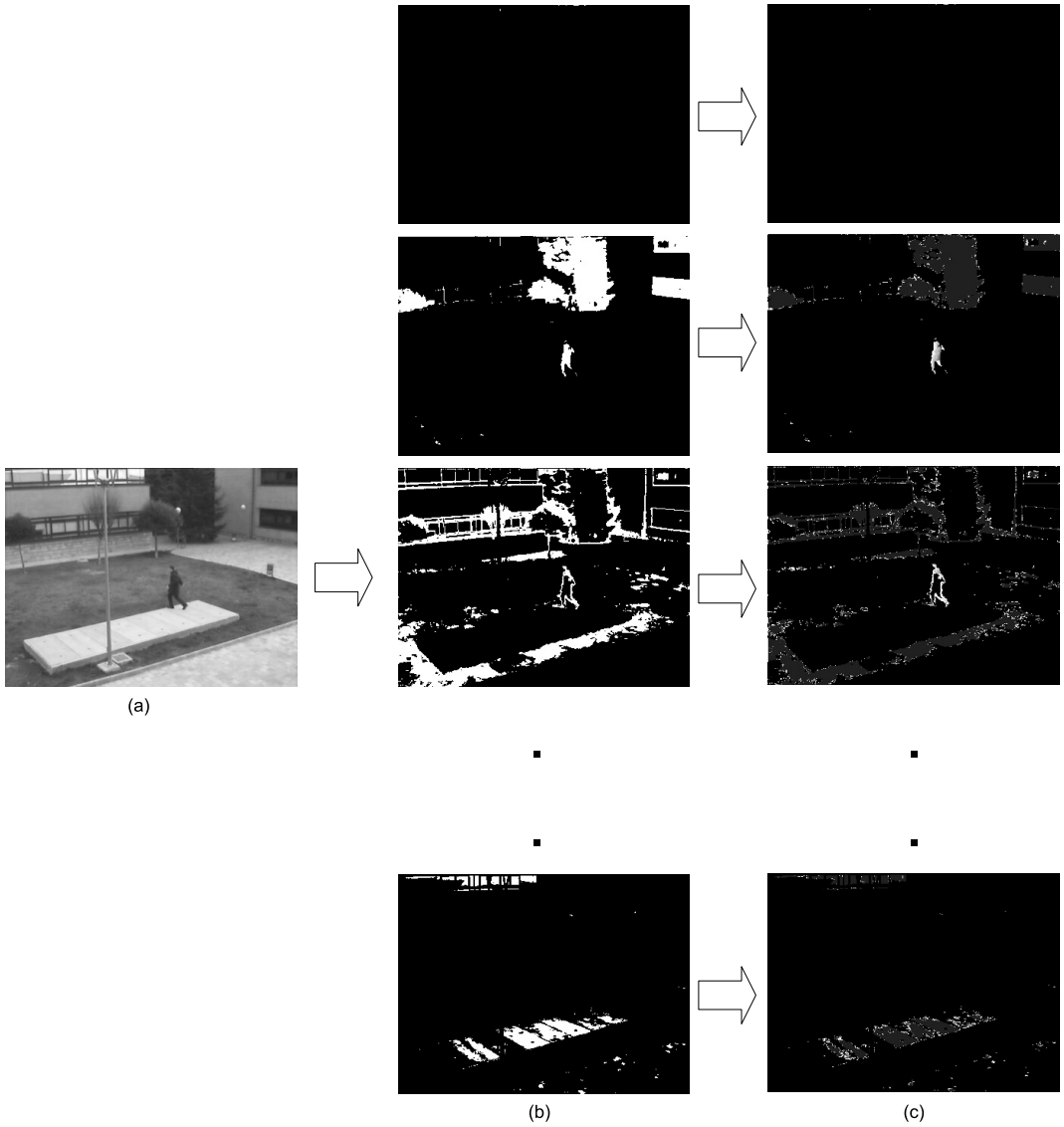


Figura 4.21: Spatial quantization and temporal motion detection in a color video sequence. (a) Original input frame. (b) Result after spatial quantization stage. (c) Result after temporal motion detection stage.

obtain the accumulated charge I_{ic}^{γ} in the 3 color components, storing the accumulative computation charge value for each pixel (x, y) at time instant t. The accumulated charge value for each band ($i = 0, 1, ..., N-1$), q_i^{γ} is calculated for each image pixel as shown in equation (4.26). The result of

this stage is depicted in Figure 4.21c.

$$I_{ic}^{\gamma}(x,y;t) = \begin{cases} v_{dis}, & \text{if } ((I_i^{\gamma}(x,y;t) = 0) \\ v_{sat}, & \text{if } ((I_i^{\gamma}(x,y;t) = 1) \ \wedge \ (I_i^{\gamma}(x,y;t-\Delta t) = 0)) \\ \text{máx}[I_i^{\gamma}(x,y;t-\Delta t) - v_{dm}, v_{dis}], & \text{otherwise} \end{cases} \tag{4.26}$$

For each pixel (x, y) we are in front of three possibilities:

1. The charge value of pixel $x, y)$ is discharged to value v_{dis} (the minimum allowed charge value) when it was not detected at band i.
2. The charge value of pixel (x, y) saturates to value v_{sat} at band i when it was detected at time instant t in that band it was not found in that band at previous instant $t - \Delta t$
3. The charge value of pixel (x, y) is decremented by a value v_{dm} when motion has been detected in consecutive time instants t and $t-\Delta t$. Of course, the permanence value cannot be discharged below the minimum value v_{dis}. Notice that the discharge of a pixel by a quantity of v_{dm} is the way to stop maintaining attention to a pixel of the image that did capture our interest in the past.

Finally, pixels with value below a threshold θ_{per} are set to 0 while the remaining pixels maintain their value as seen in equation (4.27). This way, the pixels belonging to motion which happened several frames ago are removed in order to improve later the accuracy of the shapes of the moving objects.

$$I_i^{per}(x,y;t) = \begin{cases} min, & \text{if } I_{ip}^{\gamma}(x,y;t) \leq \Theta_{per} \\ I_{ip}^{\gamma}(x,y;t), & \text{otherwise} \end{cases} \tag{4.27}$$

Spatial Fusion and Color Recomposition

During this stage, the maximum value of the i band outputs is taken, as shown in equation (4.28), to obtain the blobs associated to a moving element as obtained for each color component γ. The result of this stage is shown in Figure 4.22a for each one of the three bands r, g and b.

$$I_s^{\gamma}(x,y;t) = \arg \max_i I_i^{per}(x,y;t) \tag{4.28}$$

The final segmentation result is obtained as the logical "AND" output of the partial outputs θ, as shown in equation (4.29) and Figure 4.22b.

$$I_f(x,y,t) = I_s^r(x,y,t) \wedge I_s^g(x,y,t) \wedge I_s^b(x,y,t) \tag{4.29}$$

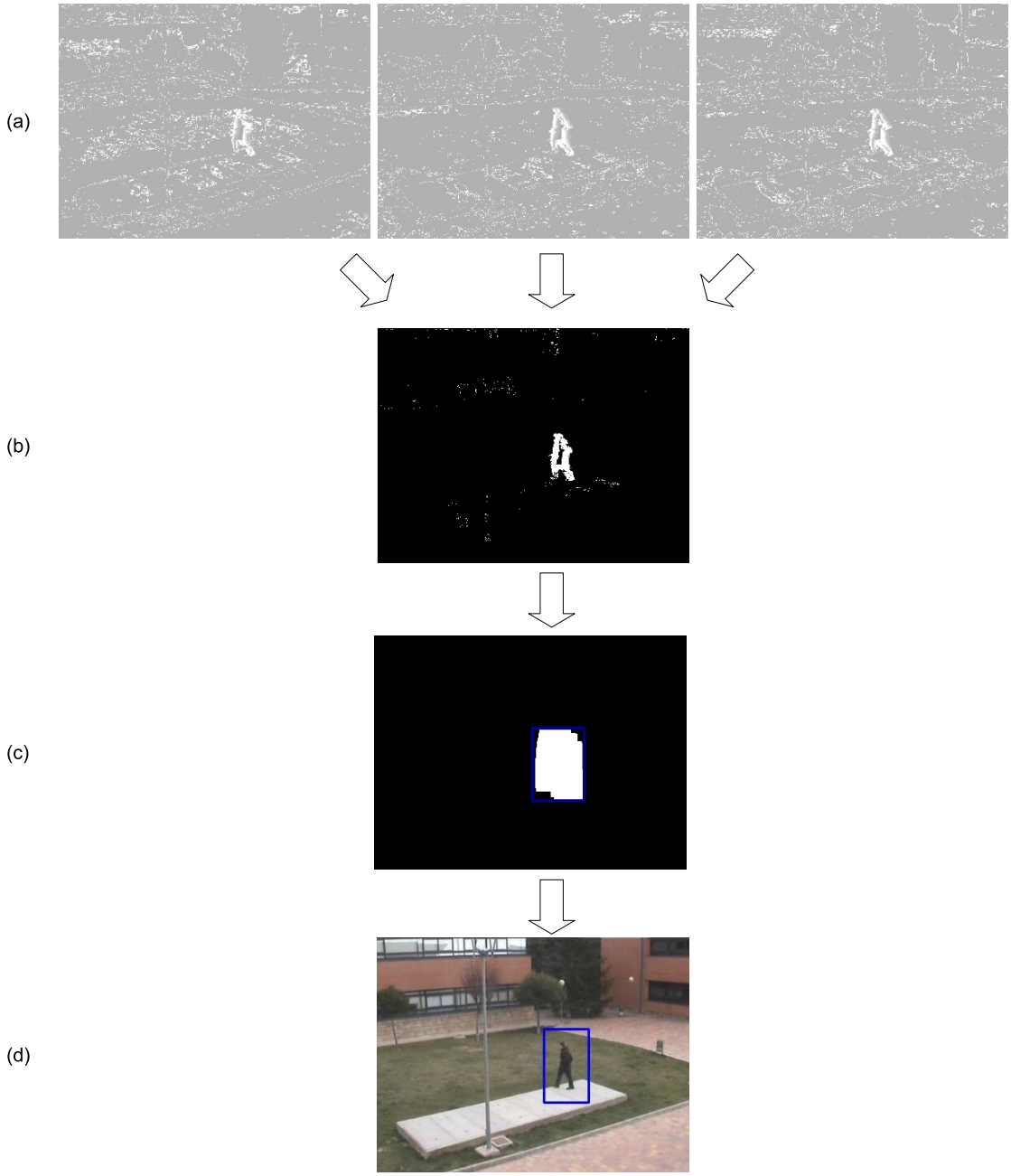


Figura 4.22: Spatial fusion, color recomposition and spatial post-processing of a frame sequence. (a) Spatial fusion for bands r, g and b. (b) Color recomposition of each band. (c) Result of the post-processing stage. (d) Final human detected.

Spatial Post-Processing

This final stage performs a binarization using spatial fusion threshold θ_{obj} (see equation (4.30)). The pixels which level is above the threshold are set to value $max = 255$, while the pixels below θ_{obj} are set to $min = 0$. Once the image has been binarized, a number of morphological operations

are performed in order to remove the image noise. Firstly, an erosion is calculated to remove small or isolated blobs which can result after the segmentation (see equation (4.31)). This is done an amount of times n_{er}. Next, a dilatation operation is performed (a number of times n_{dil}) to improve the remaining spots (see equation4.32). The result of these operations is shown in Figure 4.22c.

$$I_{sp}(x,y;t) = \begin{cases} min, & \text{if } I_f(x,y,t) \leq \Theta_{obj} \\ max, & \text{otherwise} \end{cases} \tag{4.30}$$

$$I_{er}(x,y,t) = I_{sp}(x,y,t) \ominus \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix} \tag{4.31}$$

$$I_{dil}(x,y,t) = I_{er}(x,y,t) \oplus \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix} \tag{4.32}$$

Finally, the blobs are filtered out according to their area A_{dmin} a list of blobs L_{AC} is obtained. The final result of the segmentation is shown in Figure 4.22d.

4.3.2.2. Human Detection in Color Based on Background Subtraction

A human detection system using a classic background subtraction approach is also implemented and tested, adapted from Serrano-Cuerda et al. (2013). After performing an adaptive Gaussian background subtraction, the resulting image is binarized and filtered with a series of morphological operations. Finally, blobs with human shapes are extracted from the binarized image.

Adaptive Gaussian Background Subtraction

An adaptive Gaussian background subtraction is performed on input image $I_V(t)$ obtained from the color camera, as shown in Figure 4.23a. The subtraction is based on a well-known algorithm (KaewTraKulPong and Bowden, 2002). The algorithm builds an adaptive model of the scene background based on the probabilities of a pixel to have a given color level. The authors begin their estimation of the Gaussian mixture model of the background by expected sufficient statistics update equations with a learning rate ϱ, then switch to a L-recent window version when the first L samples are processed. The expected sufficient statistics update equations provide a good estimate at the beginning before all N samples can be collected. This initial estimate improves the accuracy of the estimate and also the performance of the tracker allowing fast convergence on a stable background model. The N-recent window update equations give priority over recent data; therefore the tracker can adapt to changes in the environment.

Shadow removal is performed by the comparison of a non-background pixel against the current background components. If the difference in both chromatic and brightness components are within

some thresholds, the pixel is considered as a shadow. With this objective, an effective computational color model similar to the one proposed by Horprasert et al. (1999) is used, consisting of a position vector at the RGB mean of the pixel background, P, an expected chromaticity line, $||P||$, a chromatic distortion, CrD, and a brightness threshold, τ. For a given observed pixel value, $I(x, y)$, a brightness distortion, a, and a color distortion, ClD, from the background model can be calculated as shown in equations 4.33 and 4.34, respectively.

$$a = \arg_t \min(I(x, y) - zP)^2 \tag{4.33}$$

$$ClD = \|I(x, y) - aP\| \tag{4.34}$$

With the assumption of spherical Gaussian distribution in each mixture component, the standard deviation of the k^{th} component σ_k can be set equal to CrD. A non-background observed sample $B(x, y)$ is considered a moving shadow if $a \approx 2{,}5 \times \sigma_{B(x,y)}$ and $\tau < c < 1$.

An example of background model is shown in Figure 4.23b. A shadow detection algorithm, based on the computational color space used in the background model, is also used. After the background segmentation is performed, an initial background segmentation image (I_B is obtained as shown in Figure 4.23c.

Removal of Image Noise

However, the resulting image contains some noise which must be eliminated. Thus, an initial threshold θ_0 is applied, as shown in equation (4.35), where min is fixed to 0 (since we are obtaining binary images) and max is the maximum gray level value that a pixel can have in I_B (e.g. 255 for an 8-bit image). The value of this threshold will be experimentally fixed according to the features of the image. The result is shown in Figure 4.23d.

$$I_{Th}(x, y) = \begin{cases} \text{min, if } I_B(x, y) \leq \theta_0 \\ \text{max, otherwise} \end{cases} \tag{4.35}$$

After this operation, two morphological operations, namely opening and closing, are performed to eliminate the remaining noise of the image, obtaining I_S.

Detection of Human Candidate Blobs

Now human candidates must be extracted from I_S. Blobs with area (see equation (4.38)) lower than an area A_{minBS} are discarded, while a series of restrictions similar to the human detection based on a single frame in the infrared spectrum are imposed to the remaining ones, establishing a ROI for each blob detected. Let us remind that a ROI is defined by its coordinates (x_{start}, y_{start}) and (x_{end}, y_{end}). Since background subtraction usually extracts the humans in their entirety (like the human detection based on a single frame), similar restrictions are used, using height/width ratios as

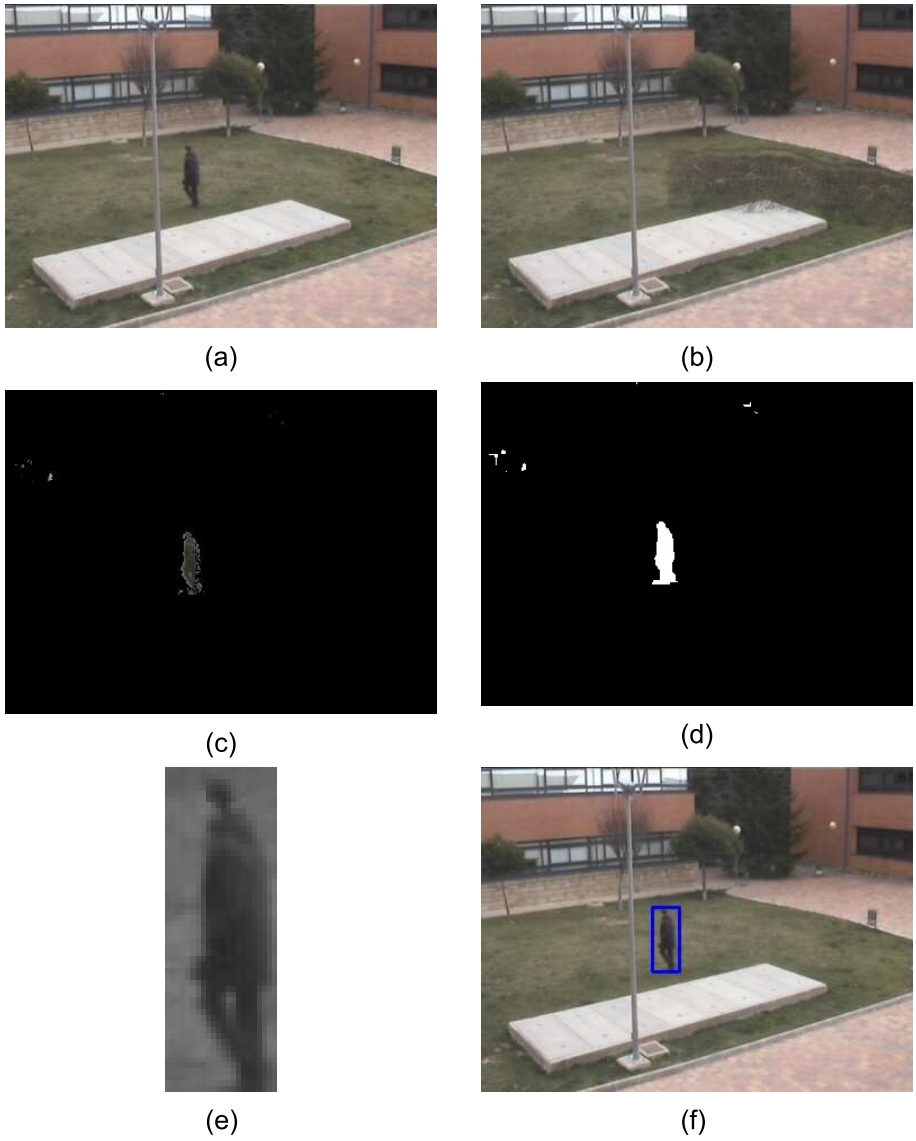


Figura 4.23: Stages of the background segmentation algorithm. (a) Original input frame. (b) Background model calculated. (c) Foreground image calculated. (d) Foreground image after binarization. (e) ROI extracted from the original image. (f) Final result.

shown in equation (4.39). A ROI that satisfies the criteria is shown in Figure 4.23e while the final result is shown in Figure 4.23f.

$$h_R = y_{end} - y_{start} \tag{4.36}$$

$$w_R = x_{end} - x_{start} \tag{4.37}$$

$$A_R = h_R \times w_R \tag{4.38}$$

$$hwR = \frac{h_R}{w_R} \tag{4.39}$$

4.4. People Fusion and Tracking

In Chapter 3 it was established that the fusion, identification and tracking levels are interconnected, so it has been decided to describe them in a common section. Now, it is important to mention here that fusion, identification and tracking levels have been implemented as a rule-based system.

Indeed, one can often represent the expertise that someone uses to do an expert task as the one we are facing with rules. A rule means a structure which has an "IF" component and a "THEN" component. The statement, or set of statements, after the word "IF" represents some pattern which you may observe. The statement, or set of statements, after the word "THEN" represents some conclusion that you can draw, or some action that you should take. A rule-based system, therefore, either identifies a pattern and draws conclusions about what it means, or identifies a pattern and advises what should be done about it, or identifies a pattern and takes appropriate action. In our case all rules will be of the action type, as shown next:

RULE rule: IF (condition)
THEN `action`

A visual overview of the proposed people fusion and tracking system is shown in Figure 4.24. The inputs to people fusion and tracking are the blob lists L_V and L_{IR} obtained from the previous *Segmentation* level, as well as the confidence degree values C_V and C_{IR} established by the *Video Acquisition* level and the tracking list $L_{HT}(t - \Delta t)$ generated in the previous time instant. First, a *Fusion of Human Candidates* is performed using the inputs previously mentioned, generating a list L_{BF}. The blobs which will be added are decided according to the confidence degree values C_V and C_{IR} as well as the coordinates from the blobs on each list. Then, each blob BF from L_{BF} is identified by a *Fusion Blobs Identification*. They can be transformed into identified humans and added to a list L_{ID} where a label is assigned to identify each human. An identified human is a blob BF from L_{BF} which has been assigned a label. This label can correspond to the nearest human from the tracking list of the previous instant $L_{HT}(t - \Delta t)$ or establish that the blob is a new detected human. Finally in a *Tracking* level, a new list $L_{HT}(t)$ corresponding to the current time instant is generated with the identified humans from L_{ID}. Those humans from $L_{HT}(t - \Delta t)$ whose labels have not been assigned to any identified human on L_{ID} are analyzed using the confidence of the spectrum where they were detected to establish if they have left the scene (and will not be included in $L_{HT}(t)$) or if they are occluded or simply have been confused with the background and have not left the scene. In that last case, their coordinates will be updated according to their speed and trajectory and will be enlisted in

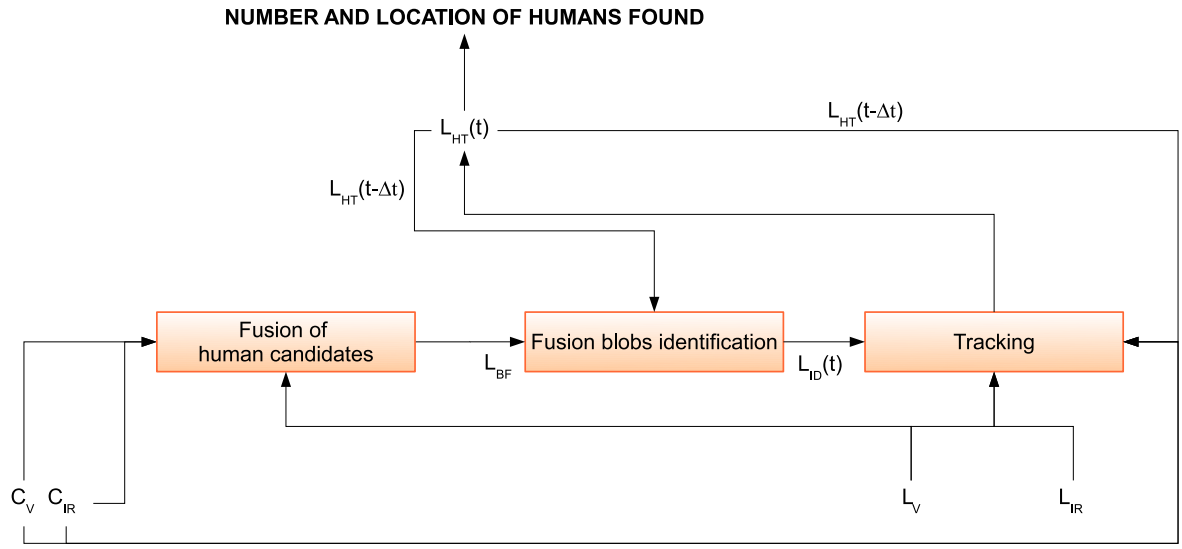


Figura 4.24: Overview of the people fusion and tracking system.

$L_{HT}(t)$.

Before the algorithm starts, the coordinate limits for the infrared camera within image I_V acquired by the visible spectrum camera must be fixed with the homography equation (4.1) explained in subsection 4.2.2. This must be done since it has been previously established that we will work with I_V because it covers a greater surface of the scenario. These coordinates are noted as $(x_{min_{IR}}, y_{min_{IR}})$ for the upper left limits and $(x_{max_{IR}}, y_{max_{IR}})$ for the lower right limits and the rectangle delimited by them will be noted as R_{IR}.

4.4.1. Fusion of Human Candidates

The first stage of the algorithm starts with the analysis of lists L_V and L_{IR} using the confidence degrees C_V for the visible spectrum and C_{IR} for the infrared spectrum. These lists L_V and L_{IR} contain the output blobs of the color and infrared segmentations, respectively. Each blob (B_V or B_{IR}) contains the upper left coordinates (x_{min}, y_{min}) and the lower right coordinates (x_{max}, y_{max}) of the ROI which delimits it. Examples of lists L_V and L_{IR} are shown in Table 4.4 and Table 4.6, respectively. A list of fusion blobs L_{BF} will be the output of this stage, containing the upper left coordinates of each blob and its lower right coordinates. For each one of these blobs, it is important to know where it was detected since this information will later be useful during the *Tracking* stage. So, the origin of each blob will be noted, set as $INFRARED$ if it is only detected by the human detection on the infrared spectrum, as $VISIBLE$ if only the human detection on the visible spectrum detects it and $COMMON$ if the blob is detected in both spectra.

Tabla 4.3: Rules to insert a blob from the visible spectrum into the list of fusion blobs.

RULE B_V0:	IF ($B_V \notin R_{IR}$) THEN `insert` (B_V,L_{BF})
RULE B_V1:	IF ($C_V > C_{IR}$ AND ($C_{IR} == LOW$) THEN `insert` (B_V,L_{BF})
RULE B_V2:	IF ($C_V > C_{IR}$) AND ($C_{IR} == MEDIUM$) AND $\nexists B_{IR} \in L_{IR} \vert B_{IR} \in ROI_{B_V}$ THEN `insert` (B_V,L_{BF})
RULE B_V3:	IF ($C_V > C_{IR}$) AND ($C_{IR} == MEDIUM$) AND $\exists! B_{IR} \in L_{IR} \vert B_{IR} \in ROI_{B_V}$ THEN `insert` (B_V,L_{BF})
RULE B_V4:	IF ($C_V == C_{IR}$) AND $\nexists B_{IR} \in L_{IR} \vert B_{IR} \in ROI_{B_V}$ THEN `insert` (B_V,L_{BF})
RULE B_V5:	IF ($C_V == C_{IR}$) AND $\exists! B_{IR} \in L_{IR} \vert B_{IR} \in ROI_{B_V}$ THEN `insert` (B_C,L_{BF})
RULE B_V6:	IF ($C_V < C_{IR}$) THEN `ignore` (B_V)

4.4.1.1. Analysis of Color Human Detection Results

First, the output L_V is scanned by applying the rules shown in Table 4.3. The main advantages and weak points of our human detection algorithms in the visible spectrum are taken into account through these rules to decide whether a blob B_V will be inserted into the fusion blobs list L_{BF}. Thus, the greater area covered by the camera in the visible spectrum will be exploited. Also, the inability for these segmentation algorithms to distinguish humans occluding themselves in a group will be addressed. A goal is to improve, when possible, the results with the help of the infrared human detection. Examples for each rule explained are shown in Figures 4.25 and 4.26.

If the limits of a blob B_V are totally or partially outside of the limits covered by the infrared camera, the blob is always inserted into L_{BF} using the action `insert` (B_V,L_{BF}) in Rule B_V0, since the color camera is our solely source in that area of the scenario. An example of this situation is shown in Figure 4.25a, where the human is directly inserted into L_{BF} since he/she is outside the area covered by the infrared camera. It will later be seen that this information can be useful even if there is no proper fusion using this rule, since there is only a camera providing information outside the limits covered by the infrared camera. However, the tracking stage will use this information later on and any

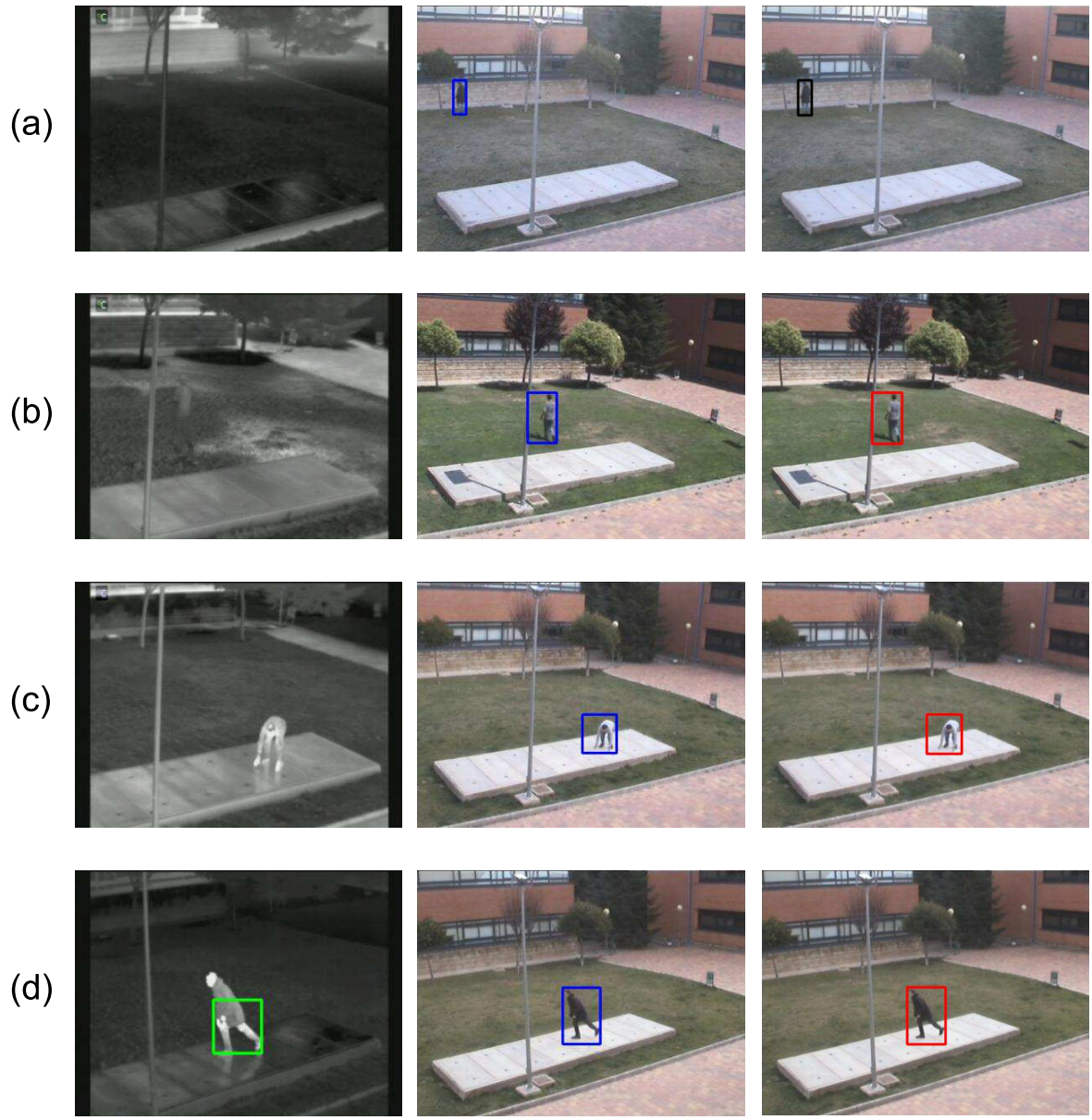


Figura 4.25: Examples of the first rules to insert a blob from the visible spectrum into the list of fusion blobs. (a) Rule B_V0 (b) Rule B_V1 (c) Rule B_V2 (d) Rule B_V3

false positive which could arise from this situation will easily be removed in later frames.

If the confidence of the visible spectrum is greater than the reliability of the infrared spectrum, four different possibilities are considered:

- If the confidence of the infrared spectrum is set to LOW, the blob B_V is inserted by Rule B_V1 into L_{BF}, since the infrared spectrum is not reliable and could affect negatively the fusion results at this stage. This could happen because the infrared spectrum has a high number of false negatives and is not able to divide groups in those conditions. It can be appreciated in Figure

4.25b that the human could not be detected in the infrared spectrum, but the detection in the visible spectrum is considered; so, the blob is inserted into the list L_{BF}.

- If C_{IR} is assigned to $MEDIUM$, the infrared spectrum reliability shows that it could still be useful to divide humans supported by the results of the visible spectrum. Thus, common detections for both spectra will be searched in an extended area from the original ROI of B_V. This ROI_{B_V} is extended in a $40\,\%$ both in the height and width of B_V and is defined by the rectangle $(B_V.x_{min} - B_V.width \times 0{,}2, B_V.y_{min} - B_V.height \times 0{,}2), (B_V.x_{max} + B_V.width \times 0{,}2, B_V.y_{max} + B_V.height \times 0{,}2)$, being $B_V.width$ and $B_V.height$ defined as shown in equations (4.40) and (4.41), respectively. Thus, a $40\,\%$ of tolerance is offered. If any single human could not be found by the infrared spectrum in ROI_{B_V}, B_V is inserted by the Rule B_V2, since the confidence on this spectrum is still higher than its value for the infrared spectrum. An example of this case is provided in Figure 4.25c, where the infrared segmentation is unable to distinguish the human from the platform where he/she is standing. Nonetheless, the human detection in the visible spectrum returns a true positive which is inserted into the final image result.

$$B_V.width = B_V.x_{max} - B_V.x_{min} \tag{4.40}$$

$$B_V.height = B_V.y_{max} - B_V.y_{min} \tag{4.41}$$

- If human detection in the infrared spectrum could only find a human in ROI_{B_V} and its confidence value C_{IR} is set to $MEDIUM$, then Rule B_V3 inserts B_V into L_{BF} with the original dimensions, because the reliability on the visible spectrum is still higher than the confidence on the infrared spectrum and that is why it will usually provide a more accurate view of the location of single humans. In 4.25d an example of this circumstance is shown. The human's head could not be detected by the infrared segmentation, but the human detection in the visible spectrum succeeds with the right dimensions. Thus, the total dimensions of the human are inserted into L_{BF}.
- The last case is as described next. The value C_{IR} is assigned to $MEDIUM$ and there is more than one human detected in this spectrum in ROI_{B_V}. When this condition happens, the blobs found by the infrared spectrum will be inserted later on when L_{IR} is analyzed, since the infrared spectrum is still reliable to divide groups when its confidence value is greater than LOW.

If the confidence for both spectra is the same, a series of different situations are again considered and infrared human detections are newly searched in the area of ROI_{B_V} in order to improve the original detection of the color segmentation. The following situations are considered:

- If the infrared human detection did not find any humans in the area of ROI_{B_V}, B_V is inserted by Rule B_V4 into L_{BF} using the action `insert` (B_V,L_{BF}) since that detection is as reliable

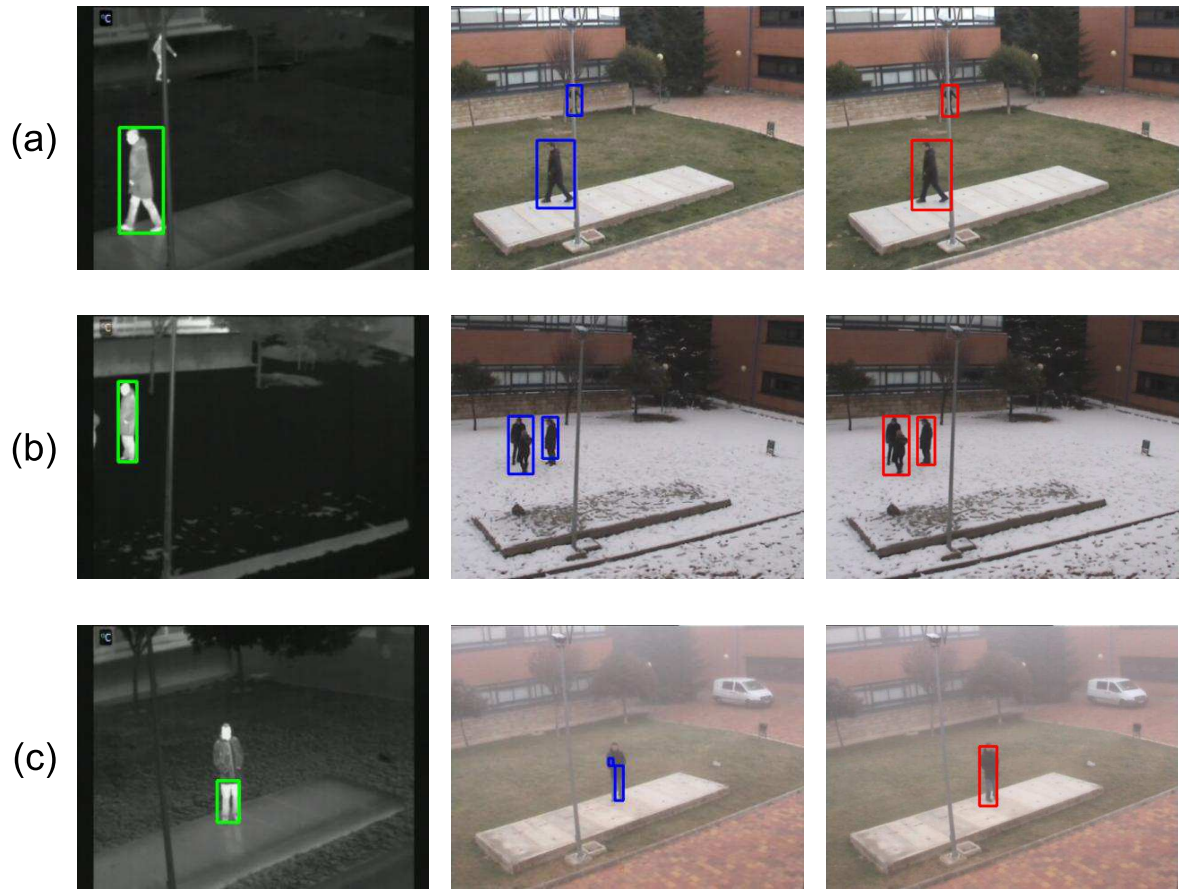


Figura 4.26: Examples of the remaining rules to insert a blob from the visible spectrum into the list of fusion blobs. (a) Rule B_V4. (b) Rule B_V5. (c) Rule B_V6.

as the results of the infrared spectrum, and the human could have not been detected by the infrared human detection in that particular case. An example of this particular situation is shown in Figure 4.26a, where the human behind the light pole is not detected by the infrared segmentation, but the human detection in the visible spectrum returns a successful detection, which is considered in the final result.

- If the infrared spectrum list L_{IR} has only detected a single blob B_{IR} in the area of B_V, a blob B_C resulting from combining the coordinates of both B_{IR} and B_V will be enlisted using the action `insert` (B_C,L_{BF}) in Rule B_V5 as detected by both spectra, with its coordinates defined by equations (4.42), (4.43), (4.44) and (4.45). This action is performed because the color segmentation could have partially detected a human, while the infrared segmentation might have detected the whole person. An example is shown in Figure 4.26b, where the human's feet were not detected by the visible spectrum segmentation but the infrared segmentation did it correctly. Thus, the right dimensions of the human are considered in the final results image. However, if more than one blob from L_{IR} is found in the area, those blobs will later be added to L_{BF} when

L_{IR} is scanned.

$$B_C.x_{min} = min(B_V.x_{min}, B_{IR}.x_{min}) \tag{4.42}$$

$$B_C.y_{min} = min(B_V.y_{min}, B_{IR}.y_{min}) \tag{4.43}$$

$$B_C.x_{max} = max(B_V.x_{max}, B_{IR}.x_{max}) \tag{4.44}$$

$$B_C.y_{max} = max(B_V.y_{max}, B_{IR}.y_{max}) \tag{4.45}$$

- Finally, if the confidence of the color segmentation C_V is lower than the reliability of the infrared human detection C_{IR}, B_V is initially "ignored" by the action `ignore` (B_V) in Rule B_V6, but it will not be completely discarded since it can still provide information in the final tracking stage, as it will be explained later. False positives which are not reflected in the final results image are shown in 4.26c in the visible spectrum.

As a running example, an L_V list depicted in Table 4.4 will be used. This table corresponds to the results of the segmentation in the visible spectrum shown in Figure 4.27. The blob marked as 2 is outside the limits ($x_{min_{IR}}, y_{min_{IR}}$) of R_{IR}, since these limits have been established as (84,69) and its x_{max} coordinate is 70, so it is totally outside the field of view of the infrared camera. Thus, Rule B_V0 is applied and the blob is enlisted into L_{BF}. The other blob is inside the limits of R_{IR} and both confidences have been set as $HIGH$, so an area delimited by the rectangle (65,95),(135,178) is scanned. Since two blobs have been found as shown in Table 4.6, the blobs will later be enlisted when L_{IR} is examined.

Tabla 4.4: Color blobs list L_V.

Blob	x_{min}	y_{min}	x_{max}	y_{max}
0	75	107	125	166
1	51	111	70	163

4.4.1.2. Analysis of Infrared Human Detection Results

Next L_{IR} is also analyzed and rules from Table 4.5 are applied. It has been previously shown that the proposed infrared human detection algorithms are able to obtain single humans and that characteristic will be used in the following rules. Since sometimes color segmentation can cover an area much bigger than the real human dimensions, a bigger search area $ROI_{B_{IR}}$ is defined as $(B_{IR}.x_{min} - B_{IR}.width \times 1{,}5, B_{IR}.y_{min} - B_{IR}.height \times 1{,}5), (B_{IR}.x_{max} + B_{IR}.width \times$

Figura 4.27: Results of the segmentation in the visible spectrum.

1,5, $B_{IR}.y_{max} + B_{IR}.height \times 1{,}5$) with $B_{IR}.width$ and $B_{IR}.height$, defined as shown in equations (4.46) and (4.47) respectively, analogue to (4.40) and (4.41). Also, a region of interest ROI_{BF} is defined, containing the rectangle which covers the limits of BF, i.e the ROI limits are $(BF.x_{min}, BF.y_{min}), (BF.x_{max}, BF.y_{max})$. Examples for each explained rule are shown in Figure 4.28.

$$B_{IR}.width = B_{IR}.x_{max} - B_{IR}.x_{min} \tag{4.46}$$

$$B_{IR}.height = B_{IR}.y_{max} - B_{IR}.y_{min} \tag{4.47}$$

If the confidence value of the infrared spectrum is greater than the reliability of the visible spectrum, each blob B_{IR} is enlisted into L_{BF} by action `insert` (B_{IR},L_{BF}) in Rule $B_{IR}0$ using an analogue case to the situation seen in the analysis of L_V in the Rule B_V1. This action is performed because in that situation the infrared spectrum is the most reliable source of information in the scene and it should always be considered, as shown by Figure 4.28a, where the scene is so dark that the visible spectrum segmentation is unable to detect the human but the infrared spectrum distinguishes him/her correctly and, thus, it is inserted into the final results image.

The action `insert` (B_{IR},L_{BF}) is also used in Rule $B_{IR}1$ when the infrared human detection has detected more than one human in a zone where the color segmentation was able to find a single detection. Now, these additional detections constitute a possible group. This case is shown in Figure 4.28b, where the infrared human detection successfully distinguished two humans where the human detection in the visible spectrum only found one, inserting these two humans into the final results. The condition of blobs B_V found in ROI_{IR} but not already inserted into L_{BF}, is added in order to avoid duplicate results from the rules for the analysis of L_V and ensure that this rule only inserts those possible groups that were not separated by the visible segmentation.

Tabla 4.5: Rules to insert a blob from the infrared spectrum into the list of fusion blobs.

RULE $B_{IR}0$:	IF ($C_{IR} > C_V$) THEN `insert` (B_{IR},L_{BF})
RULE $B_{IR}1$:	IF ($C_{IR} > LOW$ AND $C_V > C_{IR}$) AND $\exists B_V \vert (B_V \in ROI_{B_{IR}} AND B_V \notin L_{BF})$ THEN `insert` (B_{IR},L_{BF})
RULE $B_{IR}2$:	IF ($C_V == C_{IR}$) AND $\nexists BF \in L_{BF} \vert B_{IR} \in ROI_{BF}$ THEN `insert` (B_{IR},L_{BF})
RULE $B_{IR}3$:	IF ($C_{IR} = C_V$) AND $\exists B_V \vert (B_V \in ROI_{B_{IR}} AND B_V \notin L_{BF})$ THEN `insert` (B_{IR},L_{BF})
RULE $B_{IR}4$:	IF ($C_{IR} < C_V$) AND $\nexists B_V \vert (B_V \in ROI_{B_{IR}}$ THEN `ignore` (B_{IR})

Also, an equivalent application of Rule B_V2 with the action `insert` (B_{IR},L_{BF}) is used in Rule $B_{IR}2$ when the infrared human detection has made a detection which the color segmentation was unable to find and both spectra have the same confidence. In this case, detections must be considered without regarding their origin spectrum, since they are equally valuable, according to the assigned confidence values. Thus, the human detection in the visible spectrum could have suffered a punctual false negative. An example is shown in Figure4.28c, where the human could not be detected by the visible segmentation since he/she is hidden by the tree's shadow, but the infrared human detection distinguishes him/her successfully. The condition of blobs L_F not found in ROI_{IR} is newly added in order to avoid duplicated results.

Also, those blobs which did not satisfy Rule B_V3, or did not have any blob B_V previously detected, are enlisted using the same action `insert` (B_{IR},L_{BF}) by Rule $B_{IR}3$ if the confidence on both spectra have the same value. Thus, the groups divided by the infrared human detection, and not enlisted when L_V was analyzed, are included. Notice that the same condition from Rule $B_{IR}1$ is applied again to newly avoid duplicated results with blobs from L_V. An example of this rule is offered in Figure 4.28d where two people could be distinguished from the big possible human found in L_V. These humans are finally enlisted into L_{BF}.

Finally, those blobs without equivalent detections in L_V are not enlisted if the confidence on the infrared spectrum is lower than the reliability of the visible spectrum by Rule $B_{IR}3$ with action `ignore` (B_{IR}). Notice that these blobs will not be discarded as they could be useful in the later *Tracking* stage. This situation can be appreciated in Figure 4.28e, where the infrared camera has a very low visibility and only detects a false positive that is ignored by the final results of the frame.

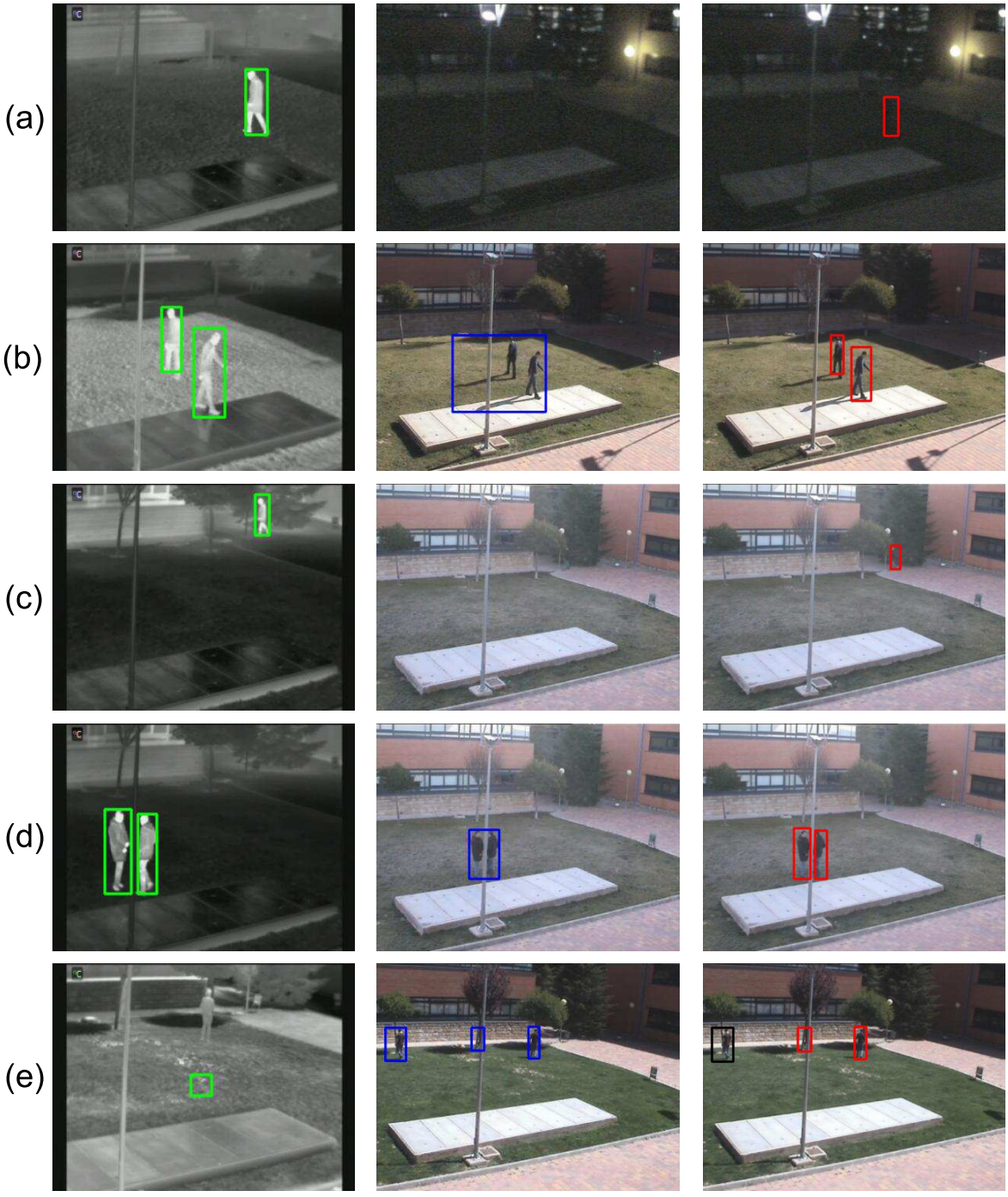


Figura 4.28: Examples of the rules to insert a blob from the infrared spectrum into the list of fusion blobs. (a) Rule $B_{IR}0$. (b) Rule $B_{IR}1$. (c)Rule $B_{IR}2$. (d)Rule $B_{IR}3$ (e)Rule $B_{IR}4$.

Now, in our running example, L_{IR} (shown in Table 4.6 with each blob associated to this table shown in Figure 4.29) will be analyzed starting with the blob marked as 0. Since the confidence in the infrared spectrum is $HIGH$ (the same value established for the visible spectrum in this example) and

Tabla 4.6: Infrared blobs list L_{IR}.

Blob	x_{min}	y_{min}	x_{max}	y_{max}
0	84	104	97	155
1	98	119	111	164

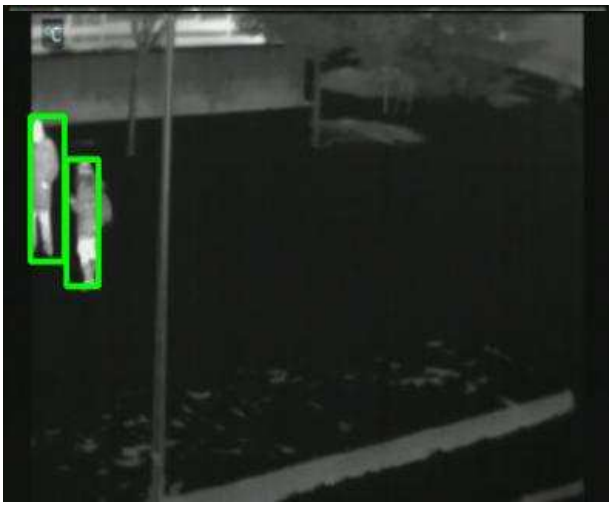


Figura 4.29: Results of the human detection in the infrared spectrum.

the blob is not in the same area as any blob in L_{BF} (remember that the blobs in L_V were analyzed earlier), the blob is enlisted. An analogue case happens with a blob marked as 1 in the table. Notice that these two blobs are inside the area of blob 0 in Table 4.4 and that blob was not enlisted when L_V was analyzed, so the two people composing that group are now enlisted separately.

The final fusion blobs list L_{BF} is shown in Table 4.7.

Tabla 4.7: Fusion blobs list L_{BF}.

Blob	x_{min}	y_{min}	x_{max}	y_{max}	Origin
0	51	111	70	163	COLOR
1	84	104	97	155	IR
2	98	119	111	164	IR

4.4.2. Fusion Blobs Identification

The next stage starts with the creation of a distance matrix d using the Euclidian distances between humans in the previous human list $L_{HT}(t - \Delta t)$ and the blobs in L_{BF}. The equation (4.48) shows how the element $d_{BF,HT}$ is calculated using the distance between a blob BF and and human HT, where (x_{c_F}, y_{c_F}) and (x_{c_T}, y_{c_T}) are the center coordinates of the blob and the human, respectively. The calculations for x_{c_F} and y_{c_F} are shown in equations (4.49) and (4.50) respectively, with analogous calculations needed for x_{c_T} and y_{c_T}. Next, d is scanned associating to each blob from L_{BF} its closest human from $L_{HT}(t - \Delta t)$. This association is made by assigning the label of the human HT to the

blob now identified as human (which will be denoted as ID). If the distance to its closest human is greater than a distance D_{max}, previously fixed based on the scene features, the blob is considered to contain a new object in the scene and a new label is assigned to that identified human. These labels consist in a numeric value which starts at 0 with the first human identified in the sequence. After that, if a new human appears at the scene its label will be 1 and, in later frames of the sequence, if the human HT previously labeled with 0 is found again, the label 0 will be newly assigned to the identified human ID to denote that it has been found in the scene.

$$d_{BF,HT} = \sqrt{(x_{c_{BF}} - x_{c_{HT}})^2 + (y_{c_{BF}} - y_{c_{HT}})^2} \tag{4.48}$$

$$BF.x_cF = \frac{BF.x_{max} - BF.x_{min}}{2} \tag{4.49}$$

$$BF.y_cF = \frac{BF.y_{max} - BF.y_{min}}{2} \tag{4.50}$$

Next, identified humans ID with an associated blob BF will be updated and enlisted into the current identification list L_{ID}. Their ($ID.x_{min}, ID.y_{min}$) and ($IDx_{max}, ID.y_{max}$) coordinates will be assigned to the equivalent coordinates of the blob BF. The center coordinates of BF will be the new center coordinates of ID and they will be enlisted in its coordinates list.

A new variable called *credibility* is introduced here. This value shows the probability that a human is actually in the scene, that we are not in front of a false positive or person who has left the monitored scenario. A greater value of credibility for a human means that it will take a longer time to remove a human from the scene when he/she is not detected during a certain amount of frames. The maximum value for credibility has been fixed to $CR_{MAX} = 5$. This number is fixed because it has been previously established that the acquisition cameras grab images at a frame rate of 5 frames per second. In the *Tracking* stage it will be explained that, when a human is not detected, credibility will be decremented by 1 until its value reaches $CR_{MIN} = 0$ to consider that he/she has left the scene, so a human will take a maximum time interval of a second without being detected to be removed. The credibility of a human ID at time instant t, that is $ID.credibility(t)$, is updated as shown in equation (4.51).

$$ID.credibility(t) = \begin{cases} CR_{MAX}, \text{if } ID.credibility(t - \Delta t) = CR_{MAX} \vee F.origin = COMMON \\ \\ ID.credibility(t - \Delta t) + 1, \text{otherwise} \end{cases} \tag{4.51}$$

Also, the average speed of a human on each axis is modified (see equations (4.52) and (4.53)), where H is the total amount of coordinates stored for the identified human ID, with a maximum value of H_{MAX}, experimentally fixed as 15. Since an acquisition speed of frame rate of 5 frames per second

is considered, the value 15 means that the average speed of a human in the last three seconds is used.

$$ID.S_X = \frac{1}{H} \times \sum_{i=1}^{H} (x_{c_i} - x_{c_i-1}) \tag{4.52}$$

$$ID.S_Y = \frac{1}{H} \times \sum_{i=1}^{H} (y_{c_i} - y_{c_i-1}) \tag{4.53}$$

Those blobs BF without any associated human HT will be enlisted into the identified humans list L_{ID} and the new identified human ID properties will be initialized with its coordinates set those of the original blob BF. The initial credibility value $ID.credibility$ of this new identified human will be set to the minimum value 0.

Finally, the identified ID will be enlisted into the final identified humans list L_{ID} shown in Table (4.8). The human outside the zone covered by the infrared spectrum was previously detected by the color segmentation, so its credibility has been updated to 3. On the other hand, the infrared camera has divided a group which previously could not be separated, so these new humans now appear with credibility CR_{min}.

Tabla 4.8: Humans identified in the current iteration of the system.

Identification	x_{min}	y_{min}	x_{max}	y_{max}	Credibility	Origin
0	51	111	70	163	3	COLOR
1	84	104	97	155	0	IR
2	98	119	111	164	0	IR

4.4.3. Tracking

Finally, those humans from the previous instant $L_{HT}(t - \Delta t)$ which were not associated to any blob BF in L_{BF} (and therefore not present in L_{ID}) will be analyzed. Before that, identified humans in L_{ID} are enlisted in $L_{HT}(t)$. Now, a new region of interest ROI_F with coordinates $((x_{min} - W_F, y_{min} - H_F)$,$(x_{max} + W_F, y_{max} + H_F)$ will be defined for each unidentified blob BF, where W_F and H_F are the width and height of BF, respectively. A human will be discarded in $L_{HT}(t)$ if a rule of the set of rules seen in Table 4.9 is satisfied, where blobs B_V from L_V and B_{IR} from L_{IR} are searched within ROI_F and a minimum credibility CR_{MIN} (usually set to 0) is used.

If Rule $T0$ was satisfied, HT is discarded by action `discard` (HT), since if it was originally a human and both spectra are equally reliable, both of them should have detected it in the current frame. Rules $T1$ and $T2$ are applied with the same action when a spectrum has more confidence than the other, establishing that HT must have been detected by at least the spectrum with greater confidence. Finally, if Rules $T0$, $T1$ and $T2$ are not satisfied, conditions in Rule $T3$ are checked. When this rule is satisfied, neither a blob B_{IR} from the infrared segmentation or a blob B_V from the segmentation

Tabla 4.9: Rules to establish if a human HT has left the scene.

RULE $T0$:	IF ($C_{IR} == C_V$ AND $HT.origin \neq COMMON$) THEN `discard` (HT)
RULE $T1$:	IF ($C_{IR} > C_V$ AND $HT.origin == COLOR$) THEN `discard` (HT)
RULE $T2$:	IF ($C_V > C_IR$ AND $HT.origin == INFRARED$) THEN `discard` (HT)
RULE $T3$:	IF (($C_{IR} == C_V$ AND $HT.origin \neq COMMON$) OR ($C_{IR} > C_V$ AND $HT.origin \neq COLOR$) OR ($C_V > C_IR$ AND $HT.origin \neq INFRARED$)) AND ($/\exists B_V \vert B_V \in ROI_F$) AND ($ID.credibility < CR_{MIN}$) AND ($\nexists B_{IR} \vert B_{IR} \in ROI_F$)) THEN `discard` ($HT$)

in the visible spectrum could not be found at ROI_F. Notice that blobs from a spectrum with lower confidence than the other and which also were originally not considered as humans are useful in this case, since these detections are taken into account independently from the confidence on their origin spectrum. The credibility of HT is also required to have a value below CR_{MIN} (usually 0) as well. Each rule performs the action `discard` (HT) which causes the human HT to not be enlisted in $L_{HT}(t)$ and it is thus considered to have left the scene.

In any other case, the human will be selected to be updated (although a correspondent blob BF was not found) and the action `update` (HT) will be realized. This operation will cause the credibility $T.credibility$ of HT to diminish as shown in equation (4.54), while its new coordinates will be assigned according to equations (4.55) and (4.56). These equations are applied analogously to its (x_{min},y_{min}) and (x_{max},y_{max}) coordinates. If any of these new coordinates places the human in the exit areas (with its coordinate x_{max} below x_{min_V}, its x_{min} coordinate greater than x_{max_V} or its coordinate y_min greater than y_{max_V}), it will also not be enlisted into $L_{HT}(t)$.

$$T.credibility(t) = \begin{cases} CR_{MIN}, \text{if } HT.credibility(t - \Delta t) > CR_{MIN} \\ \\ HT.credibility(t - \Delta t) - 1, \text{otherwise} \end{cases} \tag{4.54}$$

$$HT.x_{c_t} = HT.x_{c(t-\Delta t)} + HT.S_X \tag{4.55}$$

$$HT.y_{c_t} = HT.y_{c(t-\Delta t)} + HT.S_y \tag{4.56}$$

Finally, the location of the ROI (x_{min}, x_{max}, y_{min} and y_{max}) for each human in L_T as well as the amount of humans in the list are shown as the output of the system. The results of the people segmentation in the visible spectrum for this running example are shown in Figure 4.30a while the results on the infrared spectrum can be seen in Figure 4.30b and the humans detected by the human fusion and tracking system are shown in Figure 4.30c, where the black rectangle indicates that the human is outside the area covered by the infrared camera and the red rectangles that humans can be seen by both cameras. The final list $L_{HT}(t)$ of humans in Figure 4.30c is depicted in Table 4.10.

Tabla 4.10: List $L_{HT}(t)$ of humans in the scene.

Object	x_{min}	y_{min}	x_{max}	y_{max}	Credibility	Origin
0	51	111	70	163	3	COLOR
1	84	104	97	155	0	IR
2	98	119	111	164	0	IR

4.5. Conclusions

The system has been explained in depth in this chapter. After a brief description of each level, they are explained in depth. First, a *Video Acquisition* level grabs the frame from each camera and synchronizes both sources. It also performs a coordinate transformation assigning coordinates in the visible spectrum to the images captured from the infrared camera and establishes a confidence degree for each spectrum based on the features of the images captured by their corresponding camera.

Next, a series of *Segmentation* algorithms are developed for each spectrum. The infrared spectrum approaches start with an algorithm of human detection based on a single frame and later add information from motion of the scene and the camera. These approaches will later be tested in order to decide which algorithm suits better for a given test scenario. Two different human detection ideas are also introduced for the visible spectrum, using motion history from the moving objects detected in the scene and information from an adaptive background subtraction approach. Analogously to the infrared spectrum, these algorithms will also be later compared in our test scenario.

Finally, a *People Fusion and Tracking* mechanism is applied. It uses the information from the previous levels, combining the humans detected in both spectra to improve the robustness of the system. The fusion information is reinforced by the use of a tracking algorithm to reduce the impact of punctual segmentation failures such as false negatives using the location history of the humans previously detected to predict their future coordinates when they are not detected.

The output of the system shows the location and amount of the humans detected in the scene in the current frame.

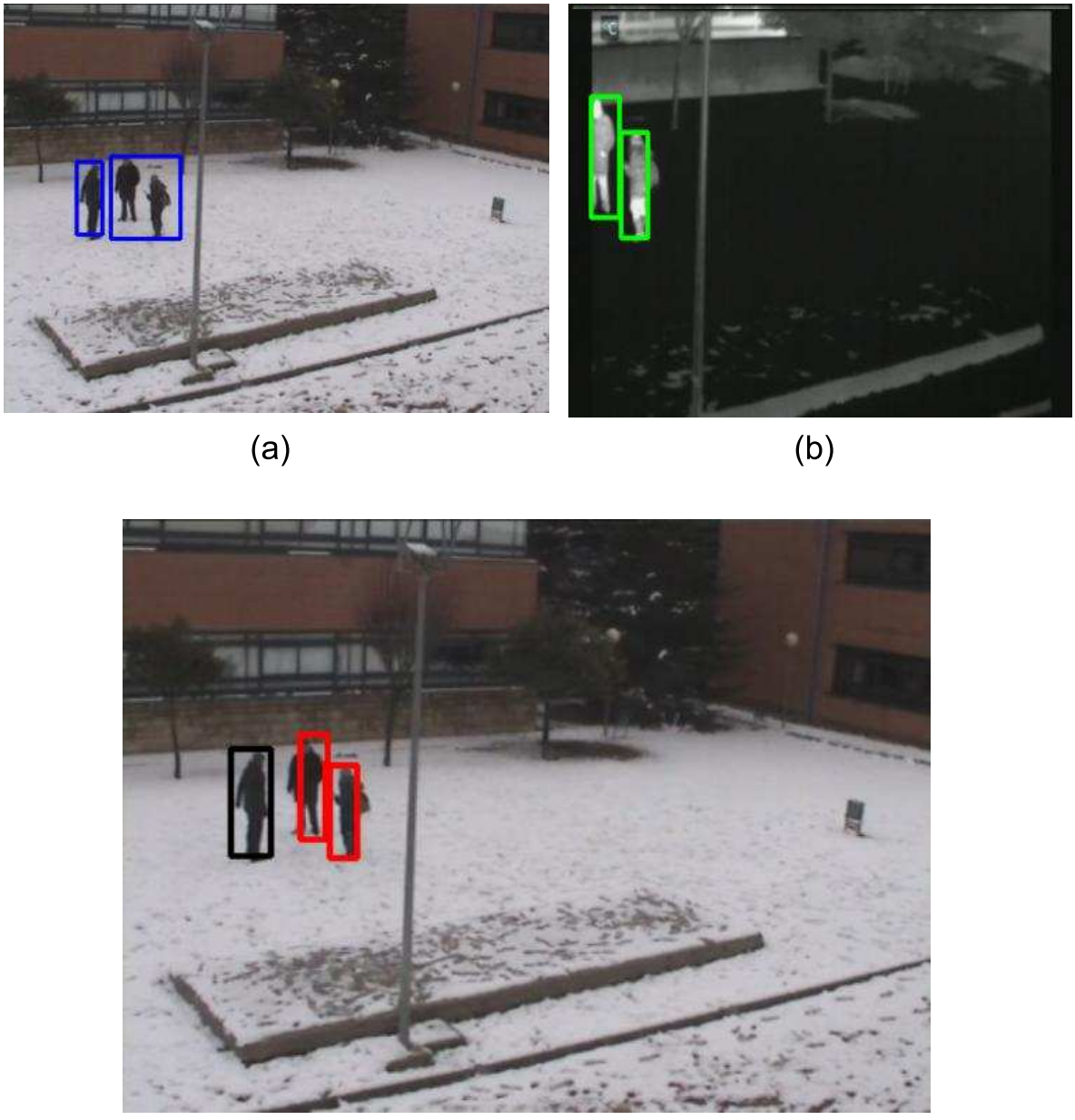


Figura 4.30: Final image results of the system (a) People segmentation in the visible spectrum (a) People segmentation in the infrared spectrum (c) People fusion and tracking

Capítulo 5

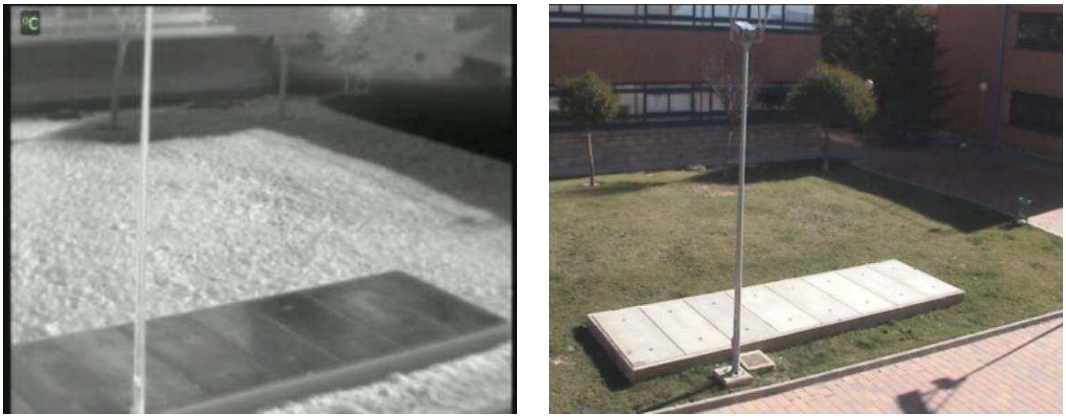
Evaluación del sistema de detección de humanos

En este capítulo se detalla el proceso de evaluación del sistema desarrollado para la detección de humanos. En primer lugar, se describe el entorno escogido para la evaluación de los diversos algoritmos de segmentación desarrollados. Este entorno también es el utilizado para la validación del algoritmo de fusión propuesto. Una vez descrito el entorno, se especifican las secuencias de prueba que se han realizado. Para cada una de estas secuencias, se explica todo lo que sucede en ella, así como las condiciones atmosféricas y de iluminación en que fue tomada.

Después de dejar claro el conjunto de pruebas que se realizarán, se procede a comparar entre sí los diversos algoritmos de segmentación desarrollados e implementados en cada espectro, con el fin de escoger los más idóneos para las condiciones establecidas previamente. Para realizar estas pruebas, con el fin de que sean reproducibles, se aporta también el valor de los diversos parámetros utilizados, así como de los umbrales de confianza usados (explicados en el anterior capítulo). Una vez que se han elegido los algoritmos para los que se realizarán las pruebas de fusión, se pasa a analizar completamente el funcionamiento de cada algoritmo por separado. También se analiza el comportamiento de la fusión en cada secuencia, aportando siempre datos cuantitativos para evaluar objetivamente los resultados de la fusión. También se incluyen fotogramas que aportan ejemplos visuales de lo que se explica en el análisis de cada caso.

5.1. Entorno de evaluación

El entorno elegido para la experimentación es un escenario de exterior, con el montaje situado en una planta de un edificio, a 6 metros de altura sobre el suelo. Se ha preferido centrar la tesis en este tipo de entornos ya que los ambientes exteriores presentan un mayor número de cambios en las condiciones de temperatura e iluminación que un entorno interior más controlado. Estos escenarios están sujetos a cambios incluso a nivel meteorológico, ya que pueden aparecer condiciones que afectan



(a) (b)

Figura 5.1: Entorno en el que se realiza la experimentación de la tesis. (a) Imagen capturada con una cámara térmica infrarroja. (b) Imagen capturada con una cámara en color

a la visibilidad como puede ser la lluvia o la nieve.

En la Figura 5.1a y Figura 5.1b se pueden ver capturas del entorno escogido en el espectro infrarrojo y en el espectro color, respectivamente. El escenario carece de puertas por las que se pueda acceder directamente a él, y los humanos pueden aparecer tanto por la parte inferior de la imagen como por la izquierda o la derecha. También pueden aparecer por la esquina superior derecha de la misma, donde, además, hay un árbol que dificulta su detección. Así mismo, presenta elementos que pueden dar lugar a oclusiones, como árboles o una farola en la parte central, además de zonas que presentan distinta iluminación respecto a la tónica general de la escena, como las sombras de los árboles ya mencionados o la farola en caso de ser de noche. En el espectro infrarrojo, aparte de las oclusiones previamente mencionadas, se añade la dificultad adicional de que hay una placa de cemento en la zona inferior de la escena que absorbe el calor del ambiente, así como el edificio que se puede ver al fondo. Este último será especialmente problemático, ya que la cámara presenta también atenuación térmica, lo que provoca que pueda distinguir peor la temperatura de los objetos conforme se alejan de ella. Por ello, las temperaturas de los humanos se pueden confundir con la del edificio cuando estos pasan cercanos a él, debido a esta pérdida de sensibilidad. Este hecho aumenta la dificultad a la hora de aislarlos del resto de la escena.

5.1.1. Secuencias de prueba

Para evaluar nuestros algoritmos, se han probado sobre una serie de secuencias a distintas temperaturas y bajo diferentes condiciones. El principal objetivo consiste en cubrir el número máximo posible de situaciones, tanto en complejidad como en variación de la temperatura. Para ello, se decidió abarcar un rango de temperaturas propias del invierno y del verano, situadas entre los -2° y los 33°. También se ha buscado trabajar bajo diversas condiciones atmosféricas, desde nieve hasta días soleados. Además se han utilizado situaciones de complejidad variable, desde un único humano paseando

por la escena hasta tres personas reuniéndose, junto a diversas acciones que pueden llevar a cabo las personas en una escena exterior. Estas acciones abarcan desde actitudes en las que los humanos son fáciles de detectar, tales como caminar o correr, a otras con mayor dificultad debido a que cambian las proporciones del espacio que ocupan, como agacharse, sentarse o incluso tumbarse en el suelo.

A continuación se describen las diferentes secuencias grabadas. Cada una de estas 12 secuencias se denomina por la temperatura a la que se capturó, seguida de las condiciones atmosféricas en el momento de la grabación.

- Secuencia *-2°Niebla*: Esta secuencia se grabó en unas circunstancias en que la niebla cubría parcialmente la escena (especialmente el fondo de la misma) y, en general, provocaba un tono difuminado en la imagen en la cámara en color, dificultando la distinción de un humano respecto al edificio y los elementos del fondo en numerosas ocasiones, tal y como se observa en la Figura 5.2a. La cámara en infrarrojo no presenta un contraste tan alto como en otras numerosas ocasiones, si bien no resulta complicado poder apreciar al humano, salvo cuando se acerca al edificio del fondo.

  En esta secuencia, un humano pasea por todo el escenario, realizando diversas acciones como agacharse, correr o sentarse en la plataforma de cemento central. La secuencia consta de 5797 fotogramas, de los cuales el humano aparece en 5155 de los mismos en la cámara en color y en 3347 de ellos se encuentra situado en la zona cubierta por ambas cámaras.

- Secuencia *2°Nevado*: Esta secuencia se grabó tras una nevada, con lo que todo el suelo aparece cubierto de nieve. Los comportamientos dentro de la secuencia presentan una gran complejidad.

  Durante el transcurso de la misma aparecen tres humanos reunidos en numerosas ocasiones (con lo que para los algoritmos resulta difícil separarlos, ya que muchas veces se ocluyen entre ellos), tal y como se aprecia en la Figura 5.2b. Se realizan diversas actividades como correr, andar, agacharse o dejar objetos en el suelo. La secuencia está compuesta por 8623 fotogramas en los que aparecen humanos en 8608 fotogramas en el espectro visible, estando situados en 8175 de los mismos en la zona cubierta por la cámara infrarroja.

- Secuencia *3°Soleado*: La secuencia empieza con una persona paseando por el entorno y realizando movimientos como agacharse. Después entra una segunda persona adoptando también diversas trayectorias. Finalmente, las dos personas cruzan sus caminos, llegando incluso a reunirse en la escena (ver Figura 5.2c).

  La secuencia está compuesta por 2863 fotogramas, en los cuales hay humanos presentes en 2707 de ellos, 2368 de los cuales se encuentran dentro de la zona cubierta por la cámara en infrarrojo.

- Secuencia *8°Noche*: Secuencia realizada con el fin de evaluar el funcionamiento de los algoritmos en una situación nocturna. Se compone de 5480 fotogramas en los cuales aparecen humanos en la cámara situada en el espectro visible en 4796 de ellos, estando los humanos en la zona cubierta por las dos cámaras en 4110 fotogramas.

Se puede observar que la visibilidad es prácticamente nula en el caso del espectro visible, mientras que en el espectro infrarrojo es también especialmente problemática debido a que los edificios aún siguen calientes por el calor que han recibido durante el día. Por ello se pueden confundir con los humanos que pasean por delante de ellos. En la Figura 5.2d puede apreciarse un ejemplo de los problemas comentados. La estructura de las acciones que se llevan a cabo es similar a la anterior, con dos personas paseando por el entorno cruzando ocasionalmente sus trayectorias.

- Secuencia *9°Nublado*: Secuencia realizada en un día nublado, en la que dos personas siguen trayectorias aleatorias por el escenario. En la Figura 5.2e se puede observar que el cielo cubierto de ese día repercute en un menor contraste en el caso de la secuencia en color, mientras que en la equivalente en el espectro infrarrojo los humanos siguen siendo fáciles de distinguir con respecto al resto del entorno.

  1753 fotogramas componen la secuencia, en los que aparecen humanos en 1673 de ellos. En la zona cubierta, también por parte de la cámara en infrarrojo pueden hallarse humanos en 1339 fotogramas.

- Secuencia *10°Nublado*: Versión más sencilla de la secuencia anterior, con una sola persona paseando por el escenario y realizando diversas acciones como agacharse y pasear por las zonas peor iluminadas del escenario, como son las sombras de los árboles. También se halla una dificultad adicional en un cambio de iluminación que tiene lugar durante la secuencia debido al paso de nubes, lo que provoca problemas al algoritmo que usa resta de fondo.

  La Figura 5.2f muestra dos fotogramas de esta secuencia, que se compone de 2121 fotogramas, apareciendo el humano en 2087 de los mismos, de los cuales está 1848 presente también en la cámara infrarroja.

- Secuencia *15°Amanece*: Secuencia rodada al amanecer, con lo que durante la escena se registran cambios graduales de iluminación y temperatura, empezando la misma en penumbra y aumentando la iluminación conforme avanza la secuencia.

  Durante la secuencia aparecen dos personas, las cuales se reúnen y cruzan continuamente tal y como se ve en la Figura 5.3a, con lo que se registran numerosas oclusiones. Está compuesta por 3013 fotogramas, en los cuales aparecen humanos en 2998 de los mismos, en 2803 de los cuales se encuentran también en la zona cubierta por la cámara en infrarrojo.

- Secuencia *15°Nublado*: Secuencia de 2688 fotogramas con una sola persona paseando por el escenario en los que aparece en la escena durante 2622 imágenes, de las cuales en 1844 se encuentra en la zona común a ambas cámaras.

  Lo que se sucede en la secuencia la convierte en más compleja que las anteriores debido a que se realizan acciones en las que la segmentación resulta más difícil a priori. Esto se debe a que el humano cambia sus dimensiones y proporciones habituales, gracias a movimientos como sentarse durante unos segundos (ver Figura 5.3b). Es importante destacar que, por el aumento

de temperatura, la radiación emitida por el humano se empieza a reflejar ya en la plataforma de cemento situada en la parte inferior de la escena.

- Secuencia *18°Soleado*: Nueva secuencia de 2381 fotogramas con grupos de personas, en la que dos personas pasean y cruzan sus trayectorias apareciendo en 2281 de ellos (en 1764 fotogramas en la cámara infrarroja).

  Se cuenta además con la dificultad añadida de que a esta temperatura aumenta la temperatura del césped y del entorno en general, haciéndose más difícil el poder distinguir a los humanos, incluso a simple vista en los fotogramas capturados en el infrarrojo. Por su parte, en el vídeo en color puede observarse que comienzan a aparecer sombras que también dificultan el funcionamiento de los algoritmos en el espectro visible. En la Figura 5.3c podemos observar un ejemplo de los cruces comentados y de la dificultad que entraña la escasa visibilidad de los humanos en el espectro infrarrojo.

- Secuencia *23°Soleado*: Secuencia compuesta por 2494 fotogramas. Esta secuencia es mucho más compleja que las anteriores, ya que, en esta ocasión, aumenta hasta tres personas la cantidad de humanos que pasean por la escena y se reúnen en diversas ocasiones, sentándose o simplemente cruzándose. En esta ocasión aparecen personas durante 2479 fotogramas, encontrándose en la zona común a ambas cámaras durante 2300 de ellos. Nuevamente, la alta temperatura hace difícil distinguir a los humanos en el espectro infrarrojo, siendo especialmente crítica la zona situada sobre la plataforma de cemento, tal y como se aprecia en la Figura 5.3d. Por otro lado, en el espectro visible se puede apreciar que todos los objetos proyectan mayor sombra, lo que también complica los resultados de los diversos algoritmos.

- Secuencia *28°Soleado*: Secuencia similar a la anterior, aunque los humanos son aún más difíciles de distinguir en el espectro infrarrojo, debido a que la temperatura ya es muy alta y se confunde la temperatura propia de los humanos con la del entorno.

  La secuencia se compone de 3752 fotogramas y se realizó para ver el funcionamiento de los algoritmos analizados en situaciones de dificultad extrema, ya sea por las temperaturas altas en el infrarrojo o por la gran presencia de sombras en el caso del color. Los humanos aparecen en el espectro visible en 3736 fotogramas, permaneciendo en la zona común a ambas cámaras durante 3637 de los mismos. La Figura 5.3e muestra un fotograma de ejemplo de esta secuencia.

- Secuencia *33°Soleado*: Secuencia grabada con mucho calor, en la que los humanos son casi imposibles de distinguir respecto al fondo en el espectro infrarrojo y aparecen siempre más fríos que el resto de su entorno tal y como se aprecia en la Figura 5.3f.

  La secuencia consta de 5376 fotogramas, en los que aparece el humano en 4828 de los mismos, estando situado en la zona común a ambas cámaras durante 3478 fotogramas.

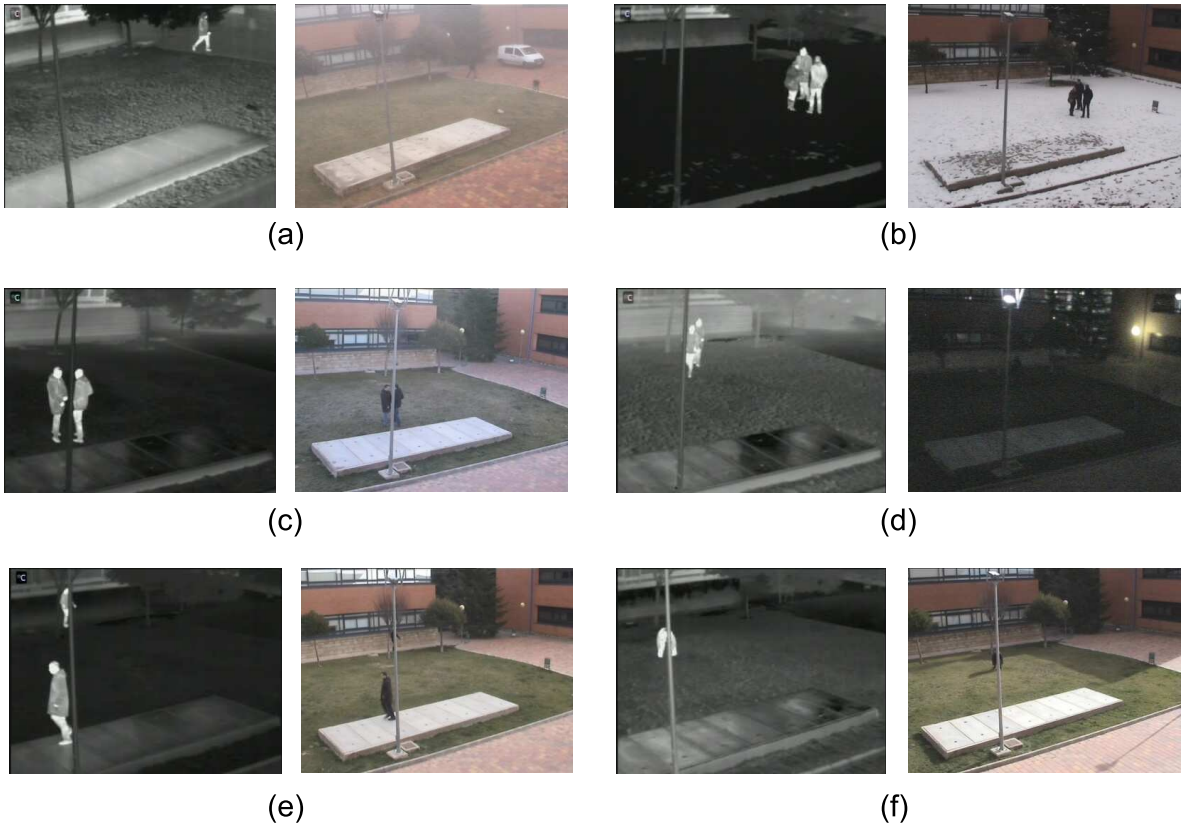


Figura 5.2: Fotogramas de ejemplo para las seis primeras secuencias utilizada en las pruebas de los algoritmos. (a) Secuencia *-2°Niebla*. (b) Secuencia *2°Nevado*. (c) Secuencia *3°Soleado*. (d) Secuencia *8°Noche*. (e) Secuencia *9°Nublado*. (f) Secuencia *10°Nublado*.

5.2. Parametrización y umbrales de confianza

Una vez establecidos el entorno y las condiciones en que se llevará a cabo la evaluación de los diversos algoritmos desarrollados e implementados, así como de la propuesta de fusión realizada, es necesario establecer los parámetros bajo los que tendrán lugar las pruebas para verificar el comportamiento de los desarrollos realizados. Puede comprobarse que estas parametrizaciones están sujetas al valor de confianza asignado a cada espectro, por lo que, una vez establecidos los valores de los distintos parámetros, se realiza un análisis de los resultados obtenidos para cada secuencia con el fin de obtener los umbrales necesarios para poder establecer la frontera entre los diversos valores de confianza $ALTA$ y $BAJA$ en el espectro visible y $ALTA$, $MEDIA$ y $BAJA$ en el espectro infrarrojo.

5.2.1. Parametrizaciones empleadas en los algoritmos

En los siguientes apartados se detallan los valores para cada parámetro de los algoritmos de detección de humanos en el espectro visible, así como los de detección de humanos en el espectro infrarrojo. Es destacable el hecho de que estos últimos, al estar basados en sucesivas ampliaciones alternativas

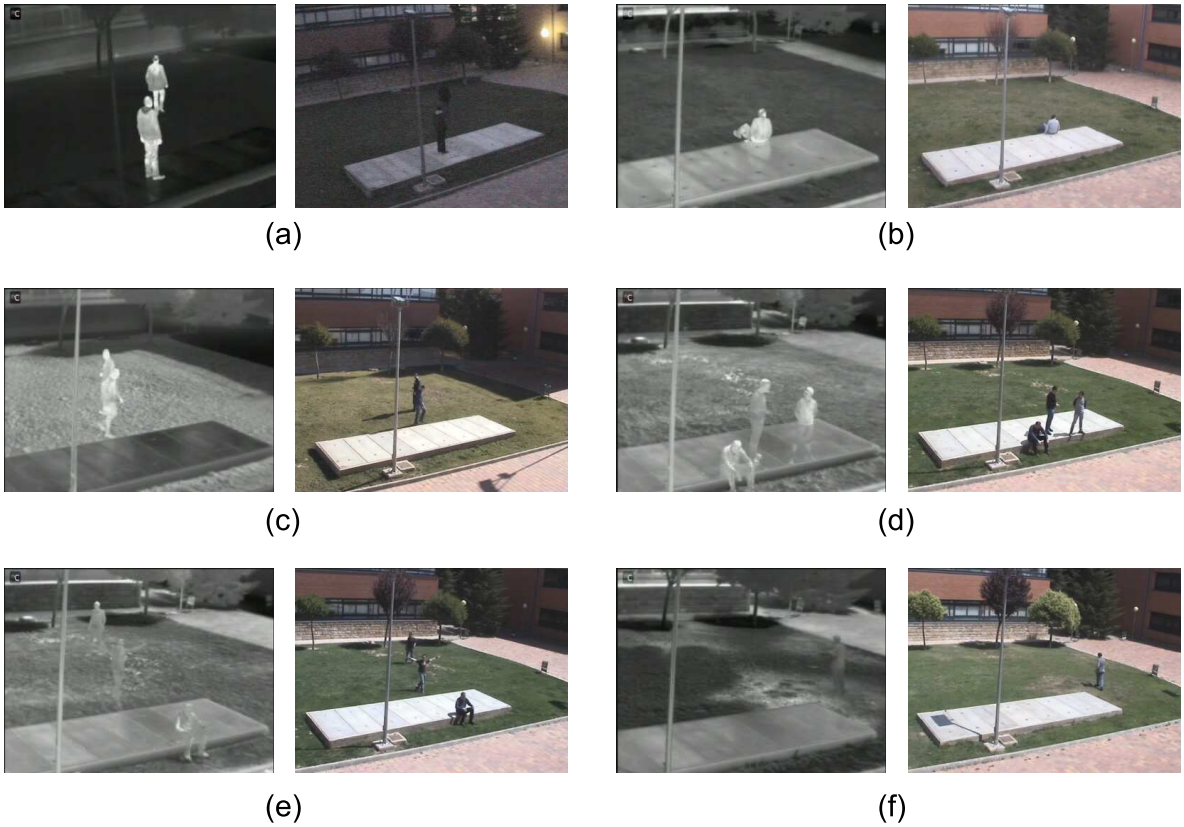


Figura 5.3: Fotogramas de ejemplo para las seis últimas secuencias utilizada en las pruebas de los algoritmos. (a) Secuencia *15°Amanece*. (b) Secuencia *15°Nublado*. (c) Secuencia *18°Soleado*. (d) Secuencia *23°Soleado*. (e) Secuencia *28°Soleado*. (f) Secuencia *33°Soleado*.

de una idea original, parten de una parametrización común.

5.2.1.1. Parámetros empleados para los algoritmos en el espectro visible

En primer lugar, se expondrán los valores de los parámetros utilizados para las diversas pruebas de los algoritmos de detección de humanos en el espectro visible. Para lograr una mayor comprensión y facilitar la lectura del presente capítulo, a partir de ahora se denominará a la detección de humanos en el espectro visible basada en la resta de fondo como *C-BS* y a la basada en computación acumulativa como *C-AC*.

En la Tabla 5.1 se observa cómo, en el caso de la confianza baja para el espectro visible, en el algoritmo *C-BS* se utiliza una apertura inicial y luego una gran cantidad de cierres, debido a que, en este caso, el mayor problema del espectro visible es más la baja sensibilidad que una precisión deficiente. Esto es así porque la cámara normalmente se encuentra cegada e interesa más obtener el mayor número de detecciones posibles, mientras que para una confianza alta un bajo número de cierres es suficiente para unir partes de humanos que puedan haber sido fragmentadas por la resta de fondo. Finalmente, puede observarse que existen una serie de parámetros dependientes de las características de la escena

Tabla 5.1: Configuración del algoritmo *C-BS* para las distintas confianzas de la detección de humanos en el espectro visible

Parámetro	Valores	
Confianza Color	$ALTA$	$BAJA$
Aperturas sobre la imagen binarizada I_S	0	1
Cierres sobre la imagen binarizada I_S	3	12
Área mínima para el humano final A_{minBS}	100	
Tasa de aprendizaje ϱ	0,0001	
Relación máxima entre la altura y la anchura del humano hwR_{max}	6	
Relación mínima entre la altura y la anchura del humano hwR_{min}	1	

Tabla 5.2: Configuración del algoritmo *C-AC* para la detección de humanos en el espectro visible

Parámetro	Valor
Número de bandas N	8
Valor de descarga v_{dis}	0
Valor de saturación v_{sat}	255
Valor de descarga media v_{dm}	32
Umbral de permanencia θ_{per}	64
Umbral de fusión espacial θ_{sp}	177
Número de erosiones n_{er}	2
Número de dilataciones n_{dil}	18
Área mínima A_{dmin}	50

o bien constantes para la tasa de fotogramas con que se ha trabajado (en este caso la velocidad de adquisición de fotogramas se encuentra experimentalmente fijada en $5 fotogramas/segundo$).

Por su parte, ya se ha explicado en la sección 4.3.2.1 que el algoritmo *C-AC* presenta una mayor invariabilidad respecto a las condiciones de la escena. Experimentalmente, se fijaron los valores de la Tabla 5.2 con 8 bandas y de forma que el historial de movimiento de un humano se actualiza cada 6 fotogramas (es decir, su movimiento más antiguo es obviado pasados 6 fotogramas del mismo), con el fin de poder representar mejor la forma obtenida del humano en las primeras etapas del algoritmo. Por otra parte, a la hora de plasmar su localización final, se conserva únicamente información de los dos últimos fotogramas, con el fin de poder tener una mayor perspectiva de dónde se encuentra situado exactamente en la escena. Finalmente, debido a que el algoritmo se encuentra más basado en el movimiento que el *C-BS*, se utiliza una serie de erosiones iniciales con el fin de eliminar pequeños falsos positivos y una gran cantidad de dilataciones con el fin de unir entre sí partes del humano que hayan hecho movimientos separados (como son las extremidades), obteniendo así una mejor idea del movimiento general del mismo.

Tabla 5.3: Configuraciones para las distintas confianzas de la detección de humanos en el espectro infrarrojo

Parámetro	Valores				
Confianza Color	ALTA			BAJA	
Confianza IR	ALTA	MEDIA	BAJA	ALTA	MEDIA
Umbral suave base γ	101	152	75	101	132
Factor ϕ sobre la desviación de la imagen	1,2	1,00	1,2	1,2	1,00
Aperturas sobre la imagen binarizada I_b	1	1	1	1	1
Cierres sobre la imagen binarizadaI_b	10	6	1	10	6
Área mínima para el humano final A_{Rmin}	100				
Relación máxima entre la altura y la anchura del humano hwR_{max}	6				
Relación mínima entre la altura y la anchura del humano hwR_{min}	1				

5.2.1.2. Parámetros empleados para los algoritmos en el espectro infrarrojo

Tal y como se ha hecho para los algoritmos en el espectro visible, con el fin de tener una mejor comprensión y distinción de los algoritmos utilizados, a partir de ahora se denominará durante el resto de este capítulo a la detección de humanos en el espectro infrarrojo basada en un único fotograma como *IR-SF*, a la basada en flujo óptico como *IR-OF* y a la basada en diferencia de fotogramas como *IR-(SF+FS)*. Debido a que los algoritmos *IR-OF* y *IR-(SF+FS)* son extensiones del algoritmo original *IR-SF*, comparten por todos sus valores paramétricos, debido a que a la hora de establecer los valores de tolerancia para la incorporación de movimiento se han utilizado constantes comunes tal y como se describió en las secciones 4.3.1.3 y 4.3.1.2.

Los niveles de confianza establecidos se corresponden con el número de configuraciones utilizadas para este espectro y se pueden ver en la Tabla 5.3. Puede observarse que se utilizan conjuntos de parámetros más restrictivos para el contraste a nivel medio (ya que es el más dado a falsos positivos que pueden resultar en sobreparticionamiento de humanos en la fusión) y con umbrales más relajados en nivel alto (ya que se tiende a detectar un mayor número de humanos en estas condiciones sin una bajada considerable de precisión) y nivel bajo (ya que las escasas detecciones en este nivel de confianza pueden ayudar a la parte de seguimiento dentro del algoritmo de fusión, reforzando la sensibilidad de los resultados). Por otro lado, tal y como se ha explicado anteriormente, se puede observar que aquí también existen parámetros independientes de la confianza del espectro, pero directamente dependientes de las condiciones de las capturas (altura de la cámara, ángulo que abarca de la escena), como son las proporciones de los humanos o que sus áreas permanecen invariables para todas las posibles confianzas.

5.2.2. Establecimiento de los umbrales de confianza

A continuación se analizan los resultados de las diversas secuencias con el fin de fijar experimentalmente los umbrales que separan los valores de confianza en cada espectro. Recordemos que en el espectro visible estos valores dependen del nivel medio de gris de la imagen, mientras que en el infrarrojo dependen del coeficiente υ_{IR} entre la media y la desviación típica del fotograma. También es

Tabla 5.4: Establecimiento de los umbrales de confianza para la detección de humanos en el espectro visible

Secuencia	Media	Sensibilidad	Precisión	*F-score*	Confianza
-2°Niebla	147	0,52	1,00	0,68	
8°Noche	49	0,08	1,00	0,15	$BAJA$
15°Amanece	59	0,23	1,00	0,37	
2°Nevado	138	1,00	0,92	0,96	
3°Soleado	133	1,00	0,98	0,99	
9°Nublado	123	1,00	0,80	0,89	
10°Nublado	126	0,95	0,96	0,96	
15°Nublado	125	1,00	0,97	0,98	$ALTA$
18°Soleado	107	0,86	0,98	0,91	
23°Soleado	116	1,00	0,99	1,00	
28°Soleado	115	1,00	0,99	1,00	
33°Soleado	125	1,00	0,99	0,99	

importante volver a hacer hincapié en el hecho de que la confianza asignada se actualizará dinámicamente con el fin de adaptarse a las condiciones cambiantes de cada escena, es decir, si la temperatura sube o baja, o bien alguien enciende o apaga un foco, la confianza del espectro correspondiente se actualizará en base a dicho suceso.

5.2.2.1. Umbrales de confianza para la detección de humanos en el espectro visible

Tal y como se explicó en el capítulo anterior, a la hora de establecer los umbrales de confianza en el espectro visible, se atendió únicamente a la media de las imágenes en nivel de gris, ya que en el espectro visible no es posible distinguir a los humanos en base a su contraste con el tono general de la escena. Por ello la desviación típica no se hace necesaria en este caso. Sin embargo, sí que se puede concluir que, generalmente, una escena con un nivel medio de gris intermedio estará mejor iluminada y, por tanto, será más fácil distinguir los humanos que habrá en la misma, ya que además habrá menores zonas oscuras. Por otra parte, una secuencia con un nivel de gris excepcionalmente alto será indicativo de que la escena presenta anomalías, como la presencia de niebla o de un foco iluminando directamente la escena. Hay que destacar que en este caso, para las secuencias, se escogieron subsecuencias donde solo estuviera presente un humano con el fin de que los resultados fueran independientes de la presencia de grupos ya que, tal y como hemos explicado previamente, estos no pueden ser separados por la detección de humanos en el espectro visible y, por tanto, no deben influir a la hora de valorar la fiabilidad de este espectro.

Monitorizando las diversas secuencias usadas, se obtuvo la Tabla 5.4, usando las configuraciones de parámetros vistas en la Tabla 5.1 para cada valor de confianza. Puesto que ya se ha establecido que el algoritmo *C-AC* tiene mayor independencia respecto a la iluminación de la escena, se usarán los resultados del algoritmo *C-BS* para obtener una mayor perspectiva de cómo afecta este valor a las condiciones de la segmentación en el espectro visible.

En este caso, observando los resultados, se ve una marcada diferencia entre las condiciones de iluminación adversas, como son la presencia de una niebla que cubre parcialmente la escena (y donde cuesta distinguir a los humanos porque estos aparecen parcial o totalmente ocultos por la niebla) y una escena nocturna o al amanecer, en las que hay zonas de la escena pobremente iluminadas en las que no es posible distinguir a los humanos. En el caso de las condiciones de iluminación cuyos fotogramas presentan un nivel medio de gris entre 95 y 135, se puede observar que los resultados son generalmente excelentes con valores de *F-score* siempre cercanos o muy superiores al $90\,\%$ y sensibilidades casi siempre en torno al $100\,\%$ de detecciones de humanos presentes en la escena. Estos valores son muy bajos para secuencias con fotogramas cuyo valor medio de nivel de gris se encuentra fuera del rango mencionado. Por tanto, se fijará experimentalmente como umbral ζ_{VL} un valor medio de nivel de gris de 95 y como umbral ζ_{VH} una media de 135.

5.2.2.2. Umbrales de confianza para la detección de humanos en infrarrojo

Ya se ha comentado en el Capítulo 4 que, a la hora de establecer umbrales de confianza en el espectro infrarrojo, se tendría en cuenta el coeficiente υ_{IR} entre la media y la desviación típica de cada imagen. Recordamos que un valor alto de υ_{IR} significa que la imagen tiene una media muy alta en relación con su desviación. Es decir, que hay pocos pixels cuyo valor se aleje de la media de la imagen. Puesto que la forma de distinguir a los humanos en el espectro infrarrojo es por cómo estos destacan respecto del fondo, ante este caso será muy difícil distinguirlos. Para establecer en qué punto hay que fijar los umbrales para poder delimitar estos casos, se analizaron los datos de todas las secuencias con el algoritmo *IR-(SF+FS)* debido a que añade información de movimiento a la detección, sin dejar de tener en cuenta las características térmicas de la escena. Así mismo, los parámetros usados son los mostrados en la Tabla 5.3. A la hora del análisis se han tenido en cuenta sus resultados en cuanto a sensibilidad, precisión y *F-score* en relación con el coeficiente comentado.

Ya se explicó también con anterioridad que la confianza de la detección de humanos en el espectro infrarrojo se basa en aquella asignada a la segmentación en el espectro visible, por lo que, para establecer estos valores, se han dividido en dos tablas, atendiendo a si corresponden a la confianza alta o baja de este espectro. Los resultados pueden apreciarse en la Tabla 5.5 y la Tabla 5.6, respectivamente.

En el caso de los resultados mostrados en la Tabla 5.6, para una confianza $BAJA$ en el espectro visible, se puede observar una división muy pronunciada entre las secuencias grabadas de niebla o al amanecer con la secuencia grabada de noche, siendo esta última secuencia la única con un *F-score* muy inferior al $90\,\%$. Puede apreciarse que la desviación típica es muy pequeña en comparación a la media, lo que determinará que los objetos son difíciles de distinguir en muchos casos respecto al fondo. Por tanto, el valor del umbral ζ_{IRN} se ha fijado en 2 en base a las pruebas realizadas.

Por su parte, debido a las parametrizaciones explicadas previamente para una confianza $ALTA$ en la detección de humanos en el espectro visible, se puede observar un menor salto de resultados entre las diversas secuencias probadas, salvo a partir de los 20° de temperatura, cuando los resultados empeoran drásticamente. Por tanto, al fijar experimentalmente como umbral ζ_{IRH} el valor $2{,}2$, a partir

Tabla 5.5: Establecimiento de los umbrales de confianza para la detección de humanos en el espectro infrarrojo para confianza $ALTA$ de la segmentación en el espectro visible

Secuencia	Media	Desviación	υ_{IR}	Sensibilidad	Precisión	*F-score*	Confianza
2°Nevado	53	36	1,47	0,71	1,00	0,83	$ALTA$
3°Soleado	62	52	1,19	0,98	0,91	0,94	
9°Nublado	49	28	1,73	0,95	0,98	0,97	
10°Nublado	68	32	2,09	0,99	0,99	0,99	$MEDIA$
15°Nublado	86	44	1,96	0,91	0,97	0,94	
18°Soleado	109	51	2,12	0,93	0,96	0,95	
33°Soleado	80	40	1,99	0,07	1,00	0,12	
23°Soleado	113	46	2,46	0,60	0,86	0,70	$BAJA$
28°Soleado	107	43	2,5	0,39	0,96	0,55	

Tabla 5.6: Establecimiento de los umbrales de confianza para la detección de humanos en el espectro infrarrojo para confianza $BAJA$ de la segmentación en el espectro visible

Secuencia	Media	Desviación	υ_{IR}	Sensibilidad	Precisión	*F-score*	Confianza
-2°Niebla	50	26	1,93	0,89	0,94	0,91	$ALTA$
15°Amanece	35	26	1,35	0,93	1,00	0,96	
8°Noche	93	42	2,22	0,81	0,86	0,83	$MEDIA$

de estas pruebas, se observa que esta frontera constituye una alternativa totalmente válida.

Así mismo, se ha probado experimentalmente que a partir de los 10° de temperatura el no exigir una configuración más restrictiva de los parámetros de los algoritmos en el espectro infrarrojo conlleva una bajada drástica de precisión.Puede observarse que el cociente entre la media y la desviación se encuentra siempre por encima de $1,9$ en estos casos, por lo que tomaremos este valor como nuestro umbral ζ_{IRL}. Tal y como pudo verse en la sección 4.4.2, este hecho se tuvo muy presente a la hora de elaborar las diversas reglas.

Finalmente, destacaremos el hecho que, en el caso de la secuencia grabada a 33°, la cámara infrarroja se encuentra prácticamente cegada, aunque caiga fuera de los valores previamente explicados. Esto se debe a que, en este caso, los humanos aparecen más oscuros que el resto del entorno, siendo casi imposibles de localizar con el enfoque tomado actualmente en los algoritmos y puede constituir un caso interesante para explorar en el futuro.

5.3. Resultados en segmentación de humanos

Con los parámetros y umbrales explicados en la sección 5.2 de este capítulo, se llevó a cabo un análisis de los resultados de los diversos algoritmos desarrollados con las secuencias y entorno detallados en la sección 5.1. El principal objetivo de estas pruebas iniciales es valorar qué algoritmos del espectro infrarrojo y visible son los más apropiados en este entorno, de cara a realizar posteriormente

una fusión entre los mismos. Por tanto, para cada secuencia se comparan sus métricas de sensibilidad, precisión y *F-score*, de las que se explicó su forma de calcularlas en el apartado 3.3. Una vez se tengan estos resultados, se plasmarán en dos tablas para poder comparar los diversos algoritmos entre sí y se escogerá el que mejor funcione en este entorno en particular. Es importante destacar que, de cara a otros entornos, pudiera ser que funcionasen mejor otros algoritmos de segmentación, aunque el algoritmo de fusión se ha desarrollado de forma que sea totalmente independiente de los algoritmos de detección utilizados como entrada a la misma.

5.3.1. Segmentación de humanos en el espectro infrarrojo

Para probar nuestros algoritmos en infrarrojo (explicados en el capítulo 4 en el apartado 4.3.1), se utilizaron las secuencias descritas en la sección 5.1. Los resultados obtenidos se muestran en la Tabla 5.7.

La primera conclusión general que se puede extraer es que, en general, el espectro infrarrojo es apropiado para detectar humanos en secuencias grabadas a temperaturas medias y bajas. Véase que la secuencia capturada a 8° presenta peores resultados, ya que fue grabada en las primeras horas de la noche y la temperatura todavía no había descendido. Sin embargo, cuando la temperatura de la escena se alza por encima de los 20°, los resultados empeoran drásticamente. Esto se debe a que la radiación térmica de los humanos es muy similar a la temperatura de los edificios, ya que el sol calienta la escena directamente afectando a los elementos de la misma. Este hecho tiene un impacto significativo en la última secuencia, en la que los humanos se encuentran totalmente "unificados" con el entorno y su distinción es prácticamente imposible, incluso para un observador humano que esté supervisando los fotogramas capturados en infrarrojo.

Por último, se compararon los tres algoritmos realizados en el espectro infrarrojo. Si bien los resultados muchas veces presentan similitudes entre los tres algoritmos, los algoritmos *IR-OF* y *IR-(SF+FS)* siempre funcionan mejor que la que únicamente utiliza la información de un solo fotograma. Es importante destacar que esto ocurre porque las dos primeras son versiones ampliadas de la última, añadiendo información nueva con el fin de mejorar la cantidad de detecciones realizadas sin descartar ningún humano encontrado por la detección de humanos basada en un solo fotograma. Este hecho se hace especialmente relevante a altas temperaturas (cuando el contraste no es muy alto) o en aquellas en las que los humanos permanecen mucho tiempo cerca del fondo. En este último caso tenemos que la temperatura de los humanos hace más difícil distinguir sus límites de la pared del fondo, lo que para la propuesta *IR-SF* muchas veces acaba como un falso negativo debido a que las detecciones hechas en esa zona se descartan por sus dimensiones o proporciones. Sin embargo, la información de movimiento hace que las propuestas *IR-OF* y *IR-(SF+FS)* sean capaces de distinguir correctamente al humano. Esta mejora es debida a que no sólo utilizan esta información, sino que además utilizan para confirmar las detecciones realizadas los píxeles destacados respecto al entorno que la propuesta basada en un solo fotograma pudo aislar solo parcialmente. Como inconveniente, la precisión es mejor en la detección de humanos basada en un único fotograma, ya que al tener menor número de detecciones también presenta una cantidad menor de falsos positivos.

Tabla 5.7: Comparación de la detección de humanos en infrarrojo basada en único fotograma, en flujo óptico y en resta de fotogramas

Secuencia	Algoritmo	VP	FP	FN	Sensibilidad	Precisión	*F-score*
2°Nevado	*IR-SF*	11588	44	5144	0,69	1,00	0,82
	IR-OF	11649	141	5063	0,70	0,99	0,82
	IR-(SF+FS)	**11928**	**147**	**4784**	**0,71**	**0,99**	**0,83**
-2°Niebla	*IR-SF*	2939	186	441	0,87	0,94	0,90
	IR-OF	3226	1270	144	0,95	0,63	0,76
	IR-(SF+FS)	**3224**	**163**	**156**	**0,95**	**0,95**	**0,95**
3°Soleado	*IR-SF*	2703	287	243	0,92	0,90	0,91
	IR-OF	2781	399	165	0,94	0,88	0,91
	IR-(SF+FS)	**2902**	**295**	**44**	**0,98**	**0,91**	**0,94**
8°Noche	*IR-SF*	4273	684	1626	0,72	0,86	0,78
	IR-OF	4261	644	1638	0,73	0,88	0,79
	IR-(SF+FS)	**4787**	**766**	**1112**	**0,81**	**0,86**	**0,83**
9°Nublado	*IR-SF*	1599	61	124	0,93	0,96	0,95
	IR-OF	1617	62	106	0,94	0,96	0,95
	IR-(SF+FS)	**1618**	**61**	**105**	**0,94**	**0,96**	**0,95**
10°Nublado	*IR-SF*	1712	127	137	0,93	0,96	0,94
	IR-OF	1713	85	136	0,93	0,95	0,93
	IR-(SF+FS)	**1827**	**12**	**22**	**0,99**	**0,99**	**0,99**
15°Amanece	*IR-SF*	3912	9	338	0,92	1,00	0,96
	IR-OF	3715	779	535	0,88	0,93	0,90
	IR-(SF+FS)	**3957**	**12**	**293**	**0,93**	**1,00**	**0,96**
15°Nublado	*IR-SF*	1321	21	523	0,72	0,98	0,83
	IR-OF	1647	203	197	0,89	0,89	0,89
	IR-(SF+FS)	**1684**	**51**	**160**	**0,91**	**0,97**	**0,94**
18°Soleado	*IR-SF*	2166	60	195	0,92	0,97	0,94
	IR-OF	2127	16	234	0,90	0,99	0,94
	IR-(SF+FS)	**2185**	**19**	**176**	**0,93**	**0,99**	**0,96**
23°Soleado	*IR-SF*	927	136	2695	0,26	0,87	0,40
	IR-OF	952	131	2670	0,31	0,92	0,47
	IR-(SF+FS)	**2174**	**363**	**1448**	**0,60**	**0,86**	**0,71**
28°Soleado	*IR-SF*	1404	121	6534	0,18	0,96	0,30
	IR-OF	1463	140	6475	0,18	0,92	0,30
	IR-(SF+FS)	**3077**	**160**	**4861**	**0,39**	**0,96**	**0,55**
33°Soleado	*IR-SF*	108	23	3408	0,03	0,82	0,06
	IR-OF	48	48	3468	0,01	0,50	0,03
	IR-(SF+FS)	**123**	**23**	**3393**	**0,03**	**0,84**	**0,04**

A continuación, analizaremos las tres propuestas por separado.

El algoritmo *IR-SF* normalmente presenta menor sensibilidad que los otras dos, puesto que solo

usa información del fotograma actual que se está procesando. Esto hace que la detección sea especialmente difícil cuando la escena tiene un contraste bajo y cuesta distinguir a los humanos del fondo. Es destacable que los resultados son mejores cuando el día está nublado, debido a que el sol ha calentado el entorno menos que en un día soleado. La sensibilidad desciende drásticamente a temperaturas cálidas, con una cantidad muy pequeña de detecciones realizadas a la temperatura más alta que se ha monitorizado.

La propuesta *IR-OF* normalmente presenta una mejora pequeña sobre la propuesta previa, debido a que añade información del movimiento de la cámara y, sin embargo, nuestra cámara está situada en una posición fija en la que el viento apenas la afecta. Aun así, muestra algo de mejora puesto que el flujo óptico detecta el movimiento de los objetos en la escena, aunque con mejor sensibilidad que en el siguiente algoritmo que se evaluará. También puede observarse que la precisión sufre un pequeño descenso ya que el movimiento de los objetos incluye también el de elementos no humanos como las hojas de los árboles.

Finalmente, la técnica *IR-(SF+FS)* normalmente una mayor mejora, al añadir directamente información obtenida a partir de la resta de imágenes entre el fotograma actual y el anterior. Además, se pueden detectar movimientos más sutiles de los objetos presentes en la escena, combinándose estas detecciones con aquellas realizadas durante por la propuesta *IR-SF*. Estas detecciones añadidas pueden llegar a su vez a provocar la aparición de un mayor número de falsos positivos, debidos a la detección ocasional de pequeños reflejos de la radiación térmica o los movimientos de partes del cuerpo como los brazos o las piernas.

Tras el análisis realizado previamente, se ha decidido que el algoritmo utilizado en el espectro infrarrojo será la **detección de humanos basada en resta de fotogramas** (*IR-(SF+FS)*), ya que su sensibilidad muestra una mejora considerable al usar la información del movimiento en la escena sin que se produzca un empeoramiento considerable de la precisión alcanzada.

5.3.2. Segmentación de humanos en el espectro visible

Para realizar las pruebas, se usaron las mismas secuencias que en el espectro infrarrojo y descritas en la sección 5.1. Sin embargo, al abarcar la cámara en color un mayor espacio del escenario, el número total de humanos y situaciones se vio incrementado.

En esta ocasión, en vez de comparar dos algoritmos de los que uno es una ampliación del otro, tenemos dos aproximaciones totalmente distintas, explicadas detalladamente en el apartado 4.3.2. Por un lado, tenemos la detección de humanos basada en la resta de fondo que se basa en la existencia de un fondo relativamente estable. Tiene mayor independencia del movimiento de los objetos en la escena, ya que presenta un periodo de absorción de los mismos al fondo. Sin embargo, ante fondos poco uniformes o con numerosos cambios de iluminación la resta de fondo es muy desaconsejable, ya que presentará numerosos falsos positivos. Por otra parte, la detección de humanos basada en computación acumulativa presenta mayor independencia respecto a las condiciones globales de la escena, al tener un enfoque basado principalmente en el movimiento. Sin embargo, al crearse “estelas”

con el historial de movimiento del objeto, es recomendable para escenarios donde no haya numerosos cruces y, especialmente, donde no se produzcan numerosos movimientos ni alejándose ni acercándose a la cámara, ya que en el eje Z las posiciones que ocupa un humano en los distintos fotogramas se solapan entre sí, lo que provoca que el algoritmo pierda efectividad.

Los resultados obtenidos se muestran en la Tabla 5.8. Se puede observar que en todos los casos el número de detecciones es mayor en la resta de fondo que en la computación acumulativa. Sin embargo, también podemos apreciar que en los casos donde el fondo está más sujeto a cambios de iluminación o sombras, la resta de fondo provoca un número sensiblemente superior de falsos positivos (salvo en la secuencia de 10° debido al viento que provoca el movimiento de las hojas de los árboles, apreciable por la computación acumulativa pero demasiado sutil para ser detectado por la resta de fondo). También vemos cómo en las secuencias de 18° y 28° aparece un mayor número de falsos positivos en el algoritmo de la resta de fondo debido a que algunas sombras no han podido ser eliminadas por el algoritmo y son detectadas como un humano aparte. Otra observación que se puede extraer es la baja sensibilidad de la computación acumulativa en escenas con poco movimiento (3°) o con movimientos en el eje Z y numerosas reuniones de humanos (23° y 28°). Finalmente, también podemos destacar que, tal y como se predijo, los resultados en la escena nocturna son bastante desfavorables, debido a que estos algoritmos se basan en el espectro visible.

Podemos concluir que, dado que las pruebas se desarrollarán en un entorno con condiciones poco variables y donde puede tener lugar una gran variedad de movimientos en todos los ejes, se trabajará con la **detección de humanos basada en la resta de fondo**.

5.4. Resultados en fusión y seguimiento

Una vez que se han elegido los algoritmos de segmentación más apropiados, es el momento de evaluar si se experimenta alguna mejora al aplicar mecanismos de fusión, especificando los puntos fuertes y débiles de cada algoritmo en cada secuencia, así como comprobando si la fusión aporta algo a la hora de potenciar las virtudes de cada espectro o atenuar sus defectos en cada conjunto de condiciones probadas. Puesto que las dos cámaras no cubren exactamente el mismo área, se ha distinguido entre la "zona común" (correspondiente a la zona que cubren tanto la cámara en infrarrojo como la cámara color) y la imagen completa (cubierta solamente por la cámara color), ya que puede resultar interesante en general el poder tener una perspectiva más global de lo que está ocurriendo en toda la escena.

Para cada secuencia se analizarán los resultados del algoritmo *IR-(SF+FS)*, los resultados del algoritmo *C-BS* tanto en la zona común como en la imagen total y, finalmente, los resultados de la fusión teniendo en cuenta ambas zonas. Así mismo, estos análisis se verán acompañados de imágenes, ejemplificando cada particularidad descrita con el fin de proporcionar un complemento visual a los datos aportados, facilitando de este modo su comprensión.

Tabla 5.8: Comparación de la detección de humanos en color basada en computación acumulativa y resta de fondo

Secuencia	Algoritmo	VP	FP	FN	Sensibilidad	Precisión	*F-score*
-2°Niebla	***C-BS***	**2045**	**14**	**3229**	**0,39**	**0,99**	**0,56**
	C-AC	1133	24	4441	0,21	0,98	0,35
2°Nevado	***C-BS***	**15707**	**142**	**8538**	**0,65**	**0,99**	**0,78**
	C-AC	17327	586	6918	0,72	0,97	0,82
3°Soleado	***C-BS***	**4549**	**187**	**318**	**0,93**	**0,96**	**0,95**
	C-AC	3211	229	1656	0,66	0,93	0,77
8°Noche	***C-BS***	**1734**	**474**	**5587**	**0,24**	**0,79**	**0,36**
	C-AC	969	147	6342	0,13	0,87	0,23
9°Nublado	***C-BS***	**2391**	**370**	**4**	**1,00**	**0,87**	**0,93**
	C-AC	1991	128	404	0,83	0,94	0,88
10°Nublado	***C-BS***	**2186**	**403**	**155**	**0,93**	**0,84**	**0,89**
	C-AC	1775	323	566	0,75	0,84	0,79
15°Amanece	***C-BS***	**1043**	**0**	**3869**	**0,21**	**1,00**	**0,35**
	C-AC	1123	24	3789	0,23	0,98	0,35
15°Nublado	***C-BS***	**2704**	**54**	**88**	**0,97**	**0,98**	**0,97**
	C-AC	1509	27	1283	0,54	0,98	0,70
18°Soleado	***C-BS***	**2836**	**329**	**640**	**0,81**	**0,84**	**0,83**
	C-AC	2584	382	892	0,73	0,87	0,80
23°Soleado	***C-BS***	**6982**	**87**	**2475**	**0,74**	**0,99**	**0,85**
	C-AC	4948	215	4509	0,52	0,96	0,68
28°Soleado	***C-BS***	**8536**	**277**	**2009**	**0,81**	**0,97**	**0,88**
	C-AC	5481	120	5064	0,52	0,98	0,68
33°Soleado	***C-BS***	**4747**	**507**	**67**	**0,99**	**0,9**	**0,94**
	C-AC	2339	63	2475	0,49	0,97	0,65

5.4.1. Secuencia -2°Niebla

Viendo en la Tabla 5.9 los resultados de esta secuencia grabada con niebla, se puede observar cómo el algoritmo *C-BS* sufre una baja sensibilidad debido a que esta condición atmosférica dificulta la distinción de humanos cuando estos se encuentran cercanos al edificio del fondo de la escena (ya que en esa zona es donde aparece mayor concentración de niebla), tal y como se observa en la Figura 5.4a. Por otro lado, aquellos humanos que se encuentran cercanos a la cámara no siempre son fáciles de detectar con claridad, ya que la niebla provoca que, en general, los colores de la escena aparezcan más atenuados y sea más difícil poder detectar a un humano aunque se elabore un modelo de fondo. Estas dificultades son ya predecibles por las características de iluminación generales de los fotogramas que la componen en el espectro visible, lo que provoca que su confianza sea asignada como $BAJA$. Por su parte, la detección de humanos en el espectro infrarrojo no presenta apenas problemas (como se aprecia en la Figura 5.4a) debido a que el contraste de los fotogramas es bastante

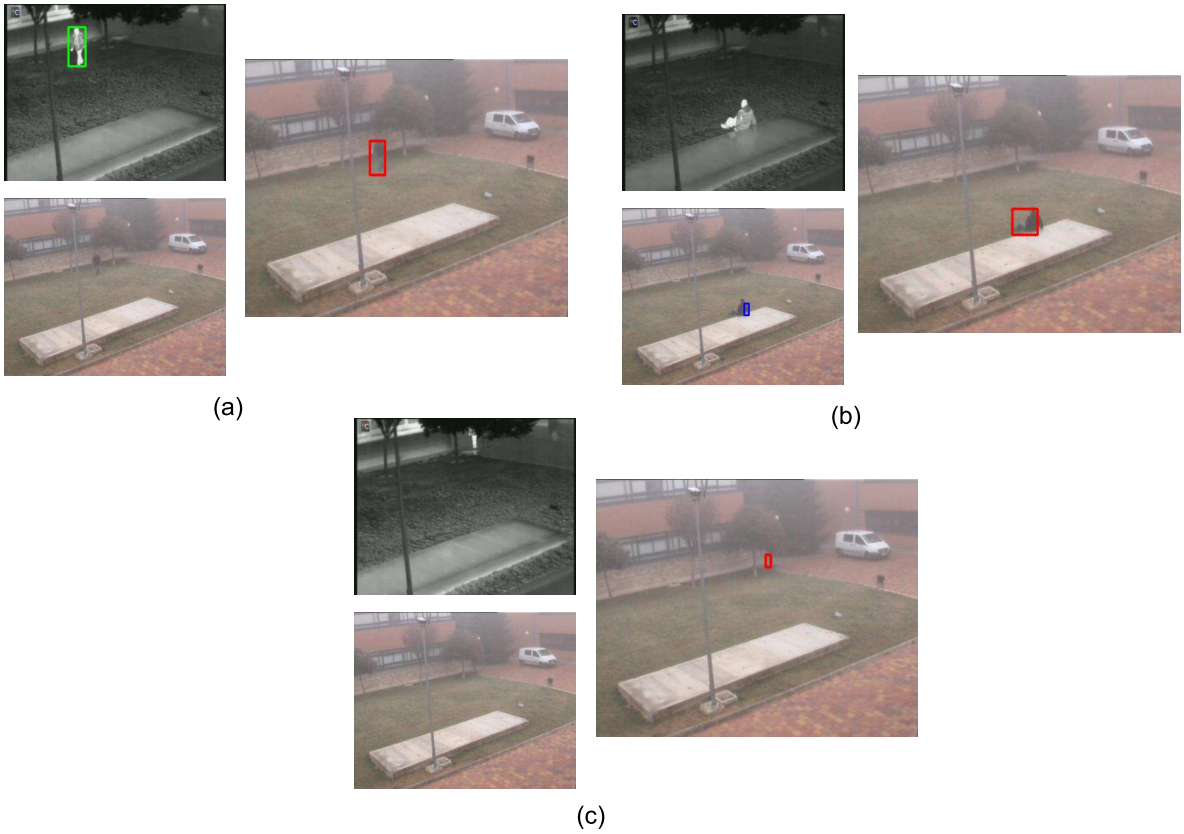


Figura 5.4: Ejemplos de resultados obtenidos en la secuencia -2°Niebla. (a) Falso negativo de la detección de humanos en el espectro visible corregido por la segmentación en infrarrojo. (b) Falso negativo de la detección de humanos en el espectro infrarrojo corregido por la segmentación en en el espectro visible y el algoritmo de fusión y seguimiento. (c) Falso negativo de las detecciones de humanos en en ambos espectros corregido por el algoritmo de fusión y seguimiento.

alto (lo que resulta en su confianza asignada como $ALTA$). Únicamente aparecen falsos negativos en momentos puntuales como el observado en la Figura 5.4b, donde la temperatura del abrigo que viste la persona coincide con la de la plataforma central. Finalmente, se puede observar que el uso de la fusión aumenta la sensibilidad con respecto a las dos detecciones de humanos utilizadas ya que, en los escasos falsos negativos del algoritmo *IR-(SF+FS)*, el seguimiento busca entonces humanos detectados en el espectro visible situados en esas zonas(obteniendo resultados como el observado en la Figura 5.4b), o directamente se estima que, dada su posición, esos humanos pueden permanecer todavía en la escena, tal y como se aprecia en la Figura 5.4c.

5.4.2. Secuencia 2°Nevado

Esta secuencia, tal y como se ha comentado antes, resulta una de las más difíciles debido a la gran cantidad de situaciones de alta complejidad que aparecen. Ambos espectros parten de condiciones lumínicas y de contraste estables, encontrándose los dos con condiciones de confianza $ALTA$,

Tabla 5.9: Resultados alcanzados en la secuencia -2°Niebla

	Sensibilidad	**Precisión**	***F-score***
IR-(SF+FS)	0,95	0,95	0,95
C-BS (Zona común)	0,38	0,97	0,54
C-BS (Imagen completa)	0,39	0,99	0,56
Fusión (Zona común)	0,96	0,95	0,96
Fusión (Imagen completa)	0,75	0,96	0,84

pudiéndose apreciar en la Tabla 5.10 que la precisión alcanzada es muy alta en los dos espectros. En la Figura 5.5a se puede observar que, cuando un humano camina solo por la escena, las condiciones de contraste e iluminación son bastante buenas, incluso óptimas en el caso del espectro infrarrojo.

Sin embargo, es inevitable que llame la atención que la sensibilidad alcanzada tanto por ambos espectros individualmente como utilizando la fusión no resulta elevada, encontrándose en valores nunca superiores al $75\,\%$ de detecciones. Este hecho se debe a que en la escena aparecen una gran cantidad de grupos en los que muchas veces los humanos se ocluyen entre sí. En ocasiones se hace difícil el apreciar que hay más de una persona en determinadas zonas, incluso para un observador humano, tal y como puede verse en la Figura 5.5b. Aun así, sí que se aprecia la habilidad de la segmentación en infrarrojo para dividir grupos, tal como que ya se ha comentado en varias ocasiones. Por ejemplo, en la Figura 5.5c, vemos cómo se ha dividido a dos personas que el algoritmo *C-BS* era incapaz de separar. En esta secuencia, esta característica resulta especialmente interesante tal y como puede apreciarse en las estadísticas, en las que aumenta un $8\,\%$ en la zona común a ambas cámaras la sensibilidad del espectro que peor funciona (el visible) y un $4\,\%$ la del que mejor (el infrarrojo), sin afectar a la precisión de forma significativa. Esto repercute en que también el *F-score* se incrementa en un $2\,\%$.

Estas estadísticas también nos demuestran cómo los dos espectros son capaces de aportar información útil en el algoritmo de fusión elaborado, ya que ambos llegan a presentar falsos negativos puntuales, ya sea por la incapacidad ya mencionada de separar grupos en la detección de humanos en color o porque sus temperaturas se confunden con las de otros elementos en la escena (como puede apreciarse en la Figura 5.5d).

Por último, echando un vistazo a las estadísticas en todo el escenario, se aprecia una tendencia similar, a pesar de que las sensibilidades del algoritmo *C-BS* y la fusión aparecen menores. Esto es debido a que hay más humanos en total en el escenario y pueden producirse falsos negativos fuera de la zona común. Aun así, la aportación positiva de la fusión sigue apreciándose, de forma que siempre las estadísticas del espectro visible se ven mejoradas al ser apoyado por la información del espectro infrarrojo.

Tabla 5.10: Resultados alcanzados en la secuencia 2°Nevado

	Sensibilidad	**Precisión**	***F-score***
IR-(SF+FS)	0,71	1,00	0,83
C-BS (Zona común)	0,65	0,97	0,78
C-BS (Imagen completa)	0,65	0,99	0,78
Fusión (Zona común)	0,73	0,99	0,84
Fusión (Imagen completa)	0,68	0,98	0,81

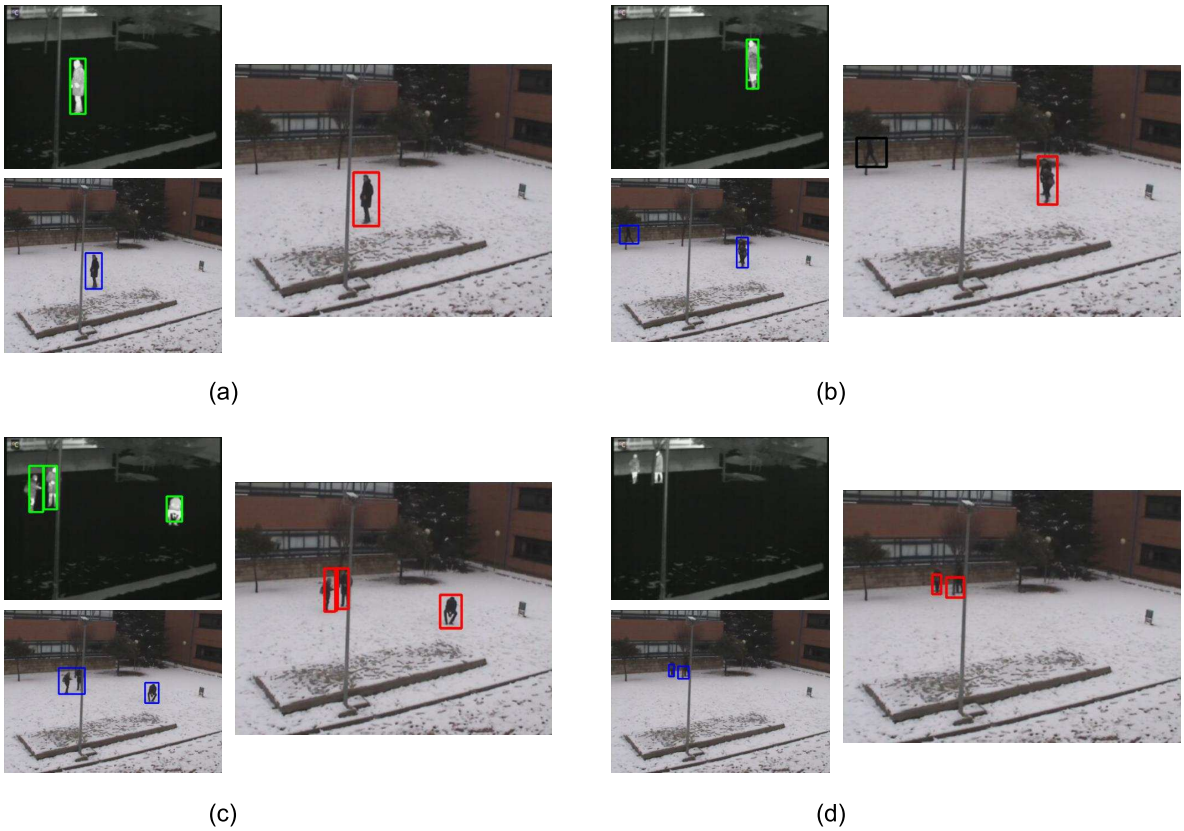


Figura 5.5: Ejemplos de resultados obtenidos en la secuencia 2°Nevado. (a) Condiciones óptimas para la segmentación, (b)Grupo difícil de distinguir a simple vista, (c) Grupo correctamente separado por la segmentación en infrarrojo respecto a la detección de humanos en color, (d) Falso negativo de la segmentación en infrarrojo corregido por la segmentación en color y reflejado en la fusión,

5.4.3. Secuencia 3°Soleado

En esta secuencia, las detecciones de los dos espectros presentan buenos resultados. Debido a las condiciones de contraste e iluminación, la confianza asignada a cada uno es $ALTA$, con ambas segmentaciones detectando generalmente a los humanos sin problemas, tal y como se aprecia en las estadísticas mostradas en la Tabla 5.11 y en la Figura 5.6a. Aunque la sensibilidad de la detección de humanos en el espectro visible sea un $6\,\%$ mejor que la realizada en el espectro infrarrojo, se

Tabla 5.11: Resultados alcanzados en la secuencia 3°Soleado

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,98	0,91	0,94
C-BS (Zona común)	0,98	0,96	0,97
C-BS (Imagen completa)	0,98	0,97	0,98
Fusión (Zona común)	1,00	0,90	0,94
Fusión (Imagen completa)	0,99	0,91	0,95

puede apreciar que ésta sirve de apoyo tanto en situaciones en el humano se encuentra cubierto por sombras en el espectro visible (como en la Figura 5.6b), como para dividir grupos como se observa en la Figura 5.6c. La detección de humanos en infrarrojo en ocasiones arroja falsos negativos debido a que el humano se encuentra cercano a la pared del edificio, dificultando su distinción con respecto a la misma. En estos casos, la segmentación en color aporta información valiosa, como se aprecia en la Figura 5.6d. Un pequeño problema que aparece es que, debido a que los dos espectros presentan la misma confianza, se combinan en ocasiones en los resultados de la fusión los falsos positivos de los dos espectros (como se observa en la Figura 5.6e), con la bajada resultante en la precisión, muy cercana a la más baja de los dos espectros.

5.4.4. Secuencia 8°Noche

Esta secuencia resulta especialmente problemática debido a las condiciones ambientales en que se grabó. Capturada en las primeras horas de la noche, los edificios no se han enfriado lo suficiente y su contraste térmico resulta elevado en relación con el resto del entorno, lo que provoca que los edificios aparezcan en la cámara térmica casi a la misma temperatura que los humanos. Por otro lado, siendo totalmente de noche, la cámara en color queda "ciega", solo apreciándose los humanos cuando se encuentran cerca de la farola que ilumina la escena en un lateral, como se puede apreciar en la Figura 5.7a. Debido a estas condiciones, la confianza en la cámara en color aparece como $BAJA$, mientras que los problemas de contraste mencionados provocan que la fiabilidad de la cámara en color solo aparezca como $MEDIA$ (el mínimo valor que puede tomar para la confianza $BAJA$ asignada a la cámara en color). La detección de humanos en infrarrojo funciona bien de todas formas cuando los humanos se encuentran alejados del fondo, tal y como se puede ver en la Figura 5.7b. Sin embargo, aparecen problemas cuando las personas se acercan a los edificios, como puede observarse en la Figura 5.7c, ya que la temperatura de sus cuerpos se confunde con la del edificio del fondo.

A pesar de las difíciles condiciones en que se ha capturado la secuencia, se puede observar en las estadísticas mostradas en la Tabla 5.12 que el algoritmo de fusión nuevamente mejora los resultados respecto a los parciales de cada cámara. Debido a que en la zona común a ambos espectros únicamente se tiene en cuenta a la cámara en infrarrojo en estas condiciones, la sensibilidad alcanzada debería ser la misma. Por contra, en este caso, esta medida aumenta gracias al algoritmo de seguimiento empleado para aportar robustez a los resultados de la fusión. Así, en la Figura 5.7d se aprecia que, aunque el humano no ha sido detectado por la segmentación en infrarrojo en este fotograma en particular,

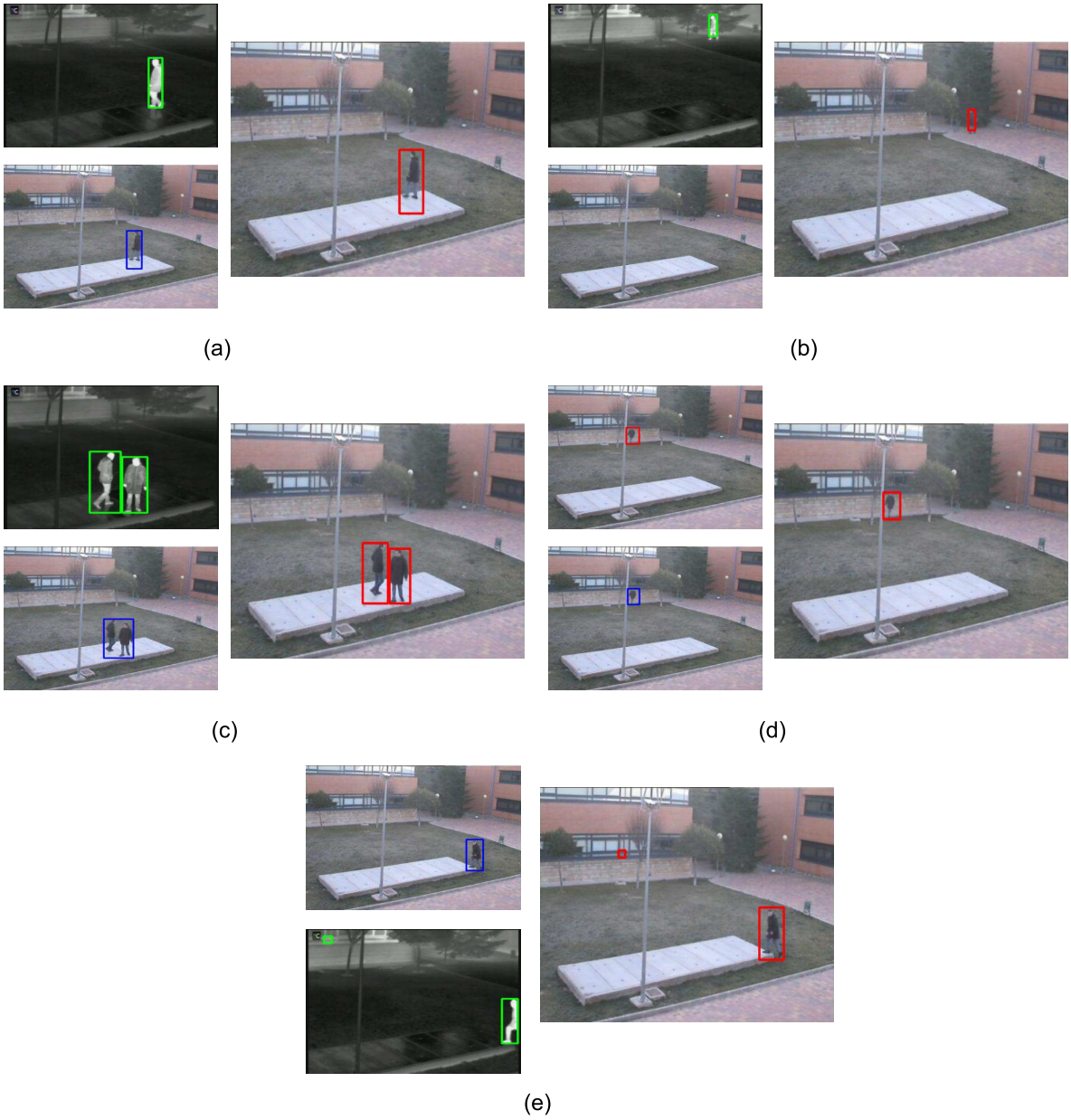


Figura 5.6: Ejemplos de resultados obtenidos en la secuencia 3°Soleado. (a) Condiciones óptimas para la segmentación. (b)Falso negativo de la segmentación en color corregido por la segmentación en infrarrojo y reflejado en la fusión. (c) Grupo correctamente separado por la segmentación en infrarrojo respecto a la detección de humanos en color. (d) Falso negativo de la segmentación en infrarrojo corregido por la segmentación en color y reflejado en la fusión. (e) Falso positivo de la segmentación en infrarrojo arrastrado a la fusión.

el hecho de que antes haya sido detectado alguien en esa posición durante varios fotogramas hace que todavía se estime que no ha abandonado la escena, estimándose correctamente su posición en la misma.

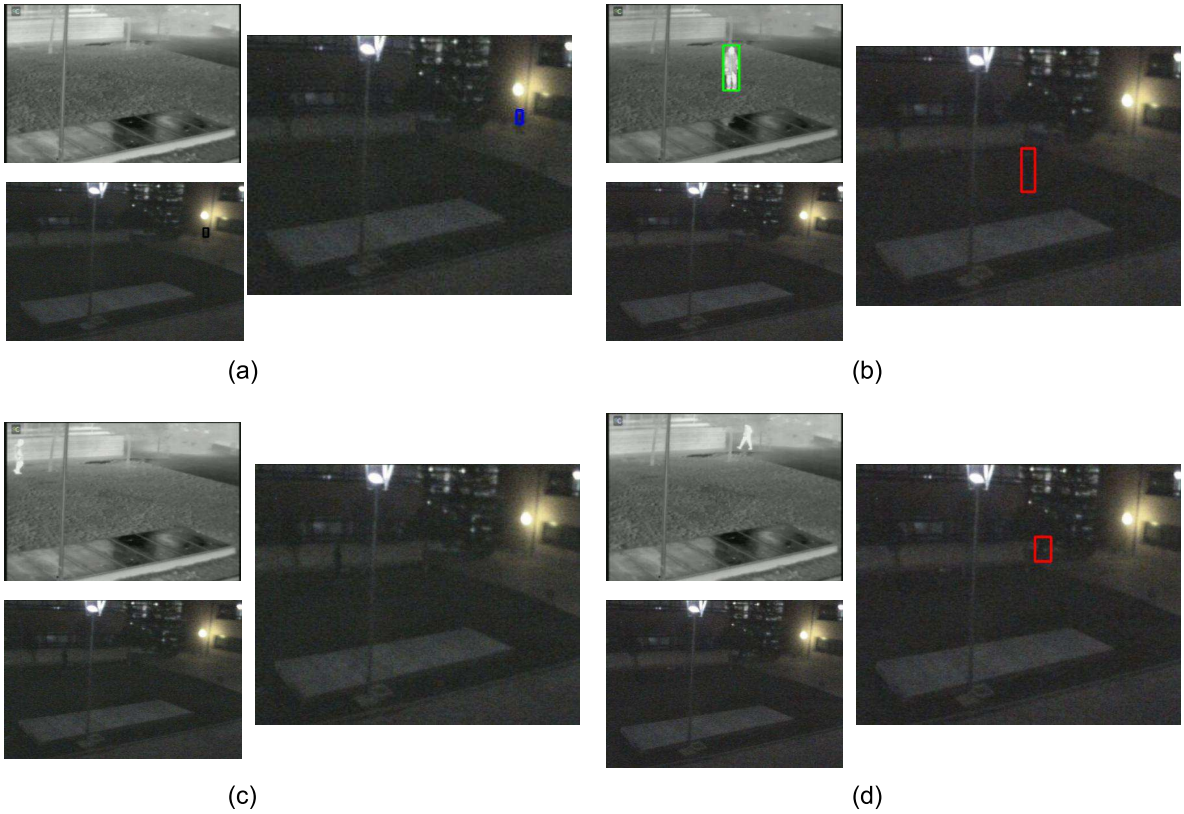


Figura 5.7: Ejemplos de resultados obtenidos en la secuencia 8°Noche. (a) Acierto de la detección de humanos en color. (b) Falso negativo de la segmentación en color corregido por la segmentación en infrarrojo y reflejado en la fusión. (c) Falso negativo de la segmentación en infrarrojo debido al bajo contraste. (d) Falso negativo de la segmentación en infrarrojo corregido por el seguimiento dentro del algoritmo de fusión.

Tabla 5.12: Resultados alcanzados en la secuencia 8°Noche

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,81	0,86	0,83
C-BS (Zona común)	0,01	0,98	0,01
C-BS (Imagen completa)	0,01	0,98	0,01
Fusión (Zona común)	0,83	0,86	0,85
Fusión (Imagen completa)	0,70	0,90	0,79

5.4.5. Secuencia 9°Nublado

Esta secuencia fue grabada en buenas condiciones de iluminación para la cámara en color y con alto contraste en la térmica-infrarroja, lo que provoca que la confianza de ambos espectros aparezca etiquetada como $ALTA$. Un ejemplo de estas condiciones puede verse en la Figura 5.8a. Sin embargo, la sensibilidad de la detección de humanos en color es de un $100\,\%$, quedando la de la segmentación en infrarrojo en un $94\,\%$ tal y como se muestra en la Tabla 5.13.

Tabla 5.13: Resultados alcanzados en la secuencia 9°Nublado

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,94	0,96	0,95
C-BS (Zona común)	1,00	0,94	0,96
C-BS (Imagen completa)	1,00	0,87	0,93
Fusión (Zona común)	1,00	0,89	0,94
Fusión (Imagen completa)	1,00	0,86	0,92

Puesto que la fusión en estas condiciones tiene en cuenta los resultados de ambos espectros, la sensibilidad de ésta se sitúa en un $100\,\%$. Este excelente dato se debe a que, cuando la detección de humanos en el espectro infrarrojo arroja falsos negativos momentáneos (tal y como se aprecia en la Figura 5.8b), la segmentación en color complementa la información. Finalmente destacaremos el problema de la bajada de precisión que se produce en la fusión. Esto es debido a la existencia de un algoritmo de seguimiento para disminuir el número de falsos negativos. En esta secuencia aparecen pocos falsos positivos en ambas segmentaciones (como se puede apreciar en las Figuras 5.8c y 5.8d). Ocurren normalmente durante un periodo de tiempo de 4 o 5 fotogramas. Además, estos falsos positivos se suelen situar en zonas donde previamente había otro humano detectado por otro espectro, lo que provoca una asignación errónea como detección común a ambos espectros, como se observa en la Figura 5.8e, del fotograma inmediatamente posterior al de la Figura 5.8c. Puesto que la credibilidad de estas detecciones es igual al número de fotogramas durante los que han aparecido previamente, éstas se mantendrán durante esa misma cantidad de fotogramas antes de ser eliminadas como detección. En otras secuencias se comprobará como, a pesar de este problema puntual, el algoritmo de seguimiento resulta beneficioso en muchas más ocasiones.

5.4.6. Secuencia 10°Nublado

Aunque esta secuencia se grabó a una temperatura relativamente baja y los resultados de la detección en infrarrojo son excelentes (tal y como se ve en la Tabla 5.14), ésta es la primera secuencia en la que la confianza de esta detección baja a $MEDIA$, debido al bajo contraste de la misma, apareciendo falsos negativos tal y como se aprecia en la Figura 5.9a. A pesar de este problema, el uso del movimiento facilita el hecho de que en esta secuencia los resultados del algoritmo *IR-(SF+FS)* sean tan buenos. También los resultados del algoritmo *C-BS* son excelentes debido a las condiciones óptimas de iluminación de la escena.

El hecho de que los dos algoritmos, cuyos resultados se están fusionando, tengan sensibilidades muy altas, hace que los resultados de la fusión sean también muy buenos. El motivo es que, aunque el algoritmo *IR-(SF+FS)* tiene una confianza menor que la del *C-BS*, su habilidad para dividir grupos y eliminar la aparición momentánea de falsos negativos hace que la sensibilidad total de la fusión alcance un $100\,\%$. Un ejemplo de esta eliminación de falsos negativos se observa en la Figura 5.9b, en la que el humano aparece entre sombras, dificultando su detección en el espectro visible. Sin embargo, recordemos que el algoritmo de seguimiento, antes de considerar que un humano ha abandonado una

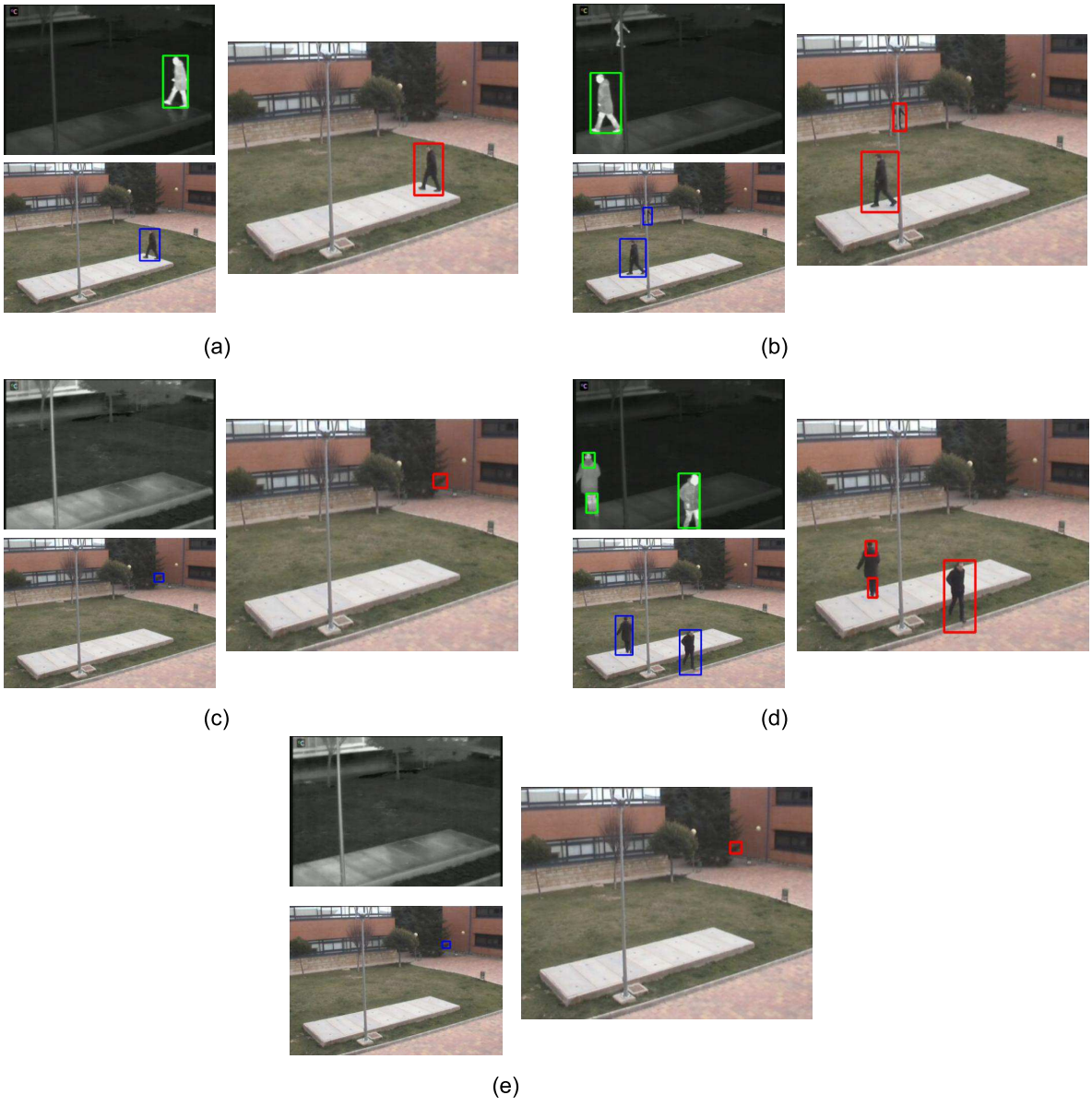


Figura 5.8: Ejemplos de resultados obtenidos en la secuencia 9°Nublado. (a) Condiciones óptimas para la segmentación. (b) Falso negativo de la segmentación en infrarrojo corregido por la segmentación en color y reflejado en la fusión. (c) Falso positivo de la segmentación en color. (d) Falso positivo de la segmentación en infrarrojo. (e) Falso positivo de la fusión.

escena, comprueba si no ha sido detectado por el espectro que tiene menor confianza. Al haberse producido en este caso dicha detección, el humano es incorporado al resultado final.

Se observa que la precisión de la fusión es un poco menor, debido a que la segmentación en el espectro infrarrojo sobreparticiona ocasionalmente a los humanos (como se aprecia en la Figura 5.9c). También aparecen puntualmente falsos positivos en la detección de humanos en color. Estos problemas apenas afectan a los resultados generales, que demuestran el buen funcionamiento general

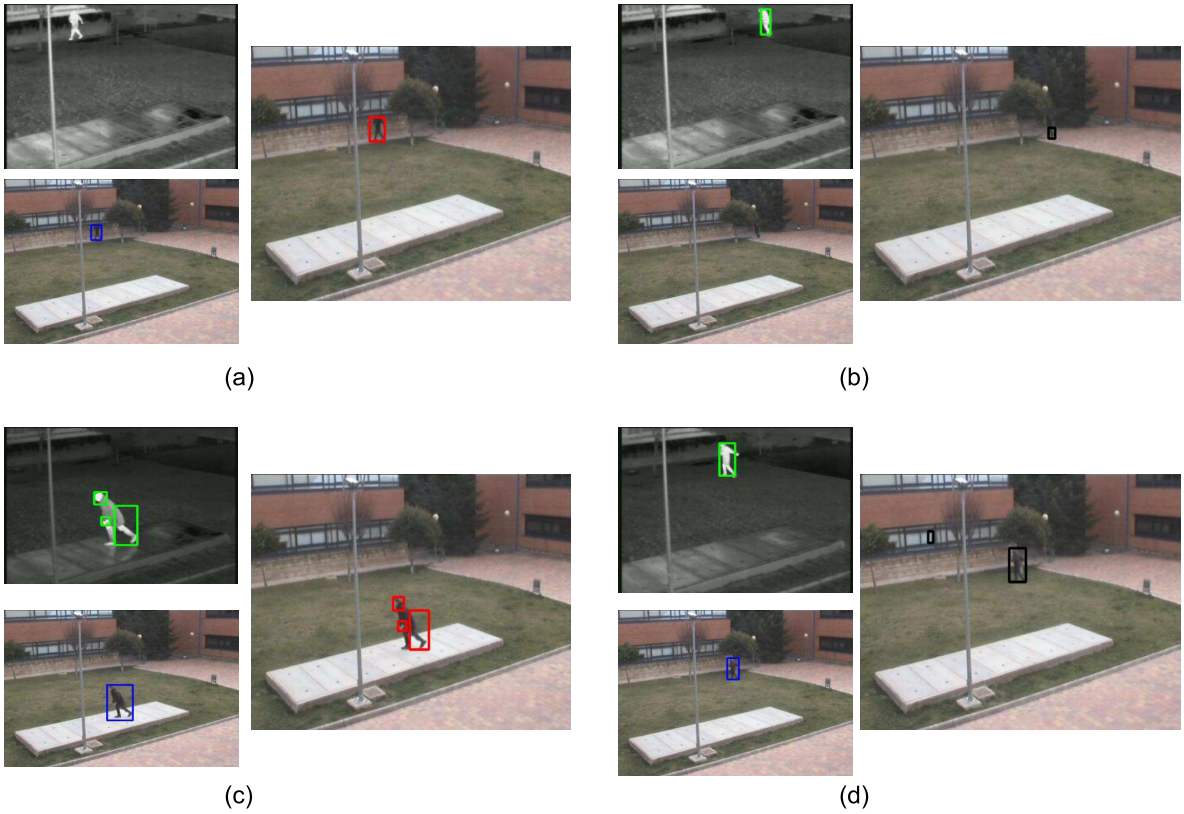


Figura 5.9: Ejemplos de resultados obtenidos en la secuencia 10°Nublado. (a) Falso negativo de la segmentación en infrarrojo corregido por la fusión. (b)Falso negativo de la segmentación en color corregido gracias a la fusión y al seguimiento. (c) Falso positivo de la detección de humanos en infrarrojo que afecta negativamente a la fusión. (d) Condiciones óptimas para la detección de humanos en ambos espectros.

Tabla 5.14: Resultados alcanzados en la secuencia 10°Nublado

Espectro	**Sensibilidad**	**Precisión**	***F-score***
IR-(SF+FS)	0,99	0,99	0,99
C-BS (Zona común)	0,98	0,97	0,97
C-BS (Imagen completa)	0,98	0,89	0,93
Fusión (Zona común)	1,00	0,93	0,96
Fusión (Imagen completa)	0,99	0,86	0,92

de los dos espectros en esta escena, tal y como muestra la Figura 5.9d.

5.4.7. Secuencia 15°Amanece

La secuencia fue grabada al comienzo del amanecer cuando todavía había poca visibilidad y el ambiente todavía se encontraba frío. A causa de esto, la confianza en el espectro infrarrojo aparece como $ALTA$ debido a su gran contraste, mientras que la del visible aparece como $BAJA$, ya que,

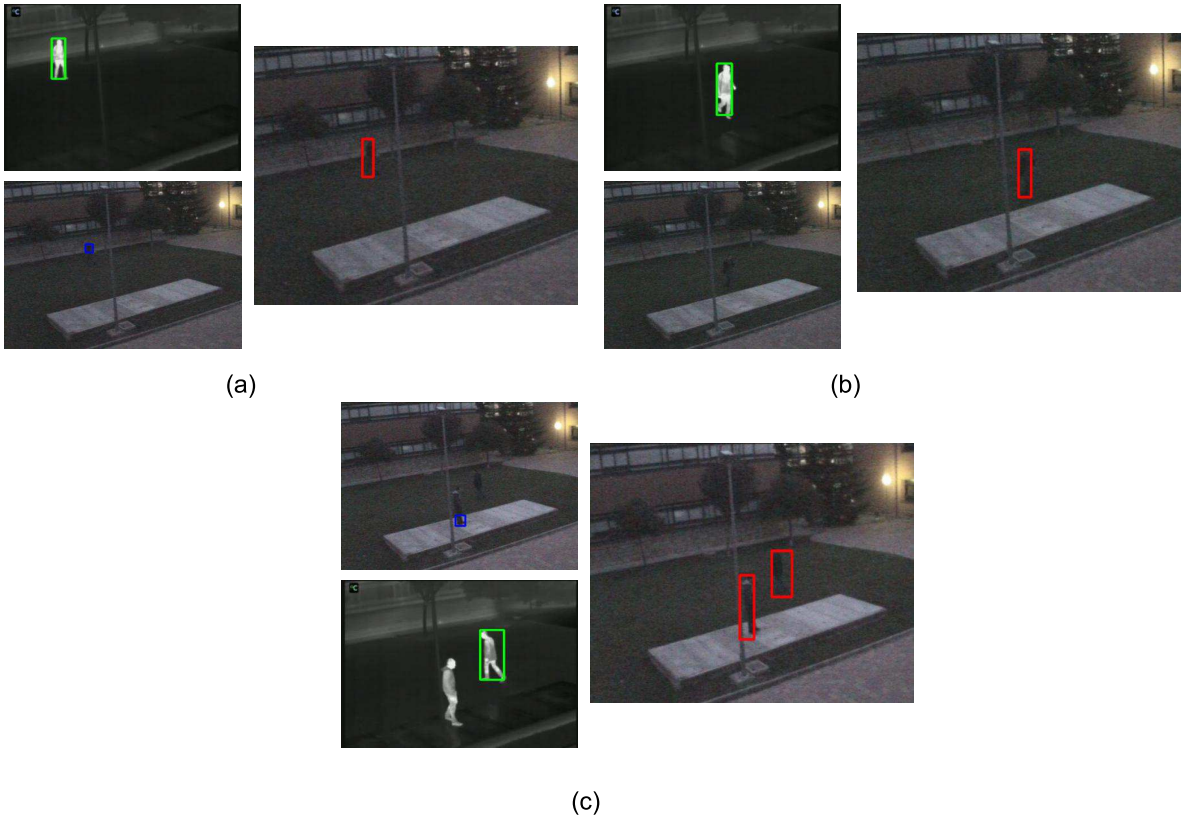


Figura 5.10: Ejemplos de resultados obtenidos en la secuencia 15°Amanece. (a) Detecciones correctas en los dos espectros usados. (b) Falso negativo de la detección de humanos en el espectro visible debido a la baja iluminación. (c) Falso negativo de la detección de humanos en infrarrojo corregido gracias a la fusión y al seguimiento.

si bien comienzan a aparecer más detecciones en las zonas mejor iluminadas del escenario respecto a la anterior secuencia nocturna (como puede apreciarse en la Figura 5.10a), sigue habiendo muchas zonas del escenario con visibilidad escasa, tal y como se ve en la Figura 5.10b. En este caso, la mayor contribución a la mejora que experimenta la sensibilidad de la fusión respecto al algoritmo *IR-(SF+FS)* la apoya la parte de seguimiento de la propuesta de fusión. Un ejemplo de este caso se puede apreciar en la Figura 5.10c, donde un humano ha sido momentáneamente no detectado por la segmentación en infrarrojo. A pesar de este falso negativo, continúa apareciendo como válido en la escena gracias a que había aparecido en anteriores ocasiones en esta segmentación. El seguimiento, por tanto, concluye que el humano tiene muchas probabilidades de encontrarse todavía ahí.

Finalmente, en la Tabla 5.15 se observa que las estadísticas en la imagen total disminuyen con respecto a la zona común, ya que la cámara en el espectro visible apenas detecta nada en esta secuencia y el total de humanos que aparecen en todo el escenario es mayor que los que aparecen únicamente en dicha zona común.

Tabla 5.15: Resultados alcanzados en la secuencia 15°Amanece

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,93	1,00	0,96
C-BS (Zona común)	0,19	1,00	0,32
C-BS (Imagen completa)	0,32	0,86	0,47
Fusión (Zona común)	0,94	0,99	0,96
Fusión (Imagen completa)	0,85	0,99	0,92

Tabla 5.16: Resultados alcanzados en la secuencia 15°Nublado

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,91	0,97	0,94
C-BS (Zona común)	1,00	0,99	0,99
C-BS (Imagen completa)	1,00	0,98	0,99
Fusión (Zona común)	1,00	0,95	0,98
Fusión (Imagen completa)	1,00	0,96	0,98

5.4.8. Secuencia 15°Nublado

En este caso las temperaturas de los humanos se asemejan mucho a las de los elementos en la escena, tal y como se aprecia en la Figura 5.11a. Sin embargo, el uso de la información de movimiento evita que los resultados de la segmentación *IR-(SF+FS)* no sean mucho más bajos, de forma similar a lo que ocurrió en la secuencia 10°Nublado. Debido a este bajo contraste, la confianza en este espectro infrarrojo se establece como $MEDIA$, mientras que la del espejo visible se asigna como $ALTA$, puesto que las condiciones de iluminación permiten apreciar a los humanos con claridad, tal y como se puede ver también en la Figura 5.11b. Así mismo, la Tabla 5.16 nos muestra que el funcionamiento del algoritmo *C-BS* es excelente en comparación al *IR-(SF+FS)*. Aparecen de todas formas falsos positivos ocasionales, debido al reflejo de las personas en la ventana, como se aprecia en la Figura 5.11c. Estos reflejos pueden llegar a tener cierto efecto negativo, también en la precisión total alcanzada en la fusión, debido a la utilización del algoritmo de seguimiento (en este caso más centrado en el espectro visible al tener éste más confianza), aunque este efecto apenas impacta a los resultados totales. Por tanto, se puede apreciar que, en este caso, los resultados de la fusión tienden a ajustarse al espectro que mejor funciona en estas circunstancias.

5.4.9. Secuencia 18°Soleado

Esta secuencia presenta un caso excepcional ya que aparecen fallos puntuales en la segmentación en el espectro visible. Estos son debidos a la existencia de numerosas sombras muy alargadas, que unidas a la gran cantidad de grupos presentes (muchas veces detectados como una sola persona), provocan que este espectro presente una sensibilidad menor que la esperable a pesar de que figure con confianza asignada como $ALTA$. Estos resultados se muestran en la Tabla 5.17. En estas circunstan-

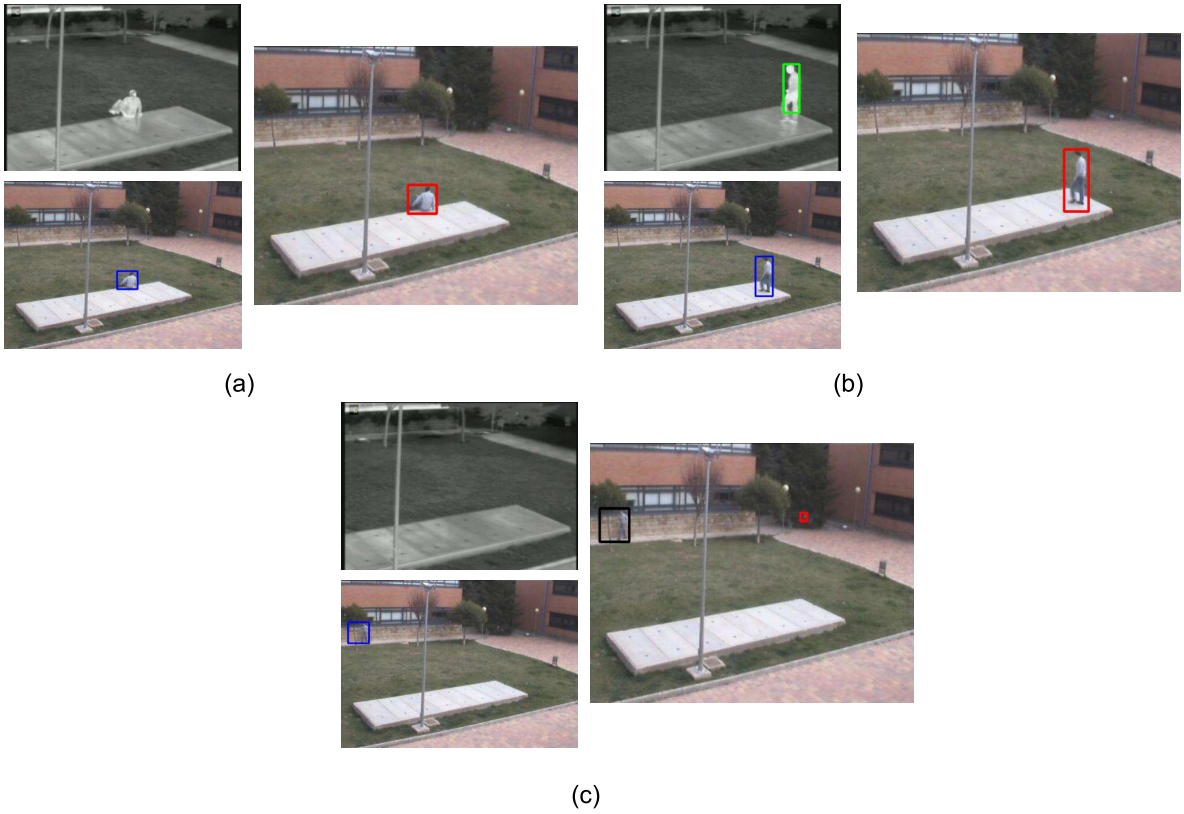


Figura 5.11: Ejemplos de resultados obtenidos en la secuencia 15°Nublado. (a) Falso negativo en la detección de humanos en infrarrojo corregido gracias a la fusión. (b) Detección correcta de humanos en los dos espectros usados. (c) Falso positivo en la fusión debido a la detección de humanos en el espectro visible.

cias es donde entra en juego el apoyo de la detección de humanos en el espectro infrarrojo a la hora de dividir los grupos erróneamente unificados, tal y como se aprecia en la Figura 5.12a.

Ya se ha especificado que los resultados del algoritmo *IR-(SF+FS)* no son totalmente descartados, sino que se mantienen como apoyo de la detección de humanos en el espectro visible. Esto sucede a pesar de que su bajo contraste (como se puede ver en el falso negativo que aparece en la Figura 5.12b) provoque que esta segmentación aparezca con confianza $MEDIA$. En esta ocasión vemos cómo su alta sensibilidad también ayuda al algoritmo de seguimiento utilizado, ya que la presencia de humanos previamente detectados por el algoritmo *C-BS* aparece reafirmada ahora por la detección de humanos en el espectro infrarrojo. En la Figura 5.12c vemos un ejemplo de como, sin el algoritmo de seguimiento, estos humanos serían descartados por no haber sido detectados en el espectro visible. Sin embargo, el algoritmo de seguimiento concluye que antes había ahí dos personas que, por su posición y credibilidad asociadas, es difícil que hayan abandonado la escena. Esta estimación provoca que se inspeccionen los resultados de la detección de humanos en el espectro infrarrojo. Al apuntar esta segmentación detecciones en esa zona, se confirma que los humanos siguen en esa posición, incorporándose a la lista definitiva de humanos en la escena.

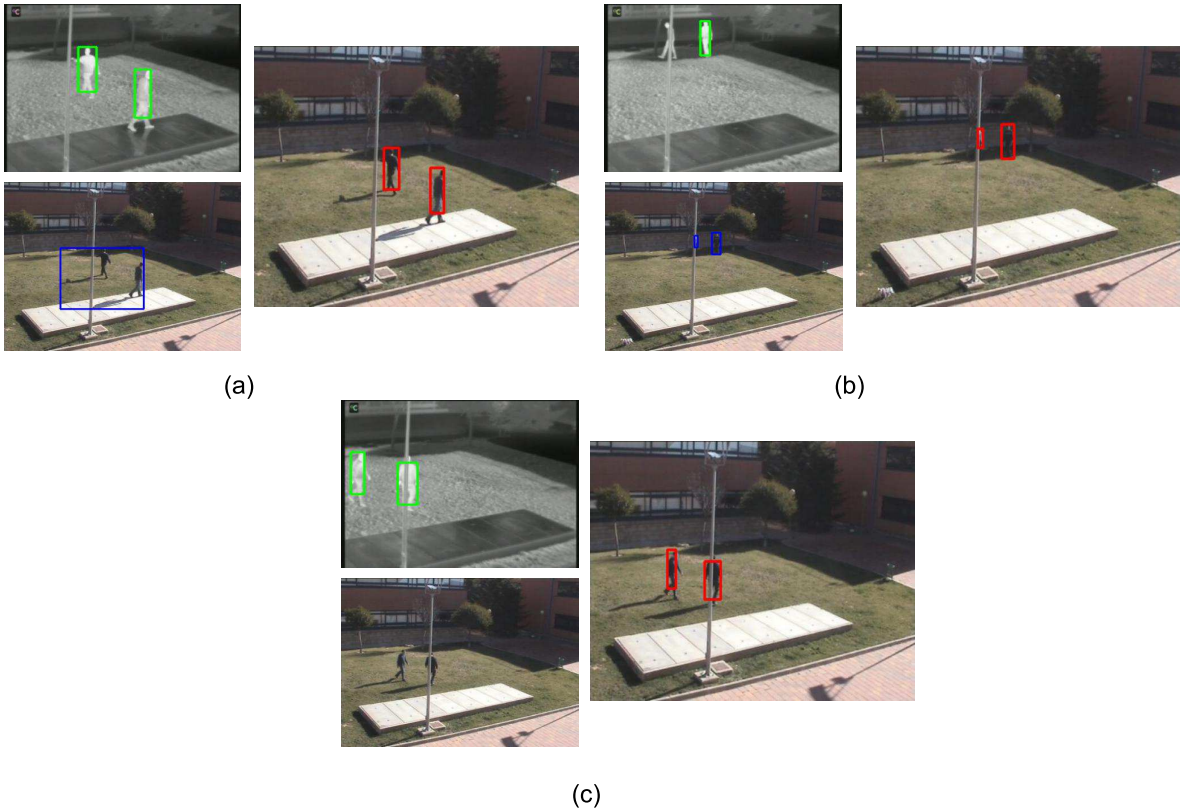


Figura 5.12: Ejemplos de resultados obtenidos en la secuencia 18°Soleado. (a) Grupo detectado en la detección de humanos en color corregido gracias a la detección de humanos en el espectro infrarrojo. (b) Falso negativo en la detección de humanos en el espectro en infrarrojo corregido por el algoritmo de fusión. (c) Falsos negativos en la detección de humanos en el espectro visible corregido gracias al trabajo del seguimiento en el algoritmo de fusión.

Tabla 5.17: Resultados alcanzados en la secuencia 18°Soleado

Espectro	**Sensibilidad**	**Precisión**	***F-score***
IR-(SF+FS)	0,93	0,94	0,94
C-BS (Zona común)	0,79	0,96	0,87
C-BS (Imagen completa)	0,79	0,96	0,87
Fusión (Zona común)	0,90	0,93	0,92
Fusión (Imagen completa)	0,93	0,87	0,90

5.4.10. Secuencia 23°Soleado

En esta secuencia, en la que ya hay una alta temperatura ambiental, la Tabla 5.18 muestra que la sensibilidad del algoritmo *IR-(SF+FS)* cae drásticamente. De hecho, es la primera donde encontramos valores situados en el $60\,\%$, ya que la temperatura de los humanos es similar a la del entorno en general. De este suceso nos da una pista el alto valor de v_{IR} que presenta. Por tanto, la confianza en el espectro infrarrojo aparece fijada siempre como $BAJA$, debido a la dificultad de distinguir a los

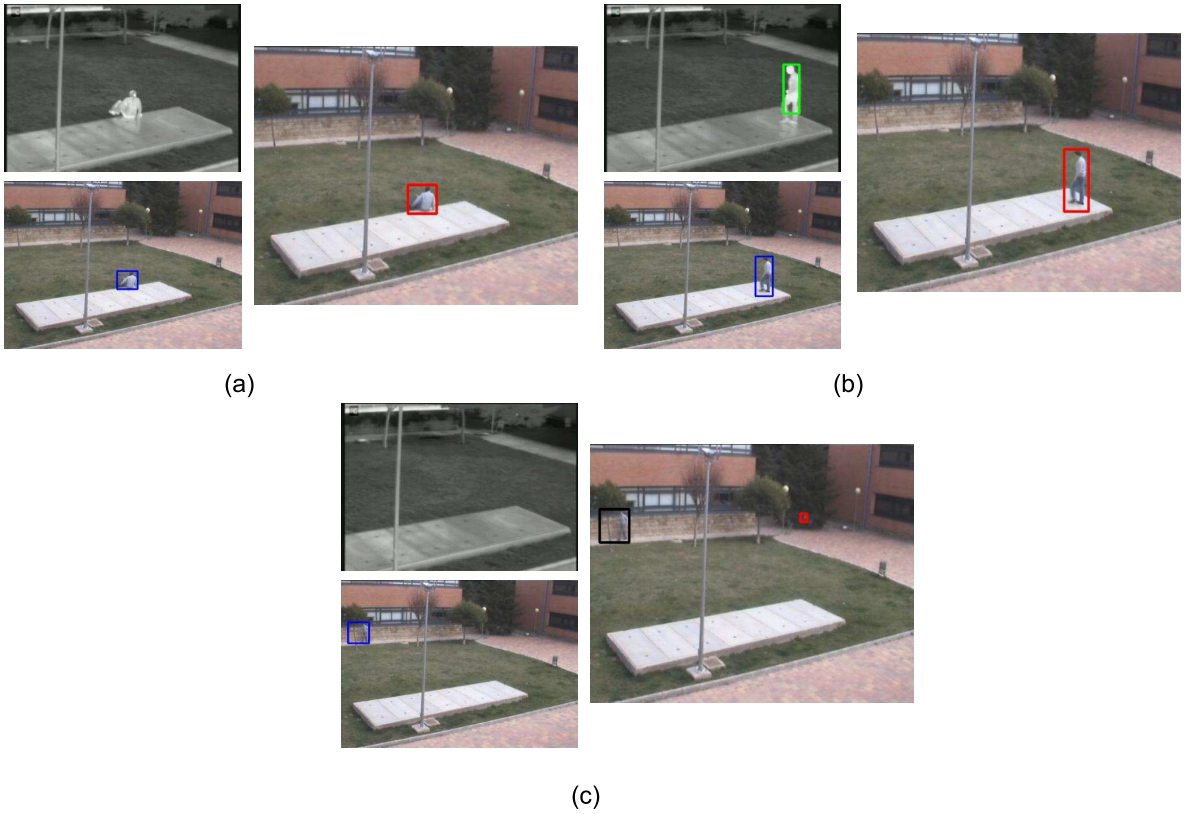


Figura 5.13: Ejemplos de resultados obtenidos en la secuencia 23°Soleado. (a) Falso negativo en la detección de humanos en el espectro infrarrojo corregido gracias a la fusión. (b) Falso negativo en la detección de humanos en el espectro visible debido a la imposibilidad de separar los componentes del grupo que aparece. (c) Falso positivo de la detección de humanos en el espectro infrarrojo omitido gracias a la fusión.

humanos en esta cámara. Se puede observar un ejemplo de esta situación en la Figura 5.13a.

Por su parte, la sensibilidad en el espectro visible no resulta tampoco excesivamente elevada, debido a la gran cantidad de grupos presentes en la escena (como se puede ver en la Figura 5.13b). De todas formas, aparecen buenos resultados cuando únicamente se encuentra presente un humano o hay varias personas en la escena aunque separadas entre sí, como se observa en la Figura 5.13c. Otro hecho destacable es que, al estar la confianza en la segmentación en infrarrojo asignada como $BAJA$, los falsos positivos de esta detección de humanos no afectan a los resultados de la fusión. En todo caso, la propuesta realizada sigue beneficiándose de los resultados alcanzados en el algoritmo de seguimiento, tal y como se ve en la Figura 5.13c.

5.4.11. Secuencia 28°Soleado

Esta secuencia presenta resultados similares a la anterior, pudiéndose observar que el algoritmo *IR-(SF+FS)* continúa arrojando peores resultados conforme aumenta la temperatura, tal y como

Tabla 5.18: Resultados alcanzados en la secuencia 23°Soleado

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,60	0,86	0,71
C-BS (Zona común)	0,82	0,98	0,90
C-BS (Imagen completa)	0,86	0,99	0,92
Fusión (Zona común)	0,83	0,98	0,90
Fusión (Imagen completa)	0,86	0,99	0,92

Tabla 5.19: Resultados alcanzados en la secuencia 28°Soleado

Espectro	Sensibilidad	Precisión	*F-score*
IR-(SF+FS)	0,39	0,96	0,55
C-BS (Zona común)	0,81	0,99	0,91
C-BS (Imagen completa)	0,80	0,97	0,88
Fusión (Zona común)	0,85	0,99	0,91
Fusión (Imagen completa)	0,84	0,98	0,90

muestra la Tabla 5.19. Nuevamente, el bajo contraste de la secuencia provoca que la confianza en dicho espectro aparezca como $BAJA$, mientras que las condiciones de iluminación en el espectro visible continúan siendo óptimas y vuelve a tener confianza $ALTA$. En este último, los humanos son detectados sin problema en caso de que se encuentren solos en la escena o separados, tal y como se observa en la Figura 5.14a.

Existe nuevamente el problema con la división de grupos del algoritmo *C-BS* previamente mencionado, tal y como se ve en la Figura 5.14b. Aparece también un nuevo problema adicional, provocado porque algunos humanos permanecen quietos en una misma zona durante un largo periodo de tiempo, lo que provoca que sean "absorbidos" por el modelo de fondo que elabora esta propuesta. En estos casos, el seguimiento vuelve a aportar información que mejora los resultados, tal y como se aprecia en la Figura 5.14c. En consecuencia, los resultados de la propuesta de fusión son superiores a los del algoritmo *C-BS* por sí solos, sin que la precisión se vea negativamente afectada en exceso.

5.4.12. Secuencia 33°Soleado

En esta secuencia, la cámara infrarrojo ya queda con una sensibilidad prácticamente nula, debido al extremo calor que hace. Sin embargo, en esta ocasión, los objetos más oscuros resultan ser los humanos, tal y como se aprecia en la Figura 5.15a. Esto repercute en que el contraste dé lugar a una confianza asignada como $MEDIA$, llegando en ocasiones a ser calificada como $ALTA$. En consecuencia, pueden aparecen falsos positivos en la fusión, como se aprecia en la Figura 5.15b. También el espectro visible, a pesar de tener una sensibilidad muy alta, arroja errores momentáneos por falsos positivos provocados por el viento, así como algunos falsos negativos (como se ve en la Figura 5.15c). En dichas ocasiones, nuevamente el algoritmo de fusión y seguimiento mejora los resultados en sensibilidad, tal y como se observa en la Figura 5.15d y en la Tabla 5.20.

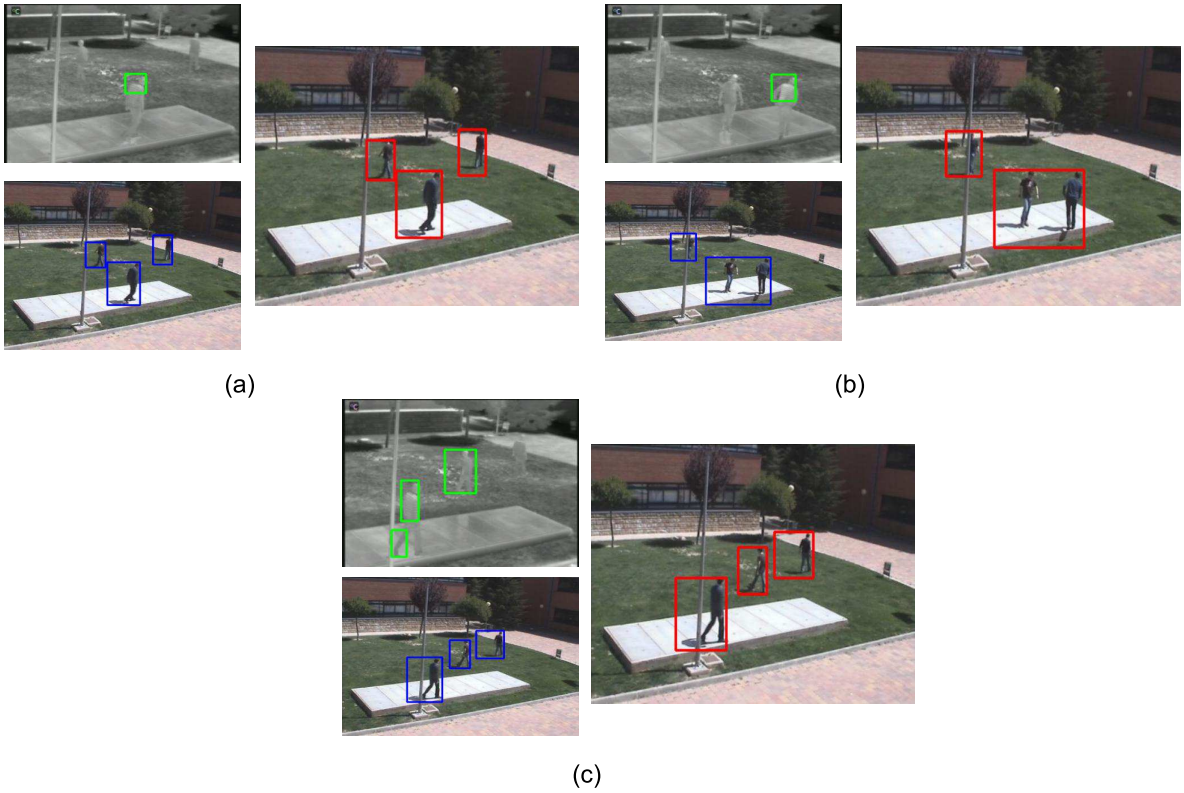


Figura 5.14: Ejemplos de resultados obtenidos en la secuencia 28°Soleado. (a) Falso negativo en la detección de humanos en el espectro infrarrojo corregido gracias a la fusión. (b) Falso negativo en la detección de humanos en el espectro visible debido a la imposibilidad de separar los componentes del grupo que aparece. (c) Falso positivo de la detección de humanos en el espectro infrarrojo omitido gracias a la fusión.

Tabla 5.20: Resultados alcanzados en la secuencia 33°Soleado

Espectro	**Sensibilidad**	**Precisión**	***F-score***
IR-(SF+FS)	0,03	0,84	0,04
C-BS (Zona común)	0,97	0,92	0,94
C-BS (Imagen completa)	0,97	0,92	0,94
Fusión (Zona común)	0,98	0,89	0,93
Fusión (Imagen completa)	0,97	0,90	0,94

5.4.13. Resumen de los resultados alcanzados

Por último, se analiza el resultado promedio de cada uno de los algoritmos probados, comparando nuevamente su sensibilidad, precisión y *F-score*. En la Tabla 5.21 se observa la conclusión de que la fusión siempre aporta mejoras con respecto a los algoritmos de segmentación por separado, ya sea en sensibilidad en particular o en *F-score* en general. Podemos apreciar que, mientras que cada espectro en solitario garantiza únicamente una sensibilidad media inferior al $80\,\%$, los resultados de la fusión

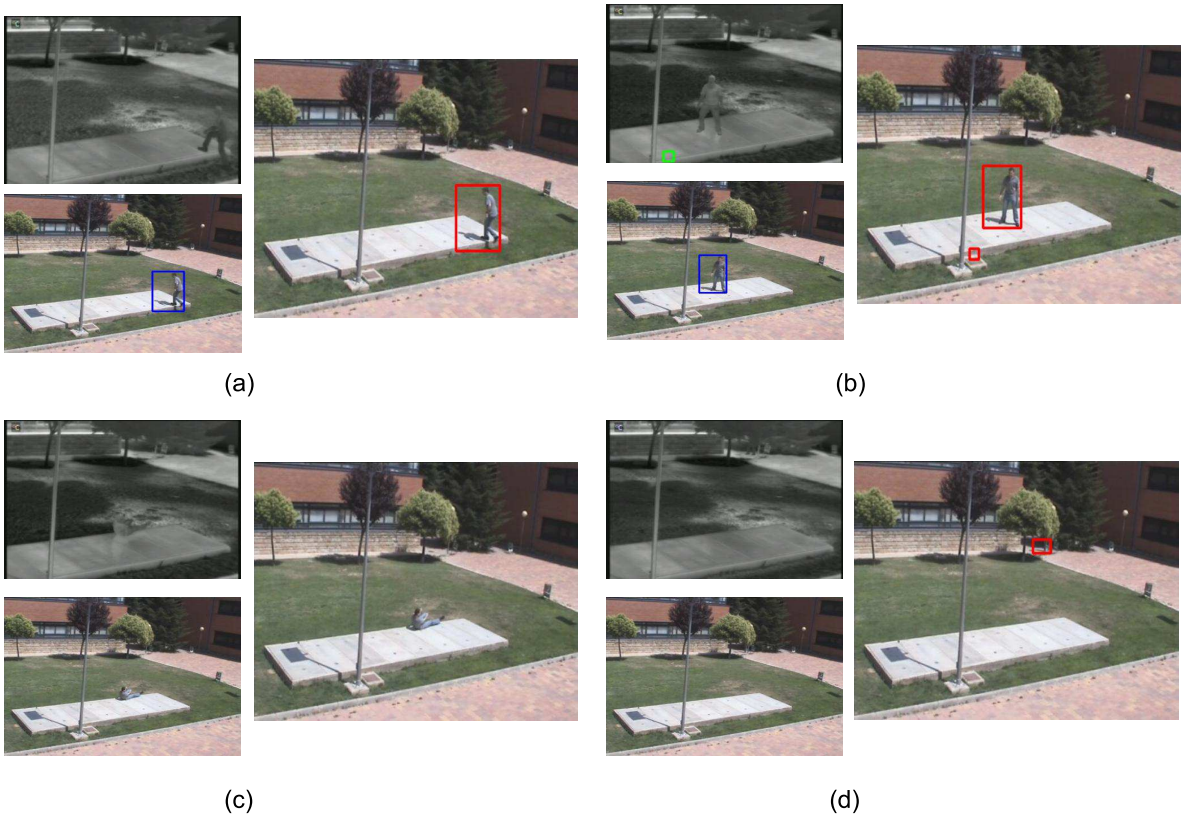


Figura 5.15: Ejemplos de resultados obtenidos en la secuencia 33°Soleado. (a) Falso negativo en la detección de humanos en el espectro infrarrojo corregido gracias a la fusión. (b) Falso negativo en la detección de humanos en el espectro visible. (c) Falso positivo en la fusión debido a la detección de humanos en el espectro infrarrojo. (d) Falso negativo de la detección de humanos en el espectro visible corregido gracias al algoritmo de fusión y seguimiento.

se alzan hasta un valor de un $92\,\%$. Este dato resulta un incremento considerable con respecto a las dos aproximaciones por separado. Por su parte, la precisión se ajusta a la de la detección que peores resultados arroja, lo que tampoco resulta excesivamente negativo, ya que en los dos algoritmos probados se elevan por encima del $90\,\%$. El *F-score* confirma la importante mejora experimentada en general al utilizar el algoritmo de fusión propuesto. Los resultados de la fusión aplicada a toda la imagen resultan algo inferiores, al tener más peso en ese caso el algoritmo *C-BS*, experimentándose aun así resultados similares. Estas estadísticas confirman que la solución propuesta es una aproximación interesante al problema de la detección de humanos dentro de un entorno de exteriores, ya que proporciona independencia con respecto a las condiciones atmosféricas de cada escena. Estas condiciones no afectan a unos resultados cuyas estadísticas en la mayoría de las secuencias traspasan el $90\,\%$, tanto en sensibilidad como en precisión y *F-score*. Estos datos normalmente se asemejan a las estadísticas obtenidas por el algoritmo de detección de humanos que mejor funciona y siempre las superan en el caso de la sensibilidad.

Tabla 5.21: Resultados medios alcanzados

Espectro	Sensibilidad	Precisión	*F-score*
IR-SF+FS	0,76	0,93	0,80
C-BS (Zona común)	0,72	0,97	0,76
C-BS (Imagen completa)	0,78	0,97	0,81
Fusión (Zona común)	0,92	0,93	0,92
Fusión (Imagen completa)	0,88	0,92	0,89

5.5. Conclusiones

En este capítulo se ha aportado información exhaustiva acerca de las pruebas que se han llevado a cabo para evaluar el funcionamiento del sistema desarrollado. Tras establecer el entorno y las condiciones en las que se ha llevado a cabo dicha evaluación, se han explicado los parámetros que se utilizaron para la evaluación de cada uno de los algoritmos desarrollados.

A continuación, se han comparado las diversas propuestas implementadas en el espectro infrarrojo, concluyendo que el algoritmo *IR-(SF+FS)* resulta el más interesante para llevar a cabo las pruebas, debido a que la información del movimiento resulta relevante, añadida a los resultados obtenidos por el algoritmo *IR-SF* sobre un único fotograma. Análogamente, se ha decidido que el algoritmo *C-BS* resulta más apropiado para la detección de humanos en el espectro visible, debido a que el tipo de movimiento y la naturaleza del escenario hacen que su funcionamiento sea mejor que la alternativa *C-AC* desarrollada.

Finalmente, se ha procedido a analizar cada secuencia grabada por separado, comparando los resultados de la detección de humanos en cada espectro con aquellos alcanzados por la fusión, aportándose datos cuantitativos y cualitativos. Se observa que en todas las secuencias se experimenta mejora al usar la fusión desarrollada con respecto a los algoritmos en solitario. Esta impresión es confirmada al analizar los resultados promedios de cada algoritmo, que revelan una mejora importante al emplear la fusión. Dichas evaluaciones confirman que la propuesta realizada es interesante debido a su independencia de las condiciones del entorno.

Chapter 6

Conclusions and Future Work

This final chapter describes the main conclusions achieved as a result of the research carried out during the course of this dissertation. It also establishes the main future research lines which have emerged from the developed work. Finally, the publications arisen as a result of the development of this thesis are also shown through describing their content, and the research projects which have contributed to financing this work are also shown.

6.1. Conclusions

Nowadays, human detection is a key topic in video surveillance. Video surveillance applications are used more frequently as the general interest in video security grows. A common approach for these applications is the exclusive use of color cameras. However, the use of color cameras is not always a versatile enough solution. A segmentation algorithm using this kind of cameras always depends on the illumination on the scene. Even though a color-based segmentation algorithm could resist sudden lighting changes, humans will be hard to detect if a scene is covered in darkness or the camera is overexposed (due to fog problems or a lighting source pointing directly to the camera and dazzling it). Any object in the scene can produce shadows or zones with worse illumination conditions where a human can be partially covered in darkness. This makes really hard to distinguish a human from the background.

On the other hand, thermal-infrared cameras are a strong alternative to perform human detection. However, they also show several problems. Humans are always highlighted from the background on scenes recorded at cold environments since their thermal readings are warmer than the scene conditions. Though, as the scenario temperature rises, humans tend to be easily confused with the background. This problem makes people very hard to distinguish at considerable heat conditions where they can even appear colder than the rest of the objects in the environment.

A basic system composed of two cameras seems to be an appropriate solution if only the results from one of them is used depending on the scene conditions. Yet, in a large number of situations, both

cameras can add useful and complementary information, so that humans not detected by one camera appear in the results from the other and viceversa. Moreover, a low detection rate and a high number of false positives are obtained when only one camera considered and precisely that camera is working under punctual adverse conditions. Thus, the study of "intermediate" cases where both cameras are working under favorable circumstances is specially interesting. It is also important to know how to solve contradictions between both cameras (such as a human detected by a camera and not found by the other) by combining the detection results in some cases or discarding information in others.

Thus, the major objective proposed in this thesis has been the "design and implementation of a robust people detection system based on fusing the information provided after human segmentation in infrared and color spectra". This objective has been divided, in turn, in the following sub-objectives described in Chapter 1 and Section 1.3:

1. **Study of algorithms widely used in the literature for motion-based object segmentation and tracking, specially focusing on those centered on human detection, facing their future evaluation depending on the monitored scene**. The main techniques of human segmentation and tracking have been studied in Chapter 2 in Section 2.1 and Section 2.3, respectively.

   Segmentation techniques have been studied and classified, with a special focus on those centered on human detection. After establishing a theoretical basis, the analysis concluded that using thermal features of the humans is an interesting approach in thermal-infrared human detection. On the other hand, the use of color information is an interesting alternative which has been seldom used and it is very interesting for the improvement of human detection performed in a scene. With these conclusions, a series of human detection algorithms were proposed and developed in both spectra. These approaches are described in Chapter 4 and Section 4.3.

   An approach based on the analysis of a single frame (described in section 4.3.1.1 was initially proposed for the infrared spectrum. Motion information from the scene was added to expand the number of detections achieved, as described in section 4.3.1.2. Finally, the use of motion inherent to the camera that acquires the frames from the scene was also considered. This approach is explained in section 4.3.1.3.

   Two algorithms were also developed in the visible spectrum using color information in different ways. On the one hand, motion information was used to propose an approach based on accumulative computation. This algorithm divides each frame into three channels (corresponding to color components R, G and B) to extract moving objects in each channel according to their motion history. This information is later unified through filtering the achieved results based on the people's area.

   On the other hand, a background subtraction algorithm was proposed, as well as an already existing mixture of Gaussian approach which creates a background model and is able to remove shadows in the scene. A series of heuristical restrictions were added to this algorithm in order to focus the results only on human detection.

   Tracking techniques were also studied and analyzed. For the current scenario it was decided that

feature-based methods were specially interesting. This study also derived in the development of a tracking algorithm. This proposal was crucial to improve the information fusion results, as seen in the following sub-objective description.

2. **Study of current image fusion techniques, specially focusing on those based on video captures taken by infrared and color cameras, and then proposal of a rule-based fusion mechanism obtained from experimentation in color and infrared spectra under different environmental conditions.** A study and classification of diverse techniques found for image fusion were introduced in Chapter 2, Section 2.2. This analysis is specially focused in those approaches taking infrared and visible camera images as an input after initially establishing a theoretical basis, as previously done for the segmentation and tracking techniques. After this study, it was decided to perform a fusion based on regions of interest, since such techniques are less dependent on the input cameras' features and they are more resistent to noise related problems.

   The developed rule-based fusion system was described in Chapter 4 under Section 4.4. In first place, it is necessary to synchronize images from both cameras and translate the thermal-infrared camera coordinates to the scene acquired by the color camera. This is due to the fact that a color camera usually covers a greater scenario area. A confidence level was assigned as well to each spectrum. This value is based on the average gray level of the image grabbed by the color camera in the visible spectrum, while on the infrared spectrum the coefficient between the average and standard deviation of each frame is used. After these initial stages (described in Section 4.2), the human detection algorithms are launched.

   When both human detections are complete, the results from each spectrum are analyzed based on the location of the detected humans and the confidence level assigned to that spectrum. Now, a decision is faced by a rule-based system for every detection at a spectrum. The system is extensively described in section 4.4.1. A human can be directly added to the final result, a more accurate alternative can be searched in the detection of the other spectrum (e.g. the visible spectrum detected only a human while the infrared spectrum was able to separate a group in that zone), detection areas from the two spectra can be combined when a human was detected in both of them. Detections can also be ignored for a while when the spectrum where they were realized did not have a high enough confidence level.

   The detections added to the final result are then associated to humans previously located in the scene by an identification process described in section 4.4.2. The main objective is to obtain a better overall perspective of what is currently going on in the scene. For every human detected in the scene we consider how long he/she has stayed in the scene as well as his/her speed on both axis. This data will be used later on to estimate if a human has left the scene when he/she has not been detected in a frame.

   The final estimation of the humans who actually remain in the scene is performed by the final tracking stage, as explained in section 4.4.3. For each human who has not been detected in the current frame several features are analyzed. These checks analyze the confidence level of the last spectrum where the human was detected to avoid possible false positives. The human's

location and speed are also checked because he/she might have been approaching the limits of the scene. Finally, detections of the first fusion stage which could have been originally discarded (because their origin spectrum did not have enough confidence level) are checked again to establish if the human could actually remain in the scene. Whether a human has left the scene of he/she is still located in the scenario is decided according to these factors. Thus, it can be affirmed that the system uses the image features of the frames grabbed by each camera and, at the same time, a feedback also takes place between previous detections and current segmentation results to reinforce the obtained detections, thus achieving a more robust system output.

3. **Proposal of a method for validating the results of human detection, and analysis of the results obtained in a real environment in order to validate the correct operation of the developed system.** In Chapter 3 under Section 3.3 the metrics used to evaluate the correct performance of the system are described, that is, sensitivity, precision and *F-score*. In order to calculate these measures, the number of humans detected in each frame is compared to the real number of humans in that frame.

   The scenario where the tests take place is described in Chapter 5 in Section 5.1, where the scenario characteristics are detailed. An outdoor scenario is used, since it is usually subject to larger variations of temperature and illumination than an indoor environment. The chosen scenario includes shadow zones as well as a building on the background which temperature varies slower than the air in the scene. Sometimes this causes that the building thermal readings are confused with the temperature of the humans on the scene when they are placed in front of it, making them very hard to be distinguished from the building. A concrete platform also appears with similar thermal features which can be confused with those humans standing on it. Other elements of interest are a lighting post and some trees which can occlude partially or totally the humans present in the environment. The trees generate dark zones as well, and sometimes humans are hidden by their shadows.

   A total amount of 12 sequences were chosen to evaluate the correct performance of the system in the selected scenario, with temperature ranging from -2° to 33° Celsius. The videos were recorded in a great variety of situations, including the environment covered by snow, fog or a day with very hot temperature. Night and sunrise sequences were also recorded. This large number of conditions contributed to elaborate an accurate description of the system's adaptability to the different atmospheric and lighting conditions which are usual in an outdoor environment.

   The recorded sequences also include a great variety of human behaviors. The videos show humans running, walking, sitting and even laying on the ground. In each sequence there appear between 1 and 3 people, who frequently cross their paths, hide behind the scenario elements or remain in a meeting during a long period. These meetings can take place in the limits of the zone covered by both cameras, resulting in a higher difficulty for the human detection in those situations. Thus, the versatility of the system is evaluated to analyze its performance in

situations of variable difficulty which can appear in a non-controlled environment.

The test results are described in Section 5.4. The achieved final outputs confirm the system's ability do provide excellent results independently of the conditions of the scene. While human detection algorithms in the infrared and visible spectra only show around 75% of sensitivity due to the fact that both spectra worsen their performance under different conditions, the use of fusion between both spectra rises this metric to 92%, guaranteeing a good system performance that is not affected by the conditions of the environment. The system precision is also rated at a 93%, with this measure at the same value of the infrared human detection measure. These data confirm the previously commented system's versatility, demonstrating that fusion is always an improvement and a solid alternative to the use of a single spectrum by its own. These measures also validate the interest of using a fusion system to overcome the different spectrum weaknesses and to enhance their strong points.

6.2. Future Work

A number of new research lines have arisen as a result of the developed work. These suggested new works involve the extension of the system and its functionalities, as well as the exploration of new approaches to test the adaptability of the system and its performance under new environments and situations.

- **Development and implementation of additional segmentation algorithms**. Different approaches for human detection have been developed during the course of this research. Yet, new segmentation proposals can be developed and implemented in the future. These algorithms can be compared to our current approaches in order to establish which approaches would be more suitable for other scenarios.

- **Addition of new test scenarios**. The current scenario was chosen due to its varying atmospherical and lighting conditions. However, it would be interesting to test our approach in different scenarios to validate it. These new scenes should include from more stable indoor environments to new outdoor situations such as crowded scenes or places which are not surrounded by buildings and, thus, where weather conditions change faster.

- **Addition of new sensors**. The current fusion approach only uses the information provided by video cameras. Since the system is based on the *INT*^{3}*-Horus* framework, it should not be difficult to add new kinds of sensors such as heat detectors or motion detectors. These sensors could provide feedback for the data provided by the cameras and help to improve the accuracy of the confidence levels as well as the parameterizations of the segmentation algorithms.

- **Enhancement of the validation methods**. As it was commented in Chapter 3, our validation system is only based on the comparison between the number of humans detected in each frame and the actual number of people in that image. An interesting study would involve a labeling of

the different humans in the scene including their location coordinates. These coordinates could be compared to the coordinates set by the system to establish if the fusion process improves the accuracy of the humans' situation in a scene. Another interesting approach is to trace the humans' silhouettes and compare them to a ground truth as well. The new ground truth should be generated in XML format to facilitate the development of an automatic validation tool.

6.3. Publications

Through the development of this dissertation, a series of publications have been achieved as a result of the different key ideas developed.

6.3.1. Journal Papers

1. Antonio Fernández-Caballero, María T. López, Juan Serrano-Cuerda & José Carlos Castillo (2013). Color video segmentation by lateral inhibition in accumulative computation. Signal, Image and Video Processing. Springer. ISSN 1863-1703. Submitted.

   Abstract: The lateral inhibition in accumulative computation (LIAC) algorithm has proved to be an efficient method for moving object segmentation in grey-level video sequences. This paper reviews the main steps and features of the LIAC algorithm, and assesses the suitability of applying the LIAC algorithm to the segmentation of color videos. Two widely used color spaces, namely RGB and HLS, are used for validating the LIAC algorithm, and a comparison is provided after performance evaluation of the algorithm in both color spaces.

2. Antonio Fernández-Caballero, Marina V. Sokolova & Juan Serrano-Cuerda (2013). Lateral inhibition in accumulative computation and fuzzy sets for human fall pattern recognition in colour and infrared imagery. The Scientific World Journal, volume 2013, article ID 935026, 10 pages, http://dx.doi.org/10.1155/2013/935026. ISSN 1537-744X.

   Abstract: Fall detection is an emergent problem in pattern recognition. In this paper a novel approach which enables to identify a type of a fall and reconstruct its characteristics is presented. The features detected include the position previous to a fall, the direction and velocity of a fall, and the post-fall inactivity. Video sequences containing a possible fall are analyzed image by image using the lateral inhibition in accumulative computation method. With this aim the region of interest of human figures are examined in each image, and geometrical and kinematic characteristics for the sequence are calculated. The approach is valid in color as well as in infrared video.

3. Antonio Fernández-Caballero, José Carlos Castillo, María T. López, Juan Serrano-Cuerda & Marina V. Sokolova (2013). INT3-Horus framework for multispectrum activity interpretation in intelligent environments. Expert Systems with Applications 40 (17), pp. 6715-6727. ISSN 0957-4174.

Abstract: The *INT*^{3}*-Horus* framework, dedicated to monitoring and activity interpretation in intelligent environments is introduced. Firstly, the paper introduces a general description of the *INT*^{3}*-Horus* approach. The following aspects of the proposal are highlighted: the framework is multisensory by nature and includes information fusion abilities; it is based on the model-view--controller paradigm; it is defined as a hybrid distributed system; it incorporates a Common Model that houses the data structures to support the exchange of information between levels of the framework. Then, the *INT*^{3}*-Horus* framework ontological model is introduced. The ontology is composed of a couple of classes, namely the Level Class and the DataType Class. The paper also describes the relations between both classes, as well as it introduces the notion of set of rules which determine the system functionality for a given domain. Lastly, a case of study on elderly fall detection is described to show the efficiency of the proposed framework.

4. José Carlos Castillo, Davide Carneiro, Juan Serrano-Cuerda, Paulo Novais, Antonio Fernández-Caballero & José Neves (2013). A multi-modal approach for activity classification and fall detection. International Journal of Systems Science. Taylor & Francis. ISSN 0020-7721.

   Abstract: The society is changing towards a new paradigm in which an increasing number of old adults live alone. In parallel, the incidence of conditions that affect mobility and independence is also rising as a consequence of a longer life expectancy. In this paper, the specific problem of falls of old adults is addressed by devising a technological solution for monitoring these users. Video cameras, accelerometers and GPS sensors are combined in a multi-modal approach to monitor humans inside and outside the domestic environment. Machine learning techniques are used to detect falls and classify activities from accelerometer data. Video feeds and GPS are used to provide location inside and outside the domestic environment. It results in a monitoring solution that does not imply the confinement of the users to a closed environment.

5. Marina V. Sokolova, Juan Serrano-Cuerda, José Carlos Castillo & Antonio Fernández-Caballero (2013). Fuzzy model for human fall detection in infrared video. Journal of Intelligent & Fuzzy Systems 24, pp. 215-228. Special Issue: Recent Advances in Intelligent & Fuzzy Systems. IOS Press. ISSN 1064-1246.

   Abstract: Fall detection, especially for elderly people, is a challenging problem which demands new products and technologies. In this paper a fuzzy model for fall detection and inactivity monitoring in infrared video is presented. The classification features proposed include geometric and kinematic parameters associated with more or less sudden changes in the tracked human-related regions of interest. A complete segmentation and tracking algorithm for infrared video as well as a fuzzy fall detection and confirmation algorithm are introduced. The proposed system is capable of identifying true and false falls, enhanced with inactivity monitoring aimed at confirming the need for medical assistance and/or care. The fall indicators used as well as their fuzzy model is explained in detail. The fuzzy model has been tested for a wide number of static and dynamic falls, demonstrating exciting initial results.

6. Antonio Fernández-Caballero, José Carlos Castillo, Juan Serrano-Cuerda & Saturnino Maldonado-Bascón (2011). Real-time human segmentation in infrared videos. Expert Systems

with Applications 38 (3), pp. 2577-2584. Elsevier Science. ISSN 0957-4174.

Abstract: In this paper, a new approach to real-time people segmentation through processing images captured by an infrared camera is introduced. The approach starts detecting human candidate blobs processed through traditional image thresholding techniques. Afterwards, the blobs are refined with the objective of validating the content of each blob. The question to be solved is if each blob contains one single human candidate or more than one. If the blob contains more than one possible human, the blob is divided to fit each new candidate in height and width.

6.3.2. Book Chapters

1. Juan Serrano-Cuerda, José Carlos Castillo, Marina V. Sokolova & Antonio Fernández-Caballero (2013). Efficient people counting from indoor overhead video camera. Advances in Intelligent and Soft Computing 221, pp. 129-137. Springer-Verlag.

   Abstract: This article introduces a system for real-time people counting. People counting is a challenging topic in the surveillance domain. The proposed system is built from *INT*^{3}*-Horus*, a multi-agent based framework for intelligent monitoring and activity interpretation. The system uses an indoor overhead video camera that detects people moving freely in a hall or room. The people counting system is flexible in detecting individuals as well as groups. Counting is independent of the trajectories and possible occlusions of the humans present in the scene. The initial results offered by the system are very promising in terms of specificity, sensitivity and F-score.

2. Marina V. Sokolova, José Carlos Castillo, Antonio Fernández-Caballero & Juan Serrano-Cuerda (2012). Intelligent monitoring and activity interpretation framework - INT3-Horus ontological model. Advances in Knowledge-Based and Intelligent Information and Engineering Systems, pp. 980-989. 10-12 Septiembre 2012, San Sebastian (Spain).

   Abstract: The *INT*^{3}*-Horus* framework, dedicated to intelligent monitoring and activity interpretation with special application in advanced surveillance systems, is introduced. *INT*^{3}*-Horus* is presented in two parts. This paper introduces the second part of the description of the framework. Here we introduce the framework ontological model. The ontology is composed of a couple of classes, namely the Level Class and the DataType Class. The paper also describes the relations between both classes, as well as it introduces the notion of set of rules which determine the system functionality for a given domain. The first part introduces the general description of the framework.

6.3.3. Publications in LNCS/LNAI series

1. Pablo Tribaldos, Juan Serrano-Cuerda, María T. López, Antonio Fernández-Caballero and Roberto J. López-Sastre (2013). People detection in color and infrared video using HOG and

linear SVM. 5th International Work-Conference on the Interplay between Natural and Artificial Computation, IWINAC 2013. Lecture Notes in Computer Science. Springer-Verlag. ISSN 0302-9743.

Abstract: This paper introduces a solution for detecting humans in smart spaces through computer vision. The approach is valid both for images in visible and infrared spectra. Histogram of oriented gradients (HOG) is used for feature extraction in the human detection process, whilst linear support vector machines (SVM) are used for human classification. A set of tests is conducted to find the classifiers which optimize recall in the detection of persons in visible video sequences. Then, the same classifiers are used to detect people in infrared video sequences obtaining excellent results.

2. Juan Serrano-Cuerda, Marina V. Sokolova, Antonio Fernández-Caballero, María T. López and José Carlos Castillo (2013). Fusion of overhead and lateral view video for enhanced people counting. 5th International Work-Conference on the Interplay between Natural and Artificial Computation, IWINAC 2013. Lecture Notes in Computer Science. Springer-Verlag. ISSN 0302-9743.

   Abstract: This article introduces a multi-camera system for real-time people counting. The proposed system is built from *INT*^{3}*-Horus* , a framework for intelligent monitoring and activity interpretation. The system uses an indoor overhead video camera and a lateral view video camera to detect people moving freely in smart spaces. The segmentation is performed from both synchronized input videos. Then, information is fused to enhance the overall efficiency. The people counting system is flexible in detecting individuals as well as groups. Also, people counting is independent of the trajectories and possible occlusions of the humans present in the smart space. The initial results offered are very promising.

3. José Manuel Gascueña, Antonio Fernández-Caballero, Elena Navarro, Juan Serrano-Cuerda & Francisco Alfonso Cano (2011). Agent-based development of multisensory monitoring systems. 4th International Work-Conference on the Interplay between Natural and Artificial Computation, IWINAC 2011. Lecture Notes in Computer Science, 6686, pp. 451-460. Springer-Verlag. ISSN 0302-9743.

   Abstract: This paper introduces the use of the VigilAgent agent methodology to develop monitoring systems. This work is based on the suitability of the specific characteristics of agency for developing monitoring systems. It is usual to develop them following an ad-hoc approach instead of using a methodology for achieving quality standards expected from commercial software. In this paper, the five phases of VigilAgent, namely System specification, Architectural Design, Detailed design, Implementation and Deployment, are introduced. The proposal is validated through the case study of controlling the access of human beings to a specific area.

4. José Carlos Castillo, Juan Serrano-Cuerda, Antonio Fernández-Caballero & María T. López (2009). Segmenting humans from mobile thermal infrared imagery. 3rd International Work-conference on the Interplay between Natural and Artificial Computation, IWINAC 2009, Lec-

ture Notes in Computer Science, 5602, pp. 225-234. Springer-Verlag. ISSN 0302-9743.

Abstract: Perceiving the environment is crucial in any application related to mobile robotics research. In this paper, a new approach to real-time human detection through processing video captured by a thermal infrared camera mounted on the indoor autonomous mobile platform mSecurit is introduced. The approach starts with a phase of static analysis for the detection of human candidates through some classical image processing techniques such as image normalization and thresholding. Then, the proposal uses Lukas and Kanade optical flow without pyramids algorithm for filtering moving foreground objects from moving scene background. The results of both phases are compared to enhance the human segmentation by infrared camera. Indeed, optical flow will emphasize the foreground moving areas gotten at the initial human candidates detection.

6.3.4. Conference Papers

1. José Carlos Castillo, Juan Serrano-Cuerda, Marina V. Sokolova and Antonio Fernández-Caballero (2012). Multispectrum video for proactive response in intelligent environments. The 8th International Conference on Intelligent Environments, IE'12. 26-29 June 2012, Guanajuato (México).

   Abstract: The exponential increase of home-bound persons that live alone and are in need of continuous monitoring requires new solutions to current problems. Most of these cases present illnesses, such as motor or psychological disabilities, that deprive them of a normal living. Abnormal situations such as forgetfulness or falls are quite common and should be prevented or dealt with. This paper presents a system able to detect dangerous situations at home, such as falls, independently from existing environment conditions. The aim of the proposed system is to proactively offer support to the citizen or to warn the emergency services when needed.

2. Antonio Fernández-Caballero, Marina V. Sokolova, Juan Serrano-Cuerda, José Carlos Castillo, Verónica Moreno, Rodrigo Castiñeira & Luis Redondo (2012). HOLDS: Efficient elderly fall detection through accelerometers and computer vision. The 8th International Conference on Intelligent Environments, IE'12. 26-29 June 2012, Guanajuato (México).

   Abstract: This paper introduces the technical description of ICT R&D project Fall Detection - HOLDS. The purpose of this project is to quickly assist elderly people or people who have a cognitive or motor disability when they suffer a fall at outdoor and indoor spaces. The project is aimed at enhancing the safety of these target groups and allows the elderly and impaired to carry out a more normal lifestyle. The HOLDS project is based on currently rising technologies, namely wireless sensors & actuators networks (WSAN) combined with advanced image processing. The HOLDS system comprises two subsystems, accelerometer-based fall detection and computer-vision-based (visible and infrared) fall detection, which are collected in a central system.

3. Juan Serrano-Cuerda, María Teresa López & Antonio Fernández-Caballero (2011). Robust

human detection and tracking in intelligent environments by information fusion of color and infrared video. The 7th International Conference on Intelligent Environments, IE'11, 25-28 July 2011, Nottingham (United Kingdom).

Abstract: This paper is related to ambient intelligence systems capable of locating and tracking humans. These are the first steps of a human-centered ambient intelligent system, ranging from data acquisition to robust tracking, for the purpose of interpreting human behaviors in monitored environments. The first objective is to improve human detection through the fusion of thermal-infrared and color video segmentation. On the level following to segmentation, the traditional tracking problems (e.g. occlusions, crossings,etc.) are faced. Finally, the use of several classifiers such as support vector machines and artificial neural networks are proposed to enhance the tracking level. The work proposes a combination of both color and thermal-infrared spectra in human tracking.

4. Francisco Alfonso Cano, José Carlos Castillo, Juan Serrano-Cuerda & Antonio Fernández-Caballero (2011). Multisensory architecture for intelligent surveillance systems - Integration of segmentation, tracking and activity analysis. 13th International Conference on Enterprise Information Systems, ICEIS 2011, 8-11 June 2011, Beijing (China), vol. AIDSS, pp. 157-162.

   Abstract: Intelligent surveillance systems deal with all aspects of threat detection in a given scene; these range from segmentation to activity interpretation. The proposed architecture is a step towards solving the detection and tracking of suspicious objects as well as the analysis of the activities in the scene. It is important to include different kinds of sensors for the detection process. Indeed, their mutual advantages enhance the performance provided by each sensor on its own. The results of the multisensory architecture offered in the paper, obtained from testing the proposal on CAVIAR project data sets, are very promising within the three proposed levels, that is, segmentation based on accumulative computation, tracking based on distance calculation and activity analysis based on finite state automaton.

6.4. Research projects

During the research of this dissertation, some projects were carried out in parallel to the development of the PhD thesis. Without their funding and the knowledge acquired during their development, this work would not have been possible.

6.4.1. National Projects

- Multisensor data fusion in complex and dynamic environments: Bio-inspired methods, intelligent agent-based architectures, and multimodal and augmented interfaces
  REFERENCE: TIN2010-20845-C03-01
  FINANCIAL ENTITY: Ministerio de Economía y Competitividad
  PARTICIPANTS: Universidad de Castilla-La Mancha

HEAD RESEARCHER UCLM: Antonio Fernández-Caballero
DURATION: 01/01/2011 - 31/12/2013
FUND: 75.700,00 €

- Fall detection system (HOLDS)
  REFERENCE: TSI-020100-2010-261
  FINANCIAL ENTITY: Ministerio de Industria, Energía y Turismo
  PARTICIPANTS: Ingeniería de Sistemas Intensivos en Software S.L., Residencia San Juan de Dios S.A., Métodos y Tecnología de Sistemas y Procesos S.L., Universidad de Castilla-La Mancha
  HEAD RESEARCHER UCLM: Antonio Fernández Caballero
  DURATION: 28/10/2010 - 31/12/2011

6.4.2. Regional Projects

- MultimodalSensors
  REFERENCE: HITO-09-106
  FINANCIAL ENTITY: Junta de Comunidades de Castilla-La Mancha (Consejería de Educación y Ciencia)
  PARTICIPANTS: SECISA Seguridad S.A., Universidad de Castilla-La Mancha
  HEAD RESEARCHER UCLM: Antonio Fernández Caballero
  DURATION: 01/05/2010 - 30/09/2010

- Model-driven engineering for the development of multisensory surveillance systems (IMSIS)
  REFERENCE: PII2I09-0069-0994
  FINANCIAL ENTITY: Junta de Comunidades de Castilla-La Mancha, Consejería de Educación y Ciencia
  PARTICIPANTS: Universidad de Castilla-La Mancha
  HEAD RESEARCHER: Antonio Fernández-Caballero
  DURATION: 01/04/2009 - 31/03/2012
  FUND: 90.000,00 €

- Real-time multisensory information fusion for advanced surveillance (FUSVIS)
  REFERENCE: PII2I09-0071-3947
  FINANCIAL ENTITY: Junta de Comunidades de Castilla-La Mancha, Consejería de Educación y Ciencia
  PARTICIPANTS: Universidad de Castilla-La Mancha
  HEAD RESEARCHER: María Teresa López Bonal
  DURATION: 01/04/2009 - 31/03/2010
  FUND: 15.655,50 €

6.4.3. Projects with Enterprizes

- Robot MR10: Intelligent surveillance systems based on thermal camera
  REFERENCE: CTR08-0133
  FINANCIAL ENTITY: MoviRobotics, S.L.
  PARTICIPANTS: Universidad de Castilla-La Mancha
  HEAD RESEARCHER: Antonio Fernández-Caballero
  DURATION: 01/07/2008 - 30/04/2009
  FUND: 22.991,53 €

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ali, A. and Aggarwal, J. (2001). Segmentation and recognition of continuous human activity. In *Proceedings of the 2001 IEEE Workshop on Detection and Recognition of Events in Video*, pages 28–35.

Ardeshir Goshtasby, A. and Nikolov, S. (2007). Guest editorial: Image fusion: Advances in the state of the art. *Information Fusion*, 8(2):114–118.

Armingol, J. M., de la Escalera, A., Hilario, C., Collado, J. M., Carrasco, J. P., Flores, M. J., Pastor, J. M., and Rodríguez, F. J. (2007). Ivvi: Intelligent vehicle based on visual information. *Robotics and Autonomous Systems*, 55(12):904–916.

Atsushi, N., Hirokazu, K., Shinsaku, H., and Seiji, I. (2002). Tracking multiple people using distributed vision systems. In *IEEE International Conference on Robotics and Automation, 2002. Proceedings. ICRA'02.*, volume 3, pages 2974–2981.

Baber, J., Afzulpurkar, N., and Satoh, S. (2013). A framework for video segmentation using global and local features. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(5):29 pages.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Surf: speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359.

Bellot, D., Boyer, A., and Charpillet, F. (2002). A new definition of qualified gain in a data fusion process: Application to telemedicine. In *Proceedings of the Fifth International Conference on Information Fusion.*

Benezeth, Y., Emile, B., Laurent, H., and Rosenberger, C. (2008). A real time human detection system based on far infrared vision. In Elmoataz, A., Lezoray, O., Nouboud, F., and Mammass, D., editors, *Image and Signal Processing*, volume 5099 of *Lecture Notes in Computer Science*, pages 76–84. Springer Berlin / Heidelberg.

Benezeth, Y., Emile, B., Laurent, H., and Rosenberger, C. (2010). Vision-based system for human detection and tracking in indoor environment. *International Journal of Social Robotics*, 2(1):41–52. 10.1007/s12369-009-0040-4.

Bertozzi, M., Broggi, A., Caraffi, C., Rose, M. D., Felisa, M., and Vezzoni, G. (2007). Pedestrian detection by means of far-infrared stereo vision. *Computer Vision and Image Understanding*, 106(2-3):194–204.

Beucher, S. (1991). The watershed transformation applied to image segmentation. In *10th Pfefferkorn Conf. on Signal and Image Processing in Microscopy and Microanalysis. Proceedings*, pages 299–314.

Birchfield, S. T., Natarajan, B., and Tomasi, C. (2007). Correspondence as energy-based segmentation. *Image and Vision Computing*, 25(8):1329–1340.

Blake, A. and Isard, M. (1998). *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer-Verlag New York, Inc.

Bobick, A. and Davis, J. (1996). Real-time recognition of activity using temporal templates. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision*, pages 39–42. IEEE.

Brunn, A., Lang, F., and Förstner, W. (1996). A procedure for segmenting surfaces by symbolic and iconic image fusion. In *Mustererkennung 1996, 18. DAGM-Symposium*, pages 11–20. Springer-Verlag.

Bugeau, A. and Pérez, P. (2008). Track and cut: simultaneous tracking and segmentation of multiple objects with graph cuts. *Journal on Image and Video Processing*, 2008:article id. 317278.

Burak Ozer, I. and Wolf, W. (2002). A hierarchical human detection system in (un)compressed domains. *IEEE Transactions on Multimedia*, 4(2):283–300.

Cabido, R., Montemayor, A., Pantrigo, J., Martínez-Zarzuela, M., and Payne, B. (2012). High-performance template tracking. *Journal of Visual Communication and Image Representation*, 23(2):271–286.

Caetano, T. and Barone, D. (2001). A probabilistic model for the human skin color. In *Proceedings of the 11th International Conference on Image Analysis and Processing*, pages 279–283.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

Cao, D., Masoud, O. T., Boley, D., and Papanikolopoulos, N. (2009). Human motion recognition using support vector machines. *Computer Vision and Image Understanding*, 113(10):1064–1075.

Capellades, M., Doermann, D., DeMenthon, D., and Chellappa, R. (2003). An appearance based approach for human and object tracking. In *International Conference on Image Processing Proceedings.*, volume 2, pages 85–88. IEEE.

Carmona, E. J., Martínez-Cantos, J., and Mira, J. (2008). A new video segmentation method of moving objects based on blob-level knowledge. *Pattern Recognition Letters*, 29(3):272–285.

Carneiro, D., Castillo, J. C., Novais, P., Fernández-Caballero, A., and Neves, J. (2012). Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications*, 39(18):13376–13389.

Castillo, J. C. (2012). *INT3-Horus: A Multilevel Framework for Intelligent Multisensor Monitoring and Activity Interpretation*. PhD thesis, Universidad de Castilla-La Mancha.

Castillo, J. C., Carneiro, D., Serrano-Cuerda, J., Novais, P., Fernández-Caballero, A., and Neves, J. (2013). A multi-modal approach for activity classification and fall detection. *International Journal of Systems Science*.

Cavallaro, A. and Ebrahimi, T. (2001). Video Object Extraction based on Adaptive Background and Statistical Change Detection. In *Proceedings of . SPIE Electronic Imaging 2001-Visual Communications and Image Processing*, volume 4310, pages 465–475. SPIE.

Chang, S.-L., Yang, F.-T., Wu, W.-P., Cho, Y.-A., and Chen, S.-W. (2011). Nighttime pedestrian detection using thermal imaging based on hog feature. In *International Conference on System Science and Engineering (ICSSE)*, pages 694–698.

Chen, H., Arora, M., and Corresponding, P. (2003). Mutual information-based image registration for remote sensing data. *International Journal of Remote Sensing*, 24(18):3701–3706.

Chouteau, J., Lerat, J., Testa, R., Moyen, B., and Banks, S. (2007). Effects of radiograph projection parameter uncertainty on tka kinematics from model-image registration. *Journal of Biomechanics*, 40(16):3744–3747.

Cielniak, G. and Duckett, T. (2004). People recognition by mobile robots. *Journal of Intelligent and Fuzzy Systems*, 15:21–27.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. (1995). Automated multi-modality image registration based on information theory. In *Information processing in medical imaging*, volume 3, pages 264–274.

Collomosse, J. and Wang, T. (2012). Probabilistic motion diffusion of labeling priors for coherent video segmentation. *IEEE Transactions on Multimedia*, 14(2):389–400.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577.

Correa, N., Li, Y.-O., Adali, T., and Calhoun, V. (2008). Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. *IEEE Journal of Selected Topics in Signal Processing*, 2(6):998–1007.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Costa, Â., Castillo, J. C., Novais, P., Fernández-Caballero, A., and Simoes, R. (2012). Sensor-driven agenda for intelligent home care of the elderly. *Expert Systems with Applications*, 39(15):12192–12204.

Cover, T., Thomas, J., Wiley, J., et al. (1991). *Elements of information theory*, volume 6. Wiley Online Library.

Cucchiara, R., Grana, C., Tardini, G., and Vezzani, R. (2004). Probabilistic people tracking for occlusion handling. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 132–135. IEEE.

da Cunha, A. L., Zhou, J., and Do, M. N. (2006). The nonsubsampled contourlet transform: Theory, design, and applications. *IEEE Transactions on Image Processing*, 15(10):3089–3101.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE Computer Society.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Davis, J. and Sharma, V. (2004). Robust detection of people in thermal imagery. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4, pages 713–716.

Davis, J. W. and Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2):162–182.

Davis, L., Philomin, V., and Duraiswami, R. (2000). Tracking humans from a moving platform. In *Proceedings of 15th International Conference on Pattern Recognition, 2000*, volume 4, pages 171–178.

Delgado, A. E., López, M. T., and Fernández-Caballero, A. (2010). Real-time motion detection by lateral inhibition in accumulative computation. *Engineering Applications of Artificial Intelligence*, 23(1):129–139.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Devarajan, D., Cheng, Z., and Radke, R. (2008). Calibrating distributed camera networks. *Proceedings of the IEEE*, 96(10):1625–1639.

DoD, U. S. D. o. D. (1991). Data fusion subpanel of the joint directors of laboratories, technical panel for c3, in: Data fusion lexicon.

DoD, U. S. D. o. D. (1994). Dsto, (defense science and technology organization) data fusion special interest group, in: Data fusion lexicon.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. (2002). Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163.

EMSI, Electromagnetic Systems Incorporated, I. (2006-2007). Tracking a large target. `http://trackertechnology.com/largetarget.html`. accessed:09-23-2013.

Fablet, R., Bouyhemy, P., and Gelgon, M. (1999). Moving object detection in color image sequences using region-level graph labeling. In *International Conference on Image Processing. Proceedings.*, volume 2, pages 939–943.

Fan, X., Xu, L., Zhang, X., and Chen, L. (2008). The research and application of human detection based on support vector machine using in intelligent video surveillance system. In *Fourth International Conference on Natural Computation*, volume 2, pages 139–143.

Fang, Y., Yamada, K., Ninomiya, Y., Horn, B., and Masaki, I. (2004). A shape-independent method for pedestrian detection with far-infrared images. *IEEE Transactions on Vehicular Technology*, 53(6):1679–1697.

Fernández-Caballero, A., Castillo, J. C., and Rodríguez-Sánchez, J. M. (2012). Human activity monitoring by local and global finite state machines. *Expert Systems with Applications*, 39(8):6982–6993.

Fernández-Caballero, A., Castillo, J. C., López, M. T., Serrano-Cuerda, J., and Sokolova, M. V. (2013). Int3-horus framework for multispectrum activity interpretation in intelligent environments. *Expert Systems with Applications*, 40(17):6715–6727.

Fernández-Caballero, A., Castillo, J. C., Martínez-Cantos, J., and Martínez-Tomás, R. (2010). Optical flow or image subtraction in human detection from infrared camera on mobile robot. *Robotics and Autonomous Systems*, 58(12):1273–1281.

Fernández-Caballero, A., Castillo, J. C., Serrano-Cuerda, J., and Maldonado-Bascón, S. (2011a). Real-time human segmentation in infrared videos. *Expert Systems with Applications*, 38(3):2577–2584.

Fernández-Caballero, A., Fernández, M. A., Mira, J., and Delgado, A. E. (2003). Spatio-temporal shape building from image sequences using lateral interaction in accumulative computation. *Pattern Recognition*, 36(5):1131–1142.

Fernández-Caballero, A., López, M. T., Carmona, E. J., and Delgado, A. E. (2011b). A historical perspective of algorithmic lateral inhibition and accumulative computation in computer vision. *Neurocomputing*, 74(8):1175–1181.

Fernández-Caballero, A., López, M. T., and Saiz-Valverse, S. (2008). Dynamic stereoscopic selective visual attention (dssva): Integrating motion and shape with depth in video segmentation. *Expert Systems with Applications*, 34(2):1394–1402.

Fernández-Caballero, A., Mira, J., Fernández, M. A., and López, M. T. (2001). Segmentation from motion of non-rigid objects by neuronal lateral interaction. *Pattern Recognition Letters*, 22(14):1517–1524.

Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282.

Fortmann, T. E., Bar-Shalom, Y., and Scheffe, M. (1983). Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, 8(3):173–184.

Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pages 23–37. Springer-Verlag.

Friedman, N. and Russell, S. (1997). Image segmentation in video sequences: a probabilistic approach. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 175–181. Morgan Kaufmann Publishers Inc.

Frintrop, S., Königs, A., Hoeller, F., and Schulz, D. (2010). A component-based approach to visual person tracking from a mobile platform. *International Journal of Social Robotics*, 2(1):53–62.

Fu, Z. and Han, Y. (2012). Centroid weighted kalman filter for visual object tracking. *Measurement*, 45(4):650–655.

Fuentes, L. M. and Velastin, S. A. (2006). People tracking in surveillance applications. *Image and Vision Computing*, 24(11):1165–1171.

Gascueña, J. M., Castillo, J. C., Navarro, E., and Fernández-Caballero, A. (2013). Engineering the development of systems for multisensory monitoring and activity interpretation. *International Journal of Systems Science*.

Gascueña, J. M. and Fernández-Caballero, A. (2011). Agent-oriented modeling and development of a person-following mobile robot. *Expert Systems with Applications*, 38(4):4280–4290.

Gascueña, J. M. and Fernández-Caballero, A. (2011). On the use of agent technology in intelligent, multisensory and distributed surveillance. *The Knowledge Engineering Review*, 26(02):191–208.

Gascueña, J. M., Fernández-Caballero, A., López, M. T., and Delgado, A. E. (2011). Knowledge modeling through computational agents: application to surveillance systems. *Expert Systems*, 28(4):306–323.

Gascueña, J. M., Navarro, E., and Fernández-Caballero, A. (2012). Model-driven engineering techniques for the development of multi-agent systems. *Engineering Applications of Artificial Intelligence*, 25(1):159–173.

Gemma and Piella (2003). A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 4(4):259–280.

Ghaeminia, M., Shabani, A., and Shokouhi, S. (2010). Adaptive motion model for human tracking using particle filter. In *20th International Conference on Pattern Recognition*, pages 2073–2076.

Gibson, J. (1950). *The Perception of the Visual World*. Riverside Press, Cambridge.

Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing 3rd Ed.* Prentice-Hall, Englewood Cliffs, NJ.

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F. Radar and Signal Processing*, 140(2):107–113.

Goubet, E., Katz, J., and Porikli, F. (2006). Pedestrian tracking using thermal infrared imaging. *Infrared Technology and Applications XXXII*, 6206:797–808.

Gouet-Brunet, V. and Lameyre, B. (2008). Object recognition and segmentation in videos by connecting heterogeneous visual features. *Computer Vision and Image Understanding*, 111(1):86–109.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Toward principles for the design of ontologies used for knowledge sharing?*, 43:907–928.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69(3):331–371.

Han, J. and Bhanu, B. (2007). Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784.

Haritaoglu, I., Beymer, D., and Flickner, M. (2002). Ghost3d: Detecting body posture and parts using stereo. In *Proceedings of the Workshop on Motion and Video Computing*, pages 175–180. IEEE Computer Society.

Haritaoglu, I., Harwood, D., and Davis, L. (2000). W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830.

Harmouche, R., Cheriet, F., Labelle, H., and Dansereau, J. (2010). Articulated model registration of mri/x-ray spine data. In *Image Analysis and Recognition*, volume 6112 of *Lecture Notes in Computer Science*, pages 20–29. Springer Berlin / Heidelberg.

Hayashi, K., Hashimoto, M., Sumi, K., and Sasakawa, K. (2004). Multiple-person tracker with a fixed slanting stereo camera. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 681–686.

Hernández-Vela, A., Reyes, M., Ponce, V., and Escalera, S. (2012). Grabcut-based human segmentation in video sequences. *Sensors*, 2012(12):15376–15393.

Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17:185–203.

Horprasert, T., Harwood, D., and Davis, L. S. (1999). A statistical approach for real-time robust background subtraction and shadow detection. In *Proceedings of the IEEE ICCV*, pages 1–19.

Hsiao, H.-H. and Leou, J.-J. (2013). Background initialization and foreground segmentation for bootstrapping video sequences. *EURASIP Journal on Image and Video Processing*, 2013(12).

Hu, M., Hu, W., and Tan, T. (2004). Tracking people through occlusions. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 724–727. IEEE.

Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., and Maybank, S. (2006). Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):663–671.

Hu, X., Chakravarty, S., She, Q., and Wang, B. (2013). A modified hierarchical graph cut based video segmentation approach for high frame rate video. *Proceedings of the SPIE*, 8661.

Huang, J., Ravi Kumar, S., Mitra, M., Zhu, W., and Zabih, R. (1999). Spatial color indexing and applications. *International Journal of Computer Vision*, 35(3):245–268.

Huwer, S. and Niemann, H. (2000). Adaptive change detection for real-time surveillance applications. In *Proceedings of the Third IEEE International Workshop on Visual Surveillance*, pages 37–46.

Image Metrology (2013). Watershed segmentation. `http://www.imagemet.com/`.

Improved Outcome Software (2013). Tutorial 9: Support vector machines. introduction. `http://www.improvedoutcomes.com/`.

Isard, M. and Blake, A. (1998). Condensation−conditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28.

Ivanov, Y., Bobick, A., and Liu, J. (2000). Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207.

Iwase, S. and Saito, H. (2004). Parallel tracking of all soccer players by integrating detected positions in multiple view images. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004*, volume 4, pages 751–754. IEEE.

Iwata, K., Satoh, Y., Yoda, I., and Sakaue, K. (2006). Hybrid camera surveillance system by using stereo omni-directional system and robust human detection. In Chang, L.-W. and Lie, W.-N., editors, *Advances in Image and Video Technology*, volume 4319 of *Lecture Notes in Computer Science*, pages 611–620. Springer Berlin / Heidelberg.

Iwata, K., Satoh, Y., Yoda, I., and Sakaue, K. (2008). Hybrid camera surveillance system using robust human detection. *Electronics and Communications in Japan*, 91(11):11–18.

Jain, A. and Ross, A. (2002). Fingerprint mosaicking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Jain, R., Kasturiy, R., and Schunck, B. G. (1995). *Machine Vision*. McGraw-Hill.

Jain, R. and Nagel, H. (1979). On the analysis of accumulative difference pictures from image sequence of real world scenes. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1(2):206–214.

Jang, J. and Ra, J. (2008). Pseudo-color image fusion based on intensity-hue-saturation color space. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2008*, pages 366–371.

Jepson, A., Fleet, D., and El-Maraghi, T. (2003). Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311.

Jharna, M. and Kiran, S. (2013). Human tracking using particle filter. *International Journal of Computer Applications*, 76(6):1–6.

KaewTraKulPong, P. and Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-Based Surveillance Systems*, pages 135–144. Springer.

Kage, H., Seki, M., Sumi, K., Tanaka, K., and Kyuma, K. (2007). Pattern recognition for video surveillance and physical security. In *SICE, 2007 Annual Conference*, pages 1823–1828.

Kale, A., Cuntoor, N., Yegnanarayana, B., Rajagopalan, A., and Chellappa, R. (2003). Gait analysis for human identification. In Kittler, J. and Nixon, M., editors, *Audio- and Video-Based Biometric Person Authentication*, volume 2688 of *Lecture Notes in Computer Science*, pages 1058–1058. Springer Berlin / Heidelberg.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering*, 82(Series D):35–45.

Kanade, T., Collins, R., Lipton, A., Burt, P., and Wixson, L. (1998). Advances in cooperative multi-sensor video surveillance. In *Darpa Image Understanding Workshop*, pages 3–24. Morgan Kaufmann.

Kang, J., Cohen, I., and Medioni, G. (2005). Persistent objects tracking across multiple non overlapping cameras. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 112–119. IEEE.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.

Kelly, P., O'Connor, N. E., and Smeaton, A. F. (2009). Robust pedestrian detection and tracking in crowded scenes. *Image and Vision Computing*, 27(10):1445–1458.

Khan, S. and Shah, M. (2000). Tracking people in presence of occlusion. In *Asian Conference on Computer Vision*, pages 1132–1137.

Kieran, D. and Yan, W. (2010). A framework for an event driven video surveillance system. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 97–102. IEEE.

Kim, J.-B. and Kim, H.-J. (2003). Multiresolution-based watersheds for efficient image segmentation. *Pattern Recognition Letters*, 24(1):473–488.

Klappstein, J., Vaudrey, T., Rabe, C., Wedel, A., and Klette, R. (2009). Moving object segmentation using optical flow and depth information. *Lecture Notes in Computer Science*, 5414:611–623.

Kohli, P., Rihan, J., Bray, M., and Torr, P. (2008). Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision*, 79(3):285–298.

Kohonen, T., Schroeder, M. R., and Huang, T. S., editors (2001). *Self-Organizing Maps*. Springer-Verlag New York, Inc., 3rd edition.

Konrad, J. (2000). Motion detection and estimation. In Bovik, A., editor, *Handbook of Image and Video Processing*, chapter 3.10, pages 207–225. Academic Press.

Koschan, A., Kang, S., Paik, J., Abidi, B., and Abidi, M. (2003). Color active shape models for tracking non-rigid objects. *Pattern Recognition Letters*, 24(11):1751–1765.

Krotosky, S. and Trivedi, M. (2006). Multimodal stereo image registration for pedestrian detection. In *Intelligent Transportation Systems Conference*, pages 109–114.

Krüger, V., Anderson, J., and Prehn, T. (2005). Probabilistic model-based background subtraction. *Image Analysis*, pages 567–576.

Kumar, P., Mittal, A., and Kumar, P. (2006). Fusion of thermal infrared and visible spectrum video for robust surveillance. In Kalra, P. and Peleg, S., editors, *Computer Vision, Graphics and Image Processing*, volume 4338 of *Lecture Notes in Computer Science*, pages 528–539. Springer Berlin / Heidelberg.

Kuno, Y., Watanabe, T., Shimosakoda, Y., and Nakagawa, S. (1996). Automated detection of human for visual surveillance system. In *ICPR '96: Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 865–869. IEEE Computer Society.

Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 878–885.

Lewis, J. J., O' Callaghan, R. J., Nikolov, S. G., Bull, D. R., and Canagarajah, N. (2007). Pixel- and region-based image fusion with complex wavelets. *Information Fusion*, 8(2):119–130.

Leykin, A. and Hammoud, R. (2010). Pedestrian tracking by fusion of thermal-visible surveillance videos. *Machine Vision and Applications*, 21(4):587–595.

Li, H., Manjunath, B. S., and Mitra, S. K. (1995). Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing*, 57(3):235–245.

Li, J. and Gong, W. (2010). Real time pedestrian tracking using thermal infrared imagery. *Journal of Computers*, 5(10).

Li, J., Gong, W., Li, W., and Liu, X. (2010). Robust pedestrian detection in thermal infrared imagery using the wavelet transform. *Infrared Physics & Technology*, 53(4):267–273.

Li, M., Zhang, Z., Huang, K., and Tan, T. (2009). Rapid and robust human detection and tracking based on omega-shape features. In *16th IEEE International Conference on Image Processing*, pages 2545–2548.

Li, S., Kang, X., Hu, J., and Yang, B. (2013). Image matting for fusion of multi-focus images in dynamic scenes. *Information Fusion*, 14(2):147–162.

Li, S. and Yang, B. (2008). Multifocus image fusion using region segmentation and spatial frequency. *Image and Vision Computing*, 26(7):971–979.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157.

López, M. T., Fernández-Caballero, A., Mira, J., Delgado, A. E., and Fernández, M. A. (2006). Algorithmic lateral inhibition method in dynamic and selective visual attention task: application to moving objects detection and labelling. *Expert Systems with Applications*, 31(3):570–594.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence-Volume 2*, pages 674–679. Morgan Kaufmann Publishers Inc.

Luo, R. C. and Kay, M. G. (1992). Data fusion and sensor integration: State of the art 1990s. In Abidi, M. A. and Gonzalez, R. C., editors, *Data Fusion in Robotics and Machine Intelligence*, pages 7–135. Academic Press, San Diego.

MacKay, D. J. C. (1998). Introduction to Monte Carlo methods. In Jordan, M. I., editor, *Learning in Graphical Models*, NATO Science Series, pages 175–204. Kluwer Academic Press.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198.

Martínez-Tomás, R., Fernández-Caballero, A., and Ferrández, J. M. (2013). Intelligent monitoring for people assistance and safety. *Expert Systems*, page In press.

McKenna, S., Jabri, S., Duric, Z., and Wechsler, H. (2000). Tracking interacting people. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 348–353.

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44:335.

Meytlis, M. and Sirovich, L. (2007). On the dimensionality of face space. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(7):1262–1267.

Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the 8th European Conference on Computer Vision*, volume 1, pages 69–82. Springer.

Mitchell, H. (2007). *Multi-Sensor Data Fusion. An Introduction*. Springer-Verlag.

Mitchell, H. B. (2010). *Image Fusion Theories, Techniques and Applications*. Springer.

Mittal, A. and Davis, L. (2003). M 2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision*, 51(3):189–203.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126.

Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:349–361.

Moreno-Garcia, J., Rodriguez-Benitez, L., Fernández-Caballero, A., and López, M. T. (2010). Video sequence motion tracking by fuzzification techniques. *Applied Soft Computing*, 10(1):318–331.

Muhammad Anwer, R., Vázquez, D., and López, A. (2011). Opponent colors for human detection. In Vitrià, J., Sanches, J., and Hernández, M., editors, *Pattern Recognition and Image Analysis*, volume 6669 of *Lecture Notes in Computer Science*, pages 363–370. Springer Berlin / Heidelberg.

Nakada, T., Kagami, S., and Mizoguchi, H. (2008). Pedestrian detection using 3d optical flow sequences for a mobile robot. In *Proceedings of the 2008 IEEE Sensors*, pages 776–779.

Nanda, H. and Davis, L. (2002). Probabilistic template based pedestrian detection in infrared videos. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 1, pages 15–20.

Navarro, E., Fernández-Caballero, A., and Martínez-Tomás, R. (2013). Intelligent multisensory systems in support of information society. *International Journal of Systems Science*.

Navon, E., Miller, O., and Averbuch, A. (2005). Color image segmentation based on adaptive local thresholds. *Image and Vision Computing*, 23(1):69–85.

Nikolov, S., Bull, D., Canagarajah, C., Halliwell, M., and Wells, P. (2000). 2-d image fusion by multiscale edge graph combination. In *Proceedings of the Third International Conference on Information Fusion.*, volume 1, pages MOD3/16–MOD3/22.

Nourbakhsh, I. R., Bobenage, J., Grange, S., Lutz, R., Meyer, R., and Soto, A. (1999). An affective mobile robot educator with a full-time job. *Artificial Intelligence*, 114(1-2):95–124.

O' Malley, R., Jones, E., and Glavin, M. (2010). Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation. *Infrared Physics and Technology*, 53(6):439–449.

Okuma, K., Taleghani, A., Freitas, N. D., Freitas, O. D., Little, J. J., and Lowe, D. G. (2004). A boosted particle filter: Multitarget detection and tracking. In *8th European Conference on Computer Vision*, pages 28–39.

OpenCV, h. p. (2013). `http://opencv.willowgarage.com/`.

Otsu, N. (1979). A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9:62–66.

Panin, G., Ladikos, A., and Knoll, A. (2006). An efficient and robust real-time contour tracking system. In *IEEE International Conference on Computer Vision Systems*, pages 44–51.

Pavón, J., Gómez-Sanz, J., Fernández-Caballero, A., and Valencia-Jiménez, J. J. (2007). Development of intelligent multisensor surveillance systems with agents. *Robotics and Autonomous Systems*, 55(12):892–903.

Pedersoli, M., Gonzàlez, J., Bagdanov, A. D., and Roca, X. (2011). Efficient discriminative multiresolution cascade for real-time human detection applications. *Pattern Recognition Letters*, 32(13):1581–1587.

Pérez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002). Color-based probabilistic tracking. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 661–675. Springer-Verlag.

Perperidis, D., Mohiaddin, R. H., and Rueckert, D. (2005). Spatio-temporal free-form registration of cardiac mr image sequences. *Medical Image Analysis*, 9(5):441–456.

Petrovic, V. and Xydeas, C. (2004). Gradient-based multiresolution image fusion. *IEEE Transactions on Image Processing*, 13(2):228–237.

Processing, S. and Communication Research Lab, L. U. (2008). System involving different types of multisensor fusion. `http://www.ece.lehigh.edu/SPCRL/IF/tank.gif`. accessed:09-23-2013.

Protégé, h. p. (2013). `http://protege.stanford.edu/`.

Pszczólkowski, S. and Soto, A. (2007). Human detection in indoor environments using multiple visual cues and a mobile robot. In *Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications*, CIARP'07, pages 350–359. Springer-Verlag.

Rabiner, L. and Juang, B. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Ratanamahatana, C. A.and Keogh, E. (2005). Three myths about dynamic time warping. In *Proceedings of the SIAM International Conference on Data Mining*.

Reenskaug, T. (1979). Thing-model-view-editor–an example from a planning system. technical note, xerox parc. Technical report.

Reid, D. (1979). An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843–854.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rivas, A., Martinez-Tomás, R., and Fernández-Caballero, A. (2011). Multiagent system for knowledge-based event recognition and composition. *Expert Systems*, 28(5):488–501.

Rockinger, O. (1996). Pixel level fusion of image sequences using wavelet frames. In *Proceedings of the 16th Leeds Annual Statistical Research Workshop*, pages 149–154. Leeds University Press.

Rodriguez, M. D. and Shah, M. (2007). Detecting and segmenting humans in crowded scenes. In *Proceedings of the 15th International Conference on Multimedia*, pages 353–356. ACM.

Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115.

Roth, D., Doubek, P., and Gool, L. (2005). Bayesian pixel classification for human tracking. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 78–83. IEEE.

Saeedi, J. and Faez, K. (2013). A classification and fuzzy-based approach for digital multi-focus image fusion. *Pattern Analysis and Applications*, 16(3):365–379.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49.

Schiele, B. (2006). Model-free tracking of cars and people based on color regions. *Image and Vision Computing*, 24(11):1172–1178.

Schwartz, W. R., Kembhavi, A., Harwood, D., and Davis, L. S. (2009). Human detection using partial least squares analysis. In *IEEE 12th International Conference on Computer Vision*, pages 24–31.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Serby, D., Meier, E., and Van Gool, L. (2004). Probabilistic object tracking using multiple features. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 184–187.

Serrano-Cuerda, J., Castillo, J., Sokolova, M., and Fernández-Caballero, A. (2013). Efficient people counting from indoor overhead video camera. In Pérez, J. B., Rodríguez, J. M. C., Fähndrich, J., Mathieu, P., Campbell, A., Suarez-Figueroa, M. C., Ortega, A., Adam, E., Navarro, E., Hermoso, R., and Moreno, M. N., editors, *Trends in Practical Applications of Agents and Multiagent Systems*, volume 221 of *Advances in Intelligent Systems and Computing*, pages 129–137. Springer International Publishing.

Serrano-Cuerda, J., Sokolova, M. V., Fernández-Caballero, A., López, M. T., and Castillo, J. C. (2012). Accumulative computation and fuzzy sets for robust fall detection in color video. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, pages 960–969.

Shafique, K. and Shah, M. (2005). A noniterative greedy algorithm for multiframe point correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):51–65.

Sharma, V. and Davis, J. W. (2009). Feature-level fusion for object segmentation using mutual information. In Hammoud, R. I., editor, *Augmented Vision Perception in Infrared*, Advances in Pattern Recognition, pages 295–320. Springer London.

Shi, J. and Malik, J. (1997). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2013). Real-time human pose recognition in parts from single depth images. In *Machine Learning for Computer Vision*, pages 119–135.

Siddiqi, M., Truc, P. T. H., Lee, S., and Lee, Y.-K. (2011). Automatic human body segmentation using level-set based active contours followed by optical flow in video surveillance. In *Proccedings of the IEEE International Conference on Intelligent Computer Communication and Processing*, pages 361–364.

Sim, R. (1998). *Mobile Robot Localisation Using Learned Landmarks*. PhD thesis, Department of Computer Sciente, MacGill University, Montréal, Canada.

Sokolova, M. V., Castillo, J. C., Fernández-Caballero, A., and Serrano-Cuerda, J. (2012). Intelligent monitoring and activity interpretation framework-int3-horus ontological model. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, pages 980–989.

Sokolova, M. V., Serrano-Cuerda, J., Castillo, J. C., and Fernández-Caballero, A. (2013). Fuzzy model for human fall detection in infrared video. *Journal of Intelligent and Fuzzy Systems*, 24(2):215–228.

Song, Y., Feng, X., and Perona, P. (2000). Towards detection of human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 810–817.

Starck, J. and Hilton, A. (2008). Model-based human shape reconstruction from multiple views. *Computer Vision and Image Understanding*, 111(2):179–194.

Suard, F., Guigue, V., Rakotomamonjy, A., and Benshrair, A. (2005). Pedestrian detection using stereo-vision and graph kernels. In *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*, pages 267–272.

Suard, F., Rakotomamonjy, A., Bensrhair, A., and Broggi, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 206–212.

Sun, S.-G. and Park, H. (2001). Segmentation of forward-looking infrared image using fuzzy thresholding and edge detection. *Optical Engineering*, 40:2638–2645.

Talukder, A. and Matthies, L. (2004). Real-time detection of moving objects from moving vehicles using dense stereo and optical flow. In *Proceedings of the 2004 International Conference on Intelligent Robots and Systems*, volume 4, pages 3718–3724.

Thompson, D. and Wettergreen, D. (2005). Multiple-object detection in natural scenes with multiple-view expectation maximization clustering. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 448–453.

Tomasi, C. and Kanade, T. (1991). *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ.

Tomasi, C. and Kanade, T. (1992). Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154.

Utsumi, A. and Tetsutani, N. (2002). Human detection using geometrical pixel value structures. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition. Proceedings*, pages 34–39.

Veenman, C., Reinders, M., and Backer, E. (2001). Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):54–72.

Vincenzo, R. and Lisa, U. (2007). An improvement of adaboost for face-detection with motion and color information. In *14th International Conference on Image Analysis and Processing*, pages 518–523.

Viola, P. and Wells III, W. (1997). Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154.

Wang, D. and Terman, D. (1997). Image segmentation based on oscillatory correlation. *Neural Computation*, 9:805–836.

Wang, J.-T., Chen, D.-B., Chen, H.-Y., and Yang, J.-Y. (2012a). On pedestrian detection and tracking in infrared videos. *Pattern Recognition Letters*, 33(6):775–785.

Wang, P., Tian, H., and Zheng, W. (2013). A novel image fusion method based on frft-nsct. *Mathematical Problems in Engineering*, 2013.

Wang, X. and Li, G. (2011). Fusion algorithm for infrared-visual image sequences. In *Sixth International Conference on Image and Graphics (ICIG)*, pages 244–248.

Wang, Z., Salah, M. B., Zhang, H., and Ray, N. (2012b). Shape based appearance model for kernel tracking. *Image and Vision Computing*, 30(4-5):332 – 344.

Weng, S.-K., Kuo, C.-M., and Tu, S.-K. (2006). Video object tracking using adaptive kalman filter. *Journal of Visual Communication and Image Representation*, 17(6):1190–1208.

Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005.*, volume 1, pages 90–97.

Wu, B. and Nevatia, R. (2006). Tracking of multiple, partially occluded humans based on static body part detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 1, pages 951–958.

Xiao, C., Gan, J., and Hu, X. (2013). Fast level set image and video segmentation using new evolution indicator operators. *The Visual Computer*, 29(1):27–39.

Xiong, Z. and Zhang, Y. (2010). A critical review of image registration methods. *International Journal of Image and Data Fusion*, 1(2):137–158.

Xu, F., Liu, X., and Fujimura, K. (2005). Pedestrian detection and tracking with night vision. *IEEE Transactions on Intelligent Transportation Systems*, 6(1):63–71.

Xu, L. and Puig, P. (2005). A hybrid blob-and appearance-based framework for multi-object tracking through complex occlusions. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 73–80.

Yang, C., Duraiswami, R., and Davis, L. (2005). Fast multiple object tracking via a hierarchical particle filter. In *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005.*, volume 1, pages 212–219.

Yang, D. B., González-Baños, H. H., and Guibas, L. J. (2003). Counting people in crowds with a real-time network of simple image sensors. In *Proceedings of the Ninth IEEE International Conference on Computer Vision-Volume 2*, ICCV '03, pages 122–129. IEEE Computer Society.

Yang, M.-T., Shih, Y.-C., and Wang, S.-C. (2004). People tracking by integrating multiple features. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4, pages 929–932.

Yeh, H., Chen, J., Huang, C., and Chen, C. (2010). An adaptive approach for overlapping people tracking based on foreground silhouettes. In *IEEE International Conference on Image Processing*, pages 3489–3492.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Computer Surveys*, 38(4).

Yilmaz, A., Li, X., and Shah, M. (2004). Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1531–1536.

Yu, Z., Wong, H.-S., and Wen, G. (2011). A modified support vector machine and its application to image segmentation. *Image and Vision Computing*, 29(1):29–40.

Zhang, L., Li, S., Yuan, X., and Xiang, S. (2007). Real-time object classification in video surveillance based on appearance learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Zhao, L. and Thorpe, C. (2000). Stereo-and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):148–154.

Zhao, T. and Nevatia, R. (2003). Bayesian human segmentation in crowded situations. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 459–466. IEEE.

Zhou, H., Yuan, Y., Zhang, Y., and Shi, C. (2009). Non-rigid object tracking in complex scenes. *Pattern Recognition Letters*, 30(2):98–102.

Zhou, J. and Hoang, J. (2005). Real time robust human detection and tracking system. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)-Workshops, 2005*, page 149.

Zhu, S. C. and Yuille, A. (1996). Region competition: Unifying snakes, region growing, and bayes/mdl for multi-band image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:884–900.

Zhu, Z. and Huang, T. (2007). *Multimodal Surveillance: Sensors, Algorithms and Systems*. Artech House Publishers.

Zitová, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000.

Zribi, M. (2010). Non-parametric and region-based image fusion with bootstrap sampling. *Information Fusion*, 11(2):85–94.