Fusión multicámara para seguimiento de objetos basada en Teoría Evidencial

por

Esteban Omar García Rodríguez
Lic. BUAP

Tesis sometida como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica
Febrero 2008
Tonantzintla, Puebla

Supervisada por:

Dr. Leopoldo Altamirano Robles
Investigador Titular del INAOE

Resumen

El uso de más de una cámara ha demostrado mejorar las capacidades de los sistemas de visión, ya sea incrementando el campo de vista o reduciendo la incertidumbre. Aún cuando en recientes trabajos se considera información como la posición donde se detiene un objeto o cuánto tiempo pasa sobre una región en particular, todavía no es común encontrar trabajo donde se utilicen múltiples cámaras para mejorar la información en los sistemas de visión.

En este trabajo se propone un modelo de fusión a nivel decisión para aprovechar cámaras distribuidas geográficamente con la finalidad de reducir la incertidumbre derivada de la perspectiva de las cámaras. Se consideran regiones previamente definidas para describir la posición de los objetos, y el proceso de integración de datos ocurre a un nivel en el que es útil para sistemas de vigilancia de alto nivel, tales como los que se usan en reconocimiento de comportamiento. En el modelo presentado, las decisiones individuales se toman usando una función de asignación de creencia básica generalizada (gbba) basada en la proyección de ejes, y se fusionan usando la regla de fusión híbrida Dezert-Smarandache (DSm). En este trabajo también se propone una forma de reducir y manejar dinámicamente el marco de discernimiento para optimizar recursos computacionales. Se realizaron pruebas sobre simulaciones animadas y secuencias reales, y los resultados son comparados contra un modelo bayesiano. Los experimentos muestran que el modelo propuesto consigue una mejora en la precisión del seguimiento a alto nivel.

Abstract

Using multiple cameras has proven to increase vision systems capabilities, mainly by extending visual field or complementing information from cameras to reduce uncertainty. While there are emerging approaches which take in consideration information such as where an object is standing or how long it has been on a particular zone, it is still no common to find work related to the use of multiple cameras to improve information useful for such surveillance systems.

In this work a decision fusion level model is proposed to take advantage of several geographically distributed cameras, in order to reduce uncertainty derived from cameras'perspective. Previously defined zones are considered to track objects position, so the stage of the processing at which data integration takes place is useful for high level surveillance systems, such those focused on behavior recognition. In our model, individual decisions are taken by means of an axis-projection-based *generalized basic belief assignment* (gbba) function and finally fused using Dezert-Smarandache (DSm) hybrid rule. It is also proposed a way to reduce and manage dynamically the frame of discernment to optimize computer resources. Results of model are presented, obtained from tests on animated simulations and real sequences, and compared to a bayesian fusion model. Experiments proved that the proposed model yields a good improvement in tracking accuracy at high level processing.

Agradecimientos

En el tiempo que pasé en el INAOE tuve la fortuna de conocer personas, que no sólo tuvieron el papel de estudiantes, profesores o científicos, sino que son, más que nada, amigos.

Agradezco al doctor Leopoldo Altamirano, quien me invitó a formar parte del Laboratorio de Visión por Computadora, un grupo de trabajo donde nacieron mis deseos por estudiar la maestría. Le agradezco su apoyo incondicional, guía y respaldo durante el tiempo que llevo de conocerlo.

A los doctores René Cumplido, Francisco Martínez y Enrique Sucar agradezco las revisiones hechas a este trabajo de tesis, así como sus útiles comentarios. Al doctor Enrique agradezco en particular sus valiosas sugerencias sobre la aplicación del modelo de fusión probabilista.

Agradezco a todos mis compañeros de maestría los ratos de diversión y compañerismo que no faltaron, aún entre las largas jornadas de estudio y desvelos. Agradezco de corazón a mis compadres: Nestor, Octavio, Elías y Gershom, y a mis mejores amigas: Paty, Sayde y Danaé.

Gracias a mis compañeros y amigos del LVC, principalmente a Janeth Cruz, Iván Olivera, Sergio Mendoza y Eduardo Rodríguez, quienes han sido un ejemplo a seguir en el Laboratorio de Visión.

Agradezco a mis padres, a mi hermana y hermano, sin quienes simplemente no habría tenido la fuerza para concluir esta etapa de mi vida.

A mis padres.

Índice de figuras

Índice de tablas

Índice general

Capítulo 1

Introducción

1.1. Uso de varios sensores

Dentro del contexto de la percepción por computadora, los sensores son una parte vital, ya que gracias a ellos es posible llevar los datos a un dominio informático.

El uso de sensores genera condiciones en las cuales se tiene que considerar la incertidumbre como un elemento presente en el procesamiento de información. Ésto debido a factores inherentes a los sensores, tales como existencia de ruido, manejo de precisión o errores de calibración, entre otros.

Actualmente la tecnología permite no usar una sola fuente o un solo tipo de información para poder analizar un evento, sino utilizar varias fuentes y en ocasiones incluso diversos tipos de información. Esta "combinación" de información no es una simple suma, unión o mezcla, sino que debe conservar un sentido y el propósito de beneficiar los resultados. En la naturaleza, un ejemplo del beneficio del uso de más de una fuente de información es la forma en la que los sentidos humanos funcionan. La información que nuestro cerebro obtiene de ellos es tratada de tal forma que se complementan y nos ayudan a discernir sobre lo que acontece a nuestro alrededor. Así por ejemplo, si escucháramos un ladrido con volumen

alto sabríamos que un perro está cerca de nosotros e inmediatamente usaríamos la vista para tener una mejor idea de la posición del perro y de sus intenciones. En este ejemplo los ojos y oídos son sensores que reciben diferente información, pero en el cerebro ambas ayudan a saber la posición del perro.

Usar varios sensores puede tener diferentes propósitos de acuerdo a su tipo, la forma en la que sean posicionados y la manera en la que se use la información obtenida de ellos, para tener un panorama general de lo que se puede hacer usando más de un sensor se puede consultar el trabajo de Brooks (6).

1.2. Motivación

Para poder realizar fusión de información, no se emplean siempre las mismas técnicas, éstas dependen de los casos de aplicación. Por ejemplo, no es lo mismo cuando se fusionan datos obtenidos de instrumentos médicos para la monitorización de signos vitales, que cuando se fusiona información proveniente de radares. En el caso de los signos vitales podría simplemente interesar que las métricas obtenidas por los sensores se encontraran dentro de determinado rango, y esto puede manejarse como simples valores booleanos, mientras que en los radares se manejan datos que describen el movimiento de los objetos detectados, como posición o velocidad. Las técnicas aplicadas dependen del propósito de la fusión y varían cuando se utilizan sensores diferentes o cuando la información representa mediciones de diferente tipo con relación al mismo fenómeno, por lo que existe un gran número de técnicas de fusión (23; 20; 12).

En el contexto de visión por computadora, una de las tareas sobre las que se ha hecho más investigación es en sistemas automatizados de vigilancia, tanto para aplicaciones militares como no militares. En tales sistemas, la fusión de varios sensores visuales (cámaras) se ha convertido en una línea de estudio, donde lo que se persigue en investigaciones recientes es obtener buenos resultados en los niveles altos de procesamiento. En estos niveles de procesamiento la información

correspondiente a los objetos o situaciones de interés se puede definir en términos de entidades o relaciones, que por lo general son la forma en la que los humanos los describiríamos. Por ejemplo, en un sistema de vigilancia de alto nivel, la salida del sistema podría ser: "El objeto azul cuadrado está sobre la región 1".

Investigaciones recientes, en cuanto a automatización de los sistemas de vigilancia (24; 11; 26; 33), buscan emplear información de alto nivel relacionada con los objetos en escena y sus interacciones, por ejemplo saber cuánto tiempo ha estado situado sobre determinada zona un objeto. En (33) un sistema de seguimiento usando regiones predefinidas es usado para analizar patrones de conducta, pero con una sola cámara. En (26) un Modelo Oculto Jerárquico de Markov se emplea para identificar actividades, basándose en el seguimiento de personas que transitan en un cuarto que se ha dividido en celdas. Aún cuando en este trabajo se utilizan dos cámaras, éstas simplemente son cambiadas de acuerdo a la que brinde una mejor perspectiva.

En los trabajos existentes hasta el momento, que funcionan con información a alto nivel, el uso de sistemas multicámaras no se está aprovechando para reducir la incertidumbre, se asume que la información de la cámara es totalmente confiable. La teoría evidencial es un modelo matemático propuesto recientemente por Jean Dezert y Florentin Smarandache (9) para combinación de información a nivel decisión, que no requiere de información previa para la asignación de creencia (simplemente la definición de una función de asignación). Su aplicación en la fusión de información de múltiples cámaras a alto nivel no ha sido explorada. Tal situación motivó a la realización de este trabajo, que plantea y prueba un modelo para la combinación de información de diferentes cámaras para tareas de vigilancia, a un nivel alto de procesamiento.

1.3. Objetivos de la investigación

Los objetivos consisten en el desarrollo e implementación de un modelo para fusión a nivel decisión, que esté basado en la Teoría Evidencial, con el objetivo de hacer seguimiento de objetos en un ambiente multicámara. Los objetivos específicos de este trabajo son:

- Definir una arquitectura de procesamiento apta para fusionar a nivel decisión.
- Definir un modelo de fusión a nivel decisión basado en teoría evidencial.
- Aplicar alineación de datos y calibración de cámaras para la problemática multicámara.
- Someter el modelo de fusión a pruebas con secuencias de imágenes sintéticas y reales, y comparar los resultados obtenidos contra los que se obtienen con una sola cámara y contra los obtenidos con un modelo de fusión bayesiano.

1.4. Contribuciones

La contribución principal de esta tesis es brindar un modelo para la fusión de información de múltiples cámaras. La arquitectura de procesamiento de la información de las cámaras permite considerar a cada cámara como un experto, y resultó ser útil tanto para la fusión con teoría evidencial DSm como para otras teorías, como la probabilista, con la que también fue probada. La información de las cámaras que se considera útil para este caso es la posición de objetos en movimiento relativa a regiones previamente definidas en la escena, con el propósito de proporcionar una base para el análisis de comportamiento y situaciones sospechosas, en lo que respecta a vigilancia automatizada.

La Teoría evidencial Dezert-Smarandache (DSmT)(9; 30) es muy reciente, y surgió como un modelo de fusión que permite resolver problemas que el mode-

lo antecesor en la Teoría Evidencial, conocido como Dempster-Shafer (DST) no resuelve. La Teoría Evidencial probada en otros ámbitos tiene la ventaja de considerar ignorancia y no requerir mediciones previas y cálculo de probabilidades *a priori*, como lo hace el modelo de fusión basado en la teoría bayesiana. La aplicación de las teorías DS y DSm para la fusión multicámara no había sido previamente estudiada, y es en este trabajo que se plantean las bases para realizar fusión de información proveniente de cámaras geográficamente distribuidas usando la teoría DSm, y se demuestra que se consigue una mejora significativa en la determinación de la posición de objetos en movimiento, considerando la incertidumbre relacionada con la perspectiva de las cámaras.

1.5. Estructura del documento

En esta sección se describe la estructura del documento:

- **Capítulo 2**. Se describe el panorama general de la investigación relacionada con fusión multicámara, partiendo desde los conceptos y aplicaciones de fusión de múltiples sensores hasta las arquitecturas y modelos más recientes.

- **Capítulo 3**. En este capítulo se describen los fundamentos de la Teoría Evidencial, abarcando tanto la Teoría Evidencial DS como la Teoría Evidencial DSm.

- **Capítulo 4**. El modelo desarrollado como trabajo de tesis es descrito en el capítulo 4. Se da a conocer la arquitectura y se explican los módulos involucrados. Los elementos teóricos del modelo son definidos y relacionados con los componentes de la arquitectura.

- **Capítulo 5**. En este capítulo se describen las pruebas y las evaluaciones realizadas a la implementación del modelo. Se presentan comparaciones sobre los resultados de las pruebas.

- **Capítulo 6**. Las conclusiones y las propuestas de trabajo futuro se abordan en el capítulo 6.

Capítulo 2

Fusión de sensores

En este capítulo se describe el escenario actual de la investigación en lo que respecta a fusión multicámara, abarcando sus aplicaciones y los métodos que se usan para fusionar información, primero desde un punto de vista general y posteriormente en el contexto de sistemas de vigilancia.

2.1. Introducción

Entre las cosas comunes de la vida cotidiana moderna están sin duda los sensores, que no son más que dispositivos que perciben señales que se generan en nuestro entorno. En este trabajo los sensores que se utilizan son los que trabajan en el espectro electromagnético que corresponde a la luz visible, y son conocidos como cámaras. Existen distintos tipos de cámaras, y pueden variar en resolución, tipos de lente, formatos de almacenamiento de información o incluso en el rango de espectro (infrarrojas, rayos X, etc.). Aún cuando en la parte fundamental de este trabajo nos centramos en el uso de cámaras como sensores, y en la fusión de la información que obtenemos con ellas, las siguientes secciones dentro de este capítulo se desarrollan desde un punto de vista general, ya que la mayoría de conceptos involucrados son aplicables a otros tipos de sensores. Será hasta la sección 2.4 que se establezca el estado del arte relacionado con la fusión de múltiples

cámaras.

Para comenzar a hablar de detalles relacionados con el campo de fusión de sensores, es necesario dar a conocer un par de definiciones de ese término. Una definición para fusión de sensores, que se apega a los propósitos de vigilancia, es dada por el Departamento de Defensa de los Estados Unidos (20): *Fusión de sensores es un proceso con múltiples interfaces, que trata con la detección automática, asociación, correlación y combinación de información de múltiples fuentes.* Esta definición, si bien no es oficial ni única, da una buena idea de lo que conlleva hablar de fusión de sensores en el contexto de vigilancia.

El uso de más de un sensor no implica necesariamente mejores resultados. Las ventajas de usar más de un sensor dependen de la manera en la que éstos se usen. Desde un punto de vista simple, aún cuando no se fusionen, usar varios sensores es útil porque se tiene redundancia, y así se tiene mayor tolerancia a fallos. Cuando no son redundantes y tampoco se fusiona la información, el uso de varios sensores tiene como ventaja la ampliación de la cobertura. Pero la herramienta más poderosa es la independencia de observaciones entre los sensores. Aún cuando sean del mismo tipo y cuando se utiliza esta característica de manera adecuada, se puede mejorar la calidad de la información.

2.2. Arquitecturas y niveles de fusión

Algunos modelos de arquitecturas[1] han sido propuestos con el propósito de estandarizar los términos usados para describir los módulos que conforman la arquitectura y la forma en que interactúan. En esta sección se hace una revisión de tales modelos con la intención de clarificar la ubicación de este trabajo en relación con las técnicas existentes .

Uno de los primeros intentos por estandarizar la terminología y la forma de

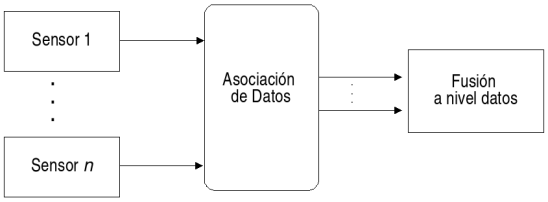
[1] Un modelo de arquitectura es considerado diferente a un modelo de fusión, ya que no describe la técnica de fusión en sí, sino solamente la forma en la que los módulos involucrados (hardware y software) interactúan

ver los sistemas de fusión fue hecho por Dasarathy (8). En su trabajo se hace un análisis de las aportaciones en el área de fusión de sensores para clasificarlas tomando en cuenta varios criterios, tales como el nivel al cual la fusión se lleva a cabo, el propósito del proceso de fusión, el dominio de la aplicación, el tipo de sensores empleados y la configuración de los sensores. La clasificación basada en el nivel de fusión ha sido una de las formas más populares para caracterizar una arquitectura. En este modelo se toman en cuenta los niveles jerárquicos de las etapas de procesamiento en el cual la integración de la información se hace. La clasificación por niveles se hace en tres clases: fusión de datos, fusión de características y fusión de decisiones. Un ejemplo de arquitectura para fusión a nivel datos es la combinación de información de dos imágenes satelitales tomando los valores numéricos pixel por pixel; una arquitectura para fusionar a nivel características tendría, por ejemplo, un módulo para extracción de bordes y la información de estos bordes sería la que se combinaría; una arquitectura que realizara una interpretación de la información por cada sensor y llevara a cabo la integración de estas interpretaciones estaría fusionando a nivel decisión. Los esquemas de fusión a nivel datos, características y decisión se muestran en la figura 2.1

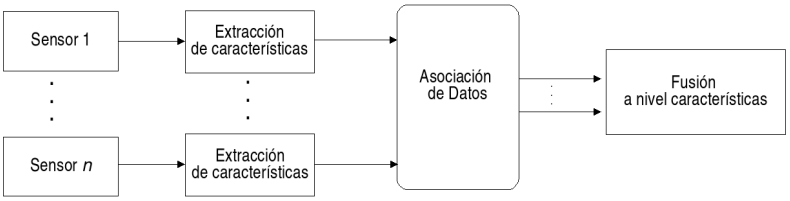
En otras fuentes, esta misma clasificación por niveles nombra de diferente manera a los niveles, pero siguen siendo lo mismo, tal como lo hace Hall (12), definiendo: fusión directa de datos de sensor, fusión de vectores característicos y fusión de decisiones o inferencias de alto nivel.

Existe otra forma de clasificar los sistemas de fusión, y también ha sido difundida por la literatura relacionada con fusión. Se trata del llamado *Modelo de Fusión de Datos JDL*, propuesto por el JDL[2] (32; 23), y cabe mencionar que a lo largo de este trabajo se maneja el término "modelo de fusión" refiriéndose a un modelo matemático, mientras que en el título de JDL simplemente se refiere a la forma en que trabajan los módulos que componen la arquitectura de fusión. En la

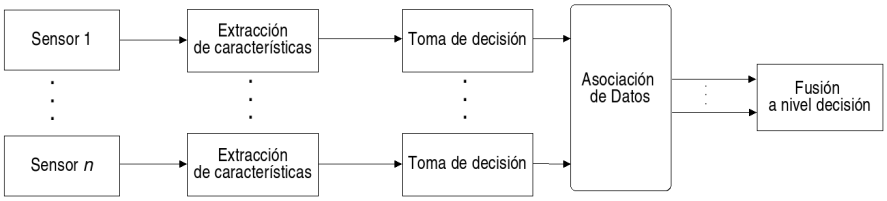
[2]El *Data Fusion Group* del *Joint Directors of Laboratories* es un comité del Departamento de Defensa de los EUA que creó el modelo de fusión denominado JDL



(a) Fusión a nivel datos



(b) Fusión a nivel características



(c) Fusión a nivel decisión

Figura 2.1: Arquitecturas generales de fusión

arquitectura JDL, al igual que en la clasificación de Dasarathy, se manejan niveles de fusión. El modelo de Fusión JDL es una descripción funcional de los elementos involucrados en el proceso de fusión. Su propósito era el de estandarizar conceptos para facilitar la descripción de los tipos de problemas para los cuales la fusión es aplicable. El modelo JDL contempla los siguientes niveles:

- Nivel 0: Estimación de estado de entidades sub-objeto (señales, características).
- Nivel 1: Estimación de estado de objetos físicos discretos (vehículos, edificios, personas).
- Nivel 2: Estimación de relaciones entre entidades (juntarse, seguir, estar sobre).
- Nivel 3: Estimación de consecuencias (resultados, objetivos).

Comparada con la clasificación de Dasaraty, la clasificación JDL contempla un nivel más, que puede servir para especificar determinadas propiedades de un sistema de fusión, sin embargo, también es criticado por ser confuso. En un reciente trabajo (23), Llinas hace una revisión de las etapas del modelo, y sugiere cambios para mejorar las fronteras en los niveles manejados en el JDL.

2.3. Modelos de fusión de información

La combinación o fusión de información es una área de estudio muy extensa. La información es frecuentemente catalogada como de bajo nivel, nivel intermedio o alto nivel, dependiendo de la etapa de procesamiento en la cual la fusión se lleva a cabo. La fusión de bajo nivel combina información cruda para producir nueva información que por lo general es más útil que la cruda. En el nivel intermedio o nivel de características, se fusionan bordes, esquinas, líneas, texturas, etc. Varias

clases de algoritmos o modelos han sido desarrollados para la fusión de información, pero no todos ellos son aplicables en todos los casos. Es decir, la aplicación de los modelos de fusión depende de la problemática. Por ejemplo, el filtro Kalman es una herramienta que se ha utilizado para la estimación de estado en sistemas de vigilancia como radares. También puede ser aplicado en la fusión de información, donde cada sensor es una fuente de medición del estado actual. Sin embargo, dado que en ese modelo el estado actual de los objetos de interés se define como una transición lineal dependiente del estado anterior, es prácticamente imposible aplicarlo en modelos de fusión a nivel decisión, donde los estados actuales no son dependientes de los anteriores en el tiempo.

En este trabajo el interés se centra en el nivel decisión, que es un nivel alto de procesamiento, es decir, la información a combinar son las decisiones provistas por expertos, cuya función será cumplida por las cámaras. A continuación se explican brevemente el modelo Probabilista y de Lógica Difusa que se han aplicado para fusión de información a alto nivel. En el capítulo siguiente se describen los modelos de fusión DS y DSm usados en la Teoría Evidencial.

2.3.1. Fusión Probabilista

Existen muchas formas de aplicar la teoría probabilista a la fusión de información. Por lo general, la regla de Bayes aplicada como una forma de calcular la probabilidad de que una hipótesis sea verdadera dada la evidencia que proviene de las fuentes de información.

Un modelo de fusión bayesiana, aplicada al proceso de identificación de objetos, fue publicada por E. Waltz y J. Llinas en (20). El modelo probabilista de fusión bayesiana simple es usado como módulo de combinación para la tarea de determinar el tipo de objeto, dada la evidencia proveniente de sensores. Cada sensor provee de una hipótesis de la identidad del objeto basada en las observaciones y en algoritmos específicos para cada sensor, lo cual corresponde al nivel

de decisión. Su funcionamiento se ilustra en la figura 2.2.

El modelo probabilista de clasificación se considera un modelo condicional

$$p(C|A_1, A_2, \ldots, A_n) \tag{2.3.1}$$

donde C es la clase y A_i son atributos. Los atributos pueden ser información proveniente de los sensores, a nivel decisión. Por ejemplo, podría ser que cada sensor realizara una identificación del objeto previo a la fusión, y las decisiones por cada uno de los sensores fueran los atributos del modelo de fusión probabilista. Esto se puede expresar usando el Teorema de Bayes en la siguiente forma:

$$p(C|S_1, \ldots, S_n) = \frac{p(C)p(S_1, \ldots, S_n|C)}{p(S_1, \ldots, S_n)} \tag{2.3.2}$$

donde S_j es la identificación hecha por el sensor j.

Para simplificar, en la práctica el denominador es despreciado, ya que se compone de una constante, y el numerador es simplificado tomando en cuenta la definición de probabilidad condicional como

$$\begin{aligned} p(S_1, \ldots, S_n|C) &= p(S_1|C)p(S_2, \ldots, S_n|C, S_1) && (2.3.3) \\ &= p(S_1|C)p(S_2|C, S_1)p(S_3, \ldots, S_n|C, S_1, S_2) && (2.3.4) \end{aligned}$$

Y dado que el modelo bayesiano simple asume independencia condicional, se cumple que

$$p(S_i|C, S_j) = p(S_i|C) \tag{2.3.5}$$

y el modelo se simplifica a

$$\begin{aligned} p(C|S_1, \ldots, S_n) &\propto p(C)p(S_1|C)p(S_2|C)p(S_3|C)\ldots && (2.3.6) \\ &= p(C)\prod_{i=1}^{n} p(S_i|C) && (2.3.7) \end{aligned}$$

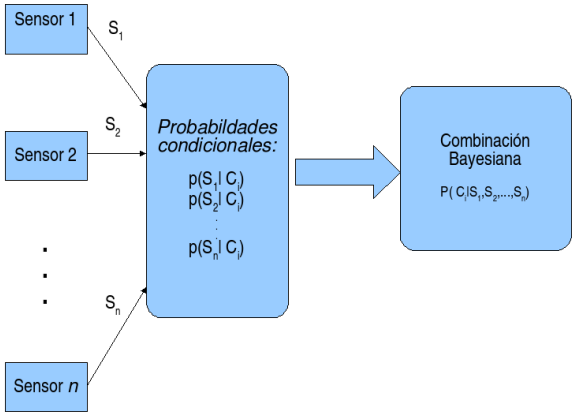


Figura 2.2: Modelo de fusión bayesiana. Cada uno de los sensores aporta una decisión (un tipo de objeto). Los tipos de objetos candidatos son evaluados condicionalmente, tomando las probabilidades condicionales, que son utilizadas finalmente para calcular una probabilidad dd C_j.

En (20) el desempeño previo de cada sensor es introducido como probabilidad *a priori*, es decir, es la probabilidad de que el sensor declare que el objeto es de cierto tipo dado que el objeto es de hecho del tipo j. Esta probabilidad se denota por $P(C|S_j)$. La probabilidad de haber observado un objeto j del conjunto de M objetos dada la declaración de evidencia S_1 del Sensor 1, declaración S_2 del Sensor 2, etc. es $P(C_j|S_1, S_2, \ldots, S_n)$.

De las probabilidades resultantes por lo general se selecciona el valor máximo, llamado probabilidad *a posteriori* máxima (MAP).

2.3.2. Fusión con Lógica Difusa

En la lógica difusa se trabaja con grados de membresía a conjuntos difusos, o también llamados distribuciones de posibilidad. Para problemas de clasificación, uno de los modelos difusos consiste en fijar para cada elemento $\mathbf{x}$ un grado de membresía a la clase C_i, de acuerdo a la fuente j. Estos modelos representan

explícitamente imprecisión en la información, así como ambigüedad entre las clases o decisiones. Los modelos de lógica difusa han mostrado gran aplicabilidad en teoría de control.

Para la combinación de información en el campo de lógica difusa, se tienen una gran variedad de operadores de combinación, que pueden ser adaptados a distintas situaciones, es decir, se amoldan fácilmente a las problemáticas. Con la lógica difusa también es posible fusionar a diferentes niveles de información. En (16) se combinan imágenes e información extraída de las imágenes satelitales multiespectrales a bajo nivel. Para el preprocesamiento se hace uso de operadores difusos o reglas difusas para la combinación de mediciones de homogeneidad locales o regionales y mediciones de contraste, para suavizar regiones homogéneas y mantener los bordes. La mayoría de las aplicaciones usan una variedad de información que puede ser extraída de una o varias imágenes y posteriormente fusionada.

Cuando se maneja información de alto nivel proveniente de varias fuentes, es posible utilizar *funciones de agregación*. Las decisiones de varias fuentes son agregadas para producir una sola etiqueta de clase. Dado un objeto $\mathbf{x}$, $d_{i,j}(\mathbf{x}) \in [0,1]$ es el grado de membresía que la fuente i aporta para la clase j. Muchas fuentes pueden producir etiquetas que son compatibles con la lógica difusa, incluso la teoría probabilista, que produce estimaciones de las probabilidades a posteriori de las clases posibles para $\mathbf{x}$. La matriz formada por elementos $\{d_{i,j}(\mathbf{x})\}$ es llamada el *perfil de decisión* para $\mathbf{x}$.

Para combinar esta información se aplican funciones de agregación, de las cuales existen una variedad para la toma de decisiones en la lógica difusa. El respaldo para la clase j es calculado usando las entradas del perfil de decisión para esa clase, el cual es la $j-$ésima columna de la matriz. Dada una función de agregación A, la salida para la clase j es $d_{e,j}(\mathbf{x}) = A(d_{1,j}(\mathbf{x}), d_{2,j}(\mathbf{x}), ..., d_{L,j}(\mathbf{x}))$

Una de las desventajas de los modelos de lógica difusa es que los operadores tienen varias definiciones, y muchos de ellos no tienen una justificación clara.

2.4. Fusión multicámara

Un entorno multicámara es descrito como un ambiente en el cual más de una cámara es utilizada para recopilar información de una escena. El término *multicámara* es generalmente usado cuando estas cámaras están geográficamente distribuidas, a diferencia de configuraciones en las que dos o más cámaras comparten exactamente la perspectiva, como en los sistemas multi-espectrales, o en la visión estereoscópica.

La investigación relacionada con múltiples cámaras se aborda desde dos enfoques que se distinguen por la cobertura que tienen las cámaras sobre la escena. El primer enfoque es cuando las cámaras se colocan de tal manera que sus campos de vista no se traslapan, cubriendo escenarios en los que los objetos son seguidos por una cámara a la vez, asociando aquellos objetos que pasan de un campo de visión en un sensor a otro campo de visión, con un sensor distinto. En el segundo enfoque, que es el que se adopta en este trabajo, se usan las cámaras traslapando sus campos de vista, lo cual parecería no tener muchas ventajas si las cámaras tienen exactamente la misma perspectiva. Sin embargo, el mayor de los beneficios de tener cámaras traslapadas es precisamente la posibilidad de tener cámaras con distintas perspectivas, y por lo tanto distinta información.

2.4.1. Detección de movimiento

La palabra *movimiento* se refiere, en un sentido estricto, al cambio de posición u orientación de un objeto dentro de un sistema de referencia. Dado que una de las tareas de los sistemas de vigilancia es la de identificar y seguir objetos en movimiento, un módulo de detección de movimiento es esencial en las primeras etapas de procesamiento.

En los sistemas de sensores de visión, el movimiento es percibido por cambios en la señal de entrada. En las cámara esta señal de entrada está conformada por secuencias de imágenes. La finalidad del proceso de detección de movimiento es

encontrar en la imagen aquellos objetos que se están moviendo con respecto al escenario.

En visión por computadora se pueden tener cámaras colocadas en posiciones estáticas y que no cambian su campo de visión, o bien cámaras móviles, que pueden estar colocadas sobre plataformas dinámicas y así cambiar su orientación (cámaras con funciones pan, tilt) o incluso posición (cámaras montadas en robots, gruas, etc.). Dado que el movimiento es relativo, una cámara en movimiento podría detectar movimiento, debido a su propio movimiento, aún cuando en la escena todos los objetos permanezcan estáticos. Existen algoritmos para hacer detección de movimiento tanto con cámaras fijas como móviles.

Las tareas más comunes en los sistemas de vigilancia multicámara se pueden llevar a cabo con cámaras estáticas, cuya principal fortaleza está en trabajar en conjunto para cubrir una área más extensa, o aprovechar diferentes perspectivas para resolver problemas de oclusión, por lo cual el uso de cámaras móviles es poco común.

Existe una cantidad muy grande de métodos para la detección de movimiento en las imágenes, como el cálculo de flujo óptico, modelo con mezclas gaussianas, diferencias acumulativas, etc. La mayoría de los métodos se basan en la comparación de imágenes, tomando la imagen actual y comparándola con la anterior o bien con una imagen definida como fondo. A continuación se mencionan algunos de los métodos usados en la detección de movimiento con campos de vista estáticos.

Cuando se tiene una secuencias de imágenes obtenida de una cámara y el campo de visión no cambia, es posible hacer un análisis de la imagen para saber cuáles pixeles pertenecen al escenario y cuáles a los objetos en movimiento. Este análisis es una comparación pixel a pixel con una imagen de fondo. En el enfoque de diferencia de imágenes se considera a una de las imágenes en la secuencia como el fondo de la escena y la siguiente imagen se toma como referencia actual en el tiempo. Cuando un objeto se encuentra en movimiento, gran parte de su área en la imagen será resaltada al calcular la diferencia entre la imagen anterior y la

actual.

Sean I_{t-1}, I_t dos imágenes obtenidas en los tiempos $t-1$ y t respectivamente. Si para el tiempo t se genera una nueva imagen $D_t(i,j) = I_t(i,j) - I_{t-1}(i,j) \forall i,j$, donde i y j son el renglón y la columna de las imágenes respectivamente, cada elemento $D(i,j)$ tendrá almacenado un pixel con el valor de la diferencia entre las imágenes I_t e I_{t-1}. Los pixeles (i,j) que no sufren un cambio en tonalidad, tendrán un valor igual o muy cercano a cero, mientras que los pixeles que cambian mucho devolverán una diferencia grande, lo cual es comúnmente un indicativo de movimiento (recordemos que también podría deberse a ruido o cambio en la iluminación).

Un paso que comúnmente se aplica a la imagen D es un proceso de umbralización, para generar una matriz binaria D_b con los pixeles que tienen cambios, esto con la finalidad de tener marcadas las regiones que podrían representar movimiento. Tal matriz binaria bien podría ser pasada posteriormente por un algoritmo de conectividad de regiones para agrupar así los pixeles que podrían representar objetos, y finalmente descartar por área los objetos muy pequeños, que comúnmente representan ruido y no movimiento.

Cuando se calcula $I_t - I_{t-1}$ y se aplica umbralización, pueden existir pixeles en D que resulten con valores negativos y que por lo tanto no pasen el umbral. Para tomar en cuenta estas áreas también, existe una variante en el algoritmo de diferencia de imágenes, que considera una diferencia absoluta. La matriz binaria que ilustra las diferencias se expresa en ese caso con la siguiente ecuación:

$$D_b(i,j) = \begin{cases} 1 : |I_t(i,j) - I_{t-1}(i,j)| \geq k \\ 0 : |I_t(i,j) - I_{t-1}(i,j)| < k \end{cases} \tag{2.4.1}$$

donde k es el valor del umbral, a partir del cual los pixeles son considerados como posible movimiento. Con este enfoque se identifican las diferencias entre las dos imágenes (no sólo de $I_t - I_{t-1}$), lo cual ilustra la posición anterior y la actual del objeto en movimiento.

(a) Imagen en $t-1$ (b) Imagen en t (c) Diferencia (d) Diferencia absoluta

Figura 2.3: Detección de movimiento por diferencia de imágenes

Al utilizarse un umbral para clasificar a un pixel como movimiento o no, el algoritmo depende de tal parámetro para poder funcionar adecuadamente, es decir, el umbral debe ser determinado con base en observaciones y análisis previo de las imágenes obtenidas.

Este algoritmo suele fallar en encontrar las partes del objeto en movimiento que quedan situadas sobre la zona que ocupaba el mismo objeto en la imagen anterior, ya que esos pixeles mantienen su tonalidad y por lo tanto no serán detectados con el algoritmo. El resultado en estos casos es que el centro del objeto no es marcado como movimiento en D_b y sólo los bordes son considerados movimiento. La figura 2.3 se muestran dos cuadros consecutivos y los resultados de detección de movimiento con diferencia y diferencia absoluta. En esta misma imagen se puede observar la influencia del ruido en la detección de movimiento, ya que no sólo la persona en movimiento es detectada, sino también se encuentran pequeños objetos inexistentes.

Si bien esta es una solución práctica y robusta a cambios de iluminación para detectar movimiento, la precisión se ve sacrificada al detectarse también las zonas donde el objeto estuvo y no sólo su posición actual.

El método de diferencia de imágenes consiste en calcular la resta pixel a pixel en las imágenes consecutivas de la secuencia, con lo que se encuentran las zonas que cambian en el tiempo. Este algoritmo, aunque es simple y rápido, también tiene algunas desventajas, entre ellas la principal es que no detecta con exactitud el contorno del objeto en movimiento.

Para resolver el problema de que en la diferencia de imágenes el objeto en movimiento forma parte de la imagen anterior y por lo tanto también es restado al momento de calcular la diferencia, una alternativa es el modelado del fondo, con lo cual se asegura que una diferencia daría únicamente el objeto en movimiento, que no forma parte del modelo del fondo.

Los métodos de extracción de fondo se basan en la generación de un modelo para cada pixel, partiendo de la observación de las imágenes. Para lo cual se requiere un bloque de la secuencia de imágenes obtenidas de las cámaras.

Dentro de este grupo de algoritmos existe una variedad de opciones que se pueden tomar, tanto para generar el modelo del fondo como para hacer que éste se vaya adaptando con el tiempo, así los problemas de cambios en la iluminación quedan resueltos.

2.4.2. Alineación de datos

Cuando se está trabajando en un entorno multicámara hay un requisito indispensable para que la información que se obtiene de las cámaras tenga coherencia, y éste es que la información obtenida de las cámaras esté relacionada entre sí, lo cual se consigue mediante sistemas de referencia espacial y temporal comunes. Es imprescindible que la información proveniente de las cámaras tenga una relación espacial y temporal. Cuando se cumplen ambas, se dice que el sistema cuenta con *Alineación de datos*. Técnicamente la alineación de datos se cumple cuando se cumplen tanto la *Alineación espacial* como la *Alineación temporal* (12). La alineación es un paso muy importante para llevar a cabo la fusión de información ya que con ella se consigue la relación entre los datos. Para conseguir esta alineación se debe contar con un método para poder asociar la información.

Alineación espacial

La alienación espacial surge por la necesidad de tener todas las fuentes de información (sensores) referenciando un sistema único de posicionamiento de los objetos en movimiento. Esto significa que al haber un objeto en movimiento éste se pueda identificar unívocamente en todos los sensores que lo tienen en su campo de visión.

En los sistemas multicámara cuyos campos de visión no se traslapan, la alineación consiste en hacer coincidir los bordes de los campos de visión que sean adyacentes, lo cual se puede solucionar con enfoques muy simples. Sin embargo, cuando los campos de visión se traslapan, como es el caso en el presente trabajo, la alineación espacial es una tarea que conlleva la definición de un sistema de referencia espacial, y una metodología para transformar la información recibida en las cámaras a dicho sistema de referencia común para todas las cámaras.

Ya sea que se trabaje con una sola cámara o se tengan un conjunto de ellas, los parámetros de calibración tanto internos como externos son necesarios para establecer una relación entre las imágenes obtenidas en la cámara y el mundo real (1) . De esta manera es posible saber características de un objeto que aparece en la imagen tomada con la cámara, como tamaño y posición, por ejemplo. Para el propósito de los sistemas de vigilancia la posición de un objeto en movimiento es la información relevante.

Una imagen obtenida con una cámara es una proyección de la escena en un plano, llamado plano imagen. Este plano imagen puede servir como sistema de referencia para ubicar el objeto, es decir, la posición de un objeto en movimiento en la imagen podría estar dada en renglones y columnas dentro de la imagen. En la figura 2.4 se muestra cómo se representan los puntos del escenario sobre la imagen plano, obtenida con la cámara.

Es bien sabido que las cámaras obtienen información de un mundo tridimensional y la almacenan en una imagen, que es una representación en dos dimensiones,

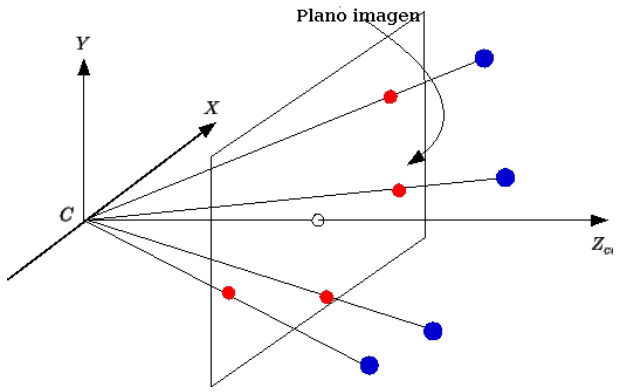


Figura 2.4: Obtención de plano imagen

y es durante este proceso que se pierde una dimensión, por lo que es prácticamente imposible volver a reconstruir el escenario con sólo una imagen.

La captura de una imagen con la cámara es un proceso que se puede explicar considerando el escenario como un espacio 3D igual a $\mathbb{R}^3$ y a la imagen como un plano proyectivo $\mathbb{P}^2$. En (13), Hartley y Zisserman plantean el proceso de captura de la imagen como un simple mapeo de $\mathbb{P}^3$ a $\mathbb{P}^2$, donde $\mathbb{P}^3$ es un plano en el espacio $\mathbb{R}^3$ escrito en términos de coordenadas homogéneas $(X, Y, Z, T)^T$, con el centro de la proyección en el origen $(0, 0, 0, 1)^T$. Ellos utilizan una cuarta coordenada para ilustrar el tiempo en el proceso de captura de un rayo de luz, aunque para fines prácticos se puede considerar que una imagen es obtenida en un instante de tiempo, por lo cual la coordenada final de (X, Y, Z, T) es irrelevante para saber dónde es mapeado un punto del escenario en la imagen, ya que se asume que esa información corresponde a un instante de tiempo. Así cada punto de $\mathbb{P}^2$ se puede definir con las coordenadas homogéneas $(X, Y, Z)^T$ y el mapeo se puede representar con una matriz $P_{3\times4} = \{I_{3\times3}|O_3\}$, donde $I_{3\times3}$ es la matriz identidad y O_3 es un vector cero. Si se quiere tener un centro de proyección diferente y un marco de coordenadas diferente en la imagen, se puede generalizar representando la matriz de proyección con dimensiones 3×4 actuando sobre las coordenadas homogéneas del punto en $\mathbb{P}^3$ mapeándolo al punto de la imagen en $\mathbb{P}^2$.

Alineación temporal

En un sistema multicámaras, y en general, en todo sistema multisensores, además de necesitar que los datos se puedan relacionar espacialmente, se deben poder relacionar también temporalmente; de otra manera, posiblemente algunos sensores estarían aportando datos desfasados en el tiempo y la información resultante de la fusión estaría distorsionada.

En los sistemas multicámara se tienen ya un buen número de mecanismos para resolver el problema de la alineación temporal. En un sistema de fusión distribuida el problema de alineación temporal es bastante complejo, sin embargo, en los sistemas centralizados, este problema suele resolverse con uno de los métodos más usados, por su sencillez y eficacia: el marcado de las imágenes por tiempo o *timestamping*.

2.4.3. Trabajo relacionado

Aunque el concepto de fusión de datos no es nuevo, el de fusión de sensores es bastante reciente, ya que es hasta la actualidad que los dispositivos de procesamiento tienen el poder necesario para aplicar técnicas que hasta hace poco serían muy costosas en tiempo de ejecución. A la par con la evolución de los dispositivos electrónicos, las aplicaciones para fusión de sensores se han ido incrementando y de la misma manera ha sucedido con las técnicas de procesamiento para cada aplicación. A continuación se presenta una breve revisión de algunos de los trabajos que utilizan múltiples cámaras para seguimiento.

El uso de múltiples cámaras puede mejorar las capacidades de un sistema de vigilancia para encontrar objetos y evitar problemas de oclusión. Actualmente se utilizan a un nivel bajo sin hacer fusión, como en el trabajo de Beynon (3), para la detección de paquetes abandonados. En su trabajo se usan cámaras que comparten campos de vista y hacen clasificación de los objetos móviles en la escena porque el interés son los objetos que se vuelven estacionarios. Otro trabajo que se

centra en la correspondencia es el de Weiming Hu (15), que hace corresponder los objetos que se observan en una cámara con los que se observan en otras, usando la detección de ejes principales. Khan es otro autor que ha trabajado con múltiples cámaras, y en (19) hace un recuento de las técnicas usadas para la correspondencia de objetos entre cámaras, mientras propone un nuevo etiquetado con miras a la autocalibración.

Estos trabajos mencionados hacen uso de múltiples cámaras para hacer seguimiento, sin embargo, el seguimiento es descrito en términos de espacios continuos en un espacio 3D o incluso sobre los planos imágenes, y aún cuando se habla de fusión en el trabajo de Hu, la información no es combinada, sino simplemente relacionada (La fusión de información implica la obtención de nueva información a partir de información de entrada). Hay pocos trabajos que hacen procesamiento de alto nivel y muchos menos que hacen fusión de información.

En (26) Nguyen presenta un modelo de aprendizaje y detección de actividades usando las trayectorias de movimiento, que aplica modelos jerárquicos ocultos de Markov. Los modelos jerárquicos ocultos de Markov, son una extensión de los Modelos Ocultos de Markov, pero incluyen la jerarquía en los estados ocultos. El reconocimiento de actividades en este trabajo se hace detectando la posición de objetos con respecto a un plano dividido en regiones con cinco puntos claves: estufa, armario, silla de comedor, refrigerador, silla de TV y puerta. La relación del trabajo de Nguyen con el modelo que se presenta aquí no reside en la identificación de actividades, sino en el uso de un sistema de múltiples cámaras y un seguimiento basado en marcas en el entorno. Las regiones en las que está dividido el piso del cuarto componen la capa más baja para la identificación de comportamientos, que en su investigación llegan a ser complejos (tal como diferenciar entre una comida normal y una botana). La identificación de posiciones que Nguyen hace está descrita en regiones y alimenta directamente a los modelos. Sin embargo, la información de las dos cámaras no es fusionada, sino que simplemente se conmuta entre sus vistas, quedando con la que se considere con mejor vista para cubrir la

secuencia. Otro trabajo que plantea también una problemática de detección de comportamiento y para el cual se usan regiones que representan zonas de interés en el escenario, es el de Wei Yan (33). En su trabajo se describe un sistema que analiza patrones de comportamiento en un espacio público, y cuenta, por ejemplo, cuántas personas se acercan a una fuente, la cual comprende una región de interés.

En (25) se presenta un sistema de vigilancia, que supervisa el tráfico de vehículos con sensores de velocidad y cámaras colocadas en una sección de carretera. 12 cámaras son colocadas sobre edificios. Las cámaras son colocadas con sus campos de vista cubriendo la misma sección que 8 sensores de flujo (sensores colocados sobre el asfalto). En este trabajo se utilizan algoritmos de visión en las cámaras, para generar las trayectorias de los vehículos. Los sensores de flujo entregan parámetros de velocidad y densidad de flujo automovilístico. En tal trabajo interesa conocer la interacción entre los vehículos usando la información tanto de las cámaras como los sensores de flujo automovilístico, tomando en cuenta que por separado cada tipo de sensores sería insuficiente (los de flujo no proveen de las trayectorias y para las cámaras las sombras y oclusiones producen errores). La fusión de información que se lleva a cabo en este trabajo ha sido altamente optimizada para la problemática que atacan: las cámaras sirven para calcular una aproximación de las trayectorias de los autos, mientras que los sensores de flujo son usados en puntos discretos para resolver problemas de oclusión y sombras. La fusión que se lleva a cabo en este trabajo de Malik y Russell es una fusión a nivel características, con una metodología muy restringida como para ser aplicada en un nivel más alto.

2.5. Resumen

La fusión de información es un término que en general se refiere a la combinación de datos, de manera que el resultado supere, en alguna medida de interés, a los datos individuales. Existen muchas formas de realizar combinación de in-

formación, dependiendo de la problemática, los tipos de datos, la calidad de la información, entre otros factores. Para describir los mecanismos de fusión se usan la *arquitectura* y el *modelo* (o algoritmo).

Las arquitecturas de fusión permiten describir la combinación de información enumerando las partes involucradas y sus relaciones, es decir, por lo general las arquitecturas son diagramas que detallan los componentes involucrados y la forma en la que interactúan, contemplando tanto los dispositivos como los módulos de procesamiento. En los componentes descritos por la arquitectura siempre existe al menos un módulo que lleva a cabo la integración de información. Para describir ese componente que conforma el módulo de fusión de información, se usan modelos matemáticos o algoritmos que varían según el propósito de la combinación, o el tipo de datos que se usen. Algunos de los modelos de fusión de información usados en la actualidad para el procesamiento de información a nivel decisión son derivados de los modelos de votación, probabilistas, teoría evidencial y la lógica difusa.

Cuando la información que se fusiona proviene de sensores, se usa el término *fusión de sensores*, y en particular, cuando los sensores son cámaras, como es el caso de este trabajo, se maneja el término *fusión de múltiples cámaras (o multicámara)*. En las tareas relacionadas con la fusión de información de múltiples cámaras, uno de los objetivos principales es el seguimiento de objetos en movimiento. Este propósito es clave en las tareas realizadas por los sistemas de vigilancia, donde se han empleado muchas de las técnicas de fusión de información conocidas, pero no es común encontrar trabajo donde realmente se combine la información de las fuentes para mejorar los resultados. Lo que parece ser común hoy en día, en cuanto al uso de múltiples cámaras, es la correspondencia de objetos en movimiento, es decir, se hacen coincidir los objetos en movimiento de una cámara con los de otra, como en los trabajos de Hu, Khan y Beynon ((15),(19),(3) respectivamente).

En los trabajos de Wei Yan (33) y de Nguyen (26) se usa información de alto nivel (relaciones entre objetos y posiciones definidas por zonas) para identificar

conductas o situaciones de interés. En estos trabajos no se fusiona la información de las cámaras, sino que se utilizan por separado.

En este trabajo se ha adaptado la teoría evidencial para la fusión de información en el contexto del seguimiento de objetos, donde la posición de los objetos es descrita de acuerdo a regiones en el plano en el que circulan. En el capítulo siguiente se presentan los fundamentos teóricos de la teoría evidencial, que se utilizará más adelante para la fusión de la información de las cámaras.

Capítulo 3

Modelo Dezert-Smarandache

3.1. Teoría Evidencial

La Teoría Evidencial o Teoría de Evidencia es una rama de las matemáticas del razonamiento con incertidumbre que permite considerar e integrar información para llegar a tomar decisiones. La teoría fue primero propuesta por Arthur Dempster y posteriormente modificada por Glen Shafer (28), por lo que también se conoce como Teoría Evidencial Dempster-Shafer, y puede ser interpretada como una generalización de la teoría de probabilidad, donde la asignación de probabilidad se hace sobre conjuntos en lugar de elementos mutuamente exclusivos (*singletons*). En algunos trabajos esta teoría es comparada con la teoría bayesiana (21; 7; Hoffman and Murphy; 17).

La aplicación de la Teoría Evidencial se ha visto reflejada en diversos campos. Por ejemplo, Guido Fioretti la compara en (10) con la teoría de Shackle para poder tomar decisiones económicas, por ejemplo cuándo invertir en farmacéutica y cuándo en biotecnología. Sin embargo, la aplicación más generalizada en los días recientes es en la fusión de información. En ese sentido, la parte principal de la DST es la *regla de fusión de Dempster*.

El modelo Dempster-Shafer demostró dar resultados útiles aplicado a fusión de

información, y sus principales ventajas sobre la teoría de probabilidad bayesiana son no requerir de conocimiento previo (probabilidades *a priori*) (7) y manejar ignorancia; además de que en la DST, la evidencia es representada con la función de creencia de Shafer, en lugar de una función de densidad de probabilidad. Sin embargo, cuando la evidencia proporcionada por las fuentes tiene ciertas condiciones, se producen deficiencias en los resultados de la combinación, como el conocido *resultado contraintuitivo*. En (29) se menciona que existe una clase infinita de casos donde la regla de Dempster puede asignar certidumbre a una opinión minoritaria. Por esto, trabajos posteriores propusieron reglas de combinación diferentes o incluso llegaron a generar nuevos modelos que no requirieran las probabilidades *a priori* del modelo bayesiano, pero que fueran robustos a cualesquiera condiciones de la evidencia. Por ello en la actualidad el término "Teoría Evidencial" no se restringe a la Teoría Dempster-Shafer.

Una alternativa reciente a los problemas encontrados en la DST, es la Teoría Dezert-Smarandache (DSmT), que toma como base a la DST, pero la extiende tanto en el modelo como en la regla de fusión para resolver problemas originados por fuentes de evidencia conflictivas, inciertas e imprecisas. Siendo un modelo matemático, éste no contempla ninguna característica relacionada con una aplicación en específico. Por ello es que para cada aplicación es necesario extender las funciones y tratar con problemas particulares. En nuestro caso se propondrá un modelo para fusión multicámara para seguimiento, que a su vez está basado en el modelo de la DSmT. Para poder explicar por qué se tomó la DSmT para modelar la problemática relacionada con la fusión multicámara y cómo es que este modelo se extiende con la finalidad de hacer fusión multicámara, primero es necesario introducir la DST y la DSmT.

3.2. Teoría Dempster-Shafer

La Teoría Demster-Shafer se basa en el modelo de Dempster-Shafer, que fue una modificación del modelo original de Dempster. En el modelo Dempster-Shafer se define un *marco de discernimiento* $\Theta = \{\theta_1, \theta_2, \cdots, \theta_n\}$ con n hipótesis θ_i que cumplen exhaustividad y exclusividad. Lo que las hipótesis θ_i representan dependen del contexto del problema que se está resolviendo.

Partiendo de Θ se define un conjunto 2^Θ, llamado *conjunto potencia de* Θ. Por ejemplo, si $\Theta = \{\theta_1, \theta_2, \theta_3\}$, entonces $2^\Theta = \{\emptyset, \theta_1, \theta_2, \theta_3, \theta_1 \cup \theta_2, \theta_1 \cup \theta_3, \theta 1 \cup \theta_2, \cup \theta_3\}$.

El modelo de Shafer considera tres funciones:
La *función de creencia básica* $m : 2^\Theta \rightarrow [0, 1]$ es asociada con un cuerpo de evidencia dado. Esto quiere decir que la evidencia nueva es representada con esta función, la cual es matemáticamente similar a la función de probabilidad.

$$m(\emptyset) = 0 \tag{3.2.1}$$

$$\sum_{A \in 2^\Theta} m(A) = 1 \tag{3.2.2}$$

Shafer definió las funciones de *creencia* y *plausibilidad* en $A \subseteq \Theta$ como

$$\mathrm{Bel}(A) = \sum_{B \in 2^\Theta, B \subseteq A} m(B) \tag{3.2.3}$$

$$\mathrm{Pl}(A) = \sum_{B \in 2^\Theta, B \cap A \neq \emptyset} m(B) = 1 - \mathrm{Bel}(A^C) \tag{3.2.4}$$

Cada observación hecha por una fuente de información provee evidencia sobre uno o más subconjuntos de todas las proposiciones de interés (2^Θ) y es posible representar ambigüedad o ignorancia asignando evidencia a Θ. Esta información es combinada mediante la regla de combinación de Dempster, lo cual es considerado algunas veces una generalización de la regla de Bayes, como en (Hoffman and Murphy) .

3.2.1. Regla de Dempster

La Regla de Dempster es la forma en la que las evidencias de diferentes fuentes se combina, según el modelo de la DST. Para k fuentes de evidencia, la regla de fusión está expresada de la siguiente manera:

$$m(\emptyset) = 0 \tag{3.2.5}$$

$$m(A) = \frac{\sum\limits_{\substack{X_1,X_2,\ldots,X_k\in 2^\Theta \\ X_1\cap X_2\cap\ldots\cap X_k=A}} \prod\limits_{i=1}^{k} m_i(X_i)}{1-\sum\limits_{\substack{X_1,X_2,\ldots,X_k\in 2^\Theta \\ X_1\cap X_2\cap\ldots\cap X_k=\emptyset}} \prod\limits_{i=1}^{k} m_i(X_i)} \qquad \forall A \neq \emptyset \in 2^\Theta \tag{3.2.6}$$

La normalización hecha con el denominador de la ecuación 3.2.6, sirve para eliminar las partes conflictivas de información entre las fuentes a combinar. El grado de conflicto entre dos fuentes está definido por

$$k_{12} = \sum_{\substack{X,Y\in 2^\Theta \\ X\cap Y=\emptyset}} m_1(X)m_2(Y) \tag{3.2.7}$$

Cuando dos fuentes tienen evidencias que no son coincidentes $k_{12} = 1$ se dice que los cuerpos de evidencia están en *completa contradicción*. Tal caso sucede cuando $\mathrm{Bel}_1(A)$=1 y además $\mathrm{Bel}_2(A^C)$=1, para algún $A \subset \Theta$. En general, cuando el conflicto entre las fuentes se vuelve alto (cuando k_{12} se acerca a 1) la combinación hecha con la regla de Dempster muestra poca lógica, a lo que se le conoce como resultado contraintuitivo. Para mayor referencia sobre los tipos de evidencia se puede consultar (27), donde además se hace un análisis detallado sobre otras reglas de combinación.

La teoría evidencial fue desarrollada como una alternativa de la teoría de probabilidad, debido a que algunas situaciones no pueden ser modeladas con las

herramientas de la probabilidad, como es el caso de la ignorancia. En la teoría evidencial los grados de creencia se asignan a subconjuntos, en lugar de a elementos del dominio de referencia o hipótesis. En su funcionamiento, la teoría evidencial no depende de probabilidades *a priori* para la combinación de fuentes de información. Para ejemplificar la diferencia entre cómo se consideran las creencias en la teoría probabilista y como se hace en la teoría evidencial, consideremos el caso en el que dos amigos apuestan sobre un partido de fútbol. Uno de ellos conoce de fútbol y sabe que la probabilidad de que gane el equipo A es mayor a la probabilidad de que gane el equipo B, pero el otro no sabe de fútbol y en un modelo probabilista se tendría por fuerza que asumir que la probabilidad de ganar es la misma para ambos equipos. En la teoría evidencial se puede representar la confianza que tiene cada participante en su apuesta, ya que se puede representar la ignorancia.

3.3. Teoría Dezert-Smarandache

Debido a las limitaciones de la Teoría Dempster-Shafer, Jean Dezert y Florentin Smarandache propusieron en el 2002 (9) un nuevo modelo que terminara con las limitaciones de la DST. Descrito de forma general, el objetivo era tener un marco de discernimiento menos restringido y una regla de fusión que tratara con fuentes conflictivas.

El resultado fue un modelo matemático con tres características que no tiene el modelo DS.

- La combinación de evidencia que refuta el tercer principio lógico o principio del tercer medio excluido en relación a los elementos pertenecientes al conjunto potencia de Θ en la DST. El tercer principio lógico establece que algo "es o no es". La DSmT no requiere de este principio, por lo que se permite considerar conceptos que se relacionan con la lógica difusa, que no tienen interpretación absoluta, como grande/pequeño, caliente/frío, etc.

- Se propone una nueva regla de combinación.
- Los elementos del marco de discernimiento no son necesariamente exhaustivos y exclusivos

En algunos problemas donde es necesario combinar información los conceptos involucrados no tienen una división marcada, tales como alto/pequeño, caliente/frío. Por ello el modelo DSmT propone un modelo donde las hipótesis o elementos del marco de discernimiento $\theta_i, i = 1, \ldots, n$ sólo tienen el requerimiento de ser exhaustivos, es decir, se pueden traslapar. A este modelo se le denomina *modelo libre DSm* y se denota como $\mathcal{M}^f(\Theta)$.

El modelo DSm además considera cuando el marco de discernimiento Θ es dinámico o varía con el tiempo. En ese caso algunos elementos de Θ podrían ser exclusivos en un momento, pero después no existir en otro. Por ello el modelo $\mathcal{M}^f(\Theta)$ puede restringirse para dar origen a un *modelo híbrido DSm* $\mathcal{M}(\Theta)$ el cual, si incluye todas las restricciones posibles se convierte en $\mathcal{M}^0(\Theta)$ que es idéntico al modelo DS.

Los fundamentos de la DSmT, en cuanto a combinación de información, son diferentes al resto de los trabajos propuestos para resolver los problemas de la DST, en cuanto a la forma de manejar incertidumbre, imprecisión y conflictos, que además promete cubrir una variedad más amplia de problemas (29).

3.3.1. Conjunto Hiperpotencia

Así como la DST tiene un conjunto potencia sobre el cual se consideran las evidencias de las fuentes, la DSmT considera un *conjunto hiperpotencia* D^Θ, que se construye a partir de un conjunto de n elementos exhaustivos $\Theta = \{\theta_1, \ldots, \theta_n\}$, llamado *marco*.

D^Θ es construido a partir de Θ siguiendo las tres reglas siguientes:

1. $\emptyset, \theta_1, \ldots, \theta_n \in D^\Theta$

2. Si $A, B \in D^{\Theta}$, entonces $A \cap B \in D^{\Theta}$ y $A \cup B \in D^{\Theta}$

3. Ningún otro elemento pertenece a D^{Θ}, excepto los obtenidos mediante 1 y 2

Esta definición del conjunto hiperpotencia D^{Θ} cumple con propiedades diferentes a las del conjunto potencia en la DST. La cardinalidad de D^{Θ}, cuándo Θ tiene n elementos, está acotada por 2^{2^n}, por lo que $\mid D^{\Theta} \mid \geq \mid 2^{\Theta} \mid$

Con estas reglas se pueden listar un par de ejemplos de conjuntos hiperpotencia:

- Para $\Theta = \{\theta_1\}$, $D^{\Theta} = \{\emptyset, \theta_1\}$.
- Para $\Theta = \{\theta_1, \theta_2\}$, $D^{\Theta} = \{\emptyset, \theta_1, \theta_2, \theta_1 \cap \theta_2, \theta 1 \cup \theta_2\}$.

3.3.2. Funciones de creencia generalizadas

Partiendo de un marco general Θ se define la función $m : D^{\Theta} \rightarrow [0, 1]$, llamada función de *asignación de creencia básica generalizada* (gbba), aplicada sobre un cuerpo de evidencia A:

$$m(\emptyset) = 0 \tag{3.3.1}$$

$$\sum_{A \in D^{\Theta}} m(A) = 1 \tag{3.3.2}$$

Las funciones de *creencia y plausibilidad generalizadas* también son usadas, tal como en la DST, pero sus definiciones se hacen usando el conjunto hiperpotencia D^{Θ} en lugar de 2^{Θ}

$$\mathrm{Bel}(A) = \sum_{\substack{B \subseteq A \\ B \in D^{\Theta}}} m(B) \tag{3.3.3}$$

$$\mathrm{Pl}(A) = \sum_{\substack{B \cap A \neq \emptyset \\ B \in D^{\Theta}}} m(B) \tag{3.3.4}$$

Los creadores del modelo DSmT definieron las funciones de creencia y plausibilidad de tal forma que fueran compatibles con las de la DST en el modelo $\mathcal{M}^0(\Theta)$, cuando D^Θ es reducido a 2^Θ. Así que se cumple que $\mathrm{Bel}(A) \leq \mathrm{Pl}(A)$, $\forall A \in D^\Theta$

3.3.3. Regla de combinación clásica DSm

Cuando se está considerando el modelo $\mathcal{M}^f(\Theta)$, la regla de combinación usada para un marco Θ con k elementos está definida como

$$m_{\mathcal{M}^f(\Theta)}(C) = \sum_{\substack{X_1, X_2, \ldots, X_k \in D^\Theta \\ X_1 \cap X_2 \cap \ldots \cap X_k = A}} \prod_{i=1}^{k} m_i(X_i) \tag{3.3.5}$$

3.3.4. Regla de combinación híbrida DSm

En algunos problemas de fusión es necesario considerar restricciones sobre los conjuntos a los que les es asignada evidencia, por lo que el modelo $\mathcal{M}^f(\Theta)$ no es suficiente y se tiene que usar en su lugar el modelo híbrido DSm $\mathcal{M}(\Theta)$ que incluye restricciones que mantienen la integridad de los resultados.

Para este modelo híbrido la regla de combinación para dos o más fuentes está definida para cualquier $A \in D^\Theta$ por las siguientes funciones:

$$m_{\mathcal{M}(\Theta)}(A) = \phi(A)\left[S_1(A) + S_2(A) + S_3(A)\right] \tag{3.3.6}$$

$$S_1(A) = \sum_{\substack{X_1, X_2, \ldots, X_k \in D^\Theta \\ X_1 \cap X_2 \cap \ldots \cap X_k = A}} \prod_{i=1}^{k} m_i(X_i) \tag{3.3.7}$$

$$S_2(A) = \sum_{\substack{X_1, X_2, \ldots, X_k \in \emptyset \\ [\mathcal{U}=A] \vee [[\mathcal{U} \in \emptyset] \wedge [A = I_t]]}} \prod_{i=1}^{k} m_i(X_i) \tag{3.3.8}$$

$$S_3(A) = \sum_{\substack{X_1, X_2, \ldots, X_k \in D^\Theta \\ X_1 \cup X_2 \cup \ldots \cup X_k = A \\ X_1 \cap X_2 \cap \ldots \cap X_k \in \emptyset}} \prod_{i=1}^{k} m_i(X_i) \tag{3.3.9}$$

donde $\phi(A)$ es la *función característica de no-vacío* de un conjunto A definida

como

$$\phi(A) = \begin{cases} 1 & \text{si } A \notin \boldsymbol{\emptyset} \\ 0 & \text{otro caso} \end{cases} \tag{3.3.10}$$

donde $\boldsymbol{\emptyset} = \{\boldsymbol{\emptyset}_{\mathcal{M}}, \emptyset\}$ y $\boldsymbol{\emptyset}_{\mathcal{M}}$ es el conjunto de todos los elementos de D^{Θ} que han sido forzados a ser vacíos debido a las restricciones del modelo $\mathcal{M}$.
Las variables $\mathcal{U}$ e I_t se definen como

$$\mathcal{U} = u(X_1) \cup u(X_2) \cup \ldots \cup u(X_k) \tag{3.3.11}$$

$$I_t = \theta_1 \cup \theta_2 \cup \ldots \cup \theta_n \tag{3.3.12}$$

donde a su vez la función $u(X)$ es la unión de todos los *singletons* θ_i que componen X.

Cada una de las funciones S tiene un propósito dentro del modelo $\mathcal{M}(\Theta)$: $S_1(A)$ corresponde a la regla clásica de combinación de la DSmT para k fuentes independientes basada en el modelo libre $\mathcal{M}^f(\Theta)$. $S_2(A)$ representa la masa de todos los conjuntos vacíos, que es transferida a las ignorancias. $S_3(A)$ transfiere la suma de los conjuntos relativamente vacíos a los conjuntos no-vacíos.

3.3.5. Ejemplo

Para ilustrar las diferencias entre la DST y la DSmT se ilustra a continuación un ejemplo donde se aplican ambos modelos para fusionar. Consideremos dos fuentes de evidencia sobre $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$, con las asignaciones de creencia siguientes:

$$\begin{array}{llll} m_1(\theta_1) = 0.998 & m_1(\theta_2) = 0 & m_1(\theta_3) = 0.001 & m_1(\theta_4) = 0.001 \\ m_2(\theta_1) = 0 & m_2(\theta_2) = 0.998 & m_2(\theta_3) = 0 & m_2(\theta_4) = 0.002 \end{array}$$

La regla de combinación de Dempster, usada por la DST obtiene un resultado

que favorece a la hipótesis θ_4:

$$
\begin{aligned}
m_DS(\theta_4) &= \frac{\displaystyle\sum_{\substack{X_1,X_2,\ldots,X_k\in 2^\Theta \\ X_1\cap X_2\cap\ldots\cap X_k=\theta_4}} \prod_{i=1}^{k} m_i(X_i)}{1-\displaystyle\sum_{\substack{X_1,X_2,\ldots,X_k\in 2^\Theta \\ X_1\cap X_2\cap\ldots\cap X_k=\emptyset}} \prod_{i=1}^{k} m_i(X_i)} \\
&= \frac{m_1(\theta_4)\cdot m_2(\theta_4)}{1-(m_1(\theta_1)\cdot(m_2(\theta_2)+m_2(\theta_4))+m_1(\theta_3)\cdot(m_2(\theta_2)+m_2(\theta_4))+m_1(\theta_4)\cdot m_2(\theta_2))} \\
&= \frac{0.001\cdot 0.002}{1-(0.998\cdot(0.998+0.002+0.001+0.001)+0.001\cdot 0.002)} \\
&= \frac{0.000002}{0.000002} = 1
\end{aligned}
$$

Este resultado, obtenido con la regla de combinación de Dempster, es una muestra clara de como la DST favorece a las hipótesis en las que las fuentes de información coinciden, aún cuando la creencia asignada no sea alta. Eso significa que la DST ignorará la creencia alta, asignada en este caso a θ_1 y θ_2 por la primera y segunda fuentes respectivamente, y como resultado $m_{DS}(\theta_4) = 1$, lo cual es uno de los resultados conocidos como contraintuitivos.

La teoría DSm tiene dos reglas de combinación disponibles. La regla de combinación clásica de la DSmT, definida en la ecuación 3.3.5, da como resultado $m_{DSm}(\theta_4) = 0.000002$, $m_{DSm}(\theta_1 \cap \theta_2) = 0.996004$, $m_{DSm}(\theta_1 \cap \theta_4) = 0.001996$, $m_{DSm}(\theta_2 \cap \theta_3) = 0.000998$, $m_{DSm}(\theta_2 \cap \theta_4) = 0.000998$ y $m_{DSm}(\theta_3 \cap \theta_4) = 0.000002$

Por otra parte, si se aplica la regla de combinación híbrida (ecuación 3.3.6), restringiendo todas las intersecciones, la redistribución de creencia daría como resultado $m_{DSmh}(\theta_4) = 0.000002$, $m_{DSmh}(\theta_1 \cup \theta_2) = 0.996004$, $m_{DSmh}(\theta_1 \cup \theta_4) = 0.001996$, $m_{DSmh}(\theta_2 \cup \theta_3) = 0.000998$, $m_{DSmh}(\theta_2 \cup \theta_4) = 0.000998$ y $m_{DSmh}(\theta_3 \cup \theta_4) = 0.000002$. La selección de la regla de combinación clásica o la híbrida dependerá de las condiciones que tenga la problemática a la que se esté aplicando el modelo.

3.4. Resumen

La teoría evidencial DSm es un modelo matemático que permite combinar información de fuentes independientes, sean sensores, observadores o expertos. Surgió como una extensión de la teoría evidencial DS, pero usa dos reglas de combinación nuevas: la regla clásica y la híbrida.

La teoría de probabilidad bayesiana suele ser considerada rival comparativa de la DST y la DSmT. Las principales ventajas de la DSm, sobre la teoría de probabilidad bayesiana, son la consideración de ignorancia y no requerir de probabilidades *a priori*. Por su parte la DSmT también tiene ventajas sobre la DST: el uso de la regla de combinación híbrida permite agregar restricciones al modelo y una de sus principales cualidades es evitar resultados contraintuitivos, que se llegan a producir con la regla de regla de Dempster de la DST. Por su parte, una de las principales ventajas que tiene la teoría probabilista sobre la teoría evidencial es que es significantemente más simple y su implementación por lo general utiliza menos recursos computacionales.

Tanto en la DST como en la DSmT se define un marco de discernimiento $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, que contiene a las hipótesis θ_i. La DST trabaja con el conjunto potencia 2^Θ, mientras que la DSmT trabaja con el conjunto hiperpotencia D^Θ. La teoría DSm puede ser restringida para funcionar como la DS, pero evitando los resultados contraintuitivos que se obtienen con la DST. En la práctica la DSmT permite resolver un mayor número de problemas, al hacer posible que se utilicen conceptos relacionados con la lógica difusa.

En el capítulo siguiente se describirá la forma en la que la teoría evidencial fue adaptada para fusionar información de múltiples cámaras a nivel decisión.

Capítulo 4

Modelo de Fusión Multicámara

La utilidad práctica de los sistemas de vigilancia automática se centra en reducir el esfuerzo humano. Por ello muchos de los trabajos recientes se enfocan en tareas como la identificación de comportamiento sospechoso o reconocimiento de conductas. En ese contexto los objetos de interés podrían ser automóviles en el caso de un sistema para vigilar violaciones al reglamento de tránsito en una intersección, o personas en el caso de un sistema de vigilancia para un centro comercial, o maletas en el caso de un sistema de monitorización de equipaje en un aeropuerto. En fin, el número de aplicaciones para sistemas de este tipo es elevado, pero en la mayoría de los casos las condiciones (áreas de vigilancia extensas, presencia de zonas de oclusión, etc.) requieren el uso de un sistema multicámara.

El propósito de este trabajo es el de realizar fusión de la información de varias cámaras a un nivel alto, como lo es el nivel decisión, con la finalidad de reducir la incertidumbre y a la vez brindar información útil para la identificación de conductas, por lo que es necesario tener información de dónde se encuentran los objetos de interés con relación a la escena.

Los capítulos anteriores nos han dado una idea del campo de investigación de fusión de información aplicada al uso de varios sensores, especialmente cámaras. También se ha dado a conocer la teoría de combinación de información DSm.

La fusión de información de cámaras puede llevarse a cabo a diversos niveles y la información que se combina puede variar tanto en su naturaleza como en su propósito. El propósito del modelo que se describirá a continuación es identificar la posición de un objeto en movimiento descrita por zonas predefinidas en el escenario, y a diferencia de otros trabajos que llevan a cabo la integración de varias cámaras tomando información 3D de la escena, como el de Weiming Hu (15) o de Sohaib Khan (19), en este trabajo la fusión se lleva a cabo a nivel decisión. La principal ventaja de realizar la fusión a nivel decisión es que el resultado de la fusión es la posición del objeto, según la información combinada, mientras que en otros niveles de fusión se requiere procesamiento posterior.

Este capítulo describe el proceso seguido para adaptación de la teoría evidencial DSm a la problemática de fusión multicámara. Como se describirá a continuación, esta tarea conlleva la solución de varios problemas intermedios, que comprenden la creación del marco de discernimiento Θ en forma dinámica, creación de la función de asignación de creencia básica generalizada y las restricciones de $\emptyset_{\mathcal{M}}$, que de no haber sido resueltos no habrían permitido la aplicación de la Teoría Evidencial a la fusión multicámara.

El modelo matemático DSm, si bien fue diseñado para la combinación de información, no había sido aplicado a problemáticas relacionadas con la fusión de información de cámaras. De hecho, hasta el momento de publicación del presente trabajo no se encontró trabajo previo relacionado con la fusión a nivel decisión para la determinación de la posición de los objetos en movimiento. Su adaptación para resolver el problema de fusión de decisiones en el seguimiento de objetos plantea varias problemáticas.

La primera de las problemáticas involucra la creación de las hipótesis θ_i que componen el marco de discernimiento, que si bien dependen del número de zonas predefinidas en la escena, deben manejarse de manera dinámica para optimizar recursos computacionales. La segunda problemática está relacionada con la asignación de creencia por parte de las cámaras, es decir, cada cámara debe

comportarse como un experto y aportar su creencia sobre las posibles posiciones que puede tener el objeto en movimiento. Además de manejar dinámicamente el marco de discernimiento, se aprovechó una de las características principales de la DSmT: la definición de las restricciones $\emptyset_{\mathcal{M}}$. Esta característica permite excluir en los cálculos aquellas hipótesis que son completamente improbables.

Junto con la adaptación del modelo, los sistemas multicámara requieren de la alineación de datos, esto significa que la información que se obtiene de todas las cámaras debe tener un sistema común de referencia espacial y temporal, así se evitan ambigüedades con la información que se maneja.

4.1. Arquitectura de fusión

Existe un gran número de arquitecturas que se han usado en problemas de fusión de información (de sensores en general) y que se encuentran descritas en (12). Para la problemática que se aborda en este trabajo fue necesario crear una arquitectura que llevara a cabo la alineación espacial y que resolviera el problema de correspondencia entre objetos en escena valiéndose de la homografía, por lo que requiere el uso de módulos específicos de procesamiento interactuando antes de realizarse la fusión.

En la figura 4.1 se muestra el diagrama de arquitectura, dividido por colores para agrupar las tareas de procesamiento. En la zona coloreada de azul se muestran los módulos para detección de movimiento, que son una tarea realizada para cada una de las imágenes obtenidas con las cámaras. La región mostrada en verde agrupa los módulos cuya tarea es, en síntesis, relacionar la información que proviene de las cámaras, tanto en tiempo como en espacio. En este grupo de módulos se lleva a cabo la alineación de datos. La zona ilustrada en amarillo engloba únicamente las tareas de fusión.

En el diagrama mostrado en la figura 4.1, la información es procesada en los módulos y transita sobre la misma línea hasta convertirse en decisiones y llegar

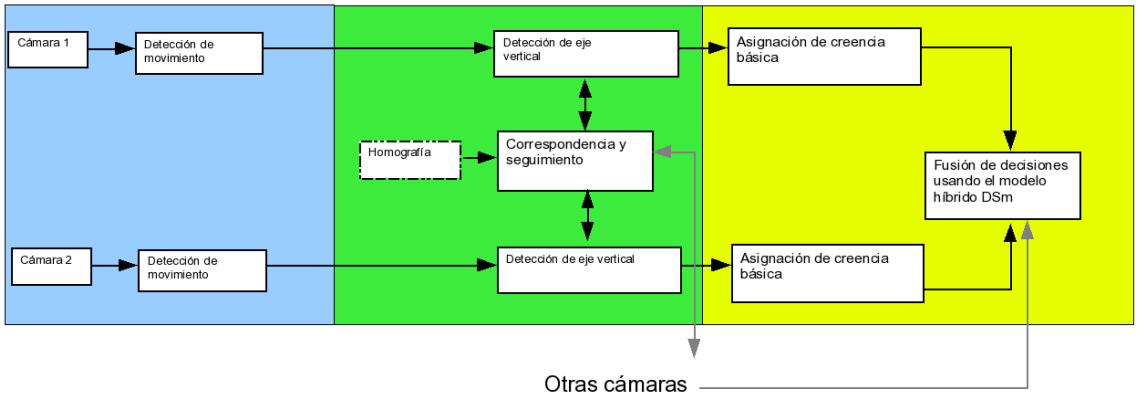


Figura 4.1: La arquitectura está dividida en tres partes: la adquisición de información y detección de movimiento; la sección de alineación de datos y la sección de fusión de información.

al módulo de fusión, y hasta ese momento la información es independiente. El módulo de correspondencia y seguimiento que se encuentra en el medio es usado únicamente para hacer corresponder los objetos en caso de que existan más de uno en escena y también para crear dinámicamente el marco de discernimiento Θ.

La arquitectura es extensible en el número de cámaras y se puede usar con tantas como sea posible en la práctica. La definición teórica no tiene restricciones en ese sentido.

Detección de movimiento

Para el módulo de detección de movimiento se seleccionó sustracción de fondo, que es uno de los algoritmos más eficientes y simples a la vez usando cámaras y campos de visión estáticos, es decir, cualquier cambio en la composición de la imagen se debe al movimiento de los objetos en la escena, o cambios en la iluminación, que por lo general son fácilmente tratados por la mayoría de los algoritmos existentes.

Una forma de obtener el modelo del fondo es con el promedio de valores de pixel a lo largo de la secuencia de entrenamiento. Supongamos que la secuencia está compuesta por N imágenes, es posible generar una matriz de promedios μ

donde cada celda representa un pixel, a partir de una secuencia de imágenes I_t, donde t es el número de imagen dentro de la secuencia. La siguiente ecuación representa la obtención del modelo del fondo.

$$\mu(i,j) = \frac{1}{N}\sum_{t=1}^{N} I_t(i,j) \tag{4.1.1}$$

Una vez que se tiene el modelo del fondo es posible calcular la diferencia, sólo que esta vez es la diferencia de la imagen actual con el fondo y no con la imagen anterior, como es el caso del algoritmo de diferencia de imágenes.

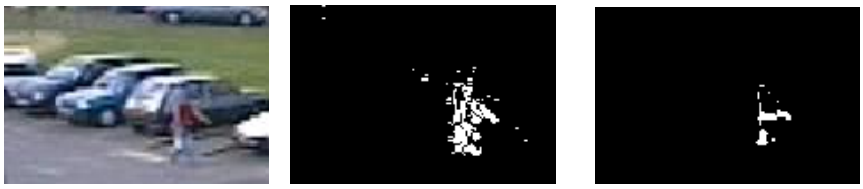
Un enfoque simple que usa el modelo de fondo y lo adapta en el tiempo es el algoritmo de Heikkila (14). En este algoritmo se define un fondo adaptable B_t para una imagen obtenida en un tiempo t. Al igual que en todos los algoritmos de extracción de fondo, se asume que los pixeles contienen movimiento si la diferencia absoluta entre la imagen I_t y el fondo B_t sobrepasa un umbral k.

Para la actualización del modelo de fondo la idea es ir integrando la información de la imagen actual al modelo del fondo, lo cual se consigue con el filtro recursivo:

$$B_{t+1} = (1-\alpha)B_t + \alpha I_t \tag{4.1.2}$$

donde α es un coeficiente de adaptación, que entre mayor sea su valor hará que los cambios en la escena se incorporen más rápido al modelo del fondo. El coeficiente de adaptación, sin embargo, no debería ser muy grande, ya que eso haría que todos los movimientos se incorporaran al fondo muy rápidamente. En la práctica α suele ser pequeño para permitir que los cambios en la iluminación y nuevos objetos que se vuelven parte de la escena se incorporen al fondo.

En la figura 4.2 es posible observar la comparación entre la detección de movimiento hecha con el algoritmo de diferencia absoluta de imágenes y la extracción de fondo adaptativa. La extracción de fondo muestra ser menos sensible al ruido, a la vez que muestra una región más acorde con las dimensiones del objeto en su



(a) Imagen en t (b) Diferencia absoluta (c) Extracción de fondo

Figura 4.2: Comparación de detección de movimiento por diferencia de imágenes y extracción de fondo

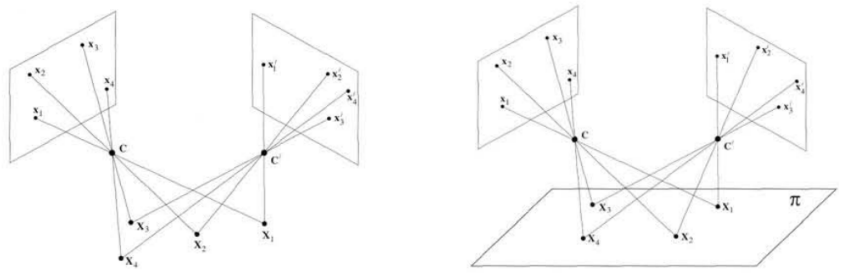
posición actual, mientras que en la de diferencia de imágenes se aprecia que la región obtenida no corresponde exactamente al contorno del objeto en movimiento en su posición actual.

4.1.1. Alineación espacial

El proceso de captura de una escena del mundo real en una imagen es expresado como

$$\begin{pmatrix} x \\ y \\ w \end{pmatrix} = P_{3\times 4} \begin{pmatrix} X \\ Y \\ Z \\ T \end{pmatrix} \tag{4.1.3}$$

En el caso de los sistemas multicámara, es necesario llevar a cabo una correspondencia entre las múltiples cámaras, para lo cual se utilizan restricciones derivadas de la homografía. Para simplificar la calibración de las cámaras y poder fijar un sistema de alienación espacial, se asume que los objetos en movimiento se desplazan sobre una superficie plana, lo cual es el caso de la mayoría de las aplicaciones de vigilancia. De esta forma la posición de los objetos observados en alguna de las imágenes obtenidas con las cámaras pueden ser relacionados con puntos en el *plano piso* y también en las imágenes obtenidas con las otras cámaras, tal como lo ilustra la figura 4.3(b). Para esta representación la ecuación 4.1.1 se simplifica



(a) Proyección de puntos en 3D (b) Proyección de puntos coplanares

Figura 4.3: Proyección de puntos en espacio 3D y coplanares.

a la transformación proyectiva

$$\begin{pmatrix} x \\ y \\ w \end{pmatrix} = H_{3\times3} \begin{pmatrix} X \\ Y \\ T \end{pmatrix} \tag{4.1.4}$$

donde la matriz $H_{3\times3}$ es la matriz de homografía

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{4.1.5}$$

Al asumir que todo lo que se observa en los planos imagen de las cámaras corresponde a la proyección del plano piso, es posible hacer una correspondencia entre los puntos del plano piso y los de los planos imagen. Esta característica nos permite relacionar los objetos en movimiento reportados por las cámaras, al mismo tiempo que nos brinda un sistema de referencia y medición común. Además de esto, asumir que se comparte un plano piso nos asegura la existencia de la homografía y es un requerimiento cumplido en la mayoría de las escenas monitorizadas. Las zonas problemáticas para emplear este modelo serían aquellas que tuvieran superficies irregulares sobre las cuales transitaran los objetos en movimiento, tales como montañas o escaleras. Sin embargo posiblemente aún sería

posible usar este modelo si se consideran subregiones en las imágenes plano, tal como se sugiere en la sección de conclusiones y trabajo futuro.

Para hacer corresponder dos puntos de los planos imagen de dos cámaras, (x_i, y_i) y (x'_i, y'_i), los puntos pueden ser asociados con $H_{3\times3}$ de la siguiente forma:

$$\begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{4.1.6}$$

En cuanto a la matriz de homografía $H_{3\times3}$, es posible recuperarla de un conjunto de puntos estáticos en el plano piso (31) o información dinámica de la escena (5). Con estos puntos se resuelve el sistema $Am = b$ calculando $m = (A^T A)^{-1} A^T b$

$$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i & y_i & 1 & 0 & 0 & 0 & -X_i x_i & -X_i y_i \\ 0 & 0 & 0 & x_i & y_i & 1 & -Y_i x_i & -Y_i y_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{pmatrix} = \begin{pmatrix} \vdots \\ X_i \\ Y_i \\ \vdots \end{pmatrix}$$

A su vez, la correspondencia entre objetos detectados en las cámaras puede ser llevada a cabo siguiendo técnicas de correspondencia por características (22) o técnicas geométricas (19; 4). En este trabajo se han usado puntos estáticos sobre el plano piso, pero el módulo de Homografía en la arquitectura de la figura 4.1 puede ser sustituido por el que más convenga.

4.1.2. Alineación temporal

En el presente trabajo se aplicó el método de marcaje por tiempo sobre las imágenes, para lo cual se requiere tener un sistema de referencia temporal sincronizado en todos los sensores, así cada vez que se toman las imágenes, éstas son marcadas con el tiempo. Finalmente cuando la información circula a lo largo de la línea de procesamiento (véase la figura 4.1) la información puede ser alineada temporalmente desechando aquellas imágenes que corresponden a tiempos pasados en el módulo que lleva a cabo la correspondencia de objetos y el seguimiento, hasta que todas se refieran al mismo instante de tiempo.

4.2. Modelo de fusión multicámara

4.2.1. Descripción general

Del diagrama de arquitectura de la figura 4.1, se han descrito ya los módulos de detección de movimiento que van después de la adquisición de imágenes, el módulo de homografía que se usa para la alineación espacial y la manera en la que se realiza la alineación temporal. Lo que queda por ser descrito corresponde ya a los módulos que hacen corresponder la información de los sensores (cuadro verde del diagrama) y finalmente la fusionan (cuadro amarillo en el diagrama).

El modelo que se propone en este trabajo contempla el seguimiento de cada uno de los objetos localmente, de manera que cada cámara brinde información sobre los objetos que se mueven sobre su campo de vista independientemente de las otras cámaras y pueda usar algoritmos propios para refinar el seguimiento según sus condiciones de visión. Sin embargo, una vez que cada cámara ha localizado los objetos en movimiento dentro de su campo de visión, esta información se debe compartir para que la correspondencia entre objetos en movimiento se lleve a cabo.

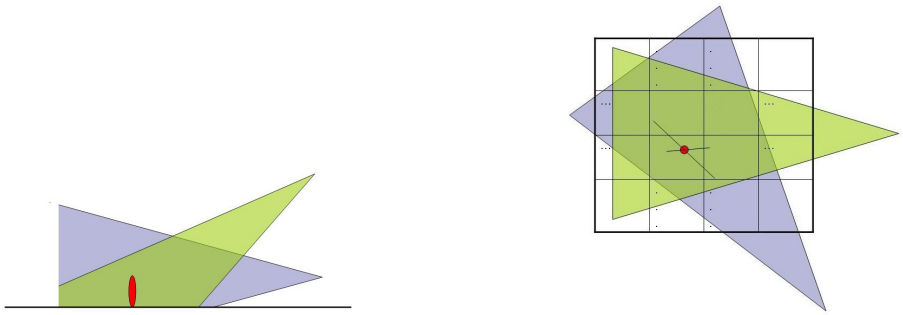
En la presencia de más de un objeto en la escena, los trabajos de Khan (19)

y de Hu (15) presentan técnicas para llevar a cabo la correspondencia de objetos entre los planos de las cámaras. En ambos trabajos el punto inferior del objeto (o pies) sirven para hacer la correspondencia, ya que es este punto el que indica unívocamente la posición del objeto sobre el plano. El algoritmo de Hu consiste en lo siguiente: se asume que en un tiempo t, M objetos con ejes verticales $L_1^i, L_2^i, \ldots, L_M^i$ son observados desde la cámara i y N objetos con ejes principales $L_1^j, L_2^j, \ldots, L_N^j$ de la cámara j. Se calcula una lista de todas las posibles correspondencias entre los pares de ejes principales. La lista de distancias es reducida quitando los pares cuya distancia sea menor a un umbral. Se calculan las posibles combinaciones con los pares restantes, y se toma la que tenga la menor suma de distancias entre ejes. Esta lista de pares es la correspondencia entre los objetos.

Enfoques más sencillos, como la correspondencia por color, tamaño, alguna otra característica o características combinadas también son funcionales, aunque Khan y Hu argumentan que son menos robustos.

Una vez hecha la correspondencia, después de la etapa de segmentación del objeto en movimiento, se pueda procesar cada objeto por separado para poder fusionar su posición con respecto a las zonas predefinidas en el plano piso. Esta parte del modelo se encuentra ilustrada por el módulo denominado *correspondencia y seguimiento* en la figura 4.1.

El eje vertical proyectado, de cada uno de los objetos, por cada cuadro de video, será usada para determinar la zona sobre la cual se encuentra el objeto con respecto al plano piso, como se muestra en la figura 4.4. Una vez hecha la correspondencia de objetos en movimiento, el módulo de correspondencia y seguimiento genera dinámicamente las hipótesis que las cámaras podrán usar para asignar sus creencias y son devueltas al módulo de detección de eje. El paso siguiente en la línea de procesamiento de cada cámara es que con las hipótesis disponibles, cada cámara asigne su creencia con respecto a la posición del objeto. La información a fusionar son las creencias de las zonas sobre las que cada objeto se encuentra, asignadas por las cámaras. En las secciones siguientes se describirá como es que



(a) Perspectiva de cámaras (b) Proyección del eje vertical del objeto

Figura 4.4: En (a) se muestran las perspectivas de las cámaras y en (b) las correspondientes proyecciones del eje vertical del objeto sobre el plano piso.

tales creencias son construidas y finalmente fusionadas.

4.2.2. Representación de incertidumbre

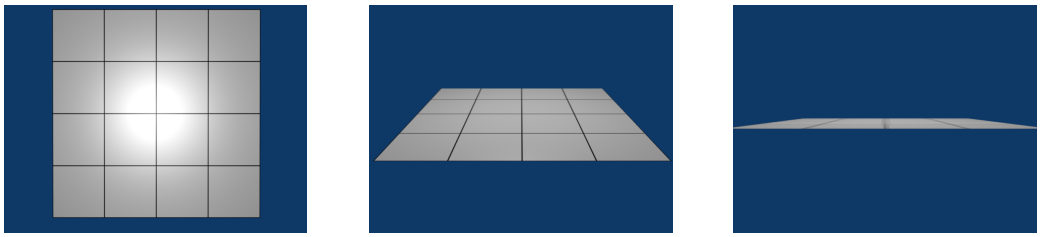
Un problema de la percepción por computadora es que se debe trabajar siempre considerando la incertidumbre que conlleva el uso de sensores. Al trabajar en visión por computadora, las cámaras son sensibles a ciertos problemas de distorsión derivados de los tipos de lentes, la calidad de los CCDs, o la posición de la cámara, que pueden ser reducidos con la calibración. Sin embargo, los parámetros de calibración no pueden llegar a eliminar algunas distorsiones que son originadas por la perspectiva cuando se tienen puntos en el espacio tridimensional como referencia, ya que la transformación del mundo 3D al 2D siempre produce una pérdida de información.

En la sección 4.1.1 se ilustró el procedimiento de adquisición de imágenes desde las cámaras. Así mismo, se mostró que usar un plano piso como referencia para los puntos sirve para el proceso de correspondencia entre cámaras y reconstrucción de las posiciones de los objetos en movimiento. Sin embargo, aún cuando se asuma que los objetos en movimiento circulan sobre el mismo plano, las posiciones de los objetos, localizados sobre ese plano son susceptibles a distorsiones relacionadas con la naturaleza de los lentes usados en las cámaras y con la perspectiva.

En la figura 4.5 se muestra un escenario generado por computadora donde el plano piso está dividido por una cuadrícula, para ilustrar las distorsiones que sufre al ser proyectado en el plano imagen cuando las imágenes son capturadas por las cámaras. Entre más cercana se coloque la cámara a un ángulo paralelo al plano piso, ésta pierde noción de profundidad con respecto al plano piso. Para poner un ejemplo de como la perspectiva afecta a la localización de objetos en movimiento, supongamos que existe una cámara posicionada en un ángulo muy cercano al del plano piso (como en la figura 4.5(c)). Si un objeto transita sobre el plano piso en dirección norte a sur o viceversa, el movimiento será casi imperceptible al ser representadas en la imagen. Un efecto aún más importante para los propósitos de este trabajo es el que surge al interpretar las imágenes para tratar de posicionar el objeto en movimiento sobre el plano piso, ya que al procesar la imagen obtenida con la cámara cualquier error en la detección de movimiento (oclusión, cambio de forma, segmentación del objeto, o cualquier otro tipo de distorsión común en el manejo de imágenes) hace parecer que el objeto se desplaza grandes distancias en la dirección norte o sur, sobre el plano, mientras que tales movimientos pueden ser inexistentes. Esto se debe a que el plano piso se encuentra representado en una zona muy pequeña de la imagen. La influencia del ángulo de la cámara es tan grande en la determinación de la posición del objeto sobre el plano piso, que en la imagen que toma la cámara puede llegar al punto de perderse toda noción de profundidad con respecto al plano piso.

El procedimiento seguido para la recuperación de la posición del objeto a partir de la imagen consiste en detectar la base del objeto en la imagen obtenida con la cámara. Este punto es después proyectado sobre el plano piso usando la transformación proyectiva correspondiente a la homografía previamente obtenida, con lo que se calcula su posición sobre el plano piso.

La influencia de la perspectiva en el procedimiento de recuperación de posición podría darnos resultados erróneos, por lo que en este trabajo se propone considerar la incertidumbre derivada de la perspectiva, en función del ángulo de la cámara



(a) Perspectiva superior (b) Perspectiva a 45 grados (c) Perspectiva baja

Figura 4.5: Influencia de la perspectiva en la incertidumbre de la información. Cuando el ángulo de la cámara se reduce también se pierde información de profundidad con respecto al objeto sobre el plano piso.

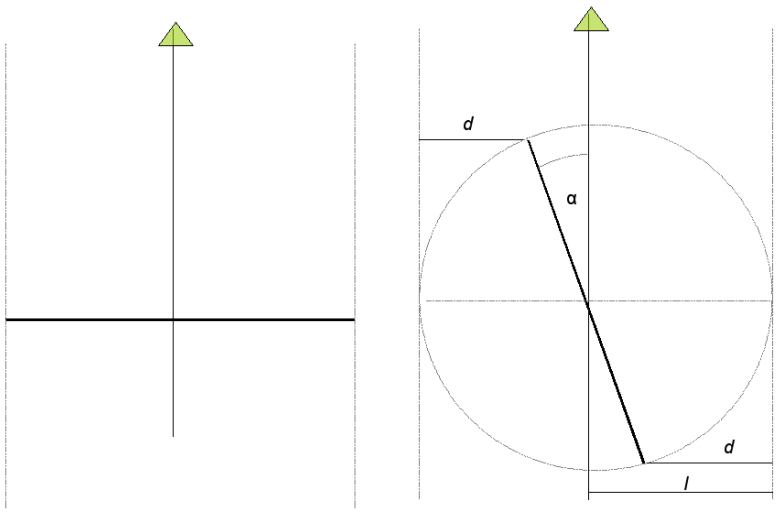
con respecto al plano piso.

Para la representación de incertidumbre, el ángulo de la cámara con respecto al plano piso juega un papel muy importante dentro del modelo. La colocación ideal de la cámara es a un ángulo recto con respecto al plano, cubriendo con una vista superior el escenario, con lo que no habría incertidumbre en la determinación de la posición del objeto.

Uno de los objetivos de este trabajo es que la fusión de información de las cámaras pueda reducir tales distorsiones de perspectiva, complementando la información de las cámaras. Una cámara pierde información de profundidad por los efectos de la perspectiva, pero conserva la información correspondiente a las otras direcciones, así la información de profundidad perdida con respecto al plano piso por una cámara puede ser compensada con la información de otras cámaras si tienen perspectivas diferentes.

En la figura 4.6 se muestran dos vistas. En la primera ilustración el plano piso está ortogonal a la cámara, lo cual correspondería a la vista en la cual no se tendría incertidumbre, mientras que en la segunda se muestra el plano piso girado, correspondiente a una cámara colocada a un ángulo α con respecto al plano piso.

La técnica que se propone en este trabajo para representar la incertidumbre es usar la proyección del eje vertical de los objetos en movimiento, y variar la longitud de la proyección de acuerdo a la perspectiva de la cámara. Con esta técnica, una



(a)Vista ideal de cámara (b) Plano girado

Figura 4.6: Representación del ángulo de la cámara con respecto al plano piso

cámara situada ortogonalmente al plano piso representa la posición de un objeto en movimiento con un punto, mientras que una cámara a un ángulo muy pequeño con respecto al plano piso proyectaría una línea, indicando la posible posición del objeto de acuerdo a su perspectiva, lo que en otras palabras significaría que el objeto en movimiento podría estar en cualquiera de los puntos de esa recta, de acuerdo a lo que esa cámara percibe.

La longitud de la proyección del eje es determinada por el ángulo de la cámara respecto al plano piso. Se establece la longitud de un eje de proyección como

$$\lambda = l\cos(\alpha) \tag{4.2.1}$$

donde l es la longitud máxima que el eje puede tomar y α es el ángulo medido, tal como lo indica la figura 4.6 . El triángulo en la parte superior de la ilustración representa la cámara. El ángulo al que está la cámara con respecto al plano puede ser visto como un giro del plano con respecto a la cámara.

Detección y proyección de eje vertical

La detección de movimiento descrita en la sección 2.4.1 sirve dentro de la arquitectura para encontrar los objetos en movimiento, pero como paso previo para la generación de las decisiones locales (antes de la fusión), es necesario localizar los ejes verticales de los objetos en movimiento y calcular la longitud de la proyección del eje, como ya se describió en la sección anterior.

Para encontrar el eje vertical de los objetos en movimiento se debe partir de la segmentación previamente hecha. Sobre las regiones segmentadas que corresponden a los objetos en movimiento, la detección de los ejes principales usada en este trabajo está basada en el trabajo de Khan (19), abarcando detección de ejes para objetos aislados y detección de ejes para objetos en grupo.

En el caso de la detección del eje principal en objetos aislados se calcula el centro del objeto promediando las coordenadas de los pixeles que lo componen y se toma únicamente la coordenada en y, lo cual es un enfoque rápido y simple. Sea I una imagen binaria obtenida en el proceso de detección de movimiento, compuesta por todos los pixeles $p_i = (x_i, y_i)$ que conforman la región del objeto detectado, donde x y y son las coordenadas horizontal y vertical respectivamente.. Sea $n = |I|$. Se calcula la región ξ correspondiente al eje del objeto como

$$\xi = \{(x, y)|x = \frac{1}{n}\sum_{k=1}^{n} x_{I_k}; y \in [y_{pies} - \lambda, y_{pies} + \lambda]\} \tag{4.2.2}$$

donde y_{pies} es el renglón más bajo de la región que corresponde al objeto en movimiento en la imagen I (los pies del objeto), y λ se aplica como se definió en la ecuación 4.2.1. Una ilustración del eje del objeto a ser proyectado se muestra en la figura 4.7.

Cuando dos o más objetos están cerca puede pasar que al hacer la segmentación de regiones se consiga una sola región para todos los objetos, por lo que también se puede hacer un análisis del histograma de conteo por columnas de los pixeles

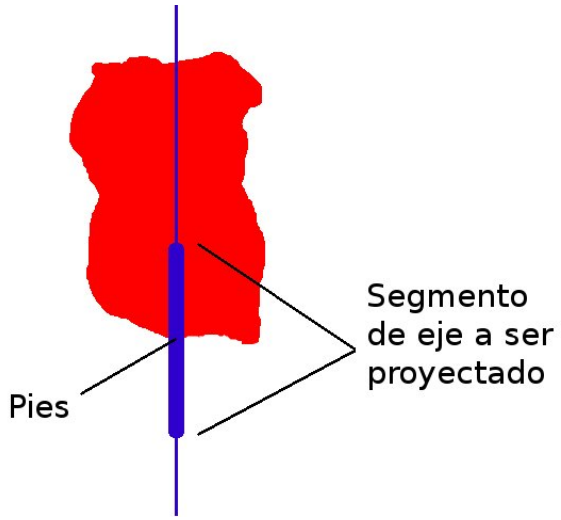


Figura 4.7: Ilustración del eje vertical y los pies del objeto. La longitud del segmento del eje vertical proyectado se calcula con el ángulo de la cámara.

en la imagen(19), para encontrar más de un eje vertical en la región.

4.2.3. Generación dinámica del marco Θ

Tal como se ha mencionado al comienzo de este capítulo, el propósito de este modelo es fusionar la información de las cámaras, a un nivel alto, por lo que las decisiones corresponden a las posiciones de objetos en movimiento descritas de acuerdo a zonas predefinidas sobre el plano piso. Si cada cámara se va a comportar como un experto, la cámara puede aportar evidencia sobre cualquiera de las zonas predefinidas, ya que tales zonas componen las hipótesis. Sin embargo, para propósitos de implementación de la Teoría Evidencial DSm, se encontró que considerar todas las regiones para fusionar no es necesario, y de hecho es perjudicial en cuanto al tiempo de ejecución.

La razón por la cual es mejor tener en Θ el menor número de elementos posibles es que su tamaño influye en el tamaño de D^{Θ}. En la sección 3.3.1 se describió la forma en la que se compone el Conjunto Hiperpotencia D^{Θ} usado para la aplicación de la regla de fusión híbrida DSm a partir de las hipótesis en Θ.

Como hipótesis de Θ deben considerarse sólo las regiones de interés para el proceso de fusión, de esta forma el tamaño de D^{Θ} se puede reducir considerablemente y por lo mismo optimizar el tiempo que tarda la fusión. A continuación se describe el procedimiento seguido para la construcción de Θ con el objetivo de tener el mínimo de elementos.

Sea $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ una partición en el plano piso, donde cada γ_x es una zona predefinida del plano, la cual puede ser de especial interés, tal como un corredor o zona de estacionamiento, ya sea un polígono regular o no. Las regiones γ_x son consideradas como candidatos a convertirse en hipótesis θ_i. Para no tener cada γ_x como hipótesis cuando no es necesario, se optimiza la construcción de Θ, tomando en cuenta sólo aquellas regiones γ_x sobre las cuales al menos una cámara asigne creencia. Eso significa que la creación de Θ se hace dinámicamente a medida que se van asignando las creencias.

Para que se pueda formar Θ, es necesario saber a cuáles regiones las cámaras les están asignando creencia. El proceso de asignación de creencia se describirá en la siguiente sección, sin embargo, es útil adelantar que la asignación de creencia se hace tomando en cuenta la longitud de la proyección del eje vertical del objeto sobre el plano piso. Teniendo esta proyección es posible encontrar todas las regiones con las cuales esta proyección se intersecta, como en la figura 4.6(b). Para esto se siguen técnicas de operación de conjuntos entre los pixeles que conforman las regiones y los que definen la proyección del eje vertical, todos ellos trabajados en el plano piso.

Sea ξ el conjunto formado por los pixeles que componen el eje vertical de un objeto, proyectado sobre el plano piso, tal como se definió en 4.2.2. Se utiliza el siguiente algoritmo para generar el marco Θ.

- $\Theta = \emptyset$
- Por cada cámara hacer
    - Por cada elemento de Γ hacer

- Si $\gamma_x \cap \xi \neq \emptyset$ entonces
  - $\Theta = \Theta \cup \{\theta_i\}$, donde i es un índice que se va incrementando y corresponde unívocamente a x

Si se cumple que $\exists \gamma_x | \gamma_x \cap \xi \neq \emptyset$ entonces se genera un θ_i correspondiente a γ_x. Al final el marco Θ estará compuesto sólo por elementos a los cuales se sabe que se hará asignación de creencias. Si ninguna cámara detecta objetos en movimiento, el marco Θ estará vacío y no se llevará a cabo la fusión, simplemente se concluye que no hay objetos en movimiento.

4.2.4. Asignación de creencias

Para cada objeto en movimiento i, es creado un marco $\Theta_i = \{\theta_1, \ldots, \theta_k\}$. El proceso de fusión de información por cada objeto puede ser llevado a cabo de manera concurrente, por lo que de aquí en adelante se tratará la problemática por cada uno de los objetos, es decir, como si hubiera sólo un objeto en movimiento en la escena.

Una de las partes más importantes del proceso de fusión es la formulación de decisiones en las cámaras. Cada cámara se comporta como un experto y proporciona una creencia a las regiones sobre las cuales el objeto puede estar situado. En esta parte se toma en cuenta el ángulo de la cámara, con el enfoque basado en la proyección del eje vertical del objeto sobre el plano piso, que se describió en la sección 4.2.2 .

Para que una cámara pueda asignar su creencia se requiere construir la función asignación de creencia básica generalizada (*gbba*). En este caso tal función nos permitirá manifestar la certidumbre que tiene una cámara sobre la posición del objeto en movimiento con respecto a las regiones predefinidas sobre el plano piso.

Una vez que se tiene el marco Θ definido, por el algoritmo descrito en la sección anterior, las cámaras pueden asignar creencia, para lo cual se utiliza el eje proyectado sobre el plano piso, ya que este depende de la perspectiva de la

cámara y por lo tanto representa la incertidumbre que ésta tiene de acuerdo a su posición (ángulo con respecto al plano piso).

Para un marco Θ general, se define una función $m_s : D^{\Theta} \rightarrow [0, 1]$, para cada cámara s. Esta función es la función de asignación de creencia básica generalizada ($gbba$).

La fusión de información multicámara es una herramienta para los sistemas de vigilancia de alto nivel. El resultado de la fusión bien podría ser sólo la zona que tiene la mayor creencia. Sin embargo, para no perder generalidad, en este trabajo la creencia resultante puede ser cualquiera de los elementos de D^{Θ}, y las entradas al mecanismo de fusión también pueden ser creencias simples o compuestas. Es decir, cada uno de los sensores puede asignar creencia a hipótesis del tipo θ_i, que son hipótesis en Θ, o elementos compuestos en D^{Θ}, tales como $\theta_i \cap \theta_j$ o $\theta_i \cup \ldots \theta_k$. Se deja abierta la posibilidad de asignar creencia a elementos compuestos del tipo $\theta_i \cap \theta_j$, ya que podrían representar intersecciones entre dos regiones en adyacentes sobre el plano piso y podrían ser usados por sistemas de interpretación de comportamientos.

En la sección 4.2.3 se mostró que debido a la construcción dinámica de Θ es fácil inferir que si ninguna cámara propone hipótesis no tiene caso ser realizada la fusión, ya que no habrá objetos en movimiento. Sin embargo, cuando $\Theta \neq \emptyset$ pero una cámara c no pudo generar el eje proyectado, esto significa que la cámara no ve al objeto, pero alguna otra sí lo hace. La cámara que no percibe al objeto se declara completamente ignorante, asignando $m_c(\theta_1 \cup \ldots \cup \theta_n) = 1$ donde $n = |\Theta|$. Con lo que dejará que sean las otras cámaras las que influyan en el resultado de la fusión.

Cuando una cámara sí puede determinar el eje del objeto, se procede a la asignación de la creencia, la cual depende de la intersección de área entre el eje proyectado y las regiones de Γ. Inicialmente, por las definiciones de la regla de fusión DSm, cada cámara tiene una unidad para asignar como creencia, que debe ser dividida entre los elementos de D^{Θ} a los que corresponda. En la figura 4.8 se

presentan gráficamente los casos considerados para la asignación de creencia. A continuación se describe el procedimiento seguido para asignar creencia en cada caso.

Dado que dentro de D^{Θ} existen tanto elementos simples (θ_i, θ_j, ...) como elementos que son compuestos (que se forman por la unión o intersección, de dos o más elementos de D^{Θ}), es necesario dividir el total de creencia a asignar, que equivale a la unidad, en dos cantidades, una para los elementos simples y otra para los compuestos. La cantidad de creencia a ser asignada a elementos compuestos en D^{Θ} es proporcional al ángulo de la cámara, lo cual hace que sea mayor cuando la cámara es perpendicular al plano.

Para los elementos simples el monto es el equivalente a $|\omega_c|$, donde ω_c es la región del eje vertical del objeto en movimiento, proyectado sobre el plano piso (medido en pixeles). Para los elementos compuestos dentro de D^{Θ}, se genera también un monto de creencia a ser distribuida, que dependerá del ángulo de la cámara. Esta cantidad de creencia υ a ser asignada a los elementos compuestos, es una fracción de $|\omega_c|$, que se encuentra con la siguiente expresión:

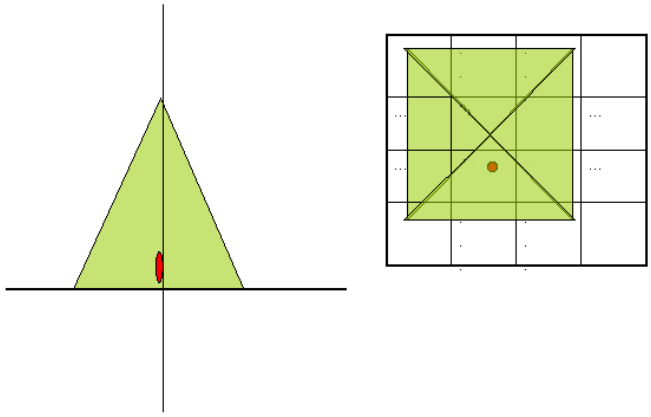
$$\upsilon = |\omega_c|\cos(\alpha_c) \tag{4.2.3}$$

Para la asignación de creencia a los elementos simples, se calcula el área de intersección entre ω_c y $\gamma_x \in \Gamma$, para la cámara c, por cada región en Γ y se normaliza dividiendo entre la suma de $\upsilon + |\omega_c|$.

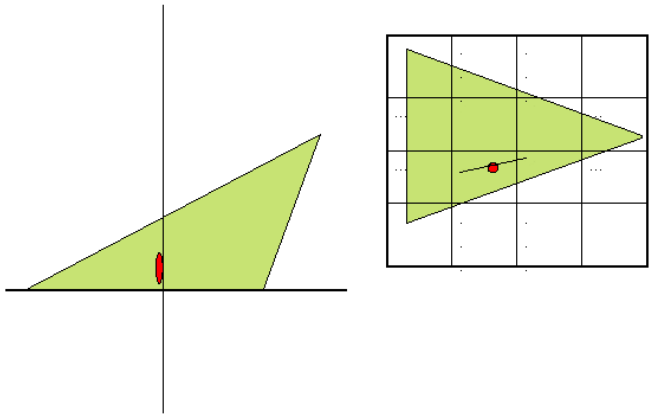
$$m_c(\theta_i) = \frac{|\omega_c \cap \gamma_x|}{\upsilon + |\omega_c|} \tag{4.2.4}$$

donde θ_i corresponde a γ_x.

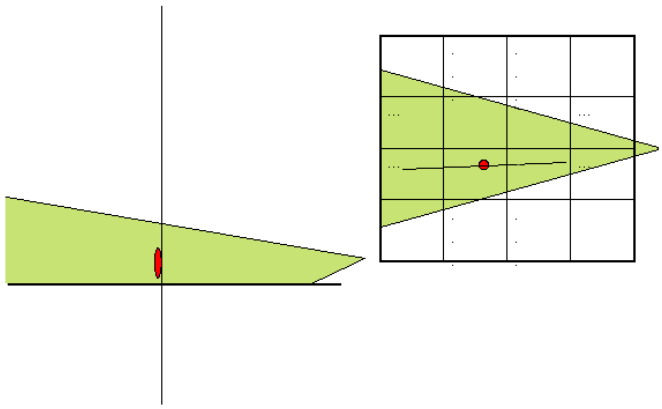
Para esta asignación directa se está dividiendo $|\omega_c|$ entre las regiones de Γ, estas áreas corresponden a los elementos θ_i en D^{Θ}. Se introduce el ángulo α_c de la cámara c ya que este ángulo representa la incertidumbre que tiene la cámara.



(a) Vista superior



(b) Vista media



(c) Vista baja

Figura 4.8: Casos de asignación de creencia por parte de la cámara. (a) Se asigna sólo a elementos simples de D^{Θ}. (b) Se hace asignación de creencia sobre elementos simples, y además, dependiendo del ángulo de la cámara se asigna a $\theta_i \cup \theta_j$ y $\theta_i \cap \theta_j$. (c) Se hace asignación de creencia sobre elementos simples y además del tipo $\theta_i \cup \theta_j \cup .. \cup \theta_k$

Entre menor sea el ángulo de la cámara, mayor será el monto a asignar a los conjuntos compuestos.

Para conservar generalidad con los sistemas de identificación de comportamiento o conductas, es posible asignar creencia a conjuntos compuestos como intersección. Se contempla este caso sólo para regiones adyacentes. Cuando sólo existen dos regiones con las cuales ω_x se intersecta (digamos que están representadas en D^Θ por θ_i y θ_j), entonces υ se puede dividir entre $\theta_i \cup \theta_j$ y $\theta_i \cap \theta_j$. La forma en la que esta división de creencias se hace es considerando nuevamente el ángulo α_c al que está la cámara, pero ahora solamente para dividir υ entre lo que corresponderá como creencia a $\theta_i \cup \theta_j$ y a $\theta_i \cap \theta_j$, mostrado en las ecuaciones 4.2.5 y 4.2.6. Esta característica es sólo opcional en el modelo, y su uso depende de la aplicación y de si la creencia resultante sobre los conjuntos compuestos del tipo $\theta_i \cup \theta_j$ es útil como decisión final. En caso de no ser útil la información, la fusión puede ser restringida para no considerarlos, al incluirlos en $\emptyset_{\mathcal{M}}$.

$$m_c(\theta_i \cup \theta_j) = \frac{\upsilon \cos(\alpha_c)}{\upsilon + |\omega_c|} = \frac{|\omega_c| \cos^2(\alpha_c)}{\upsilon + |\omega_c|} \tag{4.2.5}$$

$$m_c(\theta_i \cap \theta_j) = \frac{\upsilon(1 - \cos(\alpha_c))}{\upsilon + |\omega_c|} \tag{4.2.6}$$

En este modelo se considera la asignación de creencia a la intersección de dos hipótesis de Θ, es decir, elementos del tipo $\theta_i \cap \theta_j$ que sean elementos no restringidos de D^Θ. Tales elementos describen la creencia que tiene una cámara en que el objeto se encuentra en el borde entre las regiones γ_x y γ_y de Γ (γ_x corresponde a θ_i y γ_y corresponde a θ_j) , pero no tiene sentido hablar de intersecciones de más de tres hipótesis, ya que esto gráficamente correspondería sólo a vértices de regiones en Γ, es decir, puntos específicos sobre el plano piso, y eso es irrelevante para los objetivos de este trabajo.

Cuando ω_c se intersecta con más de dos regiones en Γ, entonces la creencia disponible para elementos compuestos se asigna a la ignorancia local, formada por

la unión de las hipótesis θ_i, θ_j, ..., θ_k

$$m_c(\theta_i \cup \theta_j \cup \ldots \cup \theta_k) = \frac{\upsilon}{\upsilon + |\omega_c|} \tag{4.2.7}$$

4.2.5. Uso de $\emptyset_{\mathcal{M}}$ para reducción de tiempo de ejecución

Como el módulo de fusión está basado en el modelo híbrido de la Teoría DSm, se aprovecha también la definición de la *función característica de no-vacío*, ϕ, que se define para cualquier $A \in D^{\Theta}$ como $\phi(A) = 1$ si $A \notin \emptyset$ y $\phi(A) = 0$ en otro caso.

El conjunto $\emptyset_{\mathcal{M}}$ es llamado *conjunto relativamente vacío*, y está formado por elementos de D^{Θ} que son forzados a ser vacíos. Por convención el conjunto vacío clásico $\emptyset \notin \emptyset_{\mathcal{M}}$. El conjunto $\emptyset = \{\emptyset, \emptyset_{\mathcal{M}}\}$ denota al conjunto de todos los elementos relativamente y absolutamente vacíos.

Para reducir el número de operaciones en la fusión, la regla de combinación híbrida incluye la función característica de no-vacío ϕ, que evalúa cuando tiene sentido calcular $S_1(A)$, $S_2(A)$ y $S_3(A)$. Eso significa que en los conjuntos restringidos no tiene caso hacer todas las operaciones. En la definición 3.3.4 de la regla de combinación híbrida del modelo DSm, el cálculo de $S_1(A)$, $S_2(A)$ y $S_3(A)$ no es necesario cuando $\phi(A) = 0$, y eso se cumple cuando $A \in \emptyset$.

Dado que los elementos de Θ están estrechamente relacionados con los elementos de Γ, es posible introducir restricciones sobre los elementos de D^{Θ} e introducirlos en $\emptyset_{\mathcal{M}}$ y por lo tanto en $\emptyset$. Estas restricciones tienen que ver con las propiedades topológicas de las regiones en Γ. Así por ejemplo, si dos regiones γ_x y γ_y no son vecinas, no tienen un borde común, por lo que no tiene sentido considerar ninguna creencia asignada a $\theta_i \cap \theta_j$ (donde θ_i corresponde a γ_x y θ_j corresponde a γ_y) , y es posible hacer $\theta_i \cap \theta_j \in \emptyset_{\mathcal{M}}$, con lo que se reducen costos computacionales considerablemente, ya que no se tendrían que calcular $S_1(\theta_i \cap \theta_j)$, $S_2(\theta_i \cap \theta_j)$ ni $S_3(\theta_i \cap \theta_j)$, porque $\phi(\theta_i \cap \theta_j) = 0$.

Las propiedades de perspectiva de las cámaras ayudan al proceso de formación de $\emptyset$, ya que solamente se aceptan las intersecciones que se hayan introducido desde un principio al asignarse las creencias básicas. Es decir, se restringen los elementos $\theta_i \cap \ldots \cap \theta_k \in D^\Theta$ para los cuales no existe una asignación directa hecha por una de las cámaras, así que son incluidos en $\emptyset_{\mathcal{M}}$. Esto se debe a que un objeto puede estar en el borde entre dos regiones de Γ solamente cuando al menos una cámara lo considera así.

Las restricciones facilitadas por $\emptyset$ son uno de los beneficios ofrecidos por la Teoría DSm y resulta muy útil para la problemática de fusión multicámara, a diferencia de la regla de combinación clásica del modelo libre DSm.

4.2.6. Fusión de decisiones

Con lo descrito hasta el momento han quedado formuladas las funciones m en cada una de las cámaras. Estas funciones conforman las decisiones que serán fusionadas.

Dado el modelo híbrido de la Teoría Evidencial DSm, se consideran las funciones para fusionar las creencias que fueron asignadas mediante m_i por parte de cada cámara i. Sean $X, A \in D^\Theta$. Sea $I_t = \theta_1 \cup \theta_2 \cup \ldots \cup \theta_n$ la ignorancia total. Sea $u(A)$ la unión de todos los elementos simples θ_i que componen A.

$$m_{\mathcal{M}(\Theta)}(A) = \phi(A)\left[S_1(A) + S_2(A) + S_3(A)\right] \tag{4.2.8}$$

donde

$$
\begin{aligned}
S_1(A) &= \sum_{\substack{X_1,X_2,\ldots,X_k\in D^\Theta \\ X_1\cap X_2\cap\ldots\cap X_k=A}} \prod_{i=1}^{k} m_i(X_i) \\
S_2(A) &= \sum_{\substack{X_1,X_2,\ldots,X_k\in \emptyset \\ [u(X_1)\cup u(X_2)\cup\ldots\cup u(X_k)=A]\vee[[u(X_1)\cup u(X_2)\cup\ldots\cup u(X_k)\in\emptyset]\wedge[A=I_t]]}} \prod_{i=1}^{k} m_i(X_i) \\
S_3(A) &= \sum_{\substack{X_1,X_2,\ldots,X_k\in D^\Theta \\ X_1\cup X_2\cup\ldots\cup X_k=A \\ X_1\cap X_2\cap\ldots\cap X_k\in\emptyset}} \prod_{i=1}^{k} m_i(X_i)
\end{aligned}
$$

Dado que se introducen algunas restricciones en $\emptyset$, el modelo híbrido $\mathcal{M}(\Theta)$ tendrá algunos elementos $A \in D^\Theta$ para los cuales $\phi(A) = 0$, por lo que no tendrá caso de ser calculado $S_1(A)$, $S_2(A)$ ni $S_3(A)$.

Para los elementos A, cuya $m_{\mathcal{M}(\Theta)}(A)$ sí necesita ser calculada, se aplica la regla de combinación clásica, que corresponde a $S_1(A)$ y posteriormente se hace la transferencia de masas de las restricciones de integridad del modelo híbrido $\mathcal{M}$, de acuerdo a la fórmula 4.2.8..

La fusión comienza siempre en el modelo libre DSm y la transferencia de masas (S_2, S_3) se aplica sólo bajo las restricciones de $\emptyset$, para obtener el resultado final.

4.2.7. Ejemplo

En esta sección se da un ejemplo sencillo para ilustrar el funcionamiento de la fusión DSm. Supóngase que se tiene un plano piso dividido en 4 regiones de interés que forman una cuadrícula, y que se tienen dos cámaras colocadas ortogonalmente. Las regiones están numeradas de izquierda a derecha y de arriba hacia abajo como $\gamma_1, \gamma_2, \gamma_3$ y γ_4.

Supongamos que un objeto se encuentra posicionado en la región γ_2. Por las posiciones de las cámaras, el eje vertical del objeto, proyectado en el plano piso sería parecido al que se muestra en la figura 4.9. De acuerdo con las proyecciones

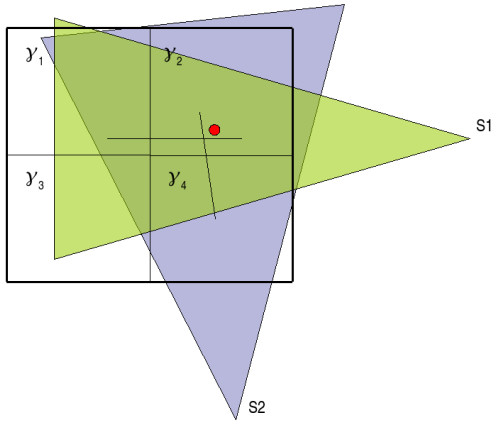


Figura 4.9: Los sensores detectan el objeto, calculan el eje vertical y lo proyectan sobre el plano piso. En esta ilustración, la posición real del objeto es marcada con un punto, pero los sensores la desconocen.

del eje vertical, las regiones sobre las que podría estar el objeto son γ_1, γ_2 o γ_4. Con estas regiones se construye el marco de discernimiento $\Theta = \theta_1, \theta_2, \theta_4$.

Para el marco de discernimiento descrito, los sensores pueden asignar creencia a los elementos $\theta_1, \theta_2, \theta_4, \theta_1 \cup \theta_2, \theta_1 \cup \theta_4, \theta_2 \cup \theta_4$ y $\theta_1 \cup \theta_2 \cup \theta_4$ de D^{Θ}.

Para algunos sistemas de reconocimiento de características (como los derivados de la lógica difusa, por ejemplo) podría ser útil que se permitieran las intersecciones de hipótesis, porque podrían representar que el objeto está en el borde entre dos regiones. Salvo por ese caso, ninguna de las intersecciones contenidas en D^{Θ} tiene sentido en las aplicaciones de los sistemas de vigilancia. En este ejemplo no tiene sentido permitir que las intersecciones sean consideradas para recibir creencia, y por eso todas ellas se incluyen en $\emptyset_{\mathcal{M}}$, con lo que no acumularán creencia.

En el modelo las creencias se asignan dependiendo del ángulo de las cámaras y del área del eje vertical proyectado. Para ejemplificar la fusión de creencias de las cámaras, a continuación se desarrolla un caso sencillo con valores de creencias hipotéticos. Supongamos que tales creencias son las siguientes:

$$m_1(\theta_1) = 0.3 \quad m_1(\theta_2) = 0.6 \quad m_1(\theta_1 \cup \theta_2) = 0.1$$
$$m_2(\theta_2) = 0.3 \quad m_2(\theta_4) = 0.3 \quad m_2(\theta_2 \cup \theta_4) = 0.4$$

El resultado de aplicar la regla de combinación híbrida, asignado a las hipótesis de Θ sería el siguiente:

$$\begin{aligned} m_{DSmh}(\theta_2) &= m_1(\theta_2) \cdot m_2(\theta_2) + m_1(\theta_2) \cdot m_2(\theta_2 \cup \theta_4) \\ &\quad + m_1(\theta_1 \cup \theta_2) \cdot m_2(\theta_2) + m_1(\theta_1 \cup \theta_2) \cdot m_2(\theta_2 \cup \theta_4) \\ &= 0.6 \cdot 0.3 + 0.6 \cdot 0.4 + 0.1 \cdot 0.3 + 0.1 \cdot 0.4 \\ &= 0.49 \end{aligned}$$

lo cual indicaría que el objeto se encuentra en la región γ_2.

4.3. Resumen

La arquitectura propuesta en este trabajo tiene el objetivo de adquirir los datos de las cámaras manteniendo una línea de procesamiento por sensor, hasta obtener una decisión por cada uno de ellos. Al mismo tiempo, es posible utilizar la información previa de alineación espacial para hacer la correspondencia de datos. La arquitectura propuesta permite la aplicación de procesamiento paralelo para los módulos de detección de movimiento y detección del eje vertical del objeto.

La detección de movimiento es una de las primeras etapas de procesamiento. Su finalidad es la de encontrar en la imagen aquellos objetos que cambian de posición. En este trabajo se seleccionó un algoritmo de detección de movimiento por extracción de fondo, que es simple y a la vez eficiente. La sustracción de fondo se hace teniendo un modelo del fondo de la escena, es decir, se parte de la observación de imágenes para estimar las partes que corresponden al fondo y por

diferencia encontrar los objetos en movimiento.

El módulo de alineación espacial permite relacionar la información de los sensores, de tal forma que se refieran al mismo tiempo y espacio. Para la alineación espacial se utiliza el plano sobre el que circulan los objetos como referencia. A este plano se le denomina plano piso. La correspondencia espacial se consigue transformando las posiciones del objeto con una matriz de homografía, para llevarlas de una cámara a un sistema de referencia común. La alineación temporal permite que la información corresponda en el tiempo. El mecanismo comúnmente usado en sistemas de fusión centralizados es el etiquetado de los datos por tiempo (*timestamping*).

Antes de llevarse a cabo la fusión de información de los sensores, las cámaras realizan un seguimiento local, independientemente de las otras cámaras. Cuando se tienen localizados los objetos en movimiento dentro del campo de visión, las cámaras comparten esa información para hacer la correspondencia de objetos. En esta etapa se hace la detección del eje vertical del objeto, que es la línea que cruza verticalmente (según el plano imagen de cada cámara) por el centro del objeto. El módulo de correspondencia y seguimiento genera dinámicamente las hipótesis del marco de discernimiento que las cámaras usarán para asignar sus creencias.

Los ejes verticales de los objetos son usados para generar las creencias sobre sus posiciones, representando la incertidumbre de las cámaras, derivada de sus perspectivas. El eje vertical es proyectado sobre el plano piso, intersectando la línea generada con las regiones del plano. La asignación de creencia es proporcional al área de intersección entre la proyección del eje y las regiones del plano piso. La cámara generará una línea de proyección más larga si está colocada paralela al plano piso. Por el contrario, si la cámara tiene una vista perpendicular al plano, la longitud de la proyección del eje vertical será pequeña, ya que una vista perpendicular corresponde a una vista superior, y es ahí cuando la cámara tiene la menor incertidumbre sobre la posición del objeto con respecto al plano. El eje vertical se proyecta sobre el plano piso teniendo como centro de la línea la posición

correspondiente a los pies del objeto, según cada cámara.

El marco de discernimiento Θ es generado dinámicamente, con el propósito de no contener a todas regiones del plano piso, sino sólo aquellas sobre las que las cámaras aporten evidencia. Los elementos θ_i son agregados a Θ solamente cuando al menos una cámara asigna creencia a su región correspondiente. Para cada objeto en movimiento es creado un marco Θ.

La función de asignación de creencia básica generalizada (gbba) permite manifestar la certidumbre que tiene una cámara sobre la posición del objeto en movimiento. Las cámaras pueden asignar creencia a hipótesis simples o compuestas, donde las compuestas pueden ser intersecciones o uniones de las hipótesis, todas ellas elementos de D^{Θ}. Cuando existe un objeto en movimiento, y al menos una cámara asigna creencia, las cámaras que no detectan al objeto en movimiento se declaran completamente ignorantes, asignando $m(\theta_1 \cup \theta_n) = 1$. dejando que las otras cámaras influyan en el resultado de la fusión.

La regla de fusión híbrida permite incluir restricciones en el modelo, mediante el uso de $\emptyset_{\mathcal{M}}$. Con estas restricciones se simplifican los cálculos al limitarse el número de intersecciones entre los conjuntos. Si el uso de intersecciones no es útil entonces todas las intersecciones se incluirán en $\emptyset_{\mathcal{M}}$. Para algunos sistemas de reconocimiento de actividades las intersecciones entre regiones podrían ser útiles, para lo cual se permite que las regiones adyacentes puedan quedar fuera de $\emptyset_{\mathcal{M}}$ y tener creencia final. Finalmente se aplica la regla de combinación híbrida para obtener una integración de las decisiones de las fuentes.

Capítulo 5

Desempeño del modelo

En este capítulo se presentan las evaluaciones del desempeño del modelo propuesto de fusión multicámara para seguimiento de objetos. También se presentan las pruebas de alineación espacial con una secuencia de imágenes tomadas de un escenario real. Se probó el modelo de fusión completo con secuencias sintéticas y reales, para medir la precisión de detección de posición de objetos en movimiento, de acuerdo a las regiones en Γ.

En la sección 5.1 se describen las secuencias de imágenes que sirvieron para realizar las pruebas. En la sección 5.2 se describen las pruebas de alineación espacial. Para finalizar, en la sección 5.3, se muestran los resultados de la implementación del modelo completo, probado sobre secuencias de imágenes sintéticas CGI (*computer generated imagery*) propias y secuencias reales del repositorio PETS (http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001 dataset.html)(Performance Evaluation of Tracking and Surveillance).

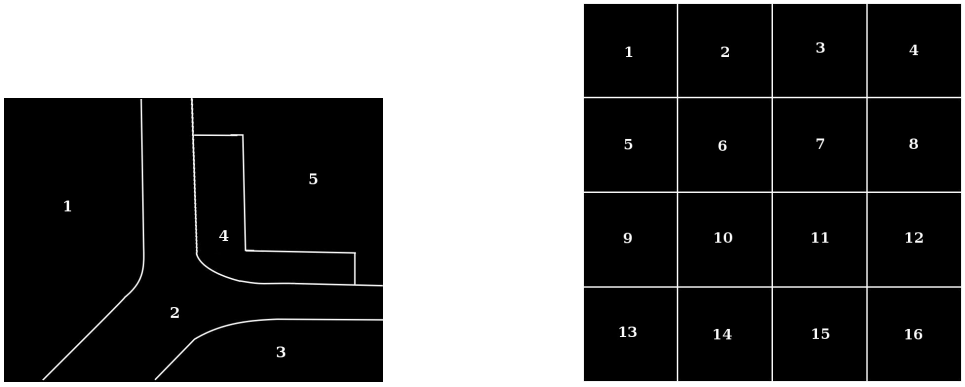
El modelo que se presenta en este trabajo considera la descripción de la posición de objetos en movimiento por regiones sobre el plano piso. Aún cuando existen otros trabajos en los que se fusiona información de múltiples cámaras, al momento de impresión de esta tesis no se encontró otro trabajo donde se hiciera fusión de información multicámara a nivel decisión con el propósito de hacer

seguimiento de objetos y que la posición de objetos fuera descrita en términos de regiones sobre el plano piso. Por esta razón no existe otro modelo en visión por computadora contra el cual se pueda comparar con fines de evaluación de desempeño. Sin embargo, dado que se desarrolló una arquitectura flexible y se tienen las decisiones obtenidas del módulo de asignación de creencias, fue posible adaptar un módulo de clasificación probabilista para que funcionara como módulo de combinación y comparar los resultados. Además se presentan comparaciones contra el seguimiento individual por cámara y contra la variación DSm-Punto, que es una modificación al modelo propuesto, y que se describirá más adelante.

5.1. Secuencias de pruebas

Las secuencias de pruebas constan de varios grupos de imágenes que corresponden a cámaras colocadas en un escenario, con campos de visión traslapados. Se usan dos tipos de secuencias: CGI y reales. Las secuencias CGI son animaciones generadas por computadora para cada una de las cámaras que cubren el escenario. Las secuencias reales corresponden al conjunto de datos PETS.

En la figura 5.1 se muestran las regiones definidas sobre el plano piso, para hacer seguimiento.



(a) Plano piso en secuencias PETS (b) Plano piso en secuencias sintéticas

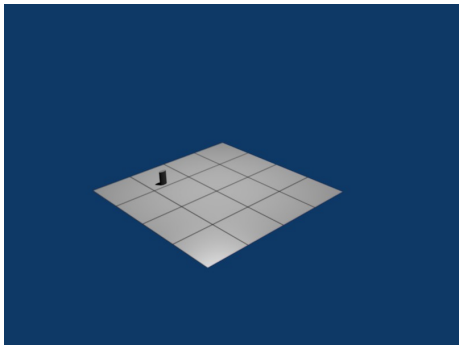
Figura 5.1: Partición Γ de los planos pisos usados para posicionamiento en pruebas.

5.1.1. Secuencias CGI

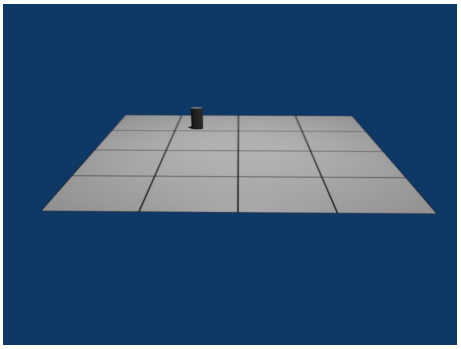
Para las secuencias CGI se consideraron tres cámaras virtuales cubriendo un escenario que contiene el plano piso, que está compuesto por 16 regiones regulares en forma de rejilla cuadrada (figura 5.2). Las imágenes fueron generadas con una resolución de 800x600 pixeles, usando el software de modelado 3D Blender (www.blender.org) y generando la secuencia con Yafray (www.yafray.org). Las secuencias sintéticas ofrecen mayor certeza de calibración, ya que los parámetros son conocidos y modificables y con ellas es posible tener los datos de seguimiento verdaderos automáticamente, para efectos de comparación. Se construyeron dos secuencias de 250 imágenes, con un objeto desplazándose por toda la cuadrícula del plano piso en diagonal y desplazándose por el centro con cambios de velocidad.

5.1.2. Secuencias reales

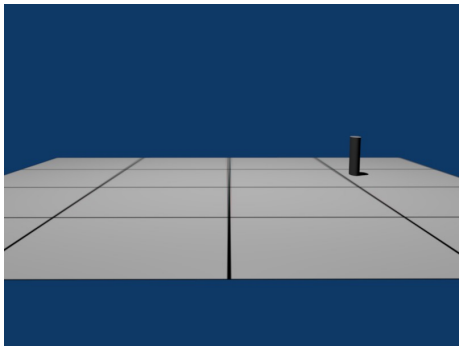
Como secuencias reales se usaron las bases de datos de PETS, que tiene en total 5752 imágenes, que es un repositorio con las secuencias multicámaras más grandes conocidas y que se utilizan en trabajos de calibración y seguimiento multicámara. Las secuencias corresponden a dos cámaras cuyos campos de visión se traslapan, pero además cubren zonas exclusivas, lo cual es útil para probar la ignorancia total en las cámaras. Las imágenes tienen una resolución de 768x576 pixeles están comprimidas en formato JPEG, y contienen una cantidad significativa de ruido, así que sirven además para probar el algoritmo de detección de movimiento. Ejemplos de esta secuencia se muestran en la figura 5.3. Se extrajeron cinco subsecuencias de condiciones variadas de iluminación, tamaños de objetos, etc. En las tres primeras se siguen personas (500, 250 y 850 imágenes respectivamente), en la cuarta un automóvil (550 imágenes) y en la última una bicicleta (340 imágenes). Las subsecuencias fueron seleccionadas de tal forma que los objetos circulan



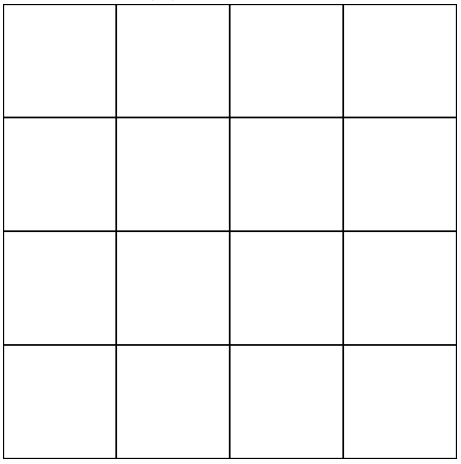
(a)Cámara 1



(b) Cámara 2



(c) Cámara 3



(d) Plano piso

Figura 5.2: Ejemplo de secuencias CGI. (a) La cámara está colocada con una vista desde la esquina del plano piso. En (b) la cámara es frontal y cubre todo el plano piso. En (c) la cámara está colocada a una altura baja, cubriendo sólo parte del plano piso.

(a)Cámara 1 (b) Cámara 2

Figura 5.3: Ejemplo de secuencias reales. Las cámaras 1 (a) está colocada a menor altura que la cámara 2 (b). Los campos de vista se traslapan.

cambiando de regiones sobre el plano piso y o en los bordes de las regiones. En un sistema de vigilancia se puede hacer seguimiento de múltiples objetos, cada uno con su propio marco de discernimiento y creencia asignada. Pero en este caso, para propósitos de pruebas, en cada secuencia se sigue un solo objeto (aún cuando siempre existe ruido derivado del movimiento de árboles y otros objetos), así es posible aplicar las métricas de evaluación y también medir tiempos de ejecución dedicados sólo a la fusión de información.

5.2. Pruebas de alineación espacial

Las pruebas de alineación espacial se realizaron sobre la secuencia real, utilizando las 5 marcas blancas colocadas sobre el asfalto en el crucero del escenario. Estas marcas fueron tomadas como puntos estáticos sobre el plano piso para el proceso de cálculo de homografía y son las usadas oficialmente para calcular los parámetros de calibración que se ofrecen junto con la secuencia de imágenes. La matriz de homografía H_{3x3} obtenida fue usada para la graficación del seguimiento de un objeto en movimiento en la secuencia real. La trayectoria del objeto es graficada en las figuras 5.4 y 5.5. Las figura 5.4 muestra la posición de los pies

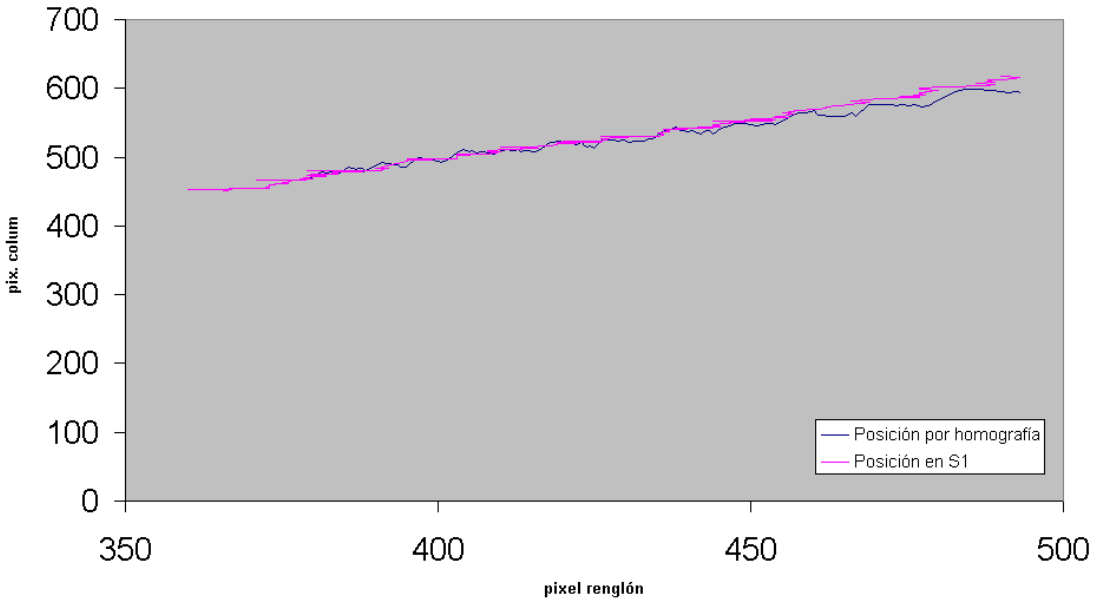


Figura 5.4: Comparación de trayectorias usando homografía de S1 a S2

del objeto en movimiento mostradas en las coordenadas del plano imagen de la cámara 1 ($S1$), en color rojo están la trayectoria obtenida por la cámara 1 y en color azul está la obtenida por la cámara 2 ($S2$), pero mapeada al plano imagen usando la homografía. La figura 5.5 muestra el caso semejante, pero la proyección hecha sobre el plano imagen de la cámara 2. En las gráficas, existe un momento en el que el objeto se encuentra fuera del campo de vista de la cámara $S2$, por lo que una de las líneas es más larga que la otra. Se obtuvo un promedio de 7.92 pixeles de diferencia entre ambas trayectorias, lo cual confirma una buena obtención de la homografía y una correcta alineación espacial. En la práctica es imposible hacer que ambas trayectorias coincidan, ya que siempre existen factores que intervienen en la detección del objeto en movimiento, tales como ruido, iluminación, etc.

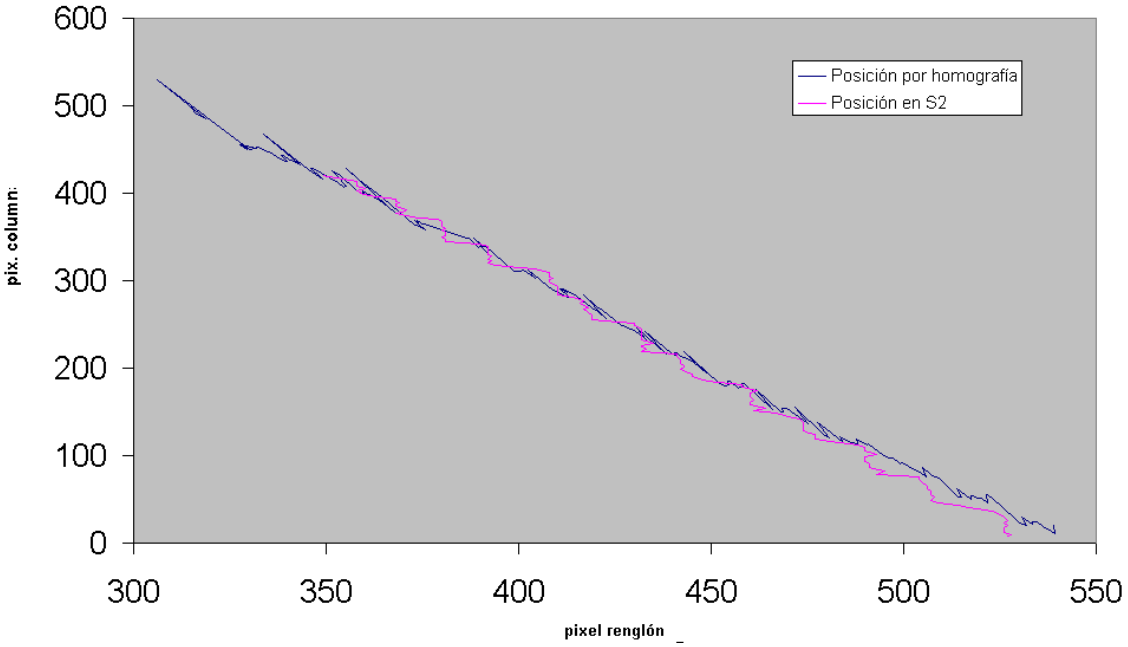


Figura 5.5: Comparación de trayectorias usando homografía de S2 a S1

5.3. Pruebas de Fusión Multicámara

Las pruebas de fusión multicámara contemplan todos los módulos de la arquitectura propuesta para hacer seguimiento, sólo se excluye el módulo de alineación espacial, el cual se ejecuta sólo para calcular la matriz de homografía y fue probado separadamente. Las pruebas de fusión multicámara fueron realizadas tanto con el conjunto de datos sintético como con el real.

Por cada una de las pruebas del DSm se hicieron tres comparaciones: comparación contra lo que se obtiene con una sola cámara (considerando cada cámara individualmente), comparación contra el modelo DSm considerando una incertidumbre baja (llamado DSm - Punto) y comparación contra el modelo probabilista.

La consideración de las cámaras individualmente es para tener una referencia de los datos de entrada al módulo de fusión. Los datos de las cámaras son tomados una vez que se ha hecho la asignación de creencias, es decir, en estos datos ya se considera la incertidumbre derivada de la perspectiva, por lo que la comparación

de éstos contra los obtenidos del módulo de fusión DSm sirven para observar los efectos de la combinación de información y los beneficios que se tienen sobre usar cada cámara individualmente.

La segunda comparación por cada prueba se hizo aplicando el modelo de fusión DSm, pero reduciendo la incertidumbre a la mitad. Esto se consiguió dividiendo entre 2 la longitud de la proyección del eje vertical y conservando la función de asignación de creencia gbba. A esta prueba se le denominó DSm-Punto, ya que gráficamente la incertidumbre es reducida y para algunas cámaras se llega a asemejar a no considerar incertidumbre en absoluto (la proyección del eje vertical se reduce a un punto). Esta prueba provee información sobre los efectos de considerar de incertidumbre a este nivel de fusión.

En fusión de información, uno de los enfoques más utilizado es el probabilista, con las distintas variantes bayesianas. En el campo de fusión de sensores, para aplicaciones de robótica principalmente, los modelos bayesianos han sido los más populares, y se han hecho comparaciones de ese modelo con modelos basados en la Teoría Evidencial (Hoffman and Murphy; 7). En el trabajo de Kettnaker (18) se fusionan probabilidades asignadas a objetos en movimiento en escenarios multicámara, pero a diferencia del modelo que se presenta en este trabajo, la fusión se aplica a niveles más bajos de procesamiento (detección de movimiento y correspondencia de objetos en tiempo, no en espacio) y con cámaras no traslapadas, lo cual hace la comparación imposible.

Dado que los modelos probabilistas pueden aplicarse para la combinación de información a niveles altos, se decidió aprovechar la arquitectura propuesta en este trabajo (figura 4.1) para crear un módulo de fusión probabilista, teniendo un clasificador bayesiano simple (Naive Bayes) por cada una de las regiones contenidas en el marco de discernimiento Θ, como se muestra en la figura 5.6. Se pudo entonces comparar también contra este modelo los resultados obtenidos con el modelo de Teoría Evidencial propuesto. Los clasificadores del módulo de fusión probabilista son similares al descrito en la sección 2.3.1. Para esta comparación, el

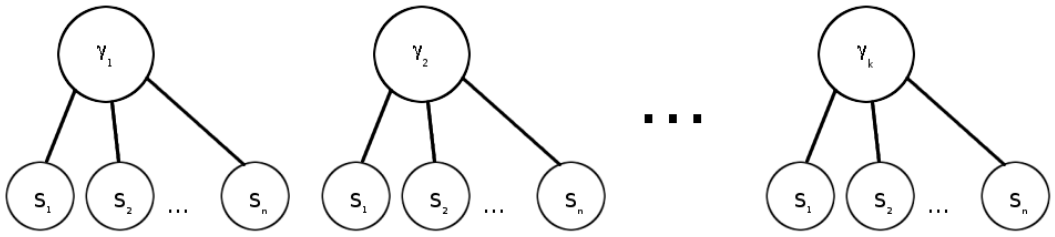


Figura 5.6: Clasificadores Bayesianos. Cada una de las k regiones del plano piso está representada por un clasificador, que recibe las probabilidades por parte de las n cámaras. Para optimizar recursos sólo se usan los clasificadores de las regiones que están representadas en el marco de discernimiento Θ.

primer paso fue tomar como base la arquitectura propuesta y cambiar el módulo de fusión con DSm por un módulo probabilista. El módulo de fusión probabilista tendría que fusionar entonces teniendo creencias como entrada, que son obtenidas con los mismos módulos de detección de movimiento, detección de eje vertical y se aplican los mismos algoritmos de correspondencia y seguimiento, bajo las mismas condiciones de alineación espacial y temporal. La creencia asignada por cada sensor es interpretada como la probabilidad condicional.

5.4. Comparaciones

5.4.1. Métricas de evaluación

Para poder cuantificar las diferencias entre las posiciones reales y las obtenidas por las diferentes fuentes (sean las cámaras individualmente o el resultado de la fusión), se usa una métrica de la calidad del seguimiento basado en regiones, partiendo de las métricas utilizadas para el seguimiento de objetos tradicional,

propuestos por Bashir y Porikli en (2). Para calcular la Taza de Detección de Seguidor y Taza de Falsas Alarmas se comparan los valores reales con los obtenidos por la fuente (cámara o fusión, según sea el caso), considerando todas las asignaciones de creencia. En las imágenes en las que la fuente y los datos reales coincidan en que existe un objeto (ambas asignan creencia), se considera como verdadero positivo (TP). Cuando los datos reales indiquen que hay un objeto en una región (creencia mayor a cero), pero la fuente no lo detecta (creencia nula), se cuenta como falso negativo. Si los datos reales indican ausencia de objetos en una región (creencia nula), pero la fuente considera que hay uno (creencia mayor a cero) entonces se cuenta como falso positivo (FP). Los valores de la tabla 5.1 y 5.2 muestran los porcentajes obtenidos de Taza de Detección de Seguidor (Tracker Detection Rate - TRDR) y Taza de Falsas Alarmas (False Alarm Rate - FAR), así como los tiempos promedio medidos en el módulo de fusión. Estas métricas fueron obtenidas promediando y midiendo en todas las pruebas realizadas. En las tablas no se incluyen los tiempos relacionados con las tareas de detección de movimiento o seguimiento, únicamente fusión.

Los valores de $TRDR$ y FAR son calculados como

$$\text{TRDR} = \frac{TP}{TG} \tag{5.4.1}$$

$$\text{FAR} = \frac{FP}{TP + FP} \tag{5.4.2}$$

$$\tag{5.4.3}$$

donde TG es el número total de regiones por cada imagen donde se encuentran objetos en movimiento de acuerdo a los datos reales. De acuerdo a estos datos, se espera que una fuente obtenga el mayor valor posible en $TRDT$, mientras que el valor de FAR sea el menor.

Los resultados de las tablas muestran que el desempeño DSm es mejor que el obtenido por cada una de las cámaras en todos los casos. Para las secuencias sintéticas, los resultados de la fusión DSm si bien no sobrepasaron a las dos prime-

Cuadro 5.1: Comparación en pruebas de secuencias sintéticas

Fuente	**TRDR**	**FAR**	**Tiempo fusión/imagen**
Cámara 1	99.5 %	52.9 %	No aplica
Cámara 2	93.9 %	43.0 %	No aplica
Cámara 3	84.4 %	45.3 %	No aplica
DSm	93.9 %	5.6 %	6.8ms
DSm - Punto	69.2 %	2.4 %	4.3ms
Probabilista	93.3 %	5.2 %	0.86ms

Cuadro 5.2: Comparación en pruebas de secuencias reales

Fuente	**TRDR**	**FAR**	**Tiempo fusión/imagen**
Cámara 1	68.1 %	21.7 %	No aplica
Cámara 2	71.0 %	2.7 %	No aplica
DSm	82.8 %	10.2 %	1.48ms
DSm - Punto	80.1 %	6.3 %	1.09ms
Probabilista	82.8 %	10.2 %	0.34ms

ras cámaras en la métrica $TRDR$ sí tuvieron un menor valor de la métrica FAR, lo cual indica que la fusión reduce el número de falsas alarmas y se aproxima al desempeño promedio de las cámaras en la métrica $TRDR$.

En las secuencias reales el $TRDR$ de la fusión DSm sí se incrementó sobre lo que ofrecen las cámaras, debido en gran medida a que las pérdidas del objeto por parte de las cámaras eran más frecuentes, haciéndose presente la ignorancia por parte de las cámaras. En este caso una cámara puede perder al objeto mucho tiempo, declarándose totalmente ignorante, pero mientras otra cámara pueda seguir al objeto el resultado de la fusión no se verá afectado. Si por ejemplo, una cámara puede seguir a un objeto la primera mitad de su trayectoria y otra lo sigue sólo la segunda mitad de tal trayectoria, cada cámara tendrá 50 % en la métrica $TRDR$, sin embargo, como resultado de la fusión, se tendrá 100 % en $TRDR$, ya que la combinación de información habría tomado la información de cada cámara para integrarla.

El modelo de fusión probabilista se desempeñó igual que el modelo DSm, excepto en una parte de las pruebas sintéticas, en las que se obtuvo una métrica $TRDR$ ligeramente menor a la que arroja el modelo DSm, pero por otro lado la taza FAR fue menor también. El comportamiento del modelo probabilista, cuando se consideran las creencias como probabilidades condicionales se comporta de manera muy semejante al modelo DSm. La única diferencia radica en que el modelo DSm, con la regla de combinación híbrida, opera sobre conjuntos y acumula también la información que aporta la ignorancia de las cámaras, mientras que en ese caso el modelo probabilista simplemente trabaja con la creencia asignada directamente a cada región, o asume probabilidades iguales para cada región cuando la cámara se declara ignorante.

El módulo de fusión probabilista fue adaptado para trabajar en las mismas condiciones que el modelo DSm que se usó como base en este trabajo. Se asume que es igualmente probable que el objeto esté en una región que en otra, es decir, las probabilidades *a priori* fueron asumidas iguales para cada región. Aún cuando este tipo de modelos puede ser pasado por etapas de entrenamiento para mejorar sus resultados, esta misma característica también constituye una desventaja en sí, al requerir de tiempo y datos de entrenamiento que el modelo DSm no requiere. Además de que, en caso de que el modelo no sea bien entrenado su desempeño empeoraría en vez de mejorar. En velocidad, el modelo probabilista confirmó que el tiempo de procesamiento es considerablemente menor, al realizar solamente multiplicaciones. Sin embargo, ambos modelos cumplen con tiempos de procesamiento que harían posible su aplicación en sistemas de tiempo real.

Para ilustrar de manera práctica las diferencias en comportamiento entre el modelo probabilista y el DSm, supongamos que se observa un objeto en movimiento sobre un plano piso dividido en dos regiones. Supongamos también que se tiene una cámara colocada con una vista alta del plano piso, es decir, la cámara tiene muy buena perspectiva para ver al objeto, por lo que asigna muy poca creencia a la ignorancia. Por otro lado se tiene otra cámara cuya perspectiva es paralela al

plano piso, por lo que genera ignorancia, además de asignar creencia a las regiones. Supongamos que las creencias dadas por la gbba son entonces las siguientes: $m_1(A) = 0.35$, $m_1(B) = 0.6$, $m_1(A \cup B) = 0.05$ $m_2(A) = 0.3$, $m_2(B) = 0.1$ y $m_2(A \cup B) = 0.6$.

El modelo probabilista tendría entonces los siguientes resultados:

$$\begin{aligned} p(A|S_1, S_2) &\propto 0.5 \cdot \frac{0.35}{0.35 + 0.6} \cdot \frac{0.3}{0.3 + 0.1} = 0.13 \\ p(B|S_1, S_2) &\propto 0.5 \cdot \frac{0.6}{0.35 + 0.6} \cdot \frac{0.1}{0.3 + 0.1} = 0.07 \end{aligned}$$

mientras que para el modelo DSm los resultados serían

$$m_{DSm}(A) = 0.35 \cdot 0.3 + 0.35 \cdot 0.6 + 0.05 \cdot 0.3 = 0.33$$
$$m_{DSm}(B) = 0.6 \cdot 0.1 + 0.6 \cdot 0.6 + 0.05 \cdot 0.1 = 0.42$$

La fusión probabilista no toma en cuenta la información de ignorancia generada por la segunda cámara, por lo que el resultado de la fusión es que el objeto está en la región representada por A, sin embargo, es claro que la información de primera cámara es más segura, ya que está mejor posicionada y por lo tanto debería tener más peso su aportación. Esto se consigue en el modelo DSm al sumar la creencia derivada de la intersección de la ignorancia de la segunda cámara con los conjuntos simples de la primera cámara. El resultado del modelo DSm es entonces que el objeto está en la región representada por B, lo cual concuerda mejor con las condiciones y la información provista por las cámaras.

Esta condición ocurre cuando las cámaras comparten una perspectiva del plano similar, pero en diferente ángulo con respecto al plano. Un ejemplo de esto podría ser un par de cámaras colocadas en el mismo edificio, pero en diferente nivel o piso. En las pruebas realizadas en este trabajo no se presentan estas condiciones,

sin embargo, este comportamiento debe ser tomado en cuenta en sistemas con cámaras que comparten vistas similares.

5.4.2. Comparación de posiciones obtenidas

En esta sección se presentan las posiciones obtenidas en segmentos de secuencias reales y una secuencia sintética, para una comparación visual de las posiciones que ofrece cada cámara y las que ofrecen los módulos de fusión: DSm, DSm-Punto y probabilista. En el eje vertical de las gráficas se representan las regiones que componen a Γ, mientras que en el horizontal se encuentra el tiempo. Las regiones están numeradas de acuerdo a como se muestra en la figura 5.1. Por cada uno de los renglones se grafica la masa que asigna la fuente (en caso de ser cámara) o la creencia resultante de la fusión por cada una de las regiones $\gamma_i \in \Gamma$. El nivel de certidumbre se muestra con la tonalidad de la barra, desde el rojo que representa una unidad completa de creencia hasta el azul, que representa cero creencia. Cuando no se detectan objetos en movimiento no se grafican barras. Junto con cada secuencia se muestran los resultados reales y es posible así comparar cuánto se asemeja el seguimiento obtenido por fusión a los datos reales y a lo que se obtendría usando las cámaras por separado.

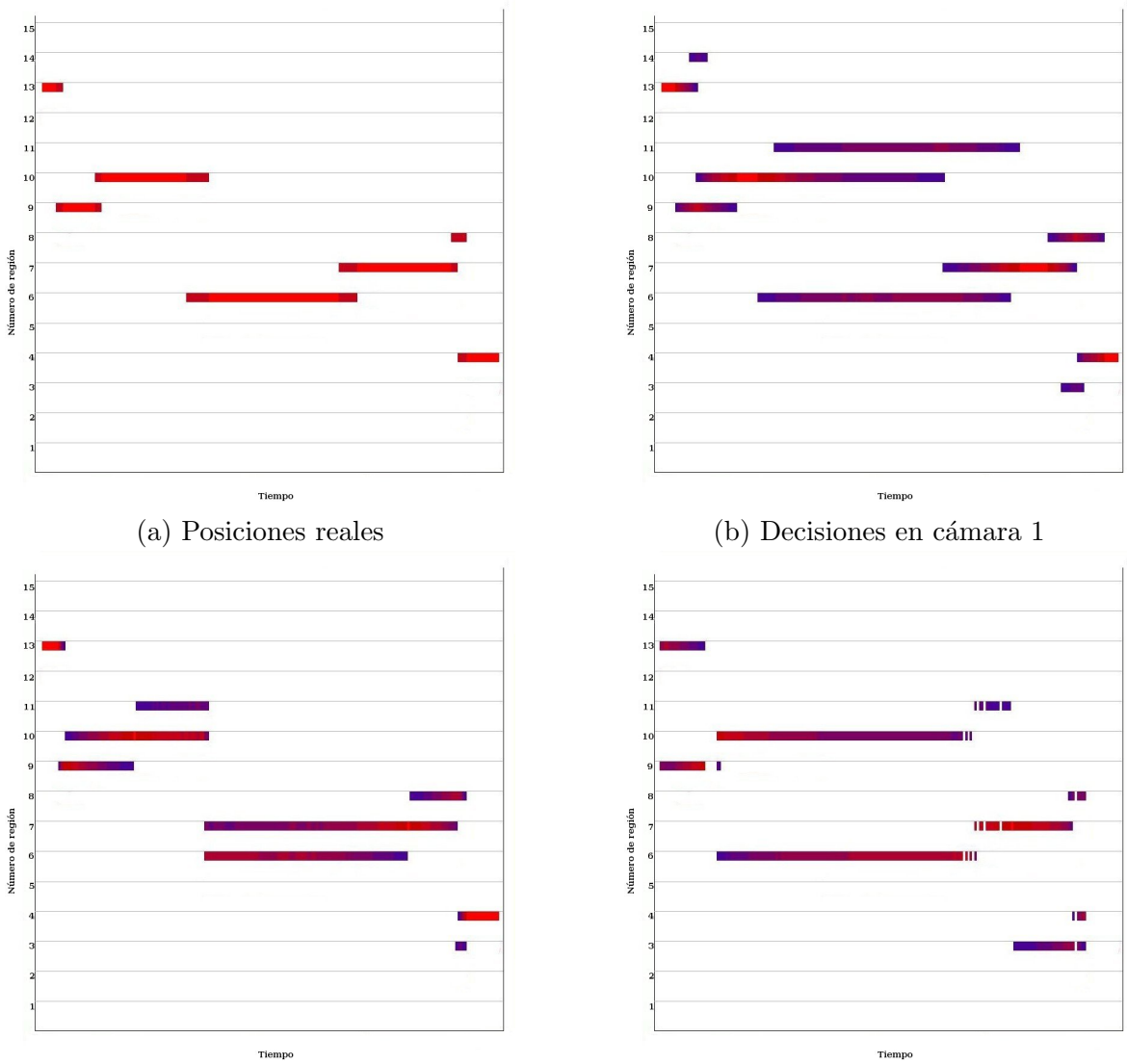
En las pruebas sintéticas se usaron 16 regiones en Γ, donde cada región tiene un tamaño casi semejante al de los objetos en movimiento lo cual no se habría podido encontrar en un caso práctico. En la figura 5.7 están las asignaciones de creencia hechas por las cámaras. Se puede ver que la cámara 3 pierde al objeto algunas veces, debido a que por su perspectiva, las líneas divisorias en el plano piso impiden hacer la detección. Algo muy marcado al tener una partición fina en Γ es que la proyección del eje vertical del objeto abarca más de una región, por lo que las cámaras tienden a asignar creencia a varias regiones y debido a que las perspectivas de las cámaras no son las mismas tampoco lo son las regiones a las que se asignan las creencias. En la figura 5.8 se muestran las asignaciones hechas

para la variante DSm-Punto, y es claro que al tener una proyección más corta del eje vertical, se intersecta con menos regiones y los valores de las creencias asignadas son mayores (se ven más rojos). Sin embargo, el hecho de disminuir la longitud del eje de proyección equivale a dejar de considerar la incertidumbre derivada de la perspectiva. En estos casos es común que una cámara proponga una región, mientras que otra cámara asigna toda su creencia a otra y el efecto es que el resultado de la fusión es asignado a la ignorancia local.

La figura 5.10 corresponde a una secuencia real en la que un objeto aparece en la escena sobre γ_2, al principio muy cercano a γ_1, y termina en γ_4. La incertidumbre derivada de las perspectivas de las cámaras se ven reflejadas claramente en la asignación de creencia. La cámara 1 genera en el inicio de la secuencia una creencia pequeña sobre γ_1 debido a que el objeto pasa muy cercano a tal región y según su perspectiva es creíble que el objeto pueda estar en esa región. La cámara 2 tiene una mejor perspectiva en el comienzo y asigna toda su creencia a γ_2. Al final de la secuencia el objeto es perdido por la cámara 2 porque el color del objeto es el mismo que el fondo. La fusión de las cámaras en esta secuencia muestra como la combinación de información toma las partes de ambas cámaras donde el objeto es visible, de esta formase mantiene información aún cuando la cámara 1 no observó bien al objeto en el inicio de la secuencia y cuando la cámara 2 pierde definitivamente al objeto . En la figura 5.11, se muestran las creencias asignadas a la misma secuencia, pero usando la fusión DSm-Punto. Los resultados son similares a los obtenidos con DSm, sin embargo, el hecho de considerar una menor incertidumbre impacta negativamente, en este caso, sobre el resultado de la fusión, principalmente porque la cámara 2 se declara ignorante (pierde el objeto) al final de las secuencia. Finalmente en la figura 5.12 se comparan los resultados de las fusiones, donde se puede ver como la fusión bayesiana asigna una pequeña probabilidad a los vecinos, debido a que aún cuando ambas cámaras coinciden en que el objeto se encuentra en la región γ_2 las probabilidades *a priori* indican una pequeña probabilidad para las regiones contiguas, por lo que la probabilidad

resultante de la fusión no es cero en las regiones vecinas.

La figura 5.13 muestra las posiciones obtenidas en otra secuencia real, donde el objeto es perdido varias veces debido a ruido en la escena. El objeto circula sobre la región γ_2 muy cercano a la región γ_4 hasta llegar a la región γ_3. En esta secuencia la cámara 1 pierde al objeto en movimiento y se comporta muy sensible al hecho de que el objeto esté cercano a la región γ_2, sin embargo el resultado de la fusión DSm es mucho mejor, ya que obtiene mucha evidencia de la cámara 2. Y sin importar que al final de la secuencia la cámara 2 pierde por completo al objeto, la fusión toma la información de la cámara 1. En síntesis, la fusión DSm mostró conservar lo mejor de ambas cámaras. En la figura 5.14 se muestran las posiciones obtenidas por fusión con DSm-Punto, y también se muestran las asignaciones de creencia hechas por las cámaras para esta secuencia. Como con DSm-Punto se reduce el eje vertical proyectado en comparación con DSm, los resultados de la cámara 1 muestran una reducción también, en la asignación de creencia a la región γ_4. La asignación de la cámara 1 es mejor, por consecuencia los resultados de la fusión son más parecidos a los datos reales al final de la fusión. En la figura 5.15, están los resultados de las fusiones, para propósitos de comparación.



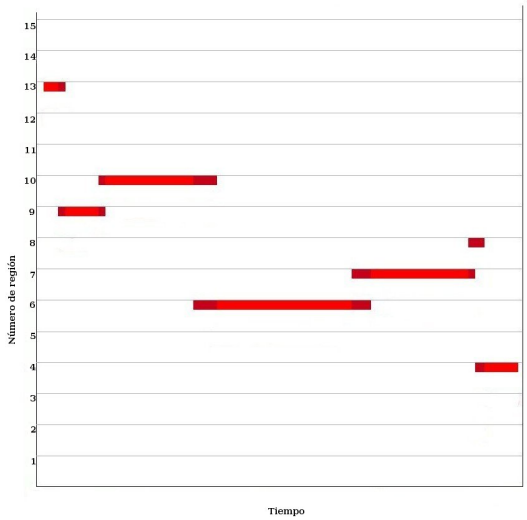
(a) Posiciones reales

(b) Decisiones en cámara 1

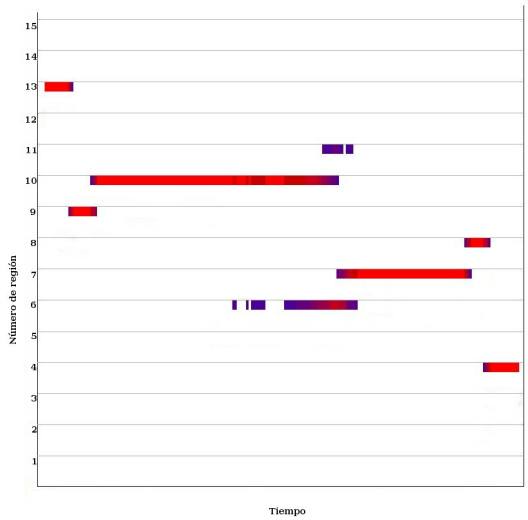
(c) Decisiones en cámara 2

(d) Decisiones en cámara 3

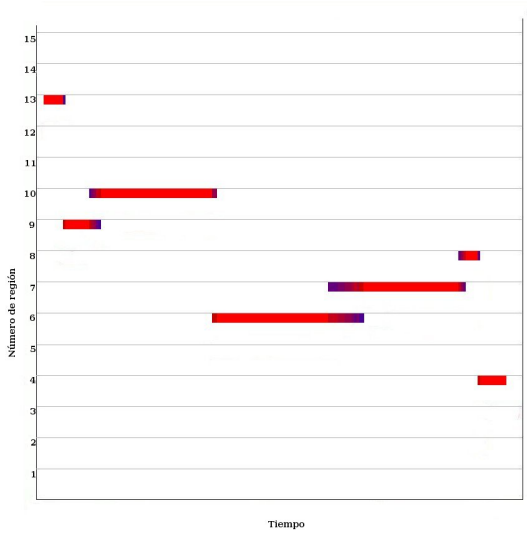
Figura 5.7: Ejemplo de posiciones obtenidas en subsecuencia sintética. Comparación de datos reales contra los obtenidos en las cámaras



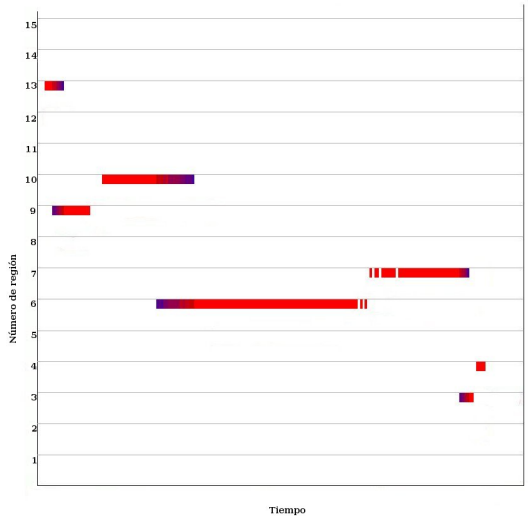
(a) Posiciones reales



(b) Decisiones en cámara 1 (DSm-Punto)

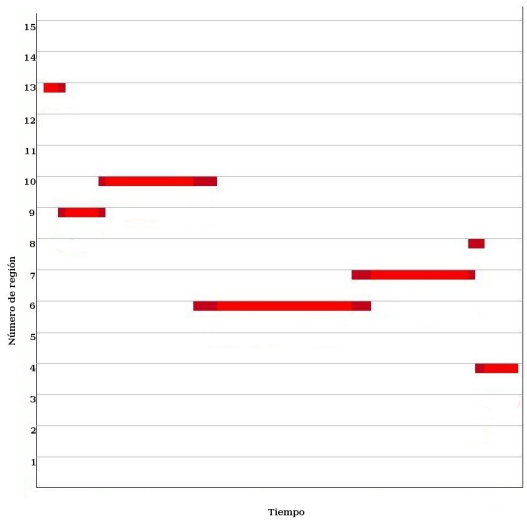


(c) Decisiones en cámara 2 (DSm-Punto)

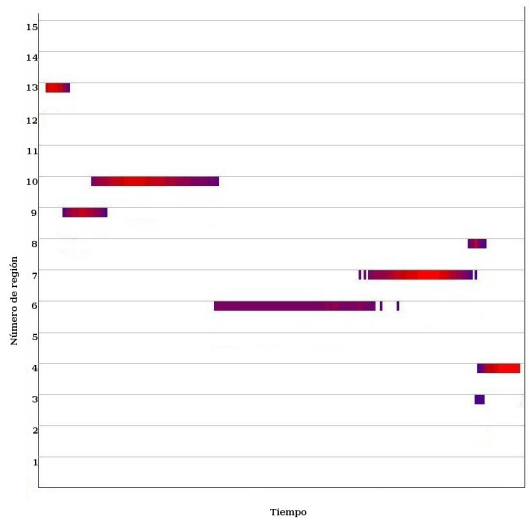


(d) Decisiones en cámara 3 (DSm-Punto)

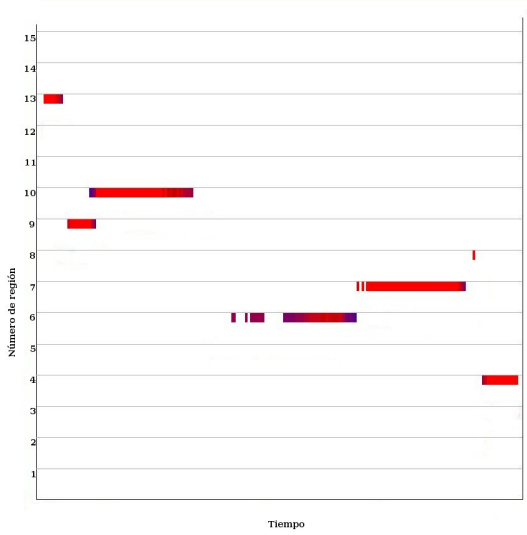
Figura 5.8: Ejemplo de posiciones obtenidas en subsecuencia sintética. Comparación de datos reales contra los obtenidos en las cámaras usando asignación para fusión DSm-Punto



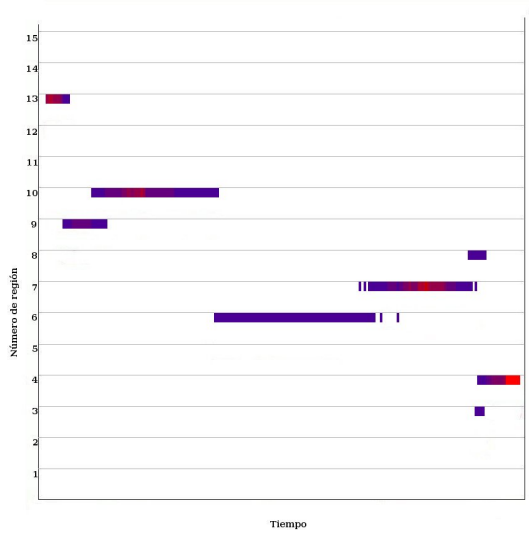
(a) Posiciones reales



(b) Decisiones por fusión DSm

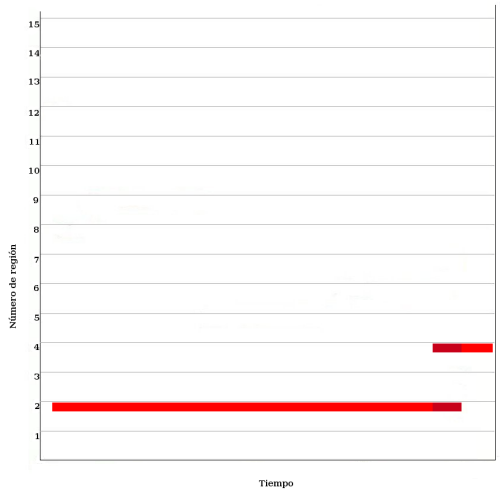


(c) Decisiones por fusión DSm - Punto

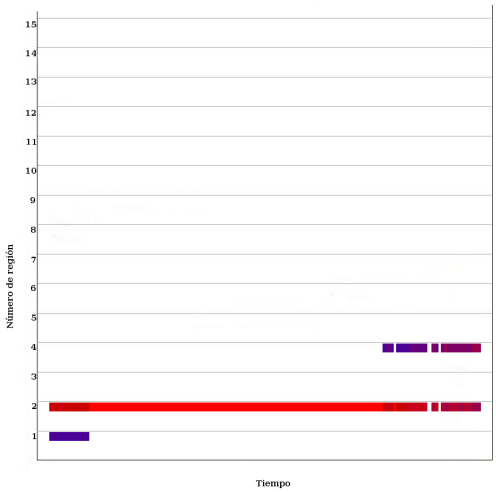


(d) Decisiones por fusión bayesiana

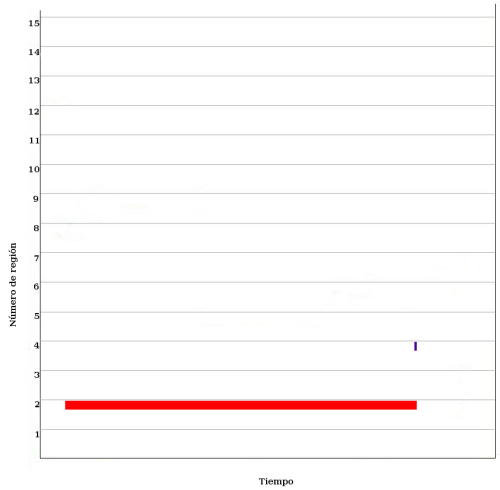
Figura 5.9: Ejemplo de posiciones obtenidas en subsecuencia sintética. Comparación de datos reales con los obtenidos por fusión DSm, DSm-Punto y bayesiana



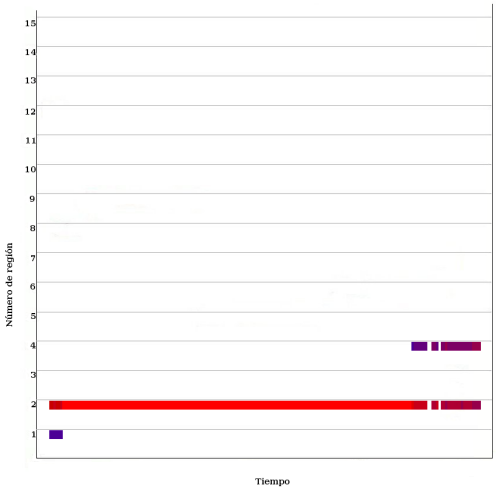
(a) Posiciones reales



(b) Decisiones en cámara 1

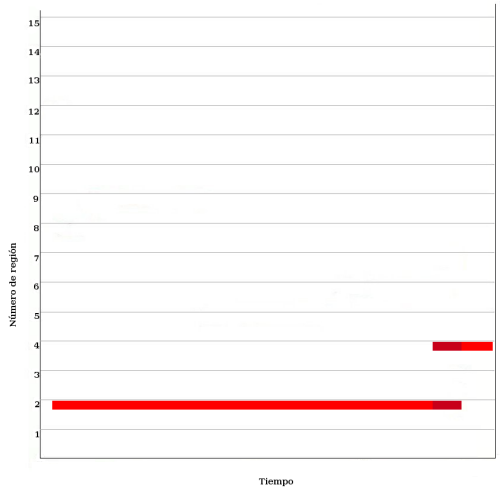


(c) Decisiones en cámara 2

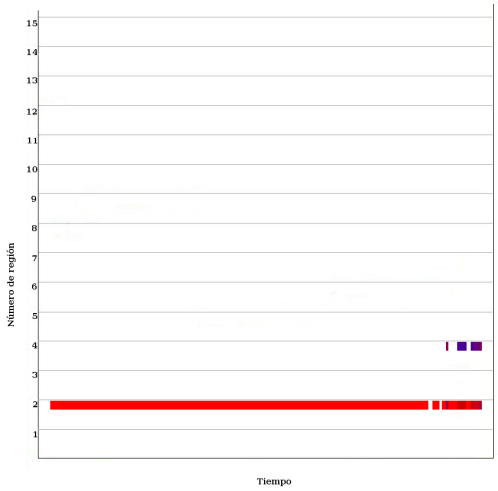


(d) Decisiones por fusión DSm

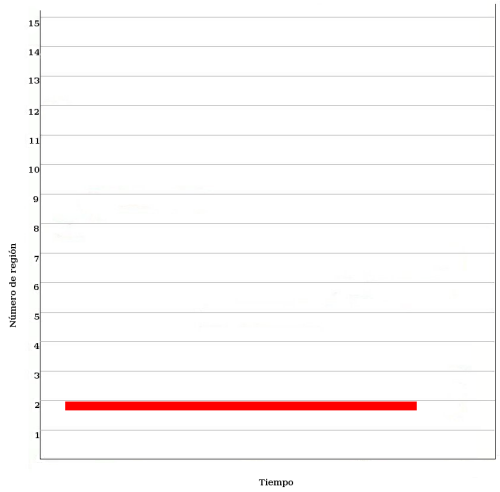
Figura 5.10: Ejemplo de posiciones obtenidas en la subsecuencia real R1. Comparación de datos reales con los obtenidos en las cámaras y por fusión DSm. Tanto en (b) como en (d) al final de la secuencia la región con mayor creencia es la 2. En este caso la fusión arroja un falso positivo derivado de la cámara 1, la cámara 2 no aporta creencia.



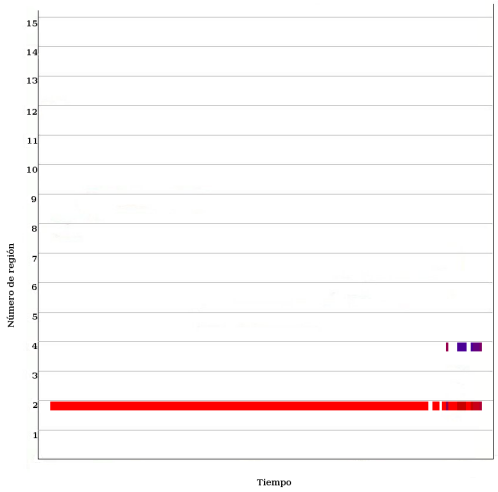
(a) Posiciones reales



(b) Decisiones en cámara 1 (DSm-Punto)

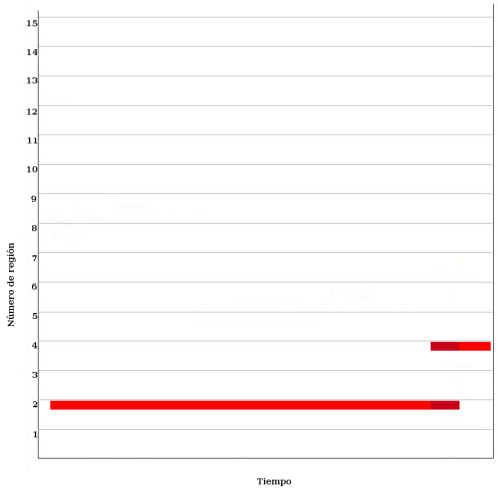


(c) Decisiones en cámara 2(DSm-Punto)

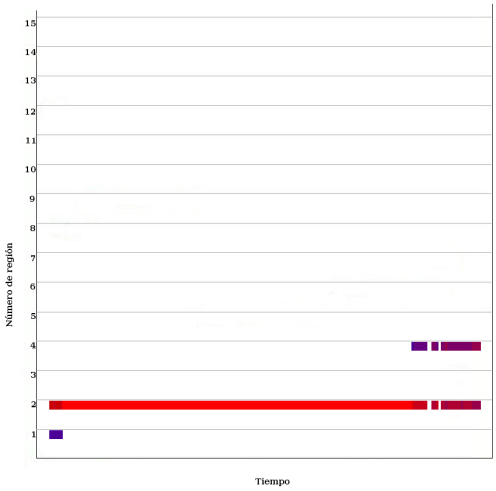


(d) Decisiones por fusión DSm-Punto

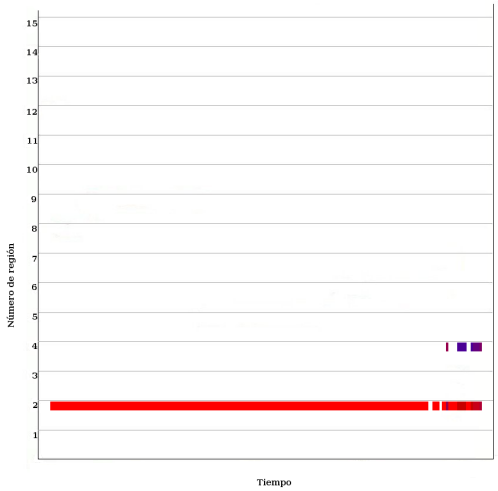
Figura 5.11: Ejemplo de posiciones obtenidas en la subsecuencia real R1. Comparación de datos reales con los obtenidos en las cámaras y por fusión DSm-Punto



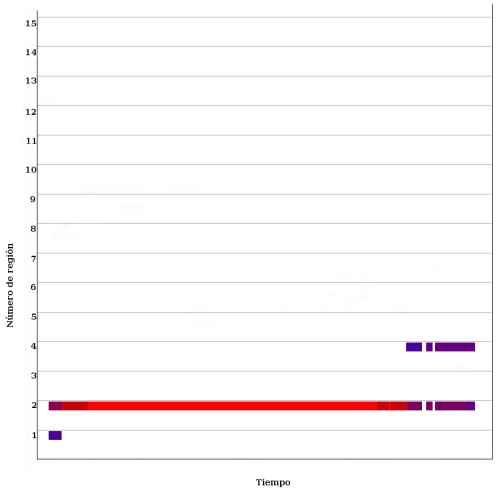
(a) Posiciones reales



(b) Decisiones por fusión DSm

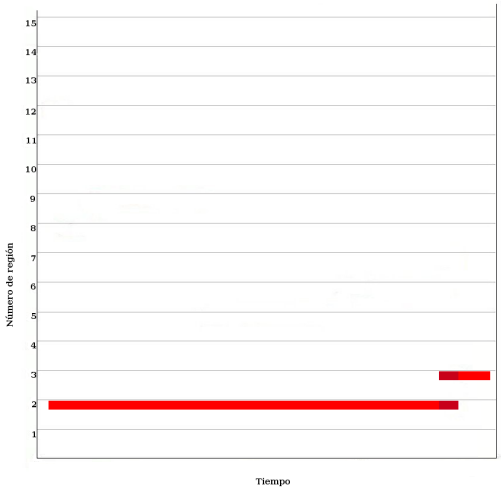


(c) Decisiones por fusión DSm - Punto

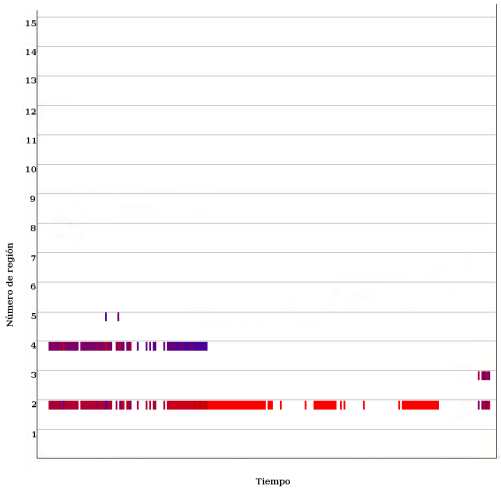


(d) Decisiones por fusión bayesiana

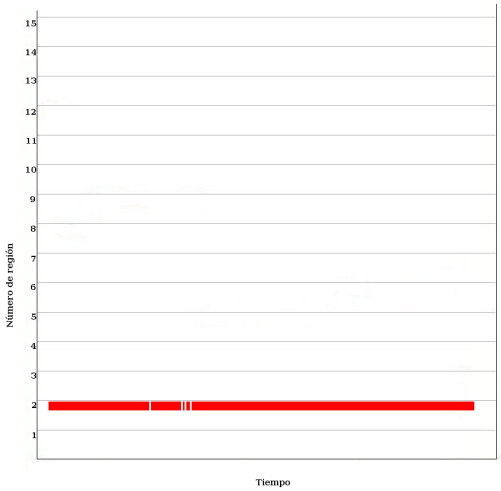
Figura 5.12: Ejemplo de posiciones obtenidas en la subsecuencia real R1. Comparación de datos reales con los obtenidos por fusión DSm, DSm-Punto y bayesiana



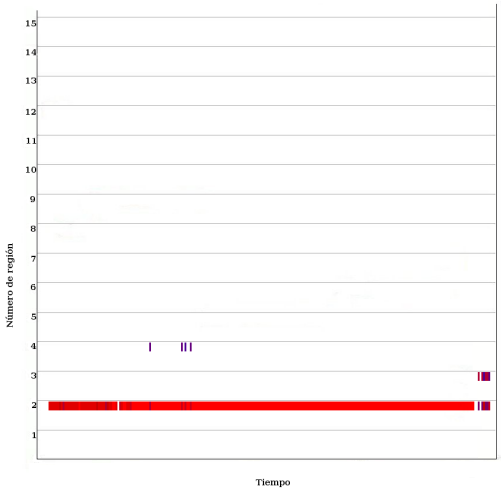
(a) Posiciones reales



(b) Decisiones en cámara 1

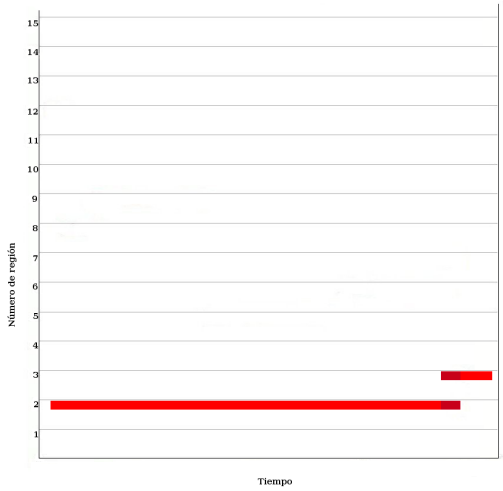


(c) Decisiones en cámara 2

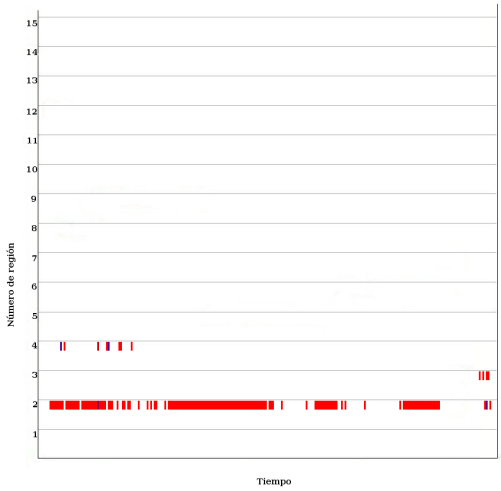


(d) Decisiones por fusión DSm

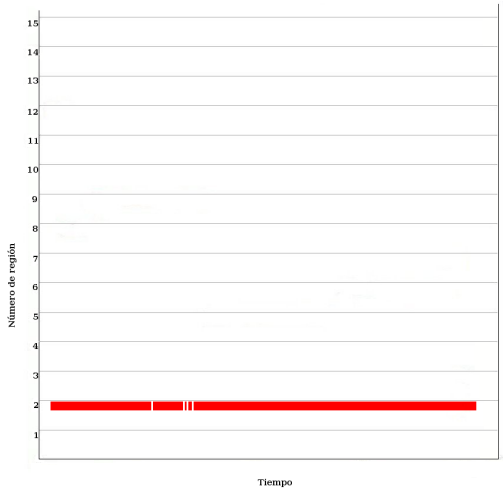
Figura 5.13: Ejemplo de posiciones obtenidas en la subsecuencia real R2. Comparación de datos reales con los obtenidos en las cámaras y por fusión DSm



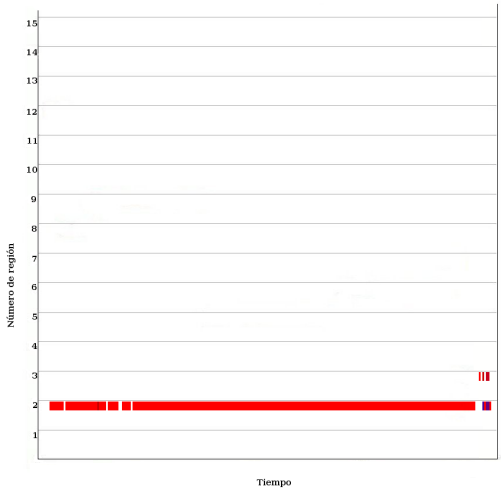
(a) Posiciones reales



(b) Decisiones en cámara 1(DSm-Punto)

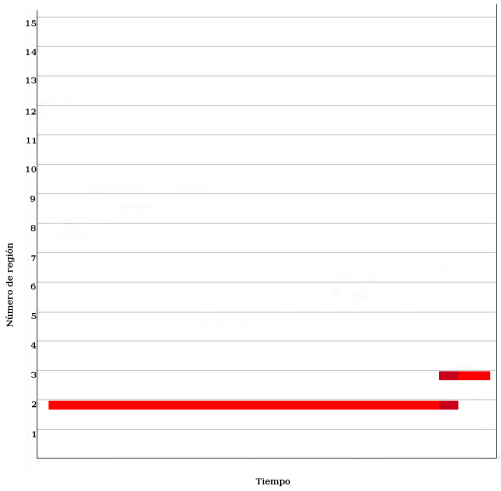


(c) Decisiones en cámara 2(DSm-Punto)

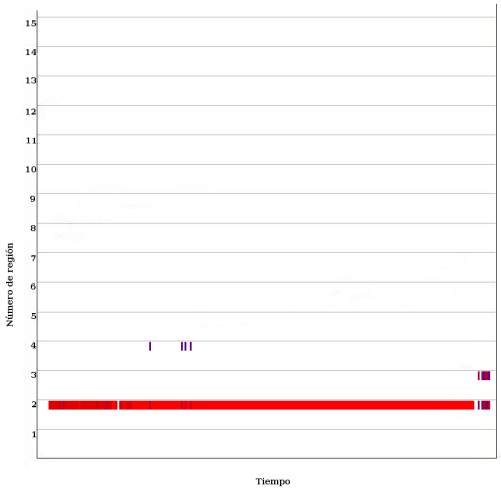


(d) Decisiones por fusión DSm-Punto

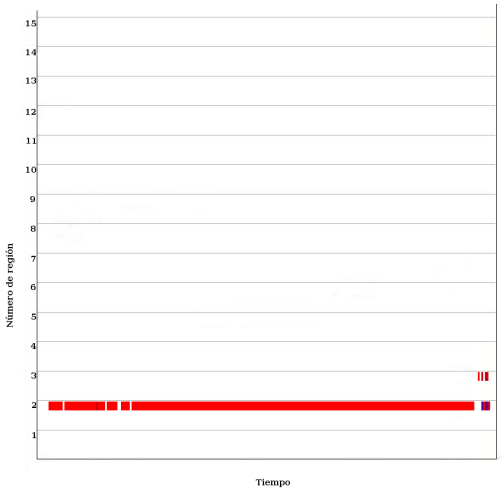
Figura 5.14: Ejemplo de posiciones obtenidas en la subsecuencia real R2. Comparación de datos reales con los obtenidos en las cámaras y por fusión DSm-Punto



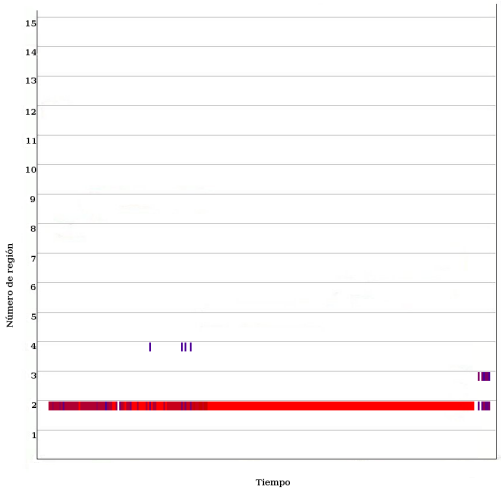
(a) Posiciones reales



(b) Decisiones por fusión DSm



(c) Decisiones por fusión DSm - Punto



(d) Decisiones por fusión bayesiana

Figura 5.15: Ejemplo de posiciones obtenidas en la subsecuencia real R2. Comparación de datos reales con los obtenidos por fusión DSm, DSm-Punto y bayesiana

5.5. Resumen

En este capítulo se presentaron las pruebas realizadas al modelo propuesto. Se probó la alineación de datos con una secuencia de imágenes reales. En las pruebas de alineación espacial se obtuvo un promedio de 7.92 pixeles de diferencia entre las trayectorias obtenidas con las cámaras en las secuencias reales.

El modelo de fusión y la arquitectura propuestos fueron probados usando el módulo de fusión DSm y un módulo de fusión probabilista adaptado. En ambos casos se usaron secuencias de imágenes sintéticas generadas por computadora y secuencias reales del repositorio multicámara PETS.

Para las pruebas de fusión de información se adaptó un módulo de fusión probabilista, considerando un clasificador bayesiano por cada una de las regiones representadas en Θ. También se agregó una modificación a la función de asignación de creencia básica generalizada, reduciendo a la mitad el eje proyectado en el plano piso por cada objeto, al cual se le llamó DSm-Punto.

Los resultados obtenidos muestran que este modelo mejora la información comparada con la que se obtiene usando las cámaras individualmente. La taza de falsas alarmas es reducida, mientras que la taza de seguimiento del objeto se mantiene en la media de las cámaras en el caso de las secuencias sintéticas y en el caso de las secuencias reales mejora. Las pruebas también demostraron que es posible utilizar un módulo de fusión probabilista con la arquitectura propuesta, considerando la creencia asignada con la gbba como probabilidad condicional y teniendo una misma probabilidad *a priori* asignada a cada región.

Si se comparan visualmente las trayectorias obtenidas por cada una de las cámaras con las obtenidas como resultado de la fusión, también es posible ver como la combinación de información obtiene una trayectoria más refinada.

Capítulo 6

Conclusiones y trabajo futuro

El creciente poder de cómputo en los dispositivos electrónicos, el auge en modelos de fusión, y el uso de cámaras inteligentes, hace posible la combinación de información para tareas de vigilancia, ya no a niveles en los que los sistemas se limitan a mostrar las posiciones de un objeto con respecto a las coordenadas de la pantalla, sino que se pueden ya expresar relaciones entre los objetos en movimiento y entidades en escena.

En este trabajo se presentó un modelo nuevo de fusión multicámara a nivel decisión para seguimiento de objetos, usando cámaras geográficamente distribuidas con campos de vista traslapados, tomando la Teoría Evidencial DSm como base para la combinación de información.

Para poder cumplir con el requerimiento de manejo de información a alto nivel, se definieron entidades llamadas *regiones* que se usan en el modelo para describir la posición de los objetos en movimiento en una escena. Tales entidades tienen un significado práctico al representar zonas de interés para el sistema de vigilancia (pasillos, zonas de estacionamiento, áreas verdes, etc.), todo ello sobre el plano en el que se mueven los objetos, denominado *plano piso*, que sirve como sistema de referencia espacial común a todas las cámaras y ayuda al proceso de alineación espacial y calibración.

En este modelo las cámaras toman en consideración su posición en el escenario con respecto al plano en el que los objetos circulan, y son consideradas para el módulo de fusión como expertos. El modelo propuesto fue probado tanto con secuencias de video sintéticas como reales, teniendo resultados que demuestran que el modelo de fusión a nivel decisión propuesto puede reducir la incertidumbre derivada de la perspectiva de las cámaras considerablemente.

La teoría evidencial DSm es muy similar a la teoría DS, sin embargo, en los casos en los que la DS derivaría en resultados contraintuitivos, la DSm no lo hace. Esto se debe a que la DSm utiliza nuevas reglas de combinación, y se ha demostrado teóricamente en los trabajos de Dezert y Smarandache (9). Por otra parte, comparada con la fusión basada en clasificadores bayesianos, los resultados son muy similares, sin embargo, existen casos en los que el uso de ignorancia permite que los sensores que tienen menor creencia asignada a ignorancia tengan mayor peso sobre el resultado, como se comentó en la sección 5.4.1. Como parte de las pruebas se desarrolló un módulo de fusión basado en un clasificador bayesiano aplicado por cada una de las regiones, que puede funcionar con la misma arquitectura que se usó para el modelo de Teoría Evidencial.

Esta contribución será de utilidad para trabajos relacionados con la identificación de comportamientos en sistemas de vigilancia y en general para investigación en el área de visión por computadora donde se requiera de un manejo e interpretación de datos a alto nivel. Aún cuando la ejecución en tiempo real no fue un objetivo en este trabajo, se consiguieron resultados que demuestran que podría ser posible alcanzar tiempos de ejecución reducidos si se aplicaran técnicas de procesamiento paralelo o cámaras inteligentes (sensores de video que realizan procesamiento con hardware dedicado). En las pruebas realizadas el comportamiento del módulo de fusión DSm tuvo resultados muy semejantes a los que se obtuvieron usando un clasificador bayesiano por cada región. Aún cuando el tiempo del módulo de fusión probabilista fue menor, tanto el módulo DSm, como el probabilista, mostraron que son aplicables en procesamiento de video en tiempo real.

En este trabajo se propuso una forma de interpretación de la información geométrica de la escena en problemas de sistemas de vigilancia. Con esta teoría cada cámara genera creencia para representar incertidumbre derivada de su perspectiva. Las cámaras, al comportarse como expertos, generan decisiones que pueden ser fusionadas por módulos de procesamiento de alto nivel. En este trabajo se adaptaron y aplicaron fusión de información con la Teoría Evidencial DSm y con múltiples clasificadores bayesianos.

6.1. Trabajo futuro

En el presente trabajo se asume que los objetos en movimiento se desplazan sobre el mismo plano (plano piso), lo cual, aunque es común en la mayoría de los casos para sistemas de vigilancia, es también una restricción, ya que algunos casos particulares comprenden la circulación de objetos en movimiento sobre planos compuestos, como es el caso de escenarios que contienen pasos a desnivel o escaleras. Estos escenarios donde los objetos circulan en varios planos podrían ser estudiados tomando diferentes criterios geométricos para la alineación espacial y correspondencia.

La incorporación de otros parámetros a la función de asignación de creencia básica generalizada (gbba) es algo que se podría mejorar los resultados. Además de considerar la incertidumbre derivada de la perspectiva, es posible considerar otras variables en la función de asignación de creencia básica generalizada, tales como características de los objetos (tamaño, forma, color, etc.) o propiedades de las cámaras (características de los lentes, resolución, zoom). Estas nuevas variables podrían servir como parámetros para la asignación de creencias y afinar el resultado de la fusión.

En este trabajo se asume que todas las cámaras poseen las mismas características, pero es posible aplicar cámaras en distintos rangos o incluso la fusión de otros tipos de sensores (sonares, radares, etc.), a un nivel de procesamiento alto.

Muchos de los algoritmos de visión por computadora están siendo llevados a dispositivos electrónicos especializados. En esta área se podrían estudiar técnicas de diseño de hardware o cómputo reconfigurable para la optimización de la asignación de creencia y fusión, con el objetivo de reducir tiempo de procesamiento.

Referencias

[1] Ballard, D. y Brown, C. (1982). *Computer Vision*. Prentice Hall.

[2] Bashir, F. y Porikli, F. (2006). *Performance Evaluation of Object Detection and Tracking Systems*. En *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

[3] Beynon, M. D., Hook, D. J. V., Seibert, M., Peacock, A., y Dudgeon, D. (2003). *Detecting Abandoned Packages in a Multi-Camera Video Surveillance System*. vol. 00, pag. 221, Los Alamitos, CA, USA. IEEE Computer Society.

[4] Black, J. y Ellis, T. (2006). *Multi camera image tracking. Image Vision Comput.*, 24(11):1256–1267.

[5] Bradshaw, K. J., Reid, I. D., y Murray, D. W. (1997). *The Active Recovery of 3D Motion Trajectories and Their Use in Prediction. IEEE Trans. Pattern Anal. Mach. Intell.*, 19(3):219–234.

[6] Brooks, R. R. y Iyengar, S. S. (1997). *Real-time distributed sensor fusion for time-critical sensor readings. Optical Engineering*, 36(3):767–779.

[7] Challa, S. y Koks, D. (2004). *Bayesian and Dempster-Shafer fusion*. En *Sadhana*, vol. 29, pag. 145–174.

[8] Dasarathy, B. V. (1997). *Sensor fusion potential exploitation-innovative architectures and illustrative applications*. En *Proceedings of the IEEE*, vol. 85, pag. 24–38.

[9] Dezert, J. (2002). *Foundations for a new theory of plausible and paradoxical reasoning. Information and Security Journal*, 9.

[10] Fioretti, G. (2002). *Evidence Theory: A Mathematical Framework for Unpredictable Hypotheses.* Experimental 0207001, EconWPA. available at http://ideas.repec.org/p/wpa/wuwpex/0207001.html.

[11] Green, R. y Guan, L. (2003). *Quantifying and Recognizing Human Movement Patterns from Monocular Video Images - Part I: A New Framework for Modeling Human Motion.*

[12] Hall, D. L. y Llinas, J., editors (2001). *Handbook of multisensor data fusion.* The Electrical Engineering and Applied Signal Processing Series. CRC Press.

[13] Hartley, R. I. y Zisserman, A. (2004). *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521540518, second edition.

[14] Heikkila, J. y Silven, O. (1999). *A Real-Time System for Monitoring of Cyclists and Pedestrians.* En *VS '99: Proceedings of the Second IEEE Workshop on Visual Surveillance*, pag. 74, Washington, DC, USA. IEEE Computer Society.

[Hoffman and Murphy] Hoffman, J. C. y Murphy, R. R. *Comparison of Bayesian and Dempster-Shafer Theory for Sensing: A Practitioner's Approach.*

[http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001 dataset.html] http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001 dataset.html. *Performance Evaluation of Tracking and Surveillance DATASET.* Referencia de Internet.

[15] Hu, W., Hu, M., Zhou, X., y Lou, J. (2006). *Principal Axis-Based Correspondence between Multiple Cameras for People Tracking. IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):663. Fellow-Tieniu Tan and Member-Steve Maybank.

[16] Hyder, A. K., Shahbazian, E., y Waltz, E. (2002). *Multisensor Fusion*. Springer.

[17] Jsang, A., Pope, S., Diaz, J., y Bouchon-Meunier, B. (2005). *Dempster's Rule as Seen by Little Coloured Balls*.

[18] Kettnaker, V. y Zabih, R. (1999). *Bayesian Multi-Camera Surveillance. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 02:pages 2253.

[19] Khan, S. y Shah, M. (2003). *Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25(10):pages 1355–1360.

[20] Klein, L. A. (1999). *Sensor and Data Fusion Concepts and Applications*, vol. TT35 de *Tutorial Texts in Optical Engineering*. SPIE.

[21] Koks, D. y Challa, S. (2003). *An Introduction to Bayesian and Dempster-Schafer Data Fusion*. Scientific and Technical Report DSTO-TR-1436, Australian Government. Department of Defense. Defence Science and Technology Organisation.

[22] Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., y Shafer, S. (2000). *Multi-camera multiperson tracking for EasyLiving*. En *Proceedings of the Third IEEE International Workshop on Visual Surveillance*, pag. 3–10.

[23] Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., y White, F. (2004). *Revisiting the JDL Data Fusion Model II*.

[24] Lv, F., Kang, J., Nevatia, R., Cohen, I., y Medioni, G. (2004). *Automatic Tracking and Labeling of Human Activities in a Video Sequence*. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance In conjunction with ECCV'04.

[25] Malik, J., Bickel, P., Rice, J., y Russell, S. (2003). *Multi-sensor traffic data fusion*. Technical report, Institute of Transportation S tudies University of California, Berkeley.

[26] Nguyen, N. T., Phung, D. Q., Venkatesh, S., y Bui, H. (2005). *Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models*. En *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pag. 955–960, Washington, DC, USA. IEEE Computer Society.

[27] Sentz, K. y Ferson, S. (2002). *Combination of Evidence in Dempster-Shafer Theory*. Technical Report SAND 2002-0835, Binghamton University, Binghamton, NY 13902-6000.

[28] Shafer, G. (2002). *Dempster-Shafer theory*. http://www.glennshafer.com/assets/downloads/articles/article48.pdf.

[29] Smarandache, F. (2004). *Advances and Applications of DSmT for Information Fusion*. American Research Press.

[30] Smarandache, F. y Dezert, J. (2006). *An Introduction to the DSm Theory for the Combination of Paradoxical, Uncertain, and Imprecise Sources of Information*.

[31] Stein, G., Lee, L., y Romano, R. (2000). *Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame. IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):758–767.

[32] Steinberg, A.Ñ., Bowman, C. L., y White, F. E. (1999). *Revisions to the JDL data fusion model*. En *SPIE AeroSense*, pag. 430–441.

[33] Yan, W. y Forsyth, D. A. (2005). *Learning the Behavior of Users in a Public Space through Video Tracking*. En *WACV-MOTION '05: Proceedings of the Seventh IEEE Workshops on Application of Computer Vision*

(WACV/MOTION'05) - Volume 1, pag. 370–377, Washington, DC, USA. IEEE Computer Society.