Detección automática y seguimiento de personas utilizando fusión de secuencias de imágenes multimodales.

por

Metzli Ramírez Martínez

Tesis
sometida como requisito parcial
para obtener el grado de

MAESTRA EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Noviembre 2013
Tonantzintla, Puebla

Supervisada por:

Dr. Leopoldo Altamirano Robles

© INAOE 2013
Derechos reservados
El autor otorga al INAOE el permiso de reproducir y distribuir copias de esta tesis en su totalidad o en partes

Resumen

Esta tesis presenta el desarrollo de un método de detección automática y seguimiento de múltiples personas para secuencias de imágenes capturadas a una tasa de 4 fps, en ambientes exteriores con iluminación no controlada, con presencia de personas uniformadas y que presentan oclusión entre sí. Este conjunto de características fue elegido debido a que actualmente se ha notado que la gran variedad de escenarios y circunstancias que pueden ocurrir en diferentes entornos presentan un gran reto en cuanto a crear métodos globales de seguimiento de personas, es decir que funcionen bajo todo tipo de circunstancias, por lo cual actualmente se está optando por generar métodos de seguimiento especializados para cada ambiente, logrando mejores resultados que si se plantea hacer un método global. Es por eso que los métodos de seguimiento en ambientes institucionales, como lo son hospitales, escuelas, laboratorios, asociaciones deportivas, campos militares, entre otros, deben ser aptos para trabajar con secuencias de imágenes con las características antes descritas. Buscando solucionar de esta forma una necesidad específica.

El método desarrollado está basado en la fusión de las secuencias de imágenes infrarrojas y visuales, lo cual implica un pre-procesamiento que incluye la calibración de ambas modalidades de cámaras y el registro de las imágenes, dicho pre-procesamiento también es descrito a lo largo de la tesis. Para la detección automática se implementó el método HOG [1] de forma independiente para cada tipo de imagen y posteriormente se fusionan ambos resultados obteniendo un resultado global. Finalmente para la etapa de seguimiento se diseñó un modelo de características a partir del cual se realiza

una comparación entre las personas en la escena, para posteriormente identificarlas y determinar su posición y trayectoria seguida a lo largo de la secuencia.

El método desarrollado obtuvo una tasa de 74.5% de verdaderos positivos para secuencias de imágenes con un alto grado de dificultad, descritas anteriormente. Se comparó con el método [2] superándolo en promedio por un 9.2%. Cabe mencionar que el método [2] probado anteriormente con secuencias de imágenes capturadas a 30 fps, en ambientes interiores, con cambios de iluminación y cambios en la posición de los objetos, mostro un tasa de perdida de tan solo 1%, lo cual muestra la diferencia de complejidad entre las secuencias de prueba.

Abstract

This thesis present the development of a method for automatic detection and tracking of multiple people for images sequences captured at 4 fps, in exterior environments, with no controlled illumination, with presence of uniformed people and with occlusion. This group of characteristics was selected because recently it has been noted that the big variety of scenes and circumstances that occur in different environments are a big challenge when we speak about to create global methods of people tracking (i.e. methods that work for all kind of circumstances). Consequently, at the present it is opting by create methods of tracking specialized for each environment; this achieves better results than with a global method. For that reason, the tracking methods for institutional environments, like hospitals, schools, laboratories, sport associations, military camps, etc., have to be suitable for working with the image sequences with the characteristics described above, looking for satisfying a specific necessity.

The developed method is based in the sequences fusion of infrared and visual images, this involve a pre-processing that include the calibration of both kind of cameras and the images registration, this pre-processing also is described along the thesis. For the automatic detection, HOG [1] method was implemented for each kind of image independently, and later both results were fused in a global result. Finally, for the tracking stage a model of characteristics was designed and with this model a comparison between the people in the scene was made, to after that identify and determine the people positions and their trajectories along the image sequence.

The developed method has a true positives rate of 74.5% for image sequences with a high grade of difficulty, described above. Our method was compared with the method [2] and we obtained better results by 9.2%. The method [2] was tested previously with images sequences captured at 30 fps, in interior environments, with light changes and changes in the objects position, and it showed a miss rate of only 1%. This show the unlike complexity between the test sequences.

Índice

Resumen II
Abstract IV
Índice VI
Lista de figuras IX
Lista de tablas XI
Capítulo 1. Introducción 1
1.1 Descripción del problema 1
1.1.1 Motivación 2
1.2 Objetivos 3
1.2.1 Objetivo general 3
1.2.2 Objetivos específicos 4
1.3 Solución propuesta 4
1.4 Alcance de la tesis 5
1.5 Organización de la tesis 6
Capítulo 2. Fundamentos teóricos 7
2. 1 Detección y seguimiento de objetos/personas 7
2. 2 Calibración de las cámaras 9
2.3 Transformaciones Geométricas 12
2.4 Registro de imágenes 14
2.5 Fusión de imágenes 17

2.6 Niveles de fusión 18
2.7 Histograma del Gradiente Orientado 19
2.8 Filtro de Kalman 22
2.9 Descriptor SURF 25
2.10 Resumen 29
Capítulo 3. Trabajos relacionados 31
3.1 Seguidores que no utilizan fusión de sensores 31
3.1.1 Aprendizaje progresivo para seguimiento visual robusto 31
3.1.2 Seguimiento utilizando métricas del tipo cross-bin 33
3.2 Seguidores que utilizan fusión de sensores 35
3.2.1 Modelo gráfico probabilístico utilizado para seguimiento de personas 35
3.2.2 Seguimiento de personas usando fusión de imágenes térmicas, visuales y mapas de profundidad. 36
3.2.3 Detector de objetos basado en fusión de múltiples sensores y seguimiento usando un modelo de forma activa. 39
3.2.4 Marco integrado de registro de imágenes, fusión de sensores y seguimiento de personas utilizando imágenes térmicas y visuales 41
3.3 Resumen 45
Capítulo 4. Método propuesto 47
4.1 Calibración de las cámaras y captura de imágenes 47
4.1.1 Calibración cámara térmica 48
4.2 Registro de las imágenes 49
4.3 Detección de personas 50
4.4 Caracterización de personas 54
4.5 Seguimiento 60
4.5.1 Comparación de características 60
4.5.2 Clasificación de personas 64

4.6 Resumen 66
Capítulo 5. Experimentos y resultados.....68
5.1 Experimentación.....68
5.1.1 Metodología de la experimentación 69
5.1.2 Evaluación.....70
5.1.2 Experimentos realizados 71
5.2 Resultados 72
5.3 Discusión 74
5.3 Resumen 75
Capítulo 6. Conclusiones y trabajo a futuro.....77
6.1 Conclusiones 77
6.2 Trabajo a futuro 79
Apéndices.....81
Apéndice A - Acrónimos.....82
Bibliografía 84

Lista de figuras

1. Método propuesto. 4

2. Ejes del mundo real con respecto a la cámara.....5

3. Geometría de las imágenes 8

4. Transformada proyectiva 13

5. Tasa de pérdida del método HOG según el cambio de tamaño de las células y los bloques 22

6. Filtros gaussianos de segundo orden de sus derivadas parcial 27

7. Filtros de Haar utilizados en el descriptor SURF. 28

8. Resultados del seguidor propuesto por Krotosky y Trivedi.....38

9. Diagrama del algoritmo de seguimiento de 40

10. Diagrama del algoritmo de [24].....42

11. Patrón de calibración de la cámara visual. 48

12. Patrón de calibración en la cámara térmica.....49

13. Grafica de la relación entre el tamaño del descriptor, la tasa de pérdida y el tamaño de los bloques y las células. 52

14. Representación del desplazamiento horizontal y vertical de las ventanas de detección. 53

15. Ilustración de las características usadas para segmentar la cabeza de una persona. 54

16. Algoritmo general del proceso de seguimiento.....61

17. Ejemplo de la matriz de similitud suponiendo once características y tres clases distintas.....63

18. Ejemplo de la extracción de los vectores de valores máximos, clases, el vector de las clases más repetidas y el vector del número de repeticiones de dichas clases.....64

19. Diagrama de flujo que representa el algoritmo de clasificación. 65

Lista de tablas

Tabla 1. Transformaciones afines básicas.....15
Tabla 2. Características de las secuencias de prueba 71
Tabla 3. Resultados del experimento con la secuencia número1.....72
Tabla 4. Resultados del experimento con la secuencia número.....72
Tabla 5. Resultados del experimento con la secuencia número.....72
Tabla 6. Resultados del experimento con la secuencia número.....73
Tabla 7. Resultados del experimento con la secuencia número.....73
Tabla 8. Promedio de los resultados de todos los experimentos.....73
Tabla 9. Desviación estándar de los resultados de todos los experimentos.....74

Capítulo 1

Introducción

En este capítulo se introducirá al tema desarrollado en esta tesis, se desea explicar cuáles fueron las motivaciones principales que inspiraron la solución que se propone, así como cuáles son los objetivos de la misma y sus limitaciones. El capítulo está formado por las siguientes secciones: descripción del problema, objetivos, solución propuesta, alcance de la tesis, organización de la tesis y discusión.

1.1 Descripción del problema

Los detectores y seguidores de personas tienen una fuerte aplicación, principalmente en las áreas de vigilancia y robótica. En robótica, los seguidores de personas se utilizan principalmente para la interacción del robot con su entorno y en específico con personas, mientras que en el área de vigilancia son útiles por la necesidad que ha surgido desde tiempos inmemorables de cuidar y resguardar la seguridad de las personas y/o los objetos de mayor valor.

Algunos ejemplos de las aplicaciones de este tipo de sistemas son los empleados en aeropuertos, estaciones de autobús o estaciones de tren, en donde si se detecta que algún individuo deja un objeto abandonado dentro de la estación este es seguido como posible sospechoso. Estos sistemas se emplean con el objetivo de prevenir algún ataque terrorista que utilice explosivos, como ha sucedido ya anteriormente. Otro ejemplo de aplicación es resguardar o monitorear la entrada y salida de individuos de laboratorios y/o áreas restringidas solo para personas autorizadas, esto se llega a emplear principalmente en empresas ya sean públicas o privadas.

Asimismo, el uso de fusión de sensores para el seguimiento de personas permite obtener una mayor cantidad de información que de otro modo no se podrían conseguir. En específico el uso de fusión de imágenes infrarrojas e imágenes visuales es muy utilizado, las razones para ellos son varias, por una parte las imágenes visuales proporcionan una gran cantidad de información en escenarios con buena iluminación, sobre detalles de los rostros y rasgos físicos de las personas, además de información de color de la vestimenta, que es muy fácil de diferencia en la mayoría de los casos y sobre todo son el tipo de cámara más común y por lo tanto más económicas del mercado[3].

Por su parte las cámaras IR también son comunes y económicas comparadas con cámaras en otros espectros, que son mucho más difíciles de conseguir. Además de que no son afectadas por la cantidad de iluminación de la escena y proporcionan información de los patrones térmicos de las personas. En general la información proporcionada por ambas imágenes se complementa haciendo el sistema más robusto y resistente a una mayor variedad de circunstancias y escenas[3].

1.1.1 Motivación

Haciendo una recopilación del estado del arte (capitulo 3), podemos notar que a pesar de existir diversos sistemas robustos de fusión de imágenes aplicadas en el seguimiento de objetivos, existen varios problemas que no han sido resueltos en su totalidad, entre ellos podemos encontrar:

- Pérdida del objetivo por oclusión: Para el caso particular de seguimiento de personas, esto ocurre cuando una persona u objeto se

interpone entre la cámara y la persona objetivo del seguimiento. Cuando la cámara pierde de vista el objetivo, en algunas ocasiones no lo vuelve a encontrar después de que la oclusión ha terminado.

- Confusión del objetivo con un objeto similar: Para el caso de seguimiento de personas, este problema es común en lugares donde hay personas uniformadas, como en campos de futbol, campos militares, escuelas, hospitales y laboratorios, aunque también llega a ocurrir espontáneamente en otro tipo de ambientes y se genera cuando dos personas tienen características similares, como sería la ropa, tal vez color de piel y/o complexión física.

- Ahorro de espacio de memoria en sistemas de vigilancia: La mayoría de los seguidores de personas no toman en cuenta la cantidad de espacio que necesita un sistema de video-vigilancia para almacenar todas las imágenes, ya que la mayoría trabajan con video a una velocidad de 24 a 30 fps, lo cual llevado a una aplicación de este tipo resulta ser poco eficiente.

Lo que se realizó en este trabajo fue utilizar la fusión de secuencias de imágenes térmicas y visuales para la detección y seguimiento de múltiples personas al mismo tiempo. El proceso de seguimiento debe ser resistente a secuencias de imágenes en ambientes con iluminación no controlada, que presenten oclusión y con personas uniformadas. Además de funcionar para secuencias de imágenes a 4 fps mínimo.

1.2 Objetivos

1.2.1 Objetivo general

Desarrollar y evaluar un método aplicado a la detección automática y seguimiento de múltiples personas en secuencias de imágenes capturadas a 4fps, que presenten oclusión y presencia de personas uniformadas, utilizando para ello fusión de secuencias de imágenes térmicas y visuales.

1.2.2 Objetivos específicos

I. Desarrollar un método para obtención del modelo de características de una persona en imágenes térmicas y visuales.

II. Proponer el algoritmo para detectar la(s) persona(s) en la escena e implementarlo dentro del entorno propuesto.

III. Desarrollar un método de seguimiento multi-objetivo útil para secuencias capturadas a 4 cuadros por segundo.

1.3 Solución propuesta

En esta tesis se propone un método para la detección automática y seguimiento de personas, en dos etapas de fusión, la primera es una fusión a nivel decisión y la segunda es una fusión a nivel característica. El método propuesto se puede apreciar en la figura 1.

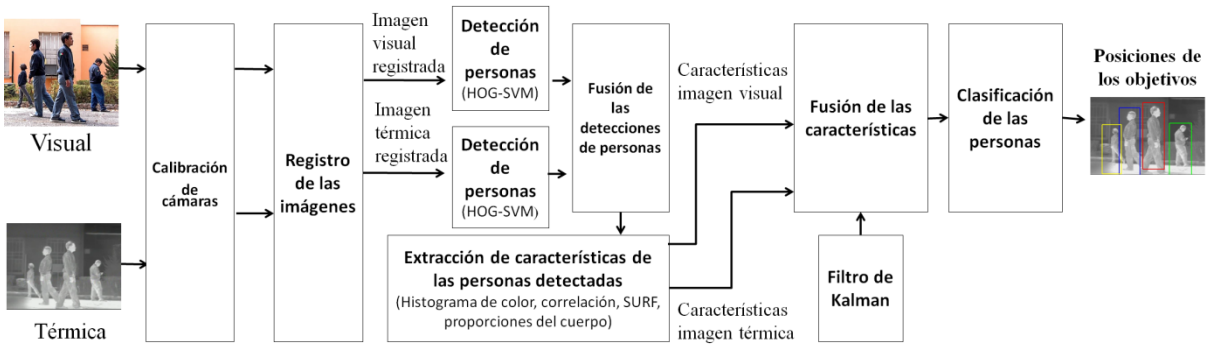


Figura 1. Método propuesto.

Los primeros dos bloques representados por dos imágenes simbolizan la captura de las imágenes provenientes tanto de la cámara térmica como la visual. Posteriormente se debe realizar la calibración de ambas cámaras para reducir la deformación que provoca el lente de cada cámara sobre sus imágenes respectivas.

Una vez lograda la calibración, ambos tipos de imagen deben ser registrados, unificando así el sistema de coordenadas, usado el sistema de las imágenes térmicas como referencia. Posteriormente, cada tipo de imagen es sometido a

un detector automático de personas de forma individual y ambos resultados son fusionados a nivel decisión, obteniendo finalmente solo un resultado global del detector automático de personas.

De las regiones de la imagen donde se han detectado la presencia de persona, se obtienen características, en ambos espectros, para la identificación y distinción entre cada persona. Estas características son fusionadas en un modelo de representación propio de cada individuo y después estos modelos son sometidos a un clasificador para identificar y seguir durante la secuencia de imágenes a cada persona.

1.4 Alcance de la tesis

- La metodología se propone para seguimiento de personas adultas y en posición erguida.
- Las personas pueden moverse sobre los ejes *x* y *z* con respecto a la posición de la cámara (ver figura 2).
- Las secuencias de imágenes son capturada en ambientes exteriores y sin control de iluminación.
- El seguidor funciona para un rango entre 2m a 6 m de distancia a partir de la cámara (ver figura 2).

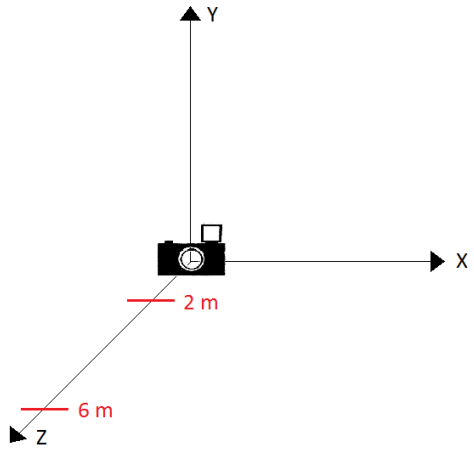


Figura 2. Ejes del mundo real con respecto a la cámara

1.5 Organización de la tesis

El panorama general de esta tesis está organizado como sigue: En el capítulo 2 se explican brevemente los fundamentos teóricos necesarios para comprender el desarrollo de la tesis. En dicho capitulo se encuentran conceptos básicos y explicaciones generales de terminología del tema de la tesis. En el capítulo 3 podemos encontrar los principales y más recientes trabajos relacionados con la tesis, los cuales se encuentran divididos en dos secciones: los seguidores que utilizan fusión de sensores y los seguidores que utilizan un solo sensor. En el capítulo 4, se encuentra una explicación detallada de la estrategia seguida para la detección automática y el seguimiento de personas, se explican puntualmente los algoritmos usados para la fusión de ambos tipos de imágenes y la clasificación de las personas. El capítulo 5 explica cómo fueron realizados los experimentos y los resultados obtenidos. Por último en el capítulo 6 tenemos las conclusiones de la tesis y el trabajo a futuro.

Capítulo 2

Fundamentos teóricos

En este capítulo se abordaran las nociones básicas necesarias para el entendimiento y comprensión de lo que se propone en esta tesis. Se explica algo de la terminología usada en el área, así como conceptos y trabajos relacionados con el tema.

2. 1 Detección y seguimiento de objetos/personas

Durante el proceso de captura de una imagen digital existen diferentes espacios coordenados, los cuales se pueden dividir en cuatro [4]:

- Coordenadas del mundo: Estas son coordenadas en 3D, fijadas en el espacio del mundo real con respecto a una esquina o punto de dicho espacio.
- Coordenadas de la proyección de la imagen: Estas coordenadas se refieren a la proyección de los objetos 3D del mundo real a una imagen plana en un espacio 2D.
- Coordenadas de la cámara: Son coordenadas en un espacio 3D, fijadas en la cámara. El origen del sistema coordenado es el centro de

proyección de la cámara. El eje z es tomado como el eje óptico de la cámara (ver figura 3).

- Coordenadas de la imagen en pixeles: Estas coordenadas se encuentran en un espacio 2D y cada pixel está representado por dos números, el eje *y* indica las filas y el eje *x* indica las columnas. El origen de este sistema se encuentra en la esquina superior izquierda de la imagen (ver figura 3).

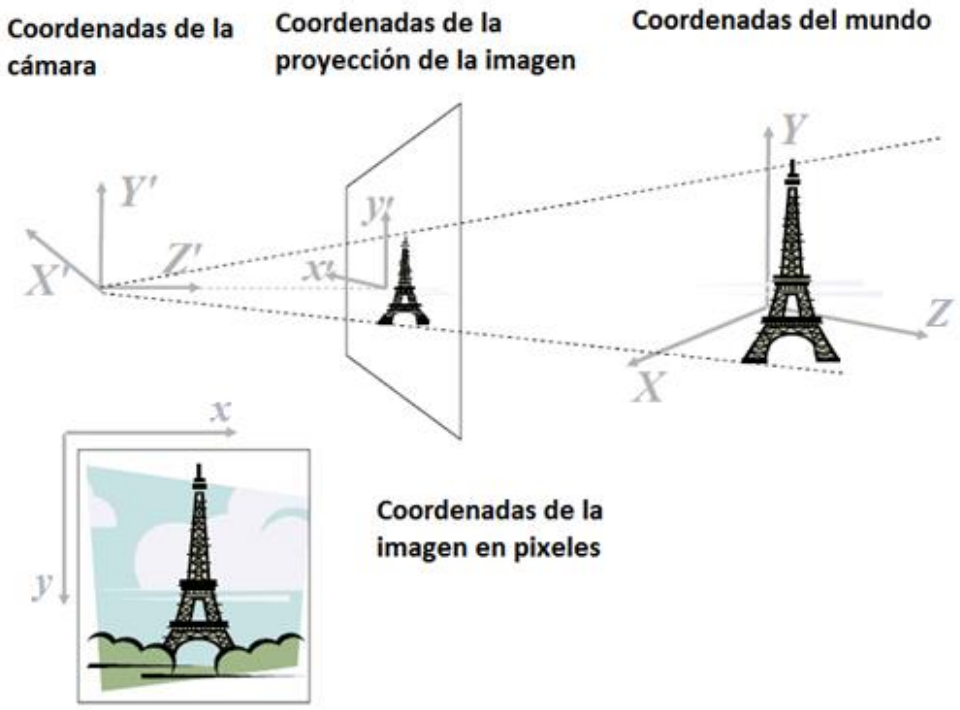


Figura 3. Geometría de las imágenes [4]

El proceso de detección de un objeto en la imagen se desarrolla dentro del espacio coordenado de la imagen en 2D. La detección consiste en extraer las mismas características tanto del objeto que se desea identificar como del resto de los diferentes objetos en la imagen, de tal forma que las características puedan ser comparables entre sí y al encontrar un grado de similitud elevado, según los métodos de evaluación del detector, se podrá asumir que se trata del mismo objeto. Por lo general, se obtienen como datos de salida la posición del objeto detectado en el sistema de coordenadas de la imagen.

Para el caso específico de la detección de personas el concepto es el mismo que con cualquier otro objeto, solo que se adecuan las características a extraer para la detección de la presencia y posición de cualquier persona en la imagen.

El seguimiento automático de personas en una secuencia de imágenes consiste en detectar la presencia y posición de una o varias personas en cada una de las imágenes, identificando a cada individuo en particular con el objetivo de determinar la trayectoria de su movimiento a lo largo de la secuencia.

2. 2 Calibración de las cámaras

En muchos de los procedimientos que implican el uso y procesamiento de imágenes es necesario realizar una calibración de la o las cámaras utilizadas para capturar dichas imágenes. Como vimos anteriormente, la captura de una imagen incluye una cámara provista de un lente que realiza una transformación entre un objeto del mundo real en 3D y una imagen plana. Esta transformación por lo regular no puede ser descrita perfectamente tanto por la perspectiva de la cámara como por la distorsión que introduce el lente y que ocurre entre los puntos reales del objeto y la localización de esos puntos en la imagen. Sin embargo, esta distorsión puede aproximarse por medio de un modelo matemático. La exactitud de dicho modelo dependerá de que tan bien sean estimados los parámetros de la cámara. Existe dos tipos de parámetros: intrínsecos y extrínsecos. Los parámetros intrínsecos o internos son los que describen el funcionamiento de una cámara y la posición de los pixeles con respecto a las coordenadas del cuadro tomado por la cámara, por ejemplo: distancia focal y centro óptico. Los parámetros extrínsecos o externos definen la posición y orientación del cuadro tomado por la cámara con respecto a la orientación de la cámara en el mundo real [5]. En general la calibración consiste en conocer las medidas de varios puntos en el mundo real y compararlas con las proporciones de los mismos puntos capturados en la imagen, para así calcular alguna transformación que acerque las proporciones de la imagen a las reales.

Los inicios de la calibración de cámaras tuvieron origen en la vigilancia de territorios, por lo que los primeros métodos fueron pensados para rangos lejanos. Sin embargo, actualmente el uso de imágenes abarca muchas más

aplicaciones con un rango cercano de enfoque y ha migrado a la era digital. En la actualidad podemos clasificar los métodos en dos tipos [6]:

- Calibración fotométrica: Es realizada a partir de la observación de un objeto en el mundo real donde su geometría en 3D es conocida con una buena precisión. Consta de tres planos ortogonales entre sí. Su principal desventaja es que se requieren aparatos costosos y un montaje elaborado.

- Auto-calibración: No se usa ningún objeto, solo se mueve la cámara en una escena estática. La escena rígida provee información que permite calcular los parámetros intrínsecos y extrínsecos de la cámara. La principal ventaja de estos métodos de calibración es que son muy flexibles, sin embargo, no están totalmente maduros pues existen parámetros que no siempre se pueden obtener.

El método de calibración que se utilizó en este trabajo fue el implementado en el *toolbox* de calibración de cámara de MATLAB 8. Este se encuentra en un punto medio entre los métodos antes mencionados y está inspirado principalmente en [6] y en [7]. La calibración usada en esta tesis consiste en fotografiar un patrón con medidas conocidas en el sistema coordenado del mundo en diferentes posiciones, para a partir de esto calcular el grado y la forma de la distorsión que la lente de la cámara produce en la imagen. Una vez obtenida una aproximación de la deformación se obtiene la transformada inversa, la cual se aplica a cada imagen reduciendo la distorsión de esta y obteniendo una imagen más fiel al mundo real.

En general la relación entre un punto 3D en el sistema de coordenadas del mundo y la proyección de dicho punto en una imagen está dada por el siguiente modelo:

$$x = KRX \qquad \text{ec.1}$$

Donde x= [x y 1] ^{T} es el vector de las coordenadas del punto en la imagen, X=[X_Y_Z_1]^{T} representa la posición del punto 3D en el sistema de coordenadas del mundo, K es la matriz intrínseca de la cámara y R son los parámetros extrínsecos de la cámara que relacionan las coordenadas del mundo

con las coordenadas de la cámara. Donde K está conformada por la distancia focal f, las coordenadas del centro óptico (u_0 v_0) y el sesgo de la imagen s.

$$K = \begin{bmatrix} f & s & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

Adentrándonos un poco en el método desarrollado e implementado en el *toolbox* de Matlab podemos resumirlo en los siguientes pasos:

1. Imprimir un patrón y pegarlo en una superficie plana.
2. Tomar algunas imagines del patrón en diferentes orientaciones moviéndolo respecto al plano de la cámara.
3. Detectar los puntos característicos en las imágenes utilizando el patrón de calibración, en el que se detectan las direcciones de las líneas encontrado aquellas que son ortogonales entre sí y los puntos que se encuentran en la intersección de la cuadricula.
4. Estimar los parámetros intrínsecos y extrínsecos usando el método de ortogonalidad de puntos de fuga. Dicho método en general cosiste en que una vez detectados los puntos ortogonales en la imagen, lo cuales llamaremos x_i, x_j y x_k, se calculan los vectores unitarios del punto en el sistema coordenado del mundo, dados por:

$$e_i = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, e_j = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, e_k = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Por lo tanto de la ec. 1:

$$x_i = KRe_i \qquad \text{ec.2}$$

$$x_j = KRe_j \qquad \text{ec.3}$$

$$x_k = KRe_k \qquad \text{ec.4}$$

Como los vectores son ortogonales entre sí:

$$e_i{}^T e_j = 0$$

Sustituyendo y dado que R es una matriz de rotación, tenemos que:

$$x_i{}^T K^{-T} R R^T K^{-1} x_i = x_i{}^T K^{-T} K^{-1} x_i = 0$$

De forma análoga se hace con los otros dos puntos y asumiendo que el sesgo de la imagen es cero se obtiene un sistema de tres ecuaciones con tres incógnitas.

Ya que se ha calculada la matriz K es fácil encontrar los valores de la matriz R, utilizando las ecuaciones 2 a 4 y sustituyendo los valores de K se obtienen tres sistemas de ecuaciones de 3x3.

5. Refinar todos los parámetros mediante el método de estimación de máxima probabilidad (MLE- Maximum Linkedhood Estimation). El MLE puede ser obtenido minimizando la siguiente ecuación:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left\|x_{ij}-\hat{x}(K,R_i,X_j)\right\|^2 \qquad \text{ec.5}$$

Donde $\hat{x}(K,R_i,X_j)$ es la proyección del punto X_j en la imagen i. La minimización es no lineal y se realiza utilizando el algoritmo de Levenberg-Marquardt (ver referencia [6]).

2.3 Transformaciones Geométricas

Las transformaciones geométricas son aquellas que se encargan de modificar la relación espacial entre los pixeles de una imagen. Estas tienen muchos usos desde modificaciones estéticas hasta corrección de distorsiones o alinear múltiples imágenes de la misma escena [8].

Principalmente las transformaciones geométricas de una imagen se pueden dividir en cinco tipos [9]:

1. De traslación en el eje *x* y/o en el eje *y*
2. De cuerpo rígido: incluye traslación en el eje *x* y/o en el eje *y* y rotación.
3. De similaridad: traslación en el eje *x* y/o en el eje *y*, rotación y escalamiento uniforme.
4. Afín: consiste en una parte lineal expresada como una matriz de multiplicación y una parte aditiva que representa una traslación. Suponiendo que las coordenadas [w z] pertenecen a la imagen original y las coordenadas [x y] pertenecen a la imagen transformada, la transformación afín se puede definir como:

$$[x\ y] = [w\ z] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + [b_1\ b_2] \qquad \text{ec.6}$$

Manipulando matemáticamente la ecuación 6, puede ser expresada como:

$$[x\ y\ 1] = [w\ z\ 1] \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ b_1 & b_2 & 1 \end{bmatrix} \qquad \text{ec.7}$$

La transformación afín incluye escalamiento en *x* y/o *y*, rotación, traslación en *x* y/o *y*, deformación en *x* y/o *y*, reflexión horizontal o vertical e identidad. En la tabla 1 se explica brevemente los casos básicos de transformaciones afines, el resto de los casos son combinaciones de estos [9].

Como se puede notar las transformaciones mencionadas en los puntos uno, dos y tres pueden tomarse como un caso particular de una transformación afín.

5. Proyectiva: esta transformación es útil para revertir la distorsión por perspectiva. Al igual que la transformada afín la transformada proyectiva usa una tercera dimensión auxiliar, sin embargo, para la transformada proyectiva esta tercera coordenada auxiliar (denominada como h en la ec. 8) no es unas constante [9]:

$$[x'y'\ h] = [w\ z\ 1] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ b_1 & b_2 & 1 \end{bmatrix} \qquad \text{ec.8}$$

Donde *x=x'/h* y *y=y'/h.*

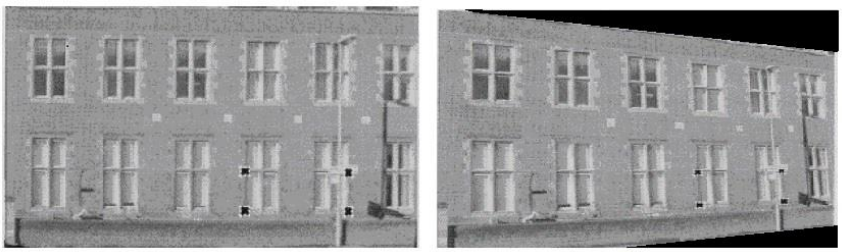


Figura 4. a) Imagen original, b) Transformada proyectiva [10]

2.4 Registro de imágenes

Una de las principales aplicaciones de las transformaciones geométricas es el registro de imágenes, el cual es un procedimiento necesario para casi cualquier tipo de fusión que involucre dos imágenes de la misma escena. El registro de imágenes es el procedimiento mediante el cual se alinean dos fotografías de tal forma que los pixeles correspondientes en ambas imágenes representen el mismo punto físico en la escena. Cuando las imágenes adquiridas provienen de un tipo de sensor diferente se le llama registro de imágenes multimodal [9].

En general el registro de imágenes consiste en tomar una imagen como referencia, encontrar los pixeles correspondientes a un mismo punto físico en ambas imágenes y posteriormente calcular la transformación que lleve de la posición del pixel de la imagen no referenciada a la posición del pixel de la imagen referenciada.

Existen métodos muy variados para realizar el registro de imágenes pero en general todos se pueden resumir en los siguientes pasos [9]:

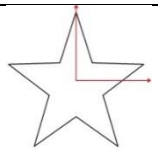
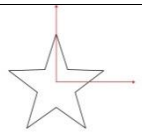
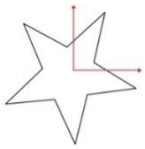
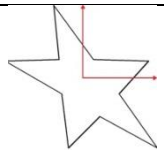
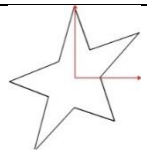
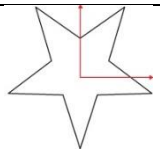
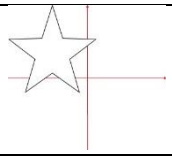
1. Detección de características: Una característica es una porción de imagen que potencialmente puede ser definida y localizada en ambas imágenes. Puede o no ser sometida a cualquier tipo de transformación o procedimiento.

2. Correspondencia de características: Utilizando uno o varios métodos cada característica de la imagen base debe ser rastreada hasta encontrar su coincidencia en la segunda imagen. Estos métodos pueden ser manuales, semi-manuales o automáticos dependiendo del algoritmo.

3. Inferencia de la transformación geométrica: Dependiendo de las posiciones de las características correspondientes, se realiza un cálculo para determinar un tipo de transformación geométrica que satisfaga a la mayoría de las correspondencias.

4. Aplicar la transformada para alinear las imágenes

Tabla 1
Transformaciones Afines básicas [9]

Tipo	Matriz de transformación	Ecuación de las coordenadas	Diagrama
Identidad	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = w$ $y = z$	
Escalamiento	$\begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = s_x\, w$ $y = s_y\, z$	
Rotación	$\begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = w\cos\theta - z\sin\theta$ $y = w\sin\theta + z\cos\theta$	
Deformación horizontal	$\begin{bmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = w + \alpha z$ $y = z$	
Deformación vertical	$\begin{bmatrix} 1 & \beta & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = w$ $y = \beta w + z$	
Reflexión vertical	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = w$ $y = -z$	
Reflexión horizontal	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x = -w$ $y = z$	Como la imagen es simétrica, quedaría igual al diagrama de la matriz identidad.
Traslación	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \delta_x & \delta_y & 1 \end{bmatrix}$	$x = w + \delta_x$ $y = z + \delta_y$	

A continuación se presenta una clasificación de algunos de los métodos para el registro de imágenes, esta clasificación solo es una de muchas otras existentes y está basada en [8].

- Métodos de correlación: Usan métricas similares a la correlación cruzada, la correlación de coeficientes o métricas definidas por la suma de las diferencias absolutas de las características de la imagen. Las principales limitaciones que presentan este tipo de métodos son que no funcionan correctamente para imágenes multimodales, el ruido de una imagen suavizada con respecto a otra puede hacer fallar el método, además que el costo computacional aumenta dependiendo del tamaño de la ventana que representa a las características.

- Métodos que utilizan transformada de Fourier*:* Este tipo de métodos explotan el hecho de que en el dominio de la frecuencia la traslación, rotación y escala tienen una clara representación. Los métodos que utilizan Fourier son apropiados para imágenes con cambios de iluminación notorios, por lo general tienen mejores resultados que los métodos de correlación, sin embargo, están restringidos solo a transformadas de traslación, rotación y escala.

- Métodos basados en características: En este tipo de métodos los parámetros de transformación son encontrados a partir de características extraídas de las imágenes. El principal problema que presentan es que dependen totalmente de que el extractor de características que se utilice sea confiable y dependiendo de este, se calcula el costo computacional, la precisión y la efectividad para imágenes con ruido o multimodales.

- Método de mapeo de puntos: Se refiere a los métodos que usan control de puntos dispersos detectados en la imagen para establecer correspondencias que permitan deducir la transformada entre ambas imágenes. Estos son útiles cuando el tipo de transformación es desconocida. Su principal desventaja es que su efectividad depende

del paso de extracción de puntos característicos y la cantidad de puntos detectados.

- Métodos que utilizan información mutua: Estos métodos están basados en la dependencia estadística o información redundante entre las imágenes en escala de grises. Esta se define como:

$$MI = \sum_{i=1}^{m} \sum_{j=1}^{n} P(g_i, g_j) \ln\left(\frac{P(g_i, g_j)}{P(g_i)P(g_j)}\right) \qquad \text{ec.9}$$

Donde m y n son el número de niveles de grises en las imágenes I y J respectivamente, $P(g_i)$ y $P(g_j)$ representan el número de pixeles en las imágenes I y J con valores g_i y g_j sobre el total de pixeles y $P(g_i, g_j)$ es el número de pixeles pares con valores (g_i, g_j) sobre el número total de pixeles.

Los métodos que usan información mutua pueden soportar variaciones substanciales entre las imágenes e inclusive algunos tipos de imágenes multimodales [8].

2.5 Fusión de imágenes

La fusión de múltiples sensores consiste en utilizar la información recabada de diferentes sensores, sean del mismo tipo o diferentes, para construir a partir de esto un sistema capaz de resolver problemas que sin la contribución de los datos de ambos sensores no sería posible resolver o la solución sería mucho más compleja o menos efectiva.

La fusión multi-sensor tiene una amplia gama de aplicaciones, de las cuales en el sector que se ha desarrollado más es en el militar. Sin embargo, también es útil en otros campos. Algunos ejemplos de sus aplicaciones en general son: vigilancia del territorio oceánico y aéreo, detección de objetivos en campos de batalla, detección de armas y sujetos armados, sensado remoto del estado de depósitos minerales, cosechas, bosques y selvas, monitoreo automático de equipo, diagnósticos médicos, aplicaciones en robótica, entre otras [11].

Como se puede notar los sensores utilizados para cada una de las aplicaciones anteriormente mencionadas pueden ser de diferentes tipos, como acústicos, electromagnéticos, de temperatura, ópticos, químicos, cámaras fotográficas o de video, mecánicos, etc. En particular el desarrollo de este trabajo se enfoca en la fusión de secuencias de imágenes tanto en el espectro visual como en el espectro infrarrojo, el motivo de la elección de estos dos tipos de imágenes radica en su bajo costo en comparación con otro tipo de imágenes (por ejemplo ondas milimétricas o rayos X), lo cual las hace factibles para aplicarse en la solución de diversos problemas [11].

2.6 Niveles de fusión

La fusión multi-sensor puede ocurrir en diferentes niveles, los cuales pueden variar según la literatura consultada, sin embargo, para propósitos de este trabajo explicaremos los niveles básicos de fusión [8]:

- Nivel señal: Consiste en combinar directamente cada señal proveniente de cada sensor con el objetivo de obtener una señal en el mismo formato que la original. Para efectos de este trabajo, definimos las imágenes de entrada como señales en 2D, de tal forma que también llamamos a este nivel de imagen o de pixel, ya que cada pixel de la imagen de salida es determinado a través de una operación entre los pixeles de las imágenes de entrada.
- Nivel característica: Primero se extraen características de cada una de las señales de entrada, posteriormente todas estas características se utilizan de manera conjunta como información para algún propósito, por ejemplo la representación de un objeto.
- Nivel decisión: Como su nombre lo indica se encarga de tomar decisiones, para ello utiliza de forma conjunta la información proveniente de uno o varios procesos de razonamiento a los que fueron sometidos los datos de cada una de las señales de entrada.

Es importante hacer notar que usualmente un sistema multi-sensor proporciona información que se puede fusionar en uno o más niveles. Es decir, los sistemas de fusión multinivel se caracterizan por realizar diferentes niveles de fusión en

diferentes etapas del proceso, como es el caso del trabajo que se presenta en esta tesis.

2.7 Histograma del Gradiente Orientado

Para detectar la presencia de una persona en una imagen existe una gran variedad de métodos, los cuales pueden ser clasificados dependiendo de diversas características. En esta tesis nos enfocaremos en una clasificación según el número de sensores que se utilice en la detección de personas, es decir si presentan algún tipo de fusión o no.

Actualmente uno de los métodos más destacados es la detección de personas utilizando histogramas del gradiente orientado (HOG por sus siglas en inglés: *Histograms of Oriented Gradient*) [1]. Este método fue publicado en el año 2005, desde entonces a la fecha (27 de octubre del 2012) ha sido citado por alrededor de 6424 artículos y se han desarrollado alrededor de 106 versiones que han desarrollado pequeñas mejoras pero en esencia el método no ha cambiado.

En un principio el método HOG fue ideado como un detector de personas para imágenes en el espectro visual, es decir un método de un solo sensor. Sin embargo, algunos trabajos como [12] han demostrado que también presenta buenos resultados en el uso de fusión de imágenes infrarrojas y visuales.

El método HOG para la detección de personas busca representar una imagen utilizando como descriptor un conjunto de histogramas de las direcciones del gradiente de los pixeles de una imagen, estas direcciones discretizadas son llamadas canales de orientación y a partir de dicha representación se realiza una clasificación SVM para determinar que imágenes enmarca a una persona. De forma general puede resumirse en los siguientes pasos [1]:

1. Ventanas de detección: La imagen que se desea procesar se divide en ventanas de detección, que en realidad son segmentos cuadrados o circulares de la imagen. Cada ventana será analizada para saber si en ella se enmarca a una persona.

2. Normalización gamma: En cada ventana de detección se aplica opcionalmente una ecualización gamma de la ley de potencia, la cual tiene un ligero efecto en el rendimiento del método. La ecualización gamma está definida por:

$$V_{out} = AV_{in}^{\gamma}$$

Donde $\gamma<1$, usualmente A=1 y $\gamma=0.5$.

3. Calculo del gradiente: Se calcula el gradiente de la ventana utilizando cualquiera de las máscaras o *kernels* ya conocidas (Sobel, Prewitt, Roberts, [-1, 1] $[-1, 1]^T$, [-1 0 1] $[-1\ 0\ 1]^T$, etc.). Se recomienda utilizar la máscara [-1 0 1] para el eje *y* y $[-1\ 0\ 1]^T$ para el eje *x*. En conjunto han demostrado trabajar mejor que el resto para este método en específico [11].

   Por cada pixel se calcula la magnitud (ec. 10) y el ángulo (ec.11) del gradiente:

$$M = \sqrt{Gx^2 + Gy^2} \qquad \text{ec.10}$$

$$\theta = \tan^{-1}\frac{Gy}{Gx} \qquad \text{ec.11}$$

   Donde *Gx* es el gradiente en *x* y *Gy* es el gradiente en *y*, para cada pixel respectivo.

4. Votación ponderada de la orientación espacial del gradiente: Se crean canales de orientación los cuales pueden ser sin signo de 0° a 360° o con signo de 0° a 180°. Aumentar el número de canales de orientación mejora significativamente el rendimiento por lo que se recomienda que mínimo sean 9 canales, después de los 9 canales el rendimiento es similar. Ejemplo: los canales pueden ser para el caso con signo: 0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5° y 180°.

   Cada pixel ejerce un voto por un canal dependiendo del que sea más cercano a su ángulo, dicho voto es ponderado por su magnitud. Los

votos ponderados son acumulados por regiones espaciales locales llamadas células, las cuales pueden ser cuadradas o redondas y el tamaño puede variar preferentemente entre 4x4 y 12x12 pixeles.

5. Normalización de los contrastes sobre los bloques espaciales sobrepuestos: las magnitudes y ángulos del gradiente varían sobre un rango amplio de cambios de iluminación y contrastes, por lo tanto una normalización efectiva de los contrastes locales resulta esencial para un buen rendimiento. Para ello se agrupan células en bloques espaciales más grandes y estos bloques se normalizan por separado. Los bloques se van sobreponiendo entre sí de tal forma que cada célula contribuye varias veces en el descriptor final. El tamaño de los bloques puede variar preferentemente entre 1x1 y 4x4 células.

   Algunas de las reglas de normalización que se recomiendan en este paso son:

$$v \rightarrow \frac{v}{\sqrt{\|v\|_2^2+\epsilon^2}} \qquad \text{ec.12}$$

ó

$$v \rightarrow \sqrt{\frac{v}{(\|v\|_1+\epsilon)}} \qquad \text{ec.13}$$

   Donde $\|v\|_1$ es la norma 1, $\|v\|_2$ la norma 2 y ϵ es una constante pequeña.

6. Recolección de los HOG en toda la ventana de detección: El descriptor final consiste en un vector donde se concatenan las respuestas normalizadas de todos los bloques de células.

7. Clasificación con SVM: Por último se utiliza un clasificador SVM para clasificar las ventanas que enmarque a una persona de las que no. El entrenamiento del clasificador en el artículo original se realizó con 2478 casos positivos (con personas) y 12180 casos negativos (libres de personas). La eficiencia obtenida varía según los parámetros del descriptor, como se aprecia en la figura 5, la mejor

eficiencia lograda es de una tasa de pérdida del 10.4%, utilizando células de 6x6 pixeles y bloques de 3x3 células.

$$Tasa\ de\ pérdida = \frac{FN}{VP + FN}$$

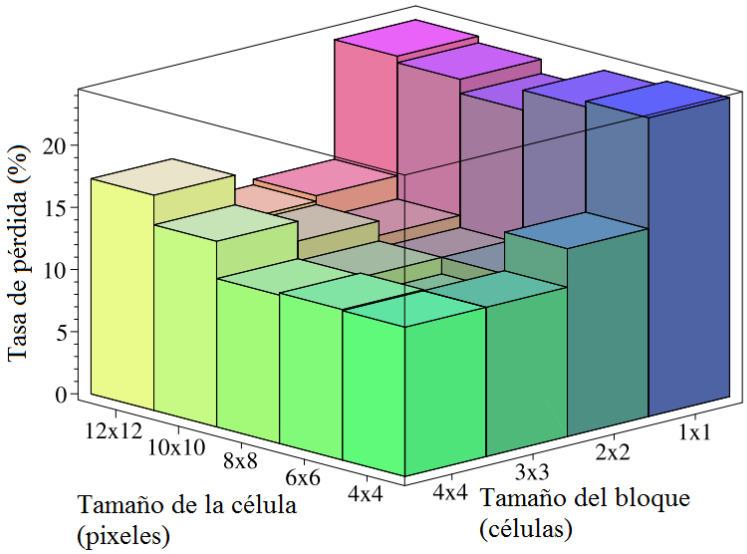


Figura 5. Tasa de pérdida del método HOG según el cambio de tamaño de las células y los bloques [1].

2.8 Filtro de Kalman

El filtro de Kalman es un algoritmo óptimo de procesamiento de datos recursivo que incorpora toda la información disponible de un sistema lineal, independientemente de su precisión, para estimar el valor actual de la variable de interés. Hace uso del modelo del comportamiento del sistema, la descripción del ruido (que debe ser gaussiano), la incertidumbre el modelo dinámico, las medidas del error y del sistema a través del tiempo y cualquier información disponible sobre las condiciones iniciales de la variable de interés [13].

Se puede decir que el filtro de Kalman es óptimo debido a que incorpora toda la información que puede ser provista, además de que combina todos los datos de las medidas disponibles, más el conocimiento anterior sobre el sistema y las

medidas anteriores, para producir un estimado de las variables de estado de tal manera que el error es minimizado en cada iteración[13].

El filtro de Kalman tiene una gran variedad de aplicaciones desde áreas tan remotas como economía hasta mecánica de fluidos, etc. En el contexto que utilizaremos el filtro de Kalman en esta tesis, es como un método probabilístico de predicción, para el seguimiento visual de personas en secuencias de imágenes.

Para realizar una predicción de una posición usando el filtro de Kalman, en general, inicialmente se observa una variable en el tiempo t, a la que llamaremos y_t y dados los estados anteriores de la variable, se realiza una predicción de cuál será su próximo estado en el tiempo t+1. Despues que la predicción es calculada, se observa la medida y_{t+1} y se corrige el error entre la predicción y la medida. El proceso se repite para t+2 y así sucesivamente.

Ahora bien, adentrándonos un poco en el método, debemos tomar en cuenta que lo primero que se asume en el filtro del Kalman, es que se está trabajando con un sistema dinámico lineal, el cual puede ser modelado por la siguiente ecuación:

$$x_{t+1} = Ax_t + w_t \qquad \text{ec.14}$$

Dicha ecuación representa la transición de un estado en el tiempo t a t+1. Dónde A es la matriz de transición del estado, x_t es la variable de estado en el tiempo t y w_t es una variable de ruido gaussiano con media cero y una matriz de covarianza Q.

También es necesario modelar la relación entre el estado y la medida, que también se asume que es lineal, por lo tanto la medida y_t puede ser expresada en términos del estado x_t:

$$y_t = Cx_t + v_t \qquad \text{ec.15}$$

Donde C es la matriz de relación entre la variable de estado y la medida, y v_t, similar a w_t, es una variable de ruido con distribución normal y una matriz de covarianza R [14].

Para aplicar el algoritmo de Kalman, este se puede dividir en dos etapas: la etapa de predicción y la etapa de corrección:

- Etapa de predicción: Consta de dos pasos, el primer paso consiste en: dado un estado en el tiempo t se hace un cálculo del valor esperado del mismo estado en el tiempo t+1,utilizando la ecuación 9, que representa el comportamiento dinámico del sistema. Posteriormente, en el Paso dos se calcula una predicción de la matriz de covarianza del error (ecuación 16).

$$P_{t+1} = AP_tA^T + Q \qquad \text{ec.16}$$

- Etapa de corrección: Consta de tres pasos, el primero consiste en realizar el cálculo de la ganancia de Kalman a partir de la matriz de covarianza del error, la matriz de relación entre la variable de estado y la medida, y la matriz de covarianza del ruido de la medida (ecuación 17).

$$K_{t+1} = P_{t+1}C^T(CP_{t+1}C^T + R)^{-1} \qquad \text{ec.17}$$

En el siguiente paso, una vez que se ha observado la medida y_{t+1}, se actualiza el valor del estado x_{t+1} utilizando para ello la ecuación 18.

$$x_{t+1} = x_{t+1} + K_{t+1}[y_{t+1} - Cx_{t+1}] \qquad \text{ec.18}$$

Por último se actualiza la matriz de covarianza del error P_{t+1} (ecuación 19).

$$P_{t+1} = [I - K_{t+1}C]P_{t+1} \qquad \text{ec.19}$$

Con los valores ya actualizados se repite la etapa de predicción y después la de corrección para el tiempo t+2 y así sucesivamente.

Cuando las secuencias de imágenes no son generadas artificialmente el ruido del proceso (w_t) y el ruido de la medida (v_t) son desconocidos, sin embargo, estos pueden ser estimados directamente de la secuencia de imágenes siempre que se conozca la velocidad o aceleración en cada imagen, no obstante estos datos no siempre están disponibles en los problemas reales.

Si la posición, velocidad y aceleración son conocidas en cada punto del tiempo, estimar el ruido del proceso es relativamente simple. El valor del ruido en cada

intervalo de tiempo puede ser calculado mediante la obtención de la diferencia entre el estado actual x_t y la transición del estado previo (Ax_{t-1}). Pero si la información no es conocida, es necesario estimar el ruido de la medida y del proceso usando alguna técnica.

En nuestro caso específico, se usó la técnica iterativa para determinar la matriz del ruido del proceso Q y la matriz del ruido de la medida R, las cuales se explican a continuación.

Para Q, inicialmente se sugiere una "buena aproximación", que puede ser propuesta a partir de la estimación de los datos de posición, velocidad y aceleración del objeto a seguir (ver capítulo 4). Después la iteración comienza siguiendo los pasos indicados continuación:

1. Realizar el algoritmo de seguimiento y guardar los estados estimados de x.
2. Determinar el ruido del proceso w_t con la siguiente ecuación:
$$w_t = x_t - Ax_{t-1}$$
3. Determinar la matriz de covarianza de w_t, que será el nuevo valor de Q y continuar iterando hasta que Q converja (ver capítulo 4).

Para la matriz R el proceso es muy similar. Inicialmente se sugiere R después de correr el algoritmo de seguimiento una vez, R es ajustada a la covarianza del error de la medida (determinado a partir de la diferencia entre la medida y la posición verdadera). Esto se repite hasta que R converge [14] (ver capítulo 4).

2.9 Descriptor SURF

El método SURF es un detector y descriptor de características en imágenes, presentado por Herbet Bay en el año 2006 [15]. Este método está parcialmente inspirado en el método SIFT [16], ambos detectan características invariantes a traslación, escala, rotación, iluminación y transformaciones afines. Las dos principales diferencias entre ambos métodos consiste en que SURF presenta una mayor velocidad de cálculo y mayor robustez ante posibles transformaciones [15].

El método SURF está basado principalmente en sumas de respuestas Haar-wavelet-2D en un espacio escala aproximado al gaussiano y hace un uso eficiente de imágenes integrales. Un espacio escala gaussiano de una imagen es el resultado de las convoluciones de funciones gaussianas-2D, con diferentes valores de σ, con la imagen original. Por lo tanto una espacio escala aproximado al gaussiano es la respuesta de las convoluciones de funciones aproximadas a una gaussiana en 2D con la imagen original [15].

De forma general el algoritmo SURF se puede describir en los siguientes pasos:

1. Se calcula la matriz Hessiana para cada valor σ, donde σ es el valor de la escala del filtro gaussiano y la matriz Hessiana H(p,σ) de un punto p=(x,y)en una imagen I(x,y) se define como:

$$H(p,\sigma) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{xy}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix}$$

   $L_{xx}(p,\sigma)$ es la convolución de la derivada parcial de segundo orden de la función gausiana con la imagen original en el punto p. Análogamente se definen $L_{xy}(p,\sigma)$ y $L_{yy}(p,\sigma)$. Sin embargo, para el caso del método SURF, para realizar las convoluciones se hace uso de aproximaciones del filtro gaussiano a un filtro tipo caja (ver figura 6) y la convolución se realiza usando imágenes integrales, las cuales son definidas por la ecuación 15:

$$I_{\Sigma}(x,y) = \sum_{i=1}^{i\leq x}\sum_{j=1}^{j\leq y} I(i,j) \qquad \text{ec.20}$$

   Donde I_{Σ} es la imagen integral e I es la imagen original. Las imágenes integrales son importantes para la aplicación del método SURF ya que permiten calcular operaciones como la convolucion con filtros tipo caja simétricos con un número fijo de sumas independiente al tamaño de la imagen. Las aproximaciones de las derivadas parciales son denotadas como D_{xx}, D_{xy} y D_{yy}.

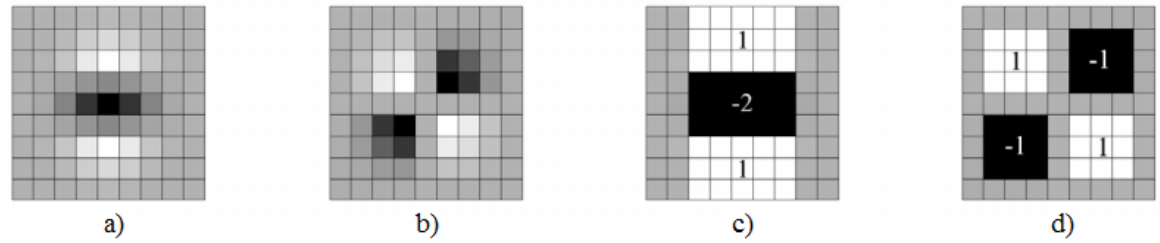


Figura 6. Filtro gaussiano de segundo orden de la derivada parcial en: a) dirección y y b) dirección xy. Aproximación del filtro gaussiano usando filtros de caja en: c) dirección y y d) dirección xy.

2. Se calcula el determinante de la matriz Hessiana, para diferentes tamaños de filtros, utilizando la siguiente ecuación.

$$\det\left(H_{aprox}\right) = D_{xx}D_{yy} - (0.9D_{xy})^2$$

   Los tamaños de los filtros van des 9x9 con una σ=1.2 que es considerada la escala inicial y van aumentando su dimensión sucesivamente por octavas. El incremento de los filtros dentro de una misma octava es el doble respecto del paso de la octava anterior, al mismo tiempo que el primero de los filtros de cada octava es el segundo de la octava predecesora. El número de octavas y el número de escalas por octava se determina experimentalmente dependiendo del tamaño de la imagen y de la aplicación.

3. Se localizan los puntos de interés en la imagen utilizando el principio de supresión no-máxima en un vecindario de 3x3x3. El principio de non-máxima supresión consiste en descartar todos los valores que no sean el máximo en una región, dejando como resultado solo el valor máximo.

4. Se calcula la respuesta Haar-wavelet en x y y en cada punto de interés usando nuevamente imágenes integrales. El área de cálculo es circular de radio 6s alrededor del punto de interés (s es la escala (equivalente a σ) de la aproximación del filtro gaussiano con el que el punto de interés fue detectado), así mismo el tamaño de los filtros Haar es de 4s.

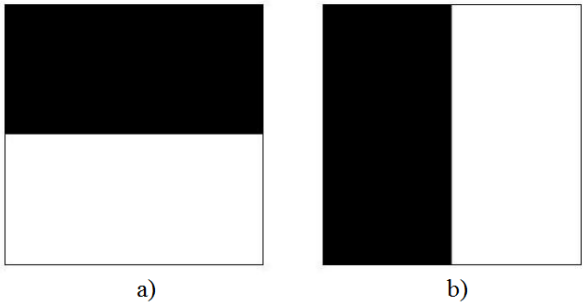


Figura 7. Filtros de Haar utilizados en el descriptor SURF, el color negro indica un valor de -1 y el color blanco un valor de 1. a) Filtro en x, b) filtro en y.

5. Una vez obtenidas las respuestas Haar, estas son ponderados utilizando una gaussiana con σ=2.5, centrada en el punto de interés.

6. Las respuestas son representadas como vectores en un espacio con la respuesta horizontal a lo largo del eje x y la respuesta vertical a lo largo del eje y.

7. La orientación dominante es estimada calculado la suma de todas las respuestas en una ventana corrediza cubriendo un ángulo de $\pi/3$. Las respuestas horizontales y verticales dentro de la ventana son sumadas. Las respuestas sumadas producen un nuevo vector. El vector más grande dentro de los 6 sectores en los que se divide el área asigna su orientación al punto de interés.

8. Se construye una región cuadrada de tamaño 20s alrededor del punto de interés orientada a lo largo de la orientación determinada en la el paso anterior. Esta región es a su vez dividida en 4 x 4 sub-regiones dentro de cada una de las cuales se calculan las respuestas de Haar de diferentes puntos con una separación de muestreo de 5 x 5 en ambas direcciones. Se consideran dx y dy las respuestas de Haar relativas a la orientación del punto de interés, en las direcciones horizontal y vertical respectivamente.

9. Las respuestas dx y dy son ponderadas por una gaussiana de valor σ=3.3s centrada en el punto de interés para obtener una mayor robustez ante deformaciones geométricas y errores de posición.

10. Cada una de la 4 sub-regiones quedan representadas por un vector definido como:

$$v = (\sum d_x , \sum d_y , \sum |d_x| , \sum |d_y|)$$

Finalmente el vector resultante engloba las 4x4 sub-regiones, lo cual nos da una longitud total de 64 elementos.

El descriptor SURF es utilizado principalmente para establecer correspondencias entre los puntos característicos en dos imágenes. La estrategia utilizada para establecer las correspondencias entre los puntos es la de "el vecino más próximo". El umbral relativo a la estrategia anterior es fijado con un valor de 0.7.

2.10 Resumen

En este capítulo se abordaron brevemente los conceptos básicos necesarios para comprender el desarrollo de esta tesis. Entre ellos se encuentran los conceptos de detección y seguimiento de personas en el ámbito de la visión artificial, se realiza una breve explicación de en qué consiste la calibración de cámaras, así como algunos de los métodos de calibración que hay. También se introduce el concepto de transformación geométrica aplicada a transformaciones en imágenes y los tipos de transformaciones que existen, además se explica el concepto de registro de imágenes junto con algunos de los tipos de métodos más utilizados. En el este capítulo se abarca el concepto de fusión de imágenes y los niveles de fusión más comunes, se explica el método HOG que se utilizó para la detección de personas, también se aborda brevemente el método de predicción filtro de Kalman y por último se da una introducción al descriptor SURF. Cada uno de estos métodos son utilizados a lo largo de la tesis.

En el próximo capítulo se comentan los trabajos más relacionados con el tema del seguimiento de personas por medio de visión artificial publicados en los

últimos años. Algunos de estos trabajos utilizan fusión de múltiples sensores, sin embargo, también se abordan algunos trabajos que utilizan un solo sensor y han sido sobresalientes en el área.

Capítulo 3

Trabajos relacionados

Existe una amplia bibliografía de trabajos sobre seguidores y detección de objetos por medio de imágenes, esto se debe a que es un tema que se ha venido abordando desde hace algunos años. Sin embargo, los diferentes retos que presenta dicha tarea aun lo hacen un problema sin una solución óptima. Los trabajos que podemos encontrar se pueden clasificar dependiendo de una gran variedad de características. En esta tesis los dividiremos principalmente en dos tipos: aquellos que usan fusión de sensores y los que utilizan un solo sensor.

3.1 Seguidores que no utilizan fusión de sensores

3.1.1 Aprendizaje progresivo para seguimiento visual robusto

Dentro del primer tipo de seguidores mencionaremos el propuesto por David Ross, Jongwoo Lim y Ming-Hsuan Yang en el año 2008 [2] , el cual es un trabajo ampliamente citado debido a los buenos resultados y la disponibilidad de los recursos necesarios para su réplica.

Dicho trabajo propone un algoritmo de seguimiento como un problema de inferencia probabilística similar al modelo de Markov y al filtro de Kalman

[17], en donde el tiempo se ve como una variable discreta que avanza un paso por cada *frame* de la secuencia de imágenes.

En cada delta de tiempo se observa una región de la imagen (denotada como F_t) y la localización del objetivo (denota como L_t). El movimiento de un objeto, de un *frame* a otro, se modela por una distribución que indique la probabilidad de que el objeto aparezca en una localidad L_t dado su antecedente en la localidad L_{t-1} y a su vez la probabilidad de que el objeto es localizado en L_t con la distribución F_t. Usando la regla de Bayes se obtiene la máxima probabilidad de la posición futura, la cual consta de cuatro parámetros: traslación en el eje *x*, traslación en el eje *y*, rotación y escalamiento. Es importante mencionar que así como se usa una distribución probabilística para modelar el proceso de movimiento, la apariencia del objetivo es modelada usando eigenbases.

Dado que el seguimiento visual es un proceso que utiliza datos que varían con el tiempo, es necesaria la continua actualización de las eiegenbases. Para calcular una nueva apariencia dada una eigenbase inicial, se utiliza el algoritmo secuencial Karhunen-Loeve, que se puede ver a más a detalle en la referencia [18].

El algoritmo en general se puede describir en los siguientes pasos [2]:

1. (Opcional)Construir una eigenbase inicial: de un conjunto de imágenes de entrenamiento del objeto u objetos similares, se crea una eigenbase inicial.
2. Elegir la localización inicial del objetivo: esto puede ser introducido manualmente o usando algún detector.
3. Buscar las localizaciones posibles del objeto: se determina un número de muestras de la distribución inicial sobre las posibles localizaciones. Para cada localización se obtiene una imagen de la región y se calculan las probabilidades de las localizaciones bajo la actual eigenbase.
4. Predecir la ubicación más probable del objetivo: se predice la ubicación futura del objetivo buscando la probabilidad máxima.
5. Cuando se decide que un número de imágenes han sido acumuladas, se actualizan incrementalmente las eigenbases, la media y el numero efectivo de observaciones.

6. Se regresa al paso tres.

El algoritmo fue probado en imágenes en el espectro visual en escala de grises, los videos fueron tomados en ambientes interiores sometidos a cambios de iluminación y cambios en la apariencia de los objetos, lo cual no afectó el desempeño ni los resultados del algoritmo. Los videos fueron capturados a 30 fps. Sin embargo, debido a que se pueden modificar algunos de sus parámetros el método se puede adaptar a una tasa diferente de fps [2]. En las pruebas realizadas por los autores del trabajo se reporta un promedio del error cuadrático medio de 5.07 pixeles por característica del individuo por *frame*. Algunas graficas presentadas también indican alrededor de un 1% de falsos positivos del objeto seguido. Cabe mencionar que en el desarrollo de esta tesis, este método fue sometido a secuencias de imágenes con condiciones más extremas y con diferentes modalidades de imágenes, con la finalidad de realizar una comparación justa entre ambos métodos. Dichos experimentos se reportan en el capítulo cinco.

3.1.2 Seguimiento utilizando métricas del tipo cross-bin

El siguiente trabajo fue desarrollado por Ido Leichter en el año 2012 [19], es uno de los trabajos más recientes en el tema. En este se hace una comparativa entre los tipos de métricas *cross-bin* y *bin-by-bin* al ser empleadas para la medición de distancias entre histogramas, mostrando que las métricas de la clase *cross-bin* son más adecuadas para el problema de seguimiento de personas en particular. Para ello se hace uso de las técnicas *Earth Mover's Distnace* (EMD), la cual utiliza una métrica del tipo *cross-bin*, y *Mean Shift* (MS), que utiliza una métrica del tipo *bin-by-bin*. Así como se propone una mezcla de ellas tanto para histogramas en una dimensión como para histogramas multidimensionales.

El algoritmo que se propone en este artículo puede resumirse en los siguientes pasos [19]:

1. Calcular el histograma del candidato actual y la distancia respecto al histograma de referencia del objetivo. Por candidato nos referimos a una porción de imagen que puede o no contener al objetivo.

2. Calcular el peso para ponderar los pixeles, dándole un mayor peso a los pixeles centrales.
3. Calcular una aproximación de la nueva localización del candidato.
4. Calcular el nuevo histograma del candidato y su distancia del histograma de referencia.
5. Si la distancia anterior es mayor a la distancia calculada actualmente, se recalcula la nueva posición promediando la posición actual con la posición del *frame* anterior.
6. Si la posición del *frame* anterior y la actual están en el mismo pixel o el número de iteraciones ha llegado al máximo establecido, el algoritmo se detiene. En otro caso, la posición del *frame* anterior se actualiza con la posición actual, al igual que la distancia calculada del *frame* anterior también se actualiza y posteriormente se regresa al paso dos.

Las diferencias fundamental que existen en aplicar los métodos *Mean Shift* (MS*), Mean Shift whit EMD* para una dimensión (MS-EMD) y *Mean Shift whit EMD* para histogramas multidimensionales (MS-EMD multidimensional), radica en la forma de calcular la ponderación de los pixeles y la distancia entre los histogramas. Mientras que en el método MS se utiliza un histograma representado por un vector unidimensional para el cálculo de la distancia y los pesos, en el método MS-EMD se utiliza un histograma acumulativo denotado por una matriz correspondiente a la acumulación de los histogramas de los diferentes recuadros. Dicho procedimiento da por resultado una matriz acumulativa que contiene las ponderaciones de los pixeles y modificaciones en el cálculo de la distancia entre histogramas [19].

Por su parte el método MS-EMD multidimensional propone un espacio de características multidimensional para medir la distancia entre los histogramas, es decir dado las matrices multidimensionales que representan los histogramas tanto del objetivo como del candidato, las ecuaciones de cálculo de distancia y de ponderación de pixeles del algoritmo original MS son acopladas al espacio multidimensional [19].

Las pruebas se realizaron con imágenes en ambientes exteriores en el espectro visual a color y en escala de grises, como resultado para ambos tipos de imágenes se obtuvo un rendimiento muy similar en la tasa de seguimiento entre

los métodos MS-EMD y MS-EMD multidimensional, mientras que el método MS original permanece ligeramente por debajo. Es importante destacar que todos los métodos tiene un costo computacional similar, determinado por O(mF), donde F es la dimensionalidad del histograma y m el número de recuadros que se propondrán como candidatos para cada búsqueda. Para MS y MS-EMD F=1 [19]. El metodo MS-EMD obtuvo un promedio de seguimiento correcto de 96.5%, mientras que el método MS original obtuvo un promedio de 73.7%.

3.2 Seguidores que utilizan fusión de sensores

El siguiente grupo de trabajos utilizan la fusión de múltiples sensores para resolver el problema de seguimiento de objetos, a continuación se mencionan alguno de los más relacionados con esta tesis.

3.2.1 Modelo gráfico probabilístico utilizado para seguimiento de personas

Este trabajo fue realizado por Siyue Chen, Wenjie Zhu y Henry Leung en el año 2008 [20]. Presenta una aproximación de un modelo gráfico probabilístico que fusiona imágenes térmicas con imágenes en el espectro visual para realizar seguimiento de personas. Cada *frame* es representado como un solo vector denominado $z_{l,n}$ donde n es el número de *frame* y l el número de pixel de la imagen original v_l y $l = (lx, ly)$.

Este trabajo comienza recolectando los componentes de apariencia de las imágenes en el espectro visual, que son segmentos de imágenes modelados con una distribución gaussiana denotada por s_1, después se calcula la probabilidad condicional del modelo s_1 dada la imagen v_1 correspondiente al objeto a seguir. La imagen se representa con un vector de intensidades de pixel. Por otra parte se realiza un corrimiento en l del cual también es calculada su probabilidad condicional [20].

Asumiendo que las imágenes han sido previamente registradas, se realiza el mismo procedimiento para las imágenes térmicas obteniendo las variables: z_2, s_2 y v_2. Como se puede notar las variables provenientes del espectro

térmico son distinguidas con un subíndice 2 mientras que las variables del espectro visual presentan un subíndice 1 [20].

Posteriormente los dos modelos se fusionan calculando la probabilidad conjunta de las variables implicadas (z_1, s_1, v_1, z_2, s_2 y v_2). Por último se estima la posición l en cada *frame* según el argumento máximo de las probabilidades en el espectro visual y en el espectro térmico. El trabajo propone un algoritmo recursivo en donde en cada iteración se actualizan las distribuciones de las variables y los parámetros estimados (ruido, media de las distribución gaussiana, matriz de precisión y probabilidad a priori) [20].

Se obtuvieron resultados de una tasa de seguimiento exitoso del 47.3% para el espectro visual, 89.6% para el espectro térmico y 93.2% para la fusión de los distintos tipos de imágenes. Las imágenes con las que fue probado fueron tomadas a 25 fps, en ambientes exteriores, con oclusión del objetivo y con objetos similares al objetivo solo en el espectro visual. Su principal desventaja es que es un seguidor de un solo objetivo y no fueron reportadas pruebas en ambientes con objetos similares en espectro térmico [20].

3.2.2 Seguimiento de personas usando fusión de imágenes térmicas, visuales y mapas de profundidad.

Este artículo fue realizado por Stephen J. Krotosky y Mohan Manubhai Trivedi en el año 2008 [12]. Propone una metodología que analiza sistemas de imágenes multimodales y multiperspectiva para el monitoreo de personas en sistemas de vigilancia. Para ello se realizó una cama de pruebas de imágenes tomadas con dos cámaras a color en el espectro visual y dos cámaras en el espectro infrarrojo.

En este trabajo se diseñó un algoritmo para la detección de personas en una escena que puede ser generalizado para incluir imágenes a color, infrarrojas o mapas de profundidad. Para ello se hace uso del método histograma del gradiente orientado (HOG) [1] y un clasificador de vectores de máquinas de soporte (SVM) para detectar la presencia de personas en las imágenes visuales e infrarrojas. Adicionalmente, se realizó una relación entre el tamaño de las personas y la profundidad de estas en la escena, para crear un detector basado en disparidad.

En una primera aproximación, el método HOG fue aplicado también a las imágenes de disparidad, sin embargo, se observó que fuera del conjunto de entrenamiento, donde los recuadros que enmarcan a las personas son recortados cuidadosamente, el método daba una cantidad de falsos positivos incluso superior a la cantidad de personas en la escena. Para encontrar una alternativa, se examinaron las imágenes de disparidad buscando características que ayudaran a diferenciar personas del resto de objetos en la imagen. Usando el conocimiento de que el tamaño de la persona está normalmente distribuido alrededor de un valor medio, se modeló la correlación lineal entre el tamaño del recuadro que enmarca a la persona y la media de la región de disparidad [12].

Por lo tanto se tienen dos clasificadores, uno para las imágenes visuales e infrarrojas y otro para los mapas de disparidad, ambos clasificadores se combinan probabilísticamente para determinar el clasificador final. El umbral de detección para determinar la presencia de una persona se fijó en un 90% [12].

Por otro lado se analizaron detectores basados en tetra-visión, el cual combina dos imágenes estéreo una proveniente de un par de imágenes visuales y otra proveniente de un par de imágenes en el espectro infrarrojo, cada una entrenada por separado. También se incorporó una aproximación trifocal que combina características de las imágenes a color e infrarrojas en un marco unificado de detección de personas. Se trabajó con dos tipos de aproximaciones: la primera incluye imágenes estéreo provenientes de las imágenes a color más imágenes infrarrojas (denominada como CSI), la segunda incluye imágenes estéreo provenientes de las imágenes infrarrojas más imágenes en el espectro visual (denominada ISC) [12].

Las secuencias de imágenes fueron evaluadas agrupándolas en: imágenes estéreo provenientes del espectro visual, aproximación trifocal CSI, imágenes estéreo provenientes del espectro IR y aproximación trifocal ISC. También se probó una aproximación usando tetra-visión, analizando por separado los mapas de profundidad provenientes de cada uno de los dos espectros utilizados y combinando sus detecciones para determinar un resultado general. Se utilizaron los operadores OR y AND para combinar dichos resultados.

Para cada persona en la escena se grafica la coordenada de su detección en cada *frame*, utilizando un color diferente por persona y con círculos amarillos se indican las detecciones perdidas. Esto muestra que las detecciones perdidas no afectan al seguimiento general de la persona, dado que la tasa de detección es mucho más alta, ver figura 8. La clasificación de que detección pertenece a que persona depende de la cercanía al resultado de un predictor de trayectoria [12].

Figura 8. Resultados del seguidor propuesto por Krotosky y Trivedi [12]

Los experimentos fueron hechos en ambientes exteriores con secuencias de 30 fps y estos demostraron un mejor rendimiento cuando se combinan las características de las imágenes multiespectrales y multimodales. Las aproximaciones trifocales aumentaron la tasa de detección correcta en un 45%, el desempeño reportado por la aproximación CSI es de un 85.9% de personas detectadas correctamente y un promedio de 0.86 falsos positivos por *frame*, mientras que la aproximación ISC es ligeramente inferior con un 80.2% de personas detectadas correctamente y un promedio de 0.69 falsos positivos por *frame*. La evaluación utilizando tetra-visión dio un rendimiento inferior a los de CSI e ISC[12].

Este método es un seguidor multiobjetivo como el que se propone en esta tesis, es decir que sigue a varios sujetos a la vez, también es resistente a cambios de iluminación. Sin embargo, al igual que varios de los trabajos anteriormente

citados asume un cambio en la posición y localización de la persona poco significativo debido a la velocidad de la captura de los *frames* por segundo [12].

3.2.3 Detector de objetos basado en fusión de múltiples sensores y seguimiento usando un modelo de forma activa.

Este trabajo propone un detector automático y seguidor de objetivos usando un modelo de forma activa (ASM, por sus siglas en inglés). Fue realizado en el año 2011 por Dongeun Lee y Sunghoon Choi [21]. El algoritmo propuesto consiste en las siguientes partes:

1. Evaluación de la imagen de entrada para la obtención de las condiciones en las que se coloca la cámara.
2. Un detector y seguidor de movimiento integrados.
3. Un seguidor de forma activa (AST), basado precisamente en el modelo de forma activa (ASM)
4. Actualización de los resultados del seguidor.

El algoritmo propuesto es considerado como un modelo físico que permite al sistema predecir con exactitud los cambios potenciales en la forma del objeto durante el paso del tiempo. El procedimiento de detección extrae objetos en movimiento mediante segmentación de movimiento entre *frames*. Las regiones son detectadas usando un detector basado en flujo óptico. La región detectada es entonces realzada usando una localización elipsoidal inicial para actualizar los conjuntos de entrenamiento. La actualización de los conjuntos de entrenamiento en la región elipsoidal es llevada a cabo usando el método SSA [22].

El modelo de la estructura local es calculado usando operadores de regularización de dirección, los cuales reflejan la orientación de los bordes y encuentra la dirección dominante del movimiento, por lo tanto, los objetos pueden ser seguidos sin ninguna información referente a la posición de la cámara. En la figura 9 se muestra un diagrama general del funcionamiento del algoritmo de seguimiento [21]:

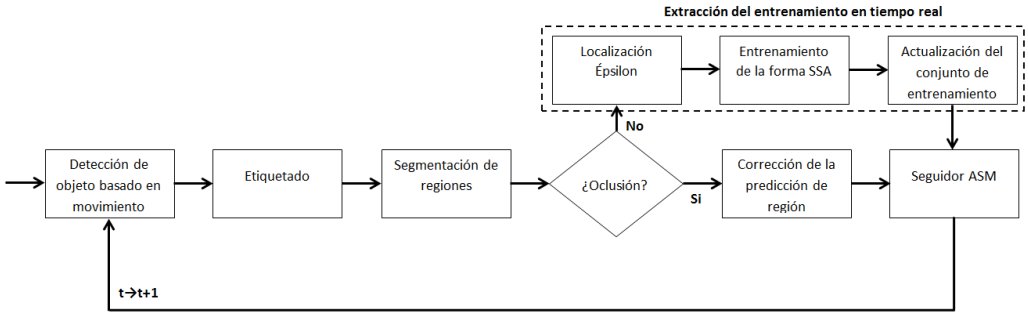


Figura 9. Diagrama del algoritmo de seguimiento de [21]

En el primer bloque podemos ver la etapa de detección del objeto, que como anteriormente se mencionó se basa en flujo óptico. La condición fundamental del flujo óptico es que la intensidad de un pixel en un objeto no cambie durante un periodo de tiempo, si hay un cambio se asume que hay un movimiento, sin embargo, puede existir ruido o cambios de iluminación que producen un cambio en la intensidad del pixel pero que no precisamente indican un movimiento. Para reducir este efecto erróneo se utilizó el método de Lucas-Kanade [23] basado en un método de cambio por bloque. Después, se extrae el movimiento de la secuencia de imágenes y se segmentan las regiones basadas en la dirección de movimiento.

Dado un *frame* de entrada, una serie de puntos deben ser colocados en el contorno del objeto, el papel de estos puntos es controlar el contorno de la forma del modelo. Como primer paso se construye una elipse en el área alrededor del centroide de la región, de la cual la forma del modelo entrenador es obtenida usando SSA [22]. Este modelo de entrenamiento debe ser constantemente actualizado dependiendo del cambio en el objeto de interés.

Una vez construido el modelo, la tarea del seguidor ASM es encontrar el conjunto de parámetros que mejor coinciden con el modelo de la imagen. Los parámetros de transformación juegan un papel importante en la inicialización del seguidor basado en ASM. La estimación inicial de la escala y los parámetros de traslación pueden ser obtenidos del módulo de la localización elipsoidal. La inicialización de la forma se refiere al proceso de estimación de la posición y el tamaño de la forma del contorno del objeto en la imagen. Una vez

que el seguidor ASM es inicializado, este usa sus propias predicciones y estados que se van almacenando en una base de datos [21].

El algoritmo fue probado utilizando una fusión de dos tipos de imágenes: térmicas y visuales. Estas fueron fusionadas a nivel señal utilizando una transformada *wavelet*. Los experimentos probados en este artículo demostraron que el algoritmo propuesto puede seguir un objeto deformable bajo condiciones de oclusión y ambientes con iluminación degradable o con poca iluminación [21].

3.2.4 Marco integrado de registro de imágenes, fusión de sensores y seguimiento de personas utilizando imágenes térmicas y visuales

Este trabajo es uno de los más recientes y completos en el área de seguimiento de personas utilizando fusión de imágenes térmicas y visuales. Fue propuesto en el año 2012 por Atousa Torabi, Guillaume Massé y Guillaume-Alexandre Bilodeau [24]. Propone un marco integrado que abarca el problema del registro termo-visual, la fusión de sensores y el seguimiento de personas en videos.

Los datos de entrada del sistema son el par de *frames* correspondientes a las imágenes visuales y térmicas. Posteriormente para separar los pixeles en primer plano del resto, se aplica un método de sustracción del fondo de la imagen propuesto por [25]. En la figura 9 se muestra el diagrama de flujo del algoritmo en general, el cual consiste principalmente en dos etapas: la inicialización y el ciclo principal para el registro de las imágenes, y la fusión de los sensores y el seguimiento.

La inicialización se realiza al principio de los videos donde para los primeros *frames*, el seguimiento es realizado de forma separada entre los videos térmicos y los videos en el espectro visual, hasta obtener suficientes puntos en la trayectoria para estimar una buena matriz de transformación. La segunda parte del algoritmo consiste en un ciclo en pares de *frames* visibles y térmicos, donde el registro de la imagen, la fusión de sensores y el seguimiento térmico-visual son realizados uno respecto del otro [24].

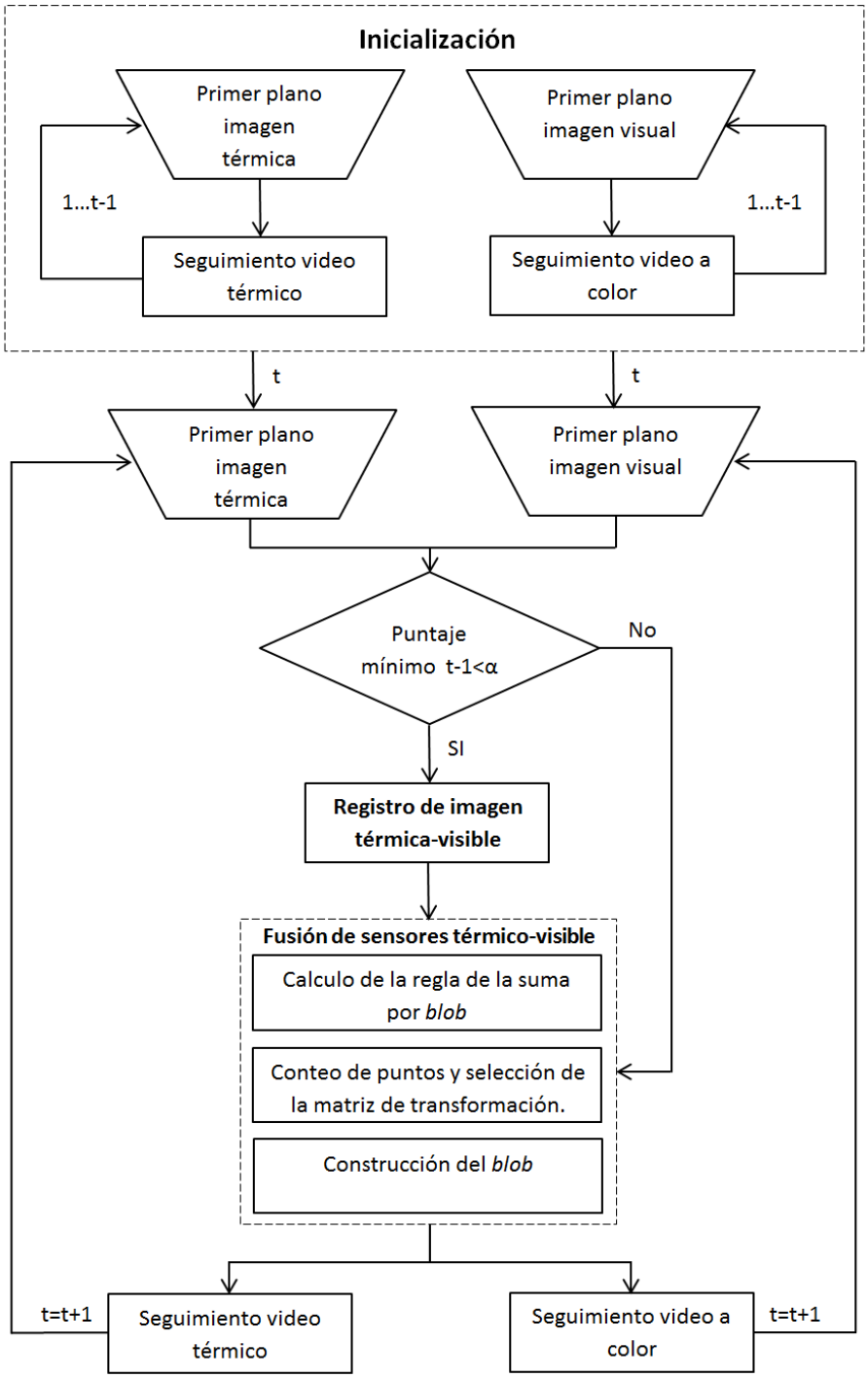


Figura 10. Diagrama del algoritmo de [24]

El registro de la imagen, estima una matriz de transformación afín (ver sección 2.3) que es usada para transformar una imagen a las coordenadas de la segunda imagen. La fusión de los sensores empalma los pixeles de los *blobs* (un *blob* es definido como todos los pixeles que pertenecen a un objeto en la imagen) de la imagen térmica y la imagen a color usando una matriz de transformación y combinando la información de ambas imágenes multimodales. En este paso, la calidad de correspondencia de los *blobs* calculados es evaluada para decidir cuándo una nueva matriz de transformación debe ser estimada o si se debe continuar con el siguiente *frame* [24].

Finalmente, el seguimiento es realizado para los videos térmicos y visibles usando una fusión de *blobs* obtenidas de la fusión de sensores. Esos nuevos puntos de la trayectoria son usados para calcular el registro de la imagen del siguiente *frame*.

El registro de la imagen es realizado alineando la imagen térmica y la imagen color usando una transformación afín, calculada mediante la correspondencia de las trayectorias de los objetos en ambas imágenes. Los puntos de las trayectorias son alineados usando el método RANSAC, el cual está basado en seleccionar aleatoriamente puntos en la trayectoria del objeto de las imágenes visuales y térmicas sincronizadas y encontrando la mejor correspondencia entre los puntos. La trasformación afín, es estimada usando la transformación lineal directa normalizada (DLT) para encontrar las solución por mínimos cuadrados [26].

La fusión de los sensores se puede resumir en los siguientes pasos [24]:

1. Cálculo de la regla de la suma para las siluetas utilizando la matriz Mn: la matriz Mn representa la matriz de transformación estimada para el registro de la imagen. A su vez una silueta es definida como una región binaria y la regla de la suma para las siluetas es una suma de probabilidades de los pixeles en primer plano en la imagen térmica y la imagen visual. En otras palabras, utilizando la matriz Mn se realiza el registro de las imágenes unificando sus sistemas de coordenadas en uno solo. Posteriormente, se obtienen las probabilidades separadas en cada imagen de los objetos que están en primer plano, si la suma de las

probabilidades sobrepasa un umbral entonces se dice que ese pixel pertenece a un espacio denominado como Sn.

2. Cálculo de la regla de la suma para las siluetas utilizando la matriz Mb: donde la matriz Mb es la mejor matriz estimada para el *frame* actual. La regla de la suma se aplica del mismo modo que en el paso anterior sólo que esta vez se utiliza la imagen térmica y la matriz Mb y el espacio Sb.
3. Cálculo del contador por *blob*: se realiza una suma de los pixeles que están en la intersección entre el espacio Sn y un *blob* determinado según la imagen térmica. Se realiza el mismo cálculo para la imagen visual y se repite este paso para el espacio Sb.
4. Cálculo global del contador calculado con la matriz Mn: se suman y normalizan los resultados provenientes de los cálculos que realizaron utilizando la matriz Mn, de cada uno de los *blobs* en la imagen, obteniendo un contador global n.
5. Cálculo global del contador calculado con la matriz Mb: se realiza los mismo que en el paso cuatro pero utilizando los resultados provenientes de los cálculos que utilizaron la matriz Mb, obteniendo el contador global b.
6. Selección de la matriz de transformación: si el contador global n es mayor que el contador global b entonces Mn remplaza a la matriz Mb.
7. Cálculo del modelo del objeto: la fusión de los sensores provee información de los pixeles pertenecientes a los *blobs* según ambas modalidades de imágenes. Estos servirán como características para resolver el problema del seguimiento. El modelo del objeto es un histograma térmico-visual de los *blobs*.

Para el seguimiento, cualquier método que calcule y actualice la trayectoria del objeto seguido *frame* por *frame* es aplicable. En este trabajo utilizaron el método de múltiples hipótesis de seguimiento (MHT) el cual se detalla en [27]. El seguimiento es calculado por separado para las imágenes visuales y térmicas usando los *blobs* construidos con los histogramas térmicos-visuales como característica.

El algoritmo de seguimiento utiliza dos grafos: un grafo de eventos para registrar los eventos y almacenar la información de apariencia mientras un objeto es seguido, y otro grafo de hipótesis para generar una hipótesis para la

asociación del manejo de datos de los objetos por separado. El algoritmo de seguimiento se puede resumir en tres pasos [24]:

1. Correspondencia de los *blobs*: El primer paso es calcular una matriz de distancia para encontrar la correspondencia entre los *blobs* del *frame* anterior y los *blobs* del *frame* actual, utilizando las similitudes de la apariencia del objeto en dos *frames* consecutivos
2. Actualización de los grafos: ambos grafos son actualizados dependiendo de los eventos detectados en el proceso de seguimiento, aumentando vértices y nodos.
3. Etiquetado de los objetos y cálculo de la trayectoria: tanto el etiquetado como el cálculo de la trayectoria son calculados procesando el gráfico de la hipótesis, tomando en cuenta para cada hipótesis la cercanía de cada *blob* con el objeto.

Las pruebas fueron hechas con videos capturados en interiores a 30fps. El número de fps es importante para asumir un grado de cambio limitado en posición y apariencia por cada *frame*, en lo cual se basa el método. Los resultados demostraron que la fusión de sensores mejora el seguimiento de objetos respecto a aplicar el mismo método solo en una modalidad de imagen (visual o infrarroja), también supera a los métodos [28] y [7] reduciendo el número de falsos positivos a solo un 1.2% y falsos negativos a un 0.2%.

3.3 Resumen

En este capítulo se explican brevemente los trabajos más sobresalientes y relacionados con el tema del seguimiento de personas utilizando visión artificial. Dichos trabajos se dividen de dos secciones: aquellos que utilizan fusión de múltiples sensores y los trabajos que utilizan un solo sensor.

Dentro de los trabajos que utilizan un solo sensor citamos dos: el primero es propuesto por D. Ross, J. Lim y M. Yang en el año 2008 [2]. Dicho trabajo utiliza eigenbases para modelar la apariencia del objeto y un modelo adaptativo probabilístico para el seguidor de personas.

El segundo trabajo fue publicado en el 2012 por I. Leichter [19]. Hace una comparativa entre las métricas del tipo *cross-bin* y *bin-by-bin* aplicado al

seguimiento de personas. Propone un método que mezcla lo mejor de ambos tipos de métricas para medir la distancia entre histogramas, obteniendo mejores resultados en su método propuesto que en los métodos por separado.

El segundo grupo consiste en los seguidores que utilizan fusión de imágenes y entre ellos encontramos los trabajos de:

- S. Chen, W. Zhu y H. Leung (2008) [20]: Propone un modelo gráfico probabilístico para seguimiento de personas en ambientes exteriores con oclusión. Logra una tasa de seguimiento exitoso del 93.2% utilizando fusión de imágenes térmicas y visuales.
- S J. Krotosky y M. M. Trivedi (2008) [12]: Compara tres modalidades de fusión: aproximación CSI (mapa de profundidad visual + imagen IR), aproximación ISC (mapa de profundidad IR + imagen visual) y tetra-visión (mapa de profundidad visual + mapa de profundidad IR).
- D. Lee y S. Choi (2011) [21]: Utiliza un modelo de forma activa (ASM) para implementar un seguidor en imágenes térmicas y visuales. Las imágenes se fusionan a nivel señal usando un transformada *wavelet.*
- Torabi, G. Massé y G. Bilodeau (2012)[24]: Propone un marco que integra el registro de imágenes, la fusión de sensores y el seguimiento de personas para imágenes visuales y térmicas.

En el próximo capítulo se explica a detalle el método que se propone en esta tesis y cada uno de los pasos que cada etapa conlleva.

Capítulo 4

Método propuesto

En este capítulo se explican los pasos que se siguieron y los algoritmos que se utilizaron para realizar la implementación del método propuesto en esta tesis. Este capítulo está dividido en las siguientes secciones: calibración de las cámaras y captura de imágenes, registro de las imágenes, detección de las personas, caracterización de personas, seguimiento y resumen. Dichas secciones, exceptuando el resumen, corresponden a cada una de las etapas de la implementación del método.

4.1 Calibración de las cámaras y captura de imágenes

La calibración de ambos tipos de cámara se realizó utilizando la herramienta de MATLAB "*camera calibration toolbox"*, esta utiliza una recopilación de dos métodos: el modelado intrínseco de la cámara está basado en [29], mientras que el resto de la calibración está inspirada en [6].

Este método fue seleccionado debido a que se puede utilizar tanto para la calibración de cámaras visuales, como para la calibración de cámaras en el espectro infrarrojo o cualquier otra modalidad, siempre y cuando el patrón de calibración sea visible en la escena, además de que la implementación del

método ha sido probada exitosamente en [12], [30] y [31], por mencionar solo algunos trabajos que lo utilizan y por último la implementación se encuentra disponible en [32], lo cual ayudo a reducir el tiempo de implementación del pre-procesamiento necesario para llegar al punto principal de esta tesis que es el seguimiento de personas.

El primer paso para realizar la calibración de una cámara, es tomar una serie de imágenes de un patrón, para el método utilizado en el *toolbox* de MATLAB se recomienda un patrón cuadriculado tipo ajedrez, como se observa en la figura 11.

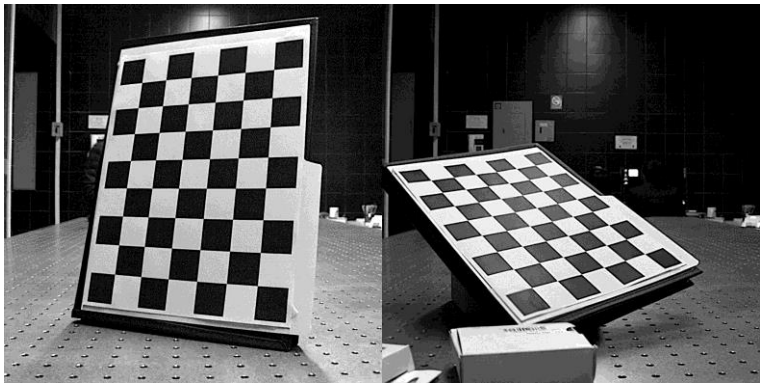


Figura 11. Patrón de calibración de la cámara visual.

Las medidas del patrón son introducidas al programa y este detecta automáticamente las esquinas del patrón en la imagen, para posteriormente calcular el modelo y los parámetros de la cámara.

Una vez calculados dichos parámetros, se le da al programa la orden de calcular la transformada correctiva de las imágenes y guardarla como una matriz. Posteriormente, ya que se han tomado las secuencias de imágenes con las que se desean probar el método propuesto, la matriz de la transformada se aplica a todas las imágenes tomadas con la cámara propia de la calibración, para disminuir la deformación que ejerce sobre sus imágenes.

4.1.1 Calibración cámara térmica

Para la calibración de la cámara térmica los pasos seguidos son exactamente los mismos que con cualquier otra cámara. Sin embargo, para hacer que el patrón

tipo ajedrez pueda distinguirse en el espectro térmico, es necesario iluminarlo con una lámpara o reflector de bulbos de halógeno de alta potencia (50 watts mínimo), la imagen del patrón obtenida se observa en la figura 12.

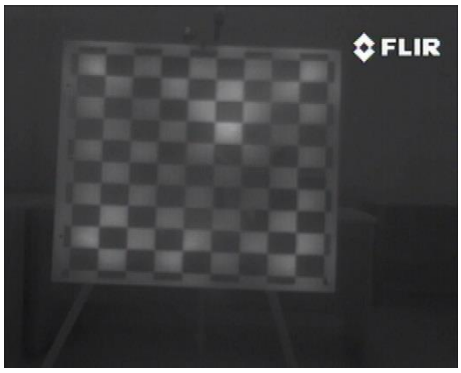


Figura 12. Patrón de calibración en la cámara térmica

4.2 Registro de las imágenes

El registro de imágenes consta principalmente de cuatro pasos: detección de características, correspondencia de características, inferencia de transformación geométrica y aplicación de la transformación (ver capítulo 2). A continuación se explica cómo se realizaron estos pasos en esta tesis.

La detección y correspondencia de características se realizó de forma manual, buscando puntos correspondientes en ambas imágenes y guardando sus coordenadas en variables. Por imagen se capturaron cuatro puntos representativos que estuvieran separados entre sí por una distancia mínima de 160 pixeles (determinada experimentalmente), dado un tamaño de imagen de 640x480 pixeles.

Se decidió usar un método manual debido a la falta de robustez, para imágenes multimodales, de la mayoría de los métodos de correspondencia de puntos y a la complejidad de los pocos métodos que reportan un rendimiento un poco mejor para el registro multimodal. Esta restricción se decidió con base en las limitaciones de tiempo y dado que esta etapa es un pre-procesamiento de las imágenes y no implica el método de detección automática, ni de seguimiento. Es decir, el objetivo de esta tesis consiste en desarrollar un método de detección

y seguimiento de personas, por lo tanto los pasos previos a esta etapa son necesarios para llegar a desarrollar el objetivo pero no son parte de la aportación del método en sí. Además de que a lo largo del pre-procesamiento de la imagen, necesario para el seguimiento de personas u otras aplicaciones, aún existen muchos problemas que no han sido resueltos en su totalidad, los cuales de forma particular pueden llegar a ser tan extensos como lo es en sí mismo la detección y el seguimiento. Lo que se propone es una aportación al problema de seguimiento en específico.

La inferencia de la transformación geométrica se realizó por medio de la función *maketform()* de MATLAB, la cual utiliza los cuatro puntos extraídos de la etapa anterior y los introduce en un sistema de ecuaciones, donde se conocen las coordenadas [x y] de la imagen térmica y las coordenadas [u v] de la imagen visible correspondientes a cada punto, pero se desconoce la matriz de transformación **A**. **A** se obtiene resolviendo el siguiente sistema de ecuaciones.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \boldsymbol{A} * \begin{bmatrix} u \\ v \end{bmatrix} \qquad \text{ec.21}$$

Finalmente la transformada **A** se aplica a todos los pixeles de la imagen visible, utilizando la función *tformarray()* de MATLAB, multiplicando **A** por la posición de los pixeles. De esta forma tenemos ambos tipos de imágenes en un mismo sistema de coordenadas.

4.3 Detección de personas

Para la detección de personas se decidió utilizar el método de detección HOG, publicado originalmente en [1] y explicado brevemente en el capítulo 2. Se eligió dicho método debido a dos razones principales, la primera de ellas consiste en que es un método ampliamente probado en varios artículos, constatando los buenos resultados reportados por los autores originales del método, lo que a la fecha lo hace uno de los métodos más utilizados en esta área. Por mencionar algunos trabajos que reportan resultados sobre el método HOG tenemos [33], [34] y [35]. El segundo motivo de su selección, pero no menos importante, cosiste en que según los resultados descritos en [12] y [36] es un método que puede aplicarse a diferentes modalidades de imágenes, en

específico a visuales y térmicas que son precisamente el tipo de imágenes con las que se trabaja en esta tesis, haciendo que no sea necesario implementar dos métodos radicalmente diferentes para la detección de personas, dado que ambas modalidades comparten características en común con información que se complementa entre sí.

Para la implementación del detector de personas el primer paso que se tuvo que ejecutar fue la recopilación y creación de las bases de datos para entrenar al clasificador SVM, para lo cual, debido a la gran cantidad de muestras requeridas, fue necesario tomar algunas muestras propias y conseguir algunas imágenes tanto térmicas como visibles de otras bases de datos, como son las usadas en [37], [38], [39] y [40].

La proporción de muestras positivas (con persona) y negativas (sin persona) se mantuvieron en 17% y 83% respectivamente, de un total de 14658 muestras para cada tipo de imagen. Se mantuvieron las mismas cantidades y proporciones de muestras que las reportadas en [1] con el objetivo de obtener una taza de clasificación correcta lo más similar posible.

Las ventanas de entrenamiento son de 64x128 pixeles y en la implementación del algoritmo para la extracción de los descriptores HOG, se utilizaron células y bloques rectangulares de 8x8 pixeles y 4x4 células. Los tamaños de las ventanas, células y bloques fueron calculados haciendo un balance entre la longitud del descriptor y la mínima tasa de pérdida, obteniendo dichos valores como los más óptimos. En la figura 12 se presenta la gráfica de la relación del tamaño del descriptor con el tamaño de los bloques y células, se tomaron únicamente en consideración los tamaños de bloques y células que presentan una menor tasa de perdida según lo reportado por los autores.

Posteriormente una vez extraídos los descriptores HOG de las imágenes de la base de datos, se procedió a entrenar con estos un clasificador SVM utilizando el comando de MATLAB *'svmtrain()'*, el cual es provisto de dos argumentos de entrada principales: un arreglo de mxn elementos que contiene cada uno de los descriptores obtenidos de la base de datos, donde m es el tamaño de cada uno de los descriptores HOG y n es el tamaño de la muestra de entrenamiento y un arreglo que indica la clase a la que pertenece cada elemento.

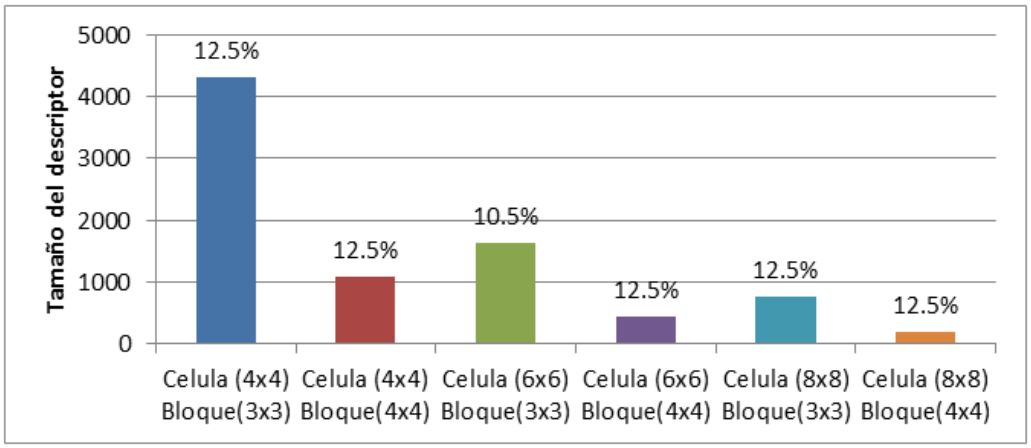


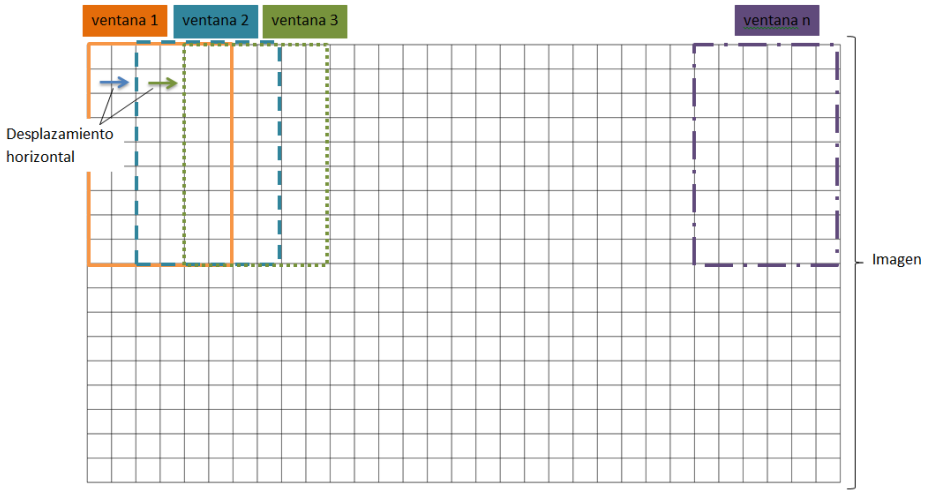
Figura 13. Grafica de la relación entre el tamaño del descriptor, la tasa de pérdida y el tamaño de los bloques y las células. (La tasa de pérdida está indicada en porcentaje en la parte superior de cada barra)

Para la detección de personas en las secuencias de imágenes de prueba, se deslizan varias ventanas de diferentes tamaños a lo largo y ancho de la imagen, comenzando de la esquina superior derecha y terminando en la esquina inferior izquierda, como se puede observar en la figura 14.

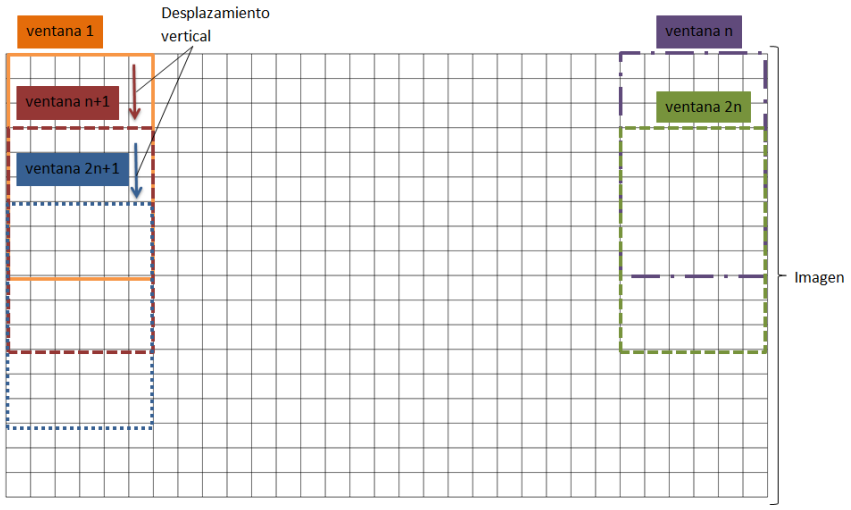
De cada ventana se extrae el descriptor HOG y utilizando el clasificador SVM previamente entrenado, se clasifica en positiva o negativa dependiendo si hay o no una persona dentro de esa región de la imagen, para ello se hizo uso de la función '*svmclassify()'*de MATLAB. El avance de la ventana tanto en x es de 1/3 del ancho de la ventana y el avance en y es 1/3 del largo. El valor del avance fue obtenido experimentalmente tomando en cuenta el número de ventanas a procesar y el evitar la pérdida o el mal encuadre de alguna persona.

Por su parte los tamaños de las ventanas van de 130x270 a 240x450 pixeles, en intervalos de aumento de 55x90 pixeles para imágenes de 640x480 pixeles. Estos valores fueron determinados tomando en cuenta los tamaños máximos y mínimos de las personas en las imágenes, lo cual se relaciona directamente con la distancia mínima y máxima que puede haber entre una persona y la cámara para que este dentro de los límites de seguimiento (de 2m a 6m), y a su vez

también se tomó en cuenta un equilibrio entre un buen ajuste del encuadre de la persona en la ventana y evitar redundancias.



a)



b)

Figura 14. Representación del desplazamiento a) horizontal y b) vertical de las ventanas de detección sobre cada imagen.

Cada ventana extraída de las imágenes se reduce a un tamaño de 64x128 pixeles para ser clasificada. La detección de personas se realiza por separado en cada tipo de imágenes, obteniendo un resultado para las imágenes térmicas y otro para las imágenes visuales.

Posteriormente, se fusionan ambos métodos utilizando una compuerta lógica del tipo OR, obteniendo de este modo un resultado global y final del detector de personas.

4.4 Caracterización de personas

Con el objetivo de poder identificar individualmente a cada persona, se extraen una serie de características de cada una de las personas detectadas en la imagen, creando con dichas características un modelo representado por un arreglo de valores.

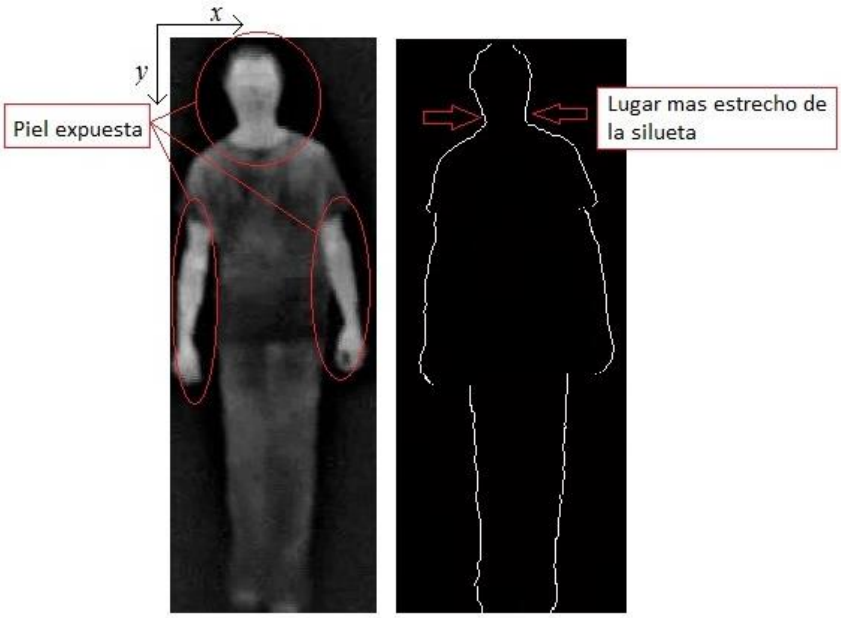


Figura 15. Ilustración de las características usadas para segmentar la cabeza de una persona.

El cuerpo de cada persona es segmentado en dos partes: la cabeza y el cuerpo. Esto se realiza utilizando la diferencia de intensidad en los pixeles que se

aprecia en las imágenes térmicas, ya que en general la temperatura reflejada de la piel expuesta es superior a la de la piel cubierta por ropa. También se hace uso de la forma de la silueta para detectar el punto más estrecho, que corresponde precisamente al cuello (ver figura 15). El seudocódigo del algoritmo elaborado para encontrar la fila que separa la cabeza del cuello se presenta a continuación:

ALGORITMO: SEGMENTACION DE LA CABEZA Y CUERPO

1. Se umbraliza la imagen IR.
   Imagen_1=Imagen>Umbral_1
   Dónde: Umbral_1=maximo- ((maximo-media)*0.30)
2. Se eligen las tres regiones con mayor área y se filtra el resto.
3. Se identifica la región más cercana a la parte superior de la imagen.
4. IF: el largo de la región superior en el eje y, es mayor a N pixeles y es menor a M pixeles (Dónde: N=altura del recuadro/9 y M=altura del recuadro/5).
   THEN: Se gurda el límite inferior de la región en la variable valor_1.
5. ELSE: se busca la siguiente región y se repite el paso 4.
6. IF: ninguna región cumple con la condición: M>largo de la región >N
   THEN: se busca el largo de la región más cercano a L y la posición del límite inferior de dicha región se guarda en la variable valor_1
   (L= altura del recuadro /7).

7. Se umbraliza la imagen IR.
   Imagen_2=Imagen>Umbral_2
   Dónde: Umbral_2=150 (parámetro experimental)
8. Se obtienen los bordes utilizando el método de Sobel.
9. Se esqueletiza el borde y se eliminan segmentos aislados.
10. Se recorre fila por fila buscando la distancia mínima entre el pixel de extrema izquierda y el pixel de extrema derecha. La coordenada *y* de la distancia mínima se guarda en la variable valor_2.

11. Se compara valor_1 con valor_2.
    Δvalor= |valor_1-valor_2|
    IF: Δvalor< altura del recuadro/5
    THEN: El resultado es el promedio de valor_1 y valor_2
12. ELSEIF: Se toma al valor más cercano a (altura del recuadro/7) y se guarda en val
    IF: val >altura del recuadro/3
    THEN: El resultado es igual al valor por default:
    altura del recuadro/7
    ELSE: Resultado = val

Una vez que se ha logrado segmentar ambas partes del cuerpo, las características extraídas de cada persona son:

1. Contorno de la cabeza: se utiliza la función de MATLAB *edge(),* dentro de la cual se eligió el método de Sobel para la extracción de bordes. El método de Sobel utiliza dos kernels, un kernel para el eje x y un kernel para el eje y.

$$Kx = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \qquad Ky = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

   Ambos kernels se convolucionan con la imagen, obteniendo el gradiente respecto a *x* y el gradiente respecto a *y*.

$$Gx = Kx * Imagen \qquad Gy = Ky * Imagen$$

   Para obtener la magnitud del gradiente respecto a ambos ejes se utiliza la siguiente norma:

$$Im_{edge} = \sqrt{Gx^2 + Gy^2}$$

   Por último la imagen se umbraliza para obtener los bordes.

2. Contorno del cuerpo: al igual que para el contorno de la cabeza se utiliza la función *edge()* con el método de Sobel.

3. Cabeza en escala de grises de la imagen térmica: para obtener la imagen de la cabeza de cada persona en escala de grises se usa la función *rgb2gray()* de MATLAB, la cual consiste en que dada una imagen a color en formato RGB, se separa cada pixel en su componente R, G y B, el valor de cada componente es multiplicado por un escalar y sumado obteniendo un único valor, de tal forma que cada componente aporte un porcentaje de información al resultado final. El estándar utilizado es:

   Im_gris= 0.2989 * R + 0.5870 * G + 0.1140 * B

   De la misma forma se extraen las características 4,5 y 6.

4. Cabeza en escala de grises de la imagen visual.

5. Cuerpo en escala de grises de la imagen térmica.

6. Cuerpo en escala de grises de la imagen visual.

7. Puntos SURF de la imagen térmica de la persona: para la extracción de los puntos SURF se utiliza la función de MATLAB *detectSURFFeatures(),* dicha función como su nombre lo indica aplica el método SURF, explicado en el capítulo 2, para extraer puntos característicos. Esta es aplicada sobre el recuadro de la imagen térmica que enmarca a cada persona y los parámetros utilizados fueron: número de octavas igual a 3 y número de escalas por octava igual a 4. Ambos parámetros se determinan experimentalmente.

8. Puntos SURF de la imagen visual de la persona: al igual que en la característica 9, se usa la función *detectSURFFeatures()* de MATLAB. Esta vez se aplica sobre los recuadros en la imagen visual.

9. Histograma de color de la imagen visual de la persona completa: en el formato estándar RGB se tienen 256 valores de rojo, 256 valores de verde y 256 valores de azul para representar un total de 16,777,216 colores diferentes. Sin embargo, debido a que un histograma tan grande provocaría un incremento en el tiempo de cálculo, se decidió reducir el histograma a 512 combinaciones de color. Esto se realiza dividendo cada componente de color en 8 bloques, donde cada uno de estos bloques engloba 32 niveles de intensidad. A todas las intensidades dentro de un bloque se les asigna un único valor. Por ejemplo a los valores de 0 a 31 se les asignara el valor de 0, a los valores de 32 a 63 se les asigna el valor de 1 y así sucesivamente hasta llegar al bloque 7.

   Una vez realizada dicha reducción por bloques, el histograma es computado utilizando la siguiente ecuación:

$$Hist\big((64 \times B(i,j)) + (8 \times G(i,j)) + R(i,j)\big) = Hist\big((64 \times B(i,j)) + (8 \times G(i,j)) + R(i,j)\big) + 1$$

Donde i, j son la posición del pixel en la imagen y estas se recorren pixel por pixel.

10. Proporción cabeza-cuerpo de cada persona: ya que se ha segmentado la cabeza del cuerpo, la proporción que existe entre ambas secciones se obtiene de forma muy sencilla utilizando una regla de tres.

11. Predicción de la posición de la persona según el filtro de Kalman: las variables de estado que describen el sistema dinámico lineal fueron modeladas tomando en cuenta las coordenadas de posición y la velocidad del objetivo, obteniendo un vector x_t de cuatro elementos:

$$x_t = \begin{bmatrix} coord\ x \\ coord\ y \\ Vx \\ Vy \end{bmatrix}$$

Se hace la consideración de que la aceleración es cero ya que en la secuencias de imágenes capturadas las personas caminan naturalmente sin un cambio brusco en la velocidad de su caminata, es decir no pasan de caminar a correr o viceversa.

Considerando la influencia de la velocidad y la posición previa del objetivo sobre la posición actual, así como una velocidad constante se obtiene el siguiente sistema de ecuaciones:

$$coord\ x_t = coord\ x_{t-1} + Vx_{t-1}$$
$$coord\ y_t = coord\ y_{t-1} + Vy_{t-1}$$
$$Vx_t = Vx_{t-1}$$
$$Vy_t = Vy_{t-1}$$

De esta forma la matriz de transición de estado está dada por:

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Por otro lado, la matriz de relación entre la variable de estado y la medida, está dada por la matriz la siguiente matriz:

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Esto se debe a que las medidas obtenidas incluyen solamente los datos de la posición en x y la posición en y.

Los vectores de ruido gaussiano w_t y v_t se obtuvieron mediante la técnica iterativa explicada en el capítulo 2, para la cual se necesita una primera "buena aproximación" para ambos vectores. Estas primeras aproximaciones se obtuvieron midiendo manualmente, calculando w_t y v_t de tres secuencias de imágenes tomadas al azar, las ecuaciones utilizadas para el cálculo son:

$$w_t = x_t - Ax_{t-1}$$
$$v_t = y_t - Cx_t$$

Los valores obtenidos fueron promediados para obtener la aproximación inicial que se propone para cada vector:

$$w_0 = \begin{bmatrix} 3 \\ 6 \\ 2 \\ -5 \end{bmatrix} \qquad v_0 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

Por último se obtienen las matrices de covarianza Q y R utilizando la función cov() de MATLAB. El método iterativo se detiene cuando se cumple la condición Q=Q y R=R para cada caso, tomando en cuenta en esta igualdad solo los valores enteros.

Las características antes mencionadas nos servirán posteriormente para poder identificar una misma persona en las imágenes siguientes de la secuencia o determinar si hay una persona nueva en la escena. Cada una de estas características fue seleccionada a partir consultar en trabajos previos de reconocimiento e identificación de personas y/u objetos las características utilizas por cada método, así como cuáles son las características más comunes en el estado del arte. También se incluyeron características de la observación de como las personas reconocen individuos en una imagen con información visual restringida.

4.5 Seguimiento

Para realizar el seguimiento de las diferentes personas en las secuencias de imágenes, el algoritmo utilizado se explica en el diagrama de la figura 16, la implementación de los códigos fue realizada durante el desarrollo de esta tesis, haciendo uso solo de algunos comandos de MATLAB que se describirán con mayor detalle más adelante (sección 4.5.1 y 4.5.2).

Al entrar una imagen nueva en el sistema se realiza la detección de cada persona en la escena, utilizando el algoritmo propuesto por Navneet Dalal y Bill Triggs en el artículo [1]. Posteriormente, de cada una de estas personas se extraen sus características (ver sección 4.4). En caso de que sea la primera imagen de la secuencia, se guardan las características de cada persona en arreglos y se crean las primeras clases, es decir una clase por persona.

Por el otro lado, si la imagen no es la primera de la secuencia, se realiza una comparación de las características de las personas de la imagen actual con las características de las personas de las clases anteriormente creadas, para después clasificarlas según su similitud en las clases ya creadas o nuevas clases. Por último, se computan los puntos de la trayectoria de cada clase y se repite el proceso con cada imagen. Los algoritmos creados para la comparación de características y la clasificación se explican en las secciones 4.5.1 y 4.5.2 respectivamente.

En general el método propuesto consiste en aplicar los procesamientos y pre-procesamientos necesarios a las imágenes tanto térmicas como visuales para poder llegar a la parte de detección automática y seguimiento de personas. La principal aportación de esta tesis se encuentra en la propuesta, implementación y evaluación de un nuevo algoritmo de seguimiento utilizando un conjunto de características que describen a cada individuo en particular y a partir de ello se clasifican e identifica cada persona en la secuencias de imágenes.

4.5.1 Comparación de características

En la comparación de cada una de las características se siguen procesos distintos entre sí dado la naturaleza particular de cada característica.

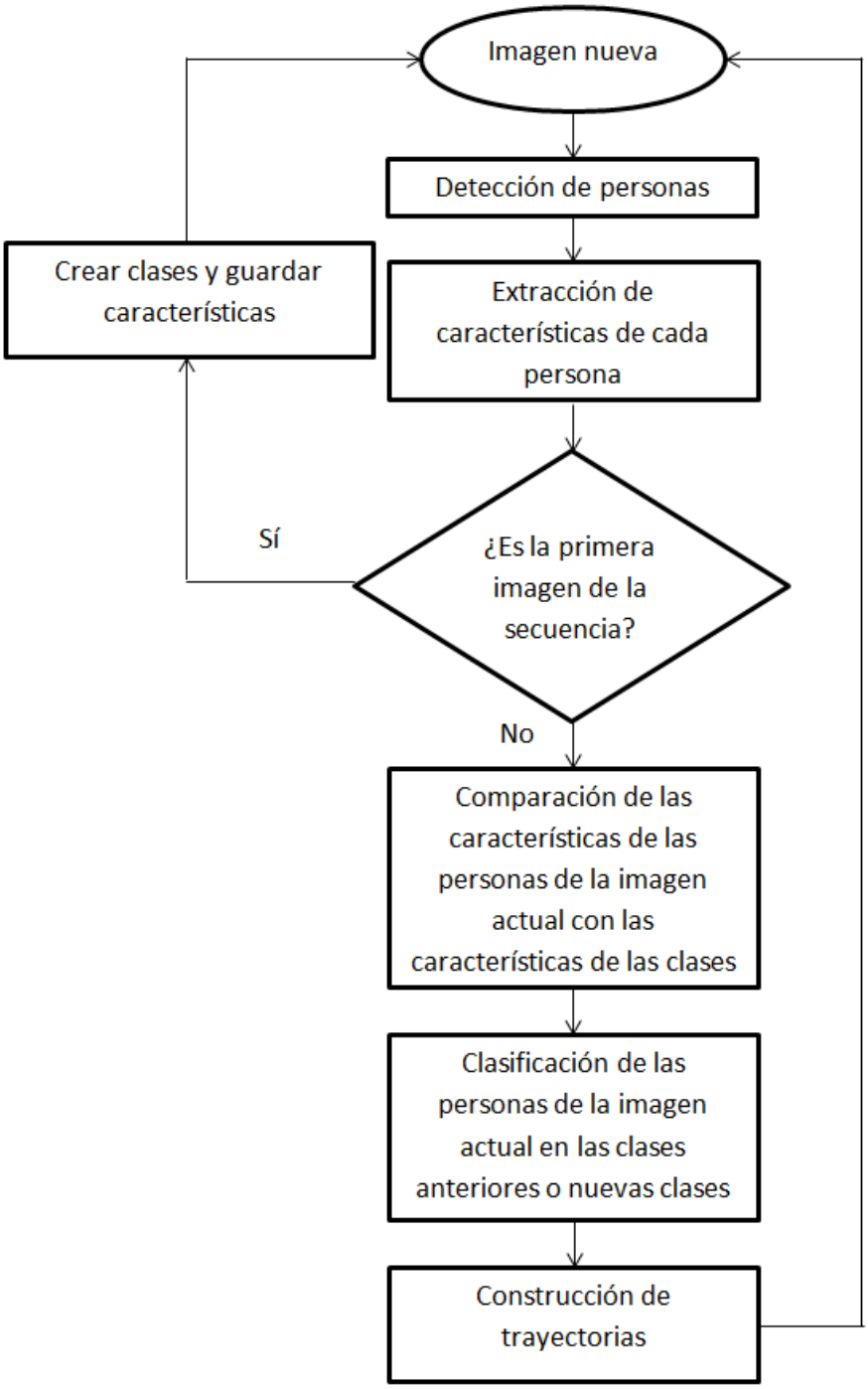


Figura 16. Algoritmo general del proceso de seguimiento

Para comparar el contorno del cuerpo, el contorno de la cabeza, la cabeza en escala de grises en ambos espectros, el cuerpo en escala de grises en ambos espectros y el histograma de color, se utiliza la correlación cruzada entre los arreglos que representan cada una de estas características en las diferentes personas a clasificar y los arreglos de las mismas características en las clases establecida anteriormente. La correlación cruzada es una medida de similitud entre dos señales que aplicada a señales discretas esta se puede definir como:

$$R[m] = \sum_{n=0}^{N-m-1} x[m+n]y^*[n]$$

Donde y^* denota el complejo conjugado de y *y* N es el tamaño de los vectores x y *y*. Para dos dimensiones la correlación cruzada se describe como:

$$R[i,j] = \sum_{m=0}^{Ma-1} \sum_{n=0}^{Na-1} A[m,n]B^*[m+i,n+j]$$

Donde $0 \leq i \leq Ma+Mb-1$, $0 \leq j \leq Na+Nb-1$, [Ma, Na] es el tamaño de la matriz A y [Mb,Nb] es el tamaño de la matriz B. Para su implementación se utilizan las funciones de MATLAB *xcorr()* y *normxcorr2(),* la primera de ellas es utilizada para la comparación de los histogramas de color y la segunda para la comparación de los recuadros del rostro y cuerpo de las personas.

La correlación cruzada es un método utilizado frecuentemente en el procesamiento de imágenes debido a su sencillez, confiabilidad y eficacia si se trata de imágenes que no rotan considerablemente [41], como es el caso de los recuadros que se desean comparar, los cambios de iluminación y escala son solucionados utilizando el método normalizado de correlación cruzada para las características que así lo requieren y unificando la escala de cada imagen a un mismo tamaño. Los valores resultantes de dichas correlaciones son guardados en una matriz que llamaremos de similitud.

Los siguientes datos que forman parte de la matriz de similitud, son el número de puntos SURF correspondientes en ambas imágenes (ver sección 2.9), para lo cual después de extraerse los puntos, se emplea el método del vecino más próximo para clasificarlos, tal como propone el autor del método [15] .Una vez

encontrados cuales son los puntos correspondientes estos se cuantifican por separado en cada espectro.

Después, se obtiene la diferencia entre ambas personas de la proporción del tamaño de la cabeza y el cuerpo de cada una, y por último, se calcula la diferencia entre el valor de la posición dada por el filtro de Kalman de la clase y la posición de la persona con la que se compara.

Las comparaciones que al dar un número menor indican mayor similitud son invertidas restándolas del valor máximo posible a obtener, con el objetivo de estandarizar que cuando una comparación dé un número mayor, esto indique una mayor similitud.

Todos los valores anteriores, son guardados en la matriz de similitud de la persona que en el momento está siendo comparada con todas la clases previamente determinadas. De tal forma que los renglones de la matriz de similitud corresponden a la característica que se está comparando y las columnas corresponden a la clase con la que se está comparando. Hay una matriz de similitud por cada persona detectada en la imagen actual. Por ejemplo en la matriz de abajo el elemento (1,1) representa la diferencia de la característica 1 de la persona que actualmente se está comparando, con respecto a la persona que representa la clase 1.

$$
\begin{array}{r c}
 & \begin{array}{ccc} clase\ 1 & clase\ 2 & clase\ 3 \end{array} \\
\begin{array}{r} carac.1 \\ carac.2 \\ carac.3 \\ carac.4 \\ carac.5 \\ carac.6 \\ carac.7 \\ carac.8 \\ carac.9 \\ carac.10 \\ carac.11 \end{array} &
\begin{bmatrix} .8 & .2 & .9 \\ 5 & 26 & 37 \\ .5 & .4 & .26 \\ 300 & 214 & 100 \\ 2 & 8 & 9 \\ 25 & 7 & 2 \\ 8 & 2 & 4 \\ .7 & .3 & 2 \\ 1.5 & 3.7 & 4.8 \\ 5 & 3.1 & 2 \\ .05 & .01 & .02 \end{bmatrix}
\end{array}
$$

Figura 17. Ejemplo de la matriz de similitud suponiendo once características y tres clases distintas

Es importante mencionar, que los valores de las clases se van actualizando realizando un promedio entre los valores de las últimas tres personas clasificadas como de dicha clase.

4.5.2 Clasificación de personas

De la matriz de similitud de cada persona, se extrae un vector con los valores máximos y otro con la clase a la que pertenecen dichos valores, como se muestra en la figura 18. Posteriormente se crea un vector donde se ordena las clases, de la que se repitió más en el vector *clases* a la que se repite menos, junto con otro vector donde indica el número de repeticiones de la clase, como se aprecia en la figura 18.

$$\begin{array}{r} \\ carac.1 \\ carac.2 \\ carac.3 \\ carac.4 \\ carac.5 \\ carac.6 \\ carac.7 \\ carac.8 \\ carac.9 \\ carac.10 \\ carac.11 \end{array} \begin{array}{c} \begin{array}{ccc} clase\ 1 & clase\ 2 & clase\ 3 \end{array} \\ \begin{bmatrix} .8 & .2 & .9 \\ .3 & .26 & .57 \\ .5 & .4 & .26 \\ .31 & .24 & .1 \\ .2 & .8 & .9 \\ .75 & .4 & .2 \\ 8 & 2 & 4 \\ 71 & 34 & 2 \\ .5 & .73 & .48 \\ 31 & 58 & 12 \\ 500 & 113 & 215 \end{bmatrix} \end{array} \rightarrow \begin{array}{c} val.max. \\ \begin{bmatrix} .9 \\ .57 \\ .5 \\ .31 \\ .9 \\ .75 \\ 8 \\ 71 \\ .84 \\ 58 \\ 500 \end{bmatrix} \end{array} \begin{array}{c} clases \\ \begin{bmatrix} 3 \\ 3 \\ 1 \\ 1 \\ 3 \\ 1 \\ 1 \\ 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} \end{array} \rightarrow \begin{array}{cc} clases\ mas\ rep. & no.de\ rep. \\ \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} & \begin{bmatrix} 6 \\ 4 \\ 1 \end{bmatrix} \end{array}$$

Figura 18. Ejemplo de la extracción de los vectores de valores máximos, clases, el vector de las clases más repetidas y el vector del número de repeticiones de dichas clases.

Con estos datos computados se procede a hacer la clasificación lo cual nos permite identificar y conocer la posición de una misma persona a lo largo de la secuencia de imágenes. Para ellos se utilizan una serie de reglas del tipo if-then. Las reglas if-then se implementaron para evitar errores de redundancia en la clasificación de personas, como por ejemplo, cuando dos personas llegan a tener un mayor grado de similitud con una misma clase que con el resto, por lo tanto dos personas diferentes en una misma imagen podrían ser clasificadas como la misma persona. Este tipo de problemas se solucionan utilizando una regla que nos permita detectar que persona tiene un mayor grado de similitud con la clase. En el diagrama de la figura 19 se puede apreciar el algoritmo.

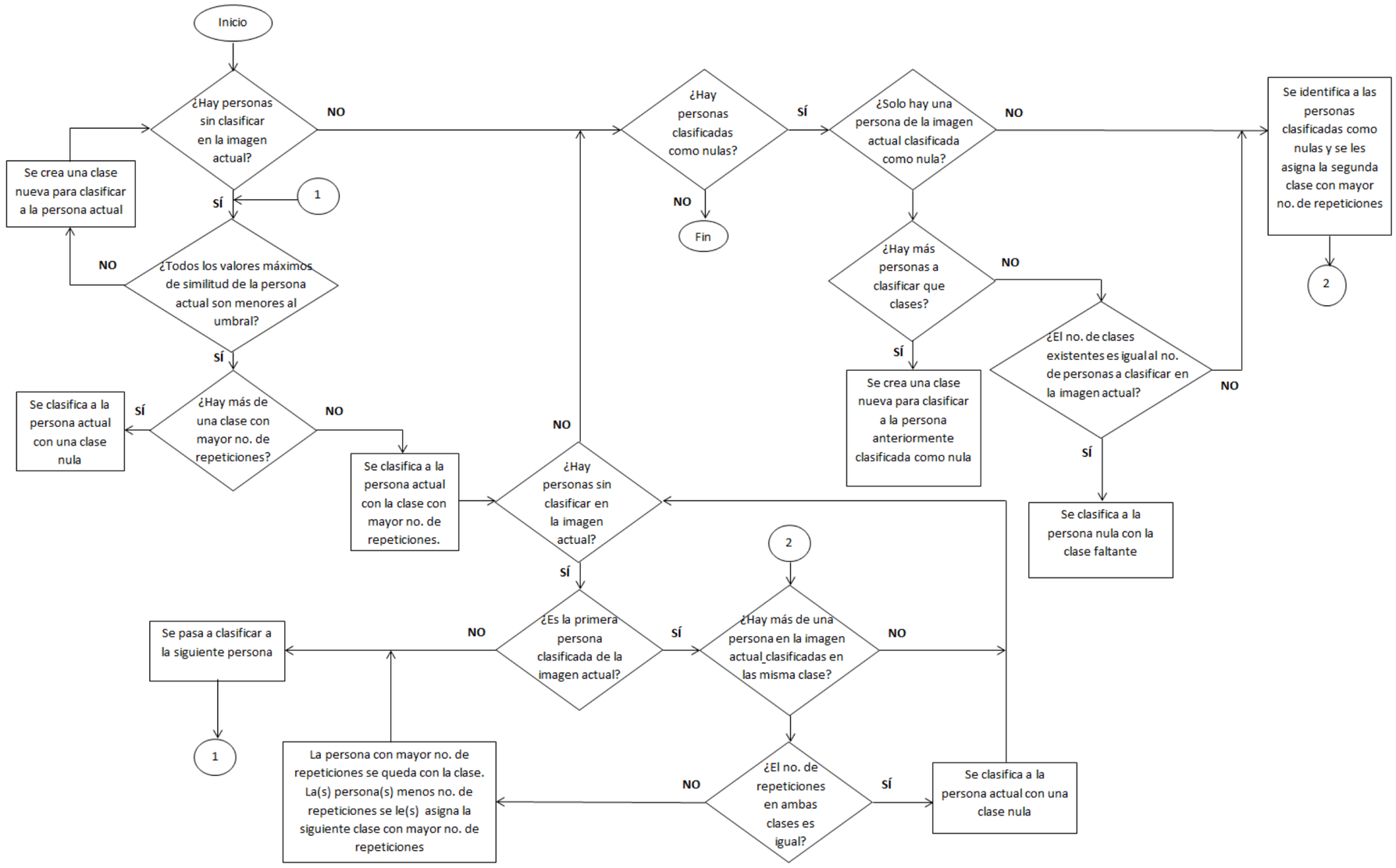


Figura 19. Diagrama de flujo que representa el algoritmo de clasificación.

El diagrama de flujo que se presenta en la figura 19 representa el funcionamiento del algoritmo que se implementó para clasificar a las personas en una imagen, es decir, este algoritmo se repite para cada imagen de la secuencia, hasta obtener la clasificación y seguimiento de todas las personas en la secuencia.

4.6 Resumen

En el capítulo actual se describieron los métodos usados, propuestos y desarrollados para la obtención del resultado final de esta tesis. El capítulo se divide en las siguientes secciones:

Captura de imágenes y calibración de cámaras: En donde se describe el procedimiento que se siguió para la captura de las imágenes y se indica la herramienta utilizada para la calibración.

Registro de imágenes: Se describen las cuatro etapas que implica el registro de imágenes. La detección de características y correspondencia de características se realizaron de forma manual, mientras que para la inferencia de transformación geométrica y aplicación de la transformación se utilizaron los comandos de MATLAB *maketform()* y *tformarray()*respectivamente.

Detección de personas: Se indican el método utilizado (HOG y SVM), se detallan los parámetros empleados en dicho método y se explican las adaptaciones del método para nuestra aplicación.

Caracterización de las personas: Se describen las características que se extrajeron para identificar a cada persona, estas son once: contorno de la cabeza, contorno del cuerpo, cabeza en escala de grises de la imagen térmica, cabeza en escala de grises de la imagen visual, cuerpo en escala de grises de la imagen térmica, cuerpo en escala de grises de la imagen visual, puntos SURF de la imagen térmica de la persona, puntos SURF de la imagen visual de la persona, histograma de color de la imagen visual de la persona completa, proporción cabeza-cuerpo de cada persona y predicción de la posición de la persona según el filtro de Kalman.

Seguimiento: Esta sección se divide en dos partes, primero en la comparación de las características entre personas, que varía según la naturaleza de la característica y por último se explica mediante un diagrama de estados el clasificador empleado.

En el siguiente capítulo se expondrán los resultados obtenidos en esta tesis en comparación con el método [2].

Capítulo 5

Experimentos y resultados

En el capítulo que se presenta, se explica detalladamente el procedimiento seguido para realizar los experimentos que sustentan el método propuesto en esta tesis. Así como también, se presentaran y discuten los resultados obtenidos de dichos experimentos.

5.1 Experimentación

Los resultados de los métodos de seguimiento de personas que actualmente existen, pueden diferir ampliamente dependiendo de la dificultad y de las características que presenten las secuencias de imágenes con las que son evaluados. Es por ello que existe un gran problema en cuanto a comparación de métodos se trata, pues para poder hacer una comparación justa es necesario utilizar el mismo grupo de secuencias. Sin embargo, actualmente no existen disponibles de forma pública secuencias de imágenes que cumplan con las características que se proponen en esta tesis, las cuales son:

- Presentar oclusión
- Presentar presencia de múltiples personas uniformadas
- Haber sido capturadas a 4fps

- Encontrar disponible cada fotograma en el espectro infrarrojo y visual

Por lo cual fue necesario crear una base de pruebas propia y buscar códigos de las implementaciones de otros métodos. El método con el que se compara es [2], se eligió por que los resultados que presenta son equiparables con los resultados de los métodos más recientes y su implementación se encuentra disponible para descargar en su sitio web.

Por su parte las características de las secuencias de imágenes mencionadas, son importantes en el sentido de que presentan un alto grado de dificultad para seguimiento de personas. Sin embargo, muchas veces son necesarias para el seguimiento de personas en instituciones como hospitales, laboratorios, campos militares o escuelas, donde las personas usan uniforme. Además de que la mayoría de las cámaras de vigilancia capturan imágenes entre 7 fps y 4 fps, con el objetivo de ahorrar espacio de memoria.

Lo que se pretende mostrar con los resultados obtenidos en la tesis es la relación que existe entre los resultados del método [2] y los resultados del método propuesto en secuencias con las características específicas que se mencionaron con anterioridad. Para de esta forma saber si el método propuesto es competitivo con el estado del arte.

5.1.1 Metodología de la experimentación

Los experimentos fueron realizados con 5 secuencias de imágenes térmicas y visuales, capturadas a 4 fps (debido a que es la menor tasa de captura de las cámaras de vigilancia que ahorran espacio de memoria). Las secuencias son de diferente duración, varían desde 15 a 25 fotogramas, con una duración de 3.75s a 6.25s.

Las secuencias de imágenes fueron capturadas en diferentes fechas y en ambientes exteriores, bajo iluminación no controlada (esto implica iluminación natural que varía dependiendo del clima).

En cada una de las secuencias se aprecian personas adultas y uniformadas, en posición erguida, caminando en diferentes direcciones. Dependiendo de la persona y la secuencia de imágenes, esta puede presentar oclusión parcial y/o

total en diferentes ocasiones, así como también puede cambiar repentinamente el rumbo de la trayectoria de movimiento.

Las personas en las secuencias de imágenes se encuentran en un rango de 2 a 6 metros de distancias a partir de la localización de las cámaras y estas presentan movimiento sobre el plano X-Z del sistema coordenado con respecto a la cámara, ver figura 2.

5.1.2 Evaluación

Cada secuencia de imágenes fue procesada por el método propuesto y por el método [2] que llamaremos ILRVT por sus siglas en inglés (ver capítulo 3). De los resultados arrojados por ambos métodos se obtuvo el porcentaje de verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y el porcentaje de confusiones del objetivo. Para el caso del seguimiento de personas en imágenes, definimos como VP al seguimiento correcto de una persona, es decir, la cantidad de veces en una secuencia que la persona es detectada y clasificada correctamente. Los FP son definidos como la cantidad de seguimiento incorrecto o el número de veces que una persona es identificada como otra. Por su parte se definen a los FN como el número de objetivos o personas no detectadas por el seguidor. Por último, el número de confusiones es la cantidad de ocasiones que el seguidor cambia al objetivo que actualmente sigue por otro objeto. Es decir, debido a que ambos métodos dependen del resultado de la clasificación previa para realizar la clasificación actual correctamente, en ocasiones si la clasificación previa erro, dicho error se continúa hasta el final de la secuencia, sin embargo, en otros casos después del primer error se sigue cambiando de objetivo. El número de veces que se cambia de objetivo es lo que se determinó como el número de confusiones.

Los resultados fueron cuantificados de forma visual, conociendo de antemano la localización de cada persona y comparándola con los resultados arrojados por los métodos. Si las coordenadas del centroide de la detección se encuentra entre la cabeza y el inicio de las piernas de la persona en el eje y y entre hombro y hombro de la persona en el eje x, se toma como VP.

5.1.2 Experimentos realizados

Como se mencionó anteriormente, las secuencias de imágenes pueden variar en algunas de sus características, por lo tanto en la siguiente tabla se marcan con una "X" las características propias de las secuencias con las que se realizaron las pruebas, así como el número de fotogramas que las conforman. Se incluyeron secuencias de imágenes con diferente conjunto de características entre sí para poder probar el método bajo diferentes circunstancias.

Tabla 2
Características de las secuencias de prueba

Secuencia / **Características**	1	2	3	4	5
Capturada en ambientes exteriores	X	X	X	X	X
Capturada con iluminación no controlada	X	X	X	X	X
Contiene múltiples objetivos	X	X	X	X	X
Objetivos similares entre sí (personas uniformadas)	X	X	X	X	X
Los objetivos presentan oclusión parcial y total	X	X	X	X	X
Movimiento de los objetivos sobre el eje X	X	X	X	X	X
Movimiento de los objetivos sobre el eje Z		X			X
Cambio repentino del rumbo de movimiento de los objetivos	X			X	X
Número de fotogramas	15	15	19	18	25

Se realizaron pruebas con las 5 secuencias, el método ILRVT fue probado empleando el codigo proporcionado por sus autores [2] . Se utilizaron para ello tres modalidades diferentes:

1. Utilizando solo secuencias de imágenes en el espectro visible.
2. Utilizando solo secuencias de imágenes en el espectro infrarrojo.
3. Utilizando fusión de los resultados de las dos modalidades anteriores.

Por su parte el método propuesto, el cual invariablemente utiliza fusión de imágenes, también fue probado con las 5 secuencias.

5.2 Resultados

A continuación se presentan los resultados de cada una de las secuencias en sus respectivas tablas:

Tabla 3
Resultados del experimento con la secuencia número 1

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	54.16%	50%	62%	**70.83%**
FP	45.85%	50%	37.5%	**29.16%**
FN	29.16%	33.33%	12.5%	**4.16%**
Confusiones	37.5%	41.66%	33.33%	**12.5%**

Tabla 4
Resultados del experimento con la secuencia número 2

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	61.53%	42.3%	73.07%	**80.76%**
FP	38.46%	57.69%	26.92%	**19.24%**
FN	23.07%	46.15%	11.53%	**3.84%**
Confusiones	34.62%	26.92%	23.07%	**11.53%**

Tabla 5
Resultados del experimento con la secuencia número 3

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	34.09%	65.9%	**68.18%**	**68.18%**
FP	65.91%	34.1%	**31.82%**	**31.82%**
FN	57.7%	28.8%	20%	**2.27%**
Confusiones	31.1%	24.4%	22.22%	**15.5%**

Tabla 6
Resultados del experimento con la secuencia número 4

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	37.8%	48.8%	60%	**75.5%**
FP	62.2%	51.2%	40%	**24.5%**
FN	55.5%	31.1%	20%	**2.2%**
Confusiones	28.8%	24.4%	17.7%	**13.3%**

Tabla 7
Resultados del experimento con la secuencia número 5

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	29.8%	45.6%	63.2%	**77.2%**
FP	70.2%	54.4%	36.8%	**22.8%**
FN	50.8%	31.5%	15.8%	**1.75%**
Confusiones	31.5%	17.5%	17.5%	**12.3%**

El promedio de los resultados para cada método se presenta en la tabla numero 8:

Tabla 8
Promedio de los resultados de todos los experimentos

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	43.48%	50.52%	65.29%	**74.49%**
FP	56.52%	49.48%	34.60%	**25.50%**
FN	43.25%	34.18%	15.97%	**2.84%**
Confusiones	32.70%	26.98%	22.76%	**13.03%**

La desviación estándar de los resultados se presenta en la tabla 9:

Tabla 9
Desviación estándar de los resultados de todos los experimentos

	ILRVT con imágenes visuales	ILRVT con imágenes IR	ILRVT con fusion de imagenes	Método propuesto.
VP	13.67	9.10	5.29	**5.02**
FP	13.67	9.10	5.22	**5.02**
FN	15.97	6.89	4.01	**1.08**
Confusiones	3.39	8.92	6.43	**1.52**

5.3 Discusión

Como se puede apreciar en la sección 5.2 el porcentaje de seguimiento correcto o porcentaje de VP, varía entre un 68% hasta superar el 80% dependiendo de la secuencia de imágenes, aunque en general se puede apreciar una tendencia a encontrarse alrededor del 75% de VP. La diferencia que puede existir entre los resultados de una secuencia y otra radica en la dificultad que esta presente, ya que como se ha mencionado estas varían en algunas de sus características. Sin embargo, la desviación estándar de los VP es de 5.02 lo que nos indica el grado de precisión del método propuesto a pesar de las diferencias entre las secuencias de imágenes.

En comparación con el método ILRVT, podemos notar que cuando se utiliza un solo tipo de imágenes (infrarrojas o visibles) los resultados obtenidos por el método propuesto, el cual utiliza fusión de ambos tipos de imágenes, son claramente superiores. Sin embargo, cuando se utiliza fusión de imágenes con el método ILRVT, los resultados obtenidos son más cercanos a los alcanzados por el método propuesto, aunque en 4 de las 5 secuencias nuestro método da resultados superiores mostrando una diferencia más significativa en el porcentaje de confusiones y de FN. Así como también muestra mejores resultados en la desviación estándar lo que implica una constancia en el rango de los resultados obtenidos.

Se puede notar, que en las secuencias de imágenes donde se aprecia una mejora más significativa, de los resultados del método propuesto en comparación con

el ILRVT, son en aquellas secuencias que presentan cambios repentinos de trayectoria de los objetivos, esto muy probablemente se deba a que el ILRVT como en muchos de los métodos de seguimiento se basan en predictores de trayectoria.

Los predictores de trayectorias, suelen ser muy útiles en los sistemas de seguimiento de personas, pues normalmente los movimientos de las personas no suelen cambiar bruscamente de un momento a otro. Sin embargo, cuando se emplean en cámaras de vigilancia reales, muchas de las cuales buscan ahorrar espacio de memoria capturando hasta 4 fps en lugar de 24 o 30 fps como el resto de las cámaras de video, las muestras son tan pocas que el movimiento de una persona puede variar considerablemente de los resultados arrojados por los predictores.

En el método que se propone en esta tesis, también se hace uso de un predictor de trayectoria (el filtro de Kalman), sin embargo, no es el único factor que se toma en cuenta para decidir el seguimiento, sino que es una de otras 10 características. Esto ayuda a un mejor desempeño bajo las circunstancias antes mencionadas.

5.3 Resumen

Este capítulo se refirió a la parte de experimentación y los resultados obtenidos. De forma general, los experimentos constaron de una comparación entre el método ILRVT [2] y el método propuesto en esta tesis, utilizando para ello un conjunto de secuencias de imagenes.

En este capítulo se detalló la metodología para realizar los experimentos, desde las características de las secuencias de imágenes hasta la descripción de las medidas utilizadas para la comparación. También se mencionaron cuales son los experimentos realizados.

En la sección 5.2 se presentan las tablas de los resultados obtenidos en la experimentación, así como el promedio de los resultados y la desviación estándar. Por último se realiza una breve discusión de los resultados

En el próximo capítulo se realiza una discusión sobre las conclusiones del trabajo que se presenta, así como también se menciona el trabajo a futuro con el que podría continuar el desarrollo de esta tesis.

Capítulo 6

Conclusiones y trabajo a futuro

En este capítulo se realizaran conjeturas sobre las fortalezas y debilidades del método propuesto, así como también se realizara un análisis del mismo, sus principales aportaciones y sus principales líneas de investigación para trabajo a futuro

6.1 Conclusiones

El seguimiento de personas, aunque es un tema ampliamente abordado, sigue siendo un problema que aún no ha sido resuelto en su totalidad, pues la amplia variedad de circunstancias que se pueden presentar es tan grande que resulta en la necesidad que desarrollar métodos para situaciones específicas y/o métodos más robustos a diferentes tipos de situaciones.

Lo que se propuso desarrollar en esta tesis fue un método aplicado a la detección automática y seguimiento de múltiples personas en secuencias de imágenes, específicamente en secuencias que presentan oclusión de el o los objetivos, presencia de personas uniformadas y que las secuencias hayan sido capturas a una tasa de 4 fps. Este tipo de restricciones son necesarias en entornos institucionales como lo son hospitales, escuelas, campos militares, etc.

Para resolver el problema se realizó una investigación sobre los trabajos previos en el seguimiento de personas y a partir de esta se concibió un método que utilizando fusión de múltiples sensores, en específico cámaras infrarrojas y visuales, pudiera seguir personas bajo las circunstancias antes mencionadas. Para ello fue necesario plantearse tres objetivo particulares, el primero de ellos consistió en hacer una recopilación de las características que diferentes métodos utilizan para realizar detección y seguimiento de objetos, tanto en el espectro térmico como en el visual, posteriormente después de analizar estas características se descartaron aquellas no aptas para las restricciones de las secuencias de imágenes y a partir del resto de características se conformó un modelo para identificar a cada persona.

El segundo objetivo también requirió de una amplia revisión de los métodos más utilizados para la detección de personas, de entre ellos se eligió un método confiable y adaptable a los diferentes espectros con los que se trabaja. Dicho método fue implementado y aplicado a nuestras secuencias de imágenes tanto térmicas como visuales.

Para el último objetivo, se concibió e implemento, según las necesidades del problema, un algoritmo capaz de realizar el seguimiento de cada persona distinguiéndola del resto, según la comparación entre los modelos de características propios de cada individuo.

Finalmente como resultado del desarrollo de la tesis se realizó un método de seguimiento de personas utilizando fusión de imágenes térmicas y visuales evaluadas con las secuencias de imágenes con las características antes mencionadas y descritas con mayor detalle en la tabla 2. Obteniendo un promedio de 74.5% de seguimiento correcto del método propuesto en comparación con un 65.3% obtenido por el método de comparación ILRVT[2].

El método ILRVT reporto una tasa de 1% de error en el seguimiento para las secuencias con las que fue probado, lo cual lo hace equiparable con muchos otros métodos de seguimiento actuales. Sin embargo, se puede apreciar una gran diferencia con los resultados arrojados en los experimentos de esta tesis, esto se debe al incremento en la complejidad de las imágenes a las que fue aplicado. Por lo tanto no se puede realizar una aseveración del rendimiento del método propuesto con respecto al resto de los métodos, pues se necesitaría

implementar cada uno de estos algoritmos y probar con las mismas imágenes para poder llegar a una afirmación, lo cual requeriría de un trabajo mucho más extenso.

Las principales contribuciones que este método aporta dentro de su área, consiste en proponer una nueva y prometedora forma de seguimiento y detección capaz de seguir personas aún si hay personas uniformadas en la escena, lo cual aumenta la dificultad, pues los objetivos son muy similares entre sí y esto puede causar que el seguidor confunda a los individuos. Otra aportación es que el seguidor funciona para secuencias de imágenes muestreadas a muy pocos cuadros por segundo (4 fps), esto debido a que muchas de las cámaras de vigilancia empleadas en aplicaciones reales funcionan de esta manera. En general el objetivo de la tesis se logró igualando o superando en diferente medida al método de comparación.

6.2 Trabajo a futuro

A continuación se enlistan algunas de las líneas de investigación y trabajo a futuro con lo que se le podría dar continuación a esta tesis:

- Analizar que otras características pueden ser integradas e integrarlas al método, con el objetivo de obtener un mejor funcionamiento e incrementar el porcentaje de seguimiento correcto. Por ejemplo la información mutua entre imágenes, descriptores en el dominio de la frecuencia u otro dominio, así como nuevos descriptores que puedan surgir para la identificación de personas etc.

- Realizar pruebas en escenarios más variados con una mayor cantidad de obstáculos, escenarios nocturnos y con mayor cantidad de personas.

- Ampliar la distancia del rango del espacio de trabajo, tanto en el campo de visión de las cámaras como en la distancia a la que se encuentran las personas de la cámara.

- Ampliar el método de detección y seguimiento de personas para que también sea capaz de detectar niños y personas en diferentes posiciones, no solo personas adultas en posición erguida.

- Debido a que el método de seguimiento se basa en reconocer a cada persona por sus características, tal vez valdría la pena probar el método y adaptarlo para reconocimiento de personas en múltiples cámaras.

- Comparar el método propuesto con una mayor cantidad de métodos, lo que implica hacer una recopilación de los métodos de seguimiento actuales, realizar su implementación y comparar utilizando las secuencias de imágenes propuestas.

Apéndices

Apéndice A - Acrónimos

2D 2 Dimensiones

3D 3 Dimensiones

ASM *Active Shape Model* (modelo de forma activa)

AST *Active Shape Traker* (seguidor de forma activa)

Ec. Ecuación

FN Falsos Negativos

FP Falsos Positivos

FPS Fotogramas por segundo

HOG *Histograms of Oriented Gradiente* (histogramas del gradiente orientado)

IR Infrarrojo

MATLAB *Matrix Laboratory* (laboratorio de matrices)

MHT Multiple Hypothesis Tracking (seguidor de múltiples hipótesis)

MLE Maximum Likelihood Estimation (estimación de máxima probabilidad)

MS Mean Shift (media de cambio)

Multidimensional Múltiples-dimensiones

Multimodal Múltiples-modalidades

Multinivel Múltiples-niveles

Multiobjetivo Múltiples-objetivos

Multisensor Multiples-sensores

Pixel *Picture-Element* (imagen-elemento)

Pre-procesamiento Previo-procesamiento

SIFT *Scale Invariant feature transform* (transformada de características invariantes a escala)

SURF *Speeded Up Robust Features* (características robustas aceleradas)

SVM *Support Vector Machines* (máquinas de vectores de soporte)

VN Verdaderos Negativos

VP Verdaderos Positivos

Bibliografía

[1] N. Dalal, B. Triggs, and D. Europe, "Histograms of Oriented Gradients for Human Detection," *Computer Vision and Pattern Recognition 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, 2005.

[2] D. a. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 125–141, Aug. 2008.

[3] G. Bhatnagar and Q. M. J. Wu, "Human Visual System Based Framework for Concealed Weapon Detection," *2011 Canadian Conference on Computer and Robot Vision*, pp. 250–256, May 2011.

[4] R. Collins, "Lecture 12: Camera Projection." Penn State University, Middletown, 2007.

[5] F. Ayres, C. E. V. Cantor, and A. L. Jimeno, *Teoría y problemas de geometría proyectiva*. New York, USA: McGraw-Hill, 1971.

[6] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, vol. 1, pp. 666–673.

[7] G.-A. Bilodeau, A. Torabi, and F. Morin, “Visible and infrared image registration using trajectories and composite foreground images,” *Image and Vision Computing*, vol. 29, no. 1, pp. 41–50, 2011.

[8] R. S. Blum and Z. Liu, *Multi-sensor image fusion and its applications*, 1st ed. CRC press, 2006.

[9] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital image processing using MATLAB*, 2nd ed., vol. 2. Mc Graw Hill, 2010.

[10] R. Collins, “Lecture 14: Prameter Estimation.” Penn State University, Middletown, 2007.

[11] D. D. L. Hall and S. A. H. McMullen, *Mathematical Techniques in Multisensor Data Fusion 2nd Ed.* Artech House Publishers, 2004.

[12] S. J. Krotosky and M. M. Trivedi, “Person Surveillance Using Visual and Infrared Imagery,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1096–1105, Aug. 2008.

[13] P. S. Maybeck, *Stochastic models , estimation, and control*, vol. 1. Academic Press, 1979.

[14] N. Funk, “A study of the kalman filter applied to visual tracking,” *University of Alberta, Project for CMPUT*, vol. 652, 2003.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded Up Robust Features,” *In Computer Vision–ECCV. Springer Berlin Heidelberg.*, vol. 3951, pp. 404–417, 2006.

[16] C. Park and S. Jung, “SIFT-based object recognition for tracking in infrared imaging system,” *2009 34th International Conference on Infrared, Millimeter, and Terahertz Waves*, pp. 1–2, Sep. 2009.

[17] M. Jordan, *Learning in graphical models (adaptive computation and machine learning)*. The MIT Press, 1998.

[18] A. Levey and M. Lindenbaum, “Sequential Karhunen-Loeve basis extraction and its application to images,” *Image Processing, IEEE Transactions on*, vol. 9, no. 8, pp. 1371–1374, 2000.

[19] I. Leichter, “Mean Shift Trackers with Cross-Bin Metrics,” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 34, no. 4, pp. 695–706, 2012.

[20] S. Chen, W. Zhu, and H. Leung, “Thermo-visual video fusion using probabilistic graphical model for human tracking,” *IEEE*, pp. 1926–1929, May 2008.

[21] D. Lee and S. Choi, “Multisensor fusion-based object detection and tracking using Active Shape Model,” *2011 Sixth International Conference on Digital Information Management*, pp. 108–114, Sep. 2011.

[22] M. Isard and A. Blake, “Active contours.” Springer-Verlag, 1998.

[23] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.

[24] A. Torabi, G. Massé, and G.-A. Bilodeau, “An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications,” *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210–221, Feb. 2012.

[25] B. Shoushtarian and H. E. Bez, “A practical adaptive approach for dynamic background subtraction using an invariant colour model and object tracking,” *Pattern Recognition Letters*, vol. 26, no. 1, pp. 5–26, 2005.

[26] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, vol. 2. Cambridge Univ Press, 2000.

[27] A. Torabi and G. Bilodeau, “A Multiple Hypothesis Tracking Method with Fragmentation Handling,” *Canadian Conference on Computer and Robot Vision*. pp. 8–15, 2009.

[28] Y. Caspi, D. Simakov, and M. Irani, “Feature-Based Sequence-to-Sequence Matching,” *International Journal of Computer Vision*, vol. 68, no. 1, pp. 53–64, 2006.

[29] J. Heikkila and O. Silven, “A four-step camera calibration procedure with implicit image correction,” in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 1106–1112.

[30] S. J. Krotosky and M. M. Trivedi, “Mutual information based registration of multimodal stereo videos for person tracking,” *Computer Vision and Image Understanding*, vol. 106, no. 2–3, pp. 270–287, May 2007.

[31] S. Krotosky and M. Trivedi, “Multimodal Stereo Image Registration for Pedestrian Detection,” *2006 IEEE Intelligent Transportation Systems Conference*, pp. 109–114, 2006.

[32] J. Y. Bouguet, “Camera calibration toolbox for maltab.” p. http://www.vision.caltech.edu/bouguetj/calib_doc/.

[33] L. Dang, B. Bui, P. D. Vo, T. N. Tran, and B. H. Le, “Improved HOG Descriptors,” *2011 Third International Conference on Knowledge and Systems Engineering*, pp. 186–189, Oct. 2011.

[34] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: an evaluation of the state of the art.,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–61, Apr. 2012.

[35] W. Choi, C. Pantofaru, S. Savarese, C. Engineering, A. Arbor, W. Garage, and M. Park, “Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion,” pp. 1076–1083, 2011.

[36] B. Wu and R. Nevatia, “Pedestrian Detection in Infrared Images based on Local Shape Features,” *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Jun. 2007.

[37] F. Barrera Campo, F. Lumbreras Ruiz, and A. D. Sappa, “Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 5, pp. 437–446, 2012.

[38] C. Ó. Conaire, N. E. O’Connor, and A. Smeaton, “Thermo-visual feature fusion for object tracking using multiple spatiogram trackers,” *Machine Vision and Applications*, vol. 19, no. 5–6, pp. 483–494, 2008.

[39] D. Tan, K. Huang, S. Yu, and T. Tan, “Efficient night gait recognition based on template matching,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, vol. 3, pp. 1000–1003.

[40] S. Munder and D. M. Gavrila, “An experimental study on pedestrian classification,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 11, pp. 1863–1868, 2006.

[41] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain,” *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.