UNIVERSIDAD REY JUAN CARLOS

Escuela Técnica Superior de Ingeniería Informática

Estudio de un Sistema de Videovigilancia con Imágenes de Baja Calidad

TESIS DOCTORAL

Autor: **D. Andrea Magadán Salazar**

Director: **Prof. Dr. D. Enrique Cabello Pardos**
Director: **Prof. Dr. D. Isaac Martín de Diego**

Madrid, Octubre 2015

UNIVERSIDAD REY JUAN CARLOS

Escuela Técnica Superior de Ingeniería Informática

Departamento de Ciencias de la Computación, Arquitectura de Computadores,
Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa

Estudio de un Sistema de Videovigilancia con Imágenes de Baja Calidad

TESIS DOCTORAL
Curso Académico 2015-2016

Autor: **D. Andrea Magadán Salazar**

Director: **Prof. Dr. D. Enrique Cabello Pardos**
Director: **Prof. Dr. D. Isaac Martín de Diego**

Madrid, Octubre 2015

Información de Contacto:
Andrea Magadán Salazar
andrea.magadan@urjc.es; magadan@cenidet.edu.mx;
magadandy@gmail.com

Impresión: Octubre de 2015

El Prof. Dr. Enrique Cabello Pardos, Profesor Titular de la Universidad del Departamento de Ciencias de la Computación, Arquitectura de Computadores, Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa, de la Universidad Rey Juan Carlos; y el Prof. Dr. Isaac Martín de Diego Profesor Interino de la Universidad del Departamento de Ciencias de la Computación, Arquitectura de Computadores, Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa, de la Universidad Rey Juan Carlos; como directores de la tesis doctoral titulada **"Estudio de un sistema de videovigilancia con imágenes de baja calidad"**, realizada por la doctorando D. **Andrea Magadán Salazar**, Maestra en Ciencias, en Ciencias de la Computación.

CERTIFICAN

Que esta Tesis Doctoral reúne los requisitos necesarios para su defensa y aprobación.

Y para que conste, expiden y firman la presente en Móstoles (Madrid), a 26 de octubre de 2015.

Fdo.: Dr. Enrique Cabello Pardos	**Fdo.: Dr. Isaac Martín de Diego**

A Caro e Isaac
Por todo el tiempo sin mamá

AGRADECIMIENTOS

A Dios que está siempre conmigo, me protege y me guía.

Deseo expresar mi profundo agradecimiento al Dr. Enrique Cabello Pardos por creer en mí y darme la oportunidad de desarrollar esta tesis bajo su asistencia. Por su soporte profesional y personal durante toda mi estancia como estudiante en el grupo de Reconocimiento Facial y Visión Artificial (FRAV), por su paciencia y confianza en todo momento; y sobre todo por su valiosa amistad.

Al Dr. Isaac Martín de Diego por su apoyo en el desarrollo de esta tesis. Por todas las sugerencias realizadas para un mejor desarrollo de esta tesis. Por enseñarme el potencial que puede tener una idea.

A la Dra. Cristina Conde Vilda por sus sugerencias en el desarrollo del presente trabajo. Por sus palabras de aliento.

A los revisores por sus consejos y sugerencias.

Al Tecnológico Nacional de México, (TecNM), por el otorgamiento de la licencia por beca comisión para la realización de mis estudios doctorales.

Al Centro Nacional de Investigación y Desarrollo Tecnológico, (Cenidet), por darme la oportunidad de continuar con mi formación académica.

Al Programa para el Desarrollo Profesional Docente, para el tipo Superior, (PRODEP), por el apoyo económico para realizar mis estudios doctorales.

Al Instituto Tecnológico de Cuautla, en especial al Dr. Gerardo Reyes y al Ing. Gerardo Jiménez por su apoyo para la creación de la base de datos ITC.

Quiero agradecer muy especialmente a Matilde Velazco Soni por sus asesorías y apoyo para la realización correcta y oportuna de los trámites ante ITNM; gracias por tu amistad. A Maira Correa, Agustín Camarillo, Erika Flores y Silvia Muñoz por su apoyo en la realización de los trámites administrativos de manera oportuna ante ITNM, CENIDET y PRODEP. A todos mis compañeros de CENIDET, que de manera directa e indirecta han ayudado a que esta tesis se haya realizado.

Mi agradecimiento infinito a Daniela Moctezuma y Oscar Siordia por su apoyo en el desarrollo de esta tesis. Por compartir sus conocimientos conmigo, por sus asesorías y palabras de aliento. Por permitirme llegar a su hogar, recibirme y cuidarme como un miembro de su familia dándome su compañía y amistad de manera generosa. Por su ayuda y sostén en infinidad de momentos y actividades diarias. A Dani por todas esas platicas diarias al ir y regresar de la Universidad. Decir "mil gracias" es poco para expresarles mi gratitud.

A mis compañeros del Grupo de Reconocimiento Facial y Visión Artificial (FRAV) de la Universidad Rey Juan Carlos: Javier, Ignacio, Jorge, Aris, Julio, Alexis y Oscar. Con sus ideas me dieron una nueva visión de mi trabajo de tesis. Por su compañía, amistad y tantas charlas sobre España y México.

A mis hermanos y sobrinos por su apoyo en cuidar a mis hijos, por todo el tiempo y cariño entregado a ellos. Ustedes son ejemplo de amor y apoyo infinito. Un agradecimiento muy especial y enorme para mi mamá y hermana Carmen. Ustedes son mi mayor soporte. Gracias!!!

A mis hijos Carolina e Isaac. Ustedes son mi mayor tesoro. Gracias por entender y soportar mis ausencias. A pesar de su edad, ustedes demostraron ser más fuertes y valientes que yo. Les amo!!!

A mi esposo Andrés, gracias por acompañarme en esta aventura, por animarme a continuar día a día, por compartir tantos momentos difíciles y alegres. Solo tú sabes lo que hemos tenido que sacrificar para que esta tesis llegara a buen término. Te amo!!!

RESUMEN

Uno de los objetivos principales de la videovigilancia inteligente es la detección de personas para tratar de entender, aprender y reconocer sus comportamientos normales y anormales. Con este propósito, en esta tesis se presenta un sistema de videovigilancia que tiene la capacidad de describir los objetos presentes en imágenes reales de escala pequeña y de baja calidad, en ambientes de exteriores; con la finalidad de localizar e identificar la presencia de seres humanos y detectar trayectorias con comportamientos diferentes.

En esta tesis se propone la modificación y mejora de dos técnicas. La primera de ellas es un detector local alternativo a los actuales, capaz de extraer características locales y representar a los objetos presentes en regiones pequeñas. La técnica propuesta se denominó detector local GSIFT ya que se basa en la combinación de los filtros Gabor y el descriptor local SIFT. La segunda técnica propuesta es un algoritmo de agrupamiento denominado pamTOK (pam Tree Out K), el cual estima de manera automática el número de categorías en que es conveniente separar el conjunto de datos analizado; encontrando los modelos correspondientes a un comportamiento normal, para detectar comportamientos anormales.

El rendimiento del detector local propuesto GSIFT se evaluó en relación a la extracción de características y su poder de descripción, en una tarea compleja como lo es la detección de la figura humana en imágenes con escala pequeña. Los resultados son validados y replicados en cuatro conjuntos de entrenamiento y prueba diferentes de ambientes urbanos no controlados, y comparados con seis de las principales técnicas de descripción local. Como soporte a las características locales extraídas, se llevó a cabo una descripción holística de los puntos de interés detectados mediante los momentos de Hu. Los resultados obtenidos de la fusión de ambos descriptores son mejores a los logrados de manera individual por la descripción local y global. Finalmente, los resultados alcanzados muestran que el detector local GSIFT es una buena opción para realizar la detección local de puntos de interés, de manera estable y robusta, en regiones con un número de píxeles menor a 30.

El algoritmo de agrupamiento pamTOK, desarrollado en esta tesis, estima el número de grupos en que es conveniente separar el conjunto de datos, mediante la especificación de un índice de agrupamiento que evalúa la relación entre homogeneidad interna de los grupos y su distancia con respecto a los otros clústeres; sin estar limitado a un número de categorías específicas. La evaluación se realizó con ocho bases de datos públicas y los resultados obtenidos demostraron que el algoritmo de agrupamiento en combinación con la distancia *Longest Common SubSequence* (LCSS) tiene un buen rendimiento; permitiendo la detección de comportamientos anormales.

Palabras clave: Videovigilancia, detector local, mejora del SIFT, filtros Gabor, detección de comportamientos sospechosos, algoritmos de agrupamiento, estimación del k óptimo.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Capítulo 1

Introducción

1.1 Introducción

En la última década, la investigación en videovigilancia se ha convertido en una importante temática en el área de Visión por Computador, debido a la necesidad de remplazar la tradicional videovigilancia pasiva por sistemas automáticos de análisis de video y a la demanda del reconocimiento automático de las actividades humanas.

De acuerdo con (Gowsikhaa, 2014) y (Ke et al. 2013), la videovigilancia es un acto de seguridad para monitorear áreas y comportamientos con ciertas características tales como aeropuertos, control de acceso en áreas especiales, vandalismo, peleas, traspaso de fronteras, flujo de personas, análisis de comportamientos anormales o sospechosos, reconocimiento de la actividad humana, sistemas de protección a los peatones, etc. Sin embargo, también estos sistemas pueden ser utilizados en sistemas de seguridad y asistencia médica en casa tales como monitorear niños, gente grande o pacientes en hospitales. Otra clase de uso es para analizar patrones de movimiento en ambientes de entretenimiento como en las actividades deportivas.

Entonces, estos sistemas de reconocimiento de las actividades humanas en video, envuelven aplicaciones de reconocimiento con una sola persona, la interacción de múltiples personas, el comportamiento de muchedumbres y la detección del reconocimiento de actividades diferentes o anormales.

No obstante, la actividad fundamental para la gran variedad de aplicaciones antes mencionadas es contar con un proceso de detección e identificación robusto de la figura humana. Definiendo la detección de personas, de acuerdo con (Barbu 2014), como la

tarea de identificar la presencia de humanos y diferenciar humanos de objetos no humanos en secuencias de video.

La detección de personas es una tarea compleja por diversos factores como la posición de la cámara de vídeo, la distancia de la cámara, la variabilidad en la apariencia de los seres humanos, la amplia gama de posturas y movimientos del cuerpo, las variaciones en el brillo, la intensidad luminosa (alta o escasa), los niveles de contraste o fondos, el número de personas, oclusiones y entornos no controlados que contienen grandes cantidades de información no relacionada que obstaculizan la tarea de detección (Barbu 2014).

1.2 Motivación

Revisando la literatura del área, se detectó que un aspecto poco tratado en la detección de peatones, es el trabajar con imágenes de baja resolución y con objetos cuyas regiones que los contienen son de tamaño pequeño; por ejemplo, en los sistemas de ayuda a la conducción, la detección e identificación de humanos a distancia es de alta prioridad, ya que dicha imágenes pueden corresponder a peatones lejanos al vehículo, primer objetivo para evitar una colisión (Pedersoli et al. 2014). (Dollar et al. 2012) menciona precisamente la necesidad de integrar en las investigaciones, imágenes de pequeñas escalas, en el rango de 30x80 pixeles, ya que la mayoría de los trabajos reportados se enfocan en regiones de peatones sobre los 100 pixeles.

Se puede argumentar que en los últimos años, gracias a los avances tecnológicos, se cuenta con cámaras de video de alta definición y su uso se ha incrementado en sitios públicos o privados para el monitoreo automático de los videos. Esto ha provocado que los esfuerzos de investigación en el desarrollo de sistemas de videovigilancia, se realice con imágenes de alta resolución, desarrollando técnicas y métodos que consideran a priori contar con la información suficiente para reemplazar o apoyar a los tradicionales sistemas de vigilancia pasiva, automatizando de manera robusta, la identificación y descripción de actividades normales o inusuales (con fines de seguridad y protección) e interesantes (para centros comerciales, eventos deportivos, etc.).

Sin embargo, la realidad es que muchos sistemas de videovigilancia ven afectado su rendimiento debido a que la adquisición de los videos se realiza a grandes distancias, la cámara tiene una perspectiva amplia de la escena; o los sistemas de monitoreo no tienen el equipamiento suficiente para la recepción de la información; es decir, la Ethernet no cuenta con el ancho de banda suficiente para la transmisión masiva de imágenes de alta calidad, o por la compresión de datos utilizada para el envío de las imágenes. Esto provoca analizar videos con menor resolución y con regiones pequeñas que generan una carencia de pixeles por objeto de interés. Esto significa que, si un objeto, en una imagen tiene un tamaño pequeño o baja resolución, la descripción del mismo no puede realizarse de manera apropiada con las técnicas de extracción de características actuales, incluso en algunos casos, el objeto puede llegar a perderse, al no poder ser descrito y con ello no poder realizar la tarea de detección y observación de su comportamiento.

Por ello, en la presente tesis se decidió realizar un sistema de videovigilancia que permita trabajar con imágenes de escala pequeña y baja calidad. El sistema está diseñado para el uso de imágenes adquiridas en largas distancias, mediante una cámara estacionaria en un ambiente externo para la detección de personas, el aprendizaje y reconocimiento de sus actividades.

1.3 Objetivos de la tesis

El objetivo general de esta tesis es diseñar e implementar un sistema de video vigilancia con capacidad de utilizar imágenes pequeñas y de baja calidad, en ambientes exteriores, con la finalidad de localizar e identificar la presencia de seres humanos y detectar trayectorias con comportamientos diferentes.

Objetivos Específicos:

1. Analizar el estado del arte en las áreas de video vigilancia, detección de peatones y detección de trayectorias anormales.
2. Detectar de manera robusta las regiones en movimiento, en videos de escenarios externos.

3. Proponer un algoritmo de detección local de puntos de interés para su uso en imágenes de escala pequeña y comparar su desempeño con los métodos de detección local más representativos.
4. Adquirir una base de imágenes en un entorno real, en un ambiente externo, cuya característica principal sea contener regiones de peatones en escala pequeña y de baja calidad.
5. Seleccionar del estado del arte, bases de datos públicas en ambientes externos, integradas de imágenes cuya ventana, que contenga al peatón, sea pequeña.
6. Revisar las distintas metodologías de detección de trayectorias anormales reportadas en el área y proponer un algoritmo de agrupamiento capaz de detectar trayectorias anómalas de peatones en ambientes externos.
7. Proponer un algoritmo de agrupamiento que de manera automática estime el número de categorías presentes en los datos analizados.
8. Seleccionar de la literatura, bases de trayectorias para evaluación del algoritmo de agrupamiento propuesto.
9. Identificar métricas que permitan mostrar la efectividad y eficacia de los algoritmos propuestos.
10. Elaboración de experimentos de detección de peatones y de detección de trayectorias anormales.

1.4 Metodología de solución propuesta

El presente trabajo está integrado por dos grandes etapas, como se puede observar en la figura 1.1. La primera etapa, tiene como objetivo, detectar las regiones en movimiento e identificar, cuáles de ellas corresponde a personas y cuáles a otro objeto. En esta fase se propuso un detector local denominado GSIFT (filtros Gabor en combinación con SIFT) que permite extraer características locales en imágenes de escala pequeña y baja resolución, para describir de manera robusta a los objetos contenidos en ellas.

La segunda sección se encarga analizar las trayectorias para detectar los comportamientos inusuales. Se trabaja bajo el supuesto de que no se cuenta con información para realizar dicho análisis. En esta fase se propuso un algoritmo de agrupamiento denominado pamTOK (PAM Tree Out K), el cual estima el número de

categorías presentes en el conjunto de datos, encontrando los modelos correspondientes a un comportamiento normal y detectar los recorridos anormales.

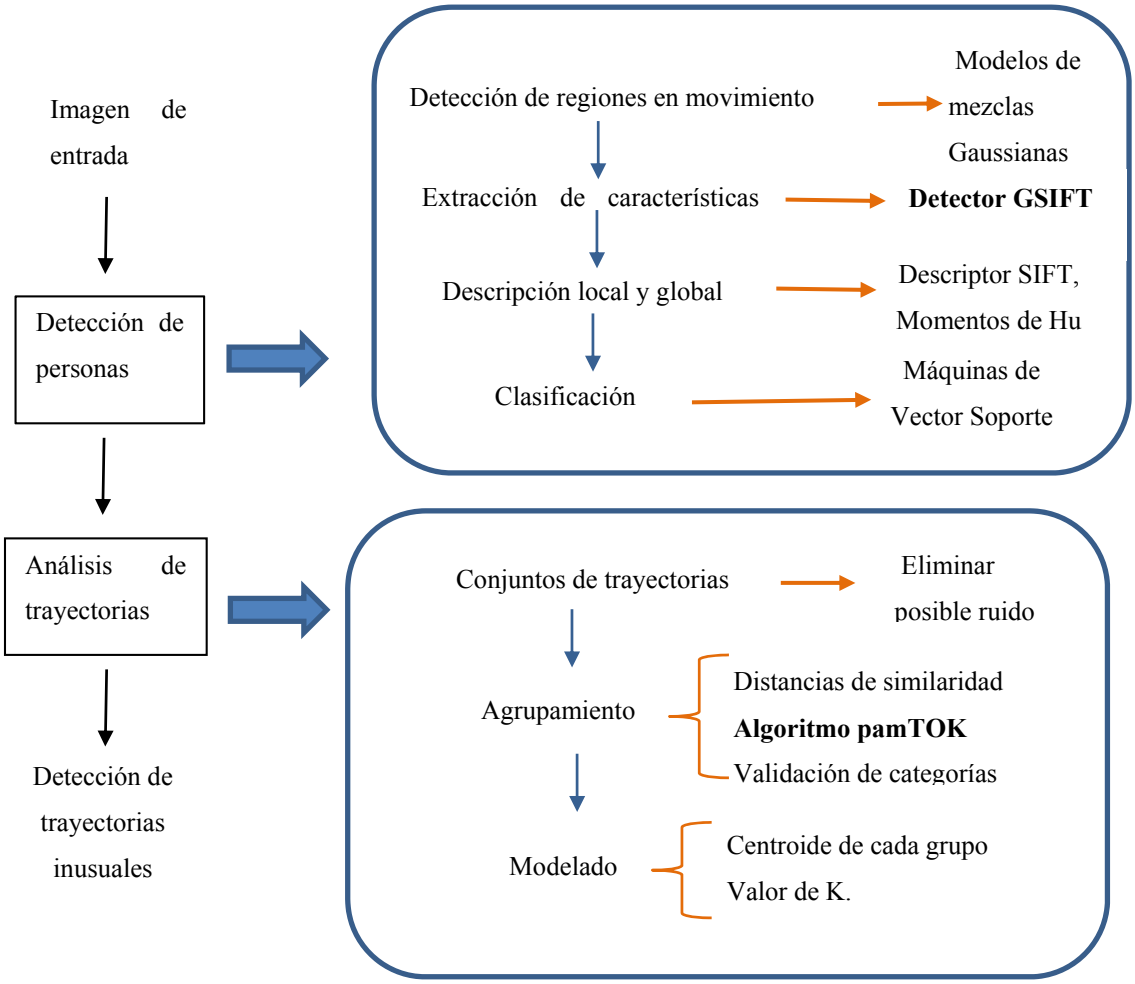


Figura 1.1 Esquema de la metodología de solución propuesta.

El módulo de la detección de personas se implementó bajo la plataforma OpenCV v2.4.8 (Bradski & Kaehler 2008), excepto el algoritmo de clasificación, lo que permitió evaluar las técnicas bajo una misma plataforma de desarrollo y eliminar el ajuste de los algoritmos. Las máquinas de vector soporte y el módulo de análisis de trayectorias se desarrollaron en el lenguaje R (Ihaka 1998). El lenguaje R es un entorno de programación que proporciona una amplia variedad de librerías para la manipulación de datos, el análisis estadístico y la creación de gráficas de alta calidad. Además, es multiplataforma y es distribuido bajo licencia GNU GLP.

A continuación, se explica la metodología seguida. La explicación en detalle de los algoritmos propuestos y aplicados se encuentra en los siguientes capítulos.

a) Detección de peatones

- *Detección de regiones en movimiento*:

El primer paso es contar con un buen modelado del fondo para localizar de manera precisa las regiones en movimiento. En este proyecto, el modelado del fondo y detección del objeto en primer plano se realiza con un modelo multimodal de mezclas gaussianas, GMM por sus siglas en inglés, (Stauffer & Grimson 1999). A cada región de movimiento se calcula el rectángulo mínimo y se pasa a niveles de gris. Para los procesos posteriores, estas regiones se consideran como la imagen de trabajo.

- *Extracción de características*:

Se propone un detector de puntos característicos locales denominado GSIFT basado en la transformada SIFT y en filtros Gabor. Este descriptor combina las ventajas del descriptor SIFT de ser invariante a la escala, rotación y a pequeños cambios en la iluminación y de los filtros Gabor proporcionar una completa representación de la imagen. A partir de los puntos de interés detectados por cada transformada, se crean imágenes binarias, y se extraen los momentos invariantes de Hu para contar con una descripción holística que capture la información de la estructura geométrica (forma) en el dominio espacial de los objetos a reconocer.

- *Clasificación de objetos:*

La clasificación se realizó mediante Máquinas de Vector Soporte (Cortes & Vapnik 1995) utilizando un kernel *Radial Base Function*. El sistema tiene como salida solo dos valores, "si o no" se detecta un peatón en la imagen. La decisión final de pertenencia, considera la respuesta de la descripción local, así como de la holística. La fusión de los resultados se realizó a nivel de *score*, mediante la media aritmética de los valores de predicción obtenidos por cada uno de los descriptores por separado. Como resultado de esta actividad se cuenta con 56 modelos de clasificación (7 técnicas y 4 bases de datos, utilizando características locales y globales). Posteriormente, se evaluaron los modelos con los conjuntos de prueba, los cuales son conjuntos disjunto de los considerados en el entrenamiento.

- *Evaluación:*

El método propuesto GSIFT fue evaluado con cuatro bases de datos en ambientes externos no controlados, que presentan diferentes condiciones como puntos de vista,

cambios de escala, ángulo, iluminación, imágenes pequeñas y de baja calidad. Dos pertenecen al *Computer Vision Center*: CVC-01 (Gerónimo et al. 2007) y CVC-02 (Gerónimo, Sappa, et al. 2010). La tercera base de datos considerada es la CBCL de peatones (Poggio 2000). Con la finalidad de contar con una base de datos propia que cumpliera con los requisitos del presente trabajo y gracias al apoyo del Instituto Tecnológico de Cuautla se realizó la adquisición de una base de datos propia denominada ITC.

Para comparar el desempeño el detector de puntos propuesto, se seleccionaron seis de las principales técnicas de descripción local: SIFT, SURF, FAST FREAK, ORB y BRISK. Se tienen resultados para cada transformada en cada base de datos. La evaluación se visualiza a través de curvas ROC para mostrar la bondad de las técnicas locales de detector y descriptor de la clasificación. Se cuenta con curvas ROC para cada base de datos.

b) Análisis de las trayectorias

El análisis de las trayectorias es una valiosa fuente de información para identificar, de manera automática, comportamientos específicos o sospechosos, llevado a cabo por los objetos de interés. Este trabajo, modela al conjunto de recorridos mediante la propuesta de un algoritmo de clasificación no supervisada, y la detección de anomalías se trata como el problema de encontrar patrones (outliers), en los datos, que no se ajustan al comportamiento esperado.

- *Medidas de distancia:*

Para este proyecto se determinó emplear medidas de similaridad que fuesen independientes a la longitud de las trayectorias, y se implementó la medida Dynamic Time Warping, DTW, (Rabiner & Juang 1993). Sin embargo, esta medida tiene el problema de la alta complejidad computacional al ser O(n2), por lo que se decidió utilizar también la medida Longest Common Subsequence, LCSS, (Vlachos, 2002) que asigna pesos a diferentes puntos. Distance with Real Penalty, ERP, propuesto por (Chen & Ng 2004) y la distancia Edit Distance on Real Sequence, EDR, (Chen et al. 2005). Como resultado de esta actividad se obtienen matrices de distancia, que es la información de entrada para analizar la semejanza o similitud de dichas trayectorias, mediante técnicas de agrupamiento.

- *Agrupamiento:*

Un problema común con el enfoque no supervisado es conocer a priori el número de grupos o clases del conjunto de datos y hacer algunas suposiciones acerca de la distribución de los datos. El presente proyecto, propone un algoritmo de agrupamiento denominado PAMTOK (PAM Tree Out K), basado en el algoritmo pam (Kaufman & Rousseeuw 1987) y el índice Dunn (Desgraupes 2013). El algoritmo estima el número óptimo de categorías, para agrupar las trayectorias de manera automática a través del valor estimado de *K*. Como salida, proporciona el valor de pertenencia de cada trayectoria a su clúster, los objetos representativos de cada grupo (medoides). Las trayectorias con una pertenencia y/o frecuencia mínima son consideradas como inusuales o anormales.

- *Evaluación:*

El método de aprendizaje fue evaluado con ocho bases de datos de trayectorias públicas. Seis bases de datos pertenecen al grupo *Computer Vision and Robotics Research Laboratory*, CVRR (Morris & Trivedi 2009a). BARD (Cancela et al. 2013) y Edimburgo (Majecka 2009).

El desempeño del algoritmo propuesto se comparó con un algoritmo de agrupamiento que pretenden estimar un valor de *K* de manera automática, denominado pamk (Christian Hennig 2014). Los resultados se exponen mediante gráficas que muestran el agrupamiento de las trayectorias, el medoide de cada grupo y las trayectorias consideradas como inusuales.

1.5 Estructura del documento

Este documento está estructurado de la siguiente manera:

El Capítulo 1 detalla la motivación del presente trabajo, los objetivos y una breve descripción de la metodología seguida.

En el Capítulo 2 se realiza un análisis de la literatura referente a la etapa de extracción y representación de la figura humana. Se revisan las técnicas de descripción local consideradas en la evaluación del detector GSIFT. Posteriormente, se analizan los

trabajos relacionados al área de análisis de trayectorias para el modelado de las actividades normales y la detección de los recorridos inusuales.

El Capítulo 3 describe las bases de datos públicas seleccionadas para la etapa de detección automática de la figura humana y posteriormente, se describen los conjuntos de datos utilizados en el análisis de trayectorias, para la detección de comportamientos normales e inusuales.

El Capítulo 4 detalla el detector de características locales propuesto GSIFT, así como también explica los objetivos y metodología de la experimentación realizada. Se presenta una comparativa del desempeño del algoritmo GSIFT con respecto a seis técnicas de descripción local clásicas. Se discuten los resultados obtenidos y se muestran de manera visual mediante gráficas ROC.

El Capítulo 5 expone el algoritmo de agrupamiento pamTOK para analizar y agrupar las trayectorias presentes en los conjuntos de datos. Se detalla la experimentación realizada con ocho bases de trayectorias de ambientes urbanos. Finalmente se realiza un análisis cuantitativo de los resultados obtenidos.

Finalmente, el Capítulo 6 resume las conclusiones del desarrollo de este proyecto, lista las contribuciones de esta tesis y presenta propuestas de trabajos futuros.

Capítulo 2

Trabajos relacionados

2.1 Introducción

La investigación en video vigilancia inteligente ha sido intensa en los últimos años, buscando obtener información de alto nivel con el fin de apoyar en la interpretación de una escena o en la detección de eventos anormales. Para ello, una etapa decisiva, consiste en tener un módulo robusto de detección de persona que permita obtener información de las trayectorias para reconocer, automáticamente, comportamientos específicos o sospechosos.

En el presente capítulo se introduce el tema de la videovigilancia, profundizando en la revisión de la literatura de dos grandes etapas de los sistemas de análisis de video inteligente. La primera parte expone trabajos sobre la extracción de características y representación de personas para su detección automática en imágenes. La segunda parte del capítulo, se integra de la revisión de trabajos relacionados al análisis de trayectorias y detección de comportamientos anormales o inusuales.

2.2 Sistemas de videovigilancia

En la literatura del área es posible encontrar excelentes revisiones de las diversas propuestas de métodos y algoritmos desarrollados para la realización de sistemas de monitoreo inteligente. Entre ellas, (Hu et al. 2004) realiza una detallada exposición sobre las distintas etapas de los sistemas de videovigilancia. (Moeslund et al. 2006) revisa los avances en el análisis y captura del movimiento humano. (Morris & Trivedi 2008) presenta una revisión basada en el análisis de la trayectoria. (Enzweiler & Gavrila 2009) revisa las técnicas relacionados a la detección de peatones en sistemas monoculares. (Chandola et al. 2009) analiza las principales técnicas en la detección de anomalías, para la detección de comportamientos normales y anormales. (Poppe 2010)

analiza el reconocimiento de las acciones humanas. (Gerónimo, Lopez, et al. 2010) presenta un estudio sobre la detección de peatones para sistemas de asistencia al manejo. (Weinland et al. 2011a) se concentra en representación de la acción, su segmentación y reconocimiento. (Liu et al. 2013) presenta los avances en términos de hardware, software y sus aplicaciones en sistemas inteligentes de video. (Ke et al. 2013) proporciona una revisión de los principales etapas de los sistemas de reconocimiento de actividades, desde el procesamiento de bajo nivel, nivel medio hasta las aplicaciones de alto nivel. (Gowsikhaa et al. 2014) revisa los trabajos relacionados al seguimiento y reconocimiento del movimiento humano en videos de video vigilancia.

De acuerdo con (Ke et al. 2013) y (Hu et al. 2004) un sistema de videovigilancia completo se integra, idealmente, de las siguientes etapas jerárquicas Sin embargo, dependiendo de la aplicación específica del sistema de monitoreo inteligente que se esté desarrollando, es necesario profundizar en alguna etapa en particular o no considerar todas las fases listadas.

1. Procesamiento de bajo nivel:

El objetivo central de este primer módulo es segmentar, identificar y seguir el objeto bajo análisis. Se integra de actividades como el modelado del fondo o entorno, detección del movimiento, extracción de características y representación, clasificación de los objetos en movimiento y seguimiento de dichos objetos.

a) *Modelado del entorno:* El propósito de esta etapa es modelar la escena sin los objeto de primer plano, para posteriormente, detectar los pixeles pertenecientes al primer plano o al objeto en movimiento en cada fotograma. Para recuperar y actualizar el fondo de la imagen, estos métodos deben tener la habilidad de actualizarse y reflejar los cambios del entorno. Hay diversos algoritmos para resolver este problema como el promedio temporal de la secuencia de imágenes (Koller et al. 1994), la estimación gaussiana adaptativa, modelos de mezclas guassianas (Permuter et al. 2006), algoritmo Wallflower (Toyama et al. 1999), modelo estadístico, modelo de fondo adaptativo para reducir la influencia de sombras y señales de color poco confiables (McKenna et al. 2000), entre otros.

b) *Detección de movimiento*: Esta etapa proporciona un foco de atención. Ayuda a detectar las regiones correspondientes a los objetos (personas, vehículos, maletas) de interés. Algunas de las propuestas para tratarlo es la resta de fondo (Cucchiara et al., 2003) técnica simple y eficiente en fondos estáticos, pero sensitiva a cambios en la escena, e iluminación; diferencia temporal entre fotogramas (Lipton et al. 1998) usada para ambientes dinámicos pero pobre para extraer pixeles relevantes; flujo óptico, caro computacionalmente y no recomendable para sistemas en tiempo real; y modelado con mezclas gaussianas (Stauffer & Grimson 1999), es el método más utilizado por sus buenos resultados en ambientes dinámicos.

c) *Extracción de características*: etapa fundamental para capturar información del objeto o región, considerando el tamaño, forma, silueta, color, movimiento, etc. Para la representación del cuerpo humano, se han usado técnicas que realizan la descripción completa de la imagen como la transformada de Fourier (Kumari & Mitra 2011), utilizan descriptores locales como SIFT (Scovanner et al. 2007), HOG (Barbu 2014) (Pedersoli et al. 2014); combinan información espacial y temporal (Javan Roshtkhari & Levine 2013). Otra propuesta es el uso de modelos, plantillas o estructuras que representen el cuerpo humano como siluetas (Hu et al. 2004), plantillas (Nguyen et al. 2009), representación de las diferentes posturas de un peatón a través de un grafo (Wang et al. 2011), o con descriptores locales (Dalal & Triggs 2005a).

d) *Clasificación de objetos y/o identificación de humanos*: Este proceso permite identificar al objeto de interés, cuando la escena contiene objetos diferentes al analizado. La selección del clasificador depende de la representación, de las características, utilizada. Sin embargo, se han usado modelos como redes neuronales, correspondencia de plantillas, métodos basados en movimiento (muy sensibles al ruido), máquinas de soporte vectorial (SVM), modelos ocultos de Markov (HMM), entre otros.

e) *Seguimiento de objetos:* El objetivo de esta etapa es obtener información de la trayectoria del objeto, para determinar su consistencia temporal, e incluso, si se desea un mayor análisis, determinar si un objeto es nuevo en la escena o ya ha sido visto y etiquetado. Los métodos de seguimiento se dividen de acuerdo a la

representación utilizada en: regiones, contornos activos, en características globales o locales; en grafos de dependencia y seguimiento basado en modelos. Las técnicas tradicionales en esta etapa son el filtro Kalman, filtro de partículas o de condensación, Redes neuronales, redes bayesianas dinámicas, método geodésico, entre otras.

2. *Procesamiento de alto nivel:*

Las técnicas de procesamiento de alto nivel incluyen fases que hacen uso de la semántica, el contexto de días y horas de los videos para poder modelar los comportamientos, acciones, eventos o interacciones de los agentes (humanos, animales, carros, etc.) bajo estudio. Este nivel envuelve también el análisis y reconocimiento de actividades para la detección de trayectorias como normales e inusuales. De las propuestas más populares para representar el comportamiento son los modelos ocultos de Markov (HMMs), redes bayesianas (BNs), o sus variantes tales como Coupled Hidden MArkov Model (CHMMs), Variable Length Markov Models (VLMMs), cyclic HMMs. Dinamic time warping (DTW), altruistic vector quantization (AVQ), autómatas finitos no determinísticos (NFA), redes neuronales autoorganizativas, filtros basados en wavelets analizados por modelos de mezclas gaussianas y agrupados por kmeans, entre otras.

A continuación se presenta una revisión de trabajos relacionados a la extracción de características y descripción de personas.

2.3 Representación de la figura humana

En la literatura del área se puede encontrar varias propuestas de extracción de características para la representación de la figura humana en 2D y 3D, una excelente revisión se puede encontrar en (Weinland et al. 2011a). Moeslund (Moeslund et al. 2006) agrupa en dos categorías los métodos aquellas que extraen información a través de ventanas y aquellas que utilizan modelos, plantillas o patrones de la figura humana. (Pedersoli et al. 2014) propone abordar la detección de peatones con dos principales familias de métodos, a través de modelos o plantillas de correspondencia y a través de bolsa de palabras (técnicas que necesitan un paso de muestreo o cuantificación después de la etapa de extracción de características). Otra manera de agrupar las técnicas de representación del cuerpo humano, es considerando si la descripción se realiza de

manera local (observaciones como una colección de regiones independientes) o global (codificando las observaciones visuales como un todo) (Poppe 2010), (Ke et al. 2013).

Realizando un análisis de los métodos, es posible decir que dichas categorías son similares. Las representaciones globales son derivadas del uso de modelos, plantillas o siluetas los cuales buscan determinar las propiedades de la figura humana como localización, dirección o pose. Estos métodos son sensitivos al ruido, oclusiones parciales y a cambios en puntos de vista. Buscando eliminar estos problemas, existen propuestas que dividen la imagen en celdas (malla) donde cada subregión describe localmente a la imagen. Sin embargo, la clasificación se realiza con una representación conjunta de todas las partes.

La tabla 2.1 muestra un resumen de trabajos relacionados a la detección de la figura humana. Cada artículo esta ordenado de manera ascendente de acuerdo al año que fue publicado. El resumen da información sobre la aportación de cada trabajo, la técnica de extracción de características aplicada y el tamaño de las imágenes utilizadas.

Tabla 2.1 Resumen de trabajos relacionados a la detección personas.

Autor	Objetivo	Aportación	Técnica de representación y clasificación	Imágenes de trabajo	Comentarios
(Wojek & Schiele 2008)	Detección de personas	Evaluación de diferentes representaciones de características.	Haar wavelets, Haar-like features, HOG, Shapelets, shape context. Clasificadores: MSV, AdaBoost y árboles de decisión.	INRIA	Buen análisis de la comparativa de técnicas. No da información sobre las características de las bases de datos.
(Sun et al. 2009)	Reconocimiento de acciones	Propone un modelado de las acciones a través de un modelo espacio-temporal	Extrae puntos con SIFT y analiza la transición de los puntos y modela la distribución con cadenas de Markov. Aplica Bolsa de palabras (BoW) y posteriormente MVS.	HOHA y LSCOM	Interesante propuesta de modelar la dinámica del movimiento.
(Wojek et al. 2009)	Detección de personas.	Evaluación de diferentes representaciones de características.	HOG, Haar wavelets y Oriented histograms on flow. Clasificadores: MVS, AdaBoost y MPLBoost	TUD-Brussels con imágenes de 720x576 y 640x480, con regiones de al menos 48 pixeles.	Buen análisis de datos. Menciona que HOG tiene problemas con imágenes de escala pequeña.
(Enzweiler & Gavrila 2009)	Detección de peatones	*Survey* y estudio experimental de técnicas de detección de peatones	HOG con MVS lineal. Haar wavelets con AdaBoost cascade. Redes neuronales con Campos locales receptivos (NN/LRF).	Daimler con imágenes de 18x36 y 48x96.	HOG tiene mejor resultados con 48x96 y Haar con la resolución baja.

Autor	Objetivo	Aportación	Técnica de representación y clasificación	Imágenes de trabajo	Comentarios
(Wang et al. 2009)	Reconocimiento de acciones	Evaluación de características espacio temporales.	Detectores Harris3D, Cuboid, Hessian, Dense sampling, HOG/HOF, HOG3D, ESURF.	KTH, UCF sport actions, Hollywood actions.	Buen trabajo y análisis de datos.
(Gerónimo, Sappa, et al. 2010)	Detección de peatones.	Fusión de características 2D y 3D con imágenes pequeñas.	Haar-wavelets y EOH (edge oriented histograms) con AdaBoost. Estimación del tamaño y límites del peatón en 3D.	CVC02 con imágenes de 13x26 a 103x206	Buenos resultados. Buena revisión de la literatura. Imágenes pequeñas.
(Wang et al. 2011)	Sistema de seguimiento de peatones.	Estructura global de la pose, construyendo un grafo que considera la adyacencia y que cubre las diferentes vistas. Con ángulo y postura se preserva la continuidad.	Aprendizaje múltiple de las diferentes posturas de un peatón, de 0 a 360 grados a través de un grafo de adyacencia (cada 10 grados). Histograma de color. Reducción de dimensionalidad de datos con OLPP.	Daimler Chrysler, de 640x480 y tamaño de peatón de 32x64	Trabajo completo. Proporciona toda la información. Idea original.
(Brehar & Nedevschi 2011)	Detección de peatones.	Estudio comparativo de la representación de Haar y HOG de manera individual y con BoF.	HOG y Haar-like features, en combinación con bolsa de características. Clasificador: AdaBoost.	Daimler con imágenes de 18x36. NICTA con regiones de 16x40 y 64x80.	Concentran sus experimentos a las imágenes de 64x80. Mejor resultado con Haar-like.
(Hsu et al. 2011)	Reconocimiento de acciones humanas.	Proponen un una propuesta denominada NWFE (extracción de características pesadas no paramétrico).	Extraen la silueta del cuerpo, extraen su curvatura, aplican PCA, después BoW para obtener los vectores de histograma. Clasificador de Bayes.	Weizmann	Alta complejidad computacional. Imágenes de alta calidad.
(Guo et al. 2012)	Detección de peatones para sistemas de transporte.	Propone la detección en dos etapas con extracción de características y clasificación en cada una de ellas.	Características Haar-like y AdaBoost para obtener candidatos positivos. Se escalan las imágenes y se obtienen Momentos de Hu y características de niveles de gris con MVS.	Imágenes propias de 16x32 y las re-escalan a 64x128.	Interesante propuesta para trabajar con imágenes pequeñas. Doble clasificador.
(Schaeffer 2012)	Detección de peatones.	Comparativa de dos descriptores nuevos como es FREAK y BRISK con SURF.	Extracción de puntos de interés con SURF y descripción con FREAK, BRISK y SURF. Clasificador: aplica BoW y MVS con RBF.	NICTA con regiones de 64x80.	FREAK obtiene un 91%, BRISK un 87% y SURF un 85%. Tienen problemas con algunas regiones con pocos puntos.
(Ouyang et al. 2012)	Detección de humanos.	Proponen una función de energía potencial de gradiente elástica pesada Comparativa con HOG multinivel y HOG-LBP.	Extracción de características de gradiente con la función de energía potencial de gradiente elástica pesada. Y su combinación con HOG. Clasificador: MVS cascada.	INRIA con imágenes de 96x160 y 64x128.	Extracción de datos interesante. Y al final la combina con HOG. Imágenes de buen tamaño.
(Sermanet et al. 2013)	Detección de peatones	Propuesta de aprendizaje de las características en múltiples capas y con aprendizaje no supervisado.	Extracción de características en capas de filtros de diferente tamaño. Aplica un CPSD (convolutional predictive sparse decomposition). Clasifican con una regresión lineal.	INRIA Daimler ETH Caltech TudBrussels	Interesante y buen trabajo. Propuesta diferente. Compara con varios tamaños de imágenes.

Autor	Objetivo	Aportación	Técnica de representación y clasificación	Imágenes de trabajo	Comentarios
(Aminian Modarres Amir Farid & Soryani 2013)	Reconocimiento de acciones humanas.	Representación basada en un nuevo grafo de la silueta del cuerpo, por medio de funciones de base elípticas, lo llaman grafo de postura corporal.	Grafo de la silueta del cuerpo humano con funciones de base elípticas. Usa vértices y bordes y corresponde al esqueleto real de la silueta. Modelos Ocultos de Markov.	Bases de datos: SINICA Academia, KTH y UCF (720x480).	Propuesta de modelado del cuerpo novedosa. Imágenes de alta definición.
(Conde et al. 2013)	Detección de personas.	Presentan un nuevo de método de caracterización denominado HOGG y su comparativa con métodos alternativos.	Método HOGG basado en HOG y los filtros Gabor.	Pets 2006, Pets 2007, Pets 2009 y CAVIAR	Buen trabajo y análisis de datos. Imágenes de alta calidad.
(Flohr & Gavrila 2013)	Segmentación de peatones	Propuesta de segmentación combinando datos de múltiples señales.	Inicializan una plantilla, que combinan con datos de color, textura y adaptan la información a un modelo de formas activa.	Penn Fudan con imágenes de 186x63 a 373x63. Daimler con imágenes de 121x34 a 468x267	Interesante propuesta de detección de peatones. Imágenes de alta calidad con buenos resultados.
(Barbu 2014)	Detección y seguimiento de personas	Método de detección basado en diferencias temporales, piel y HOG.	Correspondencia de plantillas basado en HOG y características de piel.	No proporciona información.	Proporciona porcentajes de desempeño sin información sobre los datos.
(Dou & Li 2014)	Reconocimiento de peatones.	Propuesta de fusión de descripción local y global.	Puntos de interés con SIFT 3D. Extrae MHI y MEI (imágenes de energía e histórico de movimiento) y aplica momentos de Hu. Clasifica con GMKL.	KTH con imágenes de 160x120 y Weizmann con imágenes de 180x44.	Buen trabajo, imágenes de alta resolución. Fusión de información local y global.
(Pedersoli et al. 2014)	Detección de peatones	Representación multi-resolución. Trabaja con imágenes de tamaño pequeño. Adaptación de extracción de características en tiempo real.	Proponen un modelo de plantilla deformable con una pirámide de características HOG en múltiples resoluciones. Forma un árbol jerárquico donde cada conexión es un modelo de partes. Clasificación con MVS.	CVC02: con imágenes de peatones de 70x140 a 12x24 pixeles. INRIA	Buen trabajo, presenta datos de desempeño y velocidad, compara su técnica con otras de acuerdo a lo reportado en la literatura.
(Newlinshebiah et al. 2015)	Reconocimiento de interacciones humanas.	Presenta casos de experimentación.	Descriptor SURF, aplica BoW y clasifica con MVS.	Base de datos propia. Imágenes de 720x480.	Análisis de datos superficial. Imágenes de alta calidad.

De acuerdo a los trabajos mostrados se puede ver que la extracción de características tienen una influencia crucial en el desempeño de los sistemas de reconocimiento de humanos, por lo cual, es esencial representar las características de los objetos de forma adecuada. La figura 2.1 muestra algunas de las técnicas de detección de personas más novedosas de los artículos de la tabla 2.1.

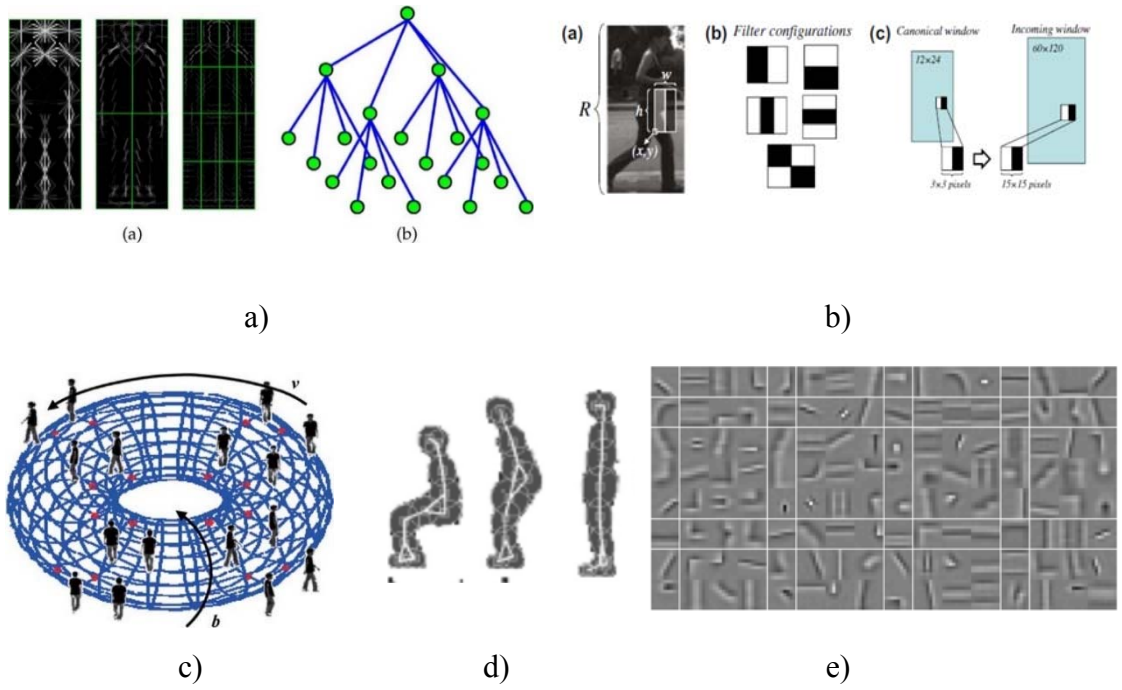


Figura 2.1 Ejemplos extracción de características.
a) Modelo de partes deformable basado en características HOG en múltiples resoluciones, (Pedersoli et al. 2014); b) Haar wavelets con imágenes de tamaño pequeño (Gerónimo, Sappa, et al. 2010); c) Estructura global del grafo de adyacencias de las posturas de peatones en 360°, (Wang et al. 2011); d) silueta del cuerpo humano con funciones elípticas (Amir Farid & Soryani 2013); e) Capas de filtros de diferente tamaño (Sermanet et al. 2013).

Las técnicas preferidas para la representación de la figura humana es HOG, Haar wavelets, Haar-like y sus combinaciones por sus buenos resultados. En la tabla 2.1 se puede observar que muchas de las investigaciones actuales centran sus desarrollos en mejorar las capacidades de HOG o utilizan esta técnica como base para nuevas e interesantes propuestas. También se puede observar que han sido y son poco exploradas las técnicas de descripción local como SIFT, SURFT, FREAK, etc., a pesar de su buen desempeño en otras áreas del reconocimiento de patrones y ser robustas a la escala, rotación, oclusiones y puntos de vista.

Los trabajos centrados al reconocimiento de acciones humanas utilizan bases de datos con imágenes de alta calidad, generalmente, en ambientes controlados con un número limitado de poses y personas. Incluso algunos artículos que realizan la detección de peatones como asistencia al manejo, utilizan bases de datos integradas por imágenes con regiones de peatones mayores de 60 pixeles, debido a la complejidad que representa la detección de personas en ambientes reales. Los artículos que reportan el uso de

imágenes con tamaños menores a 30 pixeles de altura, mencionan el bajo rendimiento que tienen las técnicas de detección tradicionales como HOG, y existen pocas propuestas para usar regiones con este tamaño.

La detección de peatones una importante área de aplicación la asistencia en manejo y debido a la complejidad que envuelve a estos tipos de sistemas inteligentes, se están desarrollando propuestas de detección mediante el uso de sensores y cámaras especializadas. La Universidad Carlos III de Madrid bajo el marco del proyecto SEGVAUTO (Catalán 2013), ha desarrollado un sistema que detecta la presencia de peatones en la vía, en condiciones de baja visibilidad (conducción nocturna) basado en cámaras térmicas. BMW cuenta con cámaras termográficas capaces de detectar personas a una distancia de unos 300 metros, ver Figura 2.2.a. Volvo con el proyecto Intellisage[1] y Ford en Mondeo[2] cuentan con un sistema de detección de peatones integrado de un radar y una cámara de alta resolución de 180 grados colocados en la parte frontal del vehículo, ver Figura 2.2.b. El radar ayuda a diferenciar el tipo de objeto y la cámara envía información a una base de datos con diferentes formas de peatones para ayudar a distinguir a personas de objetos inmóviles presentes en la carretera como señales. Panasonic está desarrollando un sistema de detección tanto de vehículos como de peatones aproximadamente a unos 40 metros, mediante un radar de onda milimétrica que permite detectar personas bajo circunstancias de baja visibilidad, como puede ser de noche, con lluvia, nieve o niebla densa.

a) b)

Figura 2.2 Ejemplos de sistemas de detección de peatones en: a) BMW b)Volvo.

[1] http://www.volvocars.com/es/acerca-de-volvo/innovaciones/coches-inteligentes
[2] http://www.ford.es/Turismos/NuevoFordMondeo/Safety-and-Security

2.3.1 Técnicas de descripción local

Una amplia variedad de técnicas de detección y descripción local se han desarrollado en las últimas décadas. Algunas de las más reconocidas son: detector de esquinas Harris (Harris & Stephens 1988), detector Harris-Laplace (Mikolajczyk & Schmid 2002), detector Harris-Affine (Mikolajczyk & Schmid 2002), MSER (Matas et al. 2004), LBP (Ojala et al. 2002), SIFT (Lowe 2004), SURF (Bay et al. 2006), FAST (Ed Rosten & Drummond 2006), BRIEF (Calonder et al. 2010), Weber local descriptor (Chen et al. 2010)[9], BRISK (Leutenegger et al. 2011), ORB (Rublee & Rabaud 2011), FREAK (Alahi et al. 2012), BRAND (do Nascimento et al. 2013), entre otras. Sin embargo, SIFT se ha convertido en el más descriptor de referencia debido a sus buenos resultados en multitud de aplicaciones en sistemas de reconocimiento de objetos (Heinly et al. 2012). Además, (Miksik 2012) y (Heinly et al. 2012) mencionan que, de acuerdo a su evaluación, descriptores binarios como BRIEF, ORB y BRISK tienen menores requerimientos de memoria, mayor velocidad de cálculo y proporcionan resultados comparables con los obtenidos por SIFT y SURF.

En la literatura se pueden encontrar diversas mejoras del SIFT, por ejemplo Ke y Sukthankar (Ke & Sukthankar 2004) propusieron el PCA-SIFT (Principal Components Analysis) para la reducción de la dimensionalidad del descriptor; Mikolajczyk y Schmid (Mikolajczyk & Schmid 2005) presentaron el descriptor GLOH (Gradient Location-Orientation Histogram), diferenciándose de éste en la utilización de una rejilla circular en el sistema de coordenadas polares para la creación del histograma de orientaciones de los puntos de interés y la utilización de PCA. Por su parte Morel y Yu (Morel & Yu 2009) propusieron un SIFT invariante a transformaciones afines; Moreno (Moreno et al. 2009) mejoró el descriptor SIFT con filtros Gabor. En (Guo et al. 2010) presentaron el descriptor invariante al reflejo de espejo denominado MIFT, en (Liao et al. 2013) optimizan el descriptor de SIFT para tener un mejor desempeño en tareas de correspondencia y recuperación. Con respecto al color, Abdel-Hakim y Farag (Abdel-Hakim & Farag 2006) introdujeron el CSIFT como un descriptor local invariante al color, y Verma en (Verma et al. 2011) extendió el descriptor SIFT a diferentes espacios de color.

Estas mejoras muestran que la comunidad científica sigue interesada en el SIFT debido a que sigue siendo uno de los extractores de características locales más atractivos para

usos prácticos y es uno de los más exitosos en las últimas décadas por su alto poder descriptivo e invariante a las trasformaciones geométricas ya mencionadas.

Precisamente por su buen desempeño encontrado en el estado del arte, las técnicas consideradas en este trabajo, para evaluar y comparar el desempeño de nuestra propuesta de descripción GSIFT, son tres técnicas ya conocidas y evaluadas como son SIFT, SURF y FAST y tres propuestas recientes de descriptores binarios como lo son ORB, BRISK y FREAK. A continuación se discuten brevemente estas técnicas y la tabla 2.2 resume sus principales características.

Tabla 2.2 Técnicas de detección y descripción local consideradas como punto de referencia en la evaluación del detector local GSIFT.

Técnica	Detector	Descriptor	Características	Desventajas
SIFT (Lowe 2004)	Espacio de Escalas Gaussiano. Detección de puntos con DoG.	Asignación de orientación. Descriptor de los puntos característicos	Robusto a escala, rotación, y a pequeños cambios en puntos de vista e iluminación.	Alto costo computacional. Vector de descripción de 128.
SURF (Bay et al. 2006)	Espacio de escalas Gaussiano mediante la Matriz Hessian. Matrices integrales.	Asigna la orientación, mediante Ondas de Haar.	Velocidad de cálculo. Vector distintivo.	Sensible a cambios en puntos de vista. Vector de tamaño 64, 36 y 128.
FAST (Ed Rosten & Drummond 2006)	Detecta un punto como esquina. Evalúa los 16 pixeles alrededor del punto de manera circular. Los evalúan considerando su intensidad mediante un umbral.	No tiene	Velocidad de cálculo.	Sensible al ruido. No robusto a la orientación ni escala. Necesita una técnica de descripción de sus puntos. Detección de características cercanas.
ORB (Rublee & Rabaud 2011)	Detector de características FAST, considerando la orientación.	Aplican un método de dirigido al descriptor BRIEF de acuerdo a la orientación de los puntos de interés.	Robusto a la orientación. Robusto al ruido. Velocidad de cálculo. Vector de tamaño 32.	Sensible a la escala. Vector de tamaño 32.
BRISK (Leutenegger et al. 2011)	Detector AGAST en un espacio de escalas continúo.	Se realiza un muestro circular alrededor de punto de interés. A partir de esta información se construye el descriptor.	Robusto a escala y rotación y ruido. Bajo costo computacional. Velocidad de cálculo.	Sensible a cambios de puntos de vista. Vector de tamaño 64.
FREAK (Alahi et al. 2012)	No tiene	Mejora del descriptor BRISK mediante un *kernel* de grueso a fino.	Velocidad de cálculo mejorada y vector de descripción.	Necesita una técnica de detección de puntos. Vector tamaño de 64.

La técnica base de todas estas técnicas es el descriptor Scale Invariant Feature Transform (SIFT) desarrollado por (Lowe 2004) es un algoritmo capaz de detectar puntos característicos estables. Es uno de los detectores de puntos más utilizado ya que sus puntos están bien localizados tanto en el dominio espacial como en la frecuencia, lo cual favorece la detección de objetos incluso en el caso que estén parcialmente ocluidos,

que aparezcan sobre un fondo muy variable o ante la presencia de ruido. El algoritmo SIFT se compone de cuatro etapas denominadas por Lowe como: 1) Detección de puntos extremos en el espacio de escalas. 2) Localización de puntos característicos. 3) Asignación de orientación. 4) Descriptor de los puntos característicos. Las dos primeras utilizadas para identificar los puntos de interés y las dos últimas para crear el descriptor, de 128 dimensiones, de cada uno de los puntos clave detectados. Estas etapas se explican en detalle en el capítulo 4.

SURF o Speeded-Up Robust Features fue desarrollado por (Bay et al. 2006), se basa en las etapas propuestas por Lowe y queriendo mejorar la velocidad de cálculo de SIFT, considera la creación del espacio de escalas mediante el determinante de la matriz Hessian y aceleran el proceso con imágenes integrales, denominándolo detector Hessian Rápido. Para la creación del descriptor obtiene la orientación dominante de cada punto de interés y calculan las Ondas de Haar en la dirección vertical y horizontal. Posteriormente, determinan el ángulo en que las ondas de Haar han obtenido una mayor respuesta. El vector de características de los puntos se calcula en base a esta orientación dominante. Por defecto, el vector de descripción SURF tiene 64 elementos. Para su uso en tiempo real se recomienda utilizar un vector de 36 de los componentes principales del vector de 64 (se realiza un análisis PCA sobre un gran conjunto de imágenes de entrenamiento). Mientras que la versión extendida tiene 128 dimensiones.

En la búsqueda de poder realizar la detección de puntos en tiempo real, (Rosten & Drummond 2006) propone al detector FAST (Features from Accelerated Segment Test). FAST fue uno de los detectores de puntos más rápido para encontrar puntos clave y hacer la correspondencia de dichas características visuales en sistemas en tiempo real, por ello fue utilizado para aplicaciones de seguimiento en paralelo y mapeo. Es compacto y robusto a la escala. Pero no ha tenido mucho éxito por la ausencia del componente de orientación.

ORB Oriented FAST and Rotated BRIEF, propuesto por (Rublee & Rabaud 2011) como su nombre lo indica es un detector de características desarrollado a partir del detector de puntos FAST y del descriptor BRIEF (Calonder et al. 2010). Rubble extendió las capacidades de FAST al cual le agregó la característica de ser invariante a la rotación mediante el cálculo de la orientación local entre el centroide y el centro de la

característica detectada. Al descriptor BRIEF, también lo mejora al obtener la orientación para para las características descritas por BRIEF maximizando la varianza del descriptor y minimizando la correlación de las características bajo ciertas orientaciones.

Y siguiendo con la búsqueda de mejorar la velocidad de cálculo (Leutenegger et al. 2011) propone al descriptor Binary Robust Invariant Scalable Keypoints, BRISK. Esta técnica tiene como base el detector de esquinas AGAST (Mair et al. 2010), el cual mejora al detector FAST en velocidad, manteniendo la misma efectividad en la detección. Sin embargo, Leutenegger mejora la detección de puntos mediante un espacio de escalas continuo para considerar la escala. Para construir su vector de descripción, se realiza un muestreo en un número de puntos detectados. Para cada punto se obtiene su distancia con el resto. Se crea el conjunto de distancia corta y otro de distancia larga. La distancia larga se utiliza para estimar la dirección del punto significativo, mientras que el subconjunto de pares de corta distancia se utiliza para construir el descriptor binario, después de girar el patrón de muestreo. El descriptor utiliza la distancia hamming. Este descriptor es robusto a la escala como a la rotación.

(Alahi et al. 2012) propone una mejora del descriptor BRISK, con FREAK (Fast Retina Keypoint). FREAK mejora el muestreo de patrones y método de selección de pares utilizado por BRISK al proponer un traslape entre los patrones que están siendo promediados y eso crea una mayor concentración de información sobre el punto clave que se ve reflejado en una mayor descripción. La forma de estos patrones Gaussianos fue inspirada por el Sistema visual humano, específicamente por la retina.

2.3.2 Combinación de técnicas

El mejorar una técnica mediante su integración en otra, no es un proceso nuevo; sin embargo, el cómo se combinan es lo que marca la diferencia. En la literatura del área se pueden encontrar trabajos que reportan la combinación de los filtros Gabor con otros descriptores locales. (Moreno et al. 2009) propuso el SIFT Gabor, desarrollo similar al trabajo implementado en esta tesis. Este algoritmo presenta una propuesta de modificación del descriptor local SIFT, usando solamente filtros de Gabor impares

como núcleos de convolución para calcular las primeras derivadas de la imagen. Ellos calculan las derivadas horizontales y verticales con un valor de orientación de $\theta = 0$ y $\theta = \pi / 2$, respectivamente, $\lambda = 6\sigma$ y un valor para el parámetro de escala $\sigma = \sigma 1 = \sigma 2$. En esta tesis, se considera un banco de cuarenta filtros para tener una mayor representación del objeto de la imagen. Moreno evaluó su propuesta con imágenes de alta calidad, base de datos clásica en la correspondencia de puntos (Mikolajczyk & Schmid 2005).

En la literatura, se pueden encontrar otras propuestas de mejora de algoritmos aplicando los filtros Gabor, por ejemplo en (Bereta et al. 2013) compararon la precisión de varios descriptores locales combinados con ondas de Gabor para el reconocimiento facial en el contexto del envejecimiento; ellos reportan que el rendimiento más alto lo obtienen con los descriptores locales ILBP LBP y MBLBP combinados con filtros de Gabor. (Liao et al. 2013) comparó su propuesta de descriptor con cinco de los descriptores más populares: SIFT, PCA-SIFT, GLOH, SIFT Gabor y los momentos Zernike. Ellos muestran que obtuvieron las puntuaciones medias más altas con su propuesta; sin embargo, ligeramente abajo se encuentran la propuesta de SIFT Gabor. Por su parte, (Wan Yussof & Hitam 2014) propusieron un detector de puntos de interés multi-escala basada en ondas de Gabor para la detección de humanos, con excelentes resultados. (Conde et al. 2013) presentó el método HoG basado en filtros de Gabor e histogramas de Gradientes Orientados. Ellos mostraron que el uso de Gabor en el procesamiento previo mejora el desempeño de HoG.

Por otra parte, la combinación del descriptor SIFT y los momentos de Hu para crear un descriptor que sea más distintivo al unir información global y local ya se ha realizado. Sin embargo, dicha combinación se realiza a partir de una extracción de características diferente y no sobre el mismo conjunto de puntos de interés como se realiza en esta tesis. Por ejemplo, (Dou & Li 2014) para el reconocimiento de acciones humanas, describe al objeto mediante una descripción SIFT 3D y obtiene los momentos de Hu a partir del gradiente obtenido con el movimiento histórico de la imagen (MHI) y el movimiento de energía de la imagen (MEI). Los vectores son posteriormente combinados.

2.4 Análisis de trayectorias

Las trayectorias de la gente moviéndose en la escena proporcionan información útil para detectar los comportamientos normales y anormales. En el mundo real, especialmente en escenarios externos con alta variabilidad, las trayectorias de las personas pueden ser muy diferentes y por lo tanto difícil de modelar. No existe una forma simple de categorizar y etiquetar trayectorias como "normales" o "anormales"; por ello, sin un conocimiento a priori del contexto, las trayectorias solo pueden ser clasificadas como normales o anormales únicamente considerando su ocurrencia estadística.

Algunos métodos utilizan una propuesta holística que considera a la escena como un todo. En ella las desviaciones de los eventos usuales y frecuentes, sin ningún conocimiento a priori, puede tener el poder de discriminar eventos raros o poco frecuentes.

De acuerdo con (Gogoi et al. 2011) la detección de *outliers* (o valor atípico) se refiere al problema de encontrar observaciones que son diferentes del resto de los datos, basados en una apropiada métrica. En muchas áreas, se trabaja la detección de los outliers debido a que la información que proporcionan puede ser crítica para la seguridad de las aplicaciones. En videovigilancia, las desviaciones pueden representar potenciales ataques a los sistemas de seguridad.

A partir de los enfoques utilizados para la detección de outliers se puede realizar la detección de comportamientos sospechosos. Entonces, dependiendo de la disponibilidad de los datos y del contexto de la problemática que se desea resolver, se tienen tres enfoques:

a) Aprendizaje no supervisado. En este caso se desea determinar el comportamiento sospechoso sin un conocimiento a priori. Los datos son procesados como una distribución estadística y los puntos más remotos o distantes son potenciales anormalidades. Este enfoque tiene la ventaja de poder detectar desviaciones que el sistema nunca había visto en un comportamiento normal.
b) Aprendizaje supervisado. En este caso, se asume la disponibilidad de contar con un conjunto de datos de entrenamiento, pre-etiquetados, que constituyan una

buena representación de los modelos de “normalidad” y “anormalidad”. Es obvio mencionar que los resultados son altamente dependientes de la calidad y representatividad de los datos de entrenamiento.

c) Aprendizaje semi-supervisado. En este caso, normalmente se cuenta con datos etiquetados para el aprendizaje del modelo de “normalidad” y todo lo que difiera de este, se considera como un comportamiento anormal.

En la siguiente revisión de la literatura, se puede observar este tipo de enfoques, donde prevalece el enfoque no supervisado. También se puede observar que la mayoría de los trabajos toman en cuenta información del contexto de la escena en el modelado de los recorridos, para obtener reglas definidas sobre las conductas de las personas y determinar mejor los comportamientos inusuales como (Leach et al. 2014), (Calderara et al. 2011). Otra propuesta es descomponer las trayectorias en partes atómicas, segmentos mínimos (Acevedo-rodr et al. 2011), subrutas (Li et al. 2013), en eventos atómicos (Jiang et al. 2011), que les permita identificar incluso pequeñas desviaciones, dentro de un recorrido completo (Cancela et al. 2013). Sin embargo, esta forma de representar la información debe contar con un método eficiente para realizar la división de la trayectoria en partes significativas; y posiblmente tendrá problemas ante recorridos no representados en el modelado.

Mientras que en el aprendizaje no supervisado, no es necesaria ninguna información de a priori, se debe contar con trayectorias completas (de inicio a fin) para poder obtener su clasificación; y si es posible, eliminar antes de iniciar, ruido o trayectorias rotas. El problema aquí consiste en contar con un métricas que permitan comparar recorridos de diversos tamaños y con algoritmo de agrupamiento que estimen el número, óptimo o cercano al óptimo, las categorías en los datos.

La tabla 2.3 muestra ejemplos de trabajos relacionados al análisis de trayectorias, para identificar comportamientos anormales. En estos artículos, se puede observar la diversidad de técnicas aplicadas para realizar el modelado de las trayectorias sin embargo, prevalece el considerar información contextual de la escena y modelar las trayectorias considerando las subrutas que la integran.

Tabla 2.3 Trabajos relacionados al análisis de trayectorias.

Autor	Aportación	Agrupamiento	Base de datos
(Kaluza et al. 2011)	Propone realizar el reconocimiento inusual a partir de la detección histórica de las acciones.	Representa las trayectorias con modelos ocultos de Markov, considera información espacial y estadística. Detecta las anomalías en base a la clasificación de los eventos individuales, con Bayes.	ESCAPES, simulador de tray., de un aeropuerto.
(Calderara et al. 2011).	Representa las trayectorias como adyacencias en un grafo pesado. Proyecta el grafo en una matriz en la cual calcula los eigenvalores.	Discretiza las trayectorias mediante el algoritmo de Voronoi. Detecta las anomalías en base a la ocurrencia estadística.	Edimburgo: 26 de agosto y 14 de julio.
(Jiang et al. 2011),	Propuesta que considera información temporal y espacial. Detecta anomalías puntuales, secuenciales y concurrentes.	Aplica el algoritmo *Clospan* para encontrar subsecuencias. Crea modelos ocultos de Markov para el reconocimiento. Detectan anomalías con el algoritmo de agrupamiento espectral.	NGSIM, trayectorias de un crucero.
(Morris & Trivedi 2011)	Método de aprendizaje de las trayectorias normales y anormales en 3 pasos.	Segmenta las trayectorias con modelo de mezclas gaussianas. Las agrupa aplicando LCSS y un algoritmo espectral. Las modela con modelos ocultos de Markov.	Trayectorias del grupo CVRR. Cross, labomni, y I5.
(Acevedo et al. 2011)	Algoritmo de agrupamiento bioinspirado, el cual no necesita conocer el número de clústeres.	Segmenta las trayectorias mediante el algoritmo Douglas-Peucker, agrupa con Red Growing Neural Gas	Lobby del laboratorio INRIA, CAVIAR
(Baiget et al. 2012)	Propone definir las anomalías en 3 tipos semánticos: suaves, intermedias y duras.	Modela el escenario y las trayectorias con modelos de mezclas gaussianas. Divide las trayectorias con cubic-spline. Detecta las anomalías dependiendo de la desviación del modelo.	Información de trayectorias propia
(Ng & Chua 2012)	Sistema de videovigilancia para un estacionamiento, considerando información dinámica y contextual.	Considera información dinámica de la trayectoria, calcula un PCA sobre los vectores y trabaja con 4 ejes. El reconocimiento lo realiza con árboles de decisión.	Base de datos propia
(Li et al. 2013)	Propuesta de detección del comportamiento anormal a través de la reconstrucción de trayectorias dispersas	Segmenta las trayectorias con cubic-B-spline curves approximation. Con las subrutas crea un diccionario. Detecta las anomalías con una reconstrucción de tray., dispersas	Crucero NGSIM. Lobby del laboratorio INRIA, CAVIAR
(Cancela et al. 2013).	Propone un algoritmo de ruta mínima. Considera información del contexto de la escena y clasifica a los objetos en la escena para determinar sus acciones.	Segmenta las trayectorias con el método Sethian Fast Marching, modela las trayectorias con contornos activos geodésicos. Aplica DTW, LCSS y una distancia propuesta llamada mapa de distancias pesada.	BARD y CANDELA
(Javan Roshtkhari & Levine 2013)	Presenta un método que no necesita información a priori, codifica el video en volúmenes espacio-temporal 3D.	Codifican los videos en composiciones espacio-temporales, mediante un muestreo denso. Crea un modelo a través de BoW. Detecta las anomalías cuando un evento no puede ser reconstruido.	UCSD pedestrian Y secuencias de metro, train y personas caminando.
(Leach et al. 2014),	Propone considerar información social de las personas y de la escena. Crea una antología de los comportamientos en base a sus características como velocidad, dirección, persistencia.	Modela el escenario en regiones. Crea una antología del comportamiento. Detecta las anomalías considerando la distancia de la trayectoria con el modelo, con el vecino más cercano y el coeficiente Bhattacharya.	Pets 2007 y Oxford.
(Yu et al. 2014)	Propone utilizar características pesadas, considerando un espacio-temporal, y son codificadas con BoF.	Caracteriza las trayectorias con los puntos de interés detectados por SIFT y trayectorias de partículas; crea un diccionario y clasifica las trayectorias con MVS.	KTH, UCF sports y TV human interaction.

Cuando el enfoque considerado para el modelado de las trayectorias, es un aprendizaje no supervisado, es importante contar con métricas adecuadas para comparar dichos recorridos. Se puede normalizar o reducir las dimensiones de las trayectorias (Morris & Trivedi 2009), de tal forma que todas tengan la misma longitud. Sin embargo, con ello se puede eliminar información que puede ser representativa. Otra opción es utilizar medidas de distancia que son independientes a la longitud tales como Dinamyc Time Warping (Rabiner & Juang 1993), Longest Commun SubSequence (Vlachos et al. 2002), distancia Piciarelli y Foresti (PF) (Piciarelli & Foresti 2006), entre otras. Las distancias más representativas y con un buen rendimiento, de acuerdo a (Zhang et al. 2006) son las distancias LCSS y DTW. Otras medidas de distancia propuestas son Distance with Real Penalty, ER, (Chen & Ng 2004) y distancia Edit Distance on Real Sequence, EDR, (Chen et al. 2005). Y de acuerdo a Chen, son robustas a factores de ruido, cambios en el tiempo y escala y son efectivas midiendo la disimilaridad entre las series de tiempo.

2.5 Conclusiones

La extracción de características es una etapa básica para para contar con una buena descripción. En la recopilación de trabajos presentados se pueden observar propuestas interesantes y novedosas para la detección de peatones; sin embargo, el tema sigue abierto y no se cuenta con una representación completa y robusta para todos los casos.

La representación global mediante el uso de plantillas, siluetas, grafos, permite tener información de la forma de la figura humana, por lo cual es el método preferido para el reconocimiento de acciones. Un problema inherente al enfoque holístico es que estas características son más sensibles a las variaciones que tenga la imagen. En cambio, las descripciones integradas con características locales proporcionan mayor robustez a la oclusión y cambios de puntos de vista. Por lo cual, muchas investigaciones utilizan la combinación de la información geométrica y las características locales para un mejor rendimiento.

Las técnicas preferidas para realizar la detección de personas son el histograma de orientación del gradiente HOG, (Dalal & Triggs 2005a) y las ondas Haar por sus excelentes resultados. En la tabla 2.1 se puede observar su uso y algunas de las mejoras

realizadas a estos dos métodos. No obstante, en los últimos años se han propuesto exitosos descriptores locales como SURF, FAST, FREAK, etc., los cuales debido a su rendimiento, se ha extendido su uso en muchas áreas del reconocimiento de patrones y robótica. Pero se estos descriptores, se han evaluado poco en el contexto de la detección de personas. En esta tesis, se consideró el utilizar estas técnicas de descripción por los excelentes resultados reportados y por su robustez ante cambios geométricos, de escala, perspectiva, etc.

Es notoria la falta de técnicas que permitan describir objetos contenidos en imágenes de tamaño pequeño y baja resolución. La técnica HOG se ha utilizado con este tipo de imágenes; sin embargo, el desempeño del algoritmo baja (Wojek et al. 2009), (Enzweiler & Gavrila 2009). En (Guo et al. 2012) se realiza un re-escalado de las imágenes para poder ser descritas con esta técnica. Normalmente, las bases de datos utilizadas para la detección de personas están integradas de imágenes de alta calidad y un número superior a los 80 pixeles de altura.

La complejidad que envuelve la detección de personas en ambientes dinámicos reales, es tal que, el desarrollo de sistemas comerciales se están apoyando en otro tipo de sensores como cámaras térmicas y radares de onda milimétrica para seres vivos en general (personas y animales) bajo circunstancias de baja visibilidad, como puede ser de noche, con lluvia, nieve o niebla densa.

Con respecto a la literatura del análisis de las trayectorias, una de las propuestas más utilizada es el modelado de las trayectorias mediante su segmentación en pequeñas unidades atómicas con significado espacial. Ese enfoque tiene la ventaja de incorporar y analizar las transiciones de los recorridos de manera más puntual; permitiendo detectar cambios o sesgos pequeños. Sin embargo, este modelado tendrá problemas ante la presencia de nuevos comportamientos a los representados.

Para un mejor rendimiento en la detección de los comportamientos anormales, los trabajos consideran información del contexto de la escena, ya que al tener información sobre las regiones integran la escena y de los objetos presentes en ellas, se establecen reglas de actividades permitidas, siendo un complemento para la determinación de las conductas normales y anormales.

El contar con información de la escena ayuda a modelar y detectar, de manera más eficiente, los tipos de actividades que pueden presentarse en un determinado ambiente. Sin embargo, no en todas las situaciones es posible contar con esta información, y sólo se tiene el comportamiento de los objetos a modelar, siendo la tarea de definir el número de categorías o grupos presentes, más compleja.

El modelado y aprendizaje de las actividades, normalmente, se realizan fuera de línea, generalmente mediante un aprendizaje no supervisado. Sin embargo, en (Morris & Trivedi 2011) se propuso actualizar el modelo con nuevas trayectorias conforme van siendo adquiridas, siendo adecuado este enfoque para aplicaciones en tiempo real. Aspecto considerado como tema de investigación actual.

Capítulo 3

Bases de datos

3.1 Introducción

La adecuada selección de los datos que serán utilizados en la etapa de entrenamiento como en la fase de evaluación es muy importante. El primer conjunto dota al sistema de la información para generar los modelos y aprender. Se recomienda se verifique sea representativo (lo más que se pueda) de las condiciones, que se pretende, el sistema de reconocimiento sea robusto. Con respecto al conjunto de prueba, este debe integrar imágenes (instancias positivas y negativas) con una amplia gama de variabiliad, de tal forma que, permita conocer, de manera objetiva y cuantitativa, el grado de rendimiento de las técnicas o métodos evaluados y/o generalización de los modelos de aprendizaje generados.

Para el desarrollo y evaluación de sistemas inteligentes en la detección de humanos y su análisis del comportamiento videos, se han propuesto una gran diversidad de bases de datos públicas. Algunas bases de datos están diseñadas para un objetivo específico y otras muestran información de un sistema de videovigilancia de manera general; pero todas permiten contar con información, rápida y objetiva, para las etapas de aprendizaje y evaluación antes mencionadas.

Los sistemas de videovigilancia se integran de las tareas de detección de personas, el seguimiento de las mismas y el análisis del movimiento para entender el comportamiento presentado. Este capítulo contiene la descripción de las cuatro bases de imágenes utilizadas en la etapa de detección de la figura humana y las ocho consideradas en la etapa del análisis de trayectorias para la detección de comportamientos inusuales. Debido a problemas en la implementación en el módulo de seguimiento de los peatones, provocados principalmente por sombras, iluminación y fallos o pérdidas en la correspondencia de los puntos; no se cuenta con datos de

trayectorias propias libres de ruido, por lo que se optó por utilizar bases públicas para el modelado del movimiento y la detección de actividades inusuales. Es decir, se utilizaron bases de datos diferentes para las dos etapas desarrolladas en esta tesis.

3.2 Bases de datos para la detección de peatones

Las bases de datos existentes para la detección de personas pueden ser agrupadas en dos tipos (Dollar et al. 2012):

1. Aquellas que se integran de imágenes que contienen a una persona posando y realizando determinadas acciones, con ciertas restricciones en una amplia gama de dominios. Las imágenes o videos de estos conjuntos de datos presentan alta calidad, con una menor variabilidad en puntos de vista, escala, entornos, y número de personas. Se utilizan principalmente para el análisis del movimiento humano y reconocimiento de acciones humanas. Excelentes revisiones de estas bases de datos se pueden encontrar en (Poppe 2010), (Chaquet et al. 2013) y (Bedagkar-Gala & Shah 2014).

2. Las bases de datos "peatonales" contiene imágenes de personas en posición vertical (de frente, lado y espaldas), posiblemente con acciones de traslado en áreas urbanas. Este tipo de base de datos principalmente (pero no las únicas) son utilizadas en la detección de peatones en general o para sistemas de asistencia al conductor (se complementan). Estas bases de datos se integran de imágenes o videos adquiridos en entornos reales de zonas urbanas, con escenarios complejos y alta variabilidad en intensidad luminosa, poses, vistas, escalas, diversos tipos de resolución, número de personas, ropa, con oclusiones, etc. Revisiones completas y detalladas se pueden encontrar en (Gerónimo, Lopez, et al. 2010), (Poppe 2010), (Weinland et al. 2011b), (Dollar et al. 2012) y (Gerónimo & López 2014). Se decidió utilizar bases de datos de este tipo, porque ellas cumplen con las condiciones especificadas en el objetivo de esta tesis: trabajar con imágenes de pequeña escala y baja resolución en condiciones reales.

La tabla 3.1 muestra las características principales de las bases de datos más utilizadas en la detección de personas; esencialmente para aplicaciones en videovigilancia de peatones en seguridad vial. Se considera indispensable el listar las direcciones de las páginas web donde se encuentran disponibles. En estas ligas se pueden encontrar también nuevos conjuntos de datos para la detección de peatones, con interesantes características como videos capturados con varias cámaras, nuevos entornos, imágenes adquiridas con sensores infrarrojos, etc.

Tabla 3.1 Bases de datos utilizadas en la detección de peatones, con énfasis en aplicaciones en seguridad vial.

Base de datos	Año	Tamaño de imágenes	Color	Ejemplos Positivos	Ejemplos Negativos	Comentarios
MIT-CBCL Pedestrian Database[1] (Poggio 2000)	2000	64x128	Niveles de gris	924	No tiene	Imágenes urbanas. Vistas frontales y de espalda.
PETS [2]	2001 2003	768x576 (imagen completa)	color	3672 frames para el entrenamiento 1452 frames para evaluar		Personas y vehículos en el exterior en un área de estacionamiento.Cámara estática
PETS 2004 o CAVIAR [3]	2004	384x288 (imagen completa)	color	6 a 12 Megabytes		Seis escenarios, realizando diversas acciones. Cámara estática
INRIA Person Dataset[4]. (Dalal & Triggs 2005b)	2005	2592x1944 y 64x128	color	1239	1218	Imágenes de diversas fuentes, alta resolución y normalizadas.Peatones estáticos
CVC01[5] (Gerónimo et al. 2007)	2007	80x160 a 20x40	color	1000	6175	Imágenes de un ambiente urbano.
NICTA Pedestrian Dataset [6] (Overett et al. 2008)	2008	Más de 40 y 80 pixeles de altura	color	25551	5207	Imágenes de alta resolución, de un ambiente urbano.
Daimler[7] (Enzweiler & Gavrila 2009)	2009	640x480 recortadas y normalizadas a 18x36	color	15660 imágenes recortadas	5124	Video de ambiente urbano de 27 minutos.
Caltech [8] (Dollár et al. 2009)	2009	640x480 con regiones de diversos tamaños de altura.	Color	6 conjuntos de 1GB cada uno.	5 conjuntos de 1GB cada uno.	Videos tomados en un ambiente urbano y capturados desde un vehículo.
TUD-Brussels [9] (Wojek et al. 2009)	2009	720x256 y 640x480 (imagen completa)	color	1092	192	Imágenes de alta resolución en ambientes urbanos de muchedumbre.
CVC02 [5] (Gerónimo, Sappa, et al. 2010)	2010	70x140 a 12x24	Color	1016	7650	Imágenes de un ambiente urbano. Videos capturados desde un vehículo.
ITC	2015	15x25 a 40x70	color	265	420	Videos tomados durante todo el día. Cámara estática.

[1]CBCL: http://cbcl.mit.edu/software-datasets.
[2]PETS: http://www.cvg.reading.ac.uk/slides/pets.html
[3]CAVIAR: http://www-prima.inrialpes.fr/PETS04/caviar_data.html
[4]INRIA: http://pascal.inrialpes.fr/data/human/
[5]CVC: http://www.cvc.uab.es/adas/databases.
[6]NICTA: https://www.nicta.com.au/category/research/computer-vision/tools/automap-datasets/
[7]Daimler Mono Pedestrian Classification Benchmark Dataset:
http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html
[8]Caltech: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/
[9]TUD-Brussels o Multi-Cue Onboard Pedestrian Detection:
https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/people-detection-pose-estimation-and-tracking/multi-cue-onboard-pedestrian-detection/

Estas bases de datos proporcionan conjuntos de prueba y de evaluación disjuntos (excepto CBCL que no tiene). Las bases de datos de los últimos años, contienen magnitudes de datos mayores, así como tienen también imágenes etiquetadas (truth label) e imágenes con oclusion de peatones. Como se puede observar, es notoria la importancia de la detección de peatones de baja resolución, sobre todo para el desarrollo de sistemas "en entornos reales" en seguridad vial. Ante ello, en la base Caltech (Dollár et al. 2009) se propone una categorización de imáges de acuerdo a su escala para su tratamiento en los sistemas, y comenta lo importante que es el desarrollo de sistemas que contemplen el uso de imágenes de las dos últimas categorías:

- Cercana: Si longitud de la región ≥ 80 pixeles.
- Media: Si longitud de la región > 30 y < 80 pixeles.
- Lejana: Si longitud de la región ≤ 30 pixeles.

A continuación se describen las cuatro bases de datos seleccionadas en este trabajo para la detección de humanos: ITC, CVC01, CVC02 y MIT-CBCL. La base de datos ITC es una base de datos propia que se adquirió para para este proyecto. Las tres últimas son bases de datos públicas que permiten comparar el desempeño de nuestro descriptor local GSIFT ante otros métodos reportados. Estas bases de datos también son evaluadas con otras seis técnicas de descripción local consideradas en la experimentación, para conocer la habilidad de estos algoritmos ante factores complejos como la variabilidad de puntos de vista, cambios de escala, ángulo, iluminación, imágenes de tamaño pequeño y baja calidad. La selección de las bases de datos se realizó considerando se integraran de imágenes de escala pequeña y pobre resolución (sin haber sido normalizada de imágenes de media y buena calidad) y una base de datos con características deseables (CBCL) por la mayoría de los sistemas de videovigilancia.

3.2.1 ITC

Se realizó la creación de una base de datos propia para contar con videos reales de videovigilancia, con vista de vuelo de pájaro, de baja resolución y escala pequeña, para complementar las bases de datos públicas que proporcionan imágenes de peatones con una vista horizontal.

La base de datos ITC se adquirió gracias al apoyo del Instituto Tecnológico de Cuautla. Se digitalizaron vídeos en una zona de estacionamiento exterior del Instituto Tecnológico de Cuautla. Se instaló una cámara IP motorizada WiFi -G día/noche NIP-02, ver figura 3.1; en el techo del edificio de la administración. Los vídeos fueron tomados en el rango de 45 metros, ver figura 3.2. El horario de la captura considerado fue de 09 a.m. a 19:00 p.m. durante dos meses, en formato AVI. La digitalización se realizó en condiciones ambientales reales, por lo que los vídeos tienen grandes cambios de iluminación en el transcurso del día y un entorno cambiante ante la presencia (entrada y salida) de diversos objetos no considerados en este estudio (bicicletas, motos, perros) y también periodos largos con escaso y nulo movimiento por parte de personas. Se cuenta con 270 videos, de 60 minutos con una imagen de 320x240 pixeles (imagen completa), a 25 cuadros por segundo.

Figura 3.1 Cámara WiFi –G,modelo NIP- 02.

Figura 3.2 Vista de la cámara de la zona de estacionamiento.

Se segmentaron 20 videos para crear los conjuntos de datos de entrenamiento y evaluación. Estas imágenes contienen peatones con una alta variabilidad en la pose, el número de personas presente, la vista, la iluminación, el tipo de ropa, entre otros. Las imágenes consideradas en la muestra de ejemplos negativos se integra de objetos como automóviles (con vistas de frente, fondo y laterales), árboles, bicicletas, pájaros, perros e incluso vistas del piso con sombras y cambios de iluminación. El conjunto de entrenamiento se compone de 165 imágenes de peatones y 200 muestras negativas. El conjunto de prueba se compone de 100 imágenes ejemplos positivos 220 muestras negativas. La imagen del video tiene una resolución de 640x480 píxeles, pero el tamaño de la región de trabajo corresponde aproximadamente de 15x25 a 40x70 píxeles. Las imágenes están en formato JPG, ver figura 3.3.

a) b

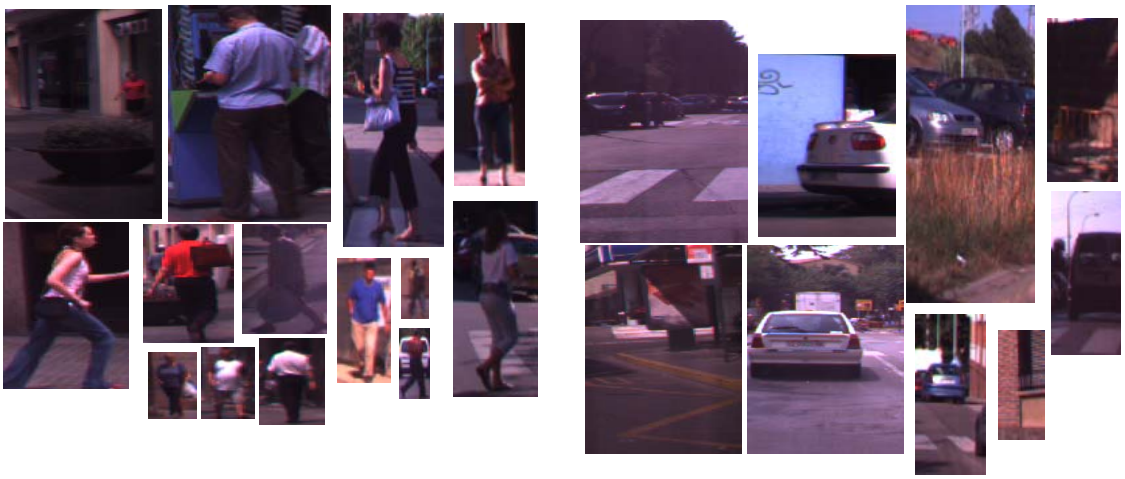
Figura 3.3 Ejemplos de base de imágenes ITC. a) Ejemplo de peatones, b) Ejemplo de negativos.

3.2.2 CVC01

La base de datos de Peatones CVC-01 (Gerónimo et al. 2007) pertenece a un escenario urbano de Barcelona. La base de imágenes consiste de 1,000 instancias positivas que presentan a personas caminando (con diferentes vistas). Se integra de 6175 ejemplos negativos tales como escenas del camino, fachadas de edificios, postes, etc. Las imágenes son a color en un formato PNG; las imágenes presentan un tamaño variable en el rango de 80x160 a 20x40 pixeles, y con cambios de intensidad luminosa fuertes.

Para este trabajo, el conjunto de entrenamiento se formó de manera aleatoria con 400 imágenes (200 imágenes de peatones y 200 de ejemplos negativos) y 600 para la fase de

evaluación (300 imágenes de peatones y 300 de ejemplos negativos), ver figura 3.4. Esta base de datos se ha utilizado en (Gerónimo et al. 2007) y (Ouyang et al. 2012).



a) b)

Figura 3.4 Ejemplo de imágenes de base CVC01. a) ejemplo de peatones, b) ejemplo de no peatones.

3.2.3 CVC02

La base de datos CVC02 (Gerónimo, Sappa, et al. 2010) consiste de tres subconjuntos, cada uno de ellos enfocados a tareas diferentes de detección de peatones, en el contexto de asistencia de manejo. Para la evaluación de este trabajo se consideró el conjunto de "entrenamiento". Este conjunto está integrado de 1016 imágenes tomadas en el rango de 0 a 50 metros; sin embargo, al localizar la región perteneciente al peatón, se tienen regiones de un tamaño aproximado de 70x140 a 12x24 pixeles; con alta variabilidad en ropa pose, iluminación y fondo. Los ejemplos negativos contienen imágenes del cielo, fachadas de edificios, de caminos, etc. Todas las imágenes son proporcionadas en formato PNG sin perdida, a color.

Para este trabajo se seleccionaron de manera aleatoria 400 imágenes para el conjunto de entrenamiento (200 imágenes de peatones y 200 de ejemplos negativos) y 600 para la fase de evaluación (300 imágenes de peatones y 300 de ejemplos negativos), ver figura 3.5. Esta base de datos se ha utilizado en (Gerónimo, Sappa, et al. 2010) y (Pedersoli et al. 2014).

a) b)

Figura 3.5 Ejemplo de imágenes de base CVC02. a) ejemplo de peatones, b) ejemplo de no peatones.

3.2.4 CBCL

La base CBCL de peatones (Papageorgiou & Poggio 2000) consiste de 924 imágenes urbanas con diferentes formas del cuerpo y con fondos sin restricción alguna. Cada imagen tiene una resolución de 64x 128 píxeles, a color, con formato de PPM. Para el conjunto de ejemplos negativos, se consideró la base de imágenes de coches del MIT[3]. Esta base de datos está integrada por 516 imágenes de coches (vista trasera o frontal) de tamaño 128x128 píxeles. Para la experimentación se consideró un conjunto de entrenamiento de 400 imágenes (200 imágenes de peatones y 200 de ejemplos negativos) y 600 instancias para la fase de prueba (300 imágenes de peatones y 300 de ejemplos negativos), como se muestra en la figura 3.6. Algunos artículos que han utilizado esta base de datos son (Fan et al. 2003), (Schauland et al. 2006), (Jung & Kim 2010) y (Farhadi et al. 2011).

[3] http://cbcl.mit.edu/cbcl/software-datasets/CarData.html

a) b

Figura 3.6 Ejemplo de imágenes de base CBCL. a) ejemplo de peatones, b) ejemplo de no peatones.

3.3 Bases de datos de trayectorias

El objetivo de estas bases de datos es el de conocer y analizar las actividades de las personas y por ello se integran de comportamientos normales y anormales, como pelearse, caerse, correr de pronto, dejar paquetes, caminar en zigzag, etc. Algunas de las bases de datos descritas en la tabla 3.1, pueden ser y son, utilizadas en esta etapa (se complementan). Ejemplo de ello se puede ver con la base de datos CAVIAR en (Acevedo-rodr et al. 2011) y (Li et al. 2013). Sin embargo, en muchas investigaciones se decide utilizar conjuntos de datos que describen las trayectorias (reales o simuladas) realizadas por humanos y/o vehículos, para evaluar propuestas de medidas de distancia y algoritmos de agrupamiento en la detección de comportamientos anormales. La tabla 3.2, lista algunos de los conjuntos de trayectorias más usados.

Tabla 3.2 Ejemplos de base de datos utilizados en la detección de comportamientos anormales.

Base de datos	Año	Tamaño de imágenes	Trayectorias normales	Trayectorias Anormales	Comentarios
CAVIAR	2005	640x480	22 grupos (gente caminado a las salidas)	19 (gente peleando, tirada y dejando paquetes)	Entrada a los laboratorios INRIA. También proporcionan videos en diferentes escenarios
NGSIM (Lankershim) dataset	2006	720x576	2,230 instancias que se integran en 8 grupos.	12 instancias	Videos de un crucero. (4 caminos transversales) de los Ángeles California. Se proporciona información de velocidad, dirección y posición de cada auto.
Pets 2007	2007	640x480	573 personas	Personas corriendo y paradas en áreas de gran movimiento.	8 videos, con 4 puntos de vista, de 24 personas en la escena.

Base de datos	Año	Tamaño de imágenes	Trayectorias normales	Trayectorias Anormales	Comentarios
Oxford data	2009	no especificado	no especificado	no especificado	Videos de 5 a 29 personas caminando. Diversos escenarios externos.
Edinburgh Informatics Forum Pedestrian (Majecka 2009)	2009	640x480	92000	no especificado	Trayectorias tomadas del pasillo de informática en la Universidad de Edimburgh
CVRR (Morris & Trivedi 2009a)	2008	no especificado	no especificado	no especificado	Proporciona 6 bases de datos de 3 escenarios diferentes.
BARD (Cancela et al. 2013)		no especificado	560	50	Escenario externo, camino peatonal.

CAVIAR: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/ o http://homepages.inf.ed.ac.uk/rbf/CAVIAR/Next Generation Simulation (NGSIM) Project: http://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm
PETS 2007: http://www.cvg.reading.ac.uk/PETS2007/data.html,
PETS: http://www.cvg.reading.ac.uk/slides/pets.html
OXFORD: http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html
Edinburgh Informatics Forum Pedestrian Database: http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/
CVRR: http://cvrr.ucsd.edu/
BARD: http://cvrr.ucsd.edu

A continuación se detallan las características de las bases de datos seleccionadas para la detección de comportamientos inusuales o anormales en esta tesis. Se decidió trabajar con las bases de datos proporcionadas por el grupo CVRR, ya que las trayectorias cuentan con el etiquetado verdadero y esto permitió evaluar de manera cuantitativa el desempeño del algoritmo de agrupamiento propuesto (pamTOK). Las otras dos bases de datos se integran de trayectorias de personas en ambientes externos y reales y son ampliamente utilizadas en la literatura. La tabla 3.3 muestra las características de los ocho conjuntos de datos considerados.

Tabla 3.3 Características de las bases de trayectorias utilizadas en este trabajo.

Base de datos	Trayectorias	Desplazamientos	Grupos	Escena
I5	806	27	8	Autopista
I5sim	800	22	8	Autopista
I5sim2	1600	141	8	Autopista
I5sim3	1600	141	16	Autopista
Cross	1900	23	19	Crucero
Labomni	209	624	15	Laboratorio
Bard	610	27	560 correctas, 50 incorrectas	Paso peatonal con áreas verdes
Edimburgo (26 agosto)	1992	1237	no especificado	Paso peatonal en la Universidad

3.3.1 Grupo CVRR

El grupo de "Computer Vision and Robotics Research Laboratory" (CVRR[4]) proporciona varios conjuntos bases de datos, entre ellas *Trajectory Clustering and Analysis Datasets* (Morris & Trivedi 2009) para la evaluación comparativa de medidas de distancia y algoritmos de agrupamiento. Las bases de datos describen las trayectorias de tres escenarios diferentes: hay 3 escenas de autopista simuladas, datos de una autopista real, una intersección simulada y una cámara omnidireccional de interior. Los datos proporcionados sólo contienen información espacial (la velocidad debe inferirse). A continuación se describen los diversos conjuntos de trayectorias.

a) Escenario de autopista, ver figura 3.7.

1. **I5:** Se integra de las trayectorias obtenidas por el seguimiento visual de vehículos de una autopista de ocho carriles.

2. **I5SIM**: Simulación de flujo libre de la autopista con cuatro carriles, con tráfico en ambas direcciones (8 carriles en total). Los puntos de las trayectorias son coordenadas del mundo real (las unidades son píxeles). El primer conjunto contiene solamente tráfico libre con una distribución de velocidad gaussiana de 70 mph con 5 mph de desviación estándar.

3. **I5_SIM2**: Escena simulada de la carretera con una distribución bimodal de velocidad (lento y rápido). El etiquetado sólo considera el carril de circulación, por ello las trayectorias I5SIM2 son etiquetadas por número de carril (considerando las coordenadas espaciales).

4. **I5_SIM3**: Escena simulada de la carretera con una distribución bimodal de velocidad (lento y rápido). El etiquetado considera el carril de circulación y la velocidad (8 carriles y 8 por flujo). Por lo que, las trayectorias en la misma línea, pero con diferente velocidad se consideran como diferentes grupos.

[4] http://cvrr.ucsd.edu/)

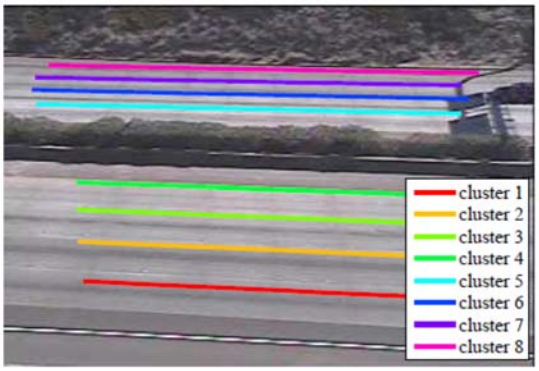


Figura 3.7 Escenario de autopista. Imagen tomada de (Morris & Trivedi 2009).

b) Escenario de una intersección o crucero de cuatro vías, ver figura 3.8:

CROSS*:* Esta base de datos describe una intersección de tráfico de cuatro vías. La escena proporciona trayectorias más complejas que las bases de datos de la autopista, incluye patrones de "continuar" y "turno", giros y vueltas en u. Las unidades son pixeles.

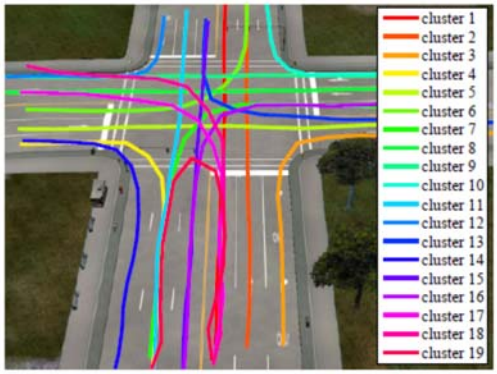


Figura 3.8 Escenario de un crucero. Imagen tomada de (Morris & Trivedi 2009).

c) Escenario de un laboratorio, ver figura 3.9:

LABOMNI: Se integra de trayectorias de humanos que caminan a través de un laboratorio, las imágenes se capturaron con una cámara omnidireccional. Este es un ambiente menos controlado que el tráfico vehicular de una autopista. Los participantes no estaban avisados de la adquisición de las imágenes, para obtener patrones de movimiento naturales. Las trayectorias tienen una larga duración y tienden a tener un cierto grado de traslape en el plano de la imagen.

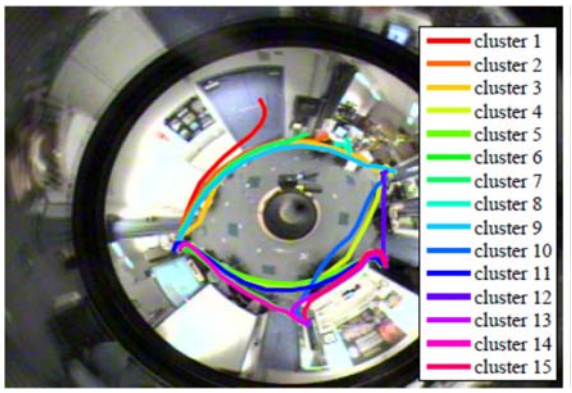


Figura 3.9 Escenario del laboratorio. Imagen tomada de (Morris & Trivedi 2009).

3.3.2 BARD

Esta base de datos fue propuesta en (Cancela et al. 2013), cuenta con 610 diferentes trayectorias, donde 50 tienen movimientos anormales. Cada ruta se integra de 27 posiciones almacenadas. La base de datos tiene claras regiones de entrada y salida. Los movimientos de la escena son sobre el pavimento, mientras que los movimientos sobre el pasto se consideran anormales. Los movimientos anormales son entre otras cosas el cruce por el pasto de algunas personas, cambios en las trayectorias por movimientos erráticos, cambios sorpresivos en la velocidad, dirección y que el objeto se detenga por largo tiempo, ver figura 3.10.

Figura 3.10 Escenario de la base de trayectorias BARD. Imagen tomada de (Cancela et al. 2013).

3.3.3 Edimburgo

La base de datos Edinburgh Informatics Forum Pedestrian Databaset (Majecka 2009) se integra de las trayectorias de personas caminando en el pasillo externo del edificio principal de la Escuela de Informática, ver figura 3.11, ubicado en la Universidad de Edimburgo, Reino Unido. Los videos fueron capturados por medio de un sistema de videovigilancia, que consistió en una cámara fija ubicada a 23m de altura, con 640x480 pixeles; en el periodo de julio 2009 a agosto 2010. El conjunto de datos disponible son más de 92,000 de trayectorias separadas por días. Ejemplos de comportamientos anormales son caminar en círculos, correr de manera inesperada, etc. La base de datos que se utilizó en la tesis es el conjunto de datos de Agosto 26, que consta de 1992 trayectorias.

Figura 3.11 Escenario de la base de trayectorias Edimburgo.
Imagen disponible en: http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING

3.4 Conclusiones

Existe una gran cantidad de bases de datos públicas para la detección de personas y el reconocimiento de acciones integradas, normalmente, de imágenes con un tamaño superior a los 100 pixeles de altura y de buena calidad. Las bases de datos utilizadas en la detección de peatones lejanos y en aplicaciones de videovigilancia en seguridad vial se integran de imágenes con menor escala. Sin embargo, las bases de datos con regiones de peatones con un tamaño superior a los 50 pixeles de altura son preferidas en los sistemas desarrollados en el área, ya que el considerar imágenes de un menor tamaño incrementa la complejidad de la detección de la figura humana.

Capítulo 4

Detección de personas

4.1 Introducción

La demanda de sistemas de reconocimiento de actividades humanas ha incrementado el desarrollo de métodos robustos de detección de humanos en ambientes reales. La extracción de características y su representación tienen una influencia crucial en el desempeño del reconocimiento de humanos, por lo cual, es esencial seleccionar o representar las características de los objetos de forma adecuada (Ke et al. 2013).

En la presente tesis, se propuso verificar si los filtros de Gabor pueden ser útiles para la extracción o la detección de puntos de interés en imágenes de baja calidad. Por ello se diseñó un nuevo descriptor denominado GSIFT (porque se basa en el descriptor local SIFT y en los filtros Gabor) para extraer características en regiones de imágenes compuestas por un reducido número de pixeles, para su descripción y clasificación.

En este capítulo se detalla el detector local propuesto, incluyendo información básica de las técnicas antecedentes. Para dar un soporte a las características locales, también se consideró el obtener información holística a partir de ellas. Entonces, se explica como a partir de los puntos de interés detectados, se obtiene información global a través de los siete momentos de Hu. La fusión de esta información se realiza a nivel de resultados en la clasificación con máquinas de vector soporte. En la figura 4.1 se puede observar la interacción de estas etapas.

Posteriormente, se presenta la evaluación del detector GSIFT que se lleva a cabo con cuatro bases públicas de datos reales, en ambientes urbanos. Tres de ellas integradas de imágenes con tamaños menores a 30 pixeles. En esta evaluación también se consideró el comparar el desempeño del detector local GSIFT, con seis técnicas de descripción local

como son SIFT, SURF, FAST FREAK, ORB y BRISK. La experimentación muestra el desempeño de las diferentes técnicas, en las cuatro bases de datos. Primero se evalúan las características locales y globales por separado y posteriormente se valida la fusión de ambos resultados. Finalmente, se discuten los resultados obtenidos y se listan las conclusiones.

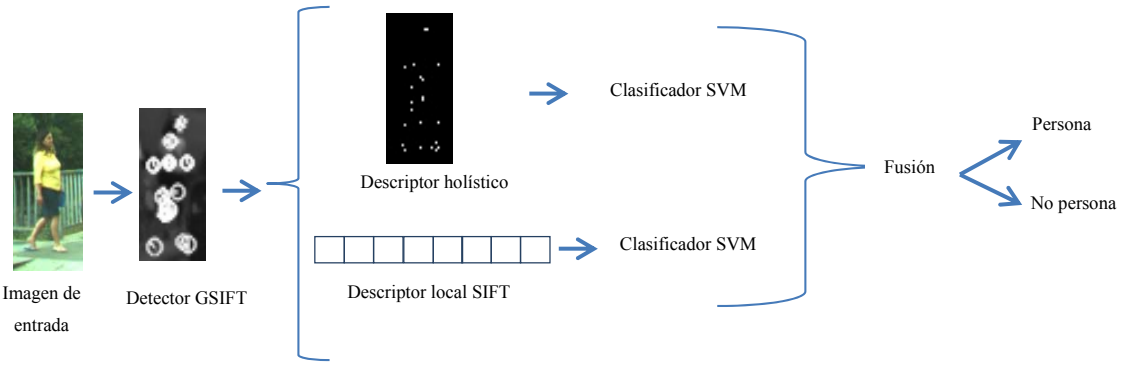


Figura 4.1 Esquema del método GSIFT para la descripción y detección de personas en pequeñas regiones.

4.2 Detector local GSIFT

Como ya se mencionó, el detector local GSIFT es una propuesta que toma como base al detector SIFT y modifica la etapa de detección local de puntos sustituyendo la función gaussiana utilizada en el espacio de escalas por un banco de filtros Gabor, como se observa en la figura 4.2. Se consideraron los filtros Gabor porque se utilizan para detectar bordes y son particularmente adecuados para representar la textura de los objetos, debido a que la frecuencia y la dirección de los filtros son similares a las representaciones del sistema visual humano; permitiendo al mismo tiempo, una buena localización espacial y la descripción de las estructuras de señal (Gabor 1946), (Moreno et al. 2009). Además, los filtros Gabor se han combinado con otros métodos de extracción de características locales, logrando una mejor detección de la información relevante (Conde et al. 2013).

A continuación, se repasa brevemente los antecedentes del detector propuesto y enseguida, se detalla el detector GSIFT.

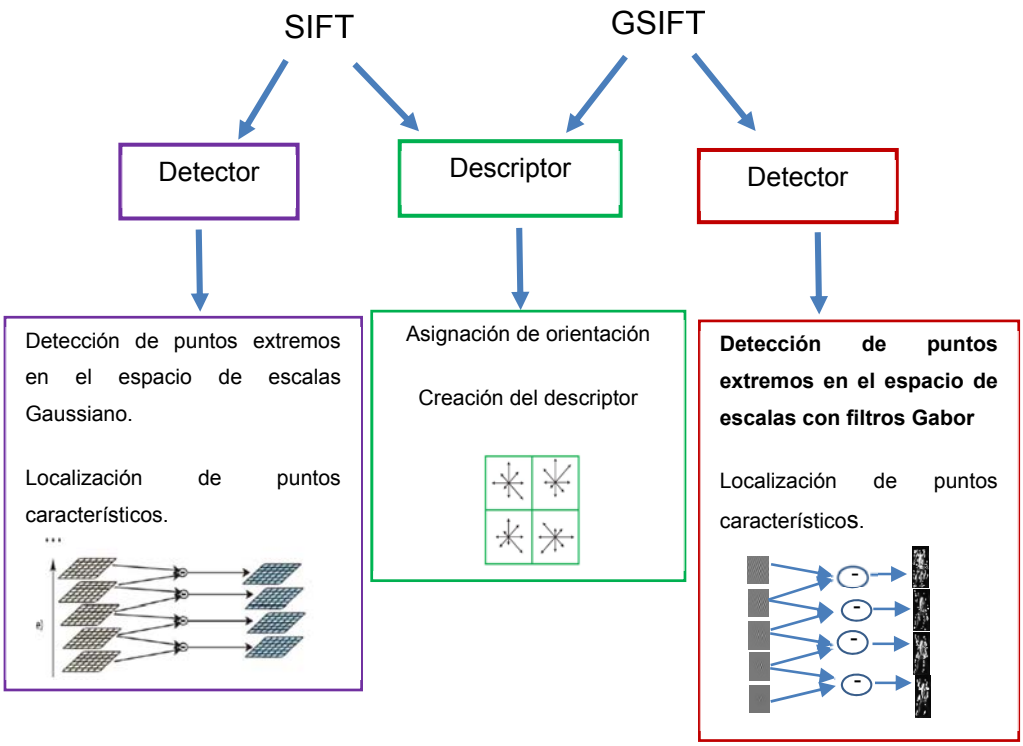


Figura 4.2 El detector GSIFT modifica la etapa de detección del SIFT incorporando filtros Gabor.

4.2.1 Descriptor local SIFT

La técnica SIFT se integra de un detector de puntos de interés y un descriptor de dichos puntos, y para su realización son necesarias las siguientes de cuatro etapas:

a) Detección de puntos extremos en el espacio de escalas.

El descriptor SIFT es construido a partir del espacio escala Gaussiano de la imagen original. Para ello, se define un espacio de escalas Gaussiano, ecuación 4.1. Donde $G(x,y,\sigma)$ es una función Gaussiana, (x,y) son las coordenadas espaciales, σ es el factor de escala y * es el operador de convolución.

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{4.1}$$

Lowe utilizó la pirámide de imágenes gaussianas junto con imágenes DoG (difference-of-Gaussian) para identificar potenciales puntos de interés, buscando los valores extremos (máximos y mínimos) del Laplaciano de la Gaussiana, ya que producen

características más estables (Mikolajczyk & Schmid 2002). La función Extrema-Local de DoG *D(x,y,σ)* se obtiene de la diferencia de dos escalas cercanas separadas por un factor multiplicativo constante, usando la ecuación 4.2.

$$D(x,y,\sigma) = \big(G(x,y,k\sigma) - G(x,y,\sigma)\big)I(x,y) = L(x,y,k\sigma) - L(x,y,\sigma) \qquad (4.2)$$

Para detectar los máximos y mínimos locales de cada punto, se realiza una comparación en las imágenes DoG del punto a analizar con el valor de sus 26 vecinos más próximos; es decir, en regiones de 3x3 en la escala actual, y adyacentes (inferior y superior). Si el valor resulta ser mayor o menor al de todos sus vecinos, se identifica el punto como máximo o mínimo local respectivamente.

b) Localización de puntos característicos

Para cada punto candidato se evalúa su estabilidad y se preservan los puntos estables Los puntos no firmemente situados sobre los bordes o aquellos con bajo contraste son bastante vulnerables al ruido y por lo tanto no podrán ser detectados bajo pequeños cambios de iluminación o variación del punto de vista de la imagen. Para eliminar los puntos con bajo contraste, se aplica un proceso de umbralización. Lowe recomienda eliminar los puntos cuyo valor sea menor a 0.03. Para la eliminación de los puntos inestables, se utiliza el cálculo de los autovalores de la matriz del Hessiano sobre la localización y escala del punto en estudio.

c) Asignación de orientación

En la tercera etapa, para cada uno de los puntos característicos se calcula la magnitud del gradiente *M*, y su orientación *θ*, mediante las ecuaciones 4.3 y 4.4.

$$M(x,y) = \sqrt{I_x(x,y)^2 + I_y(x,y)^2} \qquad (4.3)$$

$$\Theta(x,y) = \tan^{-1}\left(\frac{Iy\,(x,y)}{Ix\,(x,y)}\right) \qquad (4.4)$$

La orientación dominante se obtiene a través de la construcción de un histograma de orientaciones del gradiente de la vecindad del punto de interés. El histograma tiene 36

bins para cubrir los 360°. Cada pico en el histograma con una altura de 80% del máximo se considerará como la dirección dominante.

d) Descriptor de los puntos característicos

En esta última etapa, para cada uno de los puntos de interés se crea un vector de características. Se realiza un muestreo de las orientaciones y magnitudes del gradiente de la imagen sobre regiones de 16x16 centradas sobre cada uno de los puntos de interés. Para cada región, se realizan muestreos en sub-regiones cuadradas de 4x4 y sobre cada una de ellas se crea un histograma de orientaciones con ocho *bins* proporcionales a 45°. Al final, se tiene un vector de 128-elementos. Para dotar al vector de cierta robustez frente a cambios de iluminación, se lleva a cabo un proceso de normalización que minimiza los efectos de los cambios de iluminación.

4.2.2 Filtros Gabor

Los filtros Gabor 2D son filtros lineales cuya respuesta de impulso es una función sinusoidal multiplicada por una función gaussiana. En otras palabras, son filtros pasabanda en 2D, a los cuales si se les asigna una determinada frecuencia y dirección, realizan una reducción del ruido a la vez de preservar una dirección de la imagen original. La forma general del filtro de Gabor está dada por la ecuación 4.5:

$$G_k = \frac{1}{2\pi\sigma_x\sigma_y}\, exp\left[-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2}\right)\right] exp[2\pi\lambda(xcos\theta_k + ysin\theta_k)] \tag{4.5}$$

Donde

$$x' = (x - x_0)\, cos\, \theta + (y - y_0)\, sin\, \theta \tag{4.6}$$

$$y' = -(x - x_0)sin\theta + (y - y_0)\, cos\, \theta \tag{4.7}$$

(x,y) son las coordenadas espaciales, (x_0,y_0) definen la posición en el espacio de la onda, θ_k es la orientación del filtro Gabor, σ_x and σ_y representan la desviación estándar y λ es la longitud de la onda. La principal ventaja de los filtros Gabor es que las funciones son localizadas en el dominio especial y en el dominio de la frecuencia, por lo tanto, estas funciones son más adecuadas para representar una señal en ambos dominios.

4.2.3 Técnica GSIFT

El detector local GSIFT se integra de dos etapas. La primera fase es obtener puntos candidatos de la imagen que puedan ser identificados repetidamente bajo diferentes vistas y escalas del mismo objeto. En este caso, se propone construir una pirámide integrada de la convolución de la imagen original *I (x, y)* con el banco de filtros Gabor $g(\mathrm{x}, \mathrm{y}, \sigma, \lambda, \theta)$ para formar un "espacio de Gabor", ver ecuación 4.8. Obviamente, las imágenes de esta pirámide se integran de puntos con diferentes valores de frecuencia y de orientación.

$$\mathrm{L}\,(\mathrm{x}, \mathrm{y}, \sigma, \lambda) \;=\; g(\mathrm{x}, \mathrm{y}, \sigma, \lambda, \theta) \;*\; \mathrm{I}\,(\mathrm{x}, \mathrm{y}) \tag{4.8}$$

Donde *(x, y)* son las coordenadas espaciales, σ es el factor de escala, λ es la frecuencia de la onda y θ es la orientación.

La principal desventaja de los filtros de Gabor es que no se cuenta con valores de parámetros generales, éstos dependen de las características de la imagen (resolución, escala, etc.). La especificación de los valores para estos parámetros requiere una exhaustiva experimentación; sin embargo, en este trabajo, se emplean los valores sugeridos por (Lades et al. 1993) y (Wiskott et al. 1997) para extraer las características más importantes de la imagen: ocho orientaciones $\theta_k = (k\pi/8)$ con k=1,2,…8; cinco frecuencias entre 4 y 16 pixeles con saltos multiplicativos de $\sqrt{2}$; σ_x and $\sigma_y = \sigma = \pi$ y una fase de cambio con valor 0. Esto es, el banco de filtros considerado se integra de un total de 40 filtros Gabor que realizan una extracción de características en distintos grosores de bordes (de grueso a fino).

Se crea una pirámide integrada con ocho niveles, donde cada nivel analiza la imagen en una orientación diferente. A su vez, cada nivel se compone de imágenes resultantes de la convolución de los filtros Gabor en diferentes frecuencias, de tal forma que tengan los mejores puntos descriptivos en múltiples orientaciones y resoluciones, ver figura 4.3. Para identificar potenciales puntos de interés, se realiza la diferencia entre imágenes adyacentes función de diferencias. Este proceso es repetido en cada conjunto, usando la ecuación 4.9 donde *j = 1…8* es una constante de orientación.

$$DoGb\; D(x, y, \theta_j) = \; L\,(x, y, \lambda_1 \theta_j) \; - \; L\,(x, y, \lambda_2 \theta_j) \; \ldots - \; L\,(x, y, \lambda_5 \theta_j) \tag{4.9}$$

Como resultado de esta función de diferencias, se tienen cuatro imágenes (por conjunto) que son utilizados para detectar los valores máximos y mínimos de cada punto de la imagen, a través del método de detección local extrema. Es decir, cada pixel de la imagen actual, se compara en regiones de 3x3 respecto a las escalas adyacentes; entonces, se analiza con respecto a 26 pixeles (8 pixeles de la imagen actual, 9 pixeles de la imagen en la escala superior y 9 pixeles de la imagen en la escala inferior). Este proceso es representado en la figura 4.4.

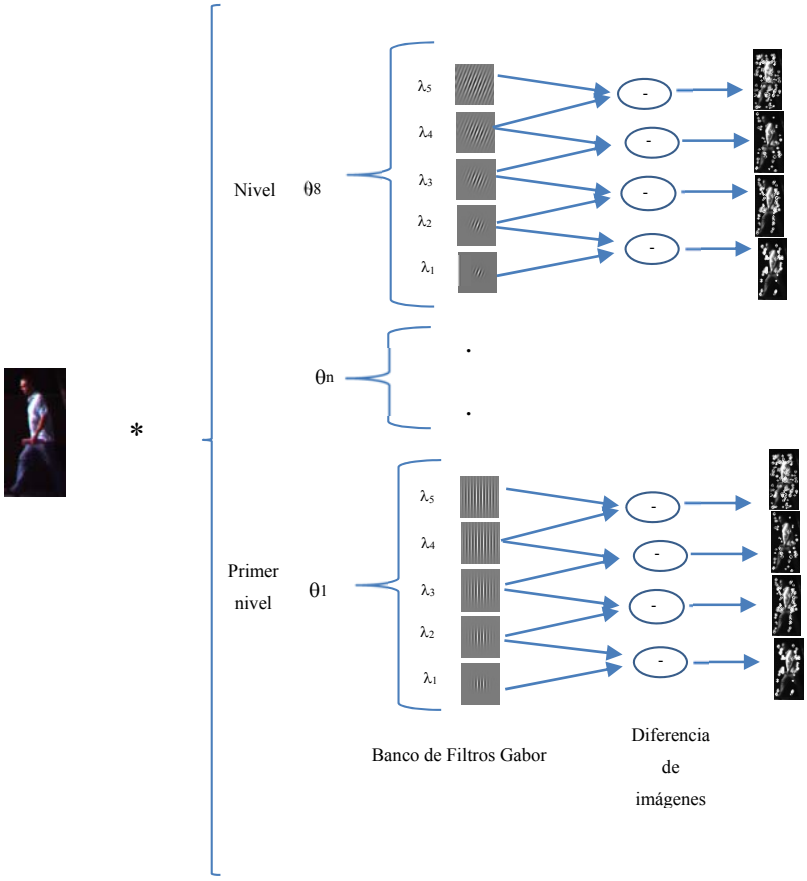


Figura 4.3 GSIFT: Detector de puntos de interés local en el espacio Gabor. La imagen de entrada es convolucionada con un banco de filtros Gabor, posteriormente se aplica una función de diferencias entre imágenes adyacentes.

Figura 4.4 Detección local extrema entre la imagen actual y las imágenes anterior y posterior.

En la segunda parte, se seleccionan los puntos en base a su estabilidad y se eliminan los puntos inestables y los repetidos. Se consideran puntos estables aquellos que cumplen con los criterios utilizados en SIFT para la localización de puntos característicos. Los puntos resultantes de este proceso de verificación en cada orientación, se unen en un único vector de características.

Después de este proceso con los puntos de interés detectados por GSIFT, se continúa con la etapa de descripción de cada uno de ellos. Se obtiene la magnitud del gradiente de la imagen y la orientación de cada subregión de la imagen (etapa tres y cuatro del descriptor SIFT), finalmente se cuenta con un vector de 128 elementos por cada punto de interés.

El detector local GSIFT propone más y mejores puntos discriminativos en múltiples resoluciones y orientaciones, y con ello, crea un vector de características más descriptivo, mejorando así el desempeño final de la detección. Es cierto que, existe un incremento en el costo computacional en la etapa de extracción de características, como en la etapa de clasificación y en la memoria al tener un vector de características más grande. Sin embargo, este proceso de detección de puntos de interés se aplica a pequeñas regiones, y no toda la imagen; es decir, el conjunto de pixeles sobre los cuales se trabaja no es grande, produciendo un número reducido de puntos de interés. Por esta razón y para evitar perder información, tampoco se aplicó una reducción en la dimensionalidad del vector de características.

4.3 Descriptor Global

Una deficiencia de las características locales es no poder resolver las ambigüedades, que pueden ocurrir, en una imagen que contenga muchas áreas que son localmente similares unas a otras. En este caso, la información global de las características es vital cuando se desea describir y realizar la correspondencia de estos puntos. Un vector de características que combina información local con holística, es más distintivo y por lo tanto eleva la solidez o robustez de la descripción al no verse afectada significativamente por el cambio (Li & Ma 2009).

Los descriptores globales resumen el contenido de la imagen en un único vector o matriz de características. Poseen la ventaja de encapsular una gran cantidad de información de la imagen requiriendo una pequeña cantidad de datos para describirla. Este tipo de descriptores han resultado ser ampliamente utilizados para diferentes tareas debido a su bajo coste computacional unido a unas prestaciones relativamente buenas.
La descripción holística de una imagen a través de los momentos fue realizada por primera vez por Hu (Hu 1962), quien propuso siete medidas invariantes a cualquier transformación de traslación, rotación y escalado de la imagen, llamados momentos invariantes. Desde entonces, se han propuesto varias familias de momentos que buscan tener el mínimo de información y mejorar su desempeño en el proceso de clasificación. En (Papakostas et al. 2013) se realiza una excelente revisión comparativa comparativo acerca de las familias más representativas de los momentos.

En este trabajo, considerando la capacidad de los momentos de Hu de realizar descripciones compactas y robustas, la descripción global se realiza a través del cálculo de los siete momentos invariantes de Hu (momentos de segundo y tercer orden), además del área de la imagen binaria, ver ecuaciones de 4.10 a 4.16, (Gonzalez 2009).

Como propuesta de este trabajo y a fin de capturar realmente la relación espacial de los puntos de interés detectados por las diferentes técnicas de detección local, se crea una imagen binaria a partir de los puntos característicos de cada transformada.

$$h_1 = \eta_{20} + \eta_{02} \tag{4.10}$$

$$h_2 = \left(\eta_{20} - \eta_{02}\right)^2 + 4\eta_{11}^2 \tag{4.11}$$

$$h_3 = \left(\eta_{30} - 3\,\eta_{12}\right)^2 + \left(3\eta_{21} - \eta_{03}\right)^2 \tag{4.12}$$

$$h_4 = \left(\eta_{30} + \eta_{12}\right)^2 + \left(\eta_{21} + \eta_{03}\right)^2 \tag{4.13}$$

$$\begin{aligned} h_5 = {} & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[\left(\eta_{30} + \eta_{12}\right)^2 - 3\left(\eta_{21} + \eta_{03}\right)^2\right] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - \left(\eta_{21} + \eta_{03}\right)^2\right] \end{aligned} \tag{4.14}$$

$$h_6 = (\eta_{20} - \eta_{02}) \left[\left(\eta_{30} + \eta_{12}\right)^2 - \left(\eta_{21} + \eta_{03}\right)^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \tag{4.15}$$

$$\begin{aligned} h_7 = {} & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) \left[\left(\eta_{30} + \eta_{12}\right)^2 - 3\left(\eta_{21} + \eta_{03}\right)^2\right] \\ & - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \left[3\left(\eta_{30} + \eta_{12}\right)^2 - \left(\eta_{21} + \eta_{03}\right)^2\right] \end{aligned} \tag{4.16}$$

4.4 Evaluación experimental

4.4.1 Clasificación

La clasificación se realizó mediante Máquinas de Vector Soporte (Cortes & Vapnik 1995) utilizando un kernel de Base Radial (RBF por sus siglas en inglés). Se llevó a cabo una búsqueda de los valores óptimos para los parámetros *gamma* y *cost* (para cada transformada y para cada base de datos) a través de la función *tune* y utilizando los rangos sugeridos en (Hsu et al. 2003) que son: $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ y $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$. Se aplicó un proceso de normalización al conjunto de momentos, para eliminar las variaciones presentes en los datos por la diferencia de valores, evitando sesgos artificiales. La normalización se realizó a los vectores de descripción tanto globales como locales mediante la media y desviación estándar de los datos de entrenamiento, de cada base de datos. Éstos valores (la media y desviación estándar) se utilizaron para normalizar a su correspondiente conjunto de prueba. El sistema tiene como salida solo dos valores, "si o no" se detecta un peatón en la imagen.

La combinación de la información obtenida por múltiples maneras obtiene un mejor rendimiento en comparación con un único sistema clasificador unimodal (Ross & Jain 2003). La fusión se puede realizar en diferentes niveles, en este caso se realiza a nivel de resultados, ya que para cada modalidad (global y local) se hace uso del vector de descripción y del clasificador correspondiente, como se muestra en la Figura 4.1. Es decir, La decisión final de pertenencia, considera la respuesta obtenida por cada clasificador y se combina para producir una puntuación única; se consideró el calcular el promedio de la media aritmética, entre otras posibilidades.

Como resultado de la etapa de entrenamiento se cuenta con 56 modelos de clasificación obtenidos para las 7 técnicas, para 4 bases de datos, utilizando características locales e información global. La siguiente actividad fue evaluar los modelos obtenidos de la etapa de aprendizaje con los conjuntos de prueba, los cuales son datos disjuntos.

4.4.2 Criterios de evaluación

La evaluación de la calidad de un sistema de verificación requiere de un detallado análisis de los posibles fallos y aciertos del sistema. Para la valoración de la calidad de las técnicas de características locales bajo estudio, se consideró el uso de las curvas

ROC (Receiver Operating Characteristic), ya que son curvas fáciles de interpretar, permiten una evaluación cuantitativa de la exactitud mediante el área bajo la curva, no requieren un nivel de decisión particular porque está incluido todo el espectro de puntos de corte y proporcionan una comparación visual directa del rendimiento, del detector local GSIFT y de las distintas técnicas de descripción local consideradas, sobre los mismos conjuntos de prueba.

Para el análisis estadístico de la curva ROC, se consideró la relación entre la razón de verdaderos positivos (TPR por sus siglas en inglés) con respecto a la razón de falsos positivos (FPR por sus siglas en inglés) con todos los posibles puntos de corte o umbral de discriminación (valor a partir del cual se considera que un caso es un positivo). La TPR indica la probabilidad de clasificar correctamente a una instancia cuyo estado real es el definido como positivo, en este caso, ser reconocido como "persona". La FPR señala cuántos resultados positivos son incorrectos, divididos entre el total de los casos negativos que integran el conjunto de evaluación. Estas medidas se obtienen de la siguiente manera, ver ecuaciones 4.17 y 4.18, en base a la matriz de confusión mostrada en la tabla 4.1.

Tabla 4.1 Matriz de confusión.

	Clase Real	
Clase Predicha	**Positiva**	**Negativa**
Positiva	Verdaderos positivos VP	Falsos Positivos FP
Negativa	Falsos negativos FN	Verdaderos negativos VN

$$TPR = \frac{VP}{VP + FN} \tag{4.17}$$

$$FPR = \frac{FP}{FP + VN} \tag{4.18}$$

A partir de la curva ROC, los estadísticos considerados a para evaluar el rendimiento de las técnicas son:

a) El punto "Equal Error Rate" (EER). Es el punto de inserción de la curva ROC con la línea convexa a la línea de discriminación. Cuanto más bajo sea el valor del EER, significa que el sistema tiene menos fallos.

b) El área bajo la curva ROC (AUC por sus siglas en inglés) es una medida de la eficacia del modelo, independientemente del punto de corte que se establezca. La exactitud máxima correspondería a un valor de AUC igual a 1 (arriba y a la izquierda) y la mínima a 0.5. Se obtuvo de acuerdo a la ecuación 4.19.

$$AUC = \frac{1 + (TPR - FPR)}{2} \tag{4.19}$$

4.4.3 Plan de pruebas

El objetivo de la experimentación realizada fue evaluar el rendimiento del detector de puntos locales GSIFT. Para comparar el desempeño del detector local GSIFT, se seleccionaron seis de las principales técnicas de descripción local: SIFT, SURF, FAST FREAK, ORB y BRISK. FAST al ser sólo una técnica de detección de puntos de interés, se combinó con las técnicas de descripción FREAK y SURF extendido.

Las imágenes para la etapa entrenamiento y de evaluación corresponden a ejemplos de peatones y ejemplos negativos, particularmente de imágenes de vistas de carros y áreas o imágenes relacionadas a ambientes urbanos. Como se comentó, las imágenes fueron seleccionadas, principalmente, por presentar una escala pequeña y con baja resolución, en un ambiente urbano real. Además, las imágenes tienen una buena representatividad con respecto a puntos de vista, aspecto de las personas, variabilidad en la intensidad luminosa y en el entorno. Incluso, en algunos casos, con oclusión o traslape entre personas.

A continuación se muestran y analizan los resultados obtenidos para cada transformada, con los conjuntos de evaluación considerados cada base de datos.

4.4.4 Clasificación con descripciones locales

El conjunto de imágenes de la base ITC, figura 4.5, es una excelente muestra del desempeño de las técnicas de descripción local con imágenes con alta variación en la intensidad luminosa, poses, imágenes de baja calidad y tamaño pequeño. El detector local GSIFT obtiene un buen resultado con 0.1031% EER, quedando sólo abajo de la técnica FASTF que logra un 0.03%. Es interesante notar que la representación realizada por el descriptor FREAK, obtiene un desempeño superior a la alcanzada por el descriptor de SURF, a partir de los mismos puntos localizados por FAST.

En esta base de datos, SURF presenta problemas con la intensidad de la luz, lo que provoca zonas homogéneas en algunas partes del cuerpo, lo que causa la no detección de puntos de interés y con ello la disminución en su rendimiento, teniendo un desempeño inferior al mostrado por SIFT, que tiene una AUC del 0.829%. Los dos descriptores locales binarios, ORB y BRISK, obtienen un conjunto mínimo de puntos en toda la imagen, siendo insuficientes para trabajar.

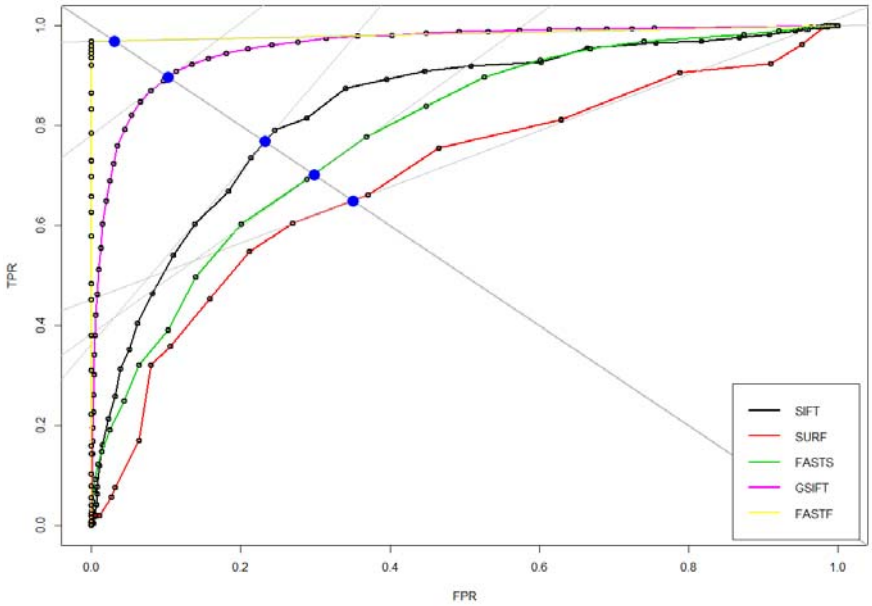


Figura 4.5 Reconocimiento de peatones en la base ITC, con descripciones locales.

La variabilidad presente en las bases de datos CVC02 y CVC01 es alta y esto se ve reflejado en el desempeño de las técnicas de detección y descripción locales. Como se puede observar en las figuras 4.6 y 4.7, es notorio el buen rendimiento del método propuesto GSIFT al alcanzar un 0.218% y 0.199% de EER respectivamente, frente a las

otras técnicas al detectar más y mejores puntos clave, ver Tabla 4.2, mejorando sustancialmente el desempeño del descriptor SIFT que alcanza un 0.316% y 0.385% de EER.

Los resultados también muestran que el algoritmo FAST es una buena alternativa para detectar puntos de interés en imágenes de baja calidad y/o con escalas pequeñas, ya que independientemente de que sea utilizada con el descriptor FREAK o con el descriptor de SURF, obtiene resultados similares a la técnica SURF, ver tablas 4.2 y 4.3 Por su parte, SURF funciona bien con imágenes de buena calidad, pero tiene problemas al trabajar con regiones con baja calidad y pocos pixeles, por ello la variabilidad de resultados alcanzando con estas bases de datos un EER de 0.308% y 0.304%, y AUC de 0.759% y 0.748%. SIFT no logra los mejores resultados, pero aun así es una buena opción, ante las variaciones geométricas de rotación y escala en la CVC01 con un EER de0.746%, obteniendo un menor resultado en la CVC02 con 0.385%. En el caso de los descriptores binarios ORB y BRISK, su rendimiento no es bueno con las imágenes pequeñas, siendo constante su mal desempeño.

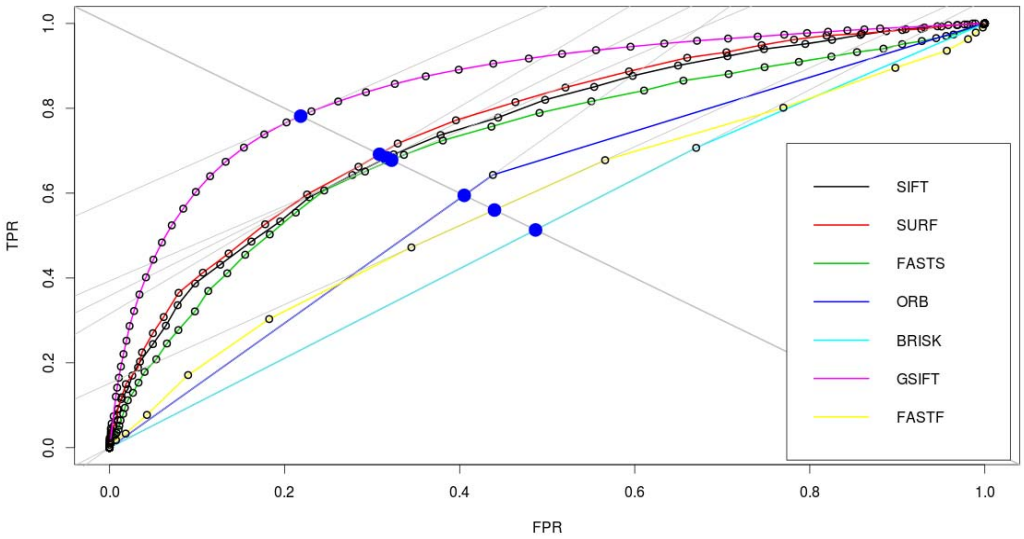


Figura 4.6 Reconocimiento de peatones en la base CVC01, con descripciones locales.

La base de imágenes CBCL permite observar los excelentes resultados que las distintas técnicas de descripción tienen con imágenes de buena calidad y un tamaño superior a los 80 pixeles. Los resultados que muestra la figura 4.8 están acordes a lo reportados en

la mayoría de los artículos de la literatura. Como se puede ver, FASTF con un 0.0%, ORB con un 0.0%, SURF con un 0.0008% y FASTS con un 0.019% de EER tienen un ejemplar desempeño. SIFT y BRISK obtienen un buen resultado con un EER de 0.136% y 0.143% respectivamente. En este caso, el método propuesto GSIFT obtiene un aceptable 0.175% de EER, siendo el método de menor desempeño. Analizando en detalle las imágenes, se encontró que se debe mejorar el filtrado de los puntos para optimizar la discriminación de los mismos y de los puntos inestables obtenidos en zonas con alto contraste.

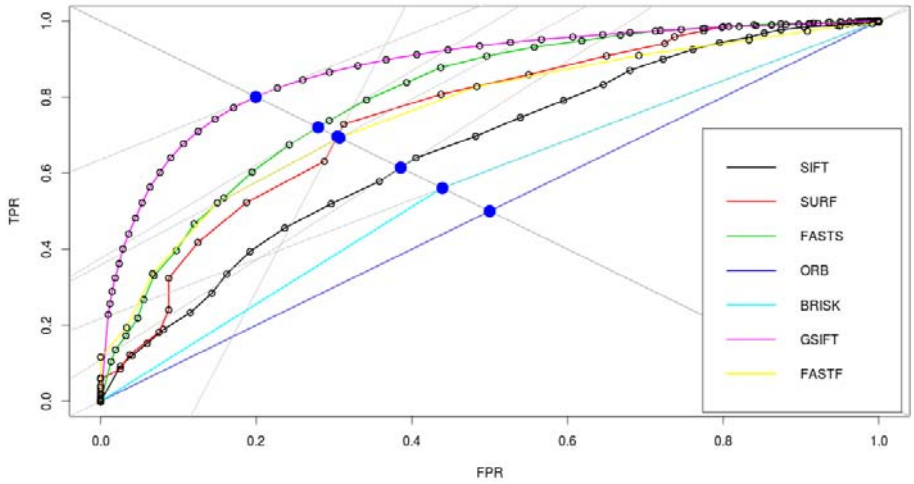


Figura 4.7 Reconocimiento de peatones en la base CVC02, con descripciones locales.

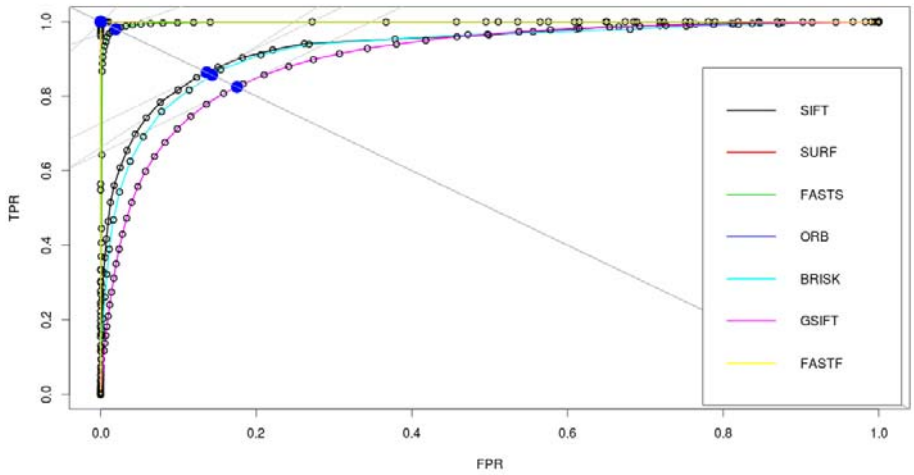


Figura 4.8 Reconocimiento de peatones en la base CBCL, con descripciones.

4.4.5 Clasificación con descripciones holísticas

En este caso de prueba, la información de entrada al clasificador (máquinas de vector soporte) es la descripción realizada a través de los momentos de Hu, de los puntos de interés detectados por cada una de las técnicas de descripción local. El conjunto de técnicas evaluadas cambia, ya que la técnica FAST se evalúa como como técnica de detección local de puntos y el método de descripción FREAK es omitido al no poder participar. Los resultados de esta evaluación se muestran en las figuras 4.9 a la 4.12.

Como se puede observar en la tabla 4.2 y la tabla 4.3 todas las técnicas mejoran significativamente sus resultados con respecto a la descripción local. Y este resultado de mejora se mantiene en las cuatro bases de datos. Por ejemplo, con la base ITC para las cuatro técnicas evaluadas se nota la diferencia en el desempeño logrado; siendo el método SURF el más beneficiado con esta descripción pasando de una AUC de 0.69% con una descripción local a un 0.99% con la descripción global. El método GSIFT también mejora su desempeño alcanzando un AUC del 0.99%, como se ve en la figura 4.9.

En la base CVC01 y de acuerdo a la figura 4.10, el detector GSIFT presenta un menor desempeño con respecto a las otras técnicas (SURF, FAST y ORB) con un EER del 0.191%. Sin embargo, si se revisan los valores de EER y AUC obtenidos por las otras técnicas de descripción, en las tablas 4.2 y 4.3, se puede ver que ellas mejoran su desempeño significativamente con la descripción global, respecto a la descripción local, siendo notoria esta diferencia. GSIFT también presenta mejores resultados, pero la diferencia es menor consiguiendo un 0.85% de AUC. No obstante, mantiene su desempeño arriba del alcanzado por la técnica SIFT que, también mejora pasando del 0.74% al 0.82% de AUC.

En la base de datos CVC02, GSIFT alcanza la mejor puntuación con el valor más bajo de EER igual a 0.036%. Siguiéndole, SURF y FAST con 0.076 y 0.093 de EER, respectivamente, como se ve en la figura 4.11. Es decir, nuevamente es posible verificar que la información geométrica considerada por los momentos de Hu, mejora sustancialmente el poder descriptivo de todas las técnicas de detección y descripción local consideradas.

En relación a la base CBCL, confirmando los resultados obtenidos en la prueba anterior utilizando los descriptores locales, todas las técnicas tienen un excelente desempeño. No obstante, en esta ocasión, GSIFT logra un EER de 0.016%, quedando ligeramente arriba de las otras técnicas de descripción local, figura 4.12.

En todas las pruebas realizadas, la técnica BRISK tiene un desempeño por debajo de lo aceptable, excepto con la base de datos CBCL. Esto se debe a que con imágenes de pequeña escala detecta pocos puntos de interés, no suficientes para contar con una descripción que le permita ser efectiva ante la variabilidad presente en las imágenes. Sin embargo, aún con imágenes de menor variabilidad y un tamaño arriba de 100 pixeles, como las que integran a esta base de datos, obtiene un menor resultado que las otras técnicas de descripción local.

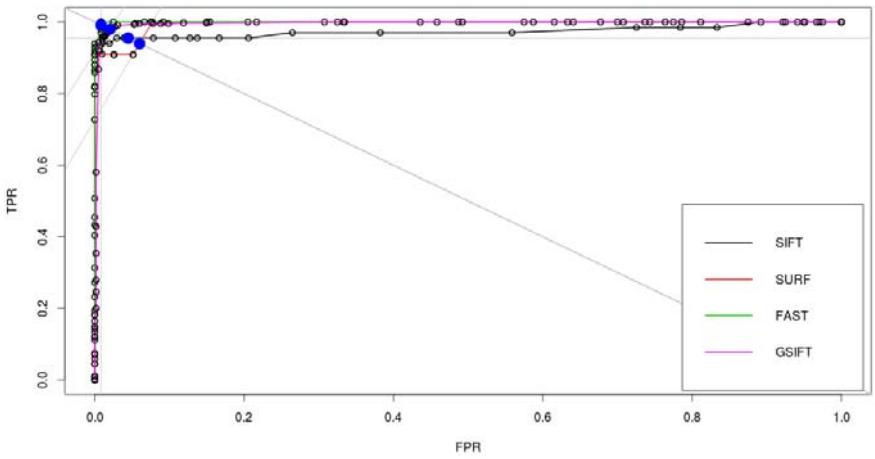


Figura 4.9 Reconocimiento de peatones en la base ITC con descripciones holísticas.

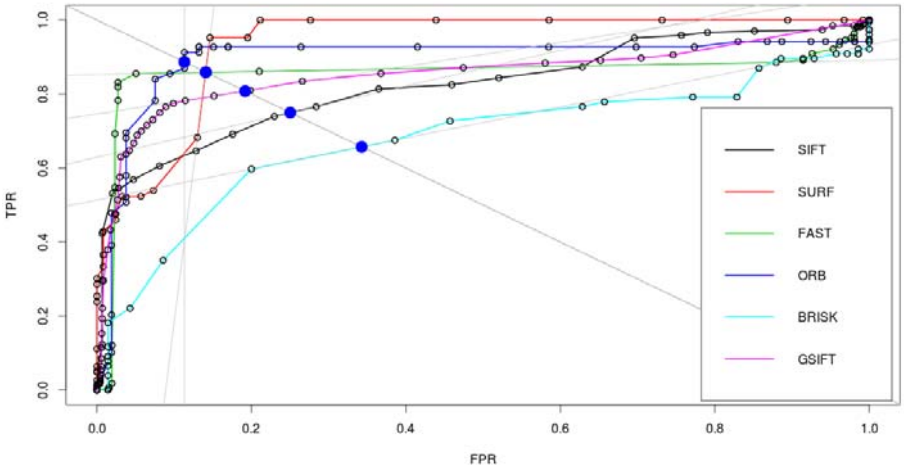


Figura 4.10 Reconocimiento de peatones en la base CVC01 con descripciones holísticas.

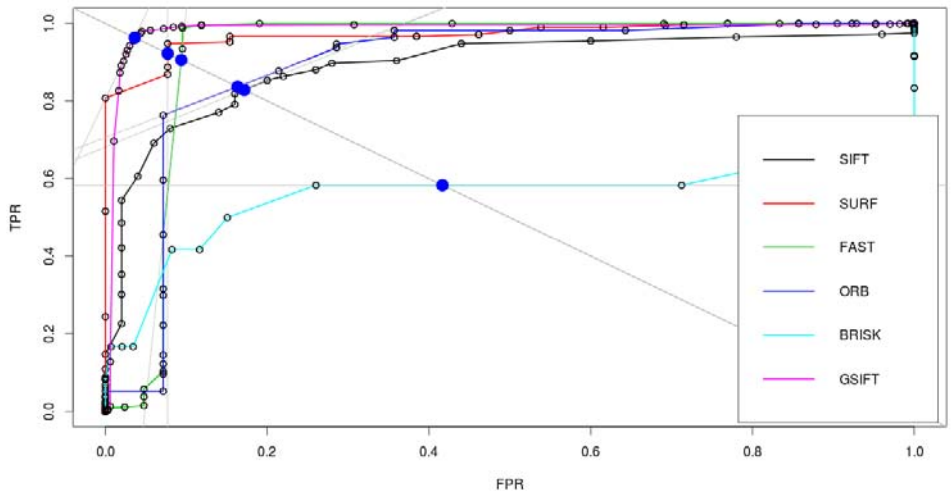


Figura 4.11 Reconocimiento de peatones en la base CVC02 con descripciones holísticas.

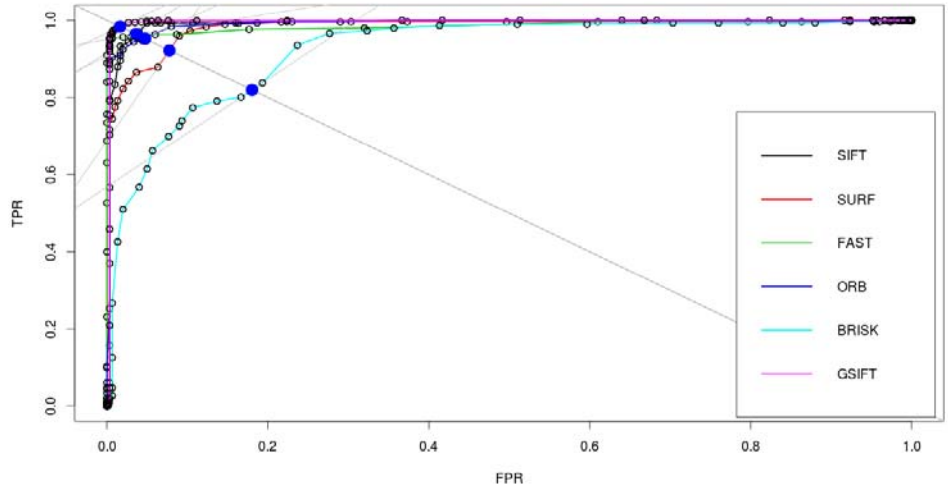


Figura 4.12 Reconocimiento de peatones en la base CBCL con descripciones holísticas.

4.4.6 Reconocimiento con fusión

El objetivo de esta prueba fue evaluar el rendimiento de combinar los resultados obtenidos de las clasificaciones anteriores. Los resultados de la fusión se llevaron a cabo a nivel de "score", calculando la media aritmética de los valores predichos obtenidos, por cada uno de los descriptores por separado.

Como puede verse en las figuras 4.13 a 4.16 y tablas 4.2 y 4.3, en esta prueba, el rendimiento de la mayoría de los métodos evaluados es mejorado sustancialmente. Es

notorio ver que la combinación de la descripción realizada por los momentos Hu en colaboración con el comportamiento local de descriptores locales, mejora significativamente las capacidades de la clasificación de los métodos originales. Y este comportamiento se mantiene en las cuatro bases de datos.

Es importante hacer notar que el detector GSIFT logra los mejores resultados frente a todas las técnicas evaluadas, en las bases de datos ITC, CVC01 y CVC02 con valores de 0.006%, 0.0452% y 0.009% de EER, respectivamente. La fusión en GSIFT tiene un mejor rendimiento en comparación con los resultados obtenidos, por separado, de las descripciones locales y globales. La combinación de la información del resultado mejora de manera uniforme a todas las técnicas de descripción local alcanzando valores de AUC altos. En la base ITC, FASTS alcanza un 0.99%, ver figura 4.13; en la base CVC01, SIFT logra un 0.93% superando los resultados alcanzados por las técnicas SURF y FAST, ver figura 4.14. Y en la base CVC02, SURF obtiene un 0.97%, ver figura 4.15

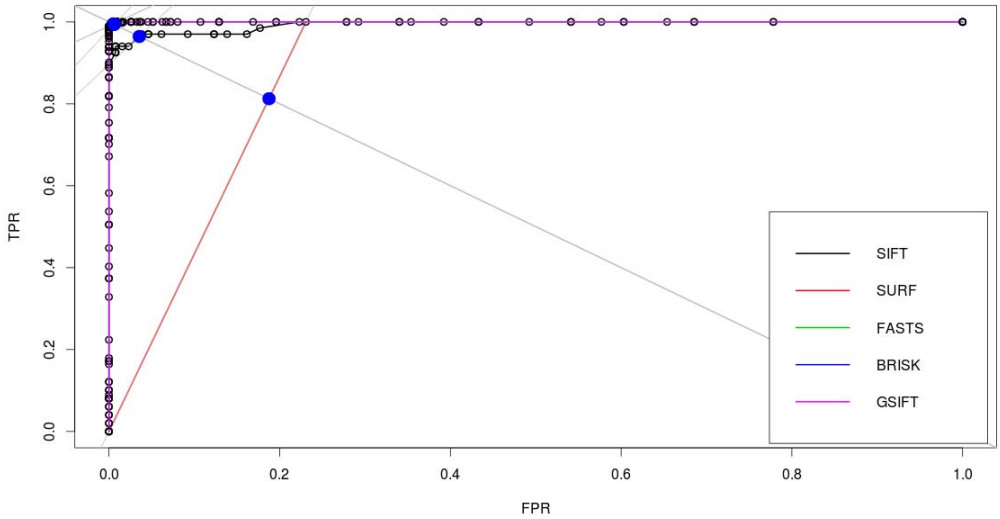


Figura 4.13 Reconocimiento de peatones en la base ITC con fusión.

Nuevamente, la base de datos CBCL presenta un mejor desempeño con todas las técnicas consideradas. Esto es comprensible, ya que se integra de imágenes de mejor calidad, con un tamaño uniforme arriba de los 100 pixeles de altura y, con menor variabilidad de iluminación y poses limitadas; logrando porcentajes de error de cero por ciento con las técnicas SURF y FASTS. También se observa el buen desempeño de las

técnicas ORB, FASTF y SIFT, cuyo valor de error es mínimo, siendo BRISK la técnica que menor eficiencia presenta, con un excelente 0.06% de EER y una AUC del 0.98%, ver tablas 4.2 y 4.3.

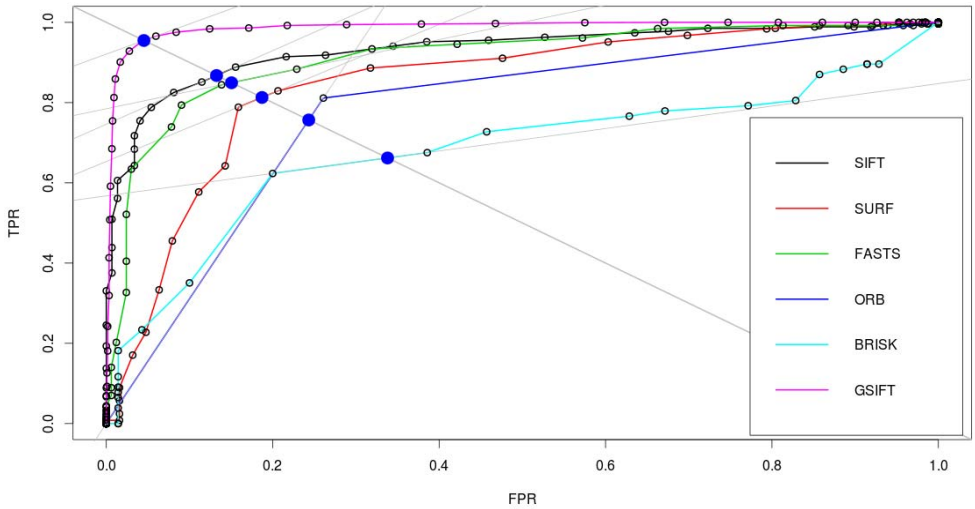


Figura 4.14 Reconocimiento de peatones en la base CVC01 con fusión.

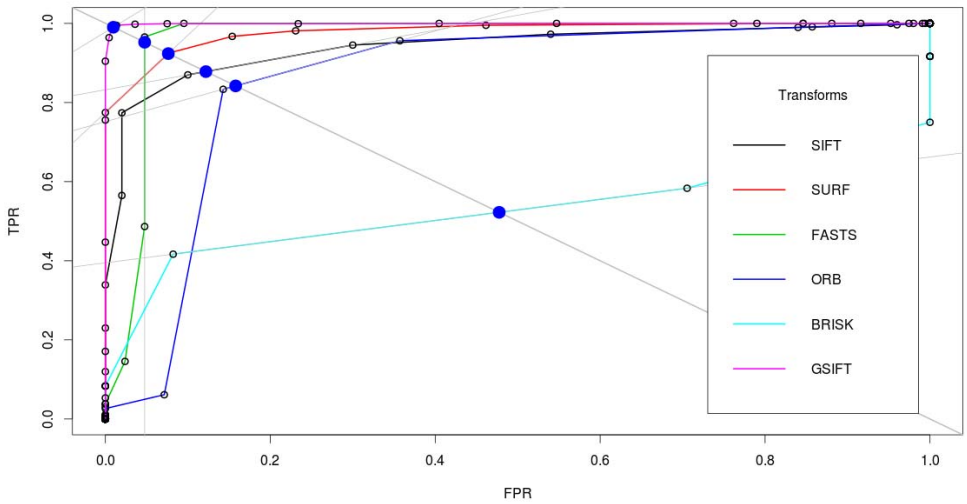


Figura 4.15 Reconocimiento de peatones en la base CVC02 con fusión.

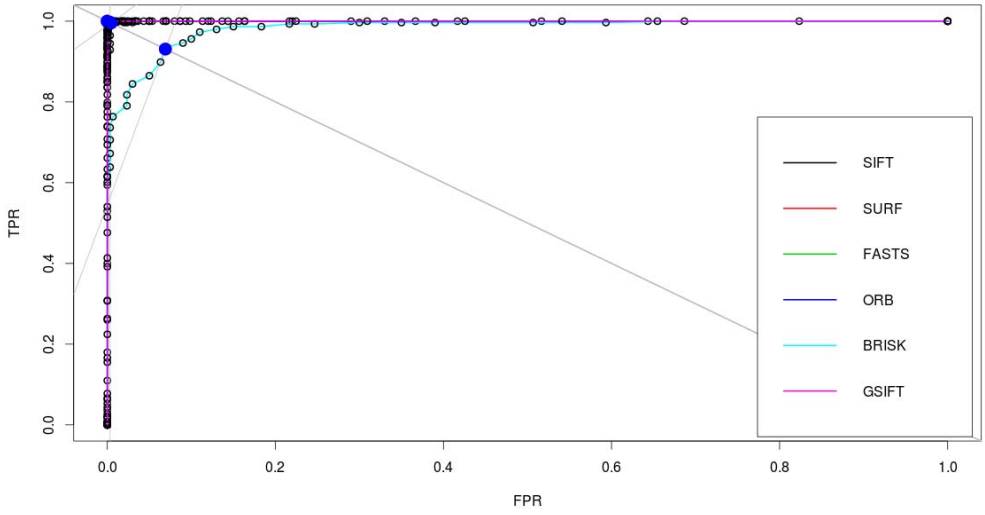


Figura 4.16 Reconocimiento de peatones en la base CBCL con fusión.

4.4.7 Análisis de resultados

Analizando los resultados obtenidos, de manera general, se puede concluir que el rendimiento de las todas las técnicas de descripción local evaluadas, se ve significativamente afectado por la escala o tamaño reducido de las imágenes. Sin embargo, este aspecto afecta más a las técnicas binarias evaluadas, provocando, la ausencia de detección de puntos de interés en un gran porcentaje de las imágenes de menor tamaño, como es el caso de ORB y BRISK.

El desempeño obtenido para la base de imágenes ITC a pesar de estar integrada de imágenes con problemas semejantes a las bases de datos anteriores, tiene ligeramente un mejor resultado que CVC02 y CVC01. Para entender los resultados de las bases de datos CVC01 y CVC02, es conveniente recordar que estas dos bases de datos contienen imágenes con mayor variabilidad en la iluminación, poses, entorno y sobre todo grandes cambios en el tamaño de las imágenes. Siendo precisamente la presencia de los grandes cambios en la escalas, el principal factor de error en la clasificación; ya que, no se logra obtener un modelo correcto y completo del objeto, siendo el clasificador incapaz de reconocerlo.

No obstante, si la base de datos se integra de imágenes de una escala pequeña, pero todas ellas normalizadas a un mismo tamaño o con una diferencia no tan grande, es posible obtener mejores modelos de aprendizaje, consiguiendo mejores resultados de detección. Una alta variabilidad en el tamaño de las imágenes, que conforman la base de datos como ocurre con las bases CVC01 y CVC02, afecta sustancialmente la etapa de aprendizaje al no obtener modelos con un porcentaje de generalización alto, consiguiendo un desempeño menor. La realidad, es que difícilmente, un sistema de videovigilancia tendrá normalizadas a un mismo tamaño, todas las imágenes a procesar.

La base CBCL permite verificar que el detector GSIFT no es la mejor opción si se cuenta con imágenes de buena calidad. Siendo una mejor opción SURF, FAST o ORB, en concordancia a lo reportado en la literatura. No obstante, GSIFT si mejora el desempeño de SIFT en este tipo de imágenes (excepto en CVC01, para descripciones locales).

Con respecto a las técnicas, el detector que muestra un mejor rendimiento general es FAST que ocupa el primer lugar en 3 de las 4 bases de datos (excepto para CVC01). Es decir, para imágenes de baja calidad y pequeñas escalas, su principal característica de ser repetible le ayuda a detectar más puntos de interés. FAST en combinación con los dos descriptores utilizados, muestra un mejor desempeño con FREAK. SURF trabaja bastante bien con imágenes de buena calidad, pero su rendimiento baja considerablemente ante la disminución de puntos detectados en imágenes de baja calidad y un menor tamaño. Los resultados también muestran que su descriptor sigue siendo competitivo pero tiene un menor desempeño ante FREAK.

Sin sorpresas, a pesar de no obtener los mejores resultados, el SIFT sigue siendo una buena opción, cumpliendo con las invarianzas geométricas prometidas. Para las técnicas binarias ORB y BRISK, esta comparativa, permite mostrar que si bien tienen una buena eficiencia ante imágenes de buena calidad, su desempeño no es el esperado ante imágenes pequeñas en las cuales obtienen pocos o nada de puntos de interés; no recomendando su uso en aplicaciones que requieran trabajar con estos aspectos.

Con el fin de evaluar el desempeño global de los resultados obtenidos, se realizó una clasificación de todos los algoritmos de detección y descripción local. En esta

comparativa se excluye la evaluación realizada a la base CBCL porque no se considera relevante debido a los excelentes resultados obtenidos por todas las técnicas evaluadas (por las características de las imágenes que la integran).

Esta comparativa se realizó siguiendo el método propuesto en (Kuncheva & Rodríguez 2007) que consiste en asignar un valor, a todos los métodos utilizados en cada una de las pruebas, de acuerdo a su posición o jerarquía de sus resultados con respecto a los obtenidos por las demás técnicas; es decir, se pondera con un valor de 1 para el mejor método y se le asigna el valor de 19 al peor método (el número total de modelos evaluados son 19). A continuación, se obtiene la desviación estándar de estas filas sobre el total de las 19 pruebas. La figura 4.17 presenta el rango promedio y los intervalos de confianza del 95% correspondiente a cada método.

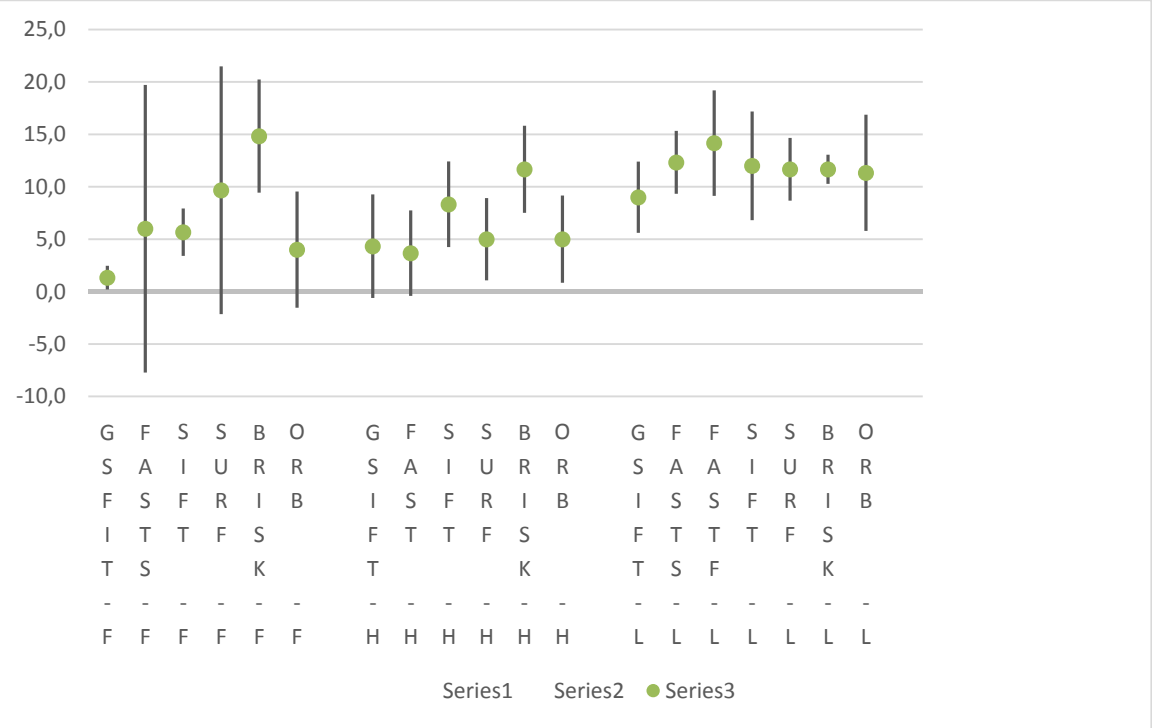


Figura 4.17 Rango promedio e intervalo de confidencia para el método propuesto GSIFT y las técnicas de descripción alternativas.

El método GSIFT logra el mejor rango promedio, de manera general. Por lo tanto, se puede resumir que el método propuesto tiene un buen desempeño, independientemente de las situaciones ambientales. El GSIFT es robusto y estable y se puede utilizar como

un detector local para imágenes de baja calidad e escalas pequeñas (aquellas con un número de pixeles menor a 30). SIFT muestra que sigue siendo competitivo, con resultados satisfactorios y estables. FAST es una buena alternativa para detectar puntos clave con el de imágenes bajo análisis. Sin embargo, como se puede ver en la figura 4.17, este método presenta la mayor variabilidad, y esto se debe a que su rendimiento depende de las condiciones de las imágenes (la iluminación, sombras, oclusión, etc.) es sensible al ruido y al método de descripción utilizado.

También, este estudio experimental muestra las propiedades de los momentos invariantes con respecto a la información local. Los momentos de Hu son capaces de obtener la relación espacial y/o geométrica de los puntos clave detectados. Como resultado, se mejora el rendimiento de la clasificación final, como se muestra en las tablas 4.2 y 4.3. Estas tablas resumen los resultados de los criterios EER y AUC, detallados en este análisis, para todas las bases de datos y técnicas evaluadas.

4.5 Conclusiones

En este capítulo se presentó un algoritmo de extracción de características locales en regiones de tamaño pequeño y baja calidad. Este algoritmo denominado GSIFT es un algoritmo que modifica la etapa de detección local de puntos del algoritmo SIFT al cambiar la función gaussiana por un banco de filtros Gabor. Los filtros Gabor han sido aplicados al análisis de imágenes, especialmente en investigaciones del rostro humano, debido a que proporcionan una completa representación de la imagen.

La experimentación realizada con cuatro bases de imágenes públicas en ambientes urbanos reales (no controlados) con imágenes de tamaño menor a 30 pixeles, permitió demostrar el buen funcionamiento del detector local propuesto. Además, se comparó su desempeño con seis técnicas de detección y descripción local que son punto de referencia por su buen desempeño en el área de Visión por Computador y en Reconocimiento de Patrones, tales como SIFT, SURF, FAST, ORB, BRISK y FREAK. Sin embargo, en esta tesis y de acuerdo a los resultados obtenidos se demostró el bajo rendimiento de estas técnicas cuando las imágenes no tienen un tamaño adecuado y una buena calidad. Un punto a resaltar es que los resultados son consistentes en las bases de

datos evaluadas, lo que sugiere que la evaluación no es dependiente de los datos utilizados.

En relación a la fusión de descriptores, la idea de combinar información global y local no es nueva, en la literatura se puede encontrar bastantes de ejemplos de cómo una representación que considera ambas características tiene un mejor desempeño en la clasificación. Sin embargo, lo novedoso de esta unión fue considerar los puntos detectados por parte de las técnicas de descripción local, y a partir de ellos, obtener dicha información holística. En esta tesis se aplicaron los momentos de Hu, pero pueden utilizarse cualquier otra técnica de descripción global.

Finalmente, y de acuerdo a los resultados obtenidos se puede concluir que la modificación del descriptor SIFT con filtros Gabor, en el detector local GSIFT, amplia y mejora la extracción de características en imágenes de escala pequeña, obteniéndose un detector local robusto y estable aun en regiones con un número de píxeles menor a 30. Sin embargo, GSIFT también tiene varias debilidades y es que para imágenes con un tamaño mayor de 80 pixeles de altura no es recomendable su uso, debido al alto costo computacional. Además, se debe realizar una experimentación más detallada en relación a los parámetros de los filtros Gabor.

Tabla 4.2 Comparación del desempeño (EER) de la descripción local, descripción global y fusión para cada método, en cada base de datos.

	ITC			CVC01			CVC02			CBCL		
	EER LOCAL	EER GLOBAL	EER FUSION	EER LOCAL	EER GLOBAL	EER FUSION	EER LOCAL	EER GLOBAL	EER FUSION	EER LOCAL	EER GLOBAL	EER FUSION
SIFT	0.2320	0.0447	0.0355	0.3166	0.2503	0.1325	0.3856	0.1715	0.1218	0.1362	0.0357	0.0033
SURF	0.3506	0.06	0.1875	0.3083	0.1407	0.1871	0.3040	0.0769	0.0763	0.0008	0.0776	**0.0**
FAST(S)	0.2982	**0.0083**	**0.0051**	0.3222	0.1411	0.1504	0.2795	0.0938	0.0476	0.0194	0.0375	**0.0**
ORB	---	---	---	0.4051	**0.1132**	0.2433	0.4999	0.1634	0.1579	**0.0**	0.0474	0.0033
BRISK	---	---	---	0.4866	0.3427	0.3380	0.4391	0.4166	0.4776	0.1432	0.1803	0.0691
FAST(F)	**0.0307**	---	---	0.4397	---	---	0.3070	---	---	**0.0**	----	---
GSIFT	0.1031	0.0203	0.0061	**0.2183**	0.1919	**0.0452**	**0.1993**	**0.0361**	**0.0096**	0.1751	**0.0163**	0.0044

Tabla 4.3 Área bajo la curva (AUC) de la descripción local, descripción global y fusión para cada método, en cada base de datos.

	ITC			CVC01			CVC02			CBCL		
	AUC LOCAL	AUC GLOBAL	AUC FUSION	AUC LOCAL	AUC GLOBAL	AUC FUSION	AUC LOCAL	AUC GLOBAL	AUC FUSION	AUC LOCAL	AUC GLOBAL	AUC FUSION
SIFT	0.8297	0.9738	0.9933	0.7463	0.8240	0.9317	0.6616	0.8910	0.9399	0.9333	0.9897	0.9998
SURF	0.6932	0.9941	0.8846	0.7590	0.8507	0.8470	0.7481	0.9709	0.9786	0.9999	0.9834	**1.0**
FAST(S)	0.7810	**0.9996**	**0.9999**	0.7172	0.8568	0.9119	0.7967	0.9234	0.9612	0.9972	0.9870	**1.0**
ORB	---	---	---	0.6024	**0.8955**	0.7752	0.500	0.8900	0.8558	**1.0**	0.9913	0.9997
BRISK	---	---	---	0.5183	0.6779	0.6822	0.5609	0.4372	0.5258	0.9250	0.9235	0.9827
FAST(F)	**0.9841**	---	---	0.5745	---	---	0.7543	---	---	**1.0**	---	---
GSIFT	0.9572	0.9963	**0.9999**	**0.8524**	0.8546	**0.9875**	**0.8736**	**0.9849**	**0.9992**	0.9013	**0.9958**	0.9998

Capítulo 5

Análisis de trayectorias

5.1 Introducción

El análisis de las trayectorias es el proceso de caracterización y entendimiento del comportamiento de cada objeto en una escena (Morris & Trivedi 2009). El reconocer lo que se considera "normal y anormal" se puede obtener a través del análisis de un conjunto representativo de trayectorias para obtener el modelo de las actividades presentes como comportamientos "normales" y asignar a cualquier instancia lejana a este modelo como un "evento anormal". Para el análisis y agrupamiento de las trayectorias, primero se define la medida de similaridad que se utilizará para comparar las trayectorias, obteniéndose una matriz de distancias. Posteriormente, se aplica un algoritmo de "clustering" para llevar a cabo el agrupamiento en *k* categorías de las trayectorias de acuerdo a su semejanza; y finalmente, se validan los clústeres obtenidos.

En este capítulo se detallan cada una de estas etapas mencionadas y se presenta el algoritmo de agrupamiento propuesto, pamTOK (pam Tree Out K). Este algoritmo estima de manera automática el número de categorías a partir del análisis de la homogeneidad interna de los clústeres y su separación espacial con respecto a los otros grupos, permitiendo identificar aquellas trayectorias poco frecuentes o inusuales.

Además, se presentan diversas pruebas con el objetivo de evaluar el rendimiento del algoritmo propuesto y para ello se verifica su desempeño en la estimación de grupos con bases de datos que cuentan con su verdadero etiquetado; posteriormente se verifica su funcionalidad respecto a la identificación de trayectorias anormales con bases de datos de trayectorias reales. También se detallan las medidas de distancia y criterios de validación utilizados en la experimentación. Todas las pruebas son realizadas y

comparadas con otro algoritmo de agrupamiento denominado pamk (Kaufman & Rousseeuw 1987).

5.2 Medidas de distancia

Debido a la naturaleza variante en el tiempo de las trayectorias, estas tienen diferente tamaño, lo que representa un problema para compararlas. Una propuesta para eliminar este inconveniente, es utilizar medidas de similaridad que sean independientes de la longitud que estas tengan. Las medidas de distancia que se seleccionaron en este trabajo son: DTW (Rabiner & Juang 1993), LCSS (Vlachos et al. 2002); ERP (Chen & Ng 2004) y EDR (Chen et al. 2005). En la selección se consideró los resultados reportados en la literatura sobre su robustez a los efectos del ruido y a su manejo con penalidad a las partes de las trayectorias sin correspondencia. Las distancias DTW y LCSS son técnicas frecuentemente utilizadas en la comparación de trayectorias, y su desempeño representa un punto de comparación para determinar la eficiencia de las distancias EDR y ERP que son menos conocidas y con ello menos utilizadas.

La trayectoria S de un objeto se define como la secuencia de pares,

$$s = [(t_1, s_1), \ldots, (t_n, s_n)]$$

la cual muestra sucesivas posiciones s_i del objeto moviéndose en un periodo de tiempo t_i. Dónde, (t_i, s_i), es un elemento de la trayectoria, n es el número de eventos que conforman S y se define como la longitud de S; s_i es un vector de dimensión d (d usualmente es igual a 2 o a 3), que indica un evento particular ocurrido en el intervalo de tiempo t_i. Por lo tanto, las trayectorias pueden ser consideradas como series en el tiempo con dimensión dos (plano x,y) o tres (plano x,y,z) (Chen et al. 2005).

La implementación de las medidas de distancia se realizó considerando las trayectorias como series bivariables. Es decir, se asume que los objetos son puntos que se mueven en un espacio de dos dimensiones *(x,y)* y que el tiempo es discreto. Así, dada una trayectoria s_i es un par (s_{ix}, s_{iy}).

Dadas dos trayectorias *R* y *S*, la distancia *Dinamic Time Warping* (DTW) se define como:

$$DTW(R,S) = \begin{cases} 0 & Sí\ m = n = 0 \\ \infty & Sí\ m = 0\ o\ n = 0 \\ dist(r_1, s_1) + min\begin{Bmatrix} DTW\big(Rest(R), Rest(S)\big), \\ DTW(Rest(R), S), DTW\big(R, Rest(S)\big) \end{Bmatrix} & de\ otra\ forma \end{cases} \tag{5.1}$$

$$dist(r_i, s_i) = \left(r_{i,x} - s_{i,x}\right)^2 + \left(r_{i,y} - s_{i,y}\right)^2 \tag{5.2}$$

La distancia *Local Common SubSecuente* (LCSS) se define como:

$$LCSS(R,S) = \begin{cases} 0 & Sí\ m = 0\ o\ n = 0 \\ LCSS\big(Rest(R), Rest(S)\big) + 1 & Sí\ \ dist(r_1, s_1) < \varepsilon \\ max\{\{LCSS(Rest(R),S), LCSS(R, Rest(S))\} & en\ otro\ caso \end{cases}$$

$$D_{LCSS} = 1 - \frac{LCSS(R,S)}{\min(m,n)} \tag{5.3}$$

Dónde *dist(r_1,s_1)* es como ecuación 5.2, y *m,n* es el número de elementos de las trayectorias R y S respectivamente.

La distancia *Edit Distance with Real Penalty (*ERP), se define como:

$$ERP(R,S) = \begin{cases} \sum_{1}^{n} dist_{erp}|r_i - g| & Sí\ m = 0 \\ \sum_{1}^{m} dist_{erp}|s_i - g| & Sí\ n = 0 \\ min\begin{Bmatrix} ERP\big(Rest(R), Rest(S)\big), +dist_{erp}(r_i, s_i), \\ ERP(Rest(R), S), +dist_{erp}(r_i, g), \\ ERP\big(R, Rest(S)\big), +dist_{erp}(s_i, g) \end{Bmatrix} & en\ otro\ caso \end{cases} \tag{5.4}$$

$$dist_{erp}(r_i, s_i) = \begin{cases} |r_i - s_i| & Sí\ r_i\ y\ s_i\ no\ son\ huecos \\ |r_i - g| & Sí\ \ s_i\ es\ un\ hueco \\ |s_i - g| & Sí\ r_i\ es\ un\ hueco \end{cases} \tag{5.5}$$

La distancia *Edit Distance on Real Sequences* (EDR) se define como:

$$EDR(R,S) = \begin{cases} n & Si\ m = 0 \\ m & Sí\ n = 0 \\ min\begin{Bmatrix} EDR\big(Rest(R), Rest(S)\big) + subcost, \\ EDR(Rest(R), S + 1, \\ EDR(R, Rest(S) + 1 \end{Bmatrix} & en\ otro\ caso \end{cases} \tag{5.6}$$

$$subcost = \begin{cases} 0 & Sí\ |r_i - s_i| \leq \varepsilon \\ 1 & Sí\ r_i\ o\ s_i\ es\ uh\ hueco \\ 1 & en\ otro\ caso \end{cases} \tag{5.7}$$

5.3 Algoritmo pamTOK

El algoritmo pam Tree Out K o pamTOK, como su nombre lo indica, se basa en algoritmo de agrupamiento pam. La idea de su funcionamiento es sencilla. Dada una matriz de distancias, el algoritmo consiste en dividir el conjunto de datos en dos partes, a través del algoritmo pam, y evaluar en cada una de ellas la homogeneidad interna del grupo y su separación espacial con respecto al grupo. En cada conjunto nuevo de datos se repetirá esta separación y evaluación hasta que la condición de división ya no sea posible, generándose un árbol binario, donde las hojas representan los grupos finales y su número el valor de K. A continuación, se detallan la información necesaria para el algoritmo, los pasos del algoritmo y en la figura 5.1 se resume en un diagrama de flujo.

5.3.1 Algoritmo pam*:*

El algoritmo Partiton Around Medoids (partición alrededor de medoides) o PAM (Kaufman & Rousseeuw 1987) busca *k* objetos representativos (medoides) que se encuentran centrados en los conglomerados que ellos definen. Su objetivo es minimizar la suma de disimilitudes de los objetos con respecto a su objeto seleccionado como medoide; en otras palabras, el medoide, es aquel objeto para el cual la diferencia promedio con todos los objetos en el conglomerado es mínima. Su funcionamiento consiste de dos etapas principales:

1. La primera parte denominada *Build-phase* tiene precisamente el objetivo de construir los grupos, especificados con el valor de K. Del conjunto de datos se seleccionan k objetos representativos para construir los k grupos. El proceso de selección de estos objetos o medoides es aquel que cumple con tener el promedio mínimo de disimilaridad con respeto a cada objeto del conjunto de datos.

2. La segunda fase, denominada SWAP, busca mejorar la calidad del grupo obtenido mediante el intercambio de los objetos representativos o medoides por un objeto (diferente al medoide) perteneciente al mismo grupo y no seleccionado antes, que reduce la función objetivo. Este proceso es iterativo y en cada paso, PAM selecciona el objeto que hace decrecer la suma de disimilitudes tanto como sea posible. Termina cuando la función objetivo deja de disminuir.

5.3.2 Obtención de coeficientes

Al finalizar el proceso de división en dos grupos mediante el algoritmo *pam*, se tienen los siguientes coeficientes:

1) Valor denominado *Silhouette* que proporciona una medida de la similaridad o disimilaridad de cada elemento con respecto a su grupo. El valor de la silueta se define mediante la ecuación 5.8.

$$s(i) = \frac{(b(i) - a(i))}{\max\{\,\mathrm{a(i)},\ \mathrm{b(i)}\}} \tag{5.8}$$

Dónde:
a(i) es el promedio de la disimilaridad del objeto *i* con respecto a los demás elementos del mismo grupo. *b(i)* es el promedio de las diferencias de *i* con todos los objetos del grupo más cercano para él. Es decir, el grupo *b* sería la segunda opción de pertenencia para el objeto *i*. *s(i)* toma un rango de valores [-1; 1] y su interpretación es la siguiente:

s(i) = 1 El elemento *i* se ha asignado al grupo apropiado.

s(i) = 0 El elemento *i* se encuentra a una distancia intermedia entre dos grupos.

s(i) = -1 El elemento fue mal agrupado.

2) *Diámetro del clúster*, para cada grupo C_k, se denota por D_k como la distancia más grande que separa a dos objetos dentro del grupo.

$$D_k = \max_{\substack{i,j \in I_k \\ i \neq j}} \left\| M_i^{\{K\}} - M_j^{\{K\prime\}} \right\| \tag{5.9}$$

3) *Separación del clúster,* se define como la menor disimilitud entre dos objetos pertenecientes a dos grupos diferentes. La distancia entre los clústeres C_k y C_{k0} es la medida de la distancia entre sus puntos más cercanos.

$$d_{kk\prime} = \min_{\substack{i \in I_k \\ j \in I_{k\prime}}} \left\| M_i^{\{K\}} - M_j^{\{K\prime\}} \right\| \tag{5.10}$$

4) *Silueta de cada grupo* (*Average Silhouette Width*, ASW). Se obtiene del promedio de todas las observaciones *s(i)* obtenidas por los elementos que pertenecen al grupo. Se tiene para cada uno de los *k* grupos.

5) *Coeficiente de Silueta* (SC) o promedio máximo de la Silueta (*Average Width),* este valor se obtiene promediando la silueta de todos los *k* grupos, en los cuales se realizó la separación de los datos. Este valor se puede utilizar como métrica para evaluar la calidad o estructura del agrupamiento.

$$SC = max_k \, S_k \tag{5.11}$$

En (Kaufman & Rousseeuw 1987) se propuso la interpretación del coeficiente SC, ver tabla 5.1.

Tabla 5.1 Interpretación de coeficiente de silueta.

SC	Interpretación
0.71 - 1.00	Se encontró una estructura fuerte.
0.51 – 0.70	Se encontró una estructura razonable
0.26 – 0.50	La estructura es débil y podría ser artificial, se debe tratar métodos adicionales con estos datos.
≤ 0.25	No hay estructura substancial.

5.3.3 Criterio de agrupamiento

Para validar la creación o no de un clúster, se debe determinar si existe o no una estructura en los datos e identificar si los grupos encontrados por el algoritmo son

"verdaderos". En la literatura, existen algunos criterios para la estimación del número de grupos presentes en el conjunto de datos, en (Desgraupes 2013) se puede encontrar gran variedad de propuestas.

Nuestro enfoque para evaluar la validez de cada clúster, considera la homogeneidad interna de los clústeres y su separación espacial con respecto a los otros grupos. Es decir, se obtiene un índice, a través de la relación de la cohesión entre los elementos de datos (como lo es la distancia máxima en el grupo) con respecto a la separación mínima de este clúster con respecto a los otros grupos.

Denotando a d_{min} como la distancia mínima entre puntos de diferentes grupos (separación del clúster) y d_{max} como la distancia más grande dentro del clúster (diámetro del clúster), se puede obtener el índice Dunn (Dunn 1974) es definido como el cociente de d_{min} y d_{max}:

$$IndDunn = \frac{d_{min}}{d_{max}} \tag{5.12}$$

A través de pruebas experimentales realizadas, se encontró que intercambiando los valores del índice (se denominó Índice Dunn Invertido, IDI), ecuación 5.13, se puede evaluar la estructura de los grupos buscando tener una estructura interna fuerte. Por ello, se propone en la tabla 5.2, la interpretación de los posibles valores a obtenerse.

$$IDI = \frac{d_{max}}{d_{min}} \tag{5.13}$$

Tabla 5.2 Interpretación de Índice Dunn invertido.

IDI	Interpretación
≥ 2.0	Clúster con estructura dispersa. Dividir nuevamente.
≥ 1.0 - < 2.0	Estructura de clúster razonable.
< 1.0	Clúster con fuerte homogeneidad interna. Ya no es necesario dividir.

5.3.4 Pasos del algoritmo PAMTOK

El algoritmo requiere los siguientes parámetros de entrada.

1. Valor de umbral para el índice Dunn Invertido (IDIε)
2. Coeficiente de Silhouette: (SCε)

3. Número mínimo de elementos por clúster, valor necesario para asegurar que es posible dividir el grupo en 2 (MIN).

El algoritmo pamTOK, ver Figura 5.1, consiste en:

1. Dada una matriz de distancias de las trayectorias *MD: n x n* y los valores de los parámetros anteriores.

2. Se aplica pam con $K=2$.

3. Sólo la primera vez que se divide, se verifica si el valor obtenido por el Coeficiente de la Silhouette es mayor o igual al umbral especificado $SC_i \geq SC\varepsilon$. Esta condición es necesaria, porque, algunas en ocasiones sucede que una simple división en 2 resulta en estructuras fuertes, siendo un resultado no óptimo. Si se cumple la condición, regresar al paso 2 con un valor mayor para dividir: $K > 2$.

4. Se evalúa para cada grupo si el índice Dunn invertido obtenido es menor al especificado como umbral: $IDI_i < IDI\varepsilon$.

5. El grupo que cumplen la condición, se almacena y etiqueta como hoja.

6. Aquel grupo o nodo que no cumpla la condición anterior y cuya cardinalidad sea mayor al número mínimo de elementos Size > MIN, se vuelve a dividir en dos, creándose el árbol binario.

7. El algoritmo es recursivo y termina cuando ya no haya grupos o nodos que dividir.

Al terminar el algoritmo, se cuenta el número de hojas obtenidas que corresponde al valor estimado de *K*, los representantes (medoides) de cada grupo y su tamaño o número de elementos, el valor de la silueta del agrupamiento y para cada elemento su pertenencia.

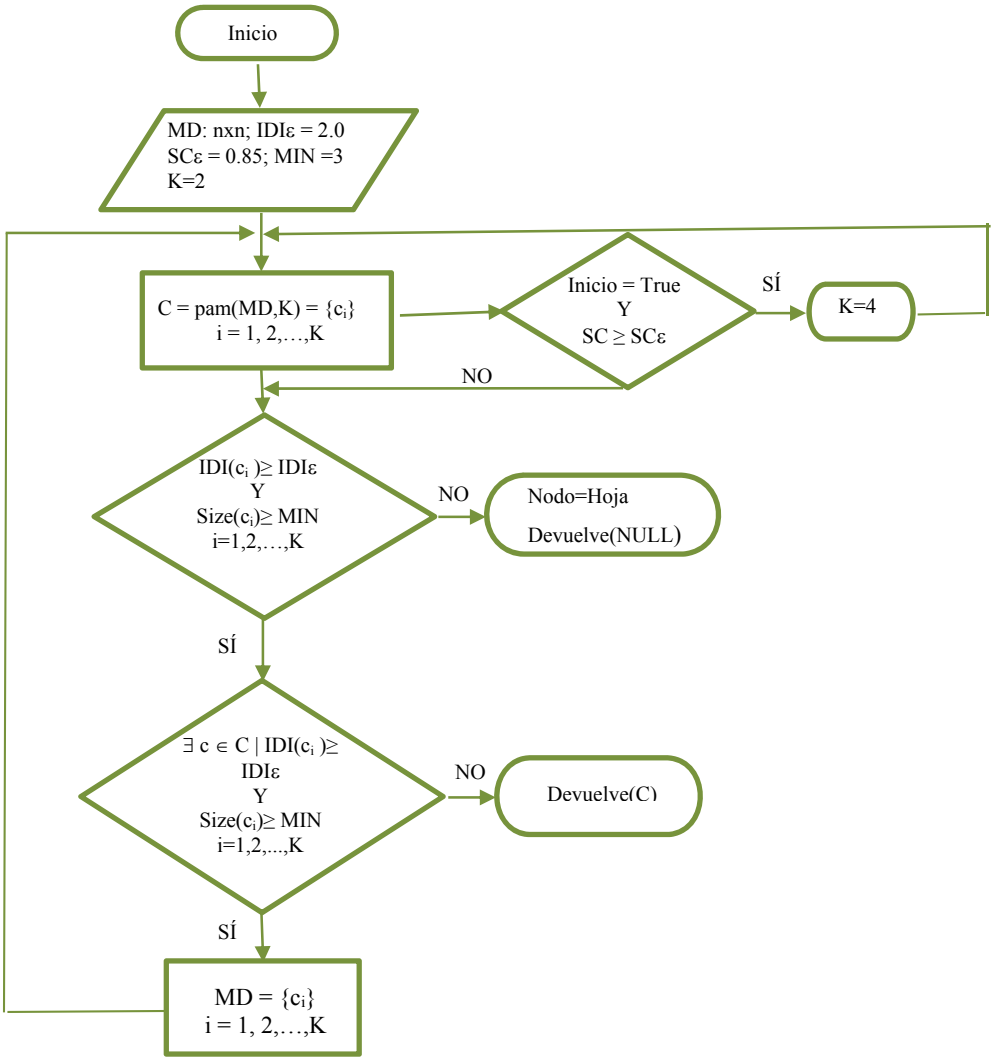


Figura 5.1 Diagrama del algoritmo pamTOK.

5.4 Experimentación

5.4.1 Plan de pruebas

Se llevaron a cabo una serie de pruebas con el objetivo de evaluar y validar los siguientes procesos:

a) Evaluar el rendimiento del algoritmo de agrupamiento propuesto pamTOK, con respecto a la estimación del número de categorías presentes en un conjunto de trayectorias.

b) Validar el desempeño del algoritmo pamTOK, en el agrupamiento de las actividades normales y la identificación de comportamientos inusuales.

c) Determinar el desempeño de las medidas de distancia, en la detección de trayectorias anormales.

Para la realización de las pruebas, el primer paso es obtener las matrices de distancia de cada una de las cuatro distancias consideradas en este estudio.

Las distancias LCSS y EDR requieren se le especifique el valor de umbral a partir del cual, se considera la similaridad y disimilaridad, de dos trayectorias. Tomando en cuenta que los datos de las trayectorias son posiciones (en pixeles) se determinaron los valores de umbral de 5, 10 y 20. Es decir, se considera que dos trayectorias tienen una ruta similar, si su diferencia (en el plano *x,y*) es menor al umbral especificado. La distancia ERP requiere solamente se especifique el peso con el cual se castigará a los huecos encontrados entre las dos trayectorias comparadas, a este peso se le dio el valor de 1. La distancia DTW no requiere de ningún tipo de parámetro de entrada. En total, se tienen 3 combinaciones de matrices de distancia para LCSS, 3 combinaciones para EDR, una para ERP y una para la distancia DWT; estas medidas se aplican a cada una de las ocho bases de trayectorias consideradas, obteniendo un total de 64 matrices de distancia. De estas matrices, 48 son consideradas en el agrupamiento de las trayectorias y 16 para la experimentación relacionada a la detección de comportamientos anormales.

Se consideró importante comparar el resultado obtenido por pamTOK con otro algoritmo de agrupamiento que estimara la cantidad de grupos. El método seleccionado fue el algoritmo de clustering *pamk* por *Partitioning around medoids with estimation of number of clusters* (Hennig 2010). El algoritmo *pamk*, también llama al algoritmo pam para realizar el agrupamiento, estimando el número de grupos a través del valor más alto de la silueta del agrupamiento (SC), en un rango de 2 a 10 grupos (valor por default). Es posible especificarse un rango mayor.

Es importante mencionar que durante el proceso de agrupamiento, los conjuntos de datos no presentan información sobre su pertenencia a algún grupo. La información del etiquetado verdadero o la información sobre el número de trayectorias inusuales se utilizan solamente en la parte de validación, para calcular los criterios que se explican a continuación.

5.4.2 Criterios de evaluación

Para validar de manera cuantitativa el desempeño del algoritmo pamTOK, se utiliza el índice de agrupamiento correcto (Correct Clustering Rate, CCR), definido en la ecuación 5.14 y la exactitud en la detección (Detection ACCuracy, DACC) definida en la ecuación 5.15. El CRR es utilizado para calcular el desempeño de la clasificación de las rutas y es utilizado con las bases de datos pertenecientes al grupo CVRR (I5sim, Cross, Labomni, etc.). La DACC es utilizada para evaluar el desempeño del algoritmo en la detección de los comportamientos: "normal-anormal" y es posible su uso con la base de datos Bard. En ambos criterios, los resultados toman un rango de valores de [0,1], donde un valor alto significa un buen desempeño (Li et al. 2013).

$$CCR = \frac{1}{N}\sum_{c=1}^{K} p_c \tag{5.14}$$

Donde N es el número total de trayectorias y p_c denota el número total de trayectorias correctamente asignadas al cluster c.

$$DACC = \frac{VP + VN}{Total\ de\ Trayectorias} \tag{5.15}$$

Donde VP (Verdaderos Positivos) es el número de trayectorias correctamente categorizadas como "comportamiento normal", y VN (Verdaderos Negativos) es el número de trayectorias correctamente detectadas en la categoría de "comportamientos anormales".

En la base de datos Bard, se conoce cuáles trayectorias tienen un comportamiento anormal o diferente, permitiendo calcular la especificidad (SPC) o razón de verdaderos negativos, que se obtiene de considerar los verdaderos negativos detectados (VN) entre el total de objetos pertenecientes al grupo de verdaderos negativos (Verdaderos Negativos más Falsos Positivos), ecuación 5.16. Se recomienda ver la tabla 4.1 de confusión, sección 4.4.2.

$$SPC = \frac{VN}{VN + FP} \tag{5.16}$$

Se propuso un criterio más para comparar el resultado de las bases de datos que no cuentan con un etiquetado verdadero como lo son las bases Bard y Edimburgo. Esta razón consiste en comparar la detección de trayectorias anormales realizada por los algoritmos pamTOK y pamk. Entonces, para validar la coincidencia de resultados obtenidos, se utiliza la razón definida en la ecuación 5.17.

$$RazónOut = \frac{\left|intersección(Outliers_{pamTOK},\ Outliers_{pamk})\right|}{\left|unión(Outliers_{pamTOK},\ Outliers_{pamk})\right|} \tag{5.17}$$

Los valores para su cálculo se obtienen de una matriz de confusión similar a la mostrada en la sección 4.4.2, tabla 4.1, y se puede observar en la figura 5.2, dónde:

- La letra *A* representa las trayectorias detectadas como anormales por ambos algoritmos.
- La letra *B* representa aquellas trayectorias detectadas como normales por pamTOK, pero anormales por pamk.
- La letra *C* representa a las trayectorias etiquetadas por pamTOK como anormales, pero pamk las considera como normales.
- La letra *D* se integra de todas las trayectorias consideradas por ambos algoritmos como normales.

Tabla 5.3 Matriz de confusión utilizada para comparar la detección de trayectorias normales y anormales, realizada por los algoritmos pamTOK y pamk.

	RazónOut	**pamTOK**	
		Outliers	**NO.outliers**
pamk	**Outliers**	A	B
	NO. outliers	C	D

5.4.3 Agrupamiento de trayectorias

Cómo ya se mencionó, el objetivo de esta prueba es evaluar desempeño del algoritmo pamTOK, estimando el número de grupos *K* óptimo, mediante el agrupamiento de trayectorias. Con esta finalidad, se seleccionaron seis bases de datos públicas de trayectorias que proporcionaran su etiquetado verdadero (*ground truth)*, de tal forma que se pueda verificar cuantitativamente el desempeño del algoritmo pamTOK. Los conjuntos de datos utilizados para este fin, pertenecen a tres escenarios distintos: las

trayectorias de vehículos en una autopista (bases de datos I5, I5sim, I5sim2, I5sim3), a un crucero de cuatro vías (base Cross) y a un laboratorio (labomni).

Los valores de los parámetros que se utilizan en pamTOK se obtuvieron de manera experimental y son: IDIε =2, SCε=0.85 y MIN=3. Para el algoritmo pamk se utilizó un rango de 2 a 20.

1. *Base de trayectorias I5*

Recordando, la base de datos I5 es una base de trayectorias reales que tienen como escenario una autopista de 8 carriles, siendo este su valor óptimo para K. La tabla 5.4 muestra los resultados del agrupamiento de los algoritmos pamTOK y pamk. En ella se puede ver el valor de K estimado y el valor de CCR obtenido en cada una de las nueve matrices de distancia.

Analizando los resultados se observa que, los grupos no son tan homogéneos como podría pensarse por el escenario. El algoritmo pamTOK, con las distancias EDR-10 y LCSS-10, obtiene un valor *k* cercano al esperado, con un CCR del 0.87% y 0.9% respectivamente. Incluso para el algoritmo pamk que evalúa cada uno de los valores *k* del rango especificado, sólo con la distancia LCSS-10 obtiene la separación correcta en ocho grupos. Llama la atención que, con las demás distancias, sólo se alcanza un agrupamiento en dos clases, es decir, proponiendo una asociación que sólo considera la orientación de las trayectorias. En la figura 5.2 se puede ver que las trayectorias superiores, se encuentran espacialmente muy cercanas, mientras que las inferiores, presentan una separación mayor.

Tabla 5.4 Resultados de agrupamiento de la base I5.

I5 /8K	**pamTOK** IDIε >= 2.5		**pamk** 2:20	
	K	**CCR**	**K**	**CCR**
DTW	11	0.599	2	0.313
EDR5	5	0.531	5	0.531
EDR10	**9**	**0.873**	11	0.844
EDR20	6	0.735	2	0.285
ERP1	12	0.614	2	0.307
LCSS5	3	0.398	5	0.641
LCSS10	**9**	**0.908**	**8**	**0.964**
LCSS20	5	0.665	6	0.781

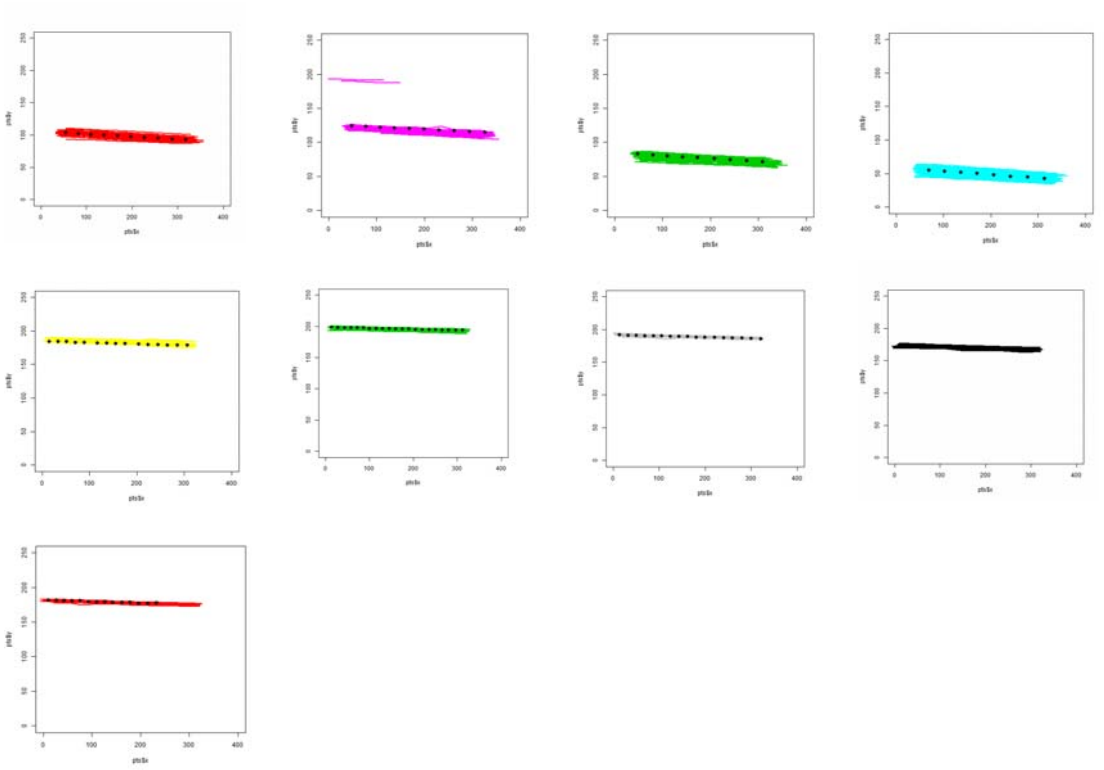


Figura 5.2 Ejemplo de agrupamiento de la base I5 con pamTOK y distancia LCSS-10.

2. *Bases de trayectorias I5sim, I5sim2 y I5sim3*

Las bases I5sim, I5sim2 y I5sim3 son conjuntos de trayectorias simuladas, de un escenario de autopista, de cuatro carriles en cada una de las dos orientaciones. La característica distintiva de I5sim, con respecto a las otras dos bases, es que es los recorridos tienen un flujo libre. El conjunto de datos para I5sim2 y I5sim3 es el mismo, realmente la diferencia se encuentra en el etiquetado, que para I5sim3 toma en cuenta la velocidad.

Las tablas 5.5 a la 5.7 muestran los resultados del agrupamiento con los algoritmos pamTOK y pamk, con respecto a las nueve matrices de distancia. La primera columna, para cada algoritmo, señala el número de grupos estimados para esa matriz de distancia. La columna etiquetada con CCR indica la calidad obtenida en la separación de los datos y la columna con etiqueta SC muestra el valor del coeficiente de silueta obtenido por el agrupamiento en su totalidad.

En la base I5sim, la tabla 5.5, el algoritmo pamTOK logra separar de manera correcta el conjunto de trayectorias, con las distancias EDR-5 y LCSS-5. Además, las mismas

distancias con un valor de umbral mayor, tienen un aceptable valor de CCR, con valores de K cercanos al deseado (excepto LCSS-20). La figura 5.3 muestra el agrupamiento de la base I5sim con pamTOK y distancia LCSS5. Los puntos en las gráficas muestran el medoide o representante de la ruta.

Considerando el valor mayor del SC, el algoritmo pamk, obtiene un excelente resultado con la mayoría de las distancias, mostrando que los grupos en esta base de datos, se encuentran bien separados. La distancia ERP obtiene un desempeño inferior en ambos algoritmos. Por otra parte, llama la atención como la distancia DTW tiene resultados tan diferentes, mostrando que el valor del umbral IDIε, no es el adecuado para ella en esta prueba.

Tabla 5.5 Resultados de agrupamiento de la base I5sim.

I5SIM /8K	**pamTOK** IDIε >= 1.2			**pamk** 2:20		
	K	**CCR**	**SC**	**K**	**CCR**	**SC**
DTW	30	0.326	0.526	8	1.0	0.778
EDR5	**8**	**1.0**	**0.847**	8	1.0	**0.847**
EDR10	10	0.916	0.732	8	1.0	0.816
EDR20	11	0.897	0.711	8	1.0	0.802
ERP1	24	0.547	0.628	2	0.251	0.765
LCSS5	**8**	**1.0**	**0.9289**	**8**	**1.0**	**0.928**
LCSS10	9	0.948	0.789	8	1.0	0.871
LCSS20	4	0.501	0.672	8	1.0	**0.881**

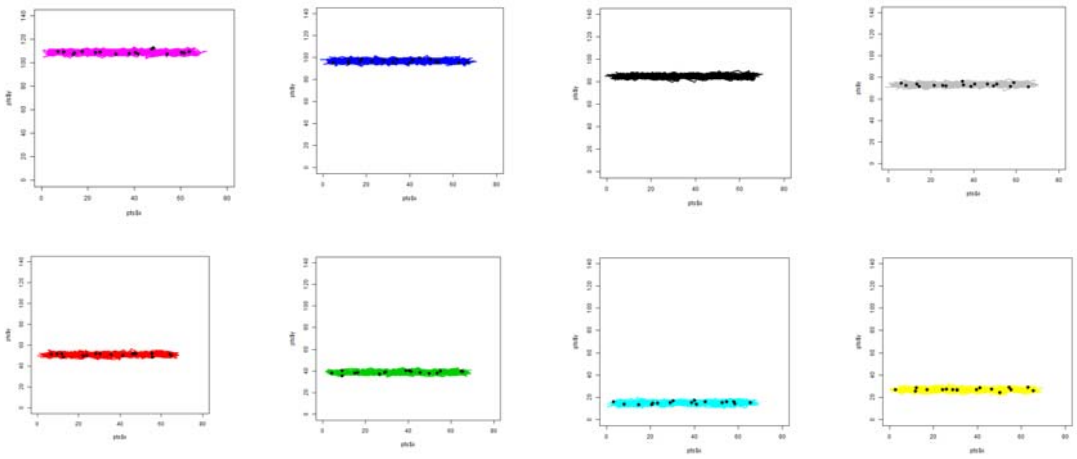


Figura 5.3 Ejemplo de agrupamiento de la base I5sim con pamTOK y distancia LCSS5.

3. *Bases de trayectorias I5sim2 y I5sim3*

Como ya se mencionó, las bases de datos I5sim2 y I5sim3 pertenecen al mismo conjunto de trayectorias. Sin embargo, el etiquetado de las trayectorias de la base I5sim3 considera las diferencias en dirección y velocidad presentes.

Los resultados obtenidos, muestran que el algoritmo pamTOK, con la distancia LCSS-5 estima el valor de *k*, considerando sólo la dirección y da como resultado los ocho grupos de la base I5sim2. Sin embargo, la misma distancia LCSS pero con un valor de umbral de 10 y 20, es capaz de encontrar las diferencias en velocidad también. En las tablas 5.6 y 5.7 se puede observar como el valor del CCR y de la silueta cambia dependiendo del etiquetado; alcanzando para la base I5sim3 con pamTOK valores arriba de 0.94%, cercanos a los valores obtenidos por pamk que logra un 1.0% de CCR con LCSS-5.

Tabla 5.6 Resultados de agrupamiento de la base I5sim2.

I5SIM2 /8K	**pamTOK** IDIε >= 1.7			**pamk** 2:20		
	K	**CCR**	**Sil-Clust**	**K**	**CCR**	**Sil-Clust**
DTW	**9**	**0.936**	0.624	2	0.250	0.658
EDR5	6	0.368	0.438	20	0.502	0.707
EDR10	9	0.507	0.353	20	0.501	0.643
EDR20	20	0.501	0.626	20	0.501	0.626
ERP1	5	0.331	0.513	2	0.188	0.623
LCSS5	**8**	**1.0**	**0.689**	16	0.5	0.833
LCSS10	17	0.5	0.745	16	0.5	0.764
LCSS20	18	0.5	0.761	16	0.5	0.787

Tabla 5.7 Resultados de agrupamiento de la base I5sim3.

I5SIM3 /16K	**pamTOK** IDIε >= 1.7			**pamk** 2:20		
	K	**CCR**	**Sil-Clust**	**K**	**CCR**	**Sil-Clust**
DTW	9	0.561	0.624	2	0.125	0.658
EDR5	6	0.361	0.438	20	0.899	0.707
EDR10	9	0.556	0.353	20	0.902	0.643
EDR20	20	0.896	0.626	20	0.896	0.626
ERP1	5	0.275	0.513	2	0.125	0.620
LCSS5	8	0.500	0.689	**16**	**1.0**	0.833
LCSS10	**17**	**0.976**	**0.745**	**16**	**0.999**	0.764
LCSS20	**18**	**0.941**	**0.761**	**16**	**0.999**	0.787

La figura 5.4 muestra las 17 rutas encontradas por pamTOK con LCSS-10. Como se puede apreciar, las rutas o grupos presentan un traslape espacial, debido a que, considerando los aspectos de la orientación y posición se tienen 8 grupos (I5sim2). Las categorías dispuestas en la misma posición, pero separadas en dos, muestran precisamente esa diferencia de velocidad presente en los datos. Las líneas rojas de las gráficas muestran recorridos etiquetados de manera errónea, de acuerdo al etiquetado verdadero.

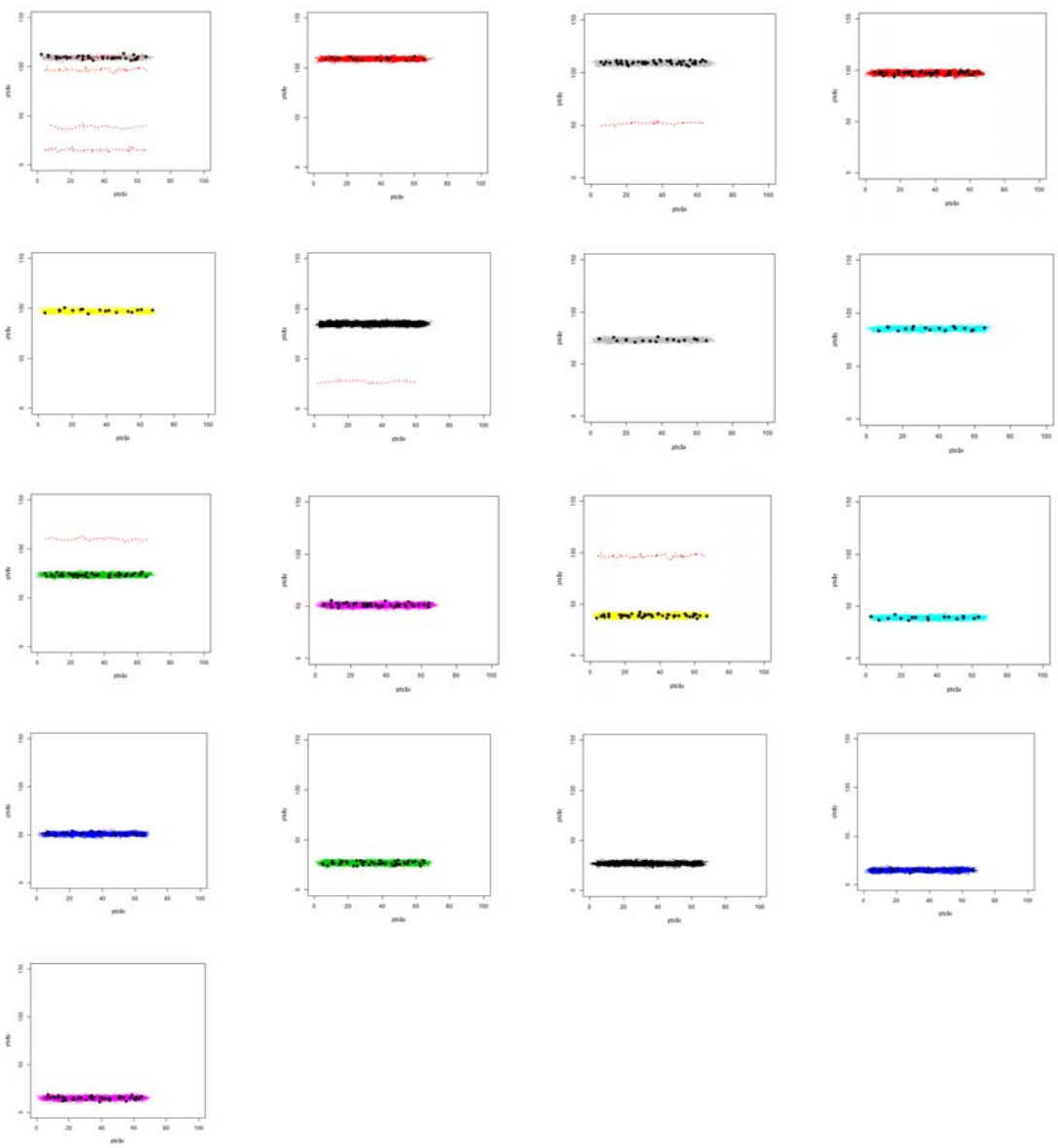


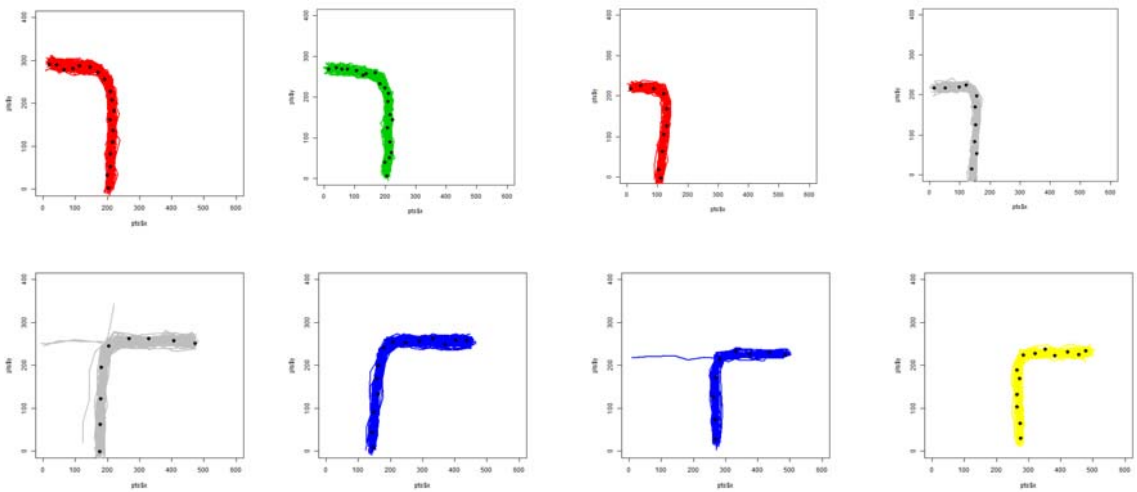
Figura 5.4 Ejemplo de agrupamiento de la base I5sim3 con pamTOK y distancia LCSS-10.

4. *Base de trayectorias Cross*

Esta base de trayectorias es más compleja que las anteriores, al presentar recorridos con giros, turnos, etc. Sin embargo, en la tabla 5.8 se puede observar que pamk obtiene el valor esperado de *k=19* con la distancia ERP. Por su parte, pamTOK, logra un 0.90% de CCR con la distancia LCSS-20, con 21 grupos. Es importante hacer notar la relación que existe entre los resultados de los dos algoritmos de agrupamiento. Ya que, a pesar de que pamk realiza una búsqueda secuencial del mejor agrupamiento, pamTOK logra acercarse a los resultados obtenidos por pamk (distancias ERP y EDR), incluso en algunos caso obtiene un mejor desempeño como se muestra con las distancias EDR y DTW, y a un menor coste computacional. La Figura 5.5 muestra los agrupamientos obtenidos con pamTOK con LCSS-20.

Tabla 5.8 Resultados de agrupamiento de la base Cross.

CROSS /19K	**pamTOK** IDIε >= 2.3			**pamk** 2:20		
	K	**CCR**	**Sil-Clust**	**K**	**CCR**	**Sil-Clust**
DTW	18	0.717	0.593	16	0.695	0.603
EDR5	5	0.168	0.211	2	0.105	0.655
EDR10	11	0.311	0.480	2	0.105	0.572
EDR20	22	0.742	0.728	2	0.105	0.486
ERP1	15	0.611	0.551	**19**	**0.663**	**0.556**
LCSS5	3	0.156	0.314	2	0.104	0.259
LCSS10	24	0.861	0.583	2	0.105	0.347
LCSS20	**21**	**0.908**	**0.728**	16	0.837	0.482



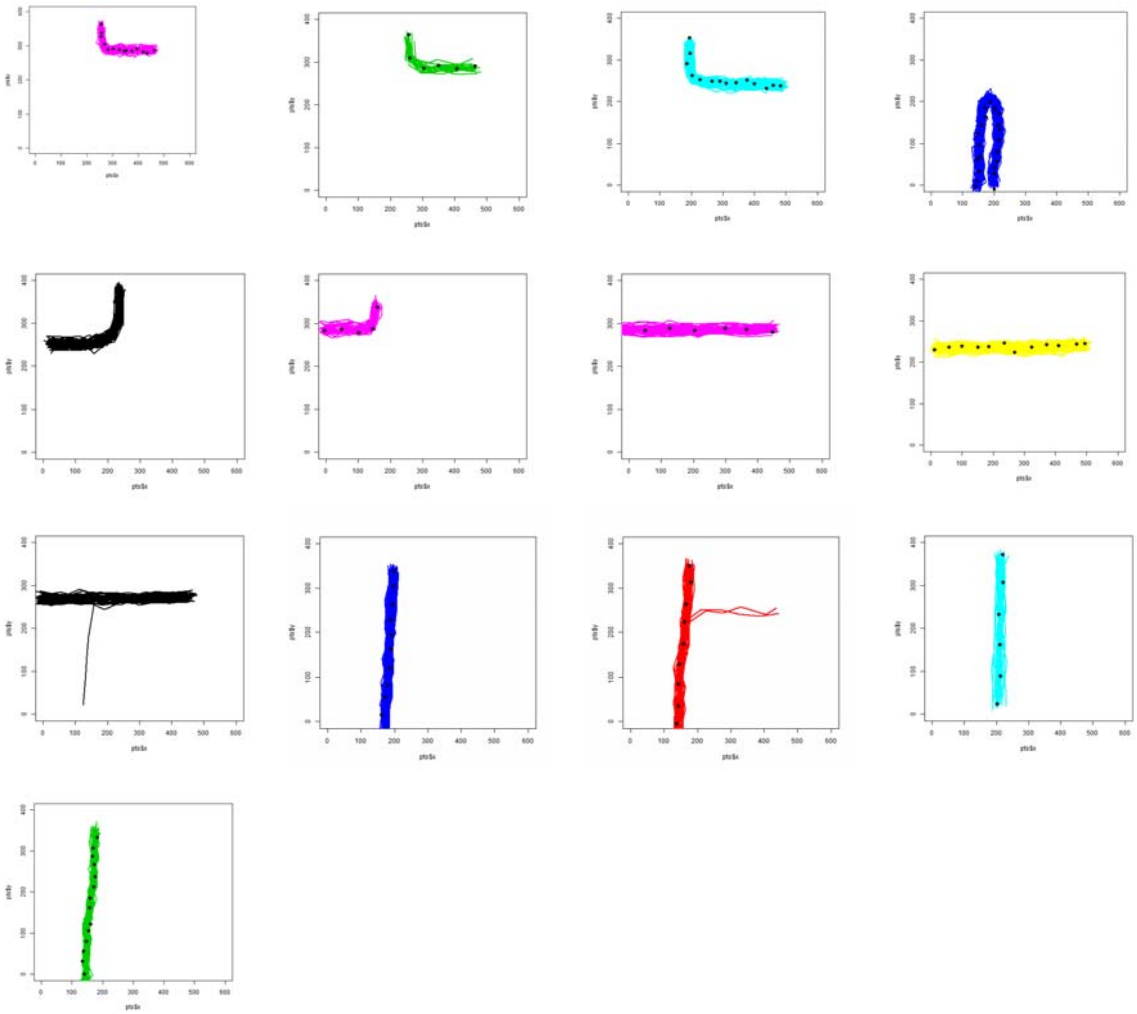


Figura 5.5 Ejemplo de agrupamiento de la base Cross con pamTOK y distancia LCSS-20.

5. *Base de trayectorias Labomni*

La base de datos Labmoni se integra de trayectorias reales de personas en un laboratorio circular. Por ello, se presentan recorridos traslapados, cambios bruscos, o trayectorias con distancias pequeñas. El resultado del algoritmo pamTOK, se acerca al etiquetado verdadero, logrando un desempeño del 0.86% de CCR y 14 grupos con LCSS-5, ver figura 5.6. Además, se puede observar que la mayoría de las distancias tienen propuestas de agrupamiento interesantes, que se acercan al número de categorías correctas. Lo contrario ocurre con el algoritmo pamk, que muestran el problema de considerar solamente el valor del coeficiente de la silueta para el agrupamiento ya que al tener un valor alto, se llega a un óptimo local, como en este caso, ver tabla 5.9.

Tabla 5.9 Resultados de agrupamiento de la base Labomni.

LABOMNI / 15K	**pamTOK** IDIε >= 1.7			**pamk** 2:20		
	K	**CCR**	**Sil-Clust**	**K**	**CCR**	**Sil-Clust**
DTW	12	0.736	0.532	6	0.631	0.734
EDR5	18	0.593	0.384	2	0.267	0.761
EDR10	22	0.612	0.408	2	0.267	0.748
EDR20	19	0.540	0.410	2	0.267	0.748
ERP1	14	0.674	0.353	2	0.267	0.753
LCSS5	**14**	**0.861**	0.503	2	0.272	0.706
LCSS10	7	0.645	0.407	2	0.272	0.691
LCSS20	20	0.712	0.509	2	0.272	0.681

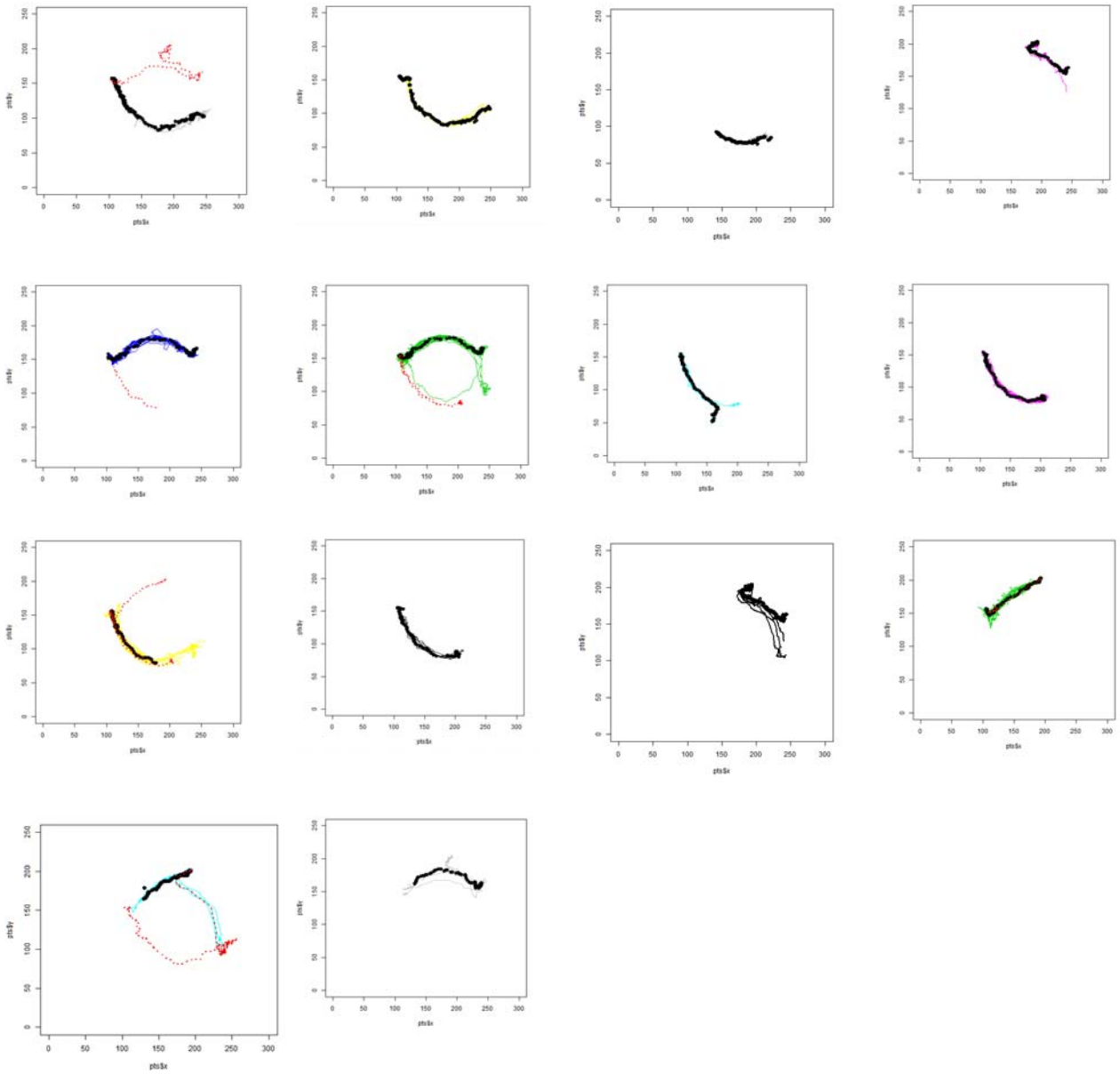


Figura 5.6 Ejemplo de agrupamiento de la base Labomni con pamTOK y distancia LCSS-5.

5.4.4 Identificación de comportamientos anormales

Para dar cumplimiento al objetivo de esta prueba, se seleccionaron las bases de datos públicas Bard y Edimburgo, integradas de trayectorias de personas y cuyo comportamiento anormal se integra de comportamientos erráticos y poco frecuentes. Para estas bases de datos no se cuenta con el etiquetado verdadero ni con el número de grupos en los cuales se deben separar.

Base de trayectorias Bard

La base de datos Bard se integra de 610 trayectorias de las cuales se tiene conocimiento que existen 50 trayectorias anormales, algunas con sólo pequeños desplazamientos al área del jardín o movimientos en zigzag. El cálculo de los criterios de validación se realiza con la información posicional de estas trayectorias en el archivo. El agrupamiento se realizó con un umbral IDIε de 1.7 y un mínimo de elementos de 3. Para pamk se utilizó un rango de 2 a 40.

La identificación de las trayectorias anormales se realizó tomando en cuenta el valor de pertenencia del objeto y frecuencia de las trayectorias. Para el factor de pertenencia, se considera el supuesto de que todos aquellos recorridos que se encuentran alejados del resto de los elementos de su grupo, son potenciales trayectorias con una ruta inusual y tienen un valor de pertenencia bajo. Para la identificación de las trayectorias con una frecuencia baja; se consideró el categorizar así aquellos grupos, con un número de trayectorias menor o igual al valor del umbral, especificado en el algoritmo de pamTOK, como el mínimo número de elementos para poder dividir. Entonces, la regla para identificar las trayectorias inusuales queda de la siguiente manera.

Son trayectorias inusuales Si:

(Pertenencia del objeto a su clase (silhouette de cada objeto) < 0.15)

O (Número de elementos en el grupo <= 3)

La tabla 5.10 muestra los resultados obtenidos para la base Bard. En las filas se encuentran listadas las distancias utilizadas con ambos algoritmos. La primera columna indica el número de grupos estimados utilizando la distancia especificada. Posteriormente las siguientes cuatro columnas hacen referencia a los valores de la figura

5.6. Las tres columnas siguientes muestran el resultado de los criterios de evaluación utilizados.

Es interesante ver como en varias distancias, ambos algoritmos, estiman un número de grupos similares o cercanos (LCSS-5, LCSS-10, LCSS-20 y DTW). Punto a favor del algoritmo pamTOK que no necesita realizar una comprobación con *K* valores (búsqueda del mejor resultado en un rango de valores para *K*) para llegar a ese resultado. Es importante hacer notar que nuevamente, la distancia LCSS es la que mejor desempeño presenta en este rubro.

Como ya se mencionó, el criterio para comparar el resultado de los dos algoritmos de agrupamiento es la razón definida en la ecuación 5.6. Esta razón consiste en comparar la detección de trayectorias anormales realizada por los algoritmos pamTOK y pamk. Para su cálculo se considera las trayectorias que ambos identifican como anormales (A) entre la unión de todos los recorridos etiquetados como anormales por pamTOK más aquellos trayectos detectados por pamk como inusuales; es decir A+B+C, de acuerdo a la figura 5.2. Para la base Bard, se pueden observar estos valores en la tabla 5.10. Revisando los resultados obtenidos en RazónOut, se observa que mientras más semejante sea el número de grupos obtenidos, la posibilidad de tener una mayor coincidencia en la detección de comportamientos inusuales crece. Logrando para la distancia LCSS resultados de 1.0% y de 0.71%; y la distancia EDR-20 un 0.44%. Sin embargo, estos buenos resultados no coinciden con la especificidad (SPC) que indica el porcentaje real de las trayectorias identificadas como anormales, lo son realmente.

Revisando los valores obtenidos en la columna del SPC de la tabla 5.10, los resultados muestran un desempeño muy inferior en la detección real de las trayectorias anormales, por ambos algoritmos. La combinación de pamTOK con la distancia EDR-10 es la que mejor desempeño tiene en este aspecto alcanzando un 0.76% y con EDR-5 de 0.52%. En términos generales, se puede ver que el algoritmo pamTOK tiene una mejor detección de las trayectorias anormales ligeramente superior o similar a la lograda por el algoritmo pamk.

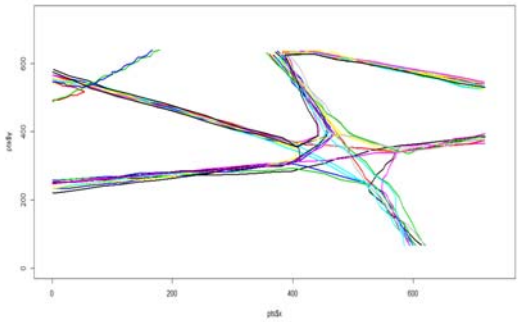
Por último, la última columna de la tabla 5.10 presenta la exactitud general de la detección (DACC), es decir, considerando en la detección las trayectorias etiquetadas

como normales y las anormales. Los resultados muestran un buen desempeño en general debido a que ambos algoritmos no tienen problemas con los trayectos normales, un ejemplo de ello, lo son nuevamente las distancias LCSS y DTW que alcanzan un 0.90% de DACC.

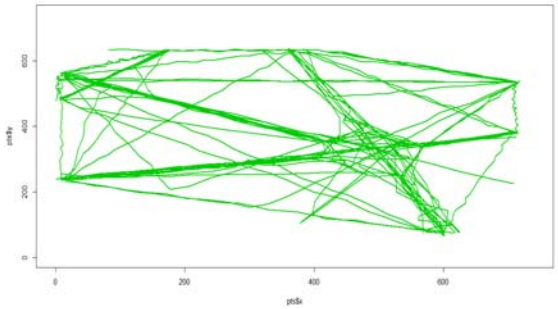
Tabla 5.10 Resultados del agrupamiento de la base Bard con los algoritmos pamTOK, pamk. Se muestra el número de grupos estimados, los valores para calcular la RazónOut y los resultados en los criterios de la especificidad y exactitud.

Algoritmo	Clases	A	B	C	D	RazónOut	SPC	DACC
pamTOK-DTW	*27*	4	12	15	579	0.129	0.06	**0.901**
pamk 2:40	*33*						0.04	0.893
pamTOK-EDR5	41	59	158	76	317	0.024	0.52	0.647
pamk 2:40	25						0.20	0.729
pamTOK-EDR10	47	16	186	26	382	0.070	0.76	0.711
pamk 2:40	24						0.12	0.868
pamTOK-EDR20	*41*	31	25	13	541	0.449	0.20	0.869
pamk 2:40	*34*						0.20	0.878
pamTOK-ERP	52	0	57	13	540	0.0	0.26	0.867
pamk 2:40	21						0.20	0.896
pamTOK-LCSS5	*32*	75	0	0	535	1.0	0.02	0.798
pamk 2:40	*32*						0.02	0.798
pamTOK-LCSS10	*34*	37	9	6	558	0.711	0.02	0.845
pamk 2:40	*35*						0.02	0.850
pamTOK-LCSS20	*35*	29	0	0	581	1.0	0.20	**0.903**
pamk 2:40	35						0.20	**0.903**

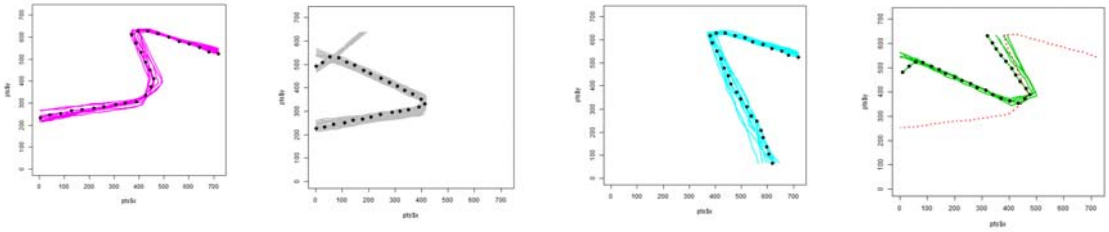
Un análisis más detallado sobre los resultados, permite observar que la base de datos contiene grupos con trayectorias normales pero con baja frecuencia y viceversa, hay varias trayectorias pertenecientes al grupo de las anormales, que tienen recorridos similares (frecuencia mayor al umbral especificado como anormal) y por lo tanto pasan desapercibidas, ver figura 5.7; de manera similar ocurre con el umbral de pertenencia al grupo. Otros recorridos presentan solamente pequeñas variaciones entre las trayectorias normales y las anormales. Esto indica que, para la detección de este tipo comportamientos anormales es mejor utilizar un modelo que realice el análisis de los recorridos en partes para identificar la zona con la desviación y detectar si representa esta actividad, un problema o no.



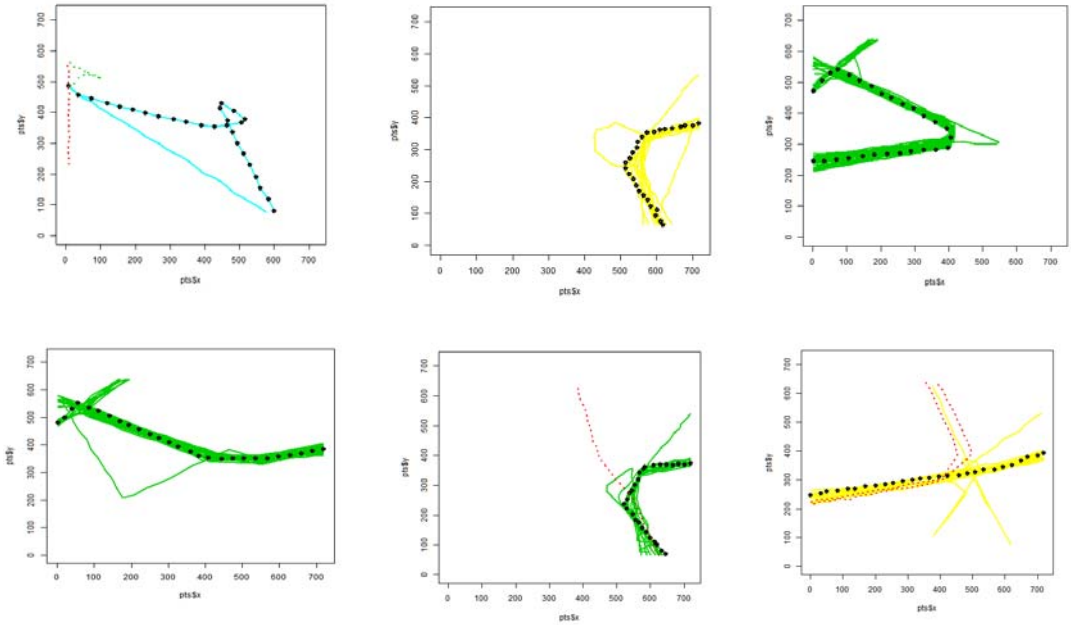
a) Medoides obtenidos



b) Trayectorias anormales



d) Grupos bien definidos



e) Ejemplos de trayectorias anormales no detectadas.

¡Error! La autoreferencia al marcador no es válida. Base de datos Bard.

a) Presenta los 35 representantes de cada grupo (medoides) obtenidos con el agrupamiento de pamTOK y distancia LCSS-20. b) Trayectorias consideradas como anormales. Algunas de estas trayectorias tienen recorridos similares a las rutas normales. c) ejemplos de grupos bien definidos con su medoide. d) Ejemplos diversos de trayectorias anormales no detectadas por tener un valor de frecuencia y pertenencia a su grupo, mayores a los umbrales especificados.

1. *Base de trayectorias Edimburgo*

Esta base de datos no proporciona información de ningún tipo, por lo que se desconoce cuál es el número de grupos correcto, ni el etiquetado de las trayectorias normales o anormales. El agrupamiento se realizó con los valores de umbral:

IDIε de 4.5 y un mínimo de elementos >= 3.
pamk: se utilizó un rango de 2 a 30.

La identificación de las trayectorias anormales se realizó, nuevamente, tomando en cuenta el valor de pertenencia del objeto a su grupo y frecuencia de las trayectorias. Los valores de la condición tienen los mismos rangos utilizados con la base Bard, es decir:

(Pertenencia del objeto a su clase (silhouette de cada objeto) < 0.15)
O (Número de elementos en el grupo <= 4)

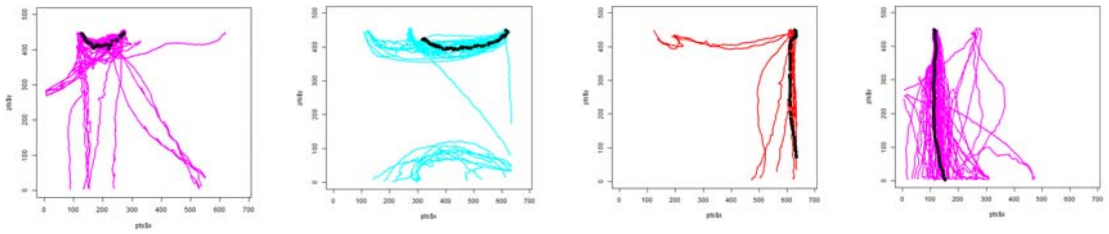
Para esta base de datos, no se cuenta con ningún tipo de información, por ello, el único criterio que se puede obtener es la RazónOut que permite conocer en qué porcentaje, los dos algoritmos de agrupamiento concuerdan respecto a la detección de trayectorias inusuales. Este resultado que se puede observar en la Tabla 5.11. La tabla también muestra el número de grupos estimados utilizando la distancia especificada y los valores para el cálculo de la razónOut.

Revisando los resultados obtenidos en la estimación del número de grupos, la combinación de pamTOK con DTW logra un agrupamiento cercano al propuesto por pamk, y en consecuencia tienen un valor de 0.587% en la RazónOut. Obviamente, entre más diferencia exista en el número estimado de grupos, por ambos algoritmos, el valor de la razón decrece, como sucede con las distancias LCSS (5,10 y 20) y EDR-20 cuyos valores de RazónOut son bajos. Esta diferencia se debe principalmente a que el algoritmo pamk estima el *k* considerando el valor de la silueta grupal o coeficiente de *silhouette*, y lamentablemente con ello queda estancado en óptimos locales. Por lo que se considera, que el algoritmo pamTOK tiene un mejor desempeño.

Tabla 5.11 Resultados del agrupamiento de la base Edimburgo con los algoritmos pamTOK, pamk. La tabla muestra el número de grupos estimados, los valores para calcular la RazónOut y el resultado de la razón.

Algoritmo	Clases	A	B	C	D	RazónOut
pamTOK-DTW	20	87	31	30	1844	**0.587**
pamk 2:40	21					
pamTOK-EDR5	10	27	101	82	1782	0.128
pamk 2:40	13					
pamTOK-EDR10	5	26	64	0	1902	0.288
pamk 2:40	4					
pamTOK-EDR20	20	12	159	63	1758	0.0512
pamk 2:40	4					
pamTOK-ERP	16	27	136	55	1774	0.123
pamk 2:40	8					
pamTOK-LCSS5	40	30	130	48	1784	0.144
pamk 2:40	6					
pamTOK-LCSS10	6	5	109	99	1779	0.023
pamk 2:40	14					
pamTOK-LCSS20	5	17	58	121	1796	0.08
pamk 2:40	15					

Como ya se mencionó, los resultados obtenidos varían dependiendo de los valores utilizados en los umbrales de las distancias, en el índice de agrupamiento (IDIε), y los utilizados para la detección de comportamientos inusuales. La Figura 5. y Figura 5.7 muestran precisamente un ejemplo de esta variación. Como se ve en las figuras mencionadas, un valor menor para IDIε significa que se desea tener clústeres homogéneos y con mayor separación con respecto a los otros grupos. Y viceversa, un valor mayor significa que se busca una estructura no tan fuerte, lo que permite tener en los grupos una mayor dispersión. La Figura 5. muestra 8 de los 20 grupos resultantes con la combinación del algoritmo pamTOK y la distancia DTW con un IDIε igual a 4.5. Mientras que la Figura 5.7 ejemplifica los grupos resultantes del agrupamiento de pamTOK y la distancia DTW con un IDIε de 2.0.



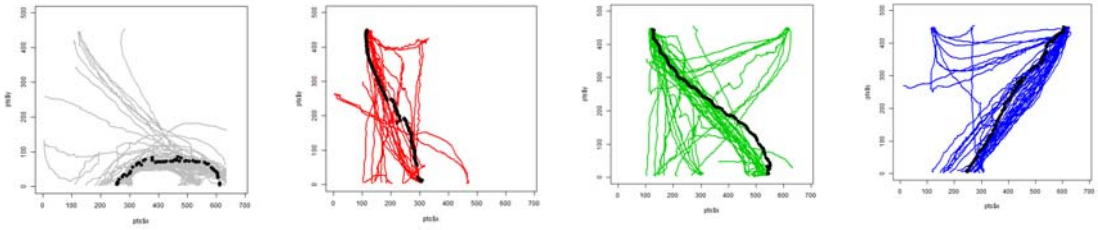


Figura 5.8 Ejemplos de grupos con un valor de umbral de agrupamiento alto (IDIε= 4.5). Base de datos Edimburgo con pamTOK y distancia DTW.

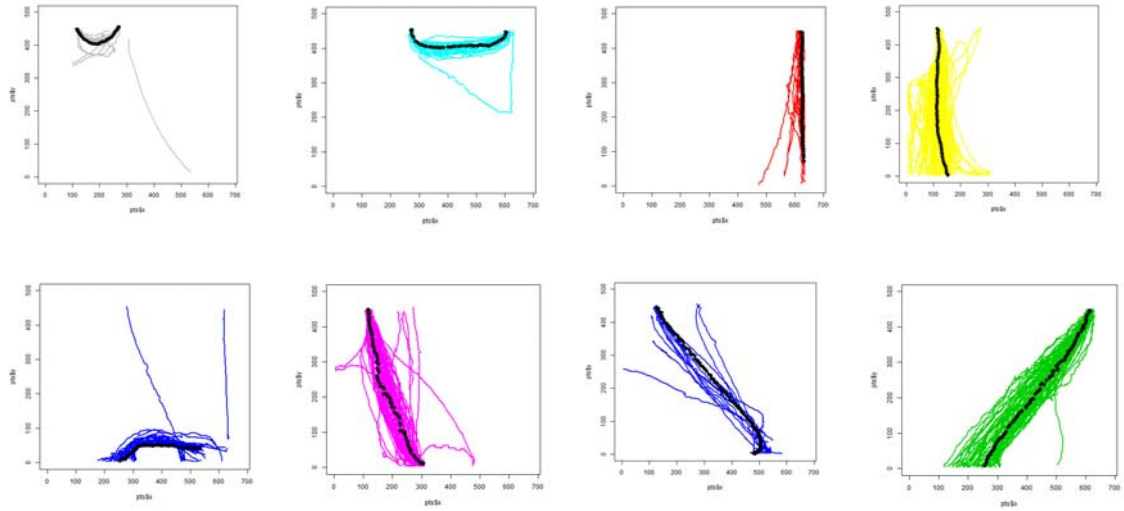


Figura 5.7 Ejemplos de grupos con un valor de umbral de agrupamiento moderado (IDIε = 2.0). Base de datos Edimburgo con pamTOK y distancia DTW.

5.4.5 Análisis de resultados

Para verificar el rendimiento de las distancias en el contexto del agrupamiento, se realizó una comparativa, asignando un valor de acuerdo a los resultados obtenidos, asignando un valor de 1 para el mejor desempeño y 8 para el menor (son ocho distancias). Para tener información objetiva, se llevó a cabo solamente con las seis bases de datos que tienen información de etiquetado verdadero, para tener una referencia real La Figura 5.8 resume de manera clara el desempeño alcanzado por cada medida de similitud. El gráfico presenta el rango promedio y los intervalos de confianza correspondiente a cada medida de distancia.

El análisis de la Figura 5.8 permite identificar que la medida de similaridad que mejor desempeño tuvo es la LCSS. Ella logró tener mejores agrupamientos en las distintas bases de datos con dos de los tres umbrales especificados (10 y 20). La distancia EDR logra buenos agrupamientos, ligeramente menores a los obtenidos por LCSS. No obstante, se nota que el umbral con valor cinco no fue correcto para ambas distancias, ya que presenta un rendimiento bajo y una dispersión grande. La distancia DTW tiene la ventaja de que no necesita se le especifique parámetro alguno y mostró un desempeño aceptable con la mayoría de las bases de datos. En esta evaluación, la distancia ERP no obtuvo buenos resultados quedando debajo de logrado por LCSS y DTW.

Estas métricas también tienen desventajas, en el caso de LCSS y EDP requieren un valor de umbral (denominado épsilon) a partir del cual, el algoritmo considera similares o diferentes dos trayectorias, el cual es un problema difícil si no se tiene un conocimiento a priori de los datos. El inconveniente de la distancia DTW es que presenta un alto costo computacional respecto a las otras distancias evaluadas, problema que comparte con ERP.

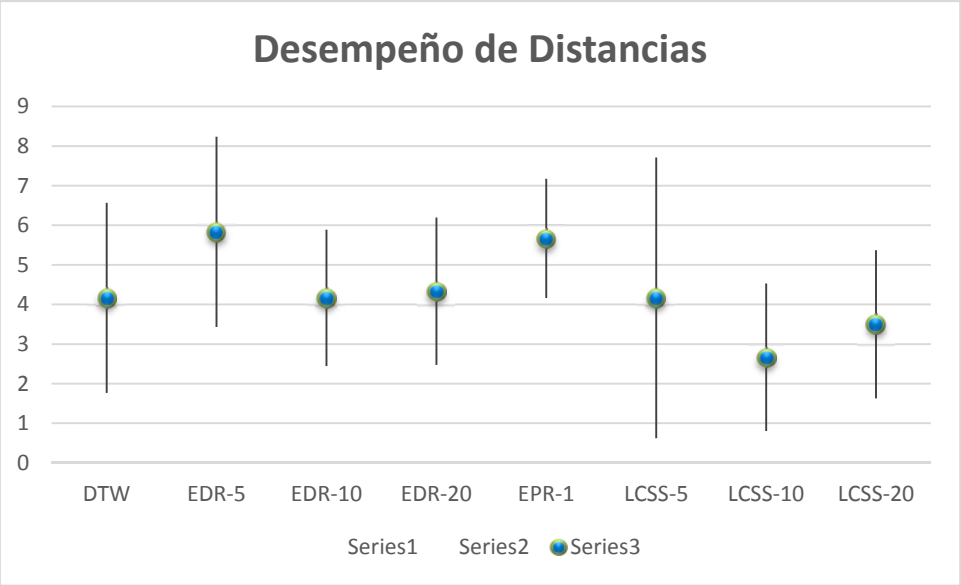


Figura 5.8 Comparativa del desempeño obtenido por las distancias en el contexto de agrupamiento de trayectorias, con las bases de datos del grupo CVRR que cuentan con etiquetado verdadero.

Analizando los resultados en relación a la estimación automática del número de grupos presentes en un conjunto de datos, el algoritmo pamTOK con la familia de bases de datos I5, I5sim, I5sim2 y I5sim3 (ver

Tabla 5.4 a Tabla 5.7) estima de manera adecuada el número de grupos presente en el etiquetado verdadero, obteniendo valores de CCR del 0.96% y 1.0%, incluso, en combinación con la distancia LCSS, es capaz de identificar las diferencias de velocidad (flujo lento y rápido) presente en los datos. Con estas bases de datos, el algoritmo pamk, también presenta un desempeño excelente, logrando realizar agrupamientos con un 1.0% de CCR. Sin embargo, este resultado lo consigue ejecutando *n* veces el algoritmo pam, y quedándose con el número de grupos que tenga el valor más alto de la silueta grupal.

Con respecto a conjuntos de datos más complejos y con información real, como los que integran las bases de datos Cross y Labomni, el algoritmo pamTOK sigue siendo eficaz, proponiendo un número de categorías similares a las óptimas (dadas por el etiquetado verdadero), alcanzando resultados de 0.90 y 0.86% de CCR, ver Tabla 5.8 y Tabla 5.9. Por su parte, el algoritmo pamk, con estas bases de datos, presenta un desempeño bajo al proponer un menor número de grupos a los esperados y esto se debe principalmente a que encuentra un óptimo local con valores altos de la silueta grupal. Estos resultados ejemplifican la desventaja de considerar este el valor para decidir el número de categorías presentes en un conjunto de datos.

En relación con la base de trayectorias Bard, se conoce cuáles trayectorias tienen un comportamiento anormal, lo que permite evaluar la especificidad (SPC) de los algoritmos de agrupamiento en la detección de estas trayectorias inusuales y la exactitud de la detección (DACC) considerando está a partir de los dos grupos presentes en los datos: trayectorias normales o verdaderos positivos y trayectorias anormales o verdaderos negativos. Como se puede observar en la Tabla 5.10, la estimación del número de grupos llevada a cabo por pamTOK y pamk con la distancia LCSS, son iguales y con un costo computacional mucho menor; lo que permite lograr resultados similares en la identificación de las trayectorias anormales, alcanzando un valor de RazónOut de 1.0%.

Los valores de DACC obtenidos también son buenos, ya que se tienen valores aceptables, alcanzando un 0.90% ambos algoritmos, demostrando que la detección con respecto a la clase de trayectorias normales es buena. Sin embargo, revisando en detalle los resultados, se pueden ver los bajos porcentajes mostrados por la especificidad que no se realiza de manera eficaz la detección de las trayectorias categorizadas como anormales. Aun así, los valores obtenidos por el algoritmo pamTOK son ligeramente superiores a los logrados por pamk.

El bajo desempeño mostrado por ambos algoritmos en la identificación de trayectorias anormales, se debe principalmente a las características que presentan estos recorridos, como que sean trayectorias semejantes a las consideradas como normales pero con pequeños sesgos (lo que da alta pertenencia a su grupo), o realizando recorridos normales en áreas no permitidas como lo son las áreas verdes, trayectorias anormales realizadas varias veces, etc. Es decir, trayectorias con pequeñas anormalidades o que necesitan información del contexto. Además, la presencia de trayectorias normales con baja frecuencia y provoca que los umbrales especificados para la detección de los comportamientos anormales, logren desempeño por debajo de lo esperado.

Finalmente, la base de datos de Edimburgo permite observar como nuevamente como el algoritmo pamk estima un menor número de grupos que los propuestos por pamTOK, y esto provoca que el porcentaje de coincidencias en la detección de las trayectorias anormales baje. También, se puede ver la influencia de los valores especificados en los umbrales tanto en las distancias utilizadas en la comparación de las trayectorias, en el índice de agrupamiento IDIε, así como en la condición para la detección de trayectorias inusuales o diferentes.

5.5 Conclusiones

En este capítulo se presentó el algoritmo de agrupamiento propuesto pamTOK, el cual no requiere se le especifique un valor especifico de categorías, ni un rango de búsqueda para determinar el mejor número de grupos. En la evaluación se consideró otro algoritmo basado en pam, que determina el número de grupos óptimo en base al valor de la silueta del agrupamiento. Ambos algoritmos (pamTOK y pamk) fueron evaluados con seis bases de datos que cuentan con el etiquetado verdadero; información que

permitió evaluar de manera objetiva el rendimiento de nuestra propuesta. Los resultados obtenidos en esta prueba, mostraron un buen desempeño del algoritmo, al obtener el número de clústeres especificado en el etiquetado verdadero o un valor cercano. Los resultados obtenidos con el algoritmo pamk, mostraron que el valor de la silueta del agrupamiento no es un buen índice de agrupamiento, ya que se pueden tener resultados correspondientes a un óptimo local. Otra desventaja de pamk, es el estar limitada su búsqueda al rango especificado.

También se evaluaron ambos algoritmo en relación a la detección de comportamientos anormales. Para ello, tomó en cuenta el valor de pertenencia del objeto a su grupo y frecuencia mínima de las trayectorias. Los resultados obtenidos muestran un rendimiento bajo debido a los umbrales especificados, sin embargo, el algoritmo pamTOK logra detectar un porcentaje mayor de comportamientos inusuales.

De acuerdo a los resultados alcanzados, se puede concluir que el rendimiento del algoritmo propuesto pamTOK es excelente respecto a decidir el número de grupos en que es conveniente separar los diversos conjuntos de datos analizados. Esta característica lo hace mejor que los algoritmos de agrupamiento clásicos que necesitan se especifique el número de grupos deseados como kmeans y pam (Kaufman & Rousseeuw 1987) o el rango de valores en el cual se debe buscar el mejor valor de *k* que minimice las diferencias en los grupos como en mclust (C. Fraley, A. E. Raftery 2012) y pamk (Christian Hennig 2014). Como resultado, se tiene un menor coste computacional.

El algoritmo pamTOK necesita se le especifique al parámetro de decisión un valor de entrada y en función de su valor, será el resultado obtenido. Sin embargo, se considera que este valor es más sencillo e intuitivo de proponer. El valor del umbral se especifica de acuerdo a la homogeneidad deseada y al número de grupos esperados.

Al final de la ejecución, el algoritmo pamtoK proporciona información que hereda del algoritmo pam, como lo es el vector de pertenencia del conjunto de datos, el valor de pertenencia de cada elemento a su grupo, el valor de la *silhouette* del agrupamiento, el número estimado de grupos y el objeto representante de cada grupo (medoide).

Capítulo 6

Conclusiones

El presente capítulo resume las actividades desarrolladas en el presente trabajo, lista las principales aportaciones obtenidas y propone futuros trabajos de investigación.

6.1 Conclusiones

En la última década, las investigaciones en el área de videovigilancia han aumentado debido al incremento en aplicaciones de monitoreo inteligente para describir, entender e identificar actividades humanas normales y anormales. El objetivo de todos estos estudios es el de perfeccionar las capacidades de las técnicas actuales, utilizadas en las distintas etapas de los sistemas de videovigilancia visual.

Con este propósito de mejora, se diseñó e implementó un sistema de video vigilancia que tuviera la habilidad de poder describir los objetos presentes en imágenes de escala pequeña y de baja calidad, en ambientes de exteriores; con la finalidad de localizar e identificar la presencia de seres humanos y detectar trayectorias con comportamientos diferentes, en aplicaciones reales.

La tesis se enfocó en el desarrollo de dos técnicas. La primera de ellas tuvo el objetivo de proponer un detector local alternativo a los actuales, que fuera capaz de extraer características locales y representar a los objetos presentes en regiones pequeñas y en entornos de baja calidad. La técnica propuesta se denominó GSIFT ya que se basa en la combinación de los filtros Gabor y el descriptor local SIFT. La segunda técnica propuesta es un algoritmo de agrupamiento denominado pamTOK (se basa en el algoritmo pam), el cual estima de manera automática el número de categorías en que es conveniente separar el conjunto de datos analizado; encontrando los modelos correspondientes a un comportamiento normal, para detectar comportamientos anormales.

El rendimiento del detector local propuesto GSIFT fue comparado con seis de los principales métodos de descripción local, utilizados en una gran variedad de trabajos en el área de reconocimiento de patrones y visión artificial. Los descriptores locales fueron evaluados en ambientes urbanos no controlados. Se evaluó su extracción de características y su poder de descripción, en un sistema de clasificación. Se compararon las fortalezas y debilidades de manera sistemática, en una tarea compleja como lo es la detección de peatones en imágenes de exteriores reales, de escala pequeña y de baja calidad.

La evaluación analizó cuatro técnicas de extracción de características que trabajan con el esquema de detección y descripción local como lo son SIFT, SURF, ORB y BRISK. Además, se realizó la combinación del detector FAST con el descriptor extendido SURF y con el descriptor FREAK, lo que resulta en una comparación de seis propuestas en la detección de peatones. Todos los resultados son validados y replicados en cuatro conjuntos de entrenamiento y prueba diferentes, lo que sugiere que los resultados obtenidos son generalizables.

Este trabajo, ofrece también una comparativa exhaustiva y objetiva de los algoritmos considerados dentro de una misma plataforma de desarrollo como lo es OpenCV V2.4.8, lo que permitió evaluarlas de manera objetiva, identificar las ventajas e inconvenientes de cada uno de ellos y eliminar el sobreajuste de los algoritmos.

Las características globales alcanzan mejores resultados obtenidos con los descriptores locales, para todos los métodos considerados. Obviamente, la fusión de los descriptores locales y globales ofrece mejores resultados que los obtenidos por métodos locales y globales de forma individual. En general, la fusión de descriptores mejora los resultados individuales.

En correspondencia con lo que se reporta en el estado del arte, todas las técnicas evaluadas tienen un notable rendimiento con imágenes de buena calidad. Sin embargo, sus resultados decaen si son expuestas a imágenes con condiciones de baja calidad y tamaño pequeño. Para este tipo de imágenes, el método FAST por tener un vector redundante, tiene mejores tasas de detección de puntos de interés en imágenes de

mediana y baja escala y con ello muestra ser la mejor alternativa en combinación con el descriptor FREAK. Con respecto a SURF, es un método eficiente y capaz de ejecutarse en un tiempo corto; no obstante, su desempeño de detección de puntos es menor al esperado ante imágenes que no presenten buena calidad (buena iluminación, contraste y zonas de textura) y un tamaño superior a los 50 pixeles. Por su parte, SIFT logra resultados satisfactorios en la mayoría de las situaciones evaluadas, su punto débil con respecto a otros métodos, es su tiempo de cómputo. Finalmente, las técnicas binarias ORB y BRISK no son una opción para aplicaciones que requieran imágenes pequeñas y de baja calidad.

Finalmente, y de acuerdo a los resultados obtenidos se puede concluir que la modificación del descriptor SIFT con filtros Gabor, en el detector local GSIFT, amplia y mejora la extracción de características en imágenes de escala pequeña, obteniéndose un detector local robusto y estable aun en regiones con un número de píxeles menor a 30.

Como se mencionó, la segunda parte de la tesis consistió en el análisis de trayectorias ya que estas son una valiosa fuente de información para reconocer automáticamente, comportamientos específicos, sospechosos o inusuales, llevado a cabo por los objetos de interés. El enfoque considerado en este trabajo fue el realizar un análisis de las trayectorias a través de su comportamiento en tiempo-espacio. Este enfoque se integra de tres etapas que son: comparar las trayectorias a través de medidas de similaridad apropiadas que permitan, posteriormente, aplicar un algoritmo de agrupamiento para modelar el conjunto de datos y conocer las rutas presentes en los recorridos. Finalmente, a partir de las actividades frecuentes, encontrar aquellas trayectorias diferentes al resto de los datos.

Para la comparar la similitud entre trayectorias se determinó emplear medidas de similaridad que fuesen independientes a la longitud de las mismas. Se seleccionó e implementaron las medidas Dynamic Time Warping (DTW), Longest Common Subsequence (LCSS), Edit Distance on Real Sequence (EDR) y Distance with Real Penalty (ERP). Estas cuatro distancias se utilizaron en combinación con los algoritmos de agrupamiento pamTOK y pamk y se aplicaron a ocho bases de trayectorias para evaluar su desempeño en la estimación automática del número de grupos en los cuáles

es conveniente separar dichos conjuntos de datos. De acuerdo a los resultados obtenidos, la distancia LCSS tiene una mejor eficiencia en este rubro, seguida por la distancia EDR y DTW, quedando en cuarto lugar la distancia ERP. Sin embargo, se considera una desventaja la necesidad de especificar el valor del umbral a partir del cual, se decidirá sobre la similaridad o disimilaridad de dos trayectorias, porque normalmente, no se cuenta con información estadística del conjunto de datos y los resultados dependen, obviamente, de la correcta selección de este valor.

El problema de estimar el número óptimo de grupos de manera automática, es un tema complejo que sigue siendo actual y de mucha investigación. En esta tesis, para la estimación automática del número de grupos presentes en un conjunto de datos, se llevó a cabo una serie de experimentos con seis bases de datos (de trayectorias reales y simuladas) cuyas trayectorias cuentan con información sobre su etiquetado verdadero. Esta información de pertenencia permitió evaluar de manera cuantitativa y objetiva (con el criterio Correct Clustering Rate, CCR) el desempeño del algoritmo de agrupamiento propuesto pamTOK y el algoritmo pamk.

Los resultados obtenidos demostraron que el algoritmo propuesto estima de manera adecuada el número de grupos presentes en el etiquetado verdadero, obteniendo valores de CCR del 0.96% y 1.0% en combinación con la distancia LCSS. Por su parte, el algoritmo pamk también presenta un desempeño excelente, y en dos bases de datos ligeramente mejor que pamTOK, logrando un 1.0% de CCR, también en combinación con la distancia LCSS. Sin embargo, es importante recordar que para estimar el número de grupos, este algoritmo se ejecuta el número de veces especificado y regresa como resultado el mejor K (número de clústeres) que consiga el valor más alto de la *silhouette* del agrupamiento. Esta condición tiene la desventaja de que el valor más alto puede pertenecer a un óptimo local, como sucede con la base Labomni.

En relación a la identificación de comportamientos anormales, se utilizaron dos bases de datos de trayectorias generadas por personas al trasladarse de un lugar a otro en la escena. Se consideró como trayectorias normales al conjunto de patrones de movimiento de mayor frecuencia de ocurrencia en la escena, y como recorridos anormales aquellas actividades que ocurren de manera inusual. Se utilizaron tres criterios de validación para evaluar el desempeño de los algoritmos de agrupamiento.

En relación a la base de datos Bard, está cuenta con información sobre cuáles trayectorias tienen un comportamiento anormal, siendo posible validar la especificidad (SPC) de los algoritmos de agrupamiento en la detección de estas trayectorias inusuales y la exactitud de la detección (DACC). Además, para comparar la detección de las trayectorias inusuales, se propuso otro criterio denominado RazónOut. Los resultados obtenidos para los algoritmos pamTOK y pamk son similares en combinación con la distancia LCSS, y eso significa que la estimación del número de grupos es fue la misma o muy cercana, alcanzando valores de DACC del 0.90% y de RazónOut de 1.0%. Y para el algoritmo pamTOK con un menor coste computacional.

Sin embargo, el criterio de especificidad muestra que la detección real del comportamiento anormal es inferior al esperado, siendo el más alto de 0.76 % por parte de pamTOK en combinación con la distancia EDR-10. El bajo desempeño mostrado por ambos algoritmos en la identificación de trayectorias anormales, se debe principalmente a que son trayectorias con pequeñas anormalidades o que necesitan información del contexto de la escena como conocer las áreas verdes que no deben pisar. Información que no se consideró en el desarrollo de este trabajo.

Las pruebas realizadas con la base de datos de Edimburgo, mostró nuevamente como la decisión de estimación del número de grupos basada en el mayor valor de la *silhouette* del agrupamiento no es eficaz, teniendo un mejor desempeño (validado de manera visual) el algoritmo pamTOK. Y una diferencia grande en el número de grupos propuestos entre los algoritmos, provoca tener una baja coincidencia en la detección de las trayectorias anormales y con ello un porcentaje de RazónOut bajo.

Finalmente, en base a los resultados obtenidos por las dos técnicas propuestas se concluye que se cumplió con el objetivo de la tesis.

6.2 Aportaciones

Las principales contribuciones de esta tesis son:

1. Se diseñó e implementó una nueva técnica de detección local de puntos de interés denominado GSIFT, adecuado para imágenes de escala pequeña y baja

resolución que no pueden ser descritas por la mayoría de las técnicas de descripción local actuales, ya que requieren de imágenes con una escala mayor.

2. El rendimiento del método GSIFT se evaluó en tres bases de datos públicas específicas para la detección de peatones en regiones pequeñas. Y se demostró que es posible lograr una mejor detección de la información relevante (mejor y más local punto de interés) mediante el uso de filtros de Gabor, mejorando el rendimiento alcanzado por SIFT.

3. El detector GSIFT es invariante a cambios de escala, rotación y traslación y robusto ante las pequeñas escalas de las imágenes.

4. Para incorporar información global a la representación local, de manera novedosa se realizó la descripción holística de los puntos de interés detectados por cada técnica, a través de los momentos de Hu, logrando niveles de clasificación más altos.

5. Se propuso el algoritmo de agrupamiento pamTOK que estima en cuántos grupos es conveniente separar el conjunto de datos, mediante la especificación de un índice de agrupamiento (IDIε) que evalúa la relación entre homogeneidad interna de los grupos y su distancia con respecto a los otros grupos; sin estar limitado a un número de grupos específicos. Las ventajas del algoritmo propuesto son:

   - El valor del índice de agrupamiento (IDIε) es intuitivo de proponer. Un valor mayor expresa que el número de grupos es menor, pero los grupos tendrán un diámetro mayor y con ello una dispersión más alta. Y viceversa, un valor menor de IDIε, producirá una mayor cantidad de grupos con una estructura interna fuerte. De acuerdo a la experimentación, el valor por default que se eligió es de dos; lo que significa que un clúster se debe dividir, si su diámetro tiene una distancia igual o mayor al doble de la distancia que separa a este conjunto con respecto a su grupo más cercano.

   - El algoritmo pamTOK tiene un costo computacional bajo comparado a otros algoritmos de agrupamiento, ya que, como resultado de la ejecución del

algoritmo pamTOK, se cuenta con una estimación buena, del número de grupos en que es conveniente separar al conjunto de datos analizados. Eliminando la necesidad de evaluar un rango *n* de valores para estimar el *K* adecuado. Además, en cada llamada, el algoritmo pam analiza la mitad del conjunto de datos anterior, requiriendo un menor coste computacional su ejecución.

6. Se propuso un criterio denominado RazónOut para validar de manera objetiva y cuantitativa el desempeño de dos algoritmos de agrupamiento en la detección de "outliers", cuando no se cuenta con información alguna sobre el número de grupos ni tampoco de los comportamientos normales y anormales.

6.3 Trabajo futuro

- Se contribuye con un detector local de puntos característicos eficaz para representar objetos presentes en imágenes de tamaño pequeño; sin embargo, se requiere un mayor análisis para optimizarlo y que sea capaz de ejecutarse en menor tiempo. Por lo que se propone reducir el coste computacional del método GSIFT paralelizando el cálculo de los filtros Gabor para cada orientación considerada (cada nivel de la pirámide). Entonces, en un mismo tiempo se estaría realizando el cálculo de las ochos orientaciones consideradas y la aplicación de DoG en las imágenes resultantes. Reduciendo a un octavo el tiempo actual.

- Se propone combinar el detector propuesto GSIFT con el descriptor local FREAK que resultó ser la mejor técnica de descripción en la evaluación realizada.

- Se propone modificar la técnica de bolsa de palabras, mediante la aplicación del algoritmo pamTOK, sustituyendo al algoritmo kmeans (que normalmente se utiliza) para obtener el vocabulario con un menor costo computacional al eliminar la etapa de búsqueda del mejor con diferentes valores los *K* grupos.

- Se propone aplicar la técnica de bolsa de palabras mejorada, al conjunto de descriptores obtenido por GSIFT.

- Realizar una experimentación más detallada en relación a los parámetros de los filtros Gabor, en el detector GSIFT.

- Realizar un comparativo del detector GSIFT con las técnicas más utilizadas en la detección de personas, como lo son HOG, y Haar wavelets.

- Se propone evaluar el desempeño del algoritmo de agrupamiento pamTOK en bases de datos[5] estándares, que tengan información sobre su pertenencia, como Zoo, soybean, mushroom, etc.

- Se propone analizar las trayectorias considerando su forma, a través de una representación simbólica que permita un análisis en trazos y tome en cuenta la orientación de la trayectoria de manera puntual. Es decir, dado un valor de ángulo a partir del cual se desea representar los movimientos, se genera el conjunto de símbolos (lenguaje) a usar en la representación, considerando el espacio de 360 grados. Para complementar la información de las trayectorias, se deberá incluir la distancia del desplazamiento y el centroide del tramo de a trayectoria cada vez que se cambie de dirección. Esta información permitirá para conocer de manera más precisa la orientación, la distancia recorrida y la localización espacial dentro de la escena, de cada trayectoria. Esta representación permitirá detectar cambios en los recorridos (incluso mínimos) y se tendrá un mejor desempeño en la detección de comportamientos anormales.

- Para el análisis de las trayectorias, tomar en cuenta información del contexto de la escena como áreas de recorrido, áreas de descanso, horarios, etc.

- Mejorar la implementación del módulo de seguimiento de la figura humana para unir las etapas desarrolladas en un sistema de videovigilancia completo.

[5] UCI repository (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/)

Bibliografía

Abdel-Hakim, A.E. & Farag, A.A., 2006. CSIFT: A SIFT descriptor with color invariant characteristics. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. pp. 1978–1983.

Acevedo-rodr, J. et al., 2011. Clustering of Trajectories in Video Surveillance Using Growing Neural Gas. , pp.461–470.

Alahi, a., Ortiz, R. & Vandergheynst, P., 2012. Freak: Fast retina keypoint. *Computer Vision and …*, pp.510–517.

Aminian Modarres Amir Farid & Soryani, M., 2013. Body posture graph: a new graph-based posture descriptor for human behaviour recognition. *IET Computer Vision*, 7(6), pp.488–499.

Baiget, P. et al., 2012. Trajectory-Based Abnormality Categorization for Learning Route Patterns in Surveillance. , pp.87–95.

Barbu, T., 2014. Pedestrian detection and tracking using temporal differencing and HOG features. *Computers & Electrical Engineering*.

Bay, H., Tuytelaars, T. & Gool, L. Van, 2006. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pp.404–417.

Bedagkar-Gala, A. & Shah, S.K., 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4), pp.270–286.

Bereta, M. et al., 2013. Local descriptors in application to the aging problem in face recognition. *Pattern Recognition*, 46(10), pp.2634–2646.

Bradski, G. & Kaehler, A., 2008. *Learning OpenCV: Computer vision with the OpenCV library*, O'Reilly Media, Inc.

Brehar, R. & Nedevschi, S., 2011. A comparative study of pedestrian detection methods using classical Haar and HoG features versus bag of words model computed from Haar and HoG features. *Proceedings - 2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing, ICCP 2011*, pp.299–306.

C. Fraley, A. E. Raftery, T.B.M. and L.S., 2012. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical*,

Calderara, S. et al., 2011. Detecting anomalies in people's trajectories using spectral graph analysis. *Computer Vision and Image Understanding*, 115(8), pp.1099–1111.

Calonder, M. et al., 2010. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*.

Cancela, B. et al., 2013. On the use of a minimal path approach for target trajectory analysis. *Pattern Recognition*, 46(7), pp.2015–2027.

Catalán, J.J.A., 2013. *Fusión de escaner láser y cámara de infrarojos para la detección y seguimiento de trayectorias de peatones.* Proyecto de fin de Carrera, Universidad Carlos III De Madrid.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), p.15.

Chaquet, J.M., Carmona, E.J. & Fernández-Caballero, A., 2013. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6), pp.633–659.

Chen, J. et al., 2010. WLD: A robust local image descriptor. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9), pp.1705–1720.

Chen, L. & Ng, R., 2004. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. pp. 792–803.

Chen, L., Özsu, M.T. & Oria, V., 2005. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. pp. 491–502.

Christian Hennig, 2014. *Package " fpc ", Flexible procedures for clustering*, Available at: http://www.homepages.ucl.ac.uk/~ucakche/.

Conde, C. et al., 2013. HoGG: Gabor and HoG-based human detection for surveillance in non-controlled environments. *Neurocomputing*, 100, pp.19–30.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273–297.

Cucchiara, Rita and Grana, Costantino and Piccardi, Massimo and Prati, A., 2003. Detecting moving objects, ghosts, and shadows in video streams. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25, pp.1337–1342.

Dalal, N. & Triggs, B., 2005. Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*. pp. 886–893.

Desgraupes, B., 2013. *Clustering Indices*.

Dollár, P. et al., 2009. Pedestrian detection: A benchmark. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp.304–311.

Dollar, P. et al., 2012. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4), pp.743–761.

Dou, J. & Li, J., 2014. Robust human action recognition based on spatio-temporal descriptors and motion temporal templates. *Optik - International Journal for Light and Electron Optics*, 125(7), pp.1891–1896.

Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1), pp.95–104.

Enzweiler, M. & Gavrila, D.M., 2009. Monocular pedestrian detection: survey and experiments. *IEEE transactions on pattern analysis and machine intelligence*, 31(12), pp.2179–95.

Fan, L., Sung, K.-K. & Ng, T.-K., 2003. Pedestrian registration in static images with unconstrained background. *Pattern Recognition*, 36(4), pp.1019–1029.

Farhadi, M., Motamedi, S.A. & Sharifian, S., 2011. Efficient Human Detection Based on Parallel Implementation of Gradient and Texture Feature Extraction Methods. *2011 7th Iranian Conference on Machine Vision and Image Processing*, pp.1–5.

Flohr, F. & Gavrila, D., 2013. PedCut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues. *Procedings of the British Machine Vision Conference 2013*, pp.66.1–66.11.

Gabor, D., 1946. Theory of communication. Part 1: The analysis of information. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26), pp.429–441.

Gerónimo, D. et al., 2007. Adaptive image sampling and windows classification for on-board pedestrian detection. In *Proceedings of the International Conference on Computer Vision Systems, Bielefeld, Germany*.

Gerónimo, D., Sappa, A.D., et al., 2010. 2D--3D-based on-board pedestrian detection system. *Computer Vision and Image Understanding*, 114(5), pp.583–595.

Gerónimo, D., Lopez, A.M., et al., 2010. Survey of pedestrian detection for advanced driver assistance systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7), pp.1239–1258.

Gerónimo, D. & López, A.M., 2014. Datasets and Benchmarking. In *Vision-based Pedestrian Protection Systems for Intelligent Vehicles*. Springer, pp. 87–93.

Gogoi, P. et al., 2011. A survey of outlier detection methods in network anomaly identification. *The Computer Journal*, p.bxr026.

Gonzalez, R.C., 2009. *Digital image processing*, Pearson Education India.

Gowsikhaa, D., Abirami, S. & Baskaran, R., 2014. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4), pp.747–765.

Guo, L. et al., 2012. Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine. *Expert Systems with Applications*, 39(4), pp.4274–4286.

Guo, X. et al., 2010. Mift: A mirror reflection invariant feature descriptor. In *Computer Vision--ACCV 2009*. Springer, pp. 536–545.

Harris, C. & Stephens, M., 1988. A combined corner and edge detector. In *Alvey vision conference*. p. 50.

Heinly, J., Dunn, E. & Frahm, J.-M., 2012. Comparative evaluation of binary features. In *Computer Vision--ECCV 2012*. Springer, pp. 759–773.

Hennig, C., 2010. Methods for merging Gaussian mixture components. *Advances in data analysis and classification*, 4(1), pp.3–34.

Hj Wan Yussof, W.N.J. & Hitam, M.S., 2014. Invariant Gabor-based interest points detector under geometric transformation. *Digital Signal Processing: A Review Journal*, 25(1), pp.190–197.

Hsu, C. et al., 2003. A practical guide to support vector classification. , (1), pp.1–16.

Hsu, F.-S., Lin, C.-H. & Lin, W.-Y., 2011. Recognizing human actions using curvature estimation and NWFE-based histogram vectors. *2011 Visual Communications and Image Processing (VCIP)*, pp.1–4.

Hu, M.-K., 1962. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2), pp.179–187.

Hu, W. et al., 2004. A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(3), pp.334–352.

Ihaka, R., 1998. R: Past and future history. *COMPUTING SCIENCE AND STATISTICS*, pp.392–396.

Javan Roshtkhari, M. & Levine, M.D., 2013. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision and Image Understanding*, 117(10), pp.1436–1452.

Jiang, F. et al., 2011. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 115(3), pp.323–333.

Jung, H.G. & Kim, J., 2010. Constructing a pedestrian recognition system with a public open database, without the necessity of re-training: an experimental study. *Pattern Analysis and Applications*, 13(2), pp.223–233.

Kaluza, B. et al., 2011. Towards Detection of Suspicious Behavior from Multiple Observations. In *Plan, Activity, and Intent Recognition*. pp. 33–40.

Kaufman, L. & Rousseeuw, P., 1987. *Clustering by means of medoids*, North-Holland.

Ke, S.-R. et al., 2013. A review on video-based human activity recognition. *Computers*, 2(2), pp.88–131.

Ke, Y. & Sukthankar, R., 2004. PCA-SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. pp. II–506.

Koller, D. et al., 1994. Towards robust automatic traffic scene analysis in real-time. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*. pp. 126–131.

Kumari, S. & Mitra, S.K., 2011. Human action recognition using DFT. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on*. pp. 239–242.

Kuncheva, L. & Rodríguez, J., 2007. An experimental study on rotation forest ensembles. *Multiple Classifier Systems*, pp.459–468.

Lades, M. et al., 1993. Distortion invariant object recognition in the dynamic link architecture. *Computers, IEEE Transactions on*, 42(3), pp.300–311.

Leach, M.J.V., Sparks, E.P. & Robertson, N.M., 2014. Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, 44, pp.71–79.

Leutenegger, S., Chli, M. & Siegwart, R.Y., 2011. BRISK: Binary Robust invariant scalable keypoints. *2011 International Conference on Computer Vision*, pp.2548–2555.

Li, C. et al., 2013. Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. *Neurocomputing*, 119, pp.94–100.

Li, C. & Ma, L., 2009. A new framework for feature descriptor based on SIFT. *Pattern Recognition Letters*, 30(5), pp.544–557.

Liao, K., Liu, G. & Hui, Y., 2013. An improvement to the SIFT descriptor for image representation and matching. *Pattern Recognition Letters*, 34(11), pp.1211–1220.

Lipton, A.J., Fujiyoshi, H. & Patil, R.S., 1998. Moving target classification and tracking from real-time video. In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*. pp. 8–14.

Liu, H., Chen, S. & Kubota, N., 2013. Intelligent Video Systems and Analytics: A Survey. *Industrial Informatics, IEEE Transactions on*, 9(3), pp.1222–1233.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), pp.91–110.

Mair, E. et al., 2010. Adaptive and generic corner detection based on the accelerated segment test. In *Computer Vision--ECCV 2010*. Springer, pp. 183–196.

Majecka, B., 2009. Statistical models of pedestrian behaviour in the forum. *Master's thesis, School of Informatics, University of Edinburgh*.

Matas, J. et al., 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10), pp.761–767.

McKenna, S.J. et al., 2000. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1), pp.42–56.

Mikolajczyk, K. & Schmid, C., 2002. An affine invariant interest point detector. In *Computer Vision-ECCV 2002*. Springer, pp. 128–142.

Mikolajczyk, K. & Schmid, C., 2005. Performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10), pp.1615–30.

Miksik, O., 2012. Evaluation of Local Detectors and Descriptors for Fast Feature Matching. , (ICPR), pp.2681–2684.

Moeslund, T.B., Hilton, A. & Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3), pp.90–126.

Morel, J.-M. & Yu, G., 2009. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2), pp.438–469.

Moreno, P., Bernardino, A. & Santos-Victor, J., 2009. Improving the SIFT descriptor with smooth derivative filters. *Pattern Recognition Letters*, 30(1), pp.18–26.

Morris, B.T. & Trivedi, M.M., 2008. A survey of vision-based trajectory learning and analysis for surveillance. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(8), pp.1114–1127.

Morris, B. & Trivedi, M., 2009. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. pp. 312–319.

Morris, B.T. & Trivedi, M.M., 2011. Trajectory Learning for Activity Understanding : , 33(11), pp.2287–2301.

Do Nascimento, E.R. et al., 2013. On the development of a robust, fast and lightweight keypoint descriptor. *Neurocomputing*, 120, pp.141–155.

Newlinshebiah, R., Sivasubbu, S.P. & Sivasankar, V., 2015. Human Interactions Recognition using Bag of Words. , (6), pp.49–52.

Ng, L.L. & Chua, H.S., 2012. Event Recognition in Parking Lot Surveillance System. , pp.875–878.

Nguyen, D.T., Ogunbona, P. & Li, W., 2009. Human detection based on weighted template matching.

Ojala, T., Pietikainen, M. & Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7), pp.971–987.

Ouyang, Y., Chen, W. & Xing, J., 2012. Human Detection Under Partial Occlusion with Weighted Edge Elastic Gradient Potential Energy of Window ⋆ Weighted Edge Elastic Gradient Potential Energy of The Window WEGPE. , 1(January), pp.267–274.

Overett, G. et al., 2008. A new pedestrian dataset for supervised learning. *IEEE Intelligent Vehicles Symposium, Proceedings*, pp.373–378.

Papageorgiou, C. & Poggio, T., 2000. A Trainable System for Object Detection. *International Journal of Computer Vision*, 38(1), pp.15–33.

Papakostas, G. a. A. et al., 2013. Moment-based local binary patterns: A novel descriptor for invariant pattern recognition applications. *Neurocomputing*, 99, pp.358–371.

Pedersoli, M. et al., 2014. Toward Real-Time Pedestrian Detection Based on a Deformable Template Model. *IEEE Transactions on Intelligent Transportation Systems (2014)*, 15(1), pp.355–364.

Permuter, H., Francos, J. & Jermyn, I., 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4), pp.695–706.

Piciarelli, C. & Foresti, G.L.G.L., 2006. On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, 27(15), pp.1835–1842.

Poggio, C.P. and T., 2000. A trainable system for objet detection. *International Journal of Computer Vision*, 38(I), pp.15–33.

Poppe, R., 2010. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), pp.976–990.

Rabiner, L.R. & Juang, B.H., 1993. Fundamentals of speech Recognition (Prentice Hall PTR. *Englewood Cliffs, New Jersey*.

Rosten, E. & Drummond, T., 2006. Machine learning for high-speed corner detection. In *Computer Vision--ECCV 2006*. Springer, pp. 430–443.

Rublee, E. & Rabaud, V., 2011. ORB: an efficient alternative to SIFT or SURF. *Computer Vision (ICCV …*, pp.2564–2571.

Schaeffer, C., 2012. A Comparison of Keypoint Descriptors in the Context of Pedestrian Detection: FREAK vs. SURF vs. BRISK. , pp.1–5.

Schauland, S. et al., 2006. Vision-Based Pedestrian Detection -- Improvement and Verification of Feature Extraction Methods and SVM-Based Classification. *2006 IEEE Intelligent Transportation Systems Conference*, pp.97–102.

Scovanner, P., Ali, S. & Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*. pp. 357–360.

Sermanet, P. et al., 2013. Pedestrian Detection with Unsupervised Multi-stage Feature Learning. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.3626–3633.

Stauffer, C. & Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*

Sun, J. et al., 2009. Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. pp. 2004–2011.

Toyama, K. et al., 1999. Wallflower: Principles and practice of background maintenance. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*. pp. 255–261.

Verma, A., Liu, C. & Jia, J., 2011. New colour SIFT descriptors for image classification with applications to biometrics. *International Journal of Biometrics*, 3(1), pp.56–75.

Vlachos, M., Kollios, G. & Gunopulos, D., 2002. Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on*. pp. 673–684.

Wang, H. et al., 2009. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*.

Wang, M., Qiao, H. & Zhang, B., 2011. A new algorithm for robust pedestrian tracking based on manifold learning and feature selection. *Intelligent Transportation Systems, IEEE Transactions on*, 12(4), pp.1195–1208.

Weinland, D., Ronfard, R. & Boyer, E., 2011a. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2), pp.224–241.

Wiskott, L. et al., 1997. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7), pp.775–779.

Wojek, C. & Schiele, B., 2008. A performance evaluation of single and multi-feature people detection. In *Pattern Recognition*. Springer, pp. 82–91.

Wojek, C., Walk, S. & Schiele, B., 2009. Multi-Cue onboard pedestrian detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp.794–801.

Yu, J., Jeon, M. & Pedrycz, W., 2014. Weighted feature trajectories and concatenated bag-of-features for action recognition. *Neurocomputing*, 131, pp.200–207.

Zhang, Z., Huang, K. & Tan, T., 2006. Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes. *18th International Conference on Pattern Recognition (ICPR'06)*, pp.1135–1138.