Videovigilancia inteligente de personas: métodos con cámaras fijas, aéreas o múltiples

Memoria explicativa de la tesis

PAU CLIMENT PÉREZ

2018-01-29

Introducción

La videovigilancia de individuos, pequeños grupos y multitudes resulta de importancia en las sociedades actuales en las que la sobrepoblación en espacios urbanos va en alza y las aglomeraciones son cada vez más comunes. Los lugares de concentración de personas, tales como aeropuertos, estaciones de tren y redes de metro, pero también recintos para conciertos y grandes manifestaciones, necesitan de una estricta vigilancia para evitar incidentes (deliberados o no) que podrían causar centenares de muertes y lesiones graves. Como consecuencia, los operadores de seguridad de todo el mundo demandan sistemas capaces de manejar estas situaciones, así como de señalizar acontecimientos extraños, e inferir conocimiento avanzado a partir de múltiples fuentes o canales de vídeo.

En los últimos años, muchos países desarrollados han visto un incremento en la instalación de cámaras en circuitos cerrados de televisión (CCTV) para estos fines (esto es, seguridad pública, o de bienes, reducción del crimen), hasta el punto de ser ubicuas. No obstante, esta gran cantidad de datos es raramente procesada por algoritmos avanzados de visión por computador, sino más bien usadas como disuasorio sobre delincuentes, así como análisis forense de incidentes una vez estos han ocurrido. En el pasado, ha habido propuestas de soluciones automáticas con una única cámara, y en menor medida, múltiples cámaras fijas conectadas en red. La utilización de múltiples cámaras es una forma efectiva de mitigar o contrarrestar los efectos de las oclusiones entre personas y objetos, ya que éstas son un factor limitante en los sistemas de cámara única. Asimismo, con la llegada y abaratamiento recientes de los vehículos aéreos no tripulados (UAV, por sus siglas en inglés), es posible el despliegue de sistemas de videovigilancia en áreas apartadas donde la instalación de cámaras fijas no es posible o deseable.

En cuanto a la naturaleza del análisis realizado por los algoritmos, ésta depende del número de personas presentes en la escena, calculada midiendo la densidad (u otras características) de la multitud. En escenarios más dispersos, las personas se pueden seguir mediante un sistema de seguimiento (*tracker*) multi-objetivo, mientras que en grandes multitudes, los métodos que tomen a la masa como una entidad por si misma resultarán de más utilidad. Por ello, los distintos niveles de aglomeración se traducen en niveles de análisis: esto es, *microscópico* y *macroscópico*, respectivamente. En un nivel intermedio habría un análisis *mesoscópico*, en el que las señales del nivel microscópico (por ejemplo la trayectoria de individuos obtenida mediante *tracking*) se utilizan para obtener información de todos los individuos que forman la muchedumbre, y por lo tanto, se infiere conocimiento de ésta como un todo.

Dentro de este contexto de métodos de videovigilancia, los **objetivos principales** de esta tesis son pues la introducción de **métodos de evaluación del nivel de granularidad de las multitudes**, que sirvan como primer paso para determinar la naturaleza de los métodos a usar a continuación. Una vez esto queda determinado, y evitando las limitaciones existentes en los métodos de cámara fija única, se presentan varios métodos de análisis: desde **cámaras instaladas en vehículos aéreos**, y desde

múltiples cámaras fijas.

Esto finaliza el primer capítulo de la tesis, que introduce el problema, los objetivos y las contribuciones principales. El resto de capítulos sigue como se detalla a continuación.

En el **capítulo 2**, el lector es familiarizado con los temas de investigación vinculados a esta tesis, más concretamente: el análisis de multitudes mediante vídeo; el seguimiento visual de individuos (*visual tracking*), el seguimiento multi-sujeto; desde múltiples cámaras u otros dispositivos; así como el seguimiento desde plataformas aéreas. Primero, se presenta en más profundidad la topología de métodos de análisis ya descrita. Asimismo, se presenta cómo los resultados de los análisis de cada nivel pueden interactuar entre sí, y usarse para el análisis a otro nivel. Finalmente, se recogen las necesidades o lagunas de investigación encontradas, ya que de éstas surgen los métodos desarrollados en los capítulos posteriores.

Clasificación de multitudes mediante una firma densidad–entropía

En el **capítulo 3** se presenta un método de análisis de multitudes a nivel *macroscópico*, para evaluar la *granularidad* de una multitud. La mayoría de los trabajos anteriores de la bibliografía se basan en la densidad como única medida para evaluar el nivel de riesgo de una multitud. Sin embargo, una multitud puede ser segura siempre y cuando sea ordenada (ausencia de violencia o brusquedad). Por tanto, en el método propuesto, junto a la densidad, se introduce una medida que utiliza la entropía para calcular el orden (o ausencia de éste) de la multitud. Muy pocos trabajos parecen utilizar la entropía de la forma en que es utilizada en este trabajo en concreto. En el método propuesto, se usan dos estimadores: para la densidad ρ y la entropía ε. Los valores de cada uno de estos estimadores conforma una *firma*, que se utiliza en un espacio $2D$ para categorizar las distintas escenas. Los resultados que se presentan muestran el potencial de este método, que asimismo es comparado con las anotaciones manuales que conforman una realidad base (*ground truth*) obtenidas mediante un proceso innovador de recogida de datos que también se describe.

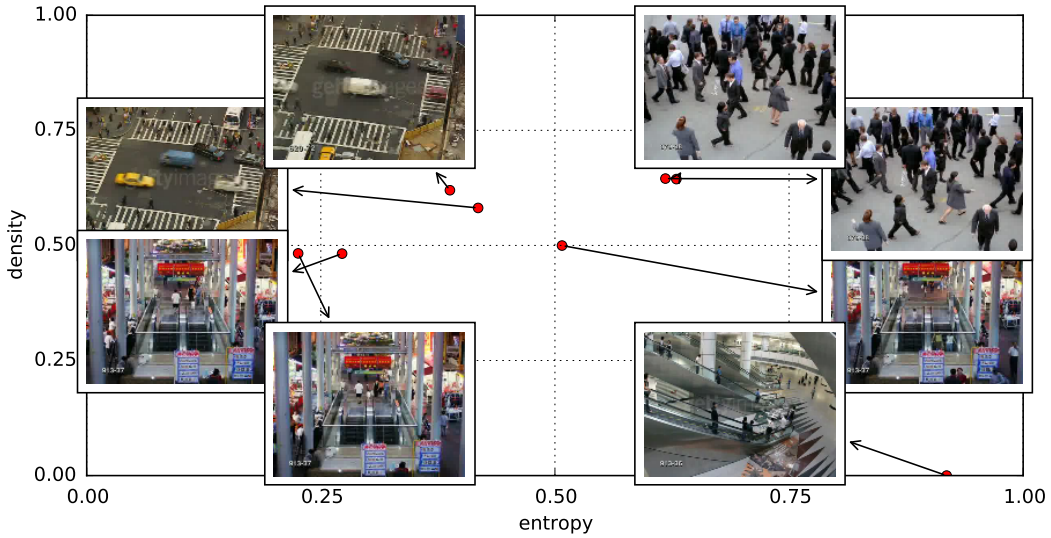

Figura1: Resultados cualitativos para la clasificación de escenas con aglomeraciones, basadas en su firma (ρ, ε). Se muestran 8 ejemplos seleccionados (2 por cuadrante). Se puede observar como en el cuadrante superior derecho las escenas tienen una mayor densidad y entropía, mientras que el resto son mucho más dispersas en una o ambas dimensiones analizadas.

Métodos de videovigilancia aérea basados en telemetría

En el **capítulo 4**, el objetivo es desarrollar métodos novedosos para realizar tareas específicas de videovigilancia aérea. Se presentan dos métodos para la compensación del movimiento basados no únicamente en imagen, sino también utilizando datos de telemetría.

En el flujo de trabajo clásico, primero se calcula la homografía entre fotogramas, y se compensa o registra el nuevo fotograma sobre el anterior, mediante un proceso de mosaico. Todo análisis posterior, depende de la exactitud de este proceso, que puede fallar en imágenes con poca textura en el fondo, y ser computacionalmente intenso. La necesidad de este proceso, sin embargo, depende de la aplicación final. Como se ve en este capítulo, no es necesario para el seguimiento de personas, aunque sea inevitable para el modelado de fondo (*background modelling*). Estos son, de hecho, los dos métodos presentados: un primer método para compensar el movimiento propio (*ego-motion*) del vehículo aéreo mediante telemetría, y de esta forma calcular la nueva posición de la ventana de búsqueda del *tracker* visual. Este método es mucho más eficiente desde el punto de vista computacional, y se puede aplicar con independencia de la textura en el fondo de la imagen. Un segundo método permite la detección de objetos en movimiento (*foreground detection*) desde UAVs, sirviéndose de datos de posicionamiento global (GPS) y de navegación inercial (INS), y un proceso de refinado global posterior.

Se demuestra que el primer método mejora las capacidades de seguimiento del sistema, puesto que

se puede acotar la ventana de búsqueda al área donde la persona debería estar según el movimiento propio al vehículo, reduciendo el coste computacional asociado a la búsqueda en un espacio mayor. Asimismo, el segundo método demuestra que la combinación de telemetría y refinado produce una mejor superposición entre fotogramas que los métodos clásicos basados en correspondencia de imágenes (*image registration*), especialmente cuando la textura del fondo es muy pobre o inexistente.

El capítulo presenta también dos conjuntos de datos (*datasets*) con varias secuencias de vídeo de referencia (*benchmark videos*) captadas con un prototipo desarrollado durante el proyecto europeo PROACTIVE. Los vídeos presentan distintas texturas de fondo. Esto es, en entornos urbanos (sobre asfalto o cemento) o rurales con poco arbolados, que son entornos comunes en el análisis de multitudes humanas, y en los que los métodos clásicos de correspondencia de fotogramas fallan.

Figura2: Fotogramas 2220, 2240 y 2260 de la secuencia `blue`. En 40 frames, el UAV sufre una fuerte rotación sobre el eje Z (viraje). El método propuesto sigue el objetivo con éxito (en verde brillante) usando una ventana de búsqueda pequeña (en rojo oscuro). El *ground truth* (verde oscuro) también se muestra para comparar, junto a la ventana que se usará en el próximo fotograma (en rojo brillante) calculada centrando una ventana alrededor del resultado actual.

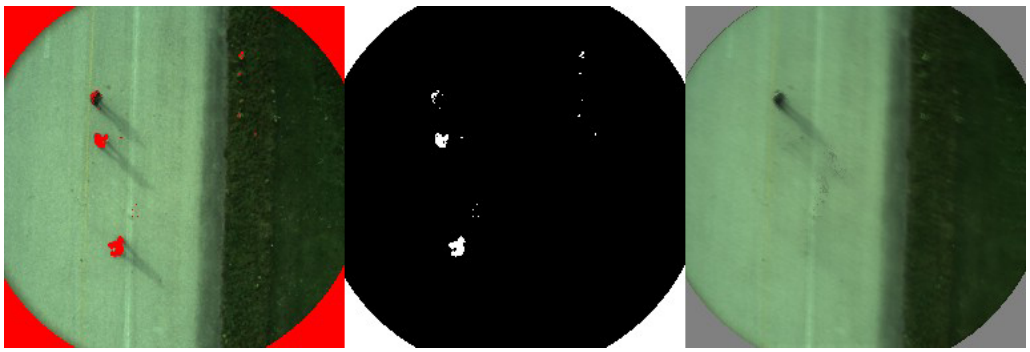

Figura3: Ejemplo de primer plano (*foreground*) segmentado 43 frames tras la inicialización. La imagen de la izquierda es el primer plano etiquetado, en el centro la máscara del primer plano, y a la derecha se observa el modelo de fondo promedio (*averaged background model*).

Análisis del comportamiento de multitudes desde la perspectiva individual

El **capítulo 5** está dedicado al análisis del comportamiento de multitudes y pequeños grupos de personas en escenas capturadas desde múltiples cámaras. Se presenta un *dataset* (véase la sección 5.3 de la tesis original), puesto que el estudio de la bibliografía muestra que ninguno de los conjuntos existentes hasta la fecha serían apropiados para la tarea planteada. Se introduce el concepto de *tracklet plots*, un diagrama de segmentos de trayectorias superpuestos, obtenidas mediante un *tracker* visual multi-sujeto durante periodos cortos de tiempo. Estos diagramas son agregados en un vector de características que describe la escena entera desde el punto de vista de cada cámara. Los vectores obtenidos de múltiples cámaras pueden combinarse, y ser utilizados para clasificar las escenas en una de las categorías predefinidas. Se utiliza un modelo de bolsa de palabras (*bag-of-words*) para caracterizar las secuencias de vídeo utilizando los vectores de características disponibles como palabras, y creando así bolsas para cada secuencia de entrenamiento. Para las secuencias a reconocer se sigue este mismo proceso, comparando la bolsa obtenida mediante el algoritmo del vecino más cercano (*nearest neighbour*).

Figura4: Ejemplo de diagrama de segmentos de trayectorias superpuestos o *tracklet plot*.

Conclusiones

Finalmente, el **capítulo 6** muestra un sumario de los hallazgos, los resultados obtenidos, algunas limitaciones y sugerencias de integración entre las contribuciones presentadas, así como de mejora mediante futuras investigaciones. A continuación se detallan algunos de los hallazgos brevemente.

Cuanto al descriptor de multitudes mediante densidad–entropía (capítulo 3), los resultados cuantitativos y cualitativos obtenidos ilustran el potencial del método. Como se observa de los resultados, cada

uno de los estimadores (densidad y entropía) funcionan generalmente bien (80 % y 73 %, respectivamente). En los cuadrantes de la Figura 1, se observan distintos tipos de escenas, clasificados según densidad y entropía. Se puede observar que las imágenes del cuadrante superior derecho muestran escenas de mayor densidad e igualmente mayor desorden en el movimiento de las personas. Esto permite señalizar este tipo de escenas como más de riesgo, e igualmente marcarlas para ser derivadas a un análisis macroscópico, ya que de otra forma se obtienen peores resultados por las oclusiones.

En cuanto a los experimentos llevados a cabo en los métodos de videovigilancia aérea (capítulo 4), para validar el primer método se utilizará una medida con fin de determinar como de contenida está la ventana de búsqueda corregida con respecto a la posición del individuo obtenida del *ground truth*. Se calcula que en el 99,7 % de los casos es así (con una desviación baja, del 1,4 %). Esto demuestra, que en condiciones de *tracking* perfecto, la corrección de la ventana de búsqueda mediante telemetría apenas fallará. En cuanto a la comparación con un *tracker* sin corrección de la ventana, y el método propuesto, se observa una mejora del 50 % (factor 1,5) sobre los valores de superposición (*overlap*) PASCAL, con mejoras puntuales que llegan a un factor de 3,31, especialmente cuando se dan rotaciones entorno al eje vertical (viraje).

Para el segundo método del capítulo 4, se compara el método propuesto de modelado del fondo (*background modelling*) mediante telemetría con refinado global tanto con información mutua (MI) o transformada discreta de Fourier (DFT) con métodos basados únicamente en imagen que utilizan detectores de esquinas (Harris, SIFT). Los resultados muestran que, independientemente de la textura del fondo, el método propuesto supera los métodos comparados. Además, si se añadiese el método de refinado a los métodos basados únicamente en imagen (y esto sería una novedad), estos mejorarían respecto al método propuesto. Sin embargo, esto último solo se da en los casos en que la textura del fondo es prominente, ya que la correspondencia entre fotogramas falla en otro caso (por estar basados solo en imagen). Aun más, el método propuesto funciona en tiempo real (frecuencia interactiva de vídeo), mientras que la mayoría de detectores de esquinas son bastante más lentos y no permiten estas altas frecuencias.

En cuanto al capítulo 5, cuando se obtienen los resultados de cada una de las vistas por separado, como método base de referencia (*baseline*), se observa que la mejor vista tiene una tasa de acierto del 82,4 % (cuando el número de palabras clave $K = 6$). Cuando se combina información de múltiples vistas el sistema puede aprovechar el resultado de la vista con mejor tasa de acierto. Esto es, la combinación de datos no resulta en un coste añadido sin beneficio, sino que sirve para que se use la mejor de las decisiones disponibles. Aunque, como la dimensionalidad del vector de características crece de forma lineal con el número de cámaras, es gracias a la reducción de la dimensionalidad aplicada, que el sistema propuesto queda justificado: el tiempo de entrenamiento será más corto (por la reducción en las dimensiones), y además el sistema podrá aprovecharse del mejor resultado disponible en cada momento (piénsese en oclusiones temporales de algunas vistas), sin conocimiento previo acerca de la cámara (vista) que está aportando la mejor respuesta.