PROYECTO FIN DE MÁSTER EN

SISTEMAS INTELIGENTES

CURSO 2015-2016

IDENTIFICACIÓN DE LA FUENTE EN VÍDEOS DE DISPOSITIVOS MÓVILES

Raquel Ramos López

Directores:

Luis Javier García Villalba

Ana Lucila Sandoval Orozco

Departamento de Ingeniería del Software e Inteligencia Artificial

Convocatoria de Septiembre

Calificación: 9.5 - Sobresaliente

MÁSTER EN INVESTIGACIÓN EN INFORMÁTICA

FACULTAD DE INFORMÁTICA

UNIVERSIDAD COMPLUTENSE DE MADRID

El presente Trabajo Fin de Máster se enmarca dentro de un proyecto de investigación titulado RAMSES aprobado por la Comisión Europea dentro del Programa Marco de Investigación e Innovación Horizonte 2020 y en el que participa el Grupo GASS del Departamento de Ingeniería del Software e Inteligencia Artificial de la Facultad de Informática de la Universidad Complutense de Madrid (Grupo de Análisis, Seguridad y Sistemas, http://gass.ucm.es, grupo 910623 del catálogo de grupos de investigación reconocidos por la UCM).

Por razones de confidencialidad del proyecto se ha omitido información del trabajo desarrollado para no infringir la normativa correspondiente.

_____

Raquel Ramos López

_____ _____

Luis Javier García Villalba Ana Lucila Sandoval Orozco

Abstract

An increasing number of mobile devices with integrated cameras has meant that most digital video comes from these devices. These digital videos can be made anytime, anywhere and for different purposes. They can also be shared on the Internet in a short period of time and may sometimes contain recordings of illegal acts. The need to reliably trace the origin becomes evident when these videos are used for forensic purposes. This work proposes an algorithm to identify the brand and model of mobile device which generated the video. Its procedure is as follows: after obtaining the relevant video information, a classification algorithm based on sensor noise and Wavelet Transform performs the aforementioned identification process.

Keywords

Digital Video, Forensics Analysis, Mark, Mobile Device, Model, Key Frame, Photo Response Non Uniformity, PRNU, Sensor Noise, Source Identification, Support Vector Machines, SVM, Wavelet Transform.

Resumen

El incesante aumento del número de dispositivos móviles con cámaras integradas ha originado que la mayoría de los vídeos digitales procedan de este tipo de dispositivos. Estos vídeos digitales pueden ser realizados en cualquier momento, en cualquier lugar y con diferentes fines, distribuyéndose en Internet en un corto período de tiempo y mostrando en ocasiones actos ilegales. La necesidad de establecer de forma fiable el origen se hace evidente cuando se utilizan estos vídeos en un contexto forense. En este trabajo se propone un algoritmo para identificar la marca y el modelo del dispositivo móvil que generó el vídeo. Su funcionamiento es como sigue: tras extraer la información relevante del vídeo, un algoritmo de clasificación, basado en el ruido del sensor y la Transformada Wavelet, realiza el proceso de identificación del dispositivo móvil.

Palabras clave

Análisis Forense, Dispositivo Móvil, Fotograma Clave, Identificación de la Fuente, Máquina de Soporte Vectorial, Marca, Modelo, PRNU, Ruido del Sensor, SVM, Transformada Wavelet, Vídeo Digital.

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todos los miembros del Grupo GASS.

Lista de acrónimos

AC	Alternative Current
CCD	Charge Coupled Device
CFA	Color Filter Array
CIELUV	CIE (L*U*V*)-space
CMM	Convex Mixture Model
CMOS	Complementary Metal Oxide Semiconductor
CMY	Cyan-Magenta-Yellow
CP	Conditional Probability
CYGM	Cyan-Yellow-Green-Magenta
CYYM	Cyan-Yellow-Yellow-Magenta
DC	Direct Current
DCT	Discrete Cosine Transform
DIP	Digital Image Processor
DSC	Digital Still Camera
DT	Distance Threshold
FPN	Fixed Pattern Noise
GOP	Group of Pictures
GRGB	Green-Red-Green-Blue
HSV	Hue Saturation Value
HSV	Human Visual System
JPEG	Joint Photographic Experts Group
MOS	Mean Opinion Score

MSE	Media Source Extensions
MPEG	Moving Picture Experts Group
PNU	Pixel Non-Uniformity
PRNU	Photo Response Non Uniformity
PSNR	Peak Signal to Noise
RBF	Radial Basis Function
RGB	Red-Green-Blue
RGBE	Red-Green-Blue-Emerald
RGBW	Red-Green-Blue-White
SOM	Self-Organizing Map
SPN	Sensor Patter Noise
SVM	Support Vector Machine

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

La creciente expansión de los teléfonos móviles se debe a la infinidad de tareas que pueden realizar: no sólo sirven para hacer o recibir llamadas sino que también permiten almacenar datos, realizar fotografías o vídeos, etc.

Al igual que las cámaras digitales han desplazado a las cámaras tradicionales, los dispositivos móviles equipados con cámara tienen un papel fundamental en la disminución del crecimiento que tenían las cámaras digitales. En un informe realizado por IC Insights [1] se pronostica que en 2016 la venta de cámaras digitales DSCs (*Digital Still Camera*) se reducirá del 47% obtenido en 2012 al 27%. De igual forma, se estima también que en 2016 las ventas de cámaras digitales integradas en teléfonos inteligentes, tabletas u ordenadores personales aumentarán al 42% frente al 31% de 2012.

Por otro lado, según el medidor de tráfico "Alexa, The web Information Company" [2], Youtube es actualmente el segundo sitio con más visitas del mundo.

Asimismo, la irrupción de los vídeos en la sociedad actual no sólo debe medirse en cifras. Un vídeo es frecuentemente utilizado en la vida diaria debido a la disponibilidad de una amplia gama de dispositivos móviles que pueden grabar y/o reproducirlos. Este uso continuado hace que en ciertos casos existan restricciones legales o limitaciones a su utilización en distintos lugares como, por ejemplo, colegios, universidades, empresas, etc. Pero al mismo tiempo, los vídeos se exhiben con mayor frecuencia en procesos judiciales como pruebas o evidencias para la aplicación de la ley [3]. Sin embargo, la manipulación del vídeo digital es cada vez más fácil debido a la aparición de nuevas y potentes herramientas de procesamiento multimedia. El papel de los expertos forenses en vídeo se torna entonces crucial para averiguar evidencias válidas a efectos legales durante la investigación de un delito.

El resto de este capítulo está organizado como sigue: el apartado 1.1 presenta el objeto de investigación de este trabajo. En el apartado 1.2 se analizan algunos trabajos relacionados. En último lugar, el apartado 1.3 resume la estructura del resto del trabajo.

1.1. Objeto de la Investigación

La mayoría de las investigaciones realizadas en los últimos años se centran en las técnicas de identificación de la fuente en imágenes, siendo casi inexistente la literatura en el caso de vídeo.

La fuente de un vídeo digital se puede identificar a través de los rasgos que impregna en él durante el proceso de creación, manifestándose en cada uno de los fotogramas estos defectos como ruido, comúnmente llamado "huella digital".

La presente investigación se centra en las técnicas de extracción de fotogramas claves de un vídeo que permitan identificar el dispositivo móvil que lo generó mediante el ruido del sensor y la Transformada Wavelet.

1.2. Contexto de la Investigación

El presente Trabajo Fin de Máster se enmarca dentro de un proyecto de investigación titulado RAMSES aprobado por la Comisión Europea dentro del Programa Marco de Investigación e Innovación Horizonte 2020 y en el que participa el Grupo GASS del Departamento de Ingeniería del Software e Inteligencia Artificial de la Facultad de Informática de la Universidad Complutense de Madrid (Grupo de Análisis, Seguridad y Sistemas, http://gass.ucm.es, grupo 910623 del catálogo de grupos de investigación reconocidos por la UCM).

Seguidamente se detallan algunos datos:

Convocatoria: H2020-FCT-2015

Tipo de Propuesta: Acción de Innovación

Número de Propuesta: 700326

Acrónimo de la Propuesta: RAMSES

Título de la Propuesta: Internet Forensic Platform for Tracking the Money Flow of Financially-Motivated Malware

Entidad financiadora: Comisión Europea, Horizonte 2020 – Programa Marco de Investigación e Innovación

Entidades participantes: Policia Judiciária (Portugal), Belgian Federal Police (Bélgica), Research Centre on Security and Crime (Italia), Politecnico di Milano (Italia), College of the Bavarian Police (Alemania), Saarland University (Alemania), Trilateral Research and Consulting (Reino Unido), University of Kent (Reino Unido), Dirección General de la Policía (España), Treelogic Telemática y Lógica Racional para la Empresa Europea S.L. (España), Universidad Complutense de Madrid (España).

Duración, desde: 01-09-2016 hasta: 31-08-2019

Investigador responsable (UCM): Luis Javier García Villalba

1.3. Estructura de la Memoria

Además de este capítulo introductorio, la memoria se estructura en 6 más:

El Capítulo 2 introduce los conceptos elementales del análisis forense de vídeos, describiendo el proceso de creación de un vídeo digital, los elementos de la cámara que sirven de base para las técnicas forenses y los componentes de

un vídeo.

El Capítulo 3 comienza revisando las principales técnicas de extracción de fotogramas claves que hay en la literatura. A continuación presenta los procedimientos más relevantes para identificar la fuente en vídeos digitales.

El Capítulo 4 presenta las contribuciones de este trabajo: un algoritmo de extracción de fotogramas representativos basado en el histograma de color y un algoritmo basado en el ruido del sensor y en la Transformada Wavelet para la identificación de la marca y modelo del dispositivo móvil fuente.

El Capítulo 5 describe la experimentación realizada para evaluar la efectividad de los algoritmos propuestos, mostrando los resultados obtenidos.

Por último, el Capítulo 6 resume las conclusiones extraídas de este trabajo así como algunas líneas futuras de trabajo.

2. CAPTURA Y GENERACIÓN DE UN VÍDEO DIGITAL

El objetivo de este capítulo es describir los pasos necesarios para generar un vídeo digital, así como los componentes que participan en este proceso. Una vez se ha creado un vídeo, es necesario estudiar de qué partes se compone. Estos conceptos serán la base de las técnicas de análisis forense descritas en los capítulos siguientes.

2.1. Creación de un Vídeo Digital

Para representar una escena real con la mayor fidelidad posible se utiliza una función de intensidad luminosa en cada punto de coordenadas (x, y). Si se elige un conjunto finito de puntos representativos de la escena real (muestreo) y se define su brillo mediante un número concreto de bits (cuantificación), el resultado es la codificación de la escena real [25].

Si se tiene en cuenta que un vídeo es una secuencia de imágenes que aparecen en pantalla en sucesión rápida produciendo sensación de movimiento, lo primero que se debe conocer son los elementos que forman parte de una imagen y cómo se genera ésta en un dispositivo digital.

Una cámara digital está compuesta por los siguientes elementos: sistema de lentes, filtros, matriz de filtros de colores (*Color Filter Array*, CFA), sensor de imagen y procesador de imagen (*Digital Image Processor*, DIP).

En la Figura 2.1 se muestran las diferentes etapas que intervienen en la generación de una imagen digital en un dispositivo móvil:

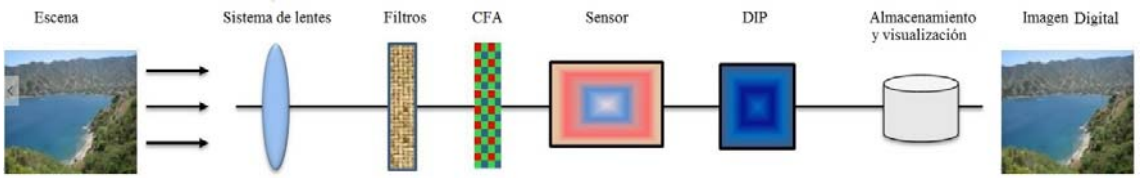

Fig. 2.1. Proceso de adquisición de imágenes en cámaras digitales.

La lente de una cámara está formada por múltiples componentes ópticos que se encargan de capturar la luz (o conjunto de fotones) de la escena controlando la exposición, el foco y la estabilidad de la imagen. La luz que entra en la cámara a través de las lentes pasa por un grupo de filtros que mejora la calidad de la imagen. A continuación, la luz pasa hacia el sensor, que se encarga de convertirla en señales eléctricas. El sensor está compuesto por una agrupación de elementos fotosensibles denominados píxeles dispuestos en celdas. Cuanto mayor sea, mayor cantidad de luz podrá recoger de la cámara y, por tanto, mayor información habrá para componer la imagen. Cada celda del sensor es un píxel (unidad mínima de una imagen) y cuanto más juntos se encuentren éstos, mayor será la calidad de la imagen que representan (concepto de resolución espacial). Cuando los fotones llegan al fotodiodo del sensor de imagen, cargan cada uno de los píxeles hasta que coincida con su nivel de luz, es decir, lo primero en una imagen (fotografía) es capturar la cantidad de luz exacta en cada uno de los millones de píxeles del sensor [24].

En la actualidad existen dos tipos de tecnologías de sensores CCD (*Charge Coupled Device*) y CMOS (*Complementary Metal Oxide Semiconductor*) y funcionan de forma similar, aunque la diferencia clave está en la forma en la que se digitalizan los píxeles y en cómo se lleva a cabo la lectura de las cargas. Los sensores CCD necesitan contar con un chip adicional para tratar la información de salida del sensor, haciendo que la fabricación del dispositivo sea más costosa y que los sensores sean más grandes. En contraste, los sensores CMOS cuentan con píxeles activos independientes, ya que ellos mismos realizan la digitalización ofreciendo velocidad, reduciendo el tamaño y el coste de los

sistemas que integran una cámara digital. Otra diferencia entre estos dos tipos de sensores es que los píxeles de una matriz CCD captan la luz simultáneamente, lo cual proporciona una salida más uniforme. Los sensores CMOS realizan la lectura generalmente como un barrido progresivo (evitando el efecto *blooming*). Los sensores CCD son muy superiores a los CMOS en rango dinámico y en términos de ruido. En contrapartida, los sensores CMOS son más sensibles a la luz y en condiciones de poca iluminación se comportan mejor. En sus inicios los sensores CMOS eran algo peor que los CCD, pero hoy día es una deficiencia que prácticamente está subsanada [4]. La tecnología CCD ha llegado a su límite y es ahora cuando se está desarrollando la tecnología CMOS, superándose sus deficiencias. Así, la mayoría de los teléfonos inteligentes contienen sensores de tipo CMOS.

Las señales almacenadas por el sensor CCD/CMOS se convierten posteriormente en una señal digital y se transmiten al procesador de imagen, quien elimina el ruido y otras anomalías introducidas.

Existen diversas fuentes de imperfecciones y de ruido introducidas en las diferentes etapas del proceso de generación de la imagen en la cámara. Incluso si se realiza una fotografía uniforme y completamente iluminada es posible observar pequeños cambios de intensidad entre los píxeles. Esto se debe al ruido del disparo que es aleatorio y, en gran parte, al patrón de ruido que es determinista y se mantiene aproximadamente igual si se capturan varias fotografías de la misma escena. El patrón de ruido en una imagen se refiere a cualquier patrón espacial que no cambia de una imagen a otra y está compuesto por el ruido espacial que es independiente de la señal o ruido de patrón fijo (*Fixed Pattern Noise*, FPN) y el ruido espacial debido a la diferencia de respuesta de cada píxel a la señal incidente o ruido de respuesta no uniforme (*Photo Response Non Uniformity*, PRNU) [4]. La estructura del patrón de ruido se ilustra en la Figura 2.2.

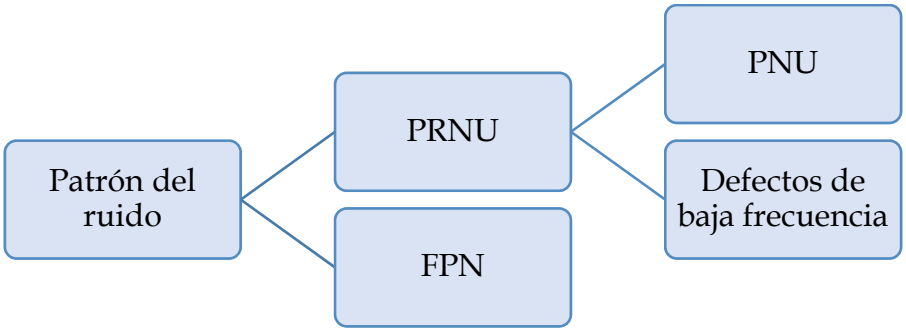

Fig. 2.2. Proceso de adquisición de imágenes en cámaras digitales.

El ruido FPN se genera por la corriente de oscuridad y también depende de la exposición y de la temperatura. Debido a que el ruido del patrón fijo es un ruido independiente aditivo, algunas cámaras lo eliminan automáticamente restando un fotograma oscuro a las imágenes que generan.

El ruido PRNU es la parte dominante del patrón de ruido de las imágenes y es un ruido dependiente multiplicativo. El ruido PRNU está formado principalmente por la uniformidad del píxel (*Pixel Non-Uniformity*, PNU) y por los defectos de baja frecuencia como la configuración del *zoom* y la refracción de la luz en las partículas de polvo y en las lentes.

El ruido PNU es la diferencia de sensibilidad a la luz entre los píxeles de la matriz del sensor. Se genera por la falta de homogeneidad de las obleas de silicio y las imperfecciones durante el proceso de fabricación del sensor. Debido a su naturaleza y origen es muy poco probable que incluso los sensores procedentes de la misma oblea presenten patrones PNU correlacionados. Este ruido no se ve afectado por la temperatura ambiente ni por la humedad.

El ruido PNU es normalmente más común, complejo y significativo en los sensores de tipo CMOS debido a la complejidad de la circuitería de la matriz de píxeles [4].

El siguiente paso es añadir el color: el sensor de imagen en la mayoría de las cámaras digitales es una matriz CFA. El sensor de la imagen por sí mismo es monocromático, es decir, detecta la intensidad de luz pero no el color. La matriz CFA superpone filtros con fotodiodos sensibles a la luz del sensor en un mosaico de colores rojo, verde y azul (*Red Green Blue*, RGB) en una cuadrícula como de ajedrez, de forma que cada lugar de la imagen (fotografía) correspondiente a un píxel recibe uno de los tres colores primarios.

El patrón RGB se conoce por el nombre de *mosaico* o *Filtro de Bayer* y está formado por un 50% de filtros verdes, un 25% de rojos y un 25% de azules. Un sensor CFA asigna el doble de píxeles al color verde en lugar de al rojo o al azul puesto que el ojo humano es más sensible a la luz verde. La visión humana depende en gran medida de las longitudes de onda en el rango del color verde para percibir detalles finos y la resolución de luminancia (o brillo de cada píxel, siendo máxima en el blanco y mínima en el negro). En la Figura 2.3 se puede apreciar que en la imagen de color verde el ojo humano es capaz de ver más detalles que en las imágenes de color azul o rojo.

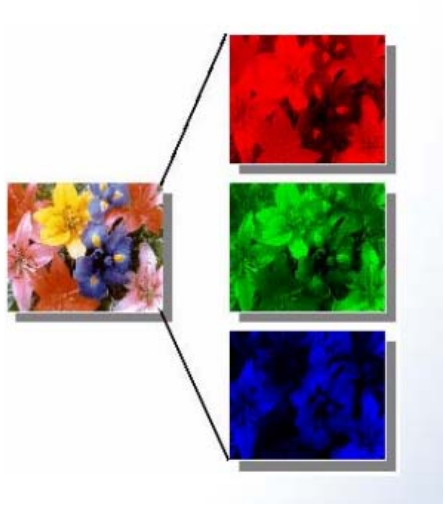

Fig. 2.3. Descomposición de una imagen en los colores primarios rojo, verde, azul (RGB).

Aparte del patrón RGB también se utilizan los siguientes modelos: *Red-Green-Blue-Emerald* (RGBE), *Cyan-Yellow-Yellow-Magenta* (CYYM), *Cyan-Yellow-Green-Magenta* (CYGM) o *Red-Green-Blue-White* (RGBW) [4].

Una vez descrito el proceso para generar la imagen dentro de un dispositivo digital, sólo quedar añadir a las imágenes una tercera dimensión, que es la temporalidad, para representar un vídeo. El conjunto de muestras será ahora tridimensional (x, y, t). Dado que el ancho de banda adecuado para la visión humana es del orden de 10 a 15 Hz, el nivel de muestreo debe estar comprendido entre 20 y 30 Hz, puesto que hay que muestrearla al menos 2 veces por ciclo (Teorema de Shannon). Esto significa que hay que usar al menos del orden de 20 a 30 fotogramas por segundo (fps) para obtener la representación de una secuencia de imágenes [25].

Teniendo en cuenta que un solo fotograma de vídeo con definición estándar utiliza casi 1 MB de almacenamiento y que un vídeo de media contiene 30 fps, un vídeo sin comprimir con una duración de 34 segundos de secuencia consume alrededor de 1 GB de almacenamiento. Evidentemente, es muy difícil tratar con estas cantidades de información, que crecen de forma vertiginosa cuando se mejora la calidad (resolución) de la señal de vídeo. La solución se encuentra en las técnicas de compresión [25].

2.2. Técnicas de Compresión Digital

En la mayoría de los formatos de vídeos se utiliza la compresión digital para disminuir el tamaño del archivo de manera que se garantice la correcta generación, transmisión, almacenamiento y visualización. Su implementación se apoya en el hecho de que la información no es aleatoria, puesto que presenta una cierta regularidad y un orden. Si dicha regularidad y orden se determinan, se puede eliminar la redundancia y la información se puede representar con los datos mínimos necesarios para su posterior reconstrucción [25].

La técnica de compresión se realiza al compilar el fichero de vídeo (*encoding*), mientras que la descompresión se realiza al proyectar el vídeo. El proceso de decodificación (*decoding*) es el proceso inverso al de la codificación. Las técnicas de compresión se basan en una serie de algoritmos que reducen la cantidad de información sin que el resultado se vea afectado (no se aprecie por el ojo humano).

Algunas técnicas de compresión no producen pérdidas (*lossless*) y otras sí (*lossy*). En los algoritmos de compresión sin pérdida o reversibles, la calidad del vídeo es idéntica a la del vídeo original, puesto que la compresión se realiza removiendo la información redundante, por lo que el proceso de compresión se puede revertir con exactitud. Sin embargo, tienen el inconveniente de que la razón de compresión es muy baja (1/2, 1/3). Este tipo de técnicas son utilizadas en el campo de la medicina, donde las imágenes digitales deben ser iguales a las originales. Estas técnicas no son aplicables a los *códecs* que utilizan los dispositivos móviles, ya que éstos usan *códecs* con pérdidas. En estos casos, la compresión no se puede revertir, si bien el usuario final no percibe la reducción de la calidad. Los algoritmos de compresión con pérdidas proporcionan tanto buena razón de compresión (desde 1/10 a 1/50 en imágenes fijas y hasta 1/200 en vídeos) como una adecuada reducción en la cantidad de información [25]. Para conseguir estos ratios tan elevados de compresión se apoyan tanto en la redundancia de datos (la información irrelevante no es necesaria y se elimina) como en las propiedades no lineales de la visión humana.

2.2.1. Redundancia en la Señal de Vídeo

El éxito de los buenos ratios de compresión en los vídeos se apoya en la disminución de las redundancias que presenta una señal de vídeo. Existen dos tipos principales, la estadística y la psicovisual, las cuales se sub-dividen para cubrir la mayor parte del espectro de posibilidades de compresión [26], tal y como se puede observar en la Figura 2.4.

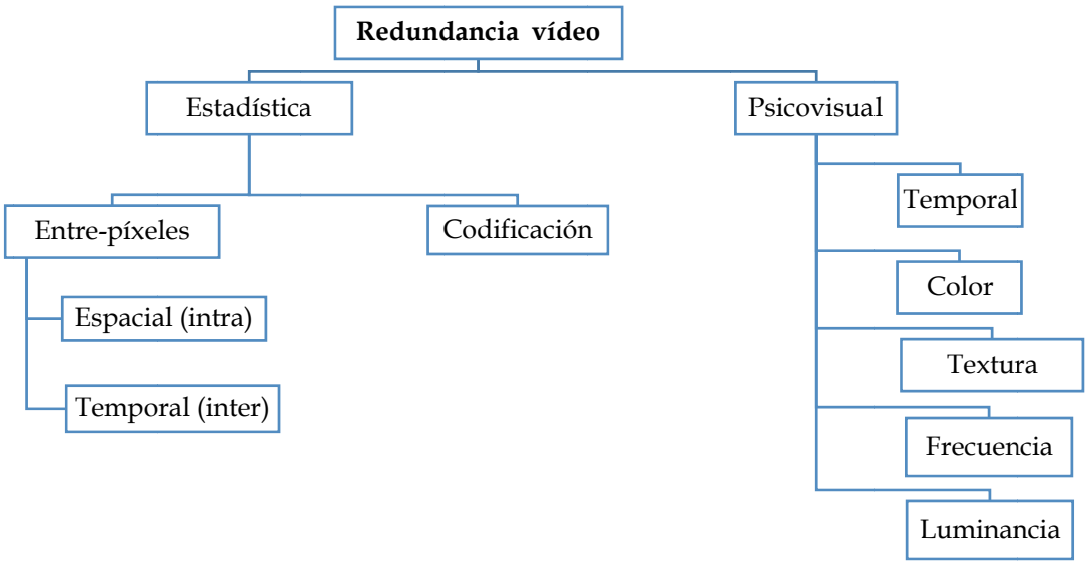

Fig. 2.4. Clasificación de la redundancia en el vídeo.

- **Redundancia estadística**: La redundancia estadística se clasifica en dos tipos: redundancia de codificación y redundancia *entre-píxeles*. La redundancia de codificación define la redundancia estadística asociada a las técnicas de codificación. La redundancia *entre-píxeles* se basa en la relación que hay entre los píxeles de un fotograma de vídeo y los píxeles de un grupo de fotogramas sucesivos puesto que no son estadísticamente independientes, ya que presentan diversos grados de dependencia (correlación). La redundancia *entre-píxeles* se divide a su vez en dos categorías, redundancia espacial y temporal.

    - **Redundancia espacial o intra**: Representa la correlación estadística entre los píxeles de un fotograma de vídeo. La redundancia espacial implica que el valor de la intensidad de un píxel puede ser estimado a partir de sus píxeles vecinos, lo que implica que los píxeles vecinos no son independientes. Uno de los primeros investigadores en estudiar las estadísticas del vídeo fue Kretzmer en los años 50, quien después de varios experimentos averiguó que la autocorrelación de los píxeles de un fotograma de vídeo en ambas direcciones horizontal y vertical

mostraban un rendimiento similar. La consecuencia directa es que, si se elimina la redundancia espacial, es posible representar un fotograma de vídeo con menos datos, pero con una eficiencia similar a la compresión de una imagen estática.

La codificación espacial se divide en codificación por predicción y codificación de la transformada del coseno (*Discrete Cosine Transform*, DCT). La transformada del coseno o DCT es un caso concreto de la Transformada de Fourier, donde la imagen transforma su representación espacial a su frecuencia equivalente. Cada elemento de la imagen es representado por determinados coeficientes de frecuencia. Las imágenes con muchos detalles se representan con coeficientes de alta frecuencia y las imágenes con poco detalle con coeficientes de baja frecuencia. Cuando la codificación *intra* o espacial trata cada imagen de forma independiente, se pueden emplear las técnicas de compresión desarrolladas para las imágenes fijas. El estándar de compresión ISO denominado JPEG se encuentra en esta categoría. Es posible codificar vídeo mediante una sucesión de imágenes codificadas en JPEG, conocida como "JPEG en movimiento".

- **Redundancia temporal o *inter***: Hace referencia a la correlación estadística entre píxeles de fotogramas consecutivos de una secuencia de vídeo. Esto permite predecir un fotograma a partir de los fotogramas vecinos en la dimensión temporal. La codificación temporal aprovecha la ventaja que existe cuando las imágenes sucesivas son similares. En lugar de enviar la información de cada imagen por separado, el codificador temporal envía la diferencia existente entre la imagen previa y la actual en forma de codificación diferencial. De esta forma se elimina la redundancia temporal, usando la información de las imágenes ya enviadas y enviando únicamente las zonas de la imagen que han cambiado de un fotograma a otro. El

codificador necesita la imagen que fue almacenada con anterioridad (*key frame)* para luego ser comparada. Un ejemplo de este tipo de compresión es el estándar *MPEG*, que se caracteriza porque puede mantener una alta calidad en la imagen a pesar de tener elevados ratios de compresión [27].

El compresor retiene toda la información de la secuencia cada cierta cantidad de fotogramas y servirá de referencia a la hora de saber qué tipo de información debe desechar hasta encontrar un nuevo fotograma de referencia. Se denominan *delta frames* a las diferencias entre la imagen no comprimida y la imagen de referencia. Una secuencia de imágenes formada por una imagen inicial (*key frame*) "I" y las siguientes imágenes "P" (*delta frames*) hasta el comienzo de otro *key frame* se denomina grupo de imágenes (*Group of Pictures*, GOP). Para elevados factores de compresión se utiliza un número grande de *delta frames*, haciendo que las GOPs aumenten de tamaño. Sin embargo, un GOP grande evita recuperar de manera eficaz una transmisión que llegue con errores.

- **Redundancia psicovisual**: Hace referencia a las características del sistema visual humano (*Human Visual System*, HVS). En el HVS la información visual se aprecia de diferentes formas. Cuando cierta información es más importante que otra, se utilizan más datos para representar la información. Bajo esta perspectiva se define qué parte de la información visual es psicovisualmente redundante, por lo que su eliminación trae consigo la compresión del vídeo. Los aspectos del HVS que están relacionados directamente con la compresión del vídeo son los dependientes de la imagen (textura y luminancia) y los independientes a ella (color, frecuencia y tiempo).

2.3. Estándares de Codificación en Dispositivos Digitales

Existen diferentes técnicas de compresión, tanto propietarias como estándares, siendo las últimas las más habituales en la mayoría de aplicaciones actuales. Los estándares son importantes para asegurar la compatibilidad y la interoperabilidad y tienen un papel especialmente relevante en la compresión de vídeo, puesto que éste se puede utilizar para varias finalidades con distintos requisitos.

En el proceso de compresión se aplica un algoritmo al vídeo original para crear un archivo comprimido listo para ser transmitido o almacenado. Para reproducir el archivo comprimido se aplica el algoritmo inverso y se crea un vídeo que incluye prácticamente el mismo contenido que el vídeo original (idealmente, el mismo). El tiempo que se tarda en comprimir, enviar, descomprimir y mostrar un archivo es lo que se denomina latencia. El par de algoritmos que funcionan conjuntamente se denomina *códec* de vídeo (**co**dificador/ **dec**odificador). Los diferentes estándares de compresión utilizan métodos distintos para reducir los datos y, en consecuencia, los resultados en cuanto a tasa de bits (*bit rate*) y latencia son diferentes. Existen dos tipos de algoritmos de compresión:

- **Compresión de imágenes**: utiliza la tecnología de codificación *intra-frame*. Los datos se comprimen fotograma a fotograma con el fin de eliminar la información innecesaria que puede ser imperceptible para el ojo humano.

- **Compresión de vídeo**: la compresión de vídeo (*inter-frame*) explota la redundancia temporal de la secuencia para aumentar el grado de compresión.

En un esfuerzo para encontrar un fondo común, el *Motion Picture Expert Group (MPEG)* desarrolla formatos de archivos estándar y algoritmos de compresión que la industria puede licenciar.

Los vídeos MPEG están diseñados en forma de estratos o capas. Cada secuencia de vídeo en MPEG se divide en uno o más GOPs. A su vez, cada grupo de imágenes contiene una o más imágenes (*picture*) de cuatro tipos posibles: I, P, B y D.

Las imágenes I se caracterizan por estar codificadas con técnicas *intraframe* de manera independiente al resto de las imágenes del GOP. Aunque ocupan un mayor tamaño benefician la sincronización y son necesarias para garantizar el acceso aleatorio. Tanto para las imágenes P como para las B se hace uso de las similitudes entre éstas y otras imágenes de referencia para su codificación. Las imágenes P obtienen predicciones de imágenes I o P anteriores, mientras que las imágenes B toman como referencia las imágenes I o P más cercanas tanto anteriores como posteriores. Existe la posibilidad de codificar este tipo de imágenes total o parcialmente sin usar ninguna predicción, es decir, con técnicas *intraframe*. Existe otro tipo de imagen definida en el estándar, las D, pero su uso no es común, se trata de una imagen en baja resolución que no puede ser usada en combinación con los otros tipos de imagen.

2.3.1. Componente DCT

La DCT permite descomponer un bloque de datos de tamaño 8x8 en una suma ponderada de frecuencias espaciales, cada una de las cuales está asociada a un coeficiente que representa la contribución de ese patrón de frecuencia al bloque analizado. Así, cada patrón de frecuencia se multiplica por su coeficiente asociado y las 64 matrices 8x8 resultantes son sumadas píxel a píxel para reconstruir el bloque original. Dependiendo de los coeficientes asociados a cada una de las frecuencias se infieren datos acerca del bloque del que provienen. Si sólo los coeficientes correspondientes a las bajas frecuencias son distintos de cero, entonces existirá poca variación entre los píxeles del bloque. Si, por el contrario, las altas frecuencias están presentes y son no nulas, la intensidad del bloque cambia rápidamente píxel a píxel.

La componente DC (*Direct Current)*, la frecuencia más baja, de cada bloque se corresponde con su valor medio e incluye la información más importante de la imagen desde un punto de vista perceptual. Las altas frecuencias espaciales AC (*Alternative Current*) contienen los detalles del bloque, por lo que, nuevamente atendiendo a criterios perceptuales, se podrían codificar con menor precisión que la componente DC. Esto se controla mediante el proceso de cuantificación. Cada uno de los coeficientes se divide por un valor entero no nulo denominado valor de cuantificación, y redondeando este cociente se obtiene finalmente el coeficiente DCT cuantificado. La tabla de cuantificación es una matriz con 64 valores, uno por cada coeficiente DCT. Existen dos tablas de cuantificación posibles para MPEG-1, una para *intra-techniques* que está en concordancia con la respuesta en frecuencia del ojo humano, y otra para *inter-techniques* con un valor fijo de 16 en todos sus componentes. Los valores de cuantificación *intra* se fijan de forma que se asocien a las altas frecuencias valores muy altos de cuantificación. Así, la precisión de estos coeficientes DCT cuantificados será menor, lo que permite al codificador descartar selectivamente valores de alta frecuencia espacial (que el ojo humano no puede apreciar fácilmente).

Con una cuantificación inteligente puede conseguirse que casi todas las altas frecuencias espaciales tengan un valor asociado muy bajo que, con el redondeo, será nulo, lo cual será de gran ayuda en la obtención de una codificación eficiente. La DCT ha demostrado poseer muchas ventajas desde el punto de vista de la compresión de datos para MPEG. La principal es el uso de los coeficientes DCT para intra *coding*. Estos coeficientes están casi completamente decorrelados, es decir, son independientes unos de otros, lo cual permite crear un algoritmo relativamente simple para codificarlos. La decorrelación es de gran interés teórico y práctico para la construcción de un modelo decodificación simple. Los coeficientes DCT cuantificados han de ser codificados con las menores pérdidas posibles para la tasa de bit requerida. De este modo, el decodificador podrá reconstruir valores más próximos a los originales [47].

3. TÉCNICAS DE ANÁLISIS FORENSE EN VÍDEOS DIGITALES

En este capítulo se describen dos temáticas para cubrir el análisis forense de un vídeo digital. En primer lugar, se estudian las principales técnicas de extracción de fotogramas claves que existen en la literatura, haciendo mención especial en aquellas que se basan en el análisis del histograma. En segundo lugar, se detallan las principales técnicas de análisis forense de vídeos digitales haciendo énfasis en las técnicas de identificación de la fuente del vídeo, ya que es la rama del análisis forense en la que se centra este trabajo.

Las técnicas de análisis forense de vídeos plantean aún muchos temas por investigar, debido a la amplia gama de posibles alteraciones que se pueden aplicar sobre éstos. Además, el análisis forense de vídeos ha demostrado ser más difícil con respecto al análisis de imágenes puesto que los datos que contienen los vídeos tienen formatos de compresión más altos, que pueden comprometer las "huellas" existentes haciendo más complicado recuperar el procesamiento de un vídeo desde su origen.

Hay que tener en cuenta que el vídeo como ente unitario no se puede clasificar en un tipo concreto de fuente, es decir, lo que identifica su fuente son las unidades mínimas de las que está compuesto un vídeo, denominadas fotogramas (*frames*). El primer paso para averiguar la fuente del vídeo digital consiste en determinar los fotogramas más representativos del vídeo, puesto que cada uno de ellos posee una "huella digital" que sirve para identificar dicha fuente.

3.1. Análisis de Contenido del Vídeo

El análisis de secuencias de vídeo orientado a la extracción automática de información referente a su contenido es un proceso genéricamente conocido

como *video parsing*. El primer paso de dicho análisis consiste habitualmente en llevar a cabo una segmentación temporal de la secuencia de vídeo, es decir, una subdivisión o estructuración en unidades homogéneas desde algún punto de vista, ya sea objetivo (luminosidad media, distribución de color, movimiento de la cámara, etc.) o subjetivo (coherencia de contenido) [9]. La segmentación identifica los límites de las capturas (*shot*) de un vídeo. El siguiente nivel de descomposición es abstraer los fotogramas claves (*key frames)* y, por último, características visuales como el color o la textura se utilizan para representar el contenido de los fotogramas más representativos. En la Figura 3.1 se presentan los tres procesos que captan diferentes niveles de información del contenido de un vídeo [5].

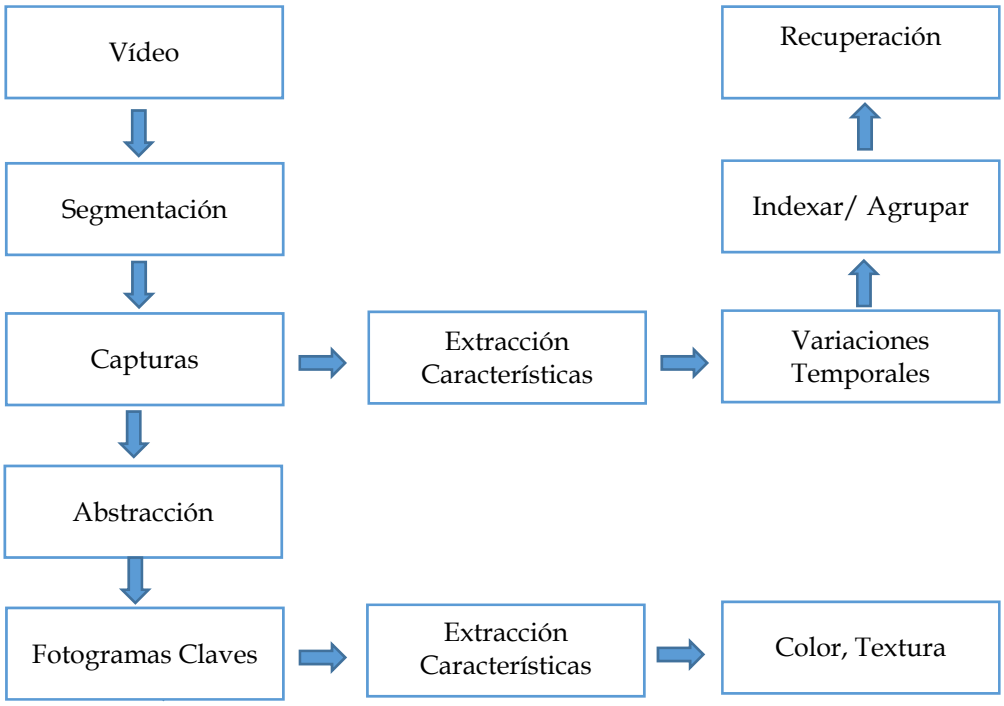

Fig. 3.1. Niveles de información del contenido del vídeo.

En [9] el vídeo se divide desde un punto de vista basado en la estructuración temporal de la secuencia siguiendo la terminología utilizada en el ámbito de la producción y la post-producción de vídeo, pudiendo subdividirse en planos,

capturas y escenas.

- Un plano es una secuencia de fotogramas caracterizada por la situación relativa entre el escenario que se desea filmar y la cámara (plano general, plano medio, primer plano, etc.)

- Una captura es un segmento ininterrumpido de una secuencia de vídeo, es decir, una secuencia de fotogramas consecutivos resultantes de una sola y continuada operación de grabación de la cámara. Una captura puede estar compuesta por varios planos. La transición entre dos capturas puede ser abrupta o gradual. La primera aparece como resultado de un efecto de corte, en el que un fotograma f_n pertenece a una captura y el fotograma siguiente f_{n+1} pertenece a la siguiente captura provocando una discontinuidad de la secuencia. En la transición gradual se ven involucrados varios fotogramas, de modo que si un fotograma f_n pertenece a una captura, el fotograma f_{n+L} pertenece a la siguiente captura, y los L-1 fotogramas intermedios representan una transformación gradual o progresiva del fotograma f_n en el f_{n+L}.

- Una escena es una sucesión de capturas adyacentes, todas ellas relacionadas con el mismo objeto o grupo de objetos, representando las mínimas subsecuencias con significado completo.

3.2. Técnicas de Extracción de Fotogramas Claves

Numerosas investigaciones relacionadas con la extracción de los fotogramas representativos se centran en las técnicas de recuperación de contenido o indexación de vídeos. Las primitivas desarrolladas para recuperación por contenido de imágenes se basan en las características de color, forma o combinaciones de ambas en el dominio transformado tras aplicar la Transformada de Wavelet. El presente trabajo se centra en las primitivas

basadas en el color para realizar la técnica de extracción de fotogramas representativos.

Para averiguar el contenido de un vídeo la cuestión que debe abordarse es la eliminación de información redundante, que reducirá significativamente la cantidad de información a procesar.

Los fotogramas claves (*key frames)* son las imágenes fijas que mejor representan el contenido de la secuencia de un vídeo de una manera abstracta. El reto en la extracción de fotogramas claves es que deben mantener el contenido y la naturaleza dinámica del vídeo, mientras se elimina toda la redundancia basada en el contenido.

3.2.1. Histogramas de Color

El interés que los histogramas de color despiertan en la comunidad científica se refleja en [41] donde realizan un exhaustivo estudio sobre los histogramas que abarca desde su cálculo hasta su comparación, pasando por un análisis sobre la resolución óptima para su utilización. Su invariancia ante transformaciones geométricas los hacen aptos para su aplicación también en dominios más concretos.

[42] usa la intersección de histogramas para comparar imágenes de logotipos.

[43] propone la utilización de histogramas compartidos para determinar la semejanza entre imágenes. Un histograma compartido se define como un histograma multidimensional en el que cada entrada cuenta el número de píxeles en la imagen que describen un conjunto particular de características como color, densidad de los bordes, textura o gradiente de magnitud.

Por su parte, en [44] los histogramas calculados sobre toda la imagen son inadecuados para representar las características locales del contenido de las

imágenes, por lo que utiliza conjuntos de histogramas asumiendo de forma implícita que cada región de la imagen incluye o bien un color dominante o bien un pequeño número de características de color.

[45] plantea un sistema de recuperación por contenido basado en el color de espacio CIELUV. Este espacio se basa en la teoría de colores complementarios de la visión humana, que aproxima mejor las diferencias de color según son percibidas por los seres humanos.

En el empeño por aproximarse al sistema humano de percepción de colores se han realizado trabajos en diversos espacios de color.

[46] calcula, en los espacios de color HSV (*Hue Saturation Value*) y CIELAB (*Lab Color Space*), vectores de coherencia de color, histogramas de color, histogramas de transición de colores, momentos de inercia de la distribución de color y composiciones de regiones de color identificadas mediante una cuantificación a 11 colores.

Hasta aquí se ha considerado el uso de histogramas de color pero en la literatura existen otras clasificaciones que se muestran a continuación:

Según [18] los métodos de extracción de fotogramas representativos son los que se basan en analizar las capturas de los vídeos, el contenido de los mismos o el movimiento y las denominadas técnicas de agrupación o *clustering*.

3.2.2. Métodos Basados en las Capturas de Vídeos

Estos métodos son los más fáciles y rápidos para extraer los fotogramas representativos.

[5] selecciona siempre el primer fotograma de cada una de las capturas como fotograma clave. Si es necesario seleccionar más, se eligen dependiendo de otros criterios como puede ser el color o el movimiento. Tras seleccionar el primer

fotograma clave, los siguientes fotogramas de la captura se comparan con el último fotograma clave en base a sus similitudes definidas por el histograma de color. Si se produce un cambio significativo de contenido entre el fotograma actual y el último fotograma clave, el fotograma actual se selecciona como nuevo fotograma clave. Este proceso se repite hasta llegar al último fotograma de la captura. Cuando se trata de vídeos comprimidos la similitud de un fotograma se calcula en base a los coeficientes DCT asociados con todos los macro-bloques en los fotogramas.

[6] propone un algoritmo para calcular de forma eficiente la similitud de capturas. Ésta se mide en términos de la intersección del flujo de formas de imágenes. La idea consiste en relacionarlo con los contenidos de las capturas de un vídeo de una forma similar a la percepción humana. En particular, la distancia entre dos capturas se define como la distancia mínima entre los fotogramas que la constituyen. La medida de similitud se basa en la diferencia normalizada de las proyecciones de luminancia de capturas. Posteriormente, se agrupan jerárquicamente los pares más similares. El dendograma resultante proporciona una representación visual de la jerárquica del clúster. Un salto repentino en el nivel proximidad del dendograma se utiliza para seleccionar automáticamente la partición adecuada de las capturas de vídeo.

Estos métodos tienen la ventaja de ser simples y poseen baja complejidad de cálculo, consiguiendo que los fotogramas extraídos tengan un significado común. Estos métodos son buenos para escenas con poco movimiento o fijas. Sin embargo, presentan el inconveniente de considerar siempre que el número de fotogramas clave se limitan a un número fijo. Además tampoco describen el contenido de movimiento de una captura de vídeo de forma eficiente.

3.2.3. Métodos Basados en el Contenido del Vídeo

Estos métodos extraen los fotogramas claves basados en los cambios de color,

textura u otra información visual de cada fotograma.

[7] divide el vídeo en cuatro capas: vídeo, episodios, capturas y fotogramas claves, siendo un episodio o conjunto de capturas la unidad semántica que describe un acto o una historia y seleccionando como capturas significativas las que aparecen repetidamente o las que son de larga duración. En base a estos dos criterios se propone la siguiente técnica: la primera captura de un episodio se incluye en una clase y a continuación se calcula la distancia entre ella y el resto de capturas, utilizándose un umbral denominado umbral de distancia DT (*Distance Threshold*). Si la distancia calculada es menor al umbral, entonces ambas capturas se fusionan. De lo contrario se clasifica en otra clase. Posteriormente, utiliza un algoritmo de clasificación basado en lógica difusa para realizar la identificación. La tasa de acierto es del 76,4% y la tasa de predicción del 95%.

[8] describe un sencillo algoritmo válido para tiempo real que unifica los algoritmos de segmentación temporal y extracción de fotogramas claves que se corresponden con el cambio de contenidos en la captura de vídeo, analizando los principales estándares de compresión de vídeo MPEG1-2 y MPEG-4. El estándar de codificación MPEG-2 comprime el vídeo dividiendo cada fotograma en bloques de tamaño fijo de 16x16 denominados *macrobloques* (MB). Cada MB contiene información sobre el tipo de su predicción temporal y sus correspondientes vectores utilizados para la compensación de movimiento. El carácter de la predicción de cada macrobloque se define en una variable llamada *MBType*. Dado que la secuencia MPEG tiene una alta redundancia temporal dentro de una captura, una referencia continua entre los fotogramas estará presente si no se producen cambios significativos de escena. La cantidad de referencias entre los fotogramas y sus cambios temporales se puede utilizar para definir una métrica. Los resultados experimentales muestran una alta robustez, tanto en la segmentación de vídeo temporal como en la extracción del fotograma clave representativo de una escena determinada. Este método es

muy simple y puede seleccionar fotogramas clave que se corresponden con el cambio de contenidos en la captura de vídeo. Pero tiene la desventaja de que los fotogramas extraídos no son siempre los más representativos y no puede indicar los cambios de información del movimiento cuantitativamente.

Estos dos últimos métodos se basan en los cambios de color, de textura o de cualquier información visual que contiene el vídeo. Cuando esta información cambia de forma significativa, el fotograma que se está procesando se elige como *key frame*. El principal inconveniente de estos métodos es que los fotogramas que extraen no son siempre, como se ha indicado anteriormente, los más representativos, existiendo cierta redundancia entre los mismos. Esto ocurre porque estos algoritmos emplean un único descriptor (color del histograma, textura, etc.) para capturar el contenido de un fotograma, no siendo suficiente un solo descriptor debido a la gran variedad de contenido visual que hay en un vídeo.

Así, un estudio reciente [11] concatena varios descriptores de la imagen proponiendo un algoritmo de agrupación de múltiples vistas ponderadas basado en CMM (*Convex Mixture Models*) que calcula de forma automática el peso de cada descriptor, reflejándose la importancia de cada uno en una secuencia de vídeo específica. Tras calcular los pesos, se crea una matriz de similitud que se construye mediante la suma ponderada de cada descriptor. Por último, se aplica un algoritmo de agrupamiento espectral utilizando la matriz de similitud para reunir los fotogramas de una captura dada, extrayéndose los fotogramas más representativos.

3.2.4. Métodos basados en la Segmentación

Existen otros algoritmos de extracción de fotogramas que se basan en la segmentación, detectando cambios significativos en términos de similitud de fotogramas sucesivos. Los más relevantes son los siguientes:

[12] selecciona los fotogramas más representativos de la siguiente forma: tras calcular el flujo óptico para cada fotograma, analiza una métrica simple del movimiento como una función del tiempo, seleccionando los mínimos. En la primera etapa utiliza el algoritmo de Horn y Schunck [10] para el flujo óptico y calcula la suma de las magnitudes de los componentes del flujo óptico para cada píxel como una métrica de movimiento *M(t)* para el fotograma *t*. En el segundo paso identifica los mínimos locales de *M(t)*.

[14] desarrolla dos algoritmos: uno de segmentación que extrae las capturas de vídeo basándose en los cambios abruptos que observa mediante un algoritmo que combina un umbral con la diferencia de fotogramas a partir de los histogramas y los descriptores de textura, y otro de selección de fotogramas claves mediante tres características visuales (histograma de color, histograma de detección del borde y estadísticas Wavelet). El primer algoritmo distingue dos tipos de capturas: *informativas* o tipo A y *no informativas* o tipo B. Los fotogramas claves sólo se extraen de las capturas de tipo A, porque en las de tipo B los fotogramas poseen imágenes de color uniforme y eso carece de sentido en términos de la información suministrada. El segundo algoritmo funciona bien en todo tipo de vídeos (ya sean comprimidos o sin comprimir). La ventaja que tiene es que el número de fotogramas claves que extrae es automático depende del vídeo y no es necesario que el usuario tenga que conocer el contenido del vídeo para ajustar dicho parámetro.

[15] extrae los fotogramas claves mediante un algoritmo de geometría computacional que divide la curva de contenido de una secuencia de vídeo en fotogramas claves que son equivalentes bajo cualquier tipo de contenido de descripción del vídeo. La idea es abordar el problema de resumir un vídeo desde diferentes puntos de vista, proponiendo tener en cuenta los siguientes principios: Distancia (*Iso-Distance*), Error (*Iso-Error*), Distorsión (*Iso-Distortion*).

Utiliza el principio de la distorsión para seleccionar los fotogramas más

representativos. El principio de distancia define que las distancias del contenido entre dos fotogramas claves deben ser iguales. Se pueden definir varios tipos de distancia (Euclídea, Manhattan, Chi-cuadrado, etc., dependiendo del contenido de los descriptores. el principio de error define que las distancias de error de contenido entre dos segmentos de línea sucesivos de la curva poligonal, producido por la interpolación lineal de los sucesivos fotogramas claves, serán iguales. En este algoritmo el número de fotogramas claves debe ser definido a priori por el usuario.

[16] divide el vídeo en segmentos o capturas (*shot*). Para cada uno de ellos elige cada fotograma y el siguiente, calculando el histograma de color de los mismos y comparando la similitud de ambos mediante la distancia Euclídea.

3.2.5. Métodos basados en Técnicas de Agrupamiento

Otros algoritmos de extracción de fotogramas clave se basan en la técnica de agrupamiento o *clustering*.

[5] establece que hay dos métodos principales de agrupamiento: organizar los datos en clústeres separados y la agrupación jerárquica en forma de árbol.

Asimismo, desarrolla una técnica de agrupación jerárquica en forma de árbol con un enfoque muy flexible de forma que distintos conjuntos de características (medidas de similitud o algoritmos de agrupamiento iterativos) se pueden aplicar a diferentes niveles. Los autores implementan dos algoritmos de agrupamiento jerárquico: el método iterativo *k-means* y el mapa organizativo o mapa de Kohonen SOM (*Self-Organizing Map*). Tras realizar pruebas con un conjunto de 700 imágenes que se clasificaron en 20 clases, el algoritmo *K-means* obtiene unos resultados del 79% de acierto y el algoritmo SOM del 84,1%.

[17] propone un método de agrupamiento no supervisado que elimina la redundancia del contenido del vídeo. El algoritmo se divide en tres etapas: en

primer lugar, todos los fotogramas de una secuencia de vídeo se dividen en clústeres donde el número de clases a priori no se conoce. En segundo lugar, el sistema trata de buscar las combinaciones óptimas de las clases obtenidas aplicando la técnica de agrupamiento jerárquico. Por último, cada una de las clases obtenidas se representa por una característica del fotograma (color, textura, forma, o bien una combinación de las anteriores). La experimentación realizada (un único vídeo de una película con 66 capturas) obtiene una tasa de acierto del 87%.

[18] presenta un algoritmo de extracción de fotogramas claves basado en la correlación temporal (*inter-frame*) que existe entre fotogramas consecutivos. El algoritmo considera que la información más importante se encuentra en el centro del fotograma y la menos relevante en las esquinas, dividiendo cada fotograma en nueve cuadrículas, cada una con pesos diferentes, asignando mayor peso a las que se encuentran en el centro.

El principal inconveniente de estos métodos es que tienen alta complejidad y prestan poca atención a los cambios de los contenidos presentados por la dinámica acumulativa.

Según [13] las principales técnicas de análisis forense relacionadas con vídeos que existen actualmente se dividen en identificación de la fuente de adquisición, detección ilegal de reproducción de vídeos y compresión de vídeos.

3.3. Técnicas de Identificación de la Fuente de Adquisición

Según [4] las tareas de análisis forense de imágenes y vídeos digitales se pueden dividir en las siguientes categorías:

- **Verificación de integridad o detección de falsificaciones**: Busca descubrir procedimientos maliciosos que se hayan aplicado a las imágenes

y vídeos como, por ejemplo, recorte o adición de objetos.

- **Recuperación de la historia de procesamiento**: Recupera la cadena de procesamiento que ha sido aplicado a una imagen o vídeo de una manera no maliciosa como, por ejemplo, recortes, filtrados, contrastes, etc.

- **Clasificación basada en la fuente**: Tiene como objetivo clasificar las imágenes y vídeos de acuerdo a su origen en cámaras digitales o escáneres.

- **Agrupación por dispositivos fuente**: Dado un grupo de imágenes o vídeos se buscan los grupos de vídeos que fueron obtenidas utilizando la misma cámara.

- **Identificación de la fuente**: Busca determinar el dispositivo que generó una imagen o vídeo determinado.

El análisis de la fuente de adquisición de vídeos es uno de los primeros problemas que han surgido en las técnicas de análisis forense. El tema de identificación de la fuente ha sido abordado desde varios enfoques: por un lado, el tipo de dispositivo que genera el contenido (cámara, escáner), y por otro, el modelo del dispositivo que genera el contenido. El objetivo básico es comprender la etapa inicial de generación de contenido multimedia. Así, una vez que se ha extraído la información más representativa de un vídeo, el siguiente paso es obtener las "huellas digitales" de los fotogramas obtenidos, para poder determinar la fuente que generó el vídeo.

Desgraciadamente, no hay mucha literatura relativa a la fuente de adquisición de vídeos.

Uno de los primeros trabajos en los que se utilizó la huella de una videocámara fue [19] que señala que el ruido térmico *dark current* de los sensores CCD se debe a la propia energía térmica del chip de silicio ya que éste

genera electrones (termoelectrones) sin que incida la luz sobre él. Estos termoelectrones son indistinguibles de los fotoelectrones producidos al incidir la luz en el sensor. Así, propone utilizar píxeles defectuosos y la propiedad *dark current* de los chips CCD para identificar la videocámara. Este enfoque es limitado porque el ruido térmico sólo puede extraerse en los fotogramas de color negro y la propiedad *dark current* es una señal débil que no sobrevive a la compresión del vídeo.

El tiempo ha demostrado que la técnica desarrollada en [20] que identifica sensores de imágenes basados en el ruido de respuesta no uniforme PRNU (*Photo-Response Non-Uniformity Noise*) proporciona una "huella digital" mucho más robusta y fiable.

Muchos de los trabajos posteriores se basan en este tipo de característica.

El patrón de ruido PRNU se produce por la variación de sensibilidad de los píxeles individuales a la luz, debido a la falta de homogeneidad e impurezas en los chips de silicio, y a las imperfecciones introducidas en el proceso de fabricación del sensor. En el caso de los vídeos puede parecer que la estimación del patrón PRNU de una cámara de vídeo a partir de una secuencia de vídeo es más sencilla que para el caso de las imágenes fijas, debido a la gran cantidad de fotogramas disponibles que hay en un vídeo. Sin embargo, esto no es cierto por dos razones principales; en primer lugar, la resolución espacial de vídeos es mucho menor que la de las imágenes fijas y, en segundo lugar, los fotogramas de vídeos generalmente tienen ratios de compresión más elevados que las imágenes comprimidas en formato JPEG.

[21] utiliza el patrón de ruido PRNU para verificar si dos videoclips de un vídeo proceden de la misma videocámara digital. El procedimiento es como sigue: en primer lugar, el patrón PRNU se estima a partir de los dos clips del vídeo utilizando el estimador de máxima verosimilitud. A continuación, los PRNUs se filtran para eliminar los defectos de formación de bloques debido a la

compresión con pérdida. Finalmente, se procesan utilizando la correlación cruzada normalizada. El pico de coeficiente de correlación de energía se utiliza para establecer el origen común de ambos PRNUs. Los experimentos se realizaron con 25 cámaras de vídeo y muestran que con sólo 40 segundos de vídeo es suficiente para tener resultados fiables. Si se disminuye la calidad del vídeo (se aumenta el ratio de compresión) y se disminuye la resolución espacial, es necesario aumentar el tiempo del videoclip para poder obtener resultados fiables. Con vídeos en formato LP de internet y una resolución de 264×352 y 150 kb/seg se obtienen buenos resultados para videoclips con una duración de 10 minutos.

En resumen, el patrón del ruido del sensor (SPN) extraído de imágenes digitales como huellas digitales del dispositivo ha demostrado ser una técnica eficaz para la identificación de dispositivos digitales.

Sin embargo, [50] observó que la extracción de ruido del sensor a partir de una sola imagen podría producir un patrón contaminado por los detalles finos y la estructura de la escena representada. Para mejorar la estimación de la huella digital propone asignar factores de ponderación que sean inversamente proporcional a la magnitud de las componentes de la señal. Para lidiar con este problema presenta un nuevo enfoque para atenuar la influencia del detalle de las escenas en el ruido del sensor mejorando la tasa de acierto. En la experimentación realizada (9 cámaras y 320 imágenes de cada una, con escenas al aire libre e interiores) se mejora la tasa de acierto un 18% con el tamaño de la imagen más pequeña (128x128) y sólo un 1% en el caso de la imagen más grande (1536x2048).

[4] realiza una comparación de los diferentes filtros que existen para la eliminación del ruido de las imágenes. Los filtros que usan la Transformada Wavelet dan los mejores resultados debido a que el ruido residual que se obtiene con este filtro contiene la menor cantidad de rasgos de la escena.

Generalmente, las áreas alrededor de los bordes se interpretan mal cuando se utilizan únicamente filtros de eliminación de ruido menos robustos, tales como el filtro de Wiener o el filtro de mediana. La experimentación realizada (14 cámaras digitales de dispositivos móviles de 7 fabricantes diferentes e imágenes con escenas reales en una dimensión de 1024x1024) que combinó el uso del patrón de ruido del sensor con la Transformada Wavelet alcanzó una tasa de éxito promedio del 87,21% para la identificación de fuente.

[22] propone un método de identificación utilizando imágenes fijas de vídeos. En las pruebas realizadas utiliza cuatro modelos diferentes de cámaras y un clasificador SVM, obteniendo en un primer experimento aplicado en el dominio del espacio con los valores de luminancia, un 82,6% de precisión. En un segundo experimento usando el mismo conjunto de vídeos, capturando el valor de luminancia, el promedio de clasificación fue del 100%. En un tercer experimento, donde se utilizaron un conjunto de vídeos con mayores cambios en las escenas, se obtuvo un 97,2% de acierto.

[23] propone un algoritmo con la información del vector de movimiento en el flujo codificado. En los experimentos realizados utiliza 100 secuencias de vídeo (20 de ellas procedentes de "Vídeo Quality Experts Group" y 80 de DVDs). Todos los vídeos fueron codificados por diferente software de edición de vídeo conocidos. El resultado fue un 74,63% de precisión en la identificación del software que se utilizó en la codificación.

[48] propone una técnica de identificación de la cámara de vídeo basada en las características de probabilidad condicional (CP). Este tipo de características fueron propuestas inicialmente [49] para propósitos de estegoanálisis. Las características CP se obtiene a partir de los valores absolutos de la matriz de coeficientes DCT. La experimentación realizada obtiene una tasa de acierto del 98,6%, 97,8% y 92.5% en la clasificación de 2, 3 y 4 teléfonos de marca iPhone respectivamente, con un recorte de imagen de 800 por 600.

3.4. Herramientas Forenses para Compresión de Vídeos

El contenido de un vídeo suele estar disponible en un formato de compresión con pérdidas. La compresión con pérdidas deja "huellas digitales" que pueden ser detectadas por el analista forense. El estudio de herramientas forenses eficaces relacionadas con vídeos comprimidos es una tarea difícil puesto que la codificación de las operaciones tiene el efecto potencial de borrar las huellas dejadas por las manipulaciones anteriores.

Por otro lado, el *códec* adoptado para comprimir una secuencia de vídeo representa un elemento connotativo distintivo. Por tanto, si el códec es detectado, puede ser útil para la identificación del dispositivo de adquisición y para revelar posibles manipulaciones.

La mayoría de las arquitecturas de codificación de vídeo se han creado sobre las herramientas de codificación de imágenes. El estándar JPEG es la técnica de codificación ampliamente adoptada para las imágenes fijas y muchos de sus principios se reutilizan para la compresión de señales de vídeo [28]. Las arquitecturas de codificación de vídeo son más complejas que las adoptadas para imágenes fijas. La mayoría de los estándares de codificación de vídeos más utilizados (MPEG-x o la familia H.26x) heredan parte del proceso de codificación de JPEG. Sin embargo, la arquitectura de MPEG es más compleja porque tiene en cuenta la codificación espacial y temporal, la interpolación de imágenes, etc. En las arquitecturas de codificación de imagen y de vídeo, la elección de los parámetros de codificación viene impulsado por las herramientas que dependen de la implementación específica del *códec* y de las características de la señal codificada.

En la compresión JPEG los parámetros de codificación definidos por el usuario se limitan a la selección de las matrices de cuantificación, que se adoptan para mejorar la eficacia de codificación basada en el análisis psicovisual de la percepción humana. Por el contrario, en el caso de compresión

de vídeo, el número de parámetros de codificación que se pueden ajustar es significativamente más amplio. Como consecuencia de ello, el analista forense debe tener en cuenta un mayor número de grados de libertad cuando se detecta la identidad del *códec*. Esta pieza de información podría permitir la identificación de las implementaciones de otros proveedores que dependen de los *códecs* de vídeo. En la literatura los métodos que estiman diferentes parámetros de codificación y elementos de sintaxis del códec adoptado, se agrupan en tres categorías principales, que se describen a continuación.

3.4.1. Técnicas de Doble Compresión de Vídeos

Cada vez que una secuencia de vídeo que previamente ha sido comprimida se edita (se recorta, se realza el contraste, brillo, etc.), se tiene que volver a comprimir. Esta es una situación típica que se produce, por ejemplo, cuando el contenido de un vídeo se descarga desde sitios web de intercambio de vídeos. Por esta razón se estudian las huellas que dejan la doble compresión de un vídeo. Las soluciones propuestas hasta ahora en la literatura se centran principalmente en el estándar de codificación de vídeo MPEG y explotan las mismas ideas usadas originalmente para la doble compresión del estándar JPEG.

[29] muestra cómo la técnica de compresión doble presenta picos característicos en el histograma, que alteran las estadísticas originales y asumen diferentes configuraciones de acuerdo a la relación entre los tamaños de las etapas del proceso de cuantificación de dos operaciones de compresión consecutivas. En el citado trabajo se destaca cómo los picos pueden ser más o menos evidentes en función de la relación que existe entre los dos tamaños de las etapas del proceso de cuantificación y se propone una estrategia para identificar la técnica de compresión doble. Su enfoque se basa en recortar la imagen reconstruida (con el fin de alterar la estructura de los bloques JPEG) y comprimirla con un conjunto de tablas de cuantificación candidatas. La imagen

se comprime luego utilizando la segunda etapa y calculando el histograma de los coeficientes de la Transformada DCT. El método propuesto elige la tabla de cuantificación de tal manera que el histograma resultante esté lo más cerca posible a la obtenida de la imagen reconstruida.

[30] aborda el problema de la estimación de la técnica de compresión doble de vídeo codificado en formato MPEG considerando dos escenarios, dependiendo de si la estructura del grupo de imágenes utilizado en la primera compresión se conserva o no. En el primer caso, cada cuadro se re-codifica en un fotograma del mismo tipo, por lo que, las tramas I, B o P permanecen, respectivamente, como tramas I, B, o P. Cuando una trama I es recodificada a una velocidad de bits diferente, los coeficientes DCT están sujetos a dos niveles de cuantificación. Por lo tanto, los histogramas de los coeficientes DCT asumen una forma característica que se desvía de la distribución original. En particular, cuando el tamaño de las etapas de cuantificación disminuye desde la primera a la segunda compresión, algunos contenedores del histograma se dejan vacíos. Por el contrario, cuando aumenta el tamaño de las etapas, el histograma se ve afectado en una forma característica. Esta última situación se presenta típicamente en el caso de la eliminación de fotogramas o ataques de inserción.

[31] propone otro método para la detección de la técnica MPEG de doble compresión, inspirándose en el método propuesto en [32].

[33] detecta la técnica de compresión doble analizando analiza un modelo de distribución de probabilidad de los coeficientes DCT de un macro-bloque en un fotograma I. Con una técnica de *estimación-maximización* (EM), la distribución de probabilidad que se produciría si un macro-bloque fue doblemente codificado se puede estimar. A continuación, dicha distribución se compara con la distribución real de los coeficientes. A partir de esta comparación, se calcula la probabilidad de si un bloque ha sido doblemente comprimido. Estas soluciones se pueden ampliar para permitir la detección de la doble compresión de vídeo

incluso en un escenario real en la que se emplean diferentes *códecs* en cada etapa de compresión.

[34] presenta un método que identifica el *códec* utilizado en la primera etapa de compresión en el caso de compresión de vídeo doble. El algoritmo propuesto se basa en el supuesto de que la cuantificación es un operador idempotente, es decir, cada vez que un cuantificador se aplica a un valor que ya ha sido previamente cuantificado y reconstruido por el mismo cuantificador, el valor de salida está altamente correlacionado con el valor de entrada. Cada vez que la secuencia de salida presenta la correlación más alta con el vídeo de entrada se infiere que la configuración de codificación adoptada corresponde a la primera compresión.

Aunque la detección de la compresión doble de las imágenes es un tema ampliamente investigado, la doble compresión del vídeo todavía resulta ser un problema abierto, debido a la complejidad y diversidad de arquitecturas de codificación de vídeo. Siempre que dos *códecs* diferentes están involucrados con parámetros similares, la detección de la compresión doble del vídeo se hace significativamente más difícil [34]. Por otra parte, la compresión múltiple es un tema actual y poco explorado a pesar del hecho de que el contenido multimedia disponible en Internet a menudo se ha codificado más de dos veces [36].

3.4.2. Identificación de las Huellas Digitales en una Red

La transmisión de un vídeo a través de un canal ruidoso deja huellas digitales características en el contenido de vídeo que ha sido reconstruido. Incluso las pérdidas de paquetes y errores podrían afectar al flujo de bits recibidos. Como consecuencia de ello, algunos de los datos codificados se perderán.

La técnica de corrección de errores está diseñada para hacerse cargo de esto, tratando de recuperar la información correcta y mitigar la distorsión que induce el canal. Sin embargo, esta operación introduce algunos elementos en el vídeo

reconstruido, que pueden ser detectados para deducir el patrón de pérdida subyacente (o error). El patrón de pérdida específico permite la identificación de las características del canal que se trabaja durante la transmisión del vídeo codificado. Es decir, es posible analizar la probabilidad de pérdida (error), la explosividad y otras estadísticas relacionadas con la distribución de los errores con el fin de identificar, por ejemplo, el protocolo o infraestructura de transmisión.

Los enfoques dirigidos a la identificación de las huellas de red están destinados a no referenciar la monitorización de la calidad, es decir, la estimación de la calidad de una secuencia de vídeo se realiza sin tener acceso a la fuente original. Estas soluciones se diseñan para proporcionar a los dispositivos de red y terminales herramientas eficaces que midan la experiencia de calidad que se ofrece al usuario final. Los enfoques propuestos se pueden dividir en dos grupos principales.

La primera clase de algoritmos de identificación de huella de la red tiene en cuenta las estadísticas de transmisión para calcular la distorsión del canal en la secuencia reconstruida.

[35] presenta un algoritmo basado en varias métricas de evaluación de calidad para estimar el deterioro de pérdida de paquetes en el vídeo reconstruido. Sin embargo, la solución propuesta adopta métricas de calidad de referencia completa que requieren la disponibilidad de la secuencia de vídeo original sin comprimir.

Otro enfoque diferente se presenta en [37], donde la distorsión del canal que afecta a la secuencia de vídeo recibida se calcula de acuerdo a tres estrategias diferentes. Una primera solución calcula la calidad final del vídeo a partir de las estadísticas de la red; una segunda solución utiliza las estadísticas de pérdida de paquetes y evalúa el impacto espacial y temporal de las pérdidas en la secuencia final; la tercera evalúa los efectos de la propagación de errores en la

secuencia. Estas soluciones se dirigen a sistemas de control utilizados por los proveedores de servicios de red, que deben controlar la calidad de las secuencias de vídeo finales sin tener acceso a la señal original.

Otra estrategia de estimación se basa en la relación señal ruido pico o PSNR (*Peak Signal to Noise Rate*) que se propone en [38]. La solución propuesta evalúa los efectos de ocultación de los errores temporal y espacial sin tener acceso a la secuencia del vídeo original y en los valores de salida presenta una buena correlación con las puntuaciones MOS (*Mean Opinion Score*). En realidad, es posible considerar este enfoque como una solución híbrida, en que se aprovecha tanto el flujo de los valores de los pixeles reconstruidos como los píxeles recibidos.

Una segunda clase de algoritmos asume que la secuencia de vídeo transmitida ha sido decodificada y que sólo los píxeles reconstruidos están disponibles. Esta situación se representa en todos los casos en los que el analista de vídeo no tiene acceso al flujo de bits.

La solución propuesta en [39] se basa en las métricas propuestas en [38], pero la estimación de calidad no referenciada se lleva a cabo sin tener en cuenta la disponibilidad del flujo de bits. Por lo tanto, la solución propuesta sólo procesa los valores de los píxeles, identificando qué porción de vídeo se ha perdido y produciendo como salida un valor de calidad que representa una buena correlación con el valor de las extensiones MSE (*Media Source Extensions*). El método supone que los trozos se corresponden con las filas de los macro bloques. Sin embargo, en los esquemas de vídeos más actuales la codificación se realiza con trozos más flexibles. [40] extiende este enfoque donde tiene en cuenta esta flexibilidad.

4. CONTRIBUCIÓN

El objetivo de este capítulo es presentar la contribución de este trabajo, a saber, una técnica de identificación de la fuente de adquisición de vídeos de dispositivos móviles que utiliza un algoritmo de extracción de fotogramas claves desarrollado íntegramente.

4.1. Consideraciones Generales

El algoritmo desarrollado de extracción de fotogramas claves presta especial atención a la naturaleza del vídeo, teniendo en cuenta que si aquellos tienen mayor variación de escena, el posterior proceso de clasificación será mejor. El algoritmo requiere además de un número determinado de fotogramas para el entrenamiento y la clasificación utilizando SVM. Una vez obtenidos los fotogramas, se realiza la extracción de las características que se obtienen del patrón de ruido del sensor y la Transformada Wavelet según el algoritmo especificado en [4] identificando la marca y el modelo del dispositivo móvil fuente de una imagen. Este último algoritmo describe el proceso de extracción de la huella del sensor de un fotograma de manera individual, la estimación del patrón del ruido del sensor cuando se cuenta con varias imágenes y la extracción de características requeridas para la identificación de la fuente.

Conviene reseñar que esta propuesta tiene como ventaja el hecho de no requerir acceso a la cámara del dispositivo móvil fuente, siempre y cuando se cuenten con vídeos procedentes del mismo.

Tras un análisis realizado sobre las diferentes técnicas de identificación de la fuente se detectó que para el caso particular de los dispositivos móviles las técnicas más adecuadas son las basadas en el ruido del sensor por el tipo de sensor que usan y las basadas en las características wavelet por su efectividad.

A la vista de las consideraciones anteriores, este trabajo identifica la fuente de adquisición de los vídeos de dispositivos móviles en dos etapas: En primer lugar, extrayendo los fotogramas más representativos teniendo en cuenta los que presentan un cambio de escena significativo porque los datos de un vídeo contienen redundancia espacial, temporal y espectral. En segundo lugar, combinando dos técnicas para la identificación: las imperfecciones en el sensor y la Transformada Wavelet.

4.2. Especificación de la Técnica

La Figura 4.2 describe la técnica propuesta en este trabajo, que se divide en cuatro etapas principales:

- Un vídeo de entrada de 12-30 fotogramas por segundo (fps) que se divide en fotogramas individuales.

- La extracción de un conjunto de fotogramas claves, esto es, con cambios de escena significativos. Para iniciar este proceso se marca como fotograma clave el primer fotograma, calculando la diferencia de histogramas entre cada fotograma y el último fotograma clave considerado mediante la correlación del histograma de color. Si la diferencia de histogramas satisface un cierto umbral, entonces el fotograma actual se selecciona como fotograma clave. Este proceso se repite para todos los fotogramas que componen el vídeo, hasta que se extraen el conjunto de fotogramas claves. Esta etapa es crucial para el resto del proceso.

- En la siguiente etapa se extrae el patrón de ruido de sensor de cada fotograma perteneciente al conjunto de fotogramas claves. Las características se obtienen utilizando la Transformada de Wavelet.

- La etapa final utiliza una máquina de vectores de soporte (SVM) para la

clasificación. Esta técnica se utiliza en escenarios cerrados. En este tipo de escenarios los vídeos, cuyas fuentes de adquisición deben ser determinadas, han de pertenecer a un grupo de dispositivos conocidos de antemano. Por lo tanto, la identificación de la fuente de adquisición está limitada a un cierto número de dispositivos conocidos. El algoritmo utiliza dos conjuntos de datos diferentes: un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El conjunto de datos de entrenamiento contiene los vídeos seleccionados a partir de los modelos conocidos de dispositivos móviles. El conjunto de datos de prueba está formado por muestras tomadas al azar de los vídeos de los dispositivos móviles que van a ser identificados.

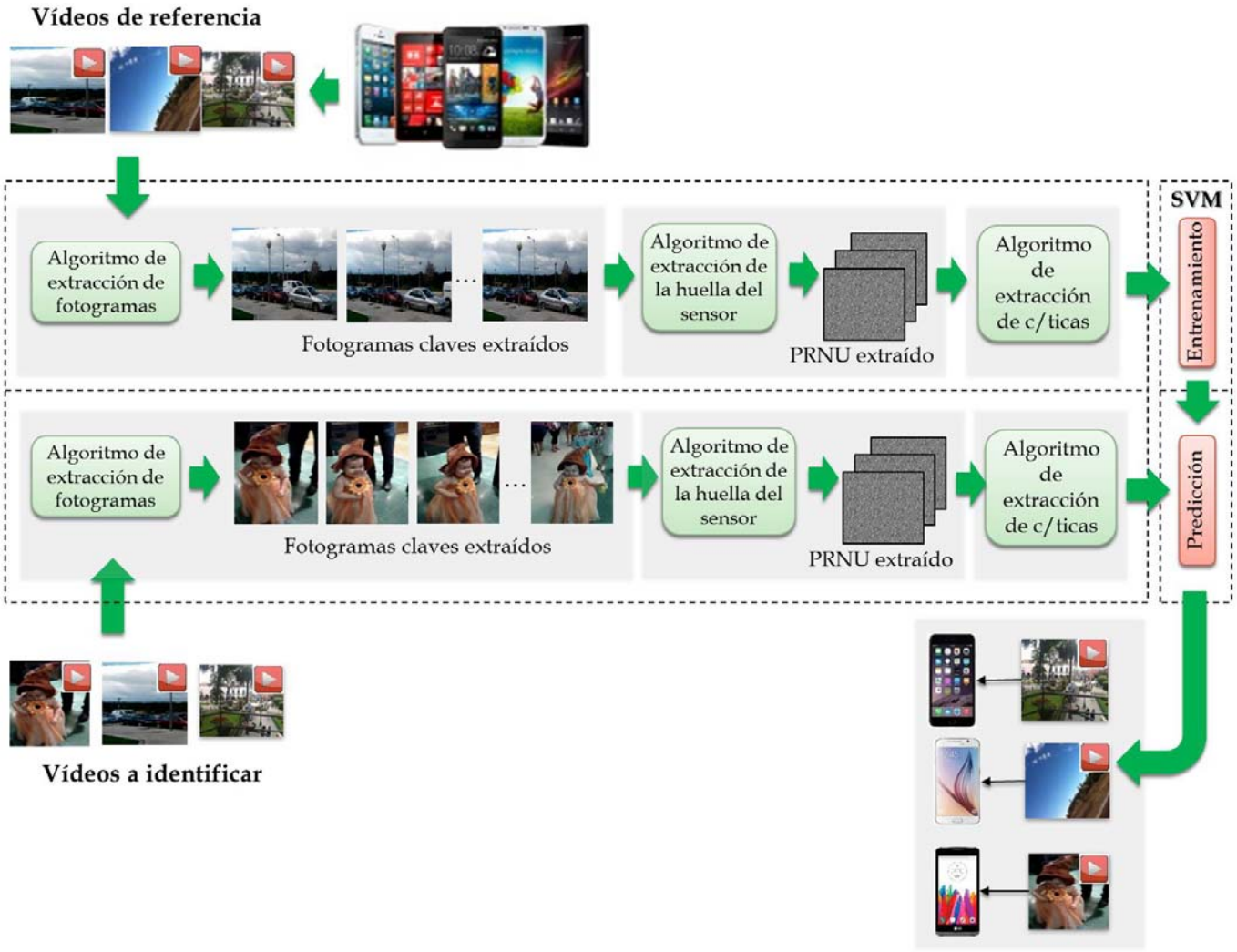

Fig. 4.2. Identificación de la fuente de adquisición de vídeos de dispositivos móviles.

A continuación, se especifica el algoritmo de extracción de fotogramas claves del vídeo así como los algoritmos que permiten extraer el patrón de ruido del sensor.

Los fotogramas seleccionados para la clasificación se obtienen mediante el Algoritmo 1, que utiliza los Algoritmos auxiliares 2, 3 y 4.

Algoritmo 1: Algoritmo de extracción de fotogramas

Input: *video: Vídeo al que extraer los fotogramas*
numFot: Número mínimo de fotogramas a extraer
umbralIni: Umbral inicial
incUmbral: Incremento del umbral
Result: *fotogramas: Vector de fotogramas extraídos del vídeo*

1. *hists ← extraerHistogramas(video)*
2. *umbral ← estimarUmbral (vídeo, hists, numFot, umbralIni, incUmbral)*
3. *fotogramas ← extraerFotogramas (vídeo, umbral, hists, true)*
4. *return fotogramas*

Algoritmo 2: Función extraerHistogramas

Input: *video: Vídeo*
Result: *hists: Vector de histogramas de todos los fotogramas del vídeo*

1. ***procedure*** *extraerHistogramas(vídeo)*
2. ***foreach*** *fotograma ϵ video* ***do***
3. *Hists ← Calcular y agregar histograma de fotograma*
4. ***end procedure***
5. *return hists*

Algoritmo 3: Función estimarUmbral

Input: *video: Vídeo al que extraer los fotogramas*
hists: Vector de histogramas
numFot: Número mínimo de fotogramas a extraer
umbralIni: Umbral inicial
incUmbral: Incremento del umbral
Result: *umbral: Umbral final calculado a partir del umbral inicial en la extracción de fotogramas*

1. ***procedure*** *estimarUmbral (hists, numFot, umbral, incrUmbral)*
2. *numFotCal ← extraerFotogramas(vídeo, umbral, hists, false)*
3. ***while*** *numFotCal ≤ numfotDeseados* ***do***
4. *umbral ← umbral + incrUmbral*
5. *numFotCal ← extraerFotogramas (video, umbral, hists, false)*
6. ***end procedure***
7. *return umbral*

Algoritmo 4: Función extracciónFotogramas

Input: *video: Vídeo al que extraer los fotogramas*
umbral: Umbral para la extracción de fotogramas
hists: Vector de histogramas
extraer: true extrae fotogramas, false no extrae

Result: *fotogramas: fotogramas extraídos del vídeo (solo si extraer=true)*
numFot: número de fotogramas extraídos del vídeo con el umbral indicado (solo si extraer=false)

1. ***procedure*** *extraerFotogramas (video, umbral, hists, extraer)*
2. *numFot ← 0*
3. ***Foreach*** *(fotograma, histograma) ϵ (video, hists)* ***do***
4. *histActual ← histograma*
5. ***if*** *histAnterior == nulo* ***then***
6. *histAnterior ← histActual*
   else
7. *correlación ← Calcular la correlación entre histActual e histAnterior usando*

$$correlación\ (H_1, H_2) = \frac{\sum_i H_1'(i)\, H_2'(i)}{\sqrt{\sum_i H_1'(i)^2\, H_2'(i)^2}}$$

8. ***if*** *correlación ≤ umbral* ***then***
9. *histAnterior ← histActual*
10. ***if*** *extraer = true* ***then***
11. *fotogramas ← Agregar fotograma*
    else
12. *numFot ← numFot +1*
13. ***end procedure***
14. *return fotogramas, numFot*

El Algoritmo 1 calcula los fotogramas claves contenidos en un vídeo, esto es, aquellos que presentan un cambio de escena significativo, y que posteriormente serán utilizados para la clasificación e identificación. Esto es así ya que, en [50] se demostró que el ruido del sensor extraído de una imagen puede estar severamente contaminado por los detalles de la escena, además de que los datos de un vídeo contienen redundancia temporal, espacial y espectral. El Algoritmo 1 requiere de 3 parámetros para su funcionamiento (el vídeo del cual serán extraídos los fotogramas, un umbral inicial que será la referencia para determinar cuando existe un cambio de escena y el valor del incremento para el umbral que se realizará en cada iteración) y de uno opcional (el número de fotogramas claves que se desea obtener). En otras palabras, el Algoritmo 1 admite dos modos de funcionamiento: un primer modo donde se puede especificar el número de fotogramas claves que se quieren extraer de cara al posterior proceso de clasificación e identificación, y un segundo modo que extrae los fotogramas claves del vídeo más relevantes.

Para comparar dos fotogramas es necesario extraer el histograma de cada uno (frecuencia de los valores de color) y mediante su correlación se puede obtener la similitud existente entre ellos. La correlación se calcula mediante la Fórmula 4.1.

$$correlación\ (H_1, H_2) = \frac{\sum_i H_1'(i)\, H_2'(i)}{\sqrt{\sum_i H_1'(i)^2\, H_2'(i)^2}} \tag{4.1}$$

donde:

$$H'_k(\mathrm{i}) = H_k(i) - \frac{1}{N}\left(\sum_j H_k(j)\right)$$

Se ha optado por el cálculo de la correlación para calcular la diferencia de histogramas de color de dos dimensiones, ya que es un vector aleatorio (variable aleatoria multidimensional) que proporciona mejores resultados que las otras medidas contempladas en la literatura [51], como la distribución de

probabilidad continua (*chi-cuadrado*) o la intersección o la distancia de *Bhattacharyya* (χ^2).

Siempre el primer fotograma del vídeo se elige como parte del conjunto de fotogramas claves seleccionados.

En el modo de funcionamiento 1 del Algoritmo 1, si la cantidad de cambios de escena obtenidos en base al umbral es menor de lo requerido, se repite el proceso de comparación, incrementando el umbral hasta que la cantidad de cambios de escena sea mayor o igual a los deseados.

Para poder determinar el umbral inicial se realizaron diversos experimentos, comprobándose que mediante la comparación de los histogramas de un vídeo, la menor correlación promedio fue de -0.27, presentando al menos 1 o 2 cambios de escena, definiéndose así el umbral inicial.

Para el valor del incremento se experimentó con diferentes valores. El valor de 0,001 fue el elegido, ya que demostró ser un valor ideal para llegar al número de fotogramas deseados en un menor tiempo. Estos incrementos se realizan porque si el umbral se encuentra más cercano al valor máximo de correlación directa (valor 1), se pueden encontrar más cambios de escena, y así extraer la cantidad de fotogramas definidos por el usuario para la clasificación e identificación.

Mediante el análisis de los trabajos de la literatura se llegó a la conclusión de que el patrón de ruido del sensor y la Transformada Wavelet ayudan a definir una huella, siendo métodos efectivos para la identificación de fuente.

Este trabajo extiende el uso del patrón de ruido del sensor y la Transformada Wavelet presentado en [4] y que obtiene vectores de 81 características para cada uno de los fotogramas obtenidos por el Algoritmo 1. Una vez obtenidas las características de cada uno de los fotogramas se utiliza una máquina SVM para la clasificación de cada fotograma independientemente.

Posteriormente, se ha de tomar un criterio que permita definir la fuente de adquisición de un vídeo como ente indivisible, en función de los resultados de la identificación de la fuente de los fotogramas que lo componen.

5. EXPERIMENTACIÓN

En la experimentación realizada se capturaron vídeos sin ninguna consideración en las características temporales o espaciales, pues debían representar casos reales. Como actualmente los teléfonos móviles presentan grandes mejoras en la calidad del vídeo, se consideró usar vídeos con calidad de 1080 p (vídeos de alta definición), es decir, con una resolución de 1920x1080 píxeles.

La Tabla 5.1 muestra las especificaciones básicas y los modelos de teléfonos móviles considerados para los experimentos.

Marca Modelo	FPS	Formato	Códec	Condiciones de Captura
Apple - iPhone 5 (M1)	24	.mov	**H.264**	Resolución: 1080p (1920x1080) Tipo de Escena: Cualquiera Orientación: Vertical Flash: Deshabilitado Luz: Natural Balance de blancos: Automático Zoom digital: 0 Tiempo de captura: 2 minutos
Nokia- 808 Pureview(M2)	30	.mp4	**MPEG-4**	
Samsung Galaxy S4 (M3)	30	.mp4	**MPEG-4**	
Wiko- Cink Slim (M4)	12	3gp	**MPEG-4**	
Zopo- ZP-980 (M5)	15	3gp	**MPEG-4**	

Tabla 5.1.Configuraciones utilizadas en cámaras digitales para dispositivos móviles.

En la Tabla 5.2 se resumen las condiciones experimentales utilizadas en los algoritmos propuestos.

Parámetro	Valor
Número de vídeos para entrenamiento por cámara	5
Número de vídeos para pruebas por cámara	5
Método de extracción	Histograma
Umbral inicial	-0,27
Incremento de umbral	0,001
Número de Fotogramas Deseados por vídeo	100

Tabla 5.2.Condiciones experimentales utilizadas en los algoritmos propuestos.

La clasificación se realizó utilizando SVM con kernel (*Gaussian*) *Radial Basis Function* (RBF). Se utilizó el paquete LibSVM que permite la clasificación de múltiples clases [53]. Es la opción mayoritariamente utilizada en los trabajos más recientes del estado del arte pues presenta buenos resultados.

Para el entrenamiento y pruebas se usó un kernel $\gamma = 2^3$ y un parámetro de coste C = 32768, que son los utilizados en [4].

El clasificador fue entrenado y probado con los vectores de características extraídos de los fotogramas.

La extracción de fotogramas se realizó en un microprocesador Intel Core i7-2670QM @ 2.20GHz y 13GB de RAM.

Para la extracción de 100 fotogramas el tiempo de ejecución varía conforme a los fotogramas por segundo que presente el vídeo. Por ejemplo, para un vídeo con una duración de 2 minutos y 30 fps, el tiempo promedio es de 125 segundos, mientras que para un vídeo con la misma duración y con 15 fps el tiempo promedio es de 70 segundos.

5.1. Experimento 1

Se realizaron 5 experimentos en los que fueron utilizados los 5 dispositivos móviles de la Tabla 5.1, empleando los parámetros de la Tabla 5.2 para cada uno de ellos. En cada experimento se utilizó un tamaño diferente en el recorte de los fotogramas seleccionados (1024x768, 800x600, 640x480, 320x240, 128x128). Estos tamaños fueron capturados ya que son resoluciones estándar, excepto el tamaño 128x128. Con respecto a los parámetros de configuración definidos para el algoritmo de extracción de las características definido en [4] se utilizaron los siguientes: *Daubechies* 8 wavelet, recorte del fotograma centrado, estimación de varianza adaptativa y no zero-meaning. Este experimento muestra los

resultados de la aplicación de la técnica a distintos tamaños de recorte de los fotogramas.

Resolución	Dispositivo					% Acierto
	M1	M2	M3	M4	M5	
1024x768	80,80%	97,40%	88,80%	85,40%	75,40%	**85,56%**
800x600	81,80%	96,80%	84,80%	86,00%	68,80%	**83,64%**
640x480	79,60%	95,60%	85,00%	85,60%	66,20%	**82,40%**
320x240	73,80%	88,20%	78,00%	78,80%	63,40%	**76,44%**
128x128	65,80%	79,20%	66,60%	75,00%	64,00%	**70,12%**

Tabla 5.3.Tasas promedio de acierto por dispositivo en función del tamaño de recorte.

En la Tabla 5.3 se muestra la media de los porcentajes de acierto para cada dispositivo para los distintos tamaños de recortes de los fotogramas. Por porcentaje de acierto se entiende el porcentaje de fotogramas de un vídeo que el clasificador identificó correctamente.

En la mayoría de los casos los porcentajes de acierto por dispositivo aumentan cuanto mayor sea el recorte de los fotogramas (esto se da para todos los casos si se tiene en cuenta la tasa de acierto promedio).

Para la mayor resolución (1024x768) se obtiene la mayor tasa de acierto promedio (85,56%).

La menor tasa de acierto promedio es del 70,12% y se logra utilizando el menor tamaño de recorte (128x128).

En ciertos casos se puede observar que la tasa de acierto aumenta cuando se pasa de una mayor resolución a una menor. Como ejemplo de este caso es el resultado obtenido en el vídeo 3 del dispositivo Apple – iPhone 5 de la Tabla 5.4. Con una resolución de 1024x768, se obtiene una tasa de acierto del 68%, mientras que para una resolución de 800x600 se obtiene una mejora con una

tasa de acierto del 73%. Estos casos son excepcionales. La Tabla 5.4 detalla la tasa de acierto por vídeo de forma individual del experimento 1.

Dispositivos	Vídeo	Resolución				
		1024x768	800x600	640x480	320x240	128x128
Apple (iPhone 5)	1	62%	63%	65%	55%	54%
	2	85%	87%	86%	75%	59%
	3	68%	73%	70%	69%	69%
	4	92%	92%	86%	83%	74%
	5	97%	94%	91%	87%	73%
Nokia (808 Pureview)	1	96%	96%	92%	84%	77%
	2	98%	98%	97%	89%	75%
	3	100%	100%	100%	92%	82%
	4	98%	98%	98%	89%	82%
	5	95%	92%	91%	87%	80%
Samsung Galaxy S4	1	86%	82%	84%	76%	67%
	2	81%	77%	78%	64%	58%
	3	95%	93%	92%	82%	72%
	4	93%	85%	84%	83%	63%
	5	89%	87%	87%	85%	73%
Wiko CinkSlim	1	93%	96%	96%	86%	86%
	2	75%	79%	78%	72%	72%
	3	89%	90%	90%	82%	79%
	4	86%	85%	83%	78%	69%
	5	84%	80%	81%	76%	69%
Zopo ZP980	1	59%	52%	51%	42%	54%
	2	91%	84%	86%	77%	66%
	3	82%	75%	71%	74%	74%
	4	67%	62%	54%	55%	56%
	5	78%	71%	69%	69%	70%

Tabla 5.4. Tasas promedio de acierto por dispositivo en función del recorte.

Como puede observarse, siempre se supera la tasa de acierto por vídeo individual del 50%. Esto indica que en todos los casos y para todos los fotogramas de un vídeo concreto de un dispositivo dado, al menos el 50% de los fotogramas son identificados correctamente.

Finalmente, la identificación de la fuente de un vídeo debe responder a la siguiente pregunta: ¿a qué fuente de adquisición pertenece ese vídeo? Lógicamente podría estimarse que el vídeo perteneciera a la fuente con el mayor número de fotogramas clasificados (mayor porcentaje de acierto) con respecto a las otras fuentes. Se podría dar el caso en el que varias fuentes tengan exactamente el mismo número de fotogramas clasificados y, a su vez, sea el mayor número con respecto a las otras fuentes. En este caso, poco habitual, se diría que la fuente del vídeo no puede ser identificada dudándose entre esas distintas fuentes. En este experimento se obtienen resultados contundentes, que no dejan lugar a dudas sobre la identificación de la fuente de adquisición del vídeo, teniendo en cuenta el criterio definido anteriormente, ya que en todos los casos el acierto supera el 50% como puede verse en la Tabla 5.4.

Asimismo, puede observarse que las tasas de acierto en muchos casos son mucho mayores (llegando hasta el 100% en algunos casos). Por tanto, según este experimento, capturando el criterio antes definido y teniendo en cuenta el vídeo como entidad unitaria (es decir, un vídeo se clasifica bien o no), se puede concluir que esta técnica identifica la fuente de un vídeo con un 100% de acierto.

5.2. Experimento 2

En este conjunto de experimentos se utiliza un mismo tamaño de recorte centrado del fotograma: 640x480. Al igual que en el conjunto de experimentos anterior se utilizaron los 5 dispositivos móviles de la Tabla 5.1, empleando los parámetros de la Tabla 5.2 para cada uno de ellos. Lo que se varía en los

distintos experimentos son los valores de los parámetros de configuración del algoritmo propuesto en [4], con el objetivo de analizar el uso de los distintos parámetros.

En la Tabla 5.5 se muestra un resumen de los experimentos realizados y los parámetros de configuración del algoritmo de extracción de las características utilizados en cada uno de ellos.

Varianza	Zero meaning	% de acierto
Adaptativa	Falso	82,4%
Adaptativa	Verdadero	82,32%
No adaptativa	Verdadero	81,56%
No adaptativa	Falso	82,96%

Tabla 5.5. Parámetros de configuración y tasa media de acierto.

La tasa promedio de acierto para cada uno de los experimentos se muestra en las Tablas 5.6, 5.7, 5.8 y 5.9. La tasa media de acierto en los distintos experimentos varía entre el 81.56% y el 82.96 %, siendo por tanto la diferencia máxima entre el mejor y peor resultado del 1.4 %.

Con los resultados anteriores se concluye que no existen un conjunto de parámetros de configuración del algoritmo de extracción de características que obtenga notablemente mejor resultado sobre los demás.

En los experimentos realizados puede verse que la tasa óptima de acierto se consigue con los parámetros de estimación de la varianza no adaptativa y no zero-meaning. Asimismo, los peores resultados en estos experimentos se obtienen con los parámetros de estimación de varianza no adaptativa y zero-meaning. Dado el estrecho margen de tasa de acierto que hay entre las diferentes configuraciones, los resultados de la configuración óptima y la restante han de tenerse en cuenta para futuros experimentos y aplicación de la técnica, pero las conclusiones no pueden extrapolarse de forma categórica.

Dispositivo	M1	M2	M3	M4	M5
M1	80%	2%	11%	1%	2%
M2	9%	96%	4%	2%	15%
M3	9%	2%	85%	0%	3%
M4	1%	0%	0%	86%	14%
M5	1%	0%	0%	11%	66%

Tabla 5.6.Tasas de acierto por dispositivo para varianza adaptativa y no zero meaning.

Dispositivo	M1	M2	M3	M4	M5
M1	77%	1%	10%	2%	2%
M2	12%	95%	5%	2%	16%
M3	8%	3%	84%	0%	2%
M4	2%	0%	0%	86%	11%
M5	2%	1%	1%	10%	69%

Tabla 5.7.Tasas de acierto por dispositivo para varianza adaptativa y zero meaning.

Dispositivo	M1	M2	M3	M4	M5
M1	75%	2%	9%	0%	1%
M2	13%	92%	4%	2%	18%
M3	8%	5%	85%	0%	2%
M4	2%	0%	0%	85%	8%
M5	3%	1%	2%	12%	70%

Tabla 5.8.Tasas de acierto por dispositivo para varianza no adaptativa y zero meaning.

Dispositivo	M1	M2	M3	M4	M5
M1	78%	1%	8%	0%	1%
M2	10%	94%	4%	1%	15%
M3	10%	5%	87%	0%	2%
M4	1%	0%	0%	86%	12%
M5	1%	0%	0%	12%	70%

Tabla 5.9.Tasas de acierto por dispositivo para varianza no adaptativa y no zero meaning.

5.3. Experimento 3

Se realizaron dos experimentos, uno con un tamaño de recorte de 1024x1024 (tamaño recomendado en [54]) y otro utilizando el tamaño completo de los fotogramas de 1920x1080.

Al igual que en experimentos anteriores se utilizaron los 5 dispositivos móviles de la Tabla 5.1, empleando los parámetros de la Tabla 5.2 para cada uno de ellos.

Debido a los resultados obtenidos en los experimentos realizados en el experimento 2, se ha elegido para este experimento la configuración de los parámetros de estimación de la varianza no adaptativa y no *zero-meaning*, ya que estos eran los que mejores resultados obtuvieron en la tasa promedio de acierto.

En la Tabla 5.10 se muestra la tasa de acierto promedio para los respectivos experimentos.

Resolución	M1	M2	M3	M4	M5	% de acierto
1024x1024	81,2%	97,2%	92,2%	88,0%	78,4%	**87,4%**
1920x1080	87,4%	98,8%	93,0%	89,8%	82,6%	**90,3%**

Tabla 5.10. Promedio de acierto por dispositivo en función del tamaño del recorte.

Se observa que existe una mejora en la tasa promedio de acierto para un tamaño de recorte de 1920x1080 del 2,92 %, con respecto al tamaño de recorte de 1024x1024.

En este experimento concreto puede verse que utilizando toda la imagen completa para todos los casos existe una mayor tasa de acierto promedio en la identificación de fuente, aunque el incremento es pequeño.

En general cuanto mayor es el tamaño del recorte mayor es la tasa de acierto.

Asimismo, con los resultados de este experimento y el experimento 1 se llega a la conclusión de que a partir de un cierto tamaño de recorte el incremento de la tasa de acierto es pequeña e incluso en algunos casos se pueden dar pequeños decrementos como se comentó en experimentos anteriores.

Finalmente, hay que tener en cuenta que a mayor tamaño de recorte mayor tiempo de ejecución del algoritmo de extracción de características.

6. CONCLUSIONES Y TRABAJO FUTURO

6.1. Conclusiones

En este trabajo se ha propuesto una técnica de identificación de la fuente de adquisición de un vídeo producido por un dispositivo móvil. Más concretamente, la marca y el modelo del mismo. Su funcionamiento es como sigue: tras extraer la información relevante (fotogramas clave) del vídeo, un algoritmo de clasificación, basado en el ruido del sensor y la Transformada Wavelet, realiza el proceso de identificación del dispositivo móvil. A la vista de la experimentación realizada se llega a la conclusión general de que esta técnica es válida pues obtiene buenos resultados.

El algoritmo desarrollado de extracción de fotogramas claves presta especial atención a la naturaleza del vídeo, teniendo en cuenta que si aquellos tienen mayor variación de escena, el posterior proceso de clasificación será mejor. El algoritmo requiere además de un número determinado de fotogramas para el entrenamiento y la clasificación utilizando SVM. Una vez obtenidos los fotogramas, se realiza la extracción de las características que se obtienen del patrón de ruido del sensor y la Transformada Wavelet según el algoritmo especificado en [4] identificando la marca y el modelo del dispositivo móvil fuente de una imagen. Este último algoritmo describe el proceso de extracción de la huella del sensor de un fotograma de manera individual, la estimación del patrón del ruido del sensor cuando se cuenta con varias imágenes y la extracción de características requeridas para la identificación de la fuente.

Cabe destacar que el vídeo como ente unitario no es clasificado en un tipo de fuente. Es decir, lo que se identifica son los fotogramas seleccionados por el algoritmo propuesto. Respecto a esto último cabe señalar que la tasa de acierto promedio varía dependiendo de los parámetros utilizados.

Teniendo en cuenta un tamaño de recorte centrado, la conclusión general a la que se llega es que a mayor tamaño de recorte, mejores son los resultados, como se demostró en el experimento 3. En este experimento se observa que tomando el tamaño completo de los fotogramas, la tasa de acierto fue mayor a la de todos los demás tamaños considerados. Todo esto constata que es necesario un tamaño de recorte lo suficientemente grande para obtener buenos resultados y que llega un momento en que el aumentar ese tamaño de recorte mejora poco los resultados o incluso en casos concretos aislados los empeora levemente.

Asimismo, mencionar que el tamaño de recorte no sólo tiene efectos en la tasa de acierto, sino que también tiene efectos en el tiempo de ejecución del algoritmo de extracción de características. Por tanto, el usuario de esta técnica tiene que tener en cuenta el aspecto del compromiso de la mejora de la tasa de acierto con respecto al tiempo de ejecución.

Según los experimentos realizados se estima que para los vídeos de alta definición, el tamaño mínimo de recorte para obtener buenos resultados es de 1024x1024, aunque puede optimizarse tomando un recorte mayor a costa de un mayor tiempo de ejecución.

Asimismo, se ha realizado un conjunto de experimentos variando los distintos parámetros de configuración definidos en [4] para un recorte centrado de fotograma fijo (640x480). En este conjunto de experimentos no se han podido obtener conclusiones categóricas y extrapolables sobre el uso de los parámetros de configuración, ya que en todos los experimentos realizados la tasa de acierto está comprendida en un margen muy pequeño.

Una vez clasificados los fotogramas seleccionados, se ha contestado a la pregunta de cuál es la fuente de adquisición del vídeo como ente unitaria. El criterio elegido ha sido que el vídeo pertenece a la fuente con mayor número de fotogramas clasificados correctamente, obteniéndose una tasa de acierto del 100%.

6.2. Trabajo Futuro

Como posibles trabajos futuros pueden señalarse los siguientes:

- Extender el algoritmo propuesto hasta llegar a la identificación del dispositivo móvil concreto dentro de cada marca y modelo.
- Robustecer los algoritmos de identificación de la fuente diseñados contra posibles ataques. Es posible inhabilitar o engañar a las técnicas de identificación de la fuente, por lo que un área a desarrollar es el diseño de estrategias para detectar cuando una huella ha sido eliminada o suplantada.
- Implementar otros algoritmos de extracción de fotogramas clave y realizar una comparativa de los mismos.

6.3. Publicaciones

De la presente investigación se ha derivado la siguiente publicación:

- Identification of smartphone brand and model via forensic vídeo analysis (with Ana Lucila Sandoval Orozco, Luis Javier García Villalba, Julio César Hernández Castro). ***Expert Systems with Applications***, Volume 55, Issue C, August 2016, pages 59-69.

  http://dx.doi.org/10.1016/j.eswa.2016.01.025.

  Esta revista posee el siguiente factor de impacto:

  F. I. (2015): 2.981 © Thomson Reuters Journal Citation Reports 2016

  COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE – 19/130 (Q1)

  ENGINEERING, ELECTRICAL & ELECTRONIC – 27/255 (Q1)

  OPERATIONS RESEARCH & MANAGEMENT SCIENCE – 6/82 (Q1)

REFERENCIAS

[1] IC Insights, "Embedded Imaging Takes Off as Stand-alone Digital Cameras Stall". http://www.icinsights.com/news/bulletins/Embedded-Imaging-Takes-Off-As-Standalone-Digital-Cameras-Stall/, 2014.

[2] Alexa Internet, Inc. "Alexa Top 500 Global Sites", http://www.alexa.com/topsites, 2015.

[3] C. Wen and K. Yang, "Image Authentication for Digital Image Evidence". Forensic Science Journal, Vol. 5, No. 1, September 2006, pp. 1–11.

[4] A. L. Sandoval Orozco, L. J. García Villaba, D. M. Arenas González, J. Rosales Corripio, J. C. Hernandez-Castro and S. J. Gibson, "Smartphone Image Acquisition Forensics using Sensor Fingerprint". IET Computer Vision, Vol. 9, Issue 5, October 2015, pp. 723-731.

[5] H. J. Zhang, J. H. Wu, D. Zhong and S. W. Smoliar, "An Integrated System for Content-Based Vídeo Retrieval and Browsing". Pattern Recognition, Vol. 30, No. 4, April 1997, pp. 643-658.

[6] M. M. Yeung and B. Liu, "Efficient Matching and Clustering of Video Shots". In Proceedings of the International Conference on Image Processing, Washington, DC, USA, October 1995, Vol. 1, pp. 338-341.

[7] Y. J. Zhang and H. B. Lu, "Hierarchical Video Organization based on Compact Representation of Video Units". In Proceedings of the International Workshop on Very Low Bitrates Video, Kioto, Japan, October 1999, pp. 67-70.

[8] J. Calic and E. Izquierdo, "Efficient Key-Frame Extraction and Video Analysis". In Proceedings of the International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, April 2002, pp. 28-33.

[9] Jesús Bescós Cano, "Segmentacion Temporal de Secuencias de Vídeo". Tesis Doctoral, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, España, Junio 2001.

[10] B. K. R. Horn, B. G. Schunck, "Determining Optical Flow". Artificial Intelligence, Vol. 17, No. 1-3, August 1981, pp. 185-204.

[11] A. Ioannidis, V. Chasanis and A. Likas, "Weighted Multi-View Key-Frame Extraction". Pattern Recognition Letters, Vol. 72, No. 1, March 2006, pp. 52-61.

[12] W. Wolf, "Key Frame Selection by Motion Analysis". In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, USA, May 1996, Vol. 2, pp. 1228-1231.

[13] P. Bestagini, M. Fontani, S. Milani, M. Barni, A. Piva, M. Tagliasacchi, S. Tubaro, "An Overview on Video Forensics". APSIPA Transactions on Signal and Information Processing, Vol. 1, 2012, http://dx.doi.org/10.1017/ATSIP.2012.2.

[14] C. Gianluigi and S. Raimondo, "An Innovative Algorithm for Key Frame Extraction in Video Summarization". Journal of Real-Time Image Processing, Vol. 1, No. 1, March 2006, pp. 69–88.

[15] C. Panagiotakis, A. D. Doulamis and G. Tziritas, "Equivalent Key Frames Selection based on ISO-Content Principles". IEEE Transactions on Circuits and Systems for Video Technology, Vol. 19, No. 3, March 2009, pp. 447–451.

[16] W. Widiarto, S. Maret, E. M. Yuniarno and M. Hariadi, "Video Summarization using a Key Frame Selection based on Shot Segmentation". In Proceedings of the International Conference on Science in Information Technology, Yogyakarta, Indonesia, October 2015, pp. 207-212.

[17] A. Hanjalic and H. J. Zhang, "An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-Validity Analysis". IEEE Transactions on Circuits and Systems for Vídeo Technology, Vol. 9, No. 8, December 1999, pp. 1280-1289.

[18] H. Liu, L. Pan and W. Meng, "Key Frame Extraction from Online Video based on Improved Frame Difference Optimization". In Proceedings of the IEEE 14th International Conference on Communication Technology, Vol. 1, No. 1, November 2012, pp. 940-944.

[19] K. Kurosawa, K. Kuroki, and N. Saitoh, "CCD Fingerprint Method Identification of a Video Camera from Videotaped Images". In Proceedings of the International Conference on Image Processing, Kobe, Japan, October 1999, Vol. 3, pp. 537–540.

[20] J. Lukáš, J. Fridrich, and M. Goljan, "Digital Camera Identification from Sensor Noise". IEEE Transactions on Information Security and Forensics, Vol. 1, No. 2, November 2006, pp. 205-214.

[21] M. Chen, J. Fridrich, M. Goljan and J. Lukás, "Source Digital Camcorder Identification using Sensor Photo Response Non-Uniformity". Security, Steganography, and Watermarking of Multimedia Contents IX, Vol. 6505, 2007, pp. 65051G.

[22] S. Yahaya, A. Ho and A. Wahab, "Advanced Video Camera Identification using Conditional Probability Features". In Proceedings of the IET Conference on Image Processing, London, UK, July 2012, pp. 1-5.

[23] Y. Su, J. Xu and B. Dong, "A Source Video Identification Algorithm based on Motion Vectors". In Proceedings of the Second International Workshop on Computer Science and Engineering, Qingdao, China, October 2009, Vol. 2, pp. 312-316.

[24] David Manuel Arenas González. "Técnicas de Identificación de la Fuente de Adquisición en Imágenes Digitales de Dispositivos Móviles", Tesis Doctoral, Facultad de Informática, Universidad Complutense de Madrid, Marzo 2015.

[25] Ana Lucila Sandoval Orozco, Luis Javier García Villalba, "Análisis Forense de Imágenes y Vídeos Digitales de Dispositivos Móviles – Parte I". Informe Técnico, GASS-UCM, Septiembre 2016.

[26] Y. Q. Shi and H. Sun, "Image and Video Compression for Multimedia Engineering". CRC Press Press, USA, 1999.

[27] Ana Lucila Sandoval Orozco, Luis Javier García Villalba, "Análisis Forense de Imágenes y Vídeos Digitales de Dispositivos Móviles – Parte II". Informe Técnico, GASS-UCM, Septiembre 2016.

[28] G. Wallace, "The JPEG Still Picture Compression Standard". IEEE Transactions on Consumer Electronics, Vol. 38, No. 1, February 1992, pp. 18–24.

[29] J. Lukás, J. Fridrich, "Estimation of Primary Quantization Matrix in Double Compressed JPEG images". In Proceedings of Digital Forensic Workshop, Cleveland, Ohio, USA, August 2003.

[30] W. Wang, H. Farid, "Exposing Digital Forgeries in Video by Detecting Double MPEG Compression". In Proceedings of the 8th ACM Workshop on Multimedia & Security, Geneva, Switzerland, September 2006, pp. 37–47.

[31] W. Luo, M. Wu, J. Huang, "MPEG Recompression Detection Based on Block Artifacts". In SPIE Conference, Vol. 6819, March 2008, 12 pages.

[32] Z. Fan, R. L. de Queiroz, "Identification of Bitmap Compression History: JPEG Detection and Quantizer Estimation". IEEE Transactions on Image Processing, Vol. 12, No. 2, February 2003, pp. 230–235.

[33] W. Wang, H. Farid, "Exposing Digital Forgeries in Video by Detecting Double Quantization". In Proceedings of the 11th ACM Workshop on Multimedia & Security, Princeton, New Jersey, September 2009, pp. 39–48.

[34] P. Bestagini, A. Allam, S. Milani, M. Tagliasacchi, S. Tubaro, "Vídeo Codec Identification". In Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing, Kioto, Japan, March 2012, pp. 2257–2260.

[35] A. R. Reibman, D. Poole, "Characterizing Packet-Loss Impairments in Compressed Video". In Proceedings of the IEEE International Conference on Image Processing, San Antonio, Texas, USA, September 2007, pp. 77–80.

[36] S. Milani, M. Tagliasacchi, M. Tubaro, "Discriminating Multiple JPEG Compression using First Digit Features". In Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing, Kioto, Japan, March 2012, pp. 2253–2256.

[37] A. R. Reibman, V. A. Vaishampayan, Y. Sermadevi, "Quality Monitoring of Video over a Packet Network". IEEE Transactions on Multimedia, Vol. 6, No. 2, February 2004, pp. 327–334.

[38] M. Naccari, M. Tagliasacchi, S. Tubaro, "No-Reference Video Quality Monitoring for H.264/AVC Coded Video". IEEE Transactions on Multimedia, Vol. 11, No. 5, May 2009, pp. 932–946.

[39] G. Valenzise, S. Magni, M. Tagliasacchi, S. Tubaro, "Estimating Channel-Induced Distortion in H.264/AVC Video without Bit Stream Information". In Proceedings of the Second International Workshop on Quality of Multimedia Experience, Trondheim, Norway, June 2010, pp. 100–105.

[40] G. Valenzise, S. Magni, M. Tagliasacchi, S. Tubaro, "No-Reference Pixel Video Quality Monitoring of Channel-Induced Distortion". IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22, No. 4, April 2012, pp. 605-618.

[41] R. Brunelli, O. Mich, "Histogram Analysis for Image Retrieval". Pattern Recognition, Vol. 34, No. 8, August 2001, pp. 1625-1637.

[42] Anil K. Jain, Aditya Vailaya, "Image Retrieval using Color and Shape". Patter Recognition, Vol. 29, No. 8, August 1996, pp. 1253-1244.

[43] G. Pass, R. Zabih. "Comparing Images using Joint Histograms". Multimedia Systems, Vol. 7, No. 3, May 1999, pp. 119-128.

[44] C. Colombo, I. Genovesi. "Color-Induced Image Representation and Retrieval". Pattern Recognition, Vol. 32, No. 10, October 1999, pp. 1685-1695.

[45] E. Binaghi, I. Gagliardi, R. Schettini, "Image Retrieval Using Fuzzy Evaluation of Color Similarity". Journal of Pattern Recognition and Artificial Intelligence, Vol. 8, No. 4, August 1994, pp. 945-958.

[46] G. Ciocca, I. Gagliardi, R. Schettini, "Quicklook2: An Integrated Multimedia System". Journal of Visual Languages & Computing, Vol. 12, No. 1, February 2001, pp. 81-103.

[47] J. L. Michell, W. B. Pennebaker, C. E. Fogg, D. J. Legal, "MPEG Video Compression Standard". Chapman & Hall, Ltd. London, UK, 1996, ISBN: 0412087715.

[48] A. Wahab, A. Ho, and S. Li, "Inter-Camera Model Image Source Identification with Conditional Probability Features". In Proceedings of IIEEJ Image Electronics and Visual Computing Workshop, Kuching, Malaysia, November 2012, Paper ID 2P-2.

[49] A. Wahab, J. Briffa, H. Schaathun, and A. T. S. Ho, "Conditional Probability Based Steganalysis for JPEG Steganography". In Proceedings of the 2009 International Conference on Signal Processing Systems, Singapore, May 2009, pp. 205–209.

[50] C. Li, "Source Camera Identification Using Enhanced Sensor Pattern Noise". IEEE Transactions on Information Forensics and Security, Vol. 5, No. 2, June 2010, pp. 280–287.

[51] E. Bashkov and N. Shozda, "Content-Based Image Retrieval Using Color Histogram Correlation". In Proceedings of the GraphiCon 2002, Nizhny Novgorod, Russia, September 2002.

[52] F. D. O. Costa, M. Eckmann, W. J. Scheirer, and A. Rocha, "Open Set Source Camera Attribution". In Proceedings of the 25th IEEE Conference on Conference on Graphics, Patterns and Images, Ouro Preto, Brazil, August 2012, pp. 71–78.

[53] C. C. Chang and C. J. Lin, LIBSVM: A Library for Support Vector Machines. Version 3.17, April 26, 2013, http://www.csie.ntu.edu.tw/cjlin/libsvm/.

[54] C. T. Li and R. Satta, “On the Location-Dependent Quality of the Sensor Pattern Noise and its Implication in Multimedia Forensics”. In Proceedings of the 4th International Conference on Imaging for Crime Detection and Prevention, London, UK, November 2011, pp. 1–6.