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].

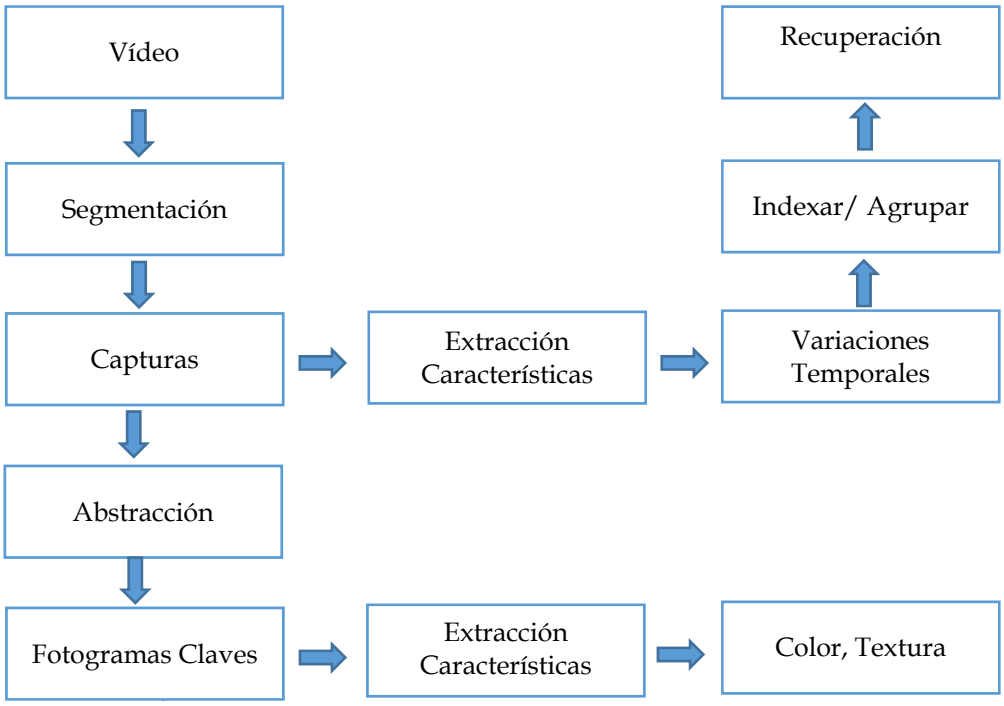

Fig. 3.1. Niveles de información del contenido del vídeo.

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.