Capítulo 3. Ataques en sistemas de sellado digital con marcas de agua

3.1 Introducción

Además de diseñar un adecuado esquema de inserción de la marca de agua, otro aspecto importante y, a menudo olvidado, consiste en la obtención de métodos de evaluación y bancos de pruebas que nos permitan comparar las características de diferentes técnicas de sellado. Esta evaluación no debe reducirse únicamente al estudio de la robustez, sino que también debe incluir una evaluación subjetiva y cuantitativa de la distorsión introducida en el documento sellado mediante la técnica de watermarking. Es importante hacer notar que los métodos bajo investigación sean puestos a prueba bajo condiciones comparables.

En el apartado 2 describimos aspectos generales, los ataques posibles a las marcas de agua son examinados en el apartado 3, algunos bancos de prueba y conclusiones se exponen en los apartados 4 y 5, respectivamente.

3.2 Definición general. Conceptos relacionados

Cuando se habla de robustez nos referimos a la capacidad de la marca de agua para resistir los diferentes procesados de señal y/o manipulaciones que pueda sufrir, con el objeto o no de eliminarla. Es decir, en el contexto de watermarking, se entiende por *ataque* todo procesado a que se vea sometida la marca de agua (y por tanto la imagen, puesto que son inseparables), ya sea para mejorar algunas características de los datos originales -como por ejemplo, uso de técnicas de mejora de la imagen para realzar o reducir el contraste, filtrado para eliminar ruido presente en la imagen, rotación de la misma para alinearla si procede de un escáner… y un sin fin de procesados diferentes que puedan ser requeridos- o bien para atacar directamente a la marca de agua, eliminándola o haciendo imposible su detección, [Heileman98], [Petitcolas98], [Kutter99], [Setyawan], [Kirovsky].

Una primera división de posibles ataques podría ser aquella que separa los ataques intencionados de los que no lo son. Los ataques no intencionados son aquellos debidos a operaciones de procesado de señal comunes realizadas por propietarios legítimos de las imágenes marcada. Los ataques intencionados

son realizados por personas con mayores conocimientos sobre las técnicas de watermarking y con mayor cantidad de recursos a su disposición para realizar el ataque.

Por tanto, conceptos relacionados de manera directa e indirecta son la robustez, la capacidad de la imagen para albergar una marca de agua (puesto que sabemos que mientras mayor sea esta capacidad, mayor podrá ser la intensidad de la marca de agua sin que ésta sea visible, lo que se traduce en una mayor robustez), las características de la imagen en general, puesto que son las que determinan su capacidad, y, obviamente, las técnicas usadas para insertar y detectar la marca de agua, que influyen también en la robustez del sistema. Como puede comprobarse, todos los procesos involucrados en los sistemas de watermarking comprenden un conjunto de conceptos relacionados unos con otros, obligando a decisiones de compromiso para la obtención de soluciones de ingeniería adecuadas a las necesidades específicas de cada aplicación.

3.3 Ataques posibles en marcas de agua

En este apartado se pretende dar respuesta a las siguientes cuestiones [Petitcolas98], [Setyawan]:

- Tipos de ataques que son posibles en los esquemas de sellado digital con marcas de agua.
- Recursos necesarios para borrar las marcas de agua completamente o para alterarlas de manera que resulten imposibles de ser detectadas/extraídas correctamente.
- Nivel de degradación de la calidad de la imagen tras sufrir ataques que tratan de eliminar la marca de agua que contiene.

3.3.1 Clasificación de ataques básicos

Una clasificación general de los diferentes tipos de ataques aplicados a los sistemas de watermarking se ilustra en el siguiente esquema [Setyawan], figura 3.1.

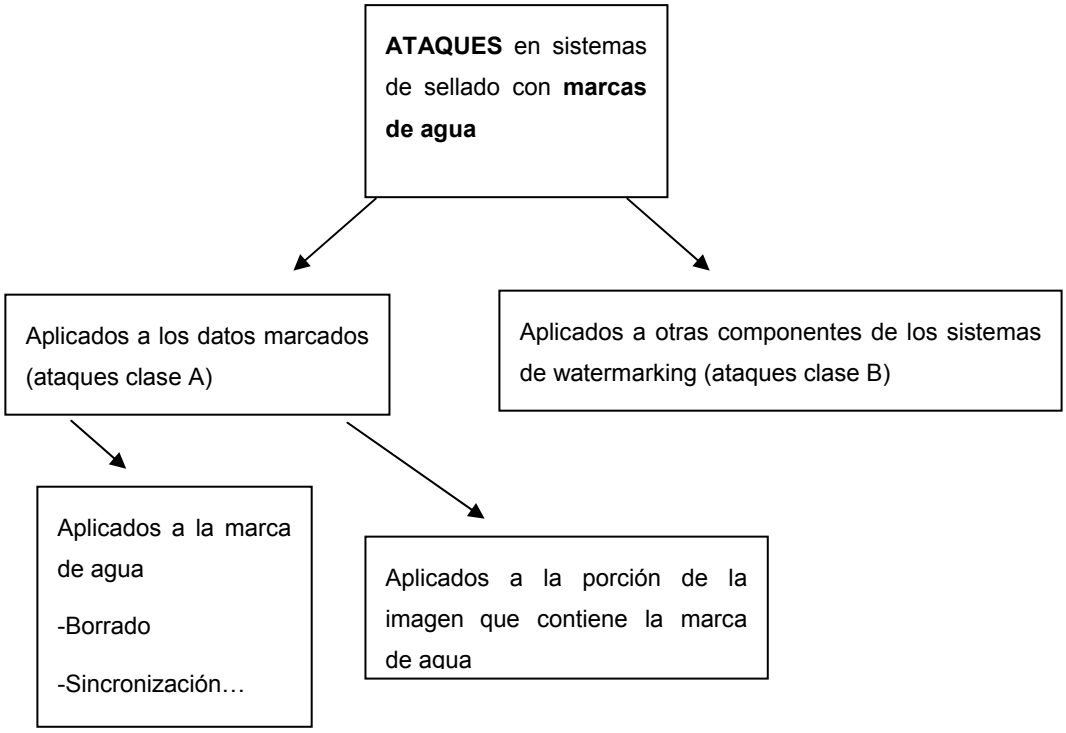

FIGURA 3.1. Clasificación general de los diferentes tipos de ataques en sistemas de watermarking.

En la figura 3.2 se ilustra la diferencia entre los ataques clase A y clase B, que básicamente, radica en los diferentes dominios en que opera cada categoría de ataques.

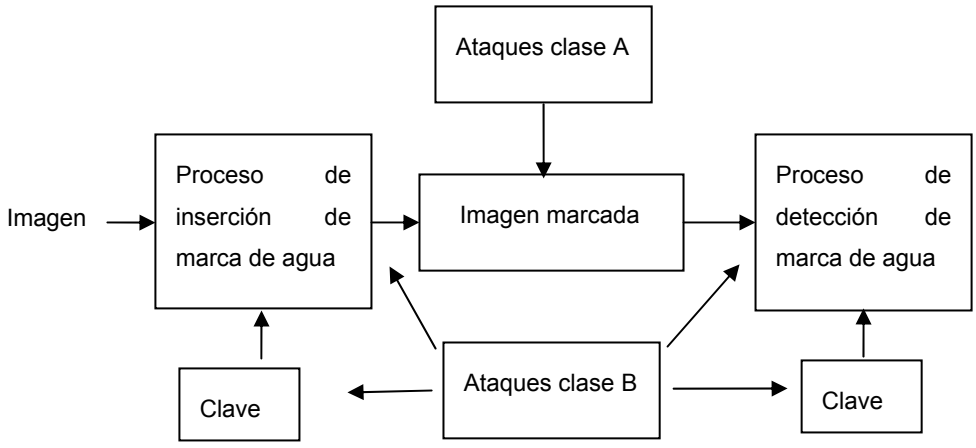

FIGURA 3.2. Representación de los dominios en que operan los ataques clase A y clase B.

Puesto que los ataques clase A operan sobre la imagen sellada, estos ataques, generalmente, involucran algunas operaciones de procesado de señal. Además, como se muestra en la figura 3.1 los ataques clase A se dividen a su

vez en dos categorías. La primera ataca directamente a la marca de agua encajada en la imagen y trata de conseguir que el detector sea incapaz de detectar cualquier marca de agua legítima fuera de una duda razonable. La segunda categoría intenta modificar, o por otro lado atacar la parte de la imagen en la que se incrusta la marca de agua, sin destruir la marca propiamente dicha.

Los ataques clase B tratan de llevarse a cabo sin prestar atención ni a la marca de agua ni a las componentes, marcadas o no, de la imagen. Por lo tanto, en este caso podría no ser necesaria una operación de procesado de señal. En su lugar, suele requerirse un conocimiento profundo de los lenguajes de programación, sistemas operativos o hardware. Estos ataques son referidos como ataques de *hacking* o *cracking* en el ámbito software y como ataques hardware si tratan sobre mencionado dominio.

3.3.1.1 Ataques clase A

Ya se ha indicado que este tipo de ataques se subdivide en dos categorías principales, los que versan sobre la marca de agua insertada y los que dirigen su efectividad sobre los datos marcados de la imagen. A partir de aquí nos referiremos a los primeros como ataques clase A-1 y a los segundos como ataques clase A-2.

3.3.1.1.1 Ataques clase A-1

Esta categoría de ataques la subdividimos en tres subclases, cada una de ellas con sus propias características distintivas, que son:

1. Ataques de borrado.
2. Ataques de sincronización.
3. Ataques de ambigüedad.

1. Ataques de borrado

La característica principal y distintiva de este tipo de ataques es su objetivo, que consiste en eliminar o degradar de forma severa la información de la marca de agua insertada en el documento sellado de tal manera que el detector no pueda determinar la presencia de la misma en la imagen bajo test. Pero estos ataques tratan de realizar lo indicado sin atentar contra la seguridad del algoritmo empleado, es decir, sin conocer la clave utilizada en el mecanismo de inserción de la marca de agua en la imagen. Frente a un ataque efectivo de

este tipo ningún procesado posterior conseguirá reconstruir la información de la marca.

Los ataques de borrado los podemos dividir, a su vez, en *ataques simples* y *ataques de análisis* que usan diferentes estrategias para conseguir un objetivo común.

1.a) Ataques de borrado simples

No intentan analizar los datos marcados para conseguir que la marca de agua sea indetectable, operan directamente sobre ellos tratando de degradar la calidad de la marca de agua lo suficiente como para que no sea detectada, o incluso de eliminarla, [Setyawan], [Kutter99], [Petitcolas98], [Barnett98], [Orúe03], [Minguillón02], [Langelaar98].

Puesto que este tipo de ataques trabajan directamente sobre la imagen marcada, tanto la imagen como la marca de agua son degradadas durante el proceso. Sin embargo estos ataques tienen en cuenta el hecho de que la energía de la señal de marca de agua es mucho menor que la de la imagen por lo que sería de esperar que la señal incrustada sea degradada por debajo de los límites de la detección correcta antes de que la calidad de la imagen se vea severamente degradada.

Hacer notar que el término simple empleado en este contexto remarca el hecho de que el atacante no trata de analizar la marca de agua insertada en la estego imagen, lo que no debería hacer pensar que los ataques de borrado simples resulten sencillos o triviales de llevar a cabo.

Ejemplos de ataques de borrado simples son:

- **Compresión con pérdidas**. Por ejemplo JPEG, que descarta porciones de la imagen que se consideran que aportan poca cantidad de información visual. La cantidad de información borrada depende de los factores de calidad y de compresión utilizados. Por otro lado la marca de agua es normalmente insertada en esas porciones desechables de los datos para conseguir que tenga el menor impacto posible en la calidad visual de la imagen. Por lo tanto dicha señal de marca de agua podría ser borrada o severamente degradada durante el proceso de compresión con pérdidas.
- **Suma de ruido**. Sumando un patrón de ruido aleatorio a una imagen marcada podría deteriorarse el patrón de la marca insertada de modo

que se hiciera irreconocible para el detector. Sin embargo, el ruido añadido tendría que tener o bien una potencia mucho mayor a la de la marca de agua o estar correlado con ella.

- **Conversión A/D y D/A**. Algunas técnicas de sellado con marcas de agua, tales como aquellas que manipulan los bits menos significativos de los datos digitales, podrían no sobrevivir a este ataque. Cuando los datos son convertidos a señales analógicas la marca de agua puede perderse. Un ejemplo sería imprimir una imagen digital y escanear el material impreso para luego obtener otra imagen digital. La imprecisión de la reproducción de la imagen escaneada debido a las limitaciones del proceso de escaneo podría conseguir eliminar la marca de agua originalmente encajada en la imagen digital.
- **Filtrado general**. Operaciones de filtrado general podrían usarse para atacar imágenes selladas. Un filtrado paso bajo, por ejemplo, podría ser capaz de borrar una marca de agua de ruido pseudo-aleatorio, puesto que la marca de agua en este caso es equivalente a un ruido de alta frecuencia.

1.b) Ataques de borrado basados en análisis

En este caso, un atacante intentaría analizar (mediante la ayuda de algún análisis estadístico) los datos sellados para encontrar o estimar la marca de agua o la imagen original no marcada. Esta información es luego utilizada para eliminar la marca de agua. Este tipo de ataques son, normalmente, bastante elaborados y realizados de forma intencionada, a diferencia de los ataques de borrado simples. El hecho de borrar la marca de agua usando ataques basados en análisis de los datos marcados, no suele afectar de forma severa a la calidad de la imagen original. Entre los ataques que pertenecen a esta categoría podemos señalar los siguientes.

- **Filtrado no lineal**. Usando este tipo de filtrado un atacante podría estimar la marca de agua insertada en una imagen. Esta estimación será usada entonces para borrar la señal de sellado. Un ejemplo de este tipo de ataques es el conocido como WRS descrito en [Langelaar98].
- **Promediado estadístico**. Para realizar este procesamiento el atacante dispone de N imágenes diferentes pero que contienen la misma marca de agua. Promediando estadísticamente estas

imágenes el atacante podría ser capaz de estimar la marca de agua contenida en ellas, información que puede usarse para conseguir borrar la señal de sellado insertada. Este ataque será particularmente exitoso si la marca de agua insertada no es dependiente de las imágenes en un grado significativo.

- **Ataque de colusión**. Podría verse como el complementario del ataque de promediado estadístico anterior. En esta proceso un conjunto de atacantes posee cada uno una copia de la misma imagen I, pero sellada con marcas de agua diferentes. Un ataque efectivo podría conseguirse promediando estas copias para estimar la imagen original sin marcar, o bien, tomando pequeñas partes de todas ellas. La resistencia a estos ataques es crítica en las aplicaciones de marcas de agua transaccionales, ya que estos sistemas incrustan marcas diferentes sobre un mismo original del producto, cuando este va dirigido a varios destinatarios [Heileman98], [Setyawan], [Cox97].

- **Observación de los dispositivos de sellado/detección**. Este método es diferente al ataque hacking que pertenece a los ataques clase B. Si el atacante tiene a su disposición el dispositivo detector, modifica las propiedades de los datos marcados (cambiando la luminancia de los píxeles, por ejemplo) y observa la respuesta del detector. Su objetivo consiste en encontrar la modificación más pequeña posible en la imagen sellada de tal forma que el detector falle en la detección de la presencia de la señal de marca de agua. Esta modificación es entonces aplicada a todas las imágenes marcadas mediante el mismo esquema o similar. Este tipo de ataque es también conocido como ataque Oracle. Si el atacante está en posesión del dispositivo que encaja la marca de agua en la imagen, éste podría ser capaz de observar los datos antes y después del proceso de inserción de marcas de agua, calculando la diferencia entre las imágenes obtendría la marca de agua. Todo lo que necesita entonces es realizar una pre-distorsión en el material no marcado restándole la diferencia obtenida (esto es, la marca de agua estimada). Cuando luego este material pase a través del dispositivo sellador (que lo que hace es insertarle la marca de agua) lo que tendrá a la salida será aproximadamente igual a la imagen original sin marca de agua

incrustada, este tipo de ataque es también conocido como ataque de protocolo.

2. Ataques de sincronización ("jitter")

La característica principal de esta categoría de ataques es que el atacante no trata de borrar la marca de agua a partir de los datos marcados, sino de eliminar la sincronización de la marca de agua de manera que no pueda ser detectada adecuadamente por el dispositivo detector [Setyawan], [Heileman98]. Tras el ataque lo más probable es que la marca de agua aún esté físicamente contenida en la imagen. Al igual que en los ataques de borrado simples el atacante no realiza ningún análisis de la imagen marcada para identificar la marca de agua. La diferencia principal entre este ataque y el ataque simple es que en el simple, la señal de marca de agua es degradada con el objetivo de hacerla indetectable, mientras que en el ataque de sincronización la marca de agua pierde únicamente su sincronización con el detector. Ataques pertenecientes a esta subcategoría pueden ser:

- **Transformación geométrica**. Simples transformaciones geométricas tales como traslación por un conjunto de píxeles, escalado de la imagen (que incluye ampliación, reducción y/o cortado de partes de la imagen sellada) o rotación de pocos grados, normalmente, consiguen que la marca de agua insertada pierda su sincronización. Estas traslaciones simples no suelen afectar a la calidad de la imagen.
- **Sustitución/eliminación de píxeles**. Este ataque se lleva a cabo, por ejemplo, borrando una fila /columna de píxeles de la imagen. Si se quiere preservar el tamaño de la imagen otra fila/columna podría ser duplicada e insertada (cerca de la fila/columna duplicada). Esta operación no provoca una degradación visible de los datos marcados.
- **Ataque de mosaico**. Este ataque es realizado mediante la división de la imagen en porciones más pequeñas. Cuando es usado en páginas web, un buscador podría reconstruir la imagen sin aparente pérdida de calidad o retraso temporal (a veces cargar una imagen completa resulta más lento que reconstruirla a partir de sus piezas). Este ataque se realiza, fundamentalmente, para prevenir que los web-crawlers diseñados para chequear las marcas de agua en imágenes almacenadas en servidores de internet puedan completar su trabajo puesto que las piezas pequeñas no contienen marcas de agua

reconocibles. Es posible, por supuesto, insertar copias individuales de la marca de agua en bloques más pequeños de la imagen original. Sin embargo los métodos de watermarking actuales son incapaces de insertar marcas de agua en piezas pequeñas de la imagen (más pequeñas de 100×100 píxeles). Por tanto todo lo que necesita hacer el atacante es dividir la imagen original en bloques más pequeños de 100×100 píxeles y los web-crawlers fallarán, muy probablemente, en la detección de la marca de agua.

3. Ataques de ambigüedad

Un atacante intenta insertar otra marca de agua en la estego imagen y de esta manera hacer difícil o imposible determinar cuál fue la primera marca de agua insertada (la legítima). Una posible contramedida frente a este ataque consistiría en insertar un sello temporal o bien manteniendo los datos marcados originales como una referencia en caso de disputa.

Una variante más sofisticada de este tipo de ataques consiste en considerar que parte de los datos marcados originariamente son una imitación de los datos originales y poner una segunda marca de agua derivada de los datos marcados legítimamente (lo que se denomina ataque de inversión). Por ejemplo, si asumimos que I es la imagen original, w la marca de agua que se inserta mediante la función $E(I,w)$ e I_w la versión sellada de I como resultado de aplicar dicha función; se ha demostrado que para la mayoría de los esquemas de sellado con marcas de agua si el atacante dispone de la imagen marcada I_w podría calcular fácilmente un patrón w', una imitación I' de la original y una función $E'(I',w')$, tal que:

$$I_w = E'(I',w') \qquad \textbf{(3.1)}$$

De manera que podría afirmar que I' es su versión original de la imagen sin marcar y w' su marca de agua, creando así cierta ambigüedad sobre la propiedad de I_w. Resolver este problema podría no ser fácil e, incluso, imposible puesto que cada parte podría afirmar rotundamente que su marca de agua está incluida en la, supuestamente, imagen original de la parte contraria (I e I'). El ataque mencionado funciona sólo en algoritmos de watermarking invertibles.

3.3.1.1.2 Ataques clase A-2

Los ataques incluidos en esta categoría pretenden modificar o estropear los datos en los que está encajada la marca de agua, sin destruirla. Ejemplos de ataques aquí enmarcados podrían ser: emborronado (desenfoque) de parte de la imagen, cambio del color de alguna zona de la imagen o incluso falseo de algunas áreas de la misma. Probar que una imagen ha sido alterada de este modo es, a veces, más un arte que una ciencia.

Estos ataques suelen funcionar frente a marcas de agua robustas, sin embargo, frente a marcas de agua frágiles, que están específicamente diseñadas para detectar alteraciones, estos ataques tienen poco que hacer.

3.3.1.2 Ataques clase B

Los ataques incluidos en esta clase pretenden deteriorar el sistema de sellado atacando cualquiera de sus componentes fundamentales, salvo la imagen marcada en sí. De este modo, podría atacar las componentes de software o las componentes de hardware del sistema de watermarking.

3.3.1.2.1 Ataques de “hacking” o “cracking”

Como ya se indicó en secciones previas, estos ataques tratan con componentes del software. Si los componentes de sellado y detección de marca de agua están implementados mediante software y disponibles con facilidad, resultan especialmente susceptibles a estos procesos.

Ataques de este tipo se realizan de la siguiente forma: un atacante (un hacker, en este contexto) obtiene el software detector o insertador de la marca de agua y accede a su código fuente localizando en él la parte que genera o detecta la marca de agua. Una vez esto es conseguido, la información obtenida se usa para cumplir el objetivo del atacante.

Por ejemplo, si encontrara la parte del código usada para generar la marca de agua podría utilizar esta información para crear marcas de agua falsas, o bien, si encontrara la parte de código del detector que chequea la presencia de marca de agua, podría modificar el código para saltarse la rutina de seguridad del esquema implementada en el detector.

Otro ejemplo sería el ***ataque por fuerza bruta*** que se usaría para encontrar la clave secreta usada en el esquema de sellado digital con marcas de agua. Este ataque se lleva a cabo generando las claves aleatorias y probando

cada una de ellas para encontrar cuál funciona. Si las posibles llaves son muchas y largas este tipo de ataque podría requerir demasiado tiempo, resultando impracticable.

3.3.1.2.2 Ataques que modifican el hardware ("hardware tampering")

Ataques que se desarrollan en las componentes hardware de los sistemas de sellado digital con marcas de agua, por ejemplo un reproductor de DVD. Un atacante podría desmontar el hardware, estudiar el funcionamiento interno del mismo y modificarlo para adaptarlo a sus necesidades. (Podría desactivar cierta circuitería del reproductor para inutilizar la característica que lo habilita para detectar o modificar marcas de agua y que en lugar de eso analizara las señales, pudiendo entonces usar su reproductor modificado para hacer copias de material protegido con derechos de copia).

Hacer notar que la clasificación de ataques aquí indicada no es única ni estándar, pudiéndose encontrar en la bibliografía existente otras enumeraciones de ataques tan válidas como la elegida, decisión basada en su aspecto sistemático.

Con el objeto de ordenar ideas y, dada la gran cantidad de ataques señalados se presenta la siguiente tabla que resume las diferentes técnicas de ataque vistas en esta sección.

<table>
<tr><td rowspan="3">CLASE 1-A</td><td rowspan="2">Ataque de borrado</td><td>Borrado simple</td><td>-Compresión con pérdidas
- Suma de ruido
-Conversión A/D y D/A
- Filtrado general</td></tr>
<tr><td>Basado en análisis</td><td>- Filtrado no lineal
- Promediado estadístico
- Ataque de colusión
- Observ. detector/insertador</td></tr>
<tr><td>Ataque de sincronización</td><td colspan="2">- Transformaciones geométricas
- Sustitución/borrado píxeles
- Ataque de mosaico</td></tr>
</table>

	Ataque de ambigüedad. Trata de crear dudas acerca de la propiedad de una imagen, insertando en ella otra marca de agua.
CLASE 2-A	- Emborronado de parte de la imagen - Cambio de color de una zona de la imagen - Falseo de áreas determinadas…
CLASE B	Ataques de "hacking" y "cracking", que modifican o utilizan el software del sistema
	Ataques que tratan de modificar el hardware del sistema

TABLA 3.1. Esquema resumen de las diferentes categorías de ataques mostradas.

3.4 Bancos de prueba. Ataques comerciales

En este apartado se comentan, brevemente, algunos de los bancos de prueba de técnicas de sellado digital de imágenes más conocidos, indicando los ataques que implementan [Petitcolas98], [Petitcolas03], [Kutter99]. Para el estudio de eficiencia de los algoritmos realizados se ha utilizado el banco de pruebas conocido como Stirmark por considerarse uno de los más adecuados para ello y ser de libre difusión, de manera que una explicación más detallada de esta herramienta puede encontrarse al final del presente documento, en el apartado dedicado a anexos (Anexo IV).

Las herramientas comerciales utilizadas para comprobar la robustez que vamos a ver son Stirmark y Optimark.

3.4.1 Stirmark

La mayoría de los esquemas de sellado digital de imágenes resisten ataques simples que pueden implementarse con herramientas estándar, pero no responden tan bien frente a combinaciones de los mismos, lo cual llevó a desarrollar Stirmark. Se trata de una herramienta para tecnologías de sellado digital de imágenes [Petitcolas98], [Kutter99]. Dada una imagen de entrada marcada, Stirmark genera un número de imágenes modificadas que pueden usarse para verificar si la marca insertada en la imagen puede aún detectarse.

El banco de pruebas de Stirmark distingue nueve categorías diferentes de ataques que son: enriquecimiento de señal, compresión, escalado, cortado o eliminación de partes de la imagen, recortado asimétrico, rotación,

transformaciones lineales, otras transformaciones geométricas y distorsiones geométricas aleatorias. En el caso de escalado, cortado de parte de la imagen, rotación, transformaciones lineales y otras transformaciones geométricas las imágenes atacadas son obtenidas con y sin compresión JPEG de un 90% de factor de calidad.

Para asegurar una comparación adecuada, Petitcolas (uno de los principales desarrolladores del banco de pruebas) impone una PSNR mínima de 38dB. Sin embargo, es una restricción discutible puesto que es una medida que tiene poca utilidad en el contexto de distorsiones geométricas.

En su versión más simple, el ataque estrella de Stirmark simula un proceso de remuestreo mediante la introducción de distorsiones geométricas aleatorias en la imagen. Por ejemplo, introduce en la imagen los mismos tipos de errores que provoca la impresión de la misma en una impresora de alta calidad que luego es escaneada por un escáner de gama alta. Aplica una distorsión geométrica menor: la imagen es ligeramente desplazada, ensanchada, estrechada y/o rotada por una cantidad aleatoria e imperceptible y entonces se vuelve a muestrear usando interpolación bilineal y/o Nyquist. En definitiva se aplica a todos los valores de las muestras un error pequeño y distribuido. Esto simula las pequeñas imperfecciones no lineales introducidas por los convertidores analógico-digitales típicamente presentes en los escáneres y dispositivos de presentación. Stirmark introduce una pérdida de calidad en la imagen prácticamente inapreciable si es aplicado una sola vez. Sin embargo, tras varias aplicaciones iteradas la distorsión se hace tremendamente visible.

Con estas distorsiones geométricas simples se podría confundir a la mayoría de los sistemas de marcado disponibles en el mercado. Más distorsiones, todavía inapreciables desde el punto de vista de la calidad visual de la imagen atacada, pueden aplicarse a una imagen. Por ejemplo, si aplicamos una torsión global a la imagen, lo cual se consigue si se aplica una ligera desviación a cada píxel, que es mayor en el centro de la imagen y prácticamente nula en los bordes. Por encima de esto, se añade un desplazamiento de alta frecuencia de la forma: $\lambda \sin(\omega_x y)\sin(\omega_y y) + n(x,y)$, donde *n* es un número aleatorio. Para que estas distorsiones sean más eficientes se utiliza una compresión JPEG al final.

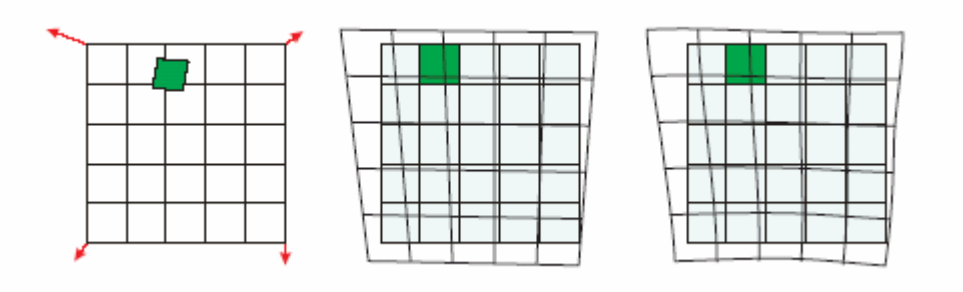

FIGURA 3.3. Representación exagerada de la distorsión introducida por Stirmark en imágenes

Stirmark ha sido desarrollada por Fabien Petitcolas durante su tesis doctoral en la Universidad de Cambridge, Reino Unido. Desde su primera publicación en 1997 ha cobrado un gran interés por la comunidad de sellado de imágenes siendo en la actualidad el banco de pruebas más utilizado aplicado a tecnologías de sellado digital de imágenes.

Las alteraciones fundamentales realizadas a la imagen, contenidas en la versión 3.1 de Stirmark, utilizada para probar la efectividad de los algoritmos implementados, se indican en el siguiente esquema.

ATAQUES IMPLEMENTADOS POR STIRMARK	
- Cortado, escalado y cambios de escala - Inversión aleatoria - Rotación con y sin escalado posterior - Filtrado FMLR, sharpening, Gaussiano	- Transformaciones lineales, torsiones/curvaturas aleatorias - Alteración de la relación de aspecto - Borrado simétrico/asimétrico de líneas - Compresión JPEG

TABLA 3.2. Conjunto de ataques que realiza el banco de pruebas Stirmark sobre la imagen bajo test.

3.4.2 Optimark

Optimark es una herramienta para algoritmos de sellado digital de imágenes estáticas que fue desarrollado en el Laboratorio de Análisis de la Información e Inteligencia Artificial del departamento de Informática de la Universidad Aristotélica de Tesalónica (Grecia).

Sus principales características se pueden resumir en las siguientes:

- Interfaz de usuario gráfica.
- Evaluación del rendimiento de la detección/decodificación usando múltiples ejemplos con diferentes claves de sellado y mensajes.
- Evaluación de las siguientes métricas del rendimiento de la detección:
    - Para detectores de marcas de agua que proporcionan una salida en números reales (flotante)
        - Curvas de las características de operación del receptor (ROC), por ejemplo, curvas de representación de la probabilidad de falsa alarma frente a la probabilidad de falso rechazo.
        - Velocidad de error igual (equal error rate).
        - Probabilidad de falsa alarma para una probabilidad de falso rechazo fija y definida por el usuario.
        - Probabilidad de falsa alarma y falso rechazo.
    - Evaluación de las siguientes métricas del rendimiento de la decodificación, para algoritmos que permiten codificación de mensajes.
        - Tasa de error de bit.
        - Porcentaje o probabilidad de decodificación perfecta de mensajes.
        - Evaluación de los tiempos medios de inserción y detección.
        - Evaluación del coste computacional del algoritmo.
        - Evaluación del límite de ruptura del algoritmo para ciertos ataques y determinados criterios de rendimiento, por ejemplo evaluación de la severidad del ataque en la que el rendimiento del algoritmo supera o cae por debajo de un cierto límite.

ATAQUES IMPLEMENTADOS POR OPTIMARK	
- Cortado	- Transformaciones lineales
- Inversión aleatoria	- Alteración de la relación de aspecto
- Rotación	- Cambios de escala
- Rotación y escalado	- Borrado de líneas
- Filtrado FMLR, sharpening, Gaussiano	- Reducción de color
- Torsiones/curvaturas aleatorias	- Compresión JPEG

Tabla 3.3. Conjunto de ataques que realiza el banco de pruebas Optimark sobre la imagen bajo test.

3.4.3 Certimark

Aunque no se trata de un banco de pruebas lo incluimos en este apartado dada la relación que existe entre el nacimiento del proyecto Certimark con la aparición de bancos de prueba aceptados de forma generalizada. Este hecho aceleró el lanzamiento del proyecto Certimark (para la certificación de técnicas de sellado digital de imágenes con marcas de agua) que comenzó en Mayo de 2002. El proyecto incluye a 15 participantes tanto académicos como industriales. Sus objetivos son:

- Diseñar, desarrollar y publicar un banco de pruebas completo, adecuado para las tecnologías de watermarking, disponiendo de diferentes escenarios de aplicación.
- Hacer este banco de pruebas como una herramienta de referencia para suministradores y clientes de diferentes sectores de tecnología.
- Establecer un proceso de certificación de algoritmos de sellado con marcas de agua.
- Concentrar recursos en encontrar las características más importantes de las técnicas de watermarking para la protección de imágenes estáticas y video a baja velocidad en entornos abiertos e inseguros como Internet.

El objetivo de Certimark es conseguir que los algoritmos de watermarking sean etiquetados con una certificación internacional. El uso de bancos de prueba

nos lleva a la obtención de una referencia reconocida internacionalmente, que permitirá a los clientes hacerse con la propiedad de los algoritmos de sellado más apropiados para sus necesidades. También se permite de este modo la competitividad entre suministradores de tecnología mientras se mantiene una calidad estándar determinada, cuando es medida por el banco de pruebas. En suma, la originalidad y atención del proyecto recae en el desarrollo en paralelo de herramientas de evaluación objetivas y de técnicas de sellado con marcas de agua robustas para poder ser certificadas y conseguir la confidencialidad de los contenidos propietarios.

3.5 Discusión y propuesta de contramedidas

Hasta el momento se han discutido diferentes formas de atacar los sistemas de watermarking, pero también pueden estudiarse posibles contramedidas que podrían realizarse para suavizar los efectos de determinados ataques.

3.5.1 Contramedidas frente a ataques básicos

Para combatir los ataques de borrado simples tendríamos que intentar que nuestro algoritmo de sellado pudiera generar marcas de agua con mayor potencia, encajándolas en la imagen de forma adecuada para no degradarla por encima de un umbral considerado, lo que se consigue explotando de forma más eficiente las propiedades del HVS. En todo caso, la utilización de señales de marca de agua más fuertes implica que los datos originales presenten una mayor degradación de su calidad visual frente a los ataques que hacen que la marca de agua sea indetectable, lo que no es nada deseable para el atacante. Sin embargo, esta medida podría no funcionar frente a los ataques de borrado basados en análisis o estimaciones, especialmente en el caso en que el atacante pudiera observar tanto la imagen marcada como la original sin marcar. La efectividad de los ataques de promediado estadístico podría reducirse haciendo que la señal de marca de agua fuera altamente dependiente del contenido de la imagen, mientras que para combatir los ataques de colusión se podría proponer usar marcas de agua anticolusión (como por ejemplo las generadas haciendo uso de los conceptos de espectro expandido), [Hart99], [Cox97] y [Setyawan].

Por su naturaleza, los ataques de sincronización no eliminan la marca de agua de la imagen sellada, lo que podría explotarse para el diseño de

contramedidas que combatan este tipo de ataques. Añadir más inteligencia en el detector podría habilitar la recuperación de la marca de agua. Otra posible contramedida sería usar algoritmos de sellado con marcas de agua que pudieran insertar algún tipo de patrón de manera que el detector pudiera estimar las modificaciones aplicadas a la imagen sellada e invertirlas. Un mecanismo que se puede utilizar para combatir las transformaciones geométricas consiste en emplear un dominio transformado para la inserción de la marca de agua que sea invariante a tales transformaciones (rotación, escalado y traslación), un dominio con estas características es el de Fourier-Mellin, [Lin01].

Usar un sello temporal seguro podría ser una solución frente a los ataques de ambigüedad simples (los que insertan una segunda marca de agua en la imagen ya marcada). Frente a ataques más sofisticados, por ejemplo el de inversión, el diseñador del algoritmo de sellado debería asegurarse de que éste fuera no invertible en sentido general.

Los ataques que modifican el contenido pueden combatirse empleando algoritmos de sellado con marcas de agua diseñados no sólo para detectar si la imagen ha sido modificada (mediante el empleo de marcas de agua frágiles, por ejemplo), sino también para mostrar si el atacante ha modificado el material.

Los ataques clase B se pueden evitar diseñando cuidadosamente los componentes software y hardware del sistema de sellado con marcas de agua. Aunque no haya garantías de que esto funcionara, se conseguiría que el ataque resultara más duro y costoso, lo que podría hacerlo impracticable en cierto sentido.

Un ataque al software, podría ser erradicado o, por lo menos, ralentizado mediante la implementación de un software de decodificación de bajo nivel que intercepta llamadas que el software detector de marcas de agua hace al sistema operativo, creando así una nueva capa entre en detector y el sistema operativo. Para mantener compatibilidad con las imágenes no selladas se podría considerar la definición de un nuevo formato para los ficheros que contengan marcas de agua. Es decir, el acceso a ficheros protegidos debería ser enrutado a este software de decodificación, mientras que los ficheros sin proteger deberían ser accedidos de la forma habitual. Este tipo de procesamiento tendría consecuencias positivas, sobre todo en lo que a velocidad se refiere.

Un ataque a un componente hardware podría ser deteriorado mediante la construcción del hardware de tal forma que cualquier modificación que se le

hiciera destruiría básicamente el hardware, sin dejar nada para que el atacante pueda aprender. Una solución menos drástica sería diseñarlo de tal forma que si se detecta alguna modificación física se destruyan o encripten todos los datos que está procesando. Este tipo de soluciones resultan difíciles, costosas y poco deseables en general, salvo en aplicaciones militares.

Cuando se intenta vencer a un atacante debemos tener en cuenta los criterios que puede considerar al desarrollar su ataque, entre estos criterios [Setyawan] suelen estar los siguientes:

- **Coste-efectividad**. El atacante puede cuestionarse si el ataque es lo suficientemente barato o efectivo como para llevarlo a cabo o no. Lógicamente, si el coste de atacar la marca de agua es mayor que el de obtener la imagen legalmente, no le saldría rentable realizar el ataque. En este caso, el coste puede no referirse a una cuestión monetaria, puede contemplar aspectos relacionados con el esfuerzo y/o el tiempo requeridos para realizar el ataque. Este criterio, depende en todo caso de la persona atacante y de sus objetivos.
- **Calidad de la degradación**. Una marca de agua bien diseñada provocará una gran degradación de la imagen original en caso de sufrir un ataque, lo que daría un escaso valor comercial a la imagen atacada. Sin embargo esto también depende del mercado del atacante. Por lo tanto, lo ideal sería conseguir que el borrado de una marca de agua provoque no una degradación de la imagen, sino que ésta sea imposible de ver.

3.5.1.1 Contramedidas básicas frente a ataques basados en estimación

Para poder resistir a este tipo de ataques debe hacerse una inserción de la marca de manera que resulte difícil estimarla. Este método se ha desarrollado en dos diferentes escenarios que se resumen brevemente a continuación.

3.5.1.1.1 Condición de espectro de potencia

Un método teórico pensado para analizar los ataques basados en la estimación considera la imagen y la marca insertada como procesos aleatorios gaussianos independientes estacionarios y coloreados. La imagen marcada es la suma de estos dos procesos, puesto que se dispone de la señal original se asume que su potencia espectral es fija, pero la potencia espectral de la marca

de agua puede ser modificada. La cuestión es, ¿cuál sería la forma más adecuada de la señal de marca de agua para resistir tales ataques? Para este escenario la estimación óptima de la forma del espectro de la marca de agua se obtiene usando filtros de Wiener.

El error cuadrático medio E entre la marca de agua original y la estimada proporciona una medida adecuada para conocer cómo de bien resiste una marca a la estimación. Se puede demostrar que E es máximo si la potencia espectral de la marca de agua es directamente proporcional a la potencia espectral de la señal original, es decir de la imagen. Este requisito es conocido como condición de espectro de potencia (PSC). Una marca cuyo espectro de potencia satisface el PSC es la más resistente frente a la estimación.

Si la distorsión se mide por la diferencia cuadrática media entre la imagen atacada y la original sin marcar, la condición de espectro de potencia tiene otra consecuencia importante: para cualquier salida del filtro adaptado, una marca que cumple la PSC provoca que el ataque mencionado incurra en una distorsión aún mayor. Para llevar la correlación a cero el ataque debe hacer la distorsión tan grande como la potencia de la imagen original, de manera que la imagen atacada se vea severamente degradada.

3.5.1.1.2 Función de visibilidad de ruido

La condición de espectro de potencia es atractiva porque se puede probar de manera rigurosa y tiene una formulación matemática precisa. Para sellado digital de imágenes con marcas de agua, la eliminación de ruido en la imagen proporciona una manera natural de desarrollar ataques basados en la estimación optimizados por los estadísticos de las imágenes, aunque de manera óptima esto podría ser difícil de probar.

La imagen sellada es tratada como una versión ruidosa de la original, donde la marca de agua representa el ruido que debería ser eliminado. De este modo la marca estimada es lo mismo que el ruido estimado.

Se aplican diferentes modelos estadísticos para las imágenes originales, tales como un proceso gaussiano no estacionario o bien un proceso gaussiano generalizado estacionario. En el primer caso el método de eliminación de ruido usa un filtro de Wiener adaptativo, mientras que en el segundo se reduce al uso de los métodos tradicionales de eliminación de ruido mediante umbralización.

Ambos métodos de eliminación de ruido producen una función de enmascaramiento de textura (TMF con valores en [0,1]) que se deriva de los estadísticos de la imagen y es, por tanto, dependiente de la misma. Para insertar una marca que resista la estimación, en el proceso de inserción de la marca debería usarse la función invertida conocida como función de visibilidad de ruido y definida como:

$$NVF = 1 - TMF \quad \textbf{(3.2)}$$

Valores de NVF cercanos a cero indican regiones de textura o borde, donde la marca debería ser amplificada. De esta manera la marca es insertada para resistir los ataques basados en la estimación haciendo uso del conocimiento de las técnicas de eliminación de ruido.

3.6 Conclusiones

Se ha mostrado una clasificación de los ataques más comunes en sistemas de watermarking, indicando algunos ejemplos, así como posibles contramedidas que pueden utilizarse para combatir los mecanismos de ataque más habituales.

Este análisis teórico de los ataques en marcas de agua nos da una idea cercana del problema de watermarking y nos permite ver los límites fundamentales de esta tecnología.

Con todo esto lo que se pretende hacer ver es la gran complejidad para conseguir algoritmos de sellado con marcas de agua robustos y eficientes, así como para disponer de estándares que permitan definir los requerimientos y propiedades de los diferentes algoritmos en función de la aplicación a la que se destinen.

La robustez a transformaciones globales afines está más o menos resuelta en la actualidad. Sin embargo, la resistencia a alteraciones globales aleatorias, ataque implementado por Stirmark, (que explota el hecho de que el sistema visual humano no es sensible a variaciones locales y transformaciones afines de modo que desplaza, escala y rota localmente los píxeles sin apreciarse distorsión en la imagen) constituye todavía un problema no resuelto satisfactoriamente.