UNIVERSIDAD MAYOR DE SAN ANDRÉS

FACULTAD DE CIENCIAS PURAS Y NATURALES

CARRERA DE INFORMÁTICA

TESIS DE GRADO

"TÉCNICAS DE MARCAS DE AGUA PARA AUTENTICACIÓN DE DOCUMENTOS"

PARA OPTAR AL TÍTULO DE LICENCIATURA EN INFORMÁTICA

MENCIÓN: INGENIERÍA DE SISTEMAS INFORMÁTICOS

POSTULANTE: SAMUEL NINA GUTIERREZ

TUTOR METODOLÓGICO: LIC. GROVER ALEX RODRIGUEZ RAMIREZ

ASESOR: M. Sc. LUCIO TORRICO DIAZ

LA PAZ – BOLIVIA

2014

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

LA CARRERA DE INFORMÁTICA DE LA FACULTAD DE CIENCIAS PURAS Y NATURALES PERTENECIENTE A LA UNIVERSIDAD MAYOR DE SAN ANDRÉS AUTORIZA EL USO DE LA INFORMACIÓN CONTENIDA EN ESTE DOCUMENTO SI LOS PROPÓSITOS SON ESTRICTAMENTE ACADÉMICOS.

LICENCIA DE USO

El usuario está autorizado a:

a) visualizar el documento mediante el uso de un ordenador o dispositivo móvil.
b) copiar, almacenar o imprimir si ha de ser de uso exclusivamente personal y privado.
c) copiar textualmente parte(s) de su contenido mencionando la fuente y/o haciendo la referencia correspondiente respetando normas de redacción e investigación.

El usuario no puede publicar, distribuir o realizar emisión o exhibición alguna de este material, sin la autorización correspondiente.

DEDICATORIA

A las personas que forman parte de mi vida.

AGRADECIMIENTOS

A Dios por darme la oportunidad de vivir y darle sentido a mi existencia.

A mis padres y hermanos por su apoyo incondicional.

Al M.Sc. Lucio Torrico Diaz, por su paciencia, comprensión, consejos y observaciones oportunas.

Al Lic. Grover Alex Rodriguez Ramirez, por paciencia y el seguimiento realizado.

A los Docentes de la Carrera que contribuyeron en mi formación.

A mis amigos por sus observaciones y por su amistad.

Contenido

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Es común contar con documentos impresos en papel ya sea este en material pre impreso o pre formateado. Debido a esto surge la necesidad de saber si los datos impresos son realmente los que dicen ser (auténticos) o en caso contrario son falsos o sufrieron alteración. Las diferentes entidades como ser instituciones del sector público, privado y otros emiten documentación importante en muchos casos con mecanismos de seguridad que son difíciles de falsificar, pero no todas incluyen estos mecanismos, dejando desprotegido los documentos ante posibles alteraciones que puedan sufrir estos, muchas de estas entidades para autenticar los datos de una copia hacen el proceso denominado legalización el cual en muchos casos es un proceso burocrático.

Con el desarrollo de la tecnología en el ámbito de la Informática tanto en software, hardware y las técnicas de marcas de agua digitales, se propone utilizar la técnica de marcas de agua para autenticar el contenido de texto de un documento físico (impreso en papel), enfocándonos en los datos de texto contenidos en este documento y no así en el material que los contiene. Para esto se propone que el vínculo entre los datos impresos en papel y los datos guardados en digital sea la marca de agua y mediante esta se realice el proceso de autenticación de forma automática e inmediata, la técnica de marcas de agua se basa en el texto impreso donde este se somete a un proceso de normalización para su aplicabilidad. En el prototipo desarrollado se establece la figura de una Autoridad Certificadora, esta presta los servicios web de registro de documentos de texto digital con su respectivo autor y servicio de autenticación de documento físico (impreso) correspondiente al digital registrado.

Las pruebas realizadas con el prototipo implementado fueron realizadas para documentos de una página y para la autenticación de varias páginas se propone hacer el proceso por separado, se obtuvieron resultados alentadores, aunque la imagen tomada con la cámara del Smartphone requiere suprimir el factor humano al momento de capturar la imagen del documento de texto, esto debe tomarse en cuenta para trabajos futuros en el área.

CAPÍTULO I: MARCO REFERENCIAL

1.1 INTRODUCCIÓN

El acceso a la tecnología tiene un constante crecimiento como es el caso de la Internet cuya difusión ha llegado a posicionarse como una herramienta de investigación, publicidad, comercio, ocio, etc. Este crecimiento se incrementa con la aparición de teléfonos móviles inteligentes y tabletas, los cuales facilitan el acceso a los diferentes servicios de la Nube (Cloud Computing).

Con los servicios de la Nube, el compartir documentos ya sea estos de texto, imágenes, audio y video es algo común. Por lo cual debido a la proliferación de estos documentos electrónicos en la red, se han propuesto diferentes mecanismos criptográficos de seguridad para proteger del acceso no autorizado a dichos documentos. Con el incremento de la capacidad de proceso de la computadora es posible descifrar la clave de los mecanismos de seguridad clásicos, dejando sin protección a los recursos o documentos electrónicos, quedando vulnerables, sujeto a modificaciones y perdiendo así el derecho de autor sobre dicho documento.

Como un complemento para los sistemas de seguridad surgen las "Marcas de agua digitales", las cuales empezaron a tener interés por parte de los investigadores a partir de los años 90 y cuya primera conferencia especializada fue organizada en 1996. El marcado de agua digital es un proceso de incrustar información, llamada "marca de agua", de manera discreta y robusta dentro el documento huésped. Una vez incrustada la marca de agua en el documento huésped, se forma otro documento el cual es resistente a los ataques que pretendan extraer la marca de agua, si se intenta degradar la marca de agua también se degrada el documento electrónico por lo que queda inservible (Barán, Gómez, & Bogarín, 2001).

Los teléfonos inteligentes son una herramienta cotidiana, de la vida moderna y de uso masivo en los países desarrollados. Entre sus características comunes está la función multitarea, el acceso a Internet vía WiFi o red 3G, función multimedia (cámara y reproductor de videos/mp3), los programas de agenda, administración de contactos, acelerómetros, GPS

y algunos programas de navegación, así como ocasionalmente la habilidad de leer documentos de negocios en variedad de formatos como PDF y Microsoft Office.

En la coyuntura de Bolivia el tener un documento o certificado es importante, ya que en ellos están impresos datos que acreditan a una persona o un bien material. Pero el proceso de autenticación de un documento es un proceso burocrático. Por lo que se realizan una serie de trámites como son las copias legalizadas significando pérdida de tiempo a los usuarios.

El presente trabajo pretende crear un mecanismo de autenticación de los documentos impresos físicamente en papel los cuales están relacionados con uno digital, con la ayuda de teléfonos inteligentes como una herramienta de escaneado del documento, se pretende autenticar los documentos impresos en tiempo real. Para ese caso el documento impreso cuenta con una marca de agua perceptible a la lente de la cámara del teléfono inteligente. Para el efecto se pretende desarrollar un prototipo en el sistema operativo Android, este prototipo autenticará el medio impreso detectando la marca de agua impresa y comparando con el que tiene el documento digital relacionado.

1.2 PROBLEMA

1.2.1 Antecedentes

Las técnicas de Watermarking insertan una señal adicional, conocida como watermark o marca de agua, directamente en los datos originales. Esta marca no es más que un mensaje idealmente imperceptible y de difícil extracción por parte de un usuario no autorizado. La información de este mensaje puede representar, por ejemplo, una secuencia binaria que contiene un número de serie, un logo, un número de tarjeta de crédito, una imagen o una firma.

La marca de agua no restringe el acceso a los datos mientras que en el cifrado clásico la idea es lograr que los mensajes sean ininteligibles para cualquier persona no autorizada. Las marcas de agua están diseñadas para residir permanentemente en los datos del anfitrión. Si la propiedad de un documento digital está en cuestión, la información del propietario del

documento se puede extraer completamente por medio del reconocimiento de la marca de agua.

Durante siglos, el tipo de criptografía que ha estado en uso es el conocido como criptografía de clave privada o criptografía de clave secreta. Este nombre viene del hecho de que el emisor y receptor en una comunicación utilizan la misma llave, la cual debe mantenerse en secreto. Este tipo de criptografía también se conoce como criptografía simétrica, debido a que es la misma clave que se utiliza en los dos lados de la comunicación. Existen varios algoritmos de clave secreta. Uno de los mejores algoritmos conocidos es el DES, que todavía se utiliza en la actualidad en aplicaciones tales como bancaria, en particular para cajeros automáticos (Barán, Gómez, & Bogarín, 2001).

Huella digital, es un mecanismo para defender los derechos de autor y combatir la copia no autorizada de contenidos, que consiste en introducir una serie de bits imperceptibles sobre un producto de soporte electrónico (CD-ROM, DVD, etc.) de forma que se puedan detectar las copias ilegales.

1.2.2 Formulación del problema

Con el incremento de la cantidad de documentos que circulan en la Internet, con esfuerzo y dedicación los criptoanalistas pueden interceptar estos documentos, de esta manera los documentos digitales quedan desprotegidos y con riesgo de sufrir alteraciones.

En el contexto Nacional hace falta estudios a profundidad de técnicas que permitan resguardar la propiedad intelectual de documentos electrónicos como ser: imágenes, videos, texto, sonido y software.

Además cuando se imprime el documento o se realizan copias de la misma se pierde la información acerca de los autores de estos documentos.

En la actualidad es necesario contar con un mecanismo que permita autenticar un documento impreso en tiempo real.

¿De qué manera contribuirán las técnicas de marcas de agua a resguardar y/o proteger el derecho de autor y autenticidad en un documento digital o impreso ya sea esto para una persona u organización, además autenticar el documento digital y/o documento impreso en papel en tiempo real y cumpliendo las características de marca de agua?

1.3 OBJETIVOS

1.3.1 Objetivo general

Establecer la técnica de marca de agua como un medio para resguardar la información sobre la propiedad intelectual de un documento digital e identificar la autenticidad de un documento impreso relacionado con el documento electrónico en tiempo real.

1.3.2 Objetivos específicos

Analizar las técnicas de watermarking apropiadas para realizar el presente trabajo.

Analizar las características de las marcas de agua apropiadas para realizar el trabajo.

Establecer a la técnica watermarking como un mecanismo que permite resguardar o proteger la información acerca de los autores de un documento.

Establecer el algoritmo de Watermaking adecuado para el presente trabajo, el cual debe cumplir las propiedades de las marcas de agua.

Lograr el desarrollo de un prototipo que permita generar, incrustar y detectar una marca de agua en el documento electrónico.

Lograr el desarrollo de un prototipo que permita incluir una marca de agua patrón al documento físico en papel derivado del documento electrónico.

Lograr el desarrollo de un prototipo que facilite el reconocimiento y autenticación de la marca de agua impresa en papel en tiempo real, con la ayuda de un Smartphone.

1.4 HIPÓTESIS

La técnica de marca de agua permite reguardar el derecho de autor y autenticidad de un documento digital o impreso, además de autenticar el documento en tiempo real.

Variables independientes:

To: texto de documento original (digital registrado).

Tr: documento de texto físico (impreso original o copia del digital).

Variables dependientes:

K: Clave, palabra con mayor frecuencia en el texto K=F(To).

W1: marca de agua robusta dependientes de texto W1=W1(To, K).

W2: marca de agua frágil W2=W2(To).

Kr: Clave, palabra con mayor frecuencia en el texto Kr=F(Tr).

Wr1: marca de agua robusta dependiente de Tr donde W1=W1(Tr, Kr).

Wr2: marca de agua frágil correspondiente a Tr donde Wr2=W2(Tr).

1.5 JUSTIFICACIONES

1.5.1 Justificación social.

El trabajo a realizarse beneficia a la población de personas e instituciones que tengan interés en proteger el derecho de autor sobre los documentos que generan.

Además a las instituciones que emiten documentos les facilitará el proceso de colocado de marcas ya que ellos podrán insertar o adecuar para colocar su propia marca.

El aporte a la comunidad científica se reflejará en la facilidad de crear, insertar e identificar al dueño de un documento electrónico el cual puede ser una publicación de un artículo científico, diseño gráfico o algún documento importante. Además el usuario de un Smartphone[1] podrá autenticar un documento impreso relacionado a su documento digital en tiempo real.

1.5.2 Justificación económica

Los mecanismos de seguridad clásicos en documentos impresos en papel requieren inversiones anticipadas para contar con ellos por lo cual se incurren en altos costos.

Para la autenticación clásica de documentos impresos, es necesario contar con personal entrenado o con equipos costosos.

Con el desarrollo del presente trabajo se disminuirá el costo en papel con características particulares de seguridad.

1.5.3 Justificación técnica

Por la tecnología creciente es oportuno crear aplicaciones que permitan utilizar los diferentes accesorios con los que vienen equipados los teléfonos inteligentes, en particular los que utilizan el sistema operativo Android.

Se cuenta con las herramientas y tecnología suficiente en el ámbito de software para el desarrollo del presente trabajo. Como ser la herramienta Matlab, lenguajes de programación y con la teoría sobre Watermarking.

En cuanto a hardware se dispone de lo necesario para desarrollar el trabajo. Como ser un equipo que soporte los programas, impresora o un dispositivo móvil con sistema Android.

[1] Teléfono inteligente, que por lo general cuenta con recursos de hardware como ser, wifi, cámara, bluetooth, acceso a internet, gps y otros dispositivos.

1.5.4 Justificación científica

Con el desarrollo del presente trabajo se mostrará la integración de las diferentes técnicas de algoritmia, funciones matemáticas utilizadas, codificación y decodificación de la marca de agua impresa y el cómo se hace para ligar un documento impreso con su contenido digital.

Se valorará los resultados obtenidos con la técnica de marcas de agua, los cuales podrán tomarse en cuenta para trabajos futuros, en el área de la seguridad, autenticidad y protección de documentos con la ayuda de la informática.

La determinación de la técnica apropiada de marca de agua y su aplicación para resguardar el derecho intelectual, y la investigación que se realiza en el presente trabajo ayudará a los investigadores locales que se interesen por realizar trabajos similares.

1.6 Alcance

El presente trabajo contempla una investigación de tipo exploratoria, por lo que el alcance se limita al desarrollo de prototipos para la gestión de marcas de agua en documentos de tipo texto. También se contempla un prototipo para el reconocimiento o autenticación de los patrones de marcas de agua insertados en estos documentos, el cual se desarrollará para Smartphones con sistema operativo Android.

CAPÍTULO II: MARCO TEÓRICO

2.1 ESTEGANOGRAFÍA

La esteganografía está definida como el arte de ocultar información en archivos de texto, imágenes, sonidos o en canales encubiertos a través de métodos y técnicas computacionales. Se encuentra en el ámbito de transportar información a través de redes informáticas (Angulo, Blandón, & Ocampo, 2007).

El proceso de la esteganografía en general consiste en colocar un mensaje oculto en algún medio de transporte o portador, llamado el transportista. El mensaje secreto está incrustado en el transportista para formar el medio de la esteganografía (Menvielle, et al., 2011).

La técnica de esteganografía se debe apoyar en dos principios básicos: el primero en seleccionar muy bien el medio en el que se desea aplicar, refiriéndose a que el archivo cubierta o portador a pesar de que pierde calidad no debe ser perceptible; el segundo aprovechando las limitaciones del hombre en cuanto a la percepción se refiere, como lo es la gama de colores que aunque varíen un poco el ojo humano no alcanzará a percibir, al igual que frecuencias que el oído no alcanza a notar (Angulo, Blandón, & Ocampo, 2007).

2.2 TÉCNICAS DE MARCAS DE AGUA DIGITALES (WATERMARKING)

La esteganografía proporciona algunas funciones muy útiles y comercialmente importantes en el mundo digital, en especial la marca de agua o filigrana digital. En esta aplicación, el autor puede incorporar un mensaje oculto en un archivo de manera que la titularidad de la propiedad intelectual puede ser afirmada más tarde y/o para garantizar la integridad del contenido. Un artista, por ejemplo, podría publicar obras de arte originales en una página Web. Si alguien roba el expediente y realiza alegaciones del trabajo como suyo, el artista puede demostrar la propiedad, ya que sólo él puede recuperar la marca de agua (Menvielle, et al., 2011).

Aunque conceptualmente es similar a la esteganografía, la filigrana digital (Marca de agua) normalmente tiene objetivos técnicos diferentes. En general, sólo una pequeña cantidad de información repetitiva se inserta dentro de la portadora, no es necesario ocultar la información de marca de agua, y es conveniente que pueda ser eliminada sin afectar la integridad de la portadora (Menvielle, et al., 2011).

Una marca de agua digital es una señal permanente integrada en los datos digitales (audio, imágenes, video y texto) que se puede detectar o extraer más tarde por medio de operaciones computacionales, con el fin de hacer afirmaciones acerca de los datos. La marca de agua se oculta en los datos del anfitrión, de tal manera que se hace resistente a muchas operaciones que no degraden el documento anfitrión. Así, por medio de marcas de agua, lo datos siguen siendo accesibles, pero marcados de forma permanente (Wu & Lui, 2004).

Un sistema de marcas de agua involucra un proceso de marcado y otro de detección que, generalmente, requieren una clave de propósito similar a la clave utilizada en los sistemas criptográficos. El nivel de disponibilidad de la clave, determinará quién o quiénes podrán leer o detectar la marca de agua. En la práctica, la mayoría de las técnicas de marcas de agua pueden considerarse como sistemas criptográficos simétricos, en los que se emplea una sola clave, variando en ellos el nivel de acceso a esa clave (Orúe, 2002).

2.2.1 Propiedades en las marcas de agua

Existe un gran número de publicaciones en las que se discuten los requisitos que deben cumplir las marcas de agua. Es bueno destacar que la seguridad de estos sistemas no debe estar basada en la ocultación de los algoritmos utilizados, sino en la fortaleza de los mismos y en la seguridad de la clave (Orúe, 2002).

Entre las propiedades deseables de un sistema de marcas de agua se encuentran la robustez, la resistencia a las manipulaciones, imperceptibilidad, el costo computacional y la baja probabilidad de error.

2.2.1.1 Robustez

Los archivos digitales de imágenes, audio y video, están expuestos a muchos tipos de modificaciones (o distorsiones): las pérdidas por compresión, los cambios producidos por el mejoramiento de imágenes, la amplificación de las señales de audio, etc. Una marca de agua se considera robusta si perdura después de esas operaciones y, en el caso de las marcas en imágenes y en video, también deben persistir después de las transformaciones geométricas (recortado, rotación, escalado). Esto quiere decir que la marca ha de estar presente en los archivos y que debe ser detectada después de las distorsiones (Barán, Gómez, & Bogarín, 2001). Para consolidar su robustez, los sistemas de marcas de agua deben insertar la marca en las regiones perceptualmente significativas de los archivos multimedia (J.Fridrich & Goljan, 1999).

Existen diversas opiniones a la hora de definir la robustez de estos sistemas, nuestro punto de vista coincide con las formuladas en los trabajos de Fridrich & Goljan(1999) entre otros, por tanto aclaramos que la valoración de esta propiedad de los sistemas de marcas de agua, no incluye los ataques basados en el conocimiento de los algoritmos de incrustado y detección de la marca, la robustez significa resistencia a ciegas frente a aquellas modificaciones producidas por las operaciones comunes a las que estarán expuestos los archivos o documentos electrónicos en particular los multimedia.

2.2.1.2 Resistencia a manipulaciones

La resistencia a manipulaciones de un sistema de marcas de agua es un aspecto que puede relacionarse con la seguridad del mismo; se refiere a su resistencia frente a los ataques hostiles basados en el total conocimiento de los algoritmos de incrustado y detección y de los archivos marcados, excepto de la clave utilizada. Se incluyen aquí los ataques a los protocolos y los ataques basados en la estimación del sistema. Según la aplicación de que se trate, unos ataques serán más importantes que otros; en general, un ataque efectivo deberá eliminar la marca de agua sin cambiar la calidad perceptual del archivo en cuestión; sin embargo, existen varias aplicaciones en las que la resistencia a determinadas manipulaciones

es un aspecto indeseable. A continuación se revisan algunos de los ataques básicos a los sistemas de marcas de agua (Orúe, 2002).

2.2.1.3 Imperceptibilidad e indetectabilidad

La imperceptibilidad y la indetectabilidad de las marcas de agua son dos conceptos que tienden a confundirse frecuentemente, aunque son muy distintos y no están relacionados entre sí.

La imperceptibilidad o transparencia de la marca tiene como base el comportamiento del sistema perceptual humano. Una marca de agua es imperceptible (transparente), si la degradación que causa en los archivos donde se ha insertado es muy difícil de apreciar. Este concepto se contrapone al de la robustez, si tenemos en cuenta que un sistema robusto debe insertar la marca en las regiones perceptualmente significativas del archivo. En algunas aplicaciones se puede aceptar una pequeña degradación de los datos, a cambio de lograr mayor robustez o menor costo del sistema (Orúe, 2002).

La indetectabilidad está relacionada con el modelo estadístico del archivo antes y después de ser marcado. Se dice que la marca es indetectable si después de haberla insertado, el archivo marcado conserva las mismas propiedades estadísticas que su original. Lo que quiere decir que una persona no autorizada no podrá detectar la presencia de la marca utilizando métodos estadísticos. Esta propiedad es muy deseable en el caso de las comunicaciones encubiertas en las que el principal objetivo es ocultar la presencia del mensaje incrustado en el archivo (Orúe, 2002).

2.2.1.4 Viabilidad del sistema

Toda tecnología que pretende ser comercializada, debe tener en cuenta varios aspectos, entre ellos: el coste computacional, el coste económico y la escalabilidad del sistema. En muchos sistemas, tales como los de audio y video, la marca debe ser insertada y/o detectada en tiempo real, lo que requiere una gran capacidad computacional de los equipos.

En algunas aplicaciones el número de equipos que insertan la marca de agua difiere de la cantidad de detectores, lo que marcará la diferencia de precio entre unos y otros de acuerdo a la aplicación concreta.

Los requerimientos computacionales exigen a los sistemas de marcas de agua simplicidad, pero ésta puede significar la reducción de la resistencia a las manipulaciones. Sin embargo, hay que tener en cuenta que la velocidad de las computadoras se dobla anualmente, de manera que un algoritmo que hoy no nos parezca razonable, podrá rápidamente convertirse en algo factible; es muy deseable diseñar sistemas de marcas de agua que sean escalables con cada generación de computadoras.

2.2.1.5 Baja probabilidad de error

En la mayoría de los sistemas de marcas de agua es muy importante distinguir entre los archivos que contienen una marca y los que no.

La probabilidad de error al detectar una marca debe ser muy pequeña. Se denomina probabilidad de falso negativo a la probabilidad de que, habiendo estado presente una marca en determinado archivo, el detector asuma que no hay tal marca. Por otro lado, la probabilidad de falso positivo es la probabilidad de que no estando la marca presente en un archivo, el detector asuma que la marca está presente (Eggers, Su, & Girod, 2001).

2.3 APLICACIONES DE LAS TÉCNICAS DE MARCAS DE AGUA

Los requisitos que deben cumplir en la práctica los algoritmos de marcas de agua deben analizarse dentro del entorno de trabajo del sistema y de acuerdo con la aplicación donde serán utilizados, dicho esto consideraremos algunas de las posibles aplicaciones de las marcas de agua y sus peculiaridades.

2.3.1 Marcas de agua como firma

Las marcas pueden utilizarse para firmar archivos multimedia. El propietario de uno de estos archivos insertará una marca de agua que lo identifique como tal. Esta aplicación puede verse en los siguientes escenarios:

2.3.1.1 Identificación de propietario

La forma usual de informar sobre el derecho de propiedad, tanto en libros, fotografías o cualquier tipo de documentos, como en las cajas de CD de música y los créditos de las películas, es una nota de copyright colocada en forma visible. Evidentemente estas notas no garantizan la protección de tales materiales, baste sólo nombrar lo fácil que resulta borrar los créditos de una película, o tirar la envoltura de un CD de música. Como complemento de las notas de copyright puede insertarse una marca de agua que formará parte del contenido del producto; pongamos por ejemplo, la información del copyright insertada dentro de una imagen fotográfica (Velasco Bautista, López Hernández, Nakano Miyatake, & Pérez Meana, 2007).

2.3.1.2 Prueba de propiedad

Los propietarios de archivos multimedia pueden usar las marcas de agua no sólo para identificar su copyright sino también para probar la propiedad que ejercen sobre estos archivos.

Entre los software de marcas de agua desarrollados con el objetivo de identificación de propietario para diversos archivos multimedia se encuentran: EIKONAmark (imagen), AudioMark, y VideoMark de la corporación Alpha Tec (Orúe, 2002).

2.3.2 Marcas de agua transaccionales (fingerprinting)

Las marcas de agua también pueden utilizarse para identificar a los compradores de los archivos multimedia, lo que puede servir para la búsqueda del infractor en el caso de distribución de copias ilegales de un archivo dado. En este caso, la marca de agua

transaccional se incrusta de manera adicional (o efectuando una nueva copia de los archivos originales) y llevará los datos del propietario y los datos del comprador. Además de usar la marca de agua (firma), para demostrar la propiedad de sus datos multimedia, el propietario podría determinar a quién atribuir la distribución ilegal de las copias que ha vendido (Orúe, 2002).

Es interesante recalcar que dentro de los requisitos de esta aplicación, el sistema ha de tener capacidad y permiso para insertar varias marcas de agua en un mismo archivo.

2.3.3 Marcas de agua para autenticación

Existen muchas aplicaciones donde la veracidad de una imagen es crucial, tal es el caso de imágenes médicas y muchas otras. Las marcas utilizadas para la autenticación contendrán la información requerida que determinará la integridad de un archivo multimedia. La marca debe ser invisible y frágil (cualquier modificación de la imagen debe alterar la marca) y es muy deseable que pueda ofrecer información sobre los cambios ocurridos en las imágenes (Eggers, Su, & Girod, 2001).

2.3.5 Control de copias

Las marcas de agua diseñadas para el control de copias contendrán la información determinada por su propietario acerca de las reglas de uso y copiado de los archivos en los que se insertan. A diferencia de las marcas de agua transaccionales, así como las marcas de aguas usadas para el monitorizado, identificación y pruebas de propiedad, que sólo sirven como herramienta para investigar a los transgresores del sistema, las marcas de agua usadas en el control de copias restringen la utilización de los archivos de acuerdo a las regla de uso y copiado que porten (Orúe, 2002).

2.3.6 Comunicaciones secretas

En esta aplicación, la marca incrustada en los archivos multimedia se utiliza por dos o más personas para comunicarse secretamente sin levantar la sospecha de terceros. Es la aplicación clásica de la esteganografía (ocultar una información dentro de otra) de

comunicación por canales subliminales. Existen varios software de dominio público que pueden utilizarse con estos fines entre ellos Steghide (Orúe, 2002).

2.4 ALGORITMO DE MARCAS DE AGUA DIGITALES

Todos los sistemas de watermarking constan de tres fases que se puede definir como (Reverte & Martínez, 2004):

- Generación de la marca.
- Inserción de la marca (algoritmo de marcado).
- Extracción o detección de la marca.

2.4.1 Generación de la marca de agua

La marca de agua es una información codificada que se inserta en un archivo de datos digital y que permite identificar el documento, pero sin afectarla. Las marcas de agua pueden ser de cualquier naturaleza como un número, un texto o incluso una imagen. No obstante dependiendo de la técnica de inserción utilizada y de la aplicación a la que se destina, las marcas deberán cumplir ciertas condiciones. Por ejemplo para el caso de marcas frágiles o visibles estas condiciones son bastante flexibles y solo habrá restricciones en cuanto al tamaño de la marca o su formato (por ejemplo si la marca es un texto, que sólo contenga caracteres ASCII o si se trata de una imagen que esté en formato BMP, etc.).

Matemáticamente podemos ver la marca de agua W como el resultado de cierta función F sobre una clave K,

$$W=F(K). \qquad \text{(Ec. 1.1)}$$

Si además se tienen en cuenta las características particulares de los datos I sobre los que se insertará la marca, tendremos

$$W=F(K,I) \qquad \text{(Ec. 1.2)}$$

2.4.2 Inserción de la marca

La figura 1 muestra el esquema general del proceso de inserción de una marca digital.

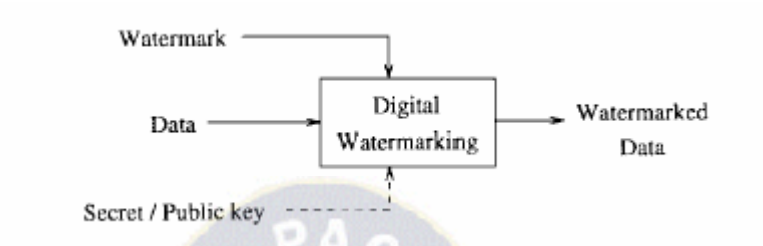


Figura 1 Proceso de inserción de marca

Fuente: (Reverte & Martínez, 2004)

Las entradas son la marca digital W, los datos digitales I que contendrán la marca y una clave privada o pública K que es opcional y que en el caso de claves privadas se utilizará para reforzar la seguridad, por ejemplo haciendo que las posiciones donde se inserta la marca sean generadas a partir de dicha clave K.

En el proceso de inserción obtenemos los datos marcados Iw a partir de los datos originales I y la marca de agua W. Matemáticamente:

$$Iw = E(I,W) \qquad \text{(Ec. 1.3)}$$

La función de inserción E dependerá de la técnica de marcado concreta que se utilice. Como se verá posteriormente para el caso de las imágenes digitales, existen muy diversas técnicas de marcado.

2.4.3 Detección o extracción de la marca

La figura 2 muestra el esquema general del proceso de detección/extracción de una marca digital.

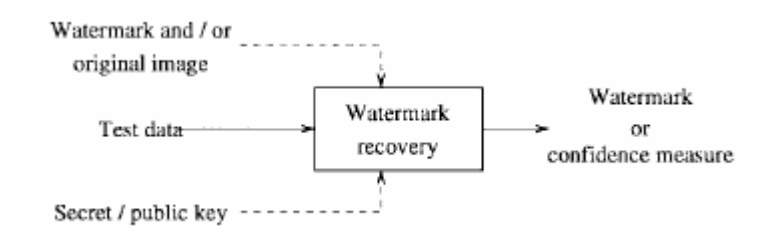


Figura 2 Proceso de verificación de marca de agua

Fuente: (Reverte & Martínez, 2004)

Las marcas deben ser detectables o extraíbles para ser útiles. Dependiendo de la forma en que la marca es insertada y dependiendo de la naturaleza del algoritmo de marcado, el método usa.

$$W' = D(Iw) \quad \text{(Ec. 1.4)}$$

En otros casos, podemos únicamente detectar cuando una marca específica está presente en la imagen. En este caso la función de decodificación D simplemente devuelve un valor de decisión Si/No acerca de la pregunta de si existe una marca determinada en los datos.

$$D(Iw, W, T) = \begin{cases} 1 & \text{si existe\_marca\_W\_en\_Iw\_con\_fuerza} > T \\ 0 & \text{en caso contrario} \end{cases} \quad \text{(Ec 1.5)}$$

La marca puede estar debilitada debido a manipulaciones sobre los datos y muchas veces no puede decirse con absoluta certeza si se halla o no presente, el valor T se conoce como threshold, y es el valor mínimo de fuerza de la marca para el que diremos que se halla presente.

Debe resaltarse que la extracción de marca puede probar la propiedad mientras que la detección solo verifica la propiedad.

2.5 TÉCNICAS DE MARCAS DE AGUA PARA DOCUMENTOS DE TEXTO

Hay diferentes clasificaciones sobre las varias características y modos de incrustación de técnicas de marcas de agua de texto. Estas técnicas son basadas en imágenes de texto, contenido, formato, funciones, sustitución de sinónimos, estructura sintáctica, siglas, nombre del verbo.

Muchos otros de los algoritmos de marca de agua de texto que dependen de varios puntos de vista (Al-Wesabi, Alsakaf, & Vasantrao, 2013).

2.5.1 Técnicas basadas en el formato

Estas técnicas están orientadas a documentos que utilizan texto con formato, donde la marca de agua tiene dependencia total del formato.

2.5.1.1 Codificación en saltos de línea

Esta técnica consiste en incrustar una marca en una página haciendo uso de los saltos de línea o el espacio de interlineado entre las líneas de la página. La marca se codifica en bits los cuales toman valores de 0 ó 1 dependiendo del desplazamiento del espacio interlineal. Además para ayudar en la decodificación se cuenta con dos líneas ya sea esto en el encabezado o en pie de página como se muestra en la figura siguiente (Brassil, Low, & Maxemchuk, 1999).

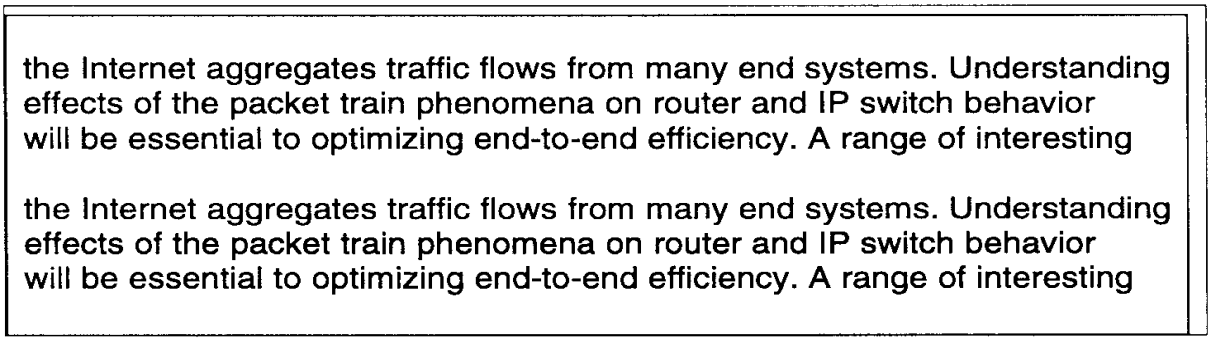


Figura 3 Técnica de interlineado

Fuente: (Brassil, Low, & Maxemchuk, 1999)

Estas líneas adyacentes inmóviles sirven como ubicaciones de referencia en el proceso de decodificación.

La mayoría de los documentos se formatean con separación uniforme entre las líneas adyacentes dentro de un párrafo. Aunque el ojo humano es particularmente hábil para darse cuenta de desviaciones de la uniformidad, nuestra experiencia sugiere que desplazamientos de líneas verticales, sean de 1/300 [pulg] y menos que pasan desapercibidos por los lectores (Brassil, Low, & Maxemchuk, 1999).

2.5.1.2 Codificación en espacio entre palabras

En esta técnica una marca se incrusta cambiando la ubicación horizontal de una palabra dentro de una línea de texto. En una implementación típica, una palabra se desplaza hacia la izquierda o derecha, mientras que las palabras inmediatamente adyacentes se dejan intactas. Estas palabras inmóviles pueden servir entonces como ubicaciones de referencia en el proceso de decodificación. Documentos con formato con texto justificado normalmente utilizan separación variable entre las palabras para distribuir el espacio en blanco de una manera agradable a la vista. Los lectores aceptan una amplia variación en la configuración de texto dentro de una línea; nuestra experiencia sugiere desplazamientos horizontales de 1/150 [pulg] y menos que fácilmente pasan desapercibidos. Dado que el espacio entre palabras en el documento original no es uniforme, la detección de una palabra requiere el conocimiento de la separación original de las palabras. Por lo tanto, las posiciones de las

palabras en el documento no marcado deben ser conocidas con el fin de extraer la información oculta. La información oculta sólo puede ser leída por la organización que posee el documento original o su agente.

2.5.1.3 Codificación en caracteres

La codificación de caracteres es la técnica que incrusta una marca mediante la alteración de una característica particular de un caracter, un cambio en la altura de un caracter o de su posición en relación con otros. Una vez más, algunos rasgos particulares no son modificados para facilitar la decodificación. Por ejemplo un algoritmo de detección puede comparar la altura de un caracter hipotéticamente alterado con la de otro no modificado del mismo caracter en otra parte de la página (Brassil, Low, & Maxemchuk, 1999).

La incrustación imperceptible de una marca por alteración de caracter requiere a menudo extremo cuidado para alterar el caracter. Es más probable que el lector observe la alteración del caracter si es idéntico. La detección de la presencia o ausencia de una marca puede o no requerir información de la imagen original, dependiendo de la técnica de marcado y la regla de selección de los caracteres que están alterados.

2.5.2 Técnica basada en el contenido

Técnicas de marcas de agua de texto basadas en el contenido son basadas en la estructura del lenguaje natural. Esta técnica utiliza la estructura sintáctica del texto del contenido para la incorporación de los bits de marca de agua, se realizan transformaciones sintácticas en un árbol sintáctico teniendo en cuenta las propiedades del lenguaje natural. También se puede sustituir sinónimos, esta técnica no altera la semántica del documento. (Brassil, Low, & Maxemchuk, 1999).

2.5.2.1 Enfoque sintáctico

En este enfoque orientado hacia la marca de agua en texto, la estructura sintáctica del texto se utiliza para incrustar la marca de agua donde se construye un árbol sintáctico y se

aplican transformaciones a ella para poder insertar la marca de agua manteniendo todas las propiedades del texto. Las técnicas de NLP[2] se utilizan para analizar la estructura sintáctica y semántica del texto en el desempeño de cualquier transformación para integrar los bits de marca de agua (Jalil, Mirza, & Sabir, 2010).

Según Jalil, Mirza, & Sabir(2010) las marcas de agua para texto que utilizan estructura sintáctica, utilizan algoritmos de procesamiento de lenguaje natural, estos son un enfoque eficaz para la protección de autenticidad del texto y protección del derecho de autor, pero el progreso en este ámbito es más lento. El NLP es un área de investigación inmadura hasta el momento, esto por el uso de algoritmos ineficientes, por lo que no se puede obtener resultados eficientes en marcas de agua para texto.

2.5.2.2 Enfoque semántico

Este enfoque consiste en incrustar la marca de agua binaria, reemplazando algunas palabras con sus sinónimos.

Las marcas de agua de texto, basadas en la semántica dependen del idioma. Las técnicas basadas en el reemplazo de sinónimos no son resistentes a los ataques de sustitución al azar. La naturaleza sensible de algunos documentos, por ejemplo, documentos legales, la poesía y citas no nos permiten realizar transformaciones semánticas al azar debido a que en estas formas de texto una transformación sencilla a veces destruye tanto la connotación semántica y el valor de texto (Jalil, Mirza, & Sabir, 2010).

2.5.2.3 Técnica basada en la estructura

Este es el enfoque más reciente utilizado para la protección de derechos de autor de los documentos de texto. Un algoritmo de marca de agua de texto para la protección de los derechos de autor del texto usando las ocurrencias de letras dobles (aa-zz) para insertar la marca de agua. El algoritmo es una mezcla de cifrado, la esteganografía y marcas de agua.

[2] NLP (Natural Language Processing) procesado de lenguaje natural, esta técnicas se basa en el lenguaje natural.

Sin embargo, los grupos se forman mediante el uso del período de parada completa[3] en este algoritmo. Los textos como la poesía, citas, contenidos web, documentos legales pueden no contener esencialmente separadores de oraciones como el punto y punto aparte; lo que hace que este algoritmo no sea aplicable a todos los tipos de texto. Para superar las deficiencias de este algoritmo, hay otro algoritmo que utiliza las preposiciones de las letras dobles para generar la marca de agua (Jalil, Mirza, & Sabir, 2010).

2.5.3 Técnica basada en imágenes binarias

Esta técnica utiliza imágenes, las cuales son la representación del documento de texto. Para incrustar la marca que es una serie de bits se determina el bit menos significativo LBS, para determinar el LBS se realizan transformaciones las cuales pueden ser: Transformadas discretas de Coseno DTC o Transformadas discretas de Wavelet DWT (Al-Wesabi, Alsakaf, & Vasantrao, 2013).

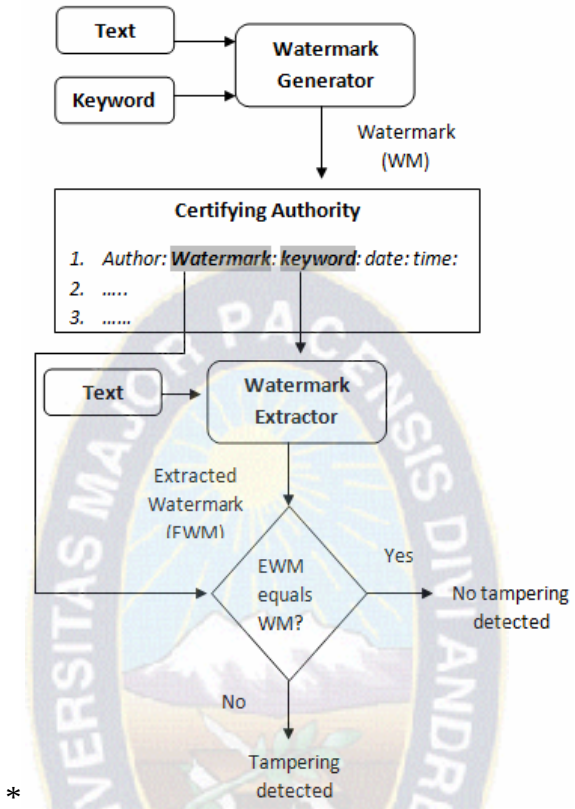
2.5.4 Técnica Zero-Watermarking

Las técnicas de marcas de agua basadas en el contenido como ser semántica y sintáctica incrustan una marca de agua en el documento de texto en sí, que influye en la calidad del texto, significado y degradan el valor del mismo. Debido a esto se propone la técnica Zero-Watermarking, la cual propone un enfoque de cero marcas de agua de manera que el documento de texto de acogida no se altera para incrustar la marca de agua, más bien se utilizan las características del texto para generar una marca de agua. La marca de agua es frágil dependiendo del algoritmo de codificación y se utiliza para autenticar documentos de texto (Jalil, Mirza, & Sabir, 2010).

La generación de marca de agua y el proceso de extracción se ilustra en la figura 4, en la cual se tiene el texto original, donde este se mantiene intacto y con la palabra clave se genera la marca de agua, por formalidad esta se ha registrado en la Autoridad de

[3] El periodo de parada completa, se refiere a las ocurrencias de signos de puntuación en el idioma inglés como ser el “.”, “,” y “;”de los párrafos que contienen letras dobles aa – zz.

Certificación (CA) y se utiliza el algoritmo de extracción para autenticar documentos de texto (Jalil, Mirza, & Sabir, 2010).



*

Figura 4 Técnica Zero-Watermarking

Fuente: (Jalil, Mirza, & Sabir, 2010)

2.6 MARCAS DE AGUA EN PAPEL

Este tipo de marca de agua se basa en el diminuto tamaño de los puntos que cubren menos de una milésima parte de la página, junto con su combinación de colores de amarillo sobre fondo blanco, los hace invisibles a simple vista. Una forma de determinar su presencia es usando en laser o una lupa, pero también se puede visualizar con la lente de una cámara utilizando el zoom (Tuohey, 2004).

Muchos otros tipos de técnicas de esteganografía se han inventado y utilizado después de la invención de la imprenta. Algunos ejemplos incluyen tinta invisible impresas en papel, micro puntos fotográficos, escritos en seda, etc. (Menvielle, et al., 2011).

La tinta invisible ha estado en uso durante siglos para la diversión de los niños y estudiantes y para el espionaje por parte de espías y terroristas. Los micro puntos y el microfilm, elementos básicos de la guerra y películas de espionaje, se introdujeron tras la invención de la fotografía (Menvielle, et al., 2011).

2.7 AUTENTICACIÓN DE DOCUMENTOS DIGITALES

El término "auténtico" se utiliza para indicar que un documento es verdadero. Sin embargo, quizá resulte engañoso utilizar el término "auténtico" al referirse a un documento digital, o, más precisamente, a un objeto digital. Esto se explica por la manera en que se crean y se preparan los objetos digitales para su visualización. Para que los datos digitales resulten inteligibles para los seres humanos, deben interpretarse. Los datos digitales se procesan mediante secuencias de comandos, por lo que un simple documento con texto escrito consistirá, por ejemplo, en varios códigos de caracteres ASCII que deben interpretarse antes de que el texto se reproduzca en la pantalla en un formato legible para las personas. Sin embargo, los datos digitales no se limitan a meros documentos de texto. El formato de los datos puede ser más elaborado, incluyendo componentes activos, como macros y lenguajes de programación, por lo que los datos pueden precisar una interpretación más compleja para que pueda leerse el texto. Además, los archivos que se abren en ordenadores distintos al que originariamente los creó pueden generar, y a menudo generan, diferentes fuentes y saltos de línea. Esta es la razón por la que difiere el formato de los archivos de documentos (Mason, 2004).

2.7.2 Autenticación

Autenticación es el acto de establecimiento o confirmación de algo (o alguien) como auténtico. La autenticación de un objeto puede significar (pensar) la confirmación de su procedencia, mientras que la autenticación de una persona a menudo consiste en verificar su identidad. La autenticación depende de uno o varios factores (Müller , 2010).

2.7.3 Autenticación con autoridad de certificación

Una de las técnicas de autenticación más conocidas en el ambiente informático es la infraestructura de clave pública (PKI) cuyos componentes son:

La autoridad de certificación (o, en inglés, CA, Certificate Authority): es la encargada de emitir y revocar certificados. Es la entidad de confianza que da legitimidad a la relación de una clave pública con la identidad de un usuario o servicio.

La autoridad de registro (o, en inglés, RA, Registration Authority): es la responsable de verificar el enlace entre los certificados (concretamente, entre la clave pública del certificado) y la identidad de sus titulares.

Los repositorios: son las estructuras encargadas de almacenar la información relativa a la PKI. Los dos repositorios más importantes son el repositorio de certificados y el repositorio de listas de revocación de certificados. En una lista de revocación de certificados (o, en inglés, CRL, Certificate Revocation List) se incluyen todos aquellos certificados que por algún motivo han dejado de ser válidos antes de la fecha establecida dentro del mismo certificado (Wikipedia, 2013).

Los usuarios y entidades finales son aquellos que poseen un par de claves (pública y privada) y un certificado asociado a su clave pública. Utilizan un conjunto de aplicaciones que hacen uso de la tecnología PKI (para validar firmas digitales, cifrar documentos para otros usuarios, etc.).

2.7.3.1 Certificado digital

En la actualidad casi todas las aplicaciones de comercio electrónico y transacciones seguras requieren un certificado digital, se ha propagado tanto su uso que se tiene ya un formato estándar de certificado digital, este es conocido como X509 v.3 Algunos de los datos más importantes de este formato son los siguientes:

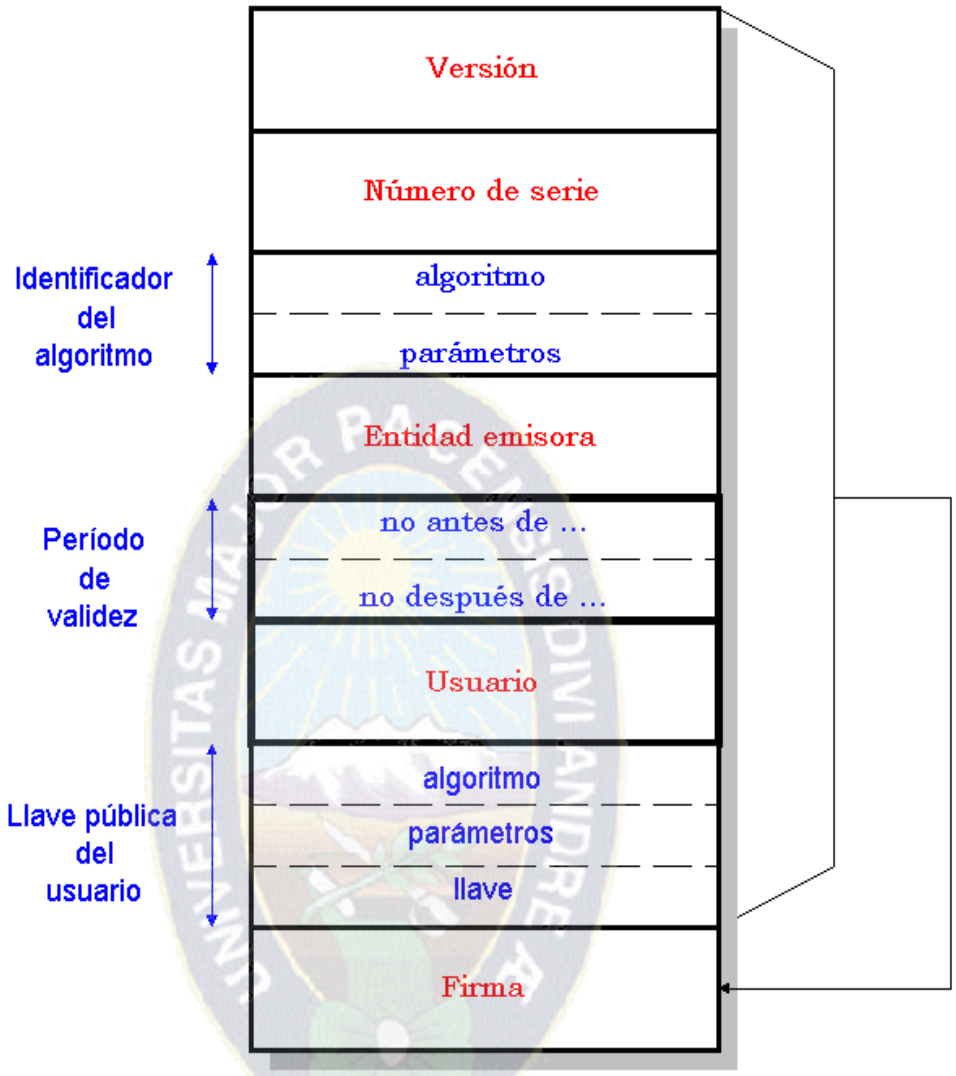


Figura 5 Estructura de un certificado digital

Fuente: (Müller , 2010)

2.7.4 Características de la autenticación

Cualquier sistema de identificación ha de poseer unas determinadas características para ser viable:

• Ha de ser fiable con una probabilidad muy elevada (podemos hablar de tasas de fallo en los sistemas menos seguros).

• Económicamente factible para la organización (si su precio es superior al valor de lo que se intenta proteger, tenemos un sistema incorrecto).

• Soportar con éxito cierto tipo de ataques.

• Ser aceptable para los usuarios, que serán al fin y al cabo quienes lo utilicen.

2.7.5 Documento digital

Un documento digital tiene información codificada en bits y para leer, visualizar o grabar la información se precisa de un dispositivo que transmita o grabe información codificada en bits. Al representarse digitalmente, los datos de entrada son convertidos en dígitos (0,1) inteligibles para la máquina y no para los sentidos humanos (Delgado Gómez, 2009).

2.7.5.1 Documento de texto plano

El texto plano es un término general que incluye cualquier material escrito, no el texto codificado, en cualquier idioma. Los e-mails, carteles, documentos de procesadores de texto, código de lengua, equipo de origen y las páginas web, son todos ejemplos de documentos de texto plano. Los datos digitales de inserción en los documentos de texto plano, se pueden lograr a través de la manipulación de las propiedades de la lengua, como las partes de la sustitución de expresión (Menvielle, et al., 2011).

Un archivo de texto llano, texto simple, texto plano, texto sencillo o texto pelado (en inglés plain text) es un archivo informático compuesto únicamente por texto sin formato, sólo caracteres, lo que lo hace también legible por humanos. Estos caracteres se pueden codificar de distintos modos dependiendo de la lengua usada. Algunos de los sistemas de codificación más usados son: ASCII, ISO-8859-1 o Latín-1 y UTF-8. (Müller , 2010).

2.8 TECNOLOGÍAS DE SOFTWARE

2.8.1 Tesseract OCR

Tesseract es un motor OCR[4] libre. Fue desarrollado originalmente por Hewlett Packard como software propietario entre 1985 y 1995. Tras diez años sin ningún desarrollo, fue liberado como código abierto en el año 2005 por Hewlett Packard y la Universidad de Nevada, Las Vegas. Tesseract es desarrollado actualmente por Google y distribuido bajo la licencia Apache, versión 2.0 (Wikipedia, 2013)

Tesseract está considerado como uno de los motores OCR libres con mayor precisión disponible actualmente.

En 1995, Tesseract era uno de los tres mejores motores OCR en cuanto a precisión, además está disponible para Linux, Windows y Mac OS X, sin embargo, sólo ha sido probado por los desarrolladores en Windows y Ubuntu. Hasta la versión 2, Tesseract sólo podía aceptar como entrada imágenes de una sola columna en formato TIFF. En estas primeras versiones no se incluía análisis de patrones, y por tanto, las imágenes con múltiples columnas o anotaciones producían resultados ilegibles. Desde la versión 3, Tesseract soporta el formato en el texto y el análisis del patrón de la página. Combinado con la biblioteca Leptonica[5] se consigue la compatibilidad con nuevos formatos de imagen, además se puede detectar si el texto es proporcional o monoespaciado. Tesseract puede procesar 60 idiomas entre los que se destacan el inglés, francés, italiano, alemán, español, portugués brasileño y neerlandés, y puede ser entrenado para funcionar con otros idiomas (Google Inc., 2005).

2.8.2 Android

Android es un sistema operativo basado en el kernel de Linux diseñado principalmente para dispositivos móviles con pantalla táctil, como teléfonos inteligentes o tabletas,

[4] Reconocimiento óptico de Caracteres por sus siglas en inglés OCR.

[5] Bibliteca Leptonica (Leptonica Image Processing Library), Es un sitio que provee software de código abierto para procesamiento y análisis de imágenes para más información ver: http://leptonica.com/

inicialmente desarrollado por Android Inc., Google Inc. respaldó económicamente y más tarde compró esta empresa en 2005. Android fue presentado en 2007 junto la fundación del Open Handset Alliance: un consorcio de compañías de hardware, software y telecomunicaciones para avanzar en los estándares abiertos de los dispositivos móviles (Wikipedia, 2013).

CAPÍTULO III: DISEÑO METODOLÓGICO

3.1 ESTRUCTURA DE DESARROLLO

Para desarrollar el presente trabajo se hace un análisis de las diferentes técnicas de watermarking para documentos de texto, se toman en cuenta las características particulares de estas técnicas, considerando las ventajas y desventajas de las mismas. Con el análisis de estas técnicas se propone una para cumplir con los requerimientos del presente trabajo.

Para cumplir satisfactoriamente con los objetivos se hace uso de la metodología de investigación exploratoria, además es necesario estructurar el desarrollo en una serie de fases. Estas fases se organizan de la siguiente manera:

Fase 1: Análisis de técnicas de watermarking.

Fase 2: Diseño de las marcas de agua.

Fase 3: Diseño de certificado de marca de agua.

Fase 4: Implementación del prototipo.

En la primera fase se hace un análisis de las ventajas y desventajas de las técnicas de marcas de agua conocidas luego de este análisis se propone la técnica para cumplir el objetivo del trabajo, el cual es establecer la técnica de marcas de agua para la protección de derechos de autor además de poder autenticar el contenido de un documento mediante su marca de agua.

En la segunda fase se hace el diseño de las marcas de agua utilizando la técnica seleccionada en la fase anterior para solucionar los problemas planteados. Además esta considera a los documentos en medio digital e impreso en papel, para este propósito se propone contar con una Autoridad de Certificación, la cual debe brindar un servicio que permita hacer registros de los documentos de texto y los datos del autor o dueño del mismo, también debe ofrecer los servicios para extraer texto plano del documento para así generar sus marcas de agua correspondiente, las cual deben almacenarse en un certificado de marca de agua.

La tercera fase, describe las características del certificado de marcas de agua propuesta y la figura de una Autoridad Certificadora.

La cuarta fase, describe las características principales del prototipo, de estas podemos mencionar su arquitectura de tipo cliente servidor. El servidor web es implementado por la Autoridad de Certificación el cual ofrece un servicio para realizar la autenticación de los documentos de texto.

3.2 ANÁLISIS DE TÉCNICAS DE WATERMARKING

La técnica de marca de agua propuesta para el presente trabajo debe permitir resguardar el derecho de autor sobre un documento de texto como ser: trabajo de investigación, página de un libro, y todo los documentos de texto sobre los cuales se reclamará derecho de autoría; a la vez debe permitir verificar la autenticidad del contenido de las palabras del texto asociado, entendiéndose por texto asociado al texto que se extrae de un documento ya sea este electrónico o físico (impreso en papel), para extraer el texto asociado se debe someter a un proceso para extraer texto sin formato o texto plano independiente de la extensión del archivo o documento electrónico.

En esta fase se trata de hacer un análisis simple de las principales características de estas técnicas.

3.2.1 Técnica basada en formatos

Esta técnica tiene dependencia directa del formato.

Características:

- El tipo de marca de agua soportado es de tipo binario.
- Para decodificación es necesario contar con el documento original.
- Trabaja con las posiciones de las palabras para codificar un bit.
- Esta técnica es derivada directamente de la esteganografía.
- No soporta texto plano.

- No cuenta con mecanismo de protección frente a la copia del contenido.

3.2.2 Técnica basada en contenido

Esta técnica se basa en la característica particular de un idioma.

- El tipo de marca de agua puede ser binario o caracteres.
- La codificación se basa en el sentido de una oración.
- Altera el contenido para crear transformaciones para la inserción de la marca de agua.
- No altera el sentido del mensaje o texto.
- Para la decodificación no es necesario el documento original.
- Altera algunas palabras, signos de puntuación.

3.2.3 Técnica basada en imagen binaria

El documento de texto se almacena como si fuera imagen con extensión de imagen como ser .jpg, .tif u otro.

A la imagen se aplican las técnicas usuales para imágenes como las transformadas, aprovechando las características de redundancia que tienen las imágenes dependiendo de la resolución de imagen se puede insertar la marca de agua binaria que representa la información del autor del documento.

- Tipo de marca de agua conjunto de bits que representa un mensaje o información del dueño y autor del documento.
- Requiere mucho espacio para almacenamiento.
- Complejidad alta para implementación.
- No es robusta frente a las operaciones de transcripción y reconocimiento óptico de caracteres OCR.

3.2.4 Técnica Zero-Watermarking

La técnica no inserta marca de agua en el documento de texto original, pero se basa en el contenido de este para generar la marca de agua, la cual se registra en una Autoridad de certificación para formalidades, entre sus principales características se destacan:

- Adecuado para texto plano.
- Robustez frente a operaciones de copia.
- Adecuado para autenticación de contenido de texto.
- No altera el documento de texto.
- Se implementa según requerimientos de robustez y fragilidad.

3.2.5 Comparación de técnicas de marcas de agua

La comparación de las técnicas descritas (en el orden presentado) se realiza en la siguiente tabla según la manipulación del documento de texto, se entiende por manipulación a los cambios o modificaciones (reducción o ampliación de tamaño y copia de texto) que no degradan el texto y la recuperación de la marca.

Técnicas	Manipulación
T1	No
T2	Si
T3	No
T4	**Si**

Tabla 1 Técnicas de marcas de agua vs manipulación de texto

Fuente: Elaboración propia

En la tabla 2 se hace una comparación para soporte de las técnicas a los tipos de documentos físicos y electrónicos como se propone en el presente trabajo. La técnica es buena si persiste en ambos medios.

Técnicas	Físico (impreso en papel)	Digital (texto plano)
T1	No	No
T2	Si	Si
T3	No	No
T4	**Si**	**Si**

Tabla 2 Técnicas vs documentos

Fuente: Elaboración propia

En la siguiente tabla se hace una comparación de las técnicas vs dependencia del idioma e integridad del mensaje. Para aplicabilidad las técnicas no deben ser dependientes de un lenguaje en particular, además deben conservar la integridad del contenido de texto.

Técnicas	Dependencia de lenguaje	Integridad de texto
T1	No	Si
T2	Si	No
T3	No	No
T4	**No**	**Si**

Tabla 3 Técnicas vs dependencia de lenguaje e integridad

Fuente: Elaboración propia

Luego de este pequeño análisis se opta por la técnica Zero-Watermarking la cual se adecua a los objetivos planteados. Con la técnica elegida se pueden implementar diferentes algoritmos para generar las marcas de agua, las cuales pueden tener diferentes grados de robustez y fragilidad.

3.3 DISEÑO DE LAS MARCAS DE AGUA

Para solucionar el problema planteado, es necesario diseñar dos marcas de agua una debe ser robusta y la otra frágil, la primera es para proteger el derecho de autor sobre el documento de texto, mientras la segunda es para verificar la autenticidad del contenido de texto.

Para el diseño de las marcas de agua es necesario obtener el texto plano del documento, con este se procede a generar las marcas de agua. La obtención de texto plano involucra un proceso como se ilustra en la figura 6.

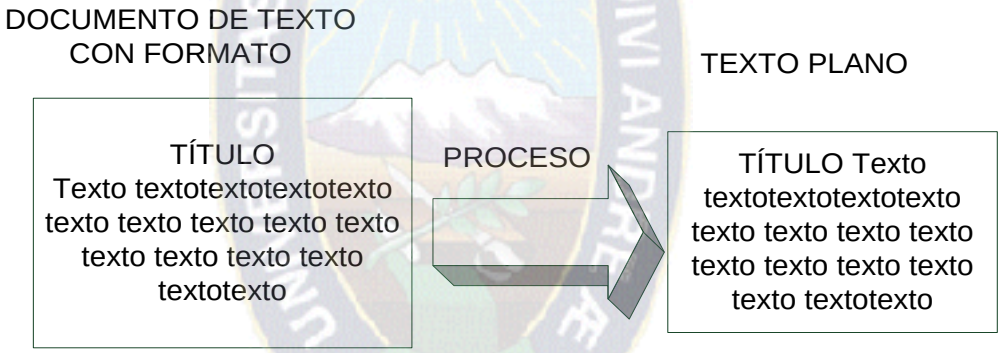


Figura 6 Proceso extracción de texto plano

Fuente: Elaboración propia

El proceso de extracción de texto consiste en extraer el texto de un documento ya sea este impreso o digital independiente de su extensión, la entrada para el proceso es el documento de texto y la salida es una cadena o texto sin formato al cual se le puede aplicar una técnica apropiada de marca de agua.

3.3.1 Algoritmo propuesto para generar marca de agua robusta

El proceso de generación de marcas de agua una vez obtenido el texto, es como se muestra en la figura siguiente que ilustra el flujo de los procesos, en el cual primero se

aplica una función o algoritmo denotado por F el cual genera la marca de agua para el texto T.

De forma general el proceso para generar la marca de agua tiene como entrada el texto original To y la palabra clave K1, el proceso devuelve como parámetro de salida la marca de agua generada para K1, como se muestra en la figura de abajo.

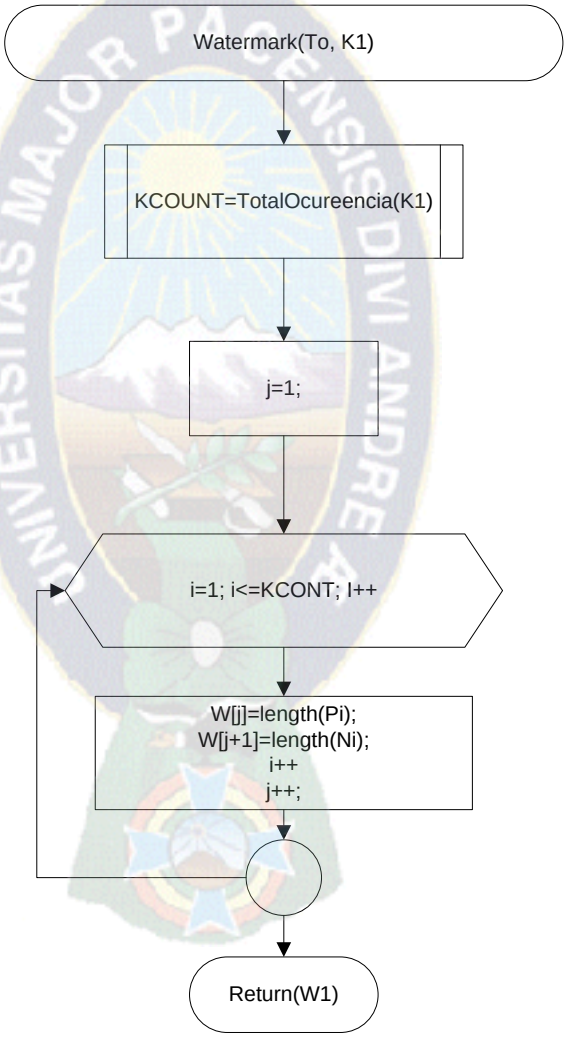


Figura 7 Proceso para generar marca de agua

Fuente: Elaboración propia

Donde los símbolos de la notación corresponden como sigue:

To= Texto original.

K1= Palabra clave.

KCOUNT= Número de ocurrencias de la palabra clave.

W1= Watermark o marca de agua robusta.

Pi= ‘Palabra anterior’ de esa ocurrencia de la palabra clave (K1).

Ni= ‘Siguiente palabra’ de esa ocurrencia de la palabra clave (K1).

La propuesta para generar la marca de agua es encontrar la palabra con más frecuencia del documento, ya que la palabra con mayor frecuencia refleja las características propias del texto. Pero este algoritmo se puede extender utilizando la clave K1, para detectar o clasificar documentos con marcas de agua similares registrados en la Autoridad Certificadora.

Una vez obtenida la palabra con más frecuencia en el texto, se procede a generar la marca de agua donde se registra la longitud de la palabra predecesora y la palabra sucesora de cada una de las ocurrencias secuencialmente separado por puntos.

3.3.2 Algoritmo propuesto para generar marca de agua frágil

Para la autenticación del documento se propone una segunda marca de agua frágil, misma que ante una pequeña alteración de una de las palabras debe quebrarse. En la investigación realizada, una de las funciones que tiene este comportamiento es el hash, esta cuenta con soporte suficiente de teoría y hay varias funciones que tienen un comportamiento parecido, de estas elegimos una en particular la cual es la función MD5.

3.3.2.1 Proceso de normalización de texto

Este proceso consiste en preparar el mensaje de texto del documento para generar las marcas de agua correspondientes, debido a que las imágenes tomadas con un Smartphone dependen de varios factores a mejorar que por el momento están fuera del alcance del

presente trabajo, lo cual repercute directamente en el procesamiento OCR de la imagen capturada por la cámara, para mejorar el proceso de reconocimiento de texto se propone normalizar el texto de un documento creando un contenido paralelo, donde este ignora los signos de puntuación y acento en las palabras, además para se colocan todas las palabras en mayúscula. Al texto normalizado se le aplican las técnicas de marcas de agua este proceso se ilustra en la siguiente figura.

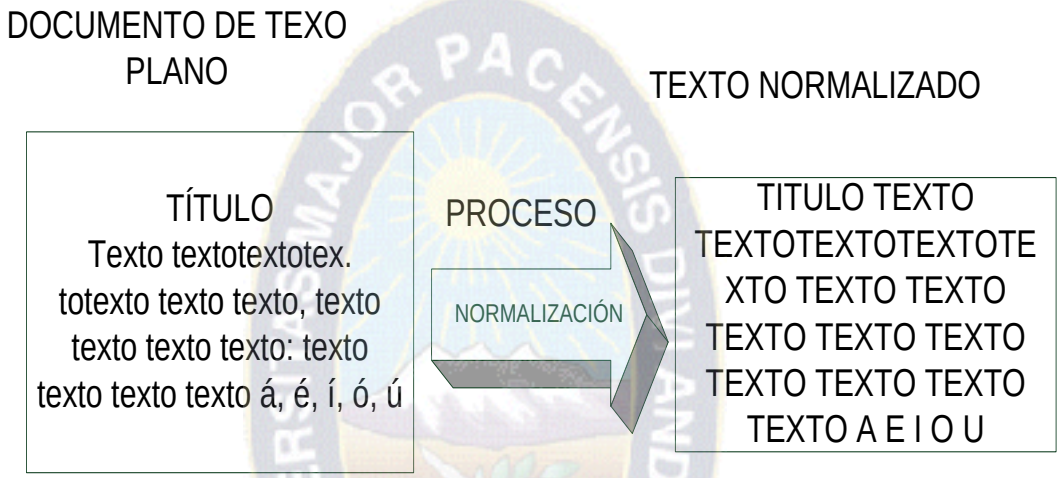


Figura 8 Proceso extracción de texto plano

Fuente: Elaboración propia

Una vez realizado el proceso de normalización de texto, se procede a diseñar el algoritmo que genera la marca de agua frágil, en la figura siguiente se muestra el algoritmo que consiste en aplicar la función hash MD5 al texto normalizado.

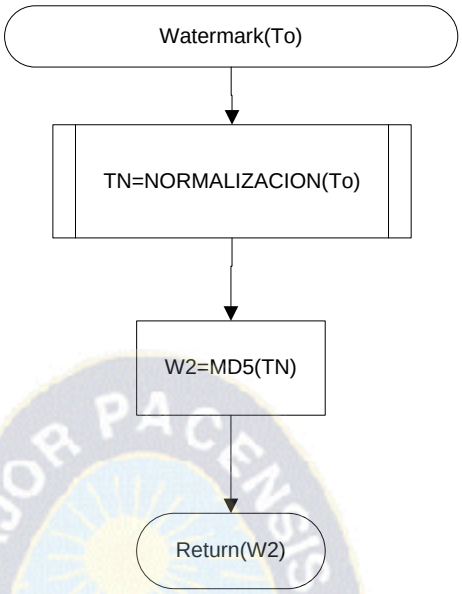


Figura 9 Proceso para generar marca de agua

Fuente: Elaboración propia

To= Texto original.

TN= Texto normalizado.

W2= Marca de agua frágil.

3.3.3 Extracción de la marca de agua y autenticación

El algoritmo que extrae la marca de agua a partir del texto es llamado algoritmo de extracción, este toma como parámetro de entrada el texto recuperado Tr y con ella se obtiene la palabra clave Kr1. El texto puede ser atacado o no. La marca de agua se genera del texto recuperado y Kr1 con el algoritmo de extracción a continuación se compara con la marca de agua original registrado en la Autoridad de Certificación.

Para la extracción de la marca de agua de un documento Tr primero se normaliza su texto plano, a partir del texto plano volvemos a generar la marca de agua robusta denotada por Wr1, luego de generar la marca buscamos en la base de datos de la Autoridad de

Certificación si hay algún documento con la misma marca W1, si es así (Wr1=W1) y no hay conflicto decidimos comparar la segunda marca de agua frágil Wr2 con la ya registrada W2, si estas son iguales revelamos el documento como auténtico. En caso de existir algún conflicto sobre la marca de agua W1 y evitar conflicto de documentos similares se agregan restricciones o reglas para decidir la autoría de un documento en base a ellas, estas reglas o restricciones se registrarán en el Certificado de marca de agua digital.

Para la detección de la marca de agua tomamos como entrada el documento recuperado denotado por Tr, de este detectamos la palabra con mayor frecuencia para establecer como clave Kr1 y con ella generamos la marca de agua Wr1, buscamos Wr1 en la base de datos si la búsqueda es exitosa realizamos el proceso de autenticación.

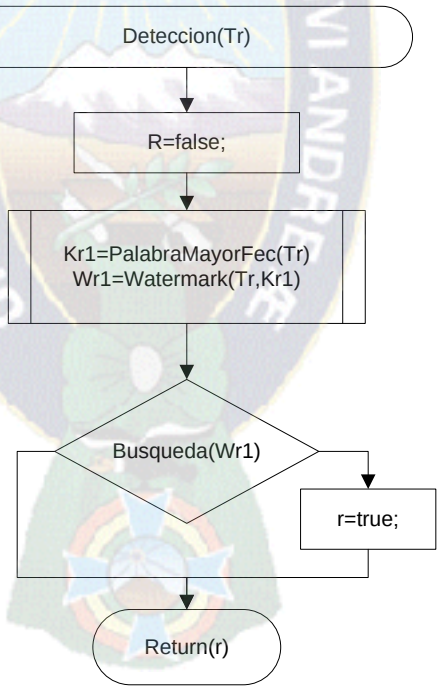


Figura 10 Proceso para detectar marca de agua

Fuente: Elaboración propia

Luego del proceso de detección, si esta resulta positiva se procede a realizar el proceso de autenticación del documento, el cual consiste en generar la segunda marca de agua Wr2 y

comparar esta con la registrada en la base de datos, este proceso se ilustra en la figura siguiente.

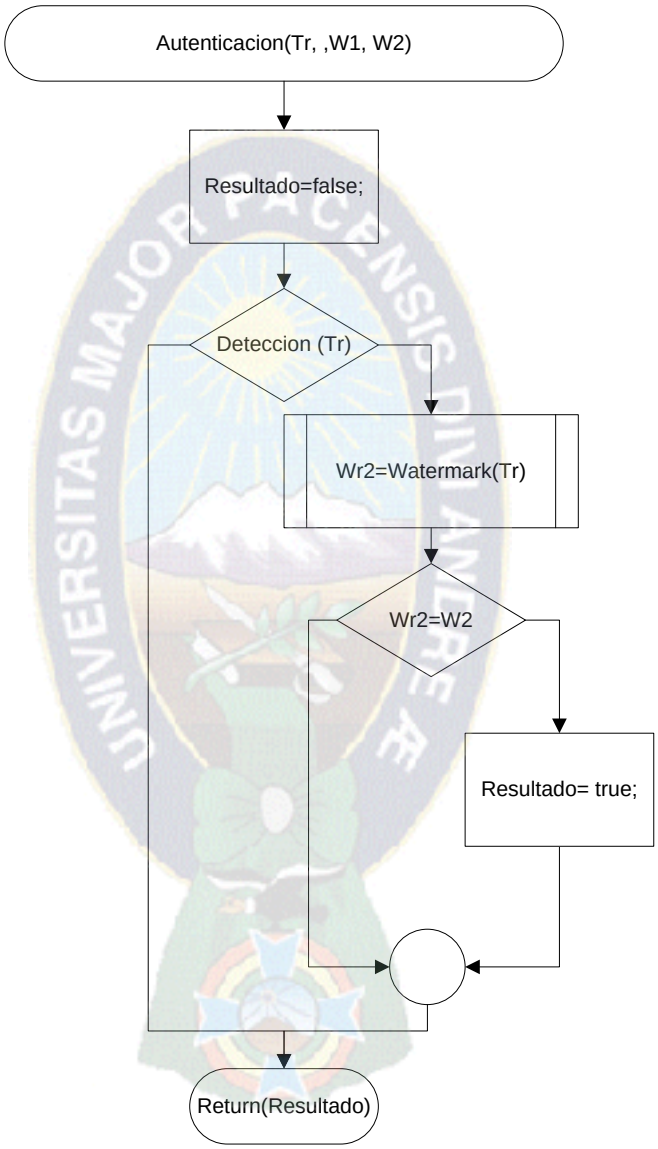


Figura 11 Proceso de autenticación

Fuente: Elaboración propia

Tr= Texto recuperado.

W1 y W2= Marca de agua del Texto original.

Wr1 = Marca de agua robusta recuperada

Wr2= Marca de agua frágil recuperada.

Resultado = Valor booleano que decide la autenticidad de un documento.

El proceso de autenticación retorna como parámetro de salida en la variable Resultado el valor booleano, con este resultado se decide si el contenido del documento es auténtico.

Para decidir si una marca de agua extraída presenta algún conflicto se procede a considerar las reglas de restricción, con ellas se procede a decidir de cual autor es realmente el documento recuperado, de esta forma resolver el problema, si más de un dueño reclama la autoría. Las reglas o restricciones se registran al momento del registro del documento donde se detectó si algún documento tiene una marca de agua similar, para así no tener conflictos al momento de la autenticación.

3.3.4 Autenticación de texto impreso en papel

En el proceso de la investigación se cuenta con técnicas de marcas de agua relacionadas con tintas invisibles y micro-puntos de color amarillo las cuales no se aplican en el presente trabajo.

Para adecuar la técnica propuesta, se propone considerar la idea de reconocer el texto existente en el documento mediante el procesado de imágenes. Donde se toma la imagen del texto y así poder extraer su texto con una herramienta OCR.

La herramienta OCR nos permitirá reconocer los caracteres impresos en el documento en papel, mediante el reconocimiento de caracteres se puede reconstruir las palabras que contiene el documento impreso en el mismo orden que fueron impresas, al recuperar el texto completo del documento impreso, se extrae el texto plano, al cual se puede aplicar la técnica de generación de marca de agua propuesta.

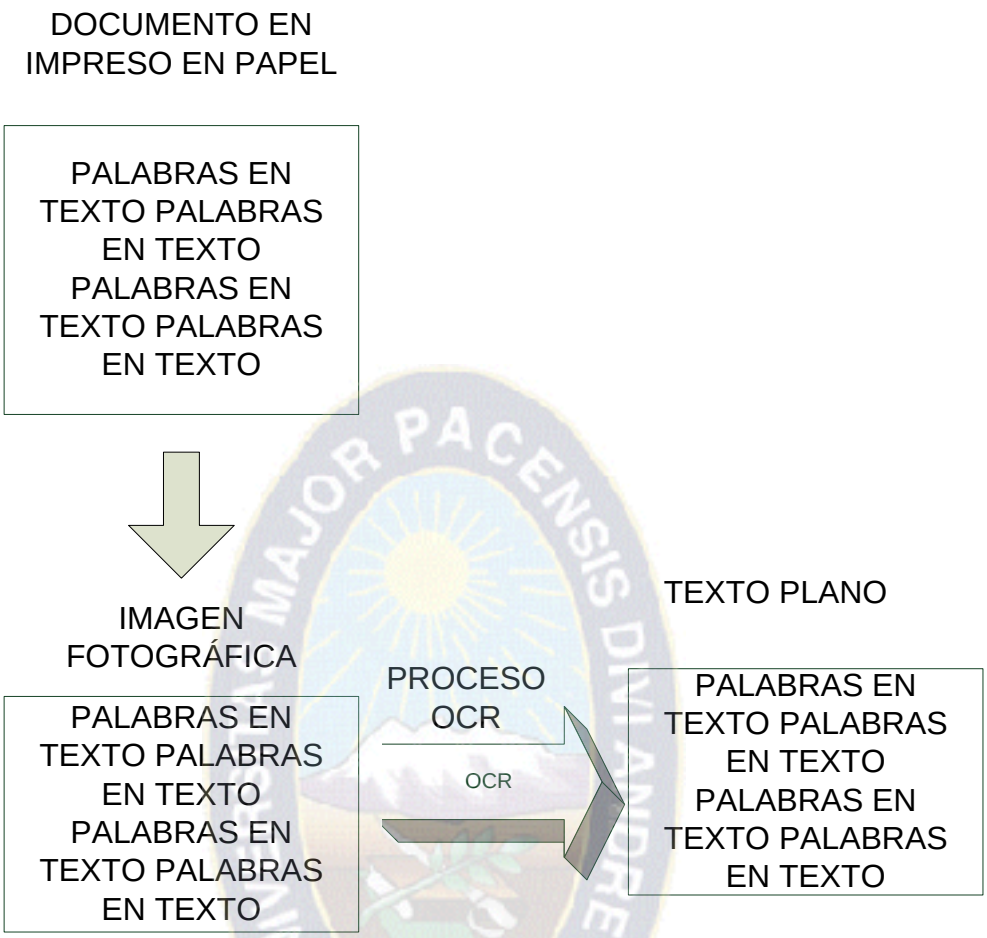


Figura 12 Proceso de extracción de texto de papel

Fuente: Elaboración propia

El proceso de la toma de la imagen fotográfica, puede realizarse mediante una cámara o escáner siempre y cuando se conserve la calidad de la imagen, ya que a esta se le procesará con un motor o herramienta OCR.

El proceso de OCR tiene por lo general como parámetro de entrada una imagen de texto y devuelve como parámetro de salida un texto en formato txt, al cual se le puede tratar como un documento de texto normal y aplicar las técnicas de marcas de agua apropiadas según lo requiera el trabajo, para cumplir con los objetivos trazados se enfoca en extraer el texto plano y a ella aplicar la técnica de marca de agua propuesta, la cual trabaja con las longitudes de las palabra vecinas de la palabra clave.

3.4 DISEÑO DE CERTIFICADO DE MARCA DE AGUA

Una vez obtenida las marcas de agua, se procede a realizar el registro de los mismos en una Autoridad de Certificación que es un tercero de confianza que acredita la validez o autenticidad de los documentos.

La Autoridad de Certificación cuenta con un proceso para realizar el registro de los datos del autor de un documento, también con el proceso para normalizar el texto y con ello generar las respectivas marcas de agua.

Figura 13 Proceso de generación de la marca de agua

Fuente: Elaboración propia

3.4.1 Registro de la marca de agua

Las marcas de agua W1 y W2 se registran en una Autoridad de Certificación, donde esta autoridad es un tercero de confianza cuya figura es reconocida, es donde se realiza el proceso de registro del documento de texto.

El proceso de registro se realiza una vez ejecutados los proceso de extracción de texto plano y generación de marca de agua para ese texto plano, en el proceso de registro se toman en cuenta los datos del autor, datos del documento, el texto plano extraído, la palabra clave K1, las marcas de agua W1 y W2, la fecha de registro, hora de registro y el algoritmo aplicado como se muestra en la figura siguiente.

Versión: 1,2 o 3
Número de Serie: 0000000….
Emisor del Certificado: VeriMex
Marca de agua 1: x.x.x.x.x.x.x....
Marca de agua 2: yyyyyyyyyy.....
Identificador del Algoritmo usado en el watermarking: xxxx
Sujeto: Autor del documento
Datos del documento: yyy
Fecha de registro:
Hora:
Algunos datos opcionales:
Periodo de Validez: De Enero 2014 a Dic 2015
Firma de la Autoridad Certificadora

Figura 14 Diseño de certificado propuesto

Fuente: Elaboración propia

Los datos del proceso de registro se almacenan en un certificado que tiene una estructura similar a los certificados de firma digital. Los certificados aparte de guardar los datos del autor de un documento sirven para el proceso de autenticación y protección del derecho de autoría de dicho documento de texto.

3.5 IMPLEMENTACIÓN DEL PROTOTIPO

Para la implementación del prototipo, se propone una arquitectura cliente servidor, donde el servidor está a cargo de la Autoridad de Certificación y el cliente es una aplicación desarrollada en el sistema operativo Android.

El administrador es uno de los funcionarios dependientes de la Autoridad de Certificación, mientras el usuario es un usuario que tiene una cuenta en el servicio implementado por la Autoridad de Certificación.

El usuario tiene acceso limitado, ya que accede solo a las tareas que permite la aplicación implementada para este propósito.

Figura 15 Arquitectura del prototipo

Fuente: Elaboración propia

Para hacer una descripción de las principales características del prototipo lo explicaremos en dos secciones, la primera se enfocará en los servicios que implementa la Autoridad de Certificación y la segunda describirá las características de la aplicación cliente.

3.5.1 Implementación de los servicios para Autoridad de Certificación

Lo básico que debe tener una Autoridad de Certificación es contar con un servidor ya sea esto en la nube o uno físico, el cual debe contar con un servicio para atender peticiones vía web.

Componente	Versión
Apache	2.x.y
MySQL	5.x.y
PHP	5.x.y
PhpMyAdmin	3.x.y
Tesseract–OCR	3.xy

Tabla 4 Componentes mínimos para servidor

Fuente: Elaboración propia

Los componentes con los que debe contar el servidor de la Autoridad de certificación se muestran en la tabla, se aconseja que sean las versiones más recientes y estables. Se eligen estos componentes por su tipo de licencia y porque son multiplataforma respecto del sistema operativo.

Con estos componentes se pretende crear un servicio web para atender las peticiones de la aplicación cliente, y dar el servicio de registro de documentos.

Para el funcionamiento de la herramienta Tesserat-OCR en el servidor se tiene que tener el servicio de Apache 2.x o superior funcionando en el servidor y también el lenguaje de programación PHP, y tener activado el módulo Open SSL.

3.5.1.1 Registro de documentos de texto

Este servicio consiste en registrar en la base de datos implementada en el gestor MySQL, los datos del documento, el autor, la marca de agua generada, la técnica o algoritmo utilizado, en caso de existir marcas de agua similar se procede a registrar reglas o restricciones.

El proceso de registro, es un simple formulario donde se llenan los datos exigidos en los campos que presenta el formulario.

3.5.1.2 Autenticación de documento impreso

Este servicio es un script elaborado en el lenguaje de programación PHP, tiene como parámetro de entrada la imagen que recibe de la aplicación cliente, la imagen se procesa con el motor OCR de la cual se obtiene un archivo en texto plano, de este archivo se extrae el texto para el cual se genera la marca de agua con la cual se realiza el proceso de autenticación que consiste en comparar la marca obtenida con la marca registrada si son similares se devuelve el resultado de que la copia es auténtica.

El medio por el cual se comunica el servidor con la aplicación cliente es por medio de JSON, para este propósito se implementa su respectivo módulo para codificación y decodificación de mensajes.

El Script para generar la marca de agua es implementado en un paquete que cumple el diseño de la técnica de generación de marca de agua. El mismo es implementado en el lenguaje PHP.

3.5.2 Aplicación cliente para autenticación de documento impreso

El prototipo que se plantea para autenticar documentos impresos en papel en tiempo real se basa en la tecnología de servicio web, el cual es brindado por una Autoridad de Certificación, el mismo implementa el servicio orientado a los teléfonos inteligentes o tabletas con sistema operativo Android.

Para este proceso se tiene en cuenta los siguientes módulos:

Toma de imagen del documento, este proceso se puede realizar con un escáner o una cámara fotográfica, en este caso se toma con la cámara del Smartphone, donde la aplicación se encarga de activar la cámara y posteriormente enviar la imagen al servidor de la Autoridad de Certificación. Para este proceso se debe tomar en cuenta la alineación del texto

ubicado en el documento de forma adecuada para que el motor OCR pueda identificar los caracteres.

Proceso de envío de imagen, en este proceso la aplicación se conecta al servidor de forma asíncrona en segundo plano, una vez conectado con el servidor se procede al envío de la imagen tomada con la cámara del dispositivo Smartphone.

Verificación de la marca de agua, este proceso también es asíncrono, se desarrolla en segundo plano, recibe la respuesta del servidor en un archivo en formato JSON, el servidor nos envía la respuesta de si el documento enviado es auténtico o no. Con la respuesta obtenida se procede a visualizar en la pantalla del Smartphone la respuesta del servidor. También se visualiza el contenido del documento de texto tal como fue registrado en el servidor, esto es para mostrar el contenido auténtico del documento de texto registrado en la Autoridad de Certificación.

Seguridad, como medida de seguridad en el prototipo se implementa contraseñas a nivel de aplicación para acceder a los servicios de la Autoridad de Certificación. Con esta medida solo pueden acceder usuarios registrados.

CAPÍTULO IV: EVALUACIÓN DE RESULTADOS

4.1 FACTORES DE EVALUACIÓN

Para obtener resultados se realizaron pruebas a las técnicas de marcas de agua propuesta para identificar el autor y autenticar el contenido de texto.

En la evaluación de las marca de agua, se consideran los siguientes factores: robustez para la primera con el que se resguarda o protege el derecho de autor y fragilidad para la segunda con la que se autentica el contenido del documento de texto.

Para la prueba de la hipótesis se consideran los resultados experimentales obtenidos con el prototipo implementado. Los textos a los cuales se aplicó la técnica se incluyen en la sección de anexos.

Texto original **To**	Número Palabras	Marca de agua **W1**	Marca de agua **W2**
Text1	40	4.7.8.5.6.7.6.10.6.5	1ac2036de2b67b96a2cc21696281dc60
Text2	46	2.4.3.6.2.6	a9fb3ba9b88edd8a0703cb11a9dd2687
Text3	81	5.4.5.11.3.5.6.3.5.4.9 .3.5.6.6.4.8.5	21d416f2be3138498edb6361134ab4d6
Text4	153	7.2.11.5.4.12.12.5.6. 13.6.2.10.2.7.13.8.13	1e070115daf5b201b37ba13f7460f96b

Tabla 5 Textos y marcas de agua

Fuente: Elaboración propia

En la tabla 5 se muestra las características de cada uno de los textos utilizados para el experimento. Donde To es el texto original registrado en la Autoridad de Certificación, W es la marca de agua correspondiente al texto vinculado al documento de texto que es el resultado de aplicar la técnica de marca de agua propuesta.

4.2.1 Evaluación de robustez

Para evaluar este factor se tomó en cuenta la afirmación de Fridrich & Goljan(1999) , quienes mencionan que para evaluar la robustez no se incluyen los ataques basados en el conocimiento de los algoritmos de incrustado y detección de la marca, la robustez significa resistencia a ciegas frente a aquellas modificaciones producidas por las operaciones comunes a las que estarán expuestos.

En la evaluación se consideró las siguientes operaciones comunes: inserción, supresión, reordenamiento de palabras o frases.

Inserción: Para evaluar este ataque que sufre el documento de texto se procede a autenticar este documento atacado Tr, los resultados son presentados en la tabla. Estos datos resultan después de insertar una palabra al azar en el documento de texto original.

Texto recuperado **Tr**	Marca de agua registrada **W1**	Marca de agua recuperada **Wr1**	Similitud %
Text1	4.7.8.5.6.7.6.10.6.5	4.7.8.5.6.7.6.10.6.5	100
Text2	2.4.3.6.2.6	2.4.3.6.2.6	100
Text3	5.4.5.1.1.3.5.6.3.5.4.9.3.5.6.6.4.8.5	5.4.5.1.1.3.5.6.3.5.4.9.3.5.6.6.4.8.4	94.7
Text4	7.2.11.5.4.1.2.1.2.5.6.1.3.6.2.10.2.7.1.3.8.1.3	7.2.11.5.4.1.2.1.2.5.6.1.3.6.2.10.2.7.1.3.8.1.3	100

Tabla 6 Resultados después de ataque de inserción

Fuente: Elaboración propia

Supresión: Los datos de este ataque se presentan en la siguiente tabla, son el resultado de suprimir una palabra al azar del texto original registrado.

Texto recuperado **Tr**	Marca de agua registrada **W1**	Marca de agua recuperada **Wr1**	Similitud Tasa de coincidencia
Text1	4.7.8.5.6.7.6.10.6.5	47856761065	Si
Text2	2.4.3.6.2.6	243626	Si
Text3	5.4.5.1.1.3.5.6.3.5.4.9.3.5.6.6.4.8.5	5451135635493566485	Si
Text4	7.2.11.5.4.1.2.1.2.5.6.1.3.6.2.10.2.7.1.3.8.1.3	5.2.7.3.4.5.9.8.4.2.1.2	No

Tabla 7 Resultado de ataque de supresión

Fuente: Elaboración propia

Reordenamiento: este tipo de ataque hace referencia a los saltos de línea, inserción de espacios entre palabras o sea cambiar las líneas originales.

Texto recuperado **Tr**	Marca de agua registrada **W1**	Marca de agua recuperada **Wr1**	Similitud
Text1	4.7.8.5.6.7.6.10.6.5	4.7.8.5.6.7.6.10.6.5	Si
Text2	2.4.3.6.2.6	2.4.3.6.2.6	Si
Text3	5.4.5.1.1.3.5.6.3.5.4.9.3.5.6.6.4.8.5	5.4.5.1.1.3.5.6.3.5.4.9.3.5.6.6.4.8.5	Si
Text4	7.2.11.5.4.1.2.1.2.5.6.1.3.6.2.10.2.7.1.3.8.1.3	7.2.11.5.4.1.2.1.2.5.6.1.3.6.2.10.2.7.1.3.8.1.3	Si

Tabla 8 Resultado de ataque de reordenamiento

Fuente: Elaboración propia

4.2.2 Evaluación de fragilidad

Para evaluar este factor se consideró la segunda marca de agua, la cual es producto de la técnica elegida para el trabajo, además esta es altamente sensible a cualquier cambio ya sea esto de una letra en las palabras del documento de texto.

Esta prueba es realizada con el prototipo en condiciones técnicas apropiadas, las cuales se incluyen en la sección recomendaciones, además supone un reconocimiento OCR del 100%.

Los resultados de esta prueba se expresan en la siguiente tabla, la cual es elaborada con los textos de prueba sin alteración alguna donde se puede ver las marcas recuperadas en la tercera columna.

Texto recuperado **Tr**	Marca de agua registrada **W2**	Marca de agua recuperada **Wr2**	**W2=Wr2**
Text1	1ac2036de2b67b96a2cc21696281dc60	1ac2036de2b67b96a2cc21696281dc60	Si
Text2	a9fb3ba9b88edd8a0703cb11a9dd2687	a9fb3ba9b88edd8a0703cb11a9dd2687	Si
Text3	21d416f2be3138498edb6361134ab4d6	21d416f2be3138498edb6361134ab4d6	Si
Text4	1e070115daf5b201b37ba13f7460f96b	1e070115daf5b201b37ba13f7460f96b	Si

Tabla 9 Recuperación de marca de agua frágil

Fuente: Elaboración propia

También se realiza la prueba en la que por los menos se altera un caracter, por alteración se entiende la modificación, supresión, borrado o inserción de una palabra o carácter. Los resultados de esta prueba se presentan en la siguiente tabla.

Texto recuperado **Tr**	Marca de agua registrada **W2**	Marca de agua recuperada **Wr2**	**W2=Wr2**
Text1	1ac2036de2b67b96 a2cc21696281dc60	5980d619fc9ad7de c531ba729147bc9e	No
Text2	5c4dabbb0cd86055 2d20da54f89ee71b	452bd90dee33e09 319a3e243814fe0e8	No
Text3	21d416f2be313849 8edb6361134ab4d6	15f0136a9d5cbbac 8a09ee886253de60	No
Text4	1e070115daf5b201 b37ba13f7460f96b	6ffedfc65d5ec45e0 8980023099a5275	No

Tabla 10 Prueba de fragilidad frente a texto alterado

Fuente: Elaboración propia

4.3 CONTRASTE DE HIPÓTESIS

Para contrastar la hipótesis planteamos la hipótesis nula H0 y la hipótesis se investigación Hi como sigue:

Hi: La técnica de marca de agua permite resguardar el derecho de autor y autenticidad de un documento digital o impreso, además de autenticar el documento en tiempo real.

H0: La técnica de marca de agua no permite reguardar el derecho de autor y autenticidad de un documento digital o impreso, además no permite autenticar el documento en tiempo real.

Para demostrar la hipótesis acudiremos los resultados experimentales obtenidos y a la técnica estadística chi- cuadrada.

Se considera como parte de la hipótesis las siguientes variables:

W1: Es la marca de agua robusta registrada en la Autoridad de Certificación.

W2: Es la marca de agua frágil registrada en la Autoridad de Certificación.

Wr1: Es la marca de agua que se recupera del documento de texto físico.

Wr2: Es la marca de agua que se genera del documento de texto físico (impreso en papel).

W1=Wr1	W1!=Wr1	Total casos observados
49	1	50

Tabla 11 Datos de recuperación de marca de agua robusta

Fuente: Elaboración propia

Los datos presentados en la tabla 11 son observaciones realizadas con las imágenes tomadas desde un escáner, estas imágenes contienen texto sin alteración alguna por lo que la recuperación de la marca de agua robusta es casi total en todos los casos.

El detalle de los resultados de la tabla 11 se muestra en la siguiente tabla para aplicar la prueba de chi-cuadrada.

Para el contraste de la hipótesis dividimos esta en 2 premisas las cuales trabajaremos de forma separada, considerando las variables W1 y Wr1 para la premisa P1; y las variables W2 y Wr2 para la premisa P2. Cuyos resultados experimentales están representados en las tablas respectivas.

P1: La marca de agua permite resguardar el derecho de autor de un documento.

P2: La marca de agua permite autenticar el texto contenido en un documento impreso en papel.

Para contrastar la premisa P1 se toman los valores de la tabla 12 para calcular $\chi^2{}_c$. El nivel de confianza tomado para el experimento es de 0.95 con (v=4-1) 3 grados de libertad cuyo valor crítico de chi-cuadrado es: $\chi^2{}_t = 0.35$.

Frecuencias	Text1	Text2	Text3	Text4	Total
Observadas(Fo)	15	10	10	14	49
Teóricas(Ft)	15	10	10	15	50

Tabla 12 Resultados experimentales vs teóricos para W1

Fuente: Elaboración propia

Hipótesis alterna para P1 H1i: $\chi^2{}_c$<=0.35, se rechaza H01.

Hipótesis nula para P1 H10: $\chi^2{}_c$>0.35, se acepta H10.

Se calcula $\chi^2{}_c$ con la fórmula: $\chi^2{}_c = \sum \frac{(Fo-Ft)^2}{Ft}$

$$\chi^2{}_c = \frac{(15-15)^2}{15} + \frac{(10-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(14-15)^2}{15} = 0.067$$

Se observa que $\chi^2{}_c < \chi^2{}_t$(0.067<0.35) este resultado rechaza H10 y se acepta H1i.

La premisa P2 es para verificación la autenticidad de un documento de texto recuperado Tr con la marca de agua frágil Wr2, los datos obtenidos son el producto del experimento realizado.

W2=Wr2	W2!=Wr2	T Total casos observados
47	3	50

Tabla 13 Recuperación de la marca de agua frágil

Fuente: elaboración propia

Los datos presentados en la tabla 13 son observaciones realizadas con las imágenes tomadas desde un escáner, estas imágenes contienen texto sin alteración alguna por lo que la recuperación de la marca de agua frágil es casi total en todos los casos, esto se muestra en la siguiente tabla.

Frecuencias	Text1	Text2	Text3	Text4	Total
Observadas(Fo)	14	9	10	14	47
Teóricas(Ft)	15	10	10	15	50

Tabla 14 Resultados teóricos vs experimentales para W2

Fuente: elaboración propia

Se toma un nivel de confianza de 0.95 con (v=2-1) 3 grados de libertad.

Hipótesis alterna para P2 H2i: $\chi^2{}_c$<=0.35, se rechaza H20.

Hipótesis nula para P2 H20: $\chi^2{}_c$>0.35, se acepta H20

Se calcula $\chi^2{}_c$ con la fórmula: $\chi^2{}_c = \sum \frac{(Fo-Ft)^2}{Ft}$

$$\chi^2{}_c = \frac{(14-15)^2}{15} + \frac{(9-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(14-15)^2}{15} = 0.23$$

Se observa que $\chi^2{}_c < \chi^2{}_t$(0.23<0.35) este resultado rechaza H21 y se acepta H2i

Para el contraste final se refuta H0 expresando en proposiciones lógicas equivalentes a las premisas P1 y P2 correspondientes a las hipótesis nulas H10 y H20 como:

"~H10 ^ ~H20 => ~H0".

Como la hipótesis nula H0 se reduce al absurdo, la hipótesis de investigación Hi propuesta se acepta como válida.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

De forma general se concluye que las técnicas de marcas de agua permiten resguardar el derecho de autor de un documento de texto, y también permiten realizar el proceso de autenticación de forma inmediata.

5.2 Estado de los objetivos específicos

- Analizar las técnicas de watermarking apropiadas para realizar el presente trabajo.

  Se realizó un análisis de las diferentes técnicas, para documento existentes. Por lo que este objetivo específico se cumple en el presente trabajo.

- Analizar las características de las marcas de agua apropiadas para realizar el trabajo.

  Se realizó un análisis de las diferentes caracteristicas de las marcas de agua, para documento existentes. Por lo que este objetivo específico se cumple en el presente trabajo.

- Establecer a la técnica watermarking como un mecanismo que permite resguardar o proteger la información acerca de los autores de un documento.

  Para cumplir con este objetivo específico se implementa un prototipo el cual permite registrar los datos del autor de un documento en una Autoridad de Certificación, donde la Autoridad de Certificación resguarda los datos del autor.

- Establecer el algoritmo de Watermaking adecuado para el presente trabajo, el cual debe cumplir las propiedades de las marcas de agua.

  Se establece el algoritmo para generar marcas de agua el cual cumple con las caracteristicas principales de los algoritmos de marcas de agua para textos existentes o publicados hasta la fecha, además muestra robustez frente a los ataques comunes para documentos de texto.

- Lograr el desarrollo de un prototipo que permita generar, incrustar y detectar una marca de agua en el documento electrónico.

  El desarrollo del prototipo para generar la marca de agua se efectuó, este prototipo funciona con la implementación de una Autoridad de Certificación, permite extraer el texto del documento electrónico, al texto plano extraído se le aplica la técnica de marca de agua.

- Lograr el desarrollo de un prototipo que permita incluir una marca de agua patrón al documento físico en papel derivado del documento electrónico.

  Por la característica de los documentos de texto y su estructura se concluye que este paso se realiza al momento de escribir dicho documento, quedando como marca de agua la característica de la misma estructura de texto.

- Lograr el desarrollo de un prototipo que facilite el reconocimiento y autenticación de la marca de agua impresa en papel en tiempo real, con la ayuda de un Smartphone.

  Para cumplir con este objetivo se implementa un prototipo de aplicación para Smartphones con sistema operativo Android, este prototipo para realizar el proceso de autenticación en tiempo real se comunica con un servidor de la Autoridad de Certificación.

  El prototipo realiza el proceso de autenticación con éxito.

5.3 Recomendaciones

Se recomienda un proceso para mejorar la imagen antes de enviar al motor OCR.

Se recomienda tomar en cuenta las diferentes técnicas de la inteligencia artificial, para agilizar el proceso de generación de marcas de agua para documentos con más contenido texto.

Se recomienda para trabajos posteriores, tomar en cuenta técnicas de marcas de agua orientadas al contenido sintáctico del texto de los documentos.

Realizar procesos para extraer texto plano de las diferentes extensiones comunes existentes en el medio, en los cuales se publican los documentos de texto.

Se recomienda tomar en cuenta la integración de herramientas tecnológicas ya que en un principio se utilizó Matlab pero en la implementación se sustituyó por herramientas del lenguaje de programación PHP, ya que este cuenta con funciones apropiadas para procesamiento de texto.

5.3.1 Recomendaciones técnicas para funcionamiento del prototipo

Recomendación para tomar la fotografía del documento de texto físico el Smartphone debe estar a una distancia aproximada de 10 a 15 cm, además en lo posible centrar la cámara alineando tanto horizontal como verticalmente.

Es recomendable tomar la imagen en un lugar apropiado como ser una oficina o un ambiente el cual debe contar con iluminación esto es para evitar sombras en la fotografía.

Requerimientos mínimos para el Smartphone

Versión de Android: 4.2.2

Camara: 5Mp (Mega Pixeles)

Procesador: 1GHz

Memoria RAM: 1GB

Conectividad: 2G/3G/HSPA o WiFi

Requerimientos para impresión de texto en papel

Tipo de papel: Bond

Color de papel: Blanco

Color de fuente: Negro

Tipo de letra: Calibri, arial

Tamaño de la letra: 11, 12, 14 pt (puntos)

Impresión del texto normalizado

Para la autenticación se recomienda organizar el texto en forma rectangular y centrar como se muestra en la figura.

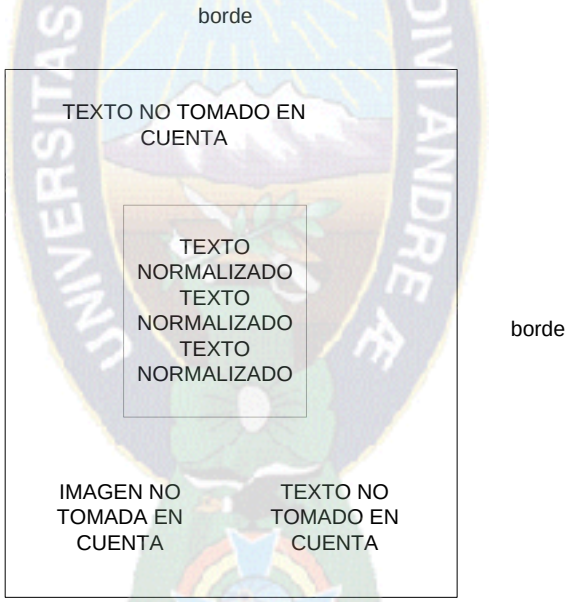


Figura 16 Organización de normalizado

Fuente: Elaboración propia

Bibliografía

Al-Wesabi, F., Alsakaf, A., & Vasantrao, K. (2013). A ZERO TEXT WATERMARKING ALGORITHM BASED ON THE PROBABILISTIC PATTERNS FOR CONTENT AUTHENTICATION OF TEXT DOCUMENTS. *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)*, 284-300.

Angulo, C. A., Blandón, L. H., & Ocampo, S. M. (Diciembre de 2007). Una Mirada a la esteganografía. *Scientia et Technica, 13*(37), 421-426.

Barán, B., Gómez, S., & Bogarín, V. (2001). Steganographic Watermarking for Documents.

Brassil, J., Low, S., & Maxemchuk, N. (1999). Copyright Protection for the Electronic Distribution of Text Documents. *Proceedings of the IEEE, vol. 87, no. 7*, 1181-1196.

Delgado Gómez, A. (2009). La conservación a largo plazo de documentos electrónicos: normativa ISO y esfuerzos nacionales e internacionales. *Revista Andaluza de Archivos*.

Eggers, J. J., Su, J., & Girod, B. (7-11 de Mayo de 2001). Blinnd Watermarking Applied to Image Authentication. Utah, Salt Lake City, USA.

Google Inc. (2005). *Google Project Hosting*. Recuperado el 28 de 02 de 2014, de https://code.google.com/p/tesseract-ocr/

INTECO. (09 de Marzo de 2010). Esteganografía, el arte de ocultar información.

J.Fridrich, & Goljan, M. (1999). Comparing Robustness of Watermarking Techniques.

Jalil, Z., Mirza, A., & Sabir, M. (2010). Content based Zero-Watermarking Algorithm for Authentication of Text Documents. *International Journal of Computer Science and Information Security*, 212-217.

Mason, S. (2004). *EL NOTARIO.* Recuperado el Febrero de 2014, de http://www.elnotariado.com/images_db/noticias_archivos/418.pdf

Menvielle, M., Garcia Neder, H. J., Groppo, M. A., Gibellini, F., Sánchez, C. B., Pozzi, J. W., y otros. (2011). Metodología para usar la esteganografía como medio de acreditar la validez de la documentación publicada electrónicamente.

Moreno López, L. (2009). Consejos de cómo crear un documento universal (formato .txt) como alternativa a documentos digitales no accesibles.

Müller , L. (2010). Sistemas de Autenticación y firmas . En *Redes de comunicación* (pág. 10).

Orúe, A. B. (2002). Marcas de agua en el mundo real.

Reverte, M. P., & Martínez, B. A. (2004). Protección del copyright de imágenes digitales.

Tuohey, J. (22 de Noviembre de 2004). *Pcworld.* Recuperado el 2013, de http://www.pcworld.com/article/118664/article.html

Velasco Bautista, C. L., López Hernández, J. C., Nakano Miyatake, M., & Pérez Meana, H. M. (2007). Esteganografía en una imagen digital en el dominio DCT. Mexico.

Wikipedia. (8 de Mayo de 2013). Recuperado el Noviembre de 2013, de http://en.wikipedia.org/wiki/Printer_steganography

Wu, M., & Lui, B. (2004). Data Hiding in Binary Image for Autentication and Annotation. *IEEE Trans on Multimedia*, 528-538.

ANEXOS

TEXTOS DE PRUEBA

Text1

LA ESTEGANOGRAFIA ESTA DEFINIDA COMO EL ARTE DE OCULTAR INFORMACION EN ARCHIVOS DE TEXTO IMAGENES SONIDOS O EN CANALES ENCUBIERTOS A TRAVES DE METODOS Y TECNICAS COMPUTACIONALES SE ENCUENTRA EN EL AMBITO DE TRASPORTAR INFORMACION A TRAVES DE REDES INFORMATICAS

Text2

UN GEN ES UN SEGMENTO CORTO DE ADN QUE LE DICE AL CUERPO COMO PRODUCIR UNA PROTEINA ESPECIFICA HAY APROXIMADAMENTE 30000 GENES EN CADA CELULA DEL CUERPO HUMANO Y LA COMBINACION DE TODOS LOS GENES CONSTITUYE EL MATERIAL HEREDITARIO PARA EL CUERPO HUMANO Y SUS FUNCIONES

Text3

UNA MARCA DE AGUA DIGITAL ES UNA SENAL PERMANENTE INTEGRADA EN LOS DATOS DIGITALES QUE SE PUEDE DETECTAR O EXTRAER MAS TARDE POR MEDIO DE OPERACIONES COMPUTACIONALES CON EL FIN DE HACER AFIRMACIONES ACERCA DE LOS DATOS LA MARCA DE AGUA SE OCULTA EN LOS DATOS DEL ANFITRION DE TAL MANERA QUE SE HACE RESISTENTE A MUCHAS OPERACIONES QUE NO DEGRADEN EL DOCUMENTO ANFITRION ASI POR MEDIO DE MARCAS DE AGUA LO DATOS SIGUEN SIENDO ACCESIBLES PERO MARCADOS DE FORMA PERMANENTE

Text 4

EL PROCESO DE LA PRODUCCION DEL CONOCIMIENTO TIENDE A REUNIR AL SUJETO COGNOSCENTE Y AL OBJETO COGNOSCIBLE AMBOS NO PUEDEN ACTUAR POR SEPARADO DEBIDO A QUE EL PRIMERO DEPENDE DEL SEGUNDO PARA QUE EMERJA LA POSIBILIDAD DE ALGUN TIPO DE CONOCIMIENTO ES POR ELLO SU DUALIDAD ESTE PLANTEAMIENTO TOMA FUERZA EN LA TEORIA DEL CONOCIMIENTO DE JOHAN HESSEN CUANDO ARGUMENTA QUE EL SUJETO SOLO ES SUJETO PARA EL OBJETO Y EL OBJETO ES SOLO OBJETO PARA EL SUJETO AMBOS TIENE GRANDES DIFERENCIAS PERO CADA UNO TIENE SUS FUNCIONES LA FUNCION DEL SUJETO ES APREHENDER DEL OBJETO Y LA FUNCION DEL OBJETO ES SER APREHENDIDO POR EL SUJETO BUNGE 1977 EN ESTE CONTEXTO LA RELACION QUE EXISTE ENTRE EL SUJETO DENOMINADO INVESTIGADOR Y EL OBJETO NOMBRADO COMO OBJETO DE INVESTIGACION ES LA REDACCION DEL OBJETO INVESTIGADO A TRAVES DE LA GENERACION DE UN INFORME DE INVESTIGACION CONOCIDO TAMBIEN COMO ARTICULO DE INVESTIGACION

Tabla de chi-cuadrada

DISTRIBUCION DE χ^2

Grados de libertad	Probabilidad										
	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,01	0,001
1	0,004	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	6,64	10,83
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	9,21	13,82
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82	11,34	16,27
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	13,28	18,47
5	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	15,09	20,52
6	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	16,81	22,46
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	18,48	24,32
8	2,73	3,49	4,59	5.53	7,34	9,52	11,03	13,36	15,51	20,09	26,12
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	21,67	27,88
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	23,21	29,59
	No significativo								Significativo		