DOI: 10.5769/C2015002 or http://dx.doi.org/10.5769/C2015002

Proteção da Prova Documental Impressa e Digitalizada com a Utilização de *Watermarking*

Felippe Pires Ferreira

***Resumo*—O artigo propõe um método para disponibilização de documentos sigilosos durante o inquérito policial, que pode ser estendido a outros documentos, introduzindo o elemento de segurança conhecido como *watermarking*. Este elemento permitirá vincular uma cópia de documento a seu destinatário inicial, e em casos de vazamento de informação permitirá identificar a origem da cópia. O método possibilitaria incluir uma *watermark* em um documento eletrônico editável e recuperá-la em documentos impressos ou digitalizados, bastando apenas um fragmento do texto. O método é baseado na semelhança entre caracteres de diferentes fontes de texto, os quais serão utilizados para criação de uma codificação identificadora da origem do documento.**

Palavras-Chave*—*Marca d'água, Cópia de Documento, Cópia Impressa, Digitalização, Inquérito Policial, Fonte de Texto.

Abstract—**This article proposes a method for available classified documents during the police investigation, which can be extended to other documents, introducing the security element known as watermarking. This element will link a document copy to its initial recipient, and in cases of information leakage will identify the origin of the copy. The method would allow include a watermark in an editable electronic document and retrieve it in printed or scanned documents, just by a fragment of the text. The method is based on the similarity between characters of different text fonts, which will be used to create a code identifying the origin of the document.**

Keywords*—*Watermarking, Document Copy, Hard-Copy, Scan, Police Investigation, Text Font.

INTRODUÇÃO

A difusão da informação com o auxílio dos avanços da tecnologia foi inicialmente aceita como grande revolução na comunicação. A facilidade de se encontrar um documento editado por uma pessoa a quilômetros de distância permite rapidez e dinamismo ao processo de comunicação. Entretanto, a crescente difusão de conhecimento também proporciona a prática de publicação de material não autorizado através da Internet. Diversos livros, artigos e documentos de trabalho também foram objeto dessa popularização e disseminação da prática de compartilhamento de documentos on-line [1].

Neste novo cenário, surgiu a necessidade de criação de um mecanismo que protegesse a produção intelectual dos autores. Entretanto, essa proteção precisa ser robusta o suficiente para que não seja removível, capaz de ser recuperável, além de identificar o material. Dessa necessidade foi criado o conceito de *watermarking*, ou termo traduzido **marca d'água** [1].

Watermarking é o processo de inserir informações sobre o objeto no próprio objeto, e esta informação pode ser extraída posteriormente para ser verificada [2]. Diferente da esteganografia, a *watermarking* não procura ser totalmente eficiente contra detecções, mas visa identificar um material, e impossibilitar sua remoção ou alteração [3].

As *watermarkings* possuem diferentes aplicações [4] [5]. Elas permitem a **identificação do proprietário**, protegendo assim a distribuição de materiais como músicas e livros com sua utilização. Muitas vezes a marca d'água não está oculta e apresenta informações sobre o proprietário do material. A **autenticidade** do material pode ser verificada e, com outros mecanismos de segurança como a criptografia, contribui para a confirmação da origem da informação e sua veracidade. Caso ocorra alguma modificação do material, a verificação da **integridade** da marca d'água pode indicar que houve manipulação não autorizada sobre este. É possível realizar o controle de cópias personalizando as marcas d'água de acordo com o destinatário do material, possibilitando o rastreio do material em caso de vazamentos.

A literatura define que a *watermarking* tem propriedades que devem estar presentes no mecanismo quando inserido em um objeto, a fim de torná-la adequada para utilização [4]. A *watermarking* deve ser **robusta**, deste modo ser resistente a manipulação do material e ainda permanecer neste. Deve ser **não perceptível** a visão humana, sendo visível apenas durante processos de extração da informação. Ser **segura**, assim apenas o proprietário do material poderá recuperá-la, alterá-la ou removê-la. E por fim, deve ser capaz de armazenar informação/mensagem em um objeto [1] [6].

Outro aspecto são os diferentes suportes/materiais em que podem ser inseridas. De acordo com o tipo de arquivo de mídia são utilizadas técnicas diferentes. É possível utilizá-la em arquivos de texto, imagens, áudio e vídeo. Cada arquivo possui a sua peculiaridade, mas todos devem buscar as propriedades que tornam a *watermarking* resiliente.

O trabalho objetiva desenvolver um algoritmo para inclusão de marcas d'água em documentos eletrônicos textuais que permanecerá no documento mesmo após sua impressão, sendo passível de recuperação em processos de digitalização ou fotografia do material. Para atingir tal objetivo, o trabalho faz uso de manipulação de fontes de caracteres para inserção de um código identificador nos documentos. Os documentos textuais devem permitir edição, não sendo aplicável a imagem de documentos.

Na seção II serão apresentadas as metodologias de utilização de *watermarking* em documentos textuais, com ênfase na metodologia Baseada em Imagem, em que a técnica proposta neste trabalho se insere. Na seção III será ilustrado o

contexto em que se deseja aplicar a marca d'água, no caso em documentos que compõem o inquérito policial. A seção IV descreverá alguns trabalhos que serviram de direcionamento para construção da técnica do artigo. A seção V descreverá a técnica proposta através de esquemas e algoritmos, ressaltando as condições em que a técnica é aplicada. Na seção VI serão apresentados os resultados obtidos com a inserção de *watermark* de tamanhos diferentes, digitalizações de texto com diferentes configurações e a qualidade de duas imagens registradas por câmeras de *smartphones*, e posteriormente, é realizada a análise dos resultados, destacando as principais diferenças dos textos com variações nos tamanhos da *watermark* e os suportes em que se encontram, além da comparação entre a técnica proposta e métodos presentes na literatura. Por fim, a seção VII apresenta a conclusão da pesquisa e expectativas para trabalhos futuros.

WATERMARKING EM DOCUMENTOS TEXTUAIS

Além de aplicação em arquivos de imagens, áudios e vídeos, outra aplicação deste elemento de segurança são os documentos textuais. Diferentes técnicas de aplicação de *watermarks* foram criadas e agrupadas em quatro principais categorias [7]:

a. Baseada em Imagem - Nesta técnica o texto é tratado como imagem, sendo comparados aspectos visuais da composição do texto, como: alinhamento, espaçamento, caracteres, etc [8];

b. Sintática - Consiste em utilizar aspectos característicos de uma linguagem como substantivos, verbos, artigos, proposições, etc. Assim construindo uma árvore sintática que por sua vez será utilizada para inserção do elemento de segurança [9];

c. Semântica - Permite realizar mudanças de palavras por sinônimos ou utilizar abreviações e acrônicos para construir a marca d'água [10];

d. Estrutural - Utiliza características próprias da linguagem como a ocorrência de letras duplas ou preposições. O texto não é alterado, mas suas características são utilizadas na criação da marca d'água [8].

A pesquisa é classificada como Baseada em Imagem, porque utiliza aspectos visuais do texto, mais especificamente dos caracteres, para inserção e recuperação da *watermark*. Dentro dessa categoria podemos listar três técnicas [10]:

a. *Line Shifting* - as linhas do texto são deslocadas verticalmente, para baixo ou para cima, de tal forma que seja pouco perceptível aos olhos humanos. As linhas ímpares não são alteradas para que funcionem como linhas de controle. Entretanto, para recuperação da *watermark* , é necessário o documento original para fins de comparação;

b. *Word-Shift Coding* - deslocamento horizontal das palavras ou linhas de acordo com a *watermark* que se pretende inserir. As palavras são classificadas em grupos, nos quais os grupos pares variam de acordo com a *watermark* e os ímpares funcionam como controle e comparação para os deslocamentos. As palavras das extremidades não podem ser alteradas para que seja preservado o alinhamento. Assim como a técnica anterior, é necessário o documento original para recuperar a informação [11];

c. *Feature Coding* - utiliza variações das características de palavras ou letras como tamanho ou espaçamento para criação da *watermark*. Também faz uso do texto original para recuperação da informação.

As duas primeiras técnicas também precisam de um texto maior que a terceira técnica, uma vez que as alterações que aquelas proporcionam são baseadas nas linhas, enquanto esta é baseada em palavras ou letras.

APLICAÇÃO DO CONCEITO DE WATERMARKING EM DOCUMENTOS DO INQUÉRITO POLICIAL

O processo inquisitorial é marcado pelo seu sigilo durante a investigação e pela quantidade de provas produzidas que irão fundamentar a investigação e o futuro processo judicial. Apesar de sigiloso, algumas peças produzidas no inquérito policial são acessadas por personalidades que compõem esse trâmite, como advogados, policiais, promotores, juízes, entre outros. Esse acesso sem um controle eficaz pode gerar vazamento de informações e comprometer a própria investigação.

Pela facilidade na disseminação da informação, a detecção do responsável pelo vazamento se torna uma tarefa complexa, porque diferentes cópias são distribuídas e ainda não há maturidade suficiente nos processos de controle de cópias que possibilitem o rastreio da informação.

A definição de uma política de controle de cópias possibilita o rastreio da informação e a responsabilização do possível agente causador do vazamento. Uma das medidas seria a introdução de *watermarking* na produção de cópias de documentos, imagens e gravações, a fim de garantir a autenticidade, integridade e controle de cópia [5].

No entanto, o acesso ao material que foi alvo de vazamento não está disponível às autoridades, apenas o material original, porque são divulgados em veículos de comunicação como jornais e Internet, portanto os arquivos eletrônicos não estão à disposição para que sejam analisados. Logo, é preciso criar uma *watermarking* que possa ser recuperável nestes cenários.

TRABALHOS CORRELATOS

Alguns trabalhos utilizavam aspectos estruturais do texto ou dos caracteres para criarem as marcas d'águas. Diferentes abordagens para criação de *watermark* relacionados às características do texto foram encontradas:

a. *Watermark* criado através das características específicas dos documentos produzidos no software *Microsoft Word*. Cada parágrafo, palavra ou letra são tratados como objetos, possuindo propriedades, as quais possuem atributos especiais em que são possíveis esconder informações. Desta forma, permite que o texto possa ser distribuído pela Internet, mas ainda manter o mecanismo de controle de cópia dentro do documento [12];

b. *Watermarking* baseado na frequência das letras em sentenças escolhidas. Para cada sentença é criada um código de letras que será concatenado com outros códigos para criação da marca d'água. Para revelar o identificador é necessário comparar o *watermark* recebido com o *watermark* obtido durante uma nova execução do algoritmo sobre o texto. A técnica auxilia a verificação de possíveis alterações no texto [13];

c. Criação de *watermark* invisível que é inserida em páginas HTML através da *tag* <*meta*>. A marca d'água é criada, submetida a uma função de *Hash*, seu

resultado é convertido para oito dígitos e inserido no documento. Os caracteres são inseridos como espaços em branco, sendo ignorados por diversos programas [14];

d. A definição da *watermarking* está baseada em características como a distância entre as palavras [15] ou analisando características de espaçamento e tamanho entre caracteres [16];

e. Aplicação sobre o texto de pequenos pontos que se tornam pouco perceptíveis durante a leitura. Para recuperação da imagem é empregada autocorrelação entre os pontos, além da aplicação de pontos de registro para corrigir distorções geométricas [17];

f. Através da alteração de características dos caracteres, como o tamanho, é possível introduzir uma mensagem. A alteração de *pixels* do caractere permite a introdução da mensagem nas letras. Caso haja poucas alterações de *pixels*, a mensagem inserida torna-se menos robusta, entretanto menos perceptível. Em situação contrária, muitas alterações de *pixels* tornam a mensagem mais robusta e mais perceptível [18].

Analisando os trabalhos correlatos, o objetivo da pesquisa é criar uma técnica de *watermarking* que seja: não perceptível; mantenha informações sobre o detentor do documento; recuperável; e permaneça no documento mesmo após sua impressão ou digitalização.

TÉCNICA PROPOSTA

Como o contexto de aplicação da técnica criada é o inquérito policial, antes da execução da técnica é preciso definir o momento em que a técnica será aplicada sobre os documentos do inquérito.

É preciso definir uma nova fase no procedimento de geração de cópias de documentos. A técnica será aplicada a documentos digitais editáveis que serão impressos ou disponibilizados eletronicamente. Antes da execução do algoritmo propriamente dito, é necessário definir as etapas anteriores que serviram de subsídio para execução da técnica proposta em uma cópia de prova documental. O algoritmo necessita de uma *watermark* que seja identificadora do detentor da cópia. O processo de emissão de cópias precisa manter um registro histórico dos documentos emitidos e vincular o identificador a seu destinatário. Além de considerar que os documentos foram obtidos originalmente de um meio eletrônico.

Nesta fase, teremos as seguintes etapas após a elaboração do documento:

1. Criação de um código identificador, representado por uma sequência de bits, para o destinatário da cópia, devendo ser armazenado e registrado historicamente a vinculação do documento a este identificador;
2. Definição de um subconjunto C com as letras do alfabeto que terão suas formas alteradas no texto;
3. Inclusão, opcional, de preâmbulo ou mecanismo de segurança à marca d'água.

O código identificador pode ser acrescentado de elementos diferentes, com finalidades distintas. A marca d'água pode vir acompanhada de um código preâmbulo; ser submetida a uma função criptográfica; ou um código de verificação de erro. No caso do preâmbulo, é definido seu tamanho e se será inserido no inicio e/ou fim do código identificador, desta forma auxiliando na identificação da *watermark* no texto. A função criptográfica será aplicada com a finalidade de ocultação da mensagem proposta, entretanto pode gerar um *watermark* de tamanho maior. A terceira forma proposta consiste na inclusão de um código de verificação de erros associado ao código identificador. Uma abordagem que poderia ser utilizada com códigos identificadores pequenos é o código de Hamming(7,4) [19], o qual possibilita a detecção de até dois bits e a recuperação de até um bit da mensagem. A aplicação de um código de verificação contribui para solução de dois problemas desse contexto de marcas d'água em documentos impressos ou digitalizados: como existe a possibilidade de ruídos, é possível que alguma informação seja perdida durante o processo de recuperação da *watermark*; e no caso do código Hamming, permite a detecção da letra que sofreu interferência ou foi alterada propositalmente.

O subconjunto descrito anteriormente é uma das bases do método. A adição de uma marca d'água não perceptível em texto utiliza variações no formato das letras dos documentos. São pequenas variações que se tornam pouco perceptíveis quando inseridos em um documento completo. As diferenças das letras podem ser impostas através de fontes novas com divergências intencionais ou através de duas fontes textuais predefinidas em editores de texto que possuem letras com formas semelhantes, como a letra 'a' da fonte *Arial* e a mesma letra 'a' da fonte *Calibri*. A escolha dos caracteres que compõem o subconjunto deverá considerar características presentes no idioma em que o texto é escrito. Devem ser analisados aspectos como:

a. Frequência de ocorrência dos caracteres nas palavras do idioma;

b. Ocorrência de dígrafos, como: rr ou ss;

c. Um subconjunto C com muitos caracteres permite a inclusão de *watermark* em um menor fragmento de texto. Entretanto, pode comprometer a marca d'água, tornando mais frequentes e perceptíveis os caracteres alterados [18];

d. O tamanho dos caracteres no documento influenciará a percepção das divergências entre letras;

e. Distorções que possam acontecer nos caracteres durante o processo de digitalização.

A figura 1 exemplifica possíveis alterações no formato dos caracteres. As elipses em vermelho realçam as diferenças das letras. Apesar de ser relativamente fácil comparar a primeira linha com a segunda, deve ser considerado que as diferenças serão pequenas, a quantidade de letras ao redor será grande e o tamanho das letras será menor que o exemplo da figura 1.

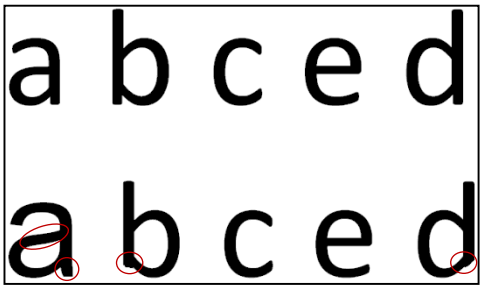

Fig. 1. Exemplo de caracteres com formas diferentes.

Definido a marca d'água, então é possível inserir o elemento de segurança ao texto a ser protegido. As figuras 2 e 3 ilustram o diagrama de inserção e recuperação das imagens nos documentos. As figuras exemplificam a utilização de um preâmbulo adicionado ao código identificador.

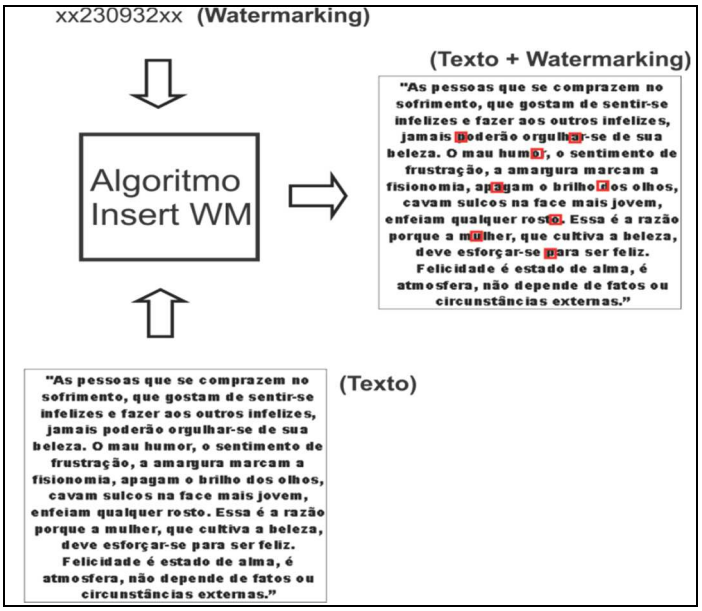

Fig. 2. Inserção da marca d'água no texto.

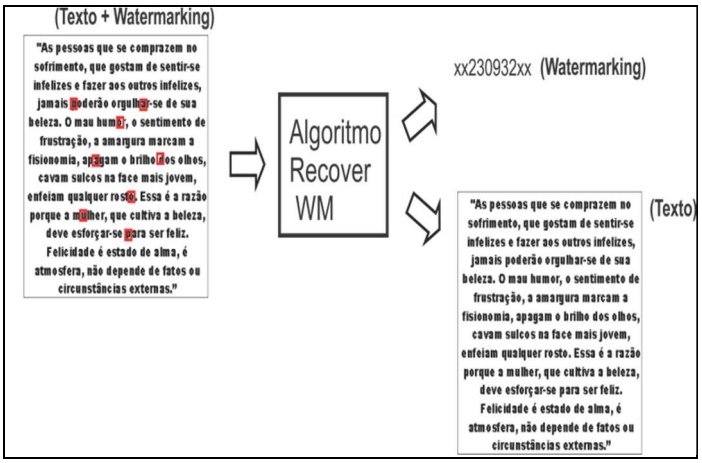

Fig. 3. Recuperação da marca d'água do texto protegido.

De posse da marca d'água e do texto, o algoritmo de inserção do elemento de segurança é iniciado (Tabela I). Como a *watermark* é uma sequência de bits, a cada representação '1' da marca d'água, o próximo caractere do texto, que está presente no subconjunto C, terá sua forma substituída. A cada representação '0', o próximo caractere do texto, que está presente no subconjunto C, terá sua forma original mantida. Desta maneira, a mudança de caracteres no texto dependerá da quantidade de números '1' presentes na *watermark*.

O algoritmo será executado sobre todo o texto, então após a execução do último dígito da *watermark* é realizada uma nova iteração da mesma marca d'água sobre os caracteres restantes do texto. Para evitar que seja percebida a alteração no texto por um leitor, algumas regras de modificação de caracteres são utilizadas:

- Caso ocorram casos de repetição de caracteres consecutivos, como: rr, ss, ee, oo, entre outras. O algoritmo ignora os caracteres;
- Caracteres em caixa-alta são ignorados;
- Títulos ou palavras em destaque que utilizem negrito, itálico ou que utilizem tamanho de fonte diferente do padrão do documento serão ignorados pelo algoritmo;
- Legendas de imagens, tabelas e notas de rodapé também são ignoradas pelo algoritmo.

Transformando essas especificações em algoritmos teríamos:

ALGORITMO PARA INSERÇÃO DE MARCA D'ÁGUA EM TEXTO

```
1. Obter Watermark;
2. Obter Subconjunto C; //Caracteres de forma variável
3. Enquanto texto não acabar Faça
4.   Próximo caractere do texto;
5.   Se letra pertence ao subconjunto C Então
6.     Se letra é um caractere válido Então //Regras
7.       Se próxima_letra_Watermark() == 1 Então
8.         Alterar forma da letra no texto;
9.       Fim_Se
10.    Fim_Se
11.  Fim_Se
12. Fim_Loop
```

ALGORITMO PARA RECUPERAÇÃO DE MARCA D'ÁGUA

```
1. Obter Subconjunto C;
2. Enquanto Texto não acabar Faça
3.   Próximo caractere do texto;
4.   Se letra pertence ao subconjunto C Então
5.     Se letra é um caractere válido Então  //Regras
6.       Se letra tem a forma alterada Então
7.         Watermark = Watermark + '1';
8.       Senão
9.         Watermark = Watermark + '0';
10.      Fim_Se
11.    Fim_Se
12.  Fim_se
13. Fim_Loop
```

Na linha 6 do algoritmo de inserção de *watermark* (Tabela I) e na linha 5 do algoritmo de recuperação de *watermark* (Tabela II), a verificação se o caractere é válido baseia-se nas regras descritas nesta seção para que o algoritmo ignore letras que são de tamanhos distintos, dígrafos, estão em caixa-alta ou fazem parte de alguma legenda.

Para ilustrar a utilização da técnica proposta foi definido um cenário de testes com duas marcas d'água com tamanhos diferentes. A primeira possui o tamanho de 16 bits, com preâmbulos de 4 bits e o código identificador de 8 bits. A segunda marca d'água utilizará uma palavra chave de 6 caracteres e será submetida ao algoritmo base64 [20], criando uma palavra de 64 bits no padrão ASCII. A função base64 é utilizada para demonstrar a possibilidade de utilização de uma função criptográfica sobre a marca d'água.

A marca d'água do primeiro teste será **1111100000101111**, enquanto a segunda marca d'água será **xx28xx**, que quando submetida à função base64 será **eHgyOHh4** em ASCII. Convertendo o texto para código binário, assumindo que cada letra é composta por oito bits, teremos **01100101 01001000 01100111 01111001 01001111 01001000 01101000 00110100**. O subconjunto de caracteres será: C = {**a, d, e, i, r**}, portanto esses caracteres poderão ser alterados no documento de acordo com as regras do algoritmo e da *watermark*

escolhida. O tamanho da fonte utilizando no texto de teste será 12 e fonte *Times New Roman*.

RESULTADOS E ANÁLISE

No exemplo, as letras do subconjunto C utilizam a fonte textual *Caladea* (Figura 4). Ressalta-se que foram utilizadas duas fontes textuais previamente definidas, o que não impossibilita que sejam utilizadas fontes novas, criadas especificamente para inserir a *watermark*.

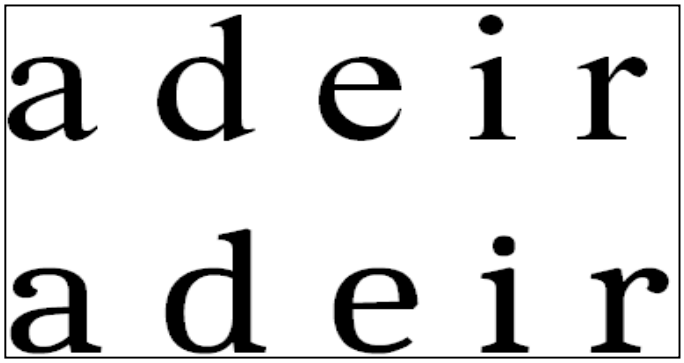

Fig. 4. Subconjunto C com os caracteres utilizados para teste. Na linha superior, os caracteres utilizam fonte *Times New Roman*. Na linha inferior utilizam fonte *Caladea*.

O texto utilizado para teste, como descrito na seção anterior, utiliza fonte *Times New Roman* de tamanho número 12, como ilustra a figura seguinte.

Balança enganosa é abominação para o SENHOR, mas o peso justo é o seu prazer. Em vindo a soberba, virá também a afronta; mas com os humildes está a sabedoria. A sinceridade dos íntegros os guiará, mas a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a justiça livra da morte. A justiça do sincero endireitará o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 5. Texto original antes da inserção da marca d'água.

Três diferentes testes foram realizados sobre o texto. Primeiramente, duas *watermarks* de tamanhos diferentes foram aplicadas a duas cópias do texto. Posteriormente, o texto com a marca d'água será exibido após ser impresso e digitalizado, exemplificando quatro resoluções diferentes no equipamento de digitalização. Por fim, o material será exibido através de duas fotografias de *Smartphones* com câmeras de 5 e 8 *megapixels*. Os resultados são exibidos nas subseções seguintes.

Watermarks de tamanhos diferentes

Como descrito anteriormente, uma marca d'água terá 16 bits (Figura 6), enquanto a outra usará 64 bits (Figura 7).

Para facilitar a visualização da aplicação das marcas d'água, a cada iteração da *watermark* no texto seu início e término será demarcada com retângulos, e as letras que forma alteradas pelo algoritmo serão representadas pelos valores binários '0' ou '1' de acordo com a correspondência da *watermark* (Figuras 8 e 9).

Balança enganosa é abominação para o SENHOR, mas o peso justo é o seu prazer. Em vindo a soberba, virá também a afronta; mas com os humildes está a sabedoria. A sinceridade dos íntegros os guiará, mas a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a justiça livra da morte. A justiça do sincero endireitará o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 6. Texto após a inclusão da *watermark* de 16 bits.

Balança enganosa é abominação para o SENHOR, mas o peso justo é o seu prazer. Em vindo à soberba, virá também a afronta; mas com os humildes está a sabedoria. A sinceridade dos íntegros os guiará, mas a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a justiça livra da morte. A justiça do sincero endireitará o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 7. Texto após a inclusão da *watermark* de 64 bits.

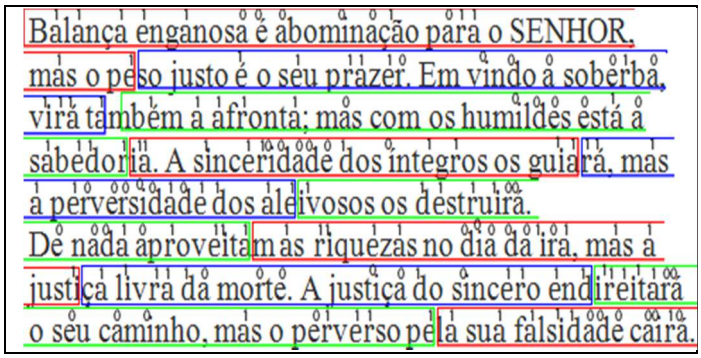

Fig. 8. Texto com a *watermark* de 16 bits e marcações.

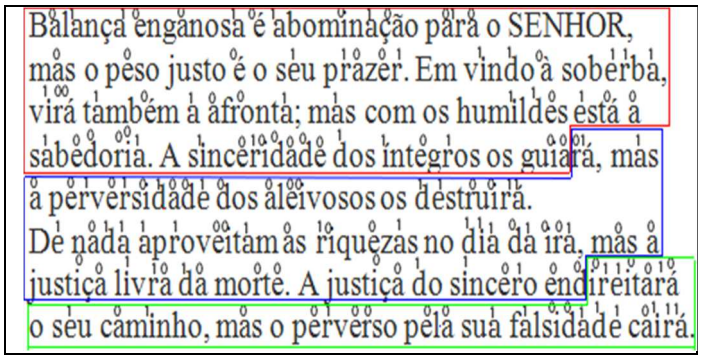

Fig. 9. Texto com a *watermark* de 64 bits e marcações.

Enquanto a marca d'água de 16 bits (Figura 8) utiliza pouco mais de uma linha para ser expressa, a segunda de 64 bits (Figura 9) utiliza quatro linhas para concluir sua primeira representação. A primeira *watermark* foi repetida nove vezes no trecho apresentado, enquanto a segunda apenas concluiu duas repetições no mesmo trecho de texto.

Os aspectos a serem destacados sobre a primeira *watermark* são:

- Capacidade maior de repetição;

- Possibilidade de recuperação da *watermark* em pequenos fragmentos de texto;
- Como esta utiliza preâmbulos, no pior caso de busca da marca d'água no texto serão verificados [2 * (tamanho em bits) - 1] caracteres pertencentes ao subconjunto C, e no melhor caso a quantidade de bits da *watermark*.

Os aspectos a serem destacados sobre a segunda *watermark* são:

- Utilização de uma função criptográfica antes da aplicação da técnica, a fim de proteger o código identificador;
- Apesar de ter um número de bits maior, isto aumenta a complexidade de recuperação da sequência binária. Além disso, a aplicação de uma função criptográfica pode retornar sequências de bits de tamanhos variados.

Digitalizações com diferentes DPI (dots per inch)

Foram selecionados quatro modos de digitalização disponíveis em uma impressora multifuncional. As opções selecionadas foram: digitalização em escala de cinza; resolução de 75 dpi, que é a menor resolução do equipamento; resolução de 200 dpi, que é a configuração padrão do dispositivo; e 600 dpi, que é a maior resolução do equipamento. Os resultados são apresentados em sequência nas figuras 10, 11, 12 e 13.

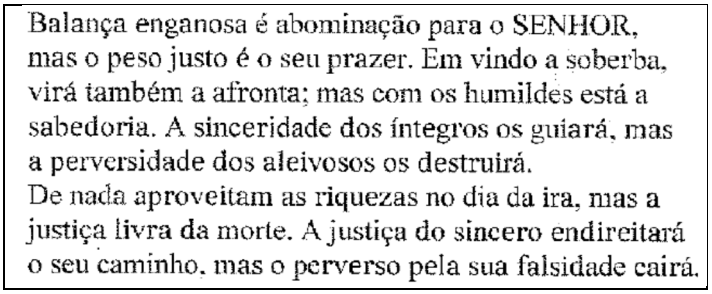
Balança enganosa é abominação para o SENHOR,
mas o peso justo é o seu prazer. Em vindo a soberba,
virá também a afronta; mas com os humildes está a
sabedoria. A sinceridade dos íntegros os guiará, mas
a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a
justiça livra da morte. A justiça do sincero endireitará
o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 10. Digitalização em escala de cinza.

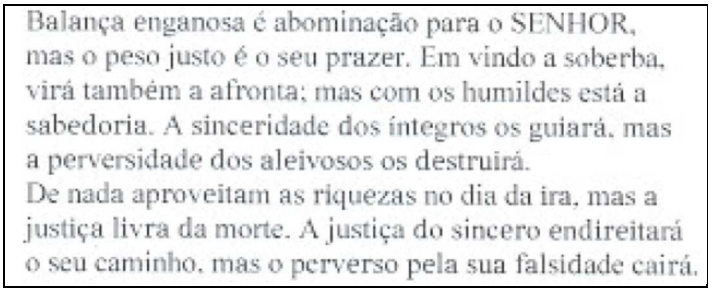
Balança enganosa é abominação para o SENHOR,
mas o peso justo é o seu prazer. Em vindo a soberba,
virá também a afronta; mas com os humildes está a
sabedoria. A sinceridade dos íntegros os guiará, mas
a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a
justiça livra da morte. A justiça do sincero endireitará
o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 11. Resolução de 75 dpi.

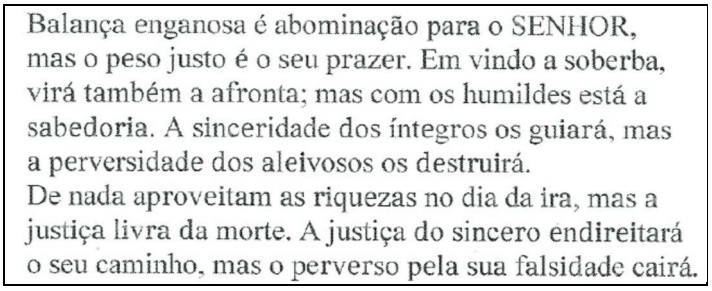
Balança enganosa é abominação para o SENHOR,
mas o peso justo é o seu prazer. Em vindo a soberba,
virá também a afronta; mas com os humildes está a
sabedoria. A sinceridade dos íntegros os guiará, mas
a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a
justiça livra da morte. A justiça do sincero endireitará
o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 12. Resolução de 200 dpi.

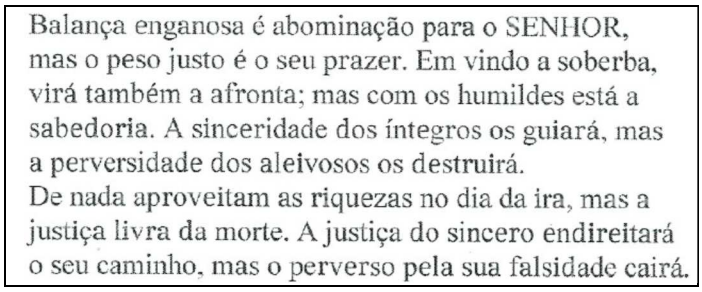
Balança enganosa é abominação para o SENHOR,
mas o peso justo é o seu prazer. Em vindo a soberba,
virá também a afronta; mas com os humildes está a
sabedoria. A sinceridade dos íntegros os guiará, mas
a perversidade dos aleivosos os destruirá.
De nada aproveitam as riquezas no dia da ira, mas a
justiça livra da morte. A justiça do sincero endireitará
o seu caminho, mas o perverso pela sua falsidade cairá.

Fig. 13. Resolução de 600 dpi.

As figuras 10, 11, 12 e 13 permitem concluir que a qualidade da digitalização de um documento pode trazer dificuldades para recuperação da *watermark*. Uma imagem muito degradada pode interferir na interpretação dos caracteres analisados. Entretanto, nos quatro cenários analisados, através da ampliação das imagens, ainda é possível verificar as diferenças entre os caracteres, como ilustra a figura 14. É perceptível que a degradação da imagem é mais acentuada com a resolução de 75 dpi, mas ainda existe a variação da silhueta da letra 'a' nos quatro cenários, o que a torna um caractere indicado para compor o subconjunto C. Propriedade esta que não é verdadeira para todos os caracteres de um alfabeto.

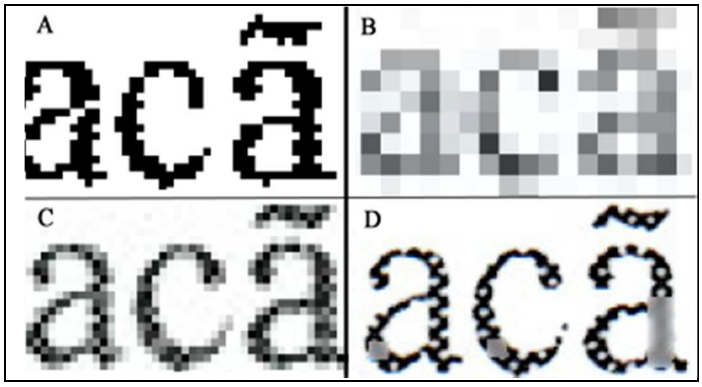

Fig. 14. Comparativo entre digitalizações do caractere 'a'. A) Digitalização em escala de cinza; B) Digitalização em resolução de 75 dpi; C) Digitalização em resolução de 200 dpi; D) Digitalização em resolução de 600 dpi.

Outro ponto a ser destacado é a distorção que algumas letras sofrem no processo de digitalização. Portanto, as variações dos caracteres selecionados para o subconjunto C devem ser avaliadas, para que não sejam ofuscadas pelo processo de digitalização.

Como a técnica de recuperação poderá ser empregada sobre imagens de baixa qualidade, o tratamento da imagem pode ser uma etapa anterior à aplicação do algoritmo. A manipulação de brilho e contraste de imagens podem acentuar as características das letras, como ilustra a figura 15, que foram submetidas a tratamento. Como é uma imagem em escala de cinza, através da equalização do histograma da imagem obtemos uma melhor distribuição de cores ao longo do histograma [21]. Assim, realçando as características das imagens e destacando as formas dos caracteres.

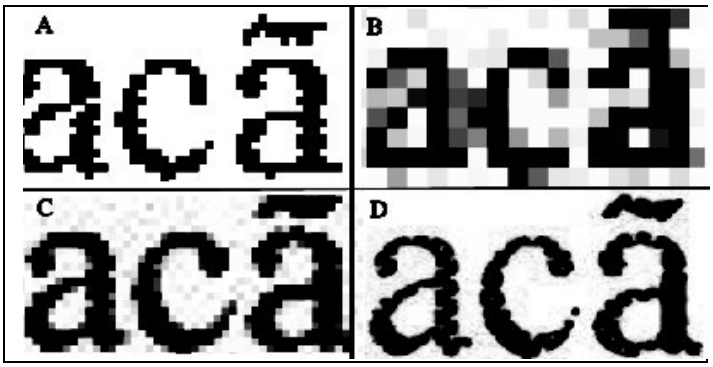

Fig. 15. Digitalizações do texto após o tratamento de brilho e contraste.

Imagem do texto através de fotografia

As imagens foram registradas por dois *Smartphones* com câmeras de 5 e 8 *megapixels*, respectivamente.

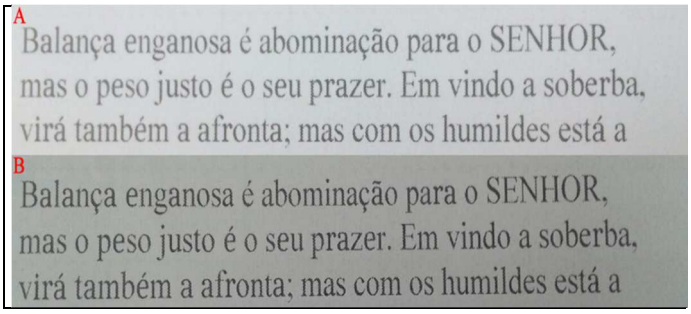

Fig. 16. Fotografia do texto após aplicação da marca d'água. A) câmera de 8 *megapixels*. B) câmera de 5 *megapixels*.

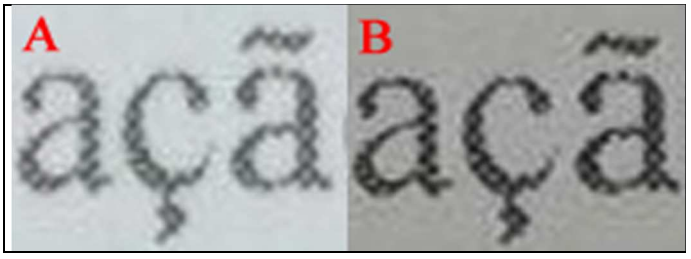

Fig. 17. Ampliação do texto fotografado após aplicação da marca d'água. A) câmera de 8 *megapixels*. B) câmera de 5 *megapixels*.

A análise da *watermark* através da fotografia permite concluir que as características identificadoras da marca d'água forma preservadas, permitindo a ampliação das imagens, mantendo os elementos identificadores dos caracteres, e obtendo resultados melhores que os textos digitalizados. Mesmo com o aumento da imagem (Figura 17), gerando distorções e ruído, ainda é possível identificar as letras que compõem a *watermark*.

Comparação entre a técnica proposta e outros métodos

Alguns critérios foram considerados, a fim de se comparar o método proposto com as técnicas encontradas na literatura. Os testes verificaram: a quantidade de informação introduzida no texto [11]; robustez contra ataques de alteração intencional de texto, deleção de sentença/linha e modificação de formatação do texto [19]; e detecção da *watermark* através de OCR. Foram selecionados os métodos *Line Shifting* e *Word-Shift Coding*.

Todas as técnicas foram aplicadas no texto da Figura 5. A Tabela III descreve a quantidade de bits inseridos no documentos com oito linhas, setenta e três palavras e trezentos e trinta e seis caracteres, desconsiderando os espaços em branco.

QUANTIDADE DE BITS POR TÉCNICA

Método	Nº de bits
Line Shifting	4
Word-Shift Coding	30
Método Proposto (subconjunto com 5 caracteres)	156

A análise quanto à robustez das técnicas, apenas o método proposto com a utilização de função criptográfica ou código de verificação de erros é capaz de detectar um ataque de alteração intencional do texto, entretanto somente utilizando o código de verificação de erros é possível recuperar o caractere alterado. Em relação à deleção de sentença/linha, os métodos *Line Shifting* e *Word-Shift Coding* não são robustos. O método proposto será robusto a depender do tamanho da *watermark* escolhida e dos caracteres do subconjunto C. Para o ataque de alteração de formatação, mudanças de alinhamento ou tamanho da fonte não comprometem a robustez do método proposto, o que não é verdade para os outros métodos testados.

Quanto à utilização de OCR para detecção das *watermarks*, foi utilizada a biblioteca *Tessaract* OCR [22], que quando executado com suas configurações padrões não detectou marcas d'água nas três técnicas.

CONCLUSÕES

A aplicação de *watermarking*, essencialmente, procura vincular uma produção, documental ou arquivo de mídia, a um autor. A técnica apresentada neste trabalho além de permitir a vinculação entre documento e autor, também possibilita a identificação do destinatário da cópia do documento. Dentro do contexto jurídico e policial, a guarda da prova documental pode afetar diretamente a condução de um processo ou inquérito. O vazamento de informações pode comprometer todo o trabalho realizado até o momento. Mas a partir do controle de cópia documental e da utilização de *watermarking*, são incluídos mais elementos fiscalizadores dentro do cenário descrito. Diferente de algumas técnicas de *watermarking* em documentos, a técnica proposta não precisa do documento original e nem de grandes fragmentos de texto para recuperação da *watermark*. Entretanto, a técnica ainda é vulnerável ao ataque de *re-typing*, que consiste em redigitar novamente o texto. Embora os vazamentos apresentados em veículos de comunicação prefiram apresentar os documentos em sua forma original com timbres, formatação e assinaturas do órgão emissor do documento.

Como a técnica utiliza imagens de textos para recuperação da marca d'água, é importante que o texto a ser analisado tenha qualidade em sua digitalização ou impressão compatível com a técnica apresentada. O texto submetido a processos de digitalização e impressão sequenciais influenciam a qualidade da informação e pode comprometer a eficiência da técnica. Uma alternativa para a baixa qualidade das imagens de texto seria o tratamento de imagem antes da aplicação da técnica.

A próxima etapa do trabalho será incluir a extração automatizada da *watermark* através de reconhecimento textual, como OCR, e melhorias na técnica para suportar ataques de *re-typing*. As variações de degradação dos caracteres durante o processo de digitalização precisam ser mensurados, a fim de

estipular um limite de qualidade do documento para ser submetido a um OCR ou mesmo ser passível de uma extração manual da marca d'água por meio da leitura do texto.

REFERÊNCIAS

[1] S. B. B. Ahmadi. *Digital Image Watermarking for Intellectal Property Protection. 4th International Scientific Conference of Iranian Academics*, 2014.

[2] S. S. Katariya. *Digital Watermarking: Review. International Journal of Engineering and Innovative Technology*, 2012.

[3] Y. Zhang e H. Qin. *A Novel Robust Text Watermarking For Word Document. 3º International Congress on Image and Signal Processing*, 2010.

[4] I. J. Cox et. Al. *Digital Watermarking and Steganography - 2º edition*. Ed. Elsevier, 2008.

[5] C. S. Woo. *Digital Image Watermarking Methods for Copyright Protection and Authentication. Thesis submitted in accordance with the regulations for Degree of Doctor of Philosophy*, 2007.

[6] S. P. Mohanty. *Digital Watermarking: A Tutorial Review. University of South Florida*, 1999.

[7] J. T. Brassil, S. Low e N. F. Maxemchuk. *Copyright Protection for the Electronic Distribution of Text Documents. Proceedings of the* IEEE, vol. 87, no. 7, pp.1181-1196, 1999.

[8] P. B. Devidas e P. N. Namdeo. *Text Watermarking Algorithm Using Structural Approach. World Congress on Information and Communication Technologies* - IEEE, 2012.

[9] M. Atallah et Al. *Natural Language Processing for Information Assurance and Security: An Overview and Implementations. Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop, Cork, Ireland*, pp. 51-65, 2000.

[10] D. Huang e H. Yan. *Interword Distance Changes Represented by Sine Waves for Watermarking Text Image. Transactions on Circuits and Systems for Video Technology* - IEEE, Vol. 11, NO. 12, 2001.

[11] R. Davarzani e K. Yaghmaie. *Farsi Text Watermarking Based on Character Coding. International Conference on Signal Processing Systems*, 2009.

[12] S. Kaur. *A Zero-Watermarking algorithm on multiple occurrences of letters for text tampering detection. Internacional Journal on Coomputer Science and Engineering*, 2013.

[13] N. Mir. *Copyright for web content using invisible text watermarking, Computers in Human Behavior*, 2014.

[14] J. Cummins, P. Diskin e S. Lau. *Steganography and Watermarking. The University of Birmingham*, 2004.

[15] D. Huang e H. Yan. *Interword Distance Changes Represented by Sine Waves for Watermarking Text Images. Transactions on Circuits and Systems for Video Technology* - IEEE, Vol. 11, NO. 12, 2001.

[16] H. Lu et al. *A New Chinese Text Digital Watermarking for Copyright Protecting Word Document. International Conference on Communication and Mobile Computing*, 2009.

[17] H. Y. Kim e J. Mayer. *Data Hiding for Binary Documents Robust to Print-Scan, Photocopy and Geometric Distortions. Computer Graphics and Image Processing*, 2007.

[18] A. L. Varna, S. Rane e A. Vetro. *Data Hiding in Hard-Copy Text Documents Robust to Print, Scan and Photocopy Operations.* ICASSP, 2009.

[19] Q. Chen et al. *Word Text Watermarking for IP Protection and Tamper Localization*. IEEE, 2011.

[20] J. Linn. *Privacy Enhancement for Internet Electronic Mail. Request for Comments* - 989, 1987.

[21] R. C. Gonzalez e R. E. Woods. Processamento Digital de Imagens. 3ª Edição. Editora Pearson, 2010.

[22]Tesseract OCR engine, disponível em: http://code.google.com/p/tesseract-ocr/.

Felippe Pires Ferreira¸ Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília-DF, Brasil, E-mail: felippe_pires@yahoo.com.br.