|  |  |  |  |
| --- | --- | --- | --- |
|  | **Universidade de Aveiro**  **Ano** 2021/2022 |  | |
| António Jorge  Ferreira Ramos | Validação da autenticidade de documentos impressos | | |
|  |  | | |
|  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Universidade de Aveiro**  **Ano 2021/2022** |  | |
| António Jorge  Ferreira Ramos | Validação da autenticidade de documentos impressos | | |
|  | Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica do Doutor André Zúquete, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro | | |
|  |  | |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
| o júri |  |
| presidente | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |
|  |  |
|  | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |
|  |  |
|  | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |
|  |  |
|  | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |
|  |  |
|  | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |
|  |  |
|  | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |
|  |  |
|  | Prof. Doutor João Antunes da Silva  professor associado da Universidade de Aveiro |

|  |  |
| --- | --- |
|  |  |
|  |  |
| agradecimentos /  acknowledgements | Agradeço a ajuda dos orientadores, amigos e familiares pelo apoio e auxílio prestado. |

|  |  |
| --- | --- |
|  |  |
|  |  |
| palavras-chave | Marca de água, códigos de barras, textual watermark, documentos impressos, segurança, processamento de imagem |
| resumo |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
| keywords | Watermark, barcodes, textual watermark, printed documents, security, image processing |
| abstract |  |

# Índice

[Índice ii](#_Toc123906441)

[Lista de tabelas iv](#_Toc123906442)

[Lista de Figuras v](#_Toc123906443)

[1. Introdução 1](#_Toc123906444)

[1.1. Descrição do sistema existente 1](#_Toc123906445)

[1.2. Objetivos 1](#_Toc123906446)

[2. Estado da Arte 2](#_Toc123906447)

[2.1. Código de barras 2](#_Toc123906448)

[2.1.1. Análise de cada tipo de código de barras 2](#_Toc123906449)

[2.2. Texto com watermark 4](#_Toc123906450)

[2.2.1. Ataque zero com mudança de texto 4](#_Toc123906451)

[2.2.1.1. Exemplo 5](#_Toc123906452)

[2.2.1.1. Vantagens 5](#_Toc123906453)

[2.2.1.2. Desvantagens 5](#_Toc123906454)

[2.2.2. Ataque zero sem mudança no texto 5](#_Toc123906455)

[2.2.2.1. Vantagens 6](#_Toc123906456)

[2.2.2.2. Desvantagens 6](#_Toc123906457)

[2.2.3. Documentos texto baseados em Eigenvalues 6](#_Toc123906458)

[2.2.3.1. Testes 6](#_Toc123906459)

[2.2.3.4. Vantagens 7](#_Toc123906460)

[2.2.3.5. Desvantagens 7](#_Toc123906461)

[2.2.4. Line-Shift Coding 7](#_Toc123906462)

[2.2.4.1. Word-Shift Coding 8](#_Toc123906463)

[2.2.4.2. Character Coding 8](#_Toc123906464)

[2.2.4.3. Vantagens dos métodos 8](#_Toc123906465)

[2.2.4.4. Desvantagens dos métodos 8](#_Toc123906466)

[2.2.5. ECL zero-based watermarking 8](#_Toc123906467)

[2.2.5.1. Vantagens 8](#_Toc123906468)

[2.2.5.2. Desvantagens 8](#_Toc123906469)

[3. Qrcode 10](#_Toc123906470)

[4. Código de barras 128 12](#_Toc123906471)

[5. Entropia documento com watermark 14](#_Toc123906472)

[6. Métodos implementados 16](#_Toc123906473)

[6.1. Primeira implementação 16](#_Toc123906474)

[6.2. Análise de quantidade de qrcode para análise forense 20](#_Toc123906475)

[6.3. Análise forense com 9 qrcode com processamento de imagem 21](#_Toc123906476)

[6. Proposta de desenvolvimento 26](#_Toc123906477)

[7. Arquitetura da solução 28](#_Toc123906478)

[7.1. Estrutura do documento 28](#_Toc123906479)

[7.2. Base de dados 28](#_Toc123906480)

[7.3. Criação do código de barras e Qrcode 29](#_Toc123906481)

[7.4. Operações a desempenhar pelo utilizador 29](#_Toc123906482)

[7.5. Processamento do documento 29](#_Toc123906483)

[7.6. Retificar documento 29](#_Toc123906484)

[7.7. Análise Forense 29](#_Toc123906485)

[8. Casos teste 31](#_Toc123906486)

[9. Conclusões e futuro trabalho 33](#_Toc123906487)

[10. Referências 35](#_Toc123906488)

[Apêndice A 37](#_Toc123906489)

[OCR 37](#_Toc123906490)

[Entropia 37](#_Toc123906491)

[Iron Software 38](#_Toc123906492)

# Lista de tabelas

[Tabela 1 - Estudo de código de barras lineares 2](#_Toc110074105)

[Tabela 2 - Possíveis escolhas 4](#_Toc110074106)

[Tabela 3 - Resultados Eigenvalues 7](#_Toc110074107)

[Tabela 6 - Classificação de erros do Qrcode 10](#_Toc110074108)

[Tabela 4 - Mudança de cor 13](#_Toc110074109)

[Tabela 5 - Performance 15](#_Toc110074110)

# Lista de Figuras

[Figura 1 – Exemplo de um ataque zero 5](#_Toc110089933)

[Figura 2 - Esquema de geração e extração da watermark 6](#_Toc110089934)

[Figura 3 - Antes do Watermark 6](#_Toc110089935)

[Figura 4 - Depois do Watermark 6](#_Toc110089936)

[Figura 5 - Comparação Watermark 7](#_Toc110089937)

[Figura 6 - Exemplo line-shift 7](#_Toc110089938)

[Figura 7 - Exemplo word-shift 8](#_Toc110089939)

[Figura 8 - Exemplo Character Coding 8](#_Toc110089940)

[Figura 9 - Funcionamento do algoritmo 8](#_Toc110089941)

[Figura 10 – Exemplo de um Qrcode 10](#_Toc110089942)

[Figura 11 - Estrutura Qrcode 10](#_Toc110089943)

[Figura 12 - Código de barra 128 10](#_Toc110089944)

[Figura 13 - Tabela ASCII 11](#_Toc110089945)

[Figura 14 - Zona livre 14](#_Toc110089946)

[Figura 15 - Zonas críticas 14](#_Toc110089947)

[Figura 16 - Ficheiro exemplar 15](#_Toc110089948)

[Figura 17 - Ficheiro processado 15](#_Toc110089949)

[Figura 18 - Resultado Qrcode 16](#_Toc110089950)

[Figura 19 - Resultado consola 16](#_Toc110089951)

[Figura 20 - Arquitetura solução proposta 20](#_Toc110089952)

# Introdução

No âmbito da tese de Mestrado de Engenharia Informática da Universidade de Aveiro, pretende-se que seja desenvolvido um módulo aplicacional a integrar com uma solução de gestão de informação classificada desenvolvida pela empresa iCreate Consulting.

Esta solução prevê a possibilidade de impressão de documentos classificados, devendo ser implementados mecanismos de marcação dos mesmos no sentido de garantir a sua autenticidade.

* 1. Descrição do sistema existente

Estes documentos vão ser tratados em áreas de segurança física destinadas para o efeito, onde para aceder o utilizador têm que ser pessoas idôneas (de confiança) e acesso a documentos com informação classificada deverá ter um certificado periódico.

Cada utilizador tem acesso a um determinado Posto de Trabalho, e aos documentos que foram destinados a esse mesmo Posto de Trabalho.

Por outro lado, esse utilizador só tem permissão para ver documentos até um nível de classificação para o qual está autorizado: por exemplo se só pode ver documentos até ao nível Confidencial, não terá acesso aos documentos Secretos ou de nível de classificação superior.

Apesar de ter acesso a documentos localizados num determinado posto, deve respeitar o princípio da “Necessidade de Conhecer", não devendo consultar indiscriminadamente os documentos a que tem acesso.

* 1. Objetivos

Para assegurar a sua autenticidade propõe-se dois métodos de segurança no documento.

O primeiro é criar um código de barras que permita validar de uma forma explicita se o documento teve origem no sistema, que possível terá mais informações que caracterize o documento, nomeadamente marcas de segurança. A ideia deste primeiro é passar uma pistola que leia o código de barras e averiguar se o código de barras é valido e se a informação do código de barras corresponde ao documento.

O segundo é criar uma watermark segura, única, por documento, gerado por um conjunto de informações do documento, que permitam auxiliar em aspetos mais forenses.

2. Estado da Arte

Este capítulo tem como objetivo agregar informações relativas ao tema da tese, sendo divido em dois subcapítulos, códigos de barras e *text* *watermarking*. O subcapítulo código de barras servirá para verificar se um documento é fidedigno, isto é comparar a assinatura do documento com a informação que estará no código de barras. O *text* *watermarking* verificará se a informação do documento foi modificada ou se é o original através de métodos de *watermarking.*

* 1. Código de barras

Código de barras é um modo de representar informação num estado visual, que possa ser visível sem de ter a necessidade de escrever texto. É classificado em duas categorias, linear (1D) e 2D. Foram analisadas as duas categorias, mas através de uma reunião com os orientadores, expôs-se que o papel impresso da empresa só tem espaço livre no *header* e no *footer*, e verificou-se que se colocasse um código de barras 2D ocupava mais espaço do que um 1D, então abordou-se com mais profundidade os tipos de códigos de barras que estão inseridos nesta categoria. Para ter uma noção mais percetível dos códigos de barras 1D foi criada uma tabela (tabela 1), cuja tem colunas que têm critérios para a comparação e porventura na seleção do mais adequado, como por exemplo as colunas com as designações nome, imagem cuja tem exemplos de códigos de barras (retiradas do Wikipédia), se o comprimento é variável ou fixo, o uso, e se é necessário aparelho próprio.

Existem duas maneiras de ler um código linear, com uma máquina própria (leitor de barras de barras) ou com um smartphone que tenha uma câmara e em alguns casos é necessário ter aplicações disponíveis na loja do sistema operativo (android e iOS) para determinados códigos de barras. Quando se realizou a reunião para expor as informações recolhidas, o coorientador informou que a empresa não permite uso de telemóveis dentro do seu estabelecimento, por questões de segurança, dando assim descartada a hipótese de sugerir a leitura de códigos de barras com uma aplicação mobile.

Em alguns códigos de barras existe um espaço reservado que não se pode alterar que se chama *checksum* ou *check* *digit*, cujo objetivo é verificar se a informação do código de barras foi bem gerada.

* + 1. Análise de cada tipo de código de barras

Este subcapítulo tem como objetivo expor a análise de vários códigos de barras lineares, cujos parâmetros de escolha foram comprimento, tamanho das barras, uso e se é necessário comprar um leitor próprio para a sua leitura. A análise recolhida está apresentada na tabela 1.

Na tabela 2 encontram-se os possíveis códigos de barras que porventura se possam utilizar. Estes códigos foram selecionados com o critério que o coorientador aconselhou e com o critério do tamanho máximo de caracteres que este pode guardar, podendo mudar no futuro de acordo com a escolha da empresa. A informação que vai se pretender guardar nos códigos de barras é a data do documento assinado, o local do posto de trabalho e um identificador único do documento.

Para ler códigos de barras é necessário ter um equipamento de leitura de código de barras, ligado a um computador que aquando do sucesso da sua leitura, abria uma página web, com o filtro do identificador do documento, que verificava as características do documento e um utilizador, que a empresa confie, ou seja uma identidade certificadora, comparava com o documento, se fosse diferente, significa que este foi danificado.

Tabela 1 - Estudo de código de barras lineares

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nome** | **Imagem** | **Fixo/**  **Variável** | **Tamanho das barras** | **Uso** | **Necessário Leitor (S/N)** |
| Post Code Austrália |  | Fixo | 4 | Correios Austrália | S |
| CodaBar |  | Fixo | 2 | Librarias | S |
| Code 25 |  | Variável | 2 | Librarias | S |
| Code 11 |  | Fixo | 2 | Telefones, já não se usa muito devido à sua antiguidade | S |
| Code 32 |  | Fixo | 2 | Farmácia | N |
| Code 39 |  | Fixo | 2 | Vários | S |
| Code 49 |  | Variável | Vários | Vários | S |
| Code 93 | Código de barras Código 93 - Simbologia nos códigos de barras | Cognex | Variável | Vários | Vários | N |
| Code 128 |  | Variável | Vários | Vários | N |
| EAN 2 |  | Variável | Vários | Revistas | N |
| EAN 5 |  | Variável | Vários | Livros | N |
| EAN-8 |  | Variável | Vários | Retail | N |
| GS1-128 |  | Variável | Vários | Vários | S |
| GS1 Databar |  | Variável | Vários | Vários | N |
| Intelligent Mail barcode |  | Fixo | 4 | Correios USA | S |
| ITF-14 |  | Variável | 2 | Encomendas | N |
| ITF-6 |  | Variável | 2 | Vários | N |
| JAN |  | Variável | Vários | Usado no Japão | S |
| Planet |  | Variável | Grande/  Pequeno | Correios USA | S |
| Plessey | MSI Plessey | Barcode Scanner SDK | Manatee Works | Variável | 2 | Catálogos, revistas, inventários | S |
| PostBar |  | Fixo | 4 | Correios Canada | S |
| PostNET |  | Fixo | Grande/  Pequeno | Correios USA | S |
| RM4SCC/KIX |  | Fixo | 4 | Correios | S |
| RM Mailmark C |  | Fixo | 4 | Correios | S |
| Universal Product Code |  | Variável | Vários | Retalhe | N |
| Telepen | Free Online Barcode Generator: Telepen Alpha | Variável | 2 | Librarias  da Inglaterra | N |

Tabela 2 - Possíveis escolhas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nome** | **Imagem** | **Tamanho máximo de dados** | **Tipos de dados** | **Vantagens** | **Desvantagens** |
| Code 39 |  | 43 | Letras e números | Ocupa menos espaço. | Não processa caracteres especiais tal como ç. |
| Code 128 |  | 48 | Letras, números e caracteres especiais | Faz encoding de strings muito eficiente.  Tem *checksum* | Precisa de espaço reservado no início, fim e *check symbols*.  Tem pouco limite de informação que pode guardar. |
| Code 93 | Código de barras Código 93 - Simbologia nos códigos de barras | Cognex | 30 | Letras, números e caracteres especiais | Open Source | Não tem checksum |

* 1. Texto com watermark

A *watermark* é um método que permite a salvaguarda de documentos originais e verificação da autenticidade de documentos ou imagens. É usada nos dias de hoje como por exemplo nas notas europeias, para impedir a impressão de notas não autorizadas. Existem muitos métodos e ataques de *watermarking*, neste documento vai ser apenas abordado o método *text* *watermark* e os métodos ataques zero devido às necessidades apresentadas pela empresa.

* + 1. Ataque zero com mudança de texto

O objetivo dos ataques zero é aumentar a fragilidade da *watermark* do documento contra ataques. Estes ataques tendem a alterar o texto do documento sem alterar a *watermark*. Estes ataques normalmente são de inserção, reordenação e remoção de texto.

Uma das soluções propostas pelos autores do artigo [1] foi ataque de substituição que explora a troca de ordem das palavras no texto. Esta solução é constituída por dois elementos. No primeiro é necessária uma lista de palavras que se pretende substituir no documento. O segundo é uma lista de palavras que vão estar depois da substituição do documento.

Em suma, escolhe-se um conjunto de palavras no documento que queremos substituir por novas palavras. Quando uma palavra é selecionada para se substituir todas as ocorrências da mesma são substituídas pela palavra nova. A percentagem do ataque determina a quantidade de modificação do documento. Se a palavra escolhida tiver várias ocorrências no documento, a percentagem vai aumentar.

A percentagem da contagem de palavras no documento é dada pela equação 1. A percentagem do ataque de substituição é dada pela expressão da equação 2.

Existem duas maneiras de escolher a lista de palavras para a substituição, normal e avançada. A normal consiste em selecionar palavras aleatórias do documento. A avançada consiste em selecionar as palavras com mesmo tamanho do que as palavras da lista de substituição.

Em conclusão, o *watermarking* textual proposto por [1] consiste em 3 passos, selecionar palavras do documento para substituição, selecionar as novas palavras e ocorrência de implementação do texto.

Text

Description automatically generated

Equação 1

Text, letter

Description automatically generated

Equação 2

* + - 1. Exemplo

Os autores do artigo [1] forneceram um exemplo de tipos de substituição como demonstra a figura 1. Podemos verificar que se substitui o pronome mais comum em inglês “The” por “close” na substituição normal e “job” na avançada. Este método permite assim ter palavras-chave ou informação principal que não se pretende perder numa maneira que só quem vai substituir as palavras sabe as originais.

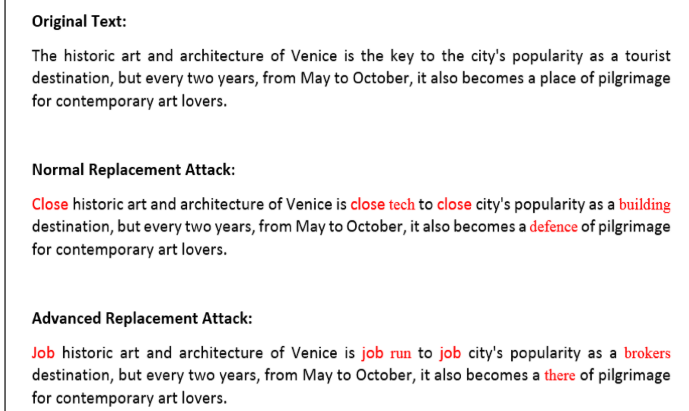


Figura 1 – Exemplo de um ataque zero

* + - 1. Vantagens

As vantagens deste método são puder alterar as palavras no texto através do utilizador e o algoritmo permite saber a zona que se alterou o documento.

* + - 1. Desvantagens

As desvantagens deste método são na leitura de texto, e não conseguir captar métodos de cópia de texto, por exemplo criar um documento totalmente novo com o mesmo texto.

* + 1. Ataque zero sem mudança no texto

A solução proposta pelos autores do artigo [2] consiste em usar as características do documento para gerar a *watermark* com o seguinte layout “autor:watermark:data:tempo” registada pela uma autoridade certificada (CA). O objetivo principal deste método é usar as palavras que tenham um tamanho maior do que quatro caracteres para proteger contra possíveis ataques.

A escolha de palavras maiores do que quatro caracteres é devido a generalidade dos ataques serem direcionados a palavras com tamanho menor do que quatro caracteres.

Para gerar a *watermark*, é necessário saber todas as frases de um documento, e assim percorrer cada palavra achando as palavras maiores do que quatro caracteres, quando acabar, guardar o primeiro caracter de cada dessas palavras e gerar a *watermark*.

Por exemplo na seguinte frase: “O José gosta muito de ler” a *watermark* ficava JGM (José Gosta Muito). Depois de percorrida todas as frases juntasse os *watermark* de cada frase, dando o *watermark* final.

Na figura 2 demonstra-se como é que a watermark é gerada e extraída. Deteta-se um ataque quando o *pattern* da *watermark* tem pelo menos ser 70% igual ao documento comparativo.

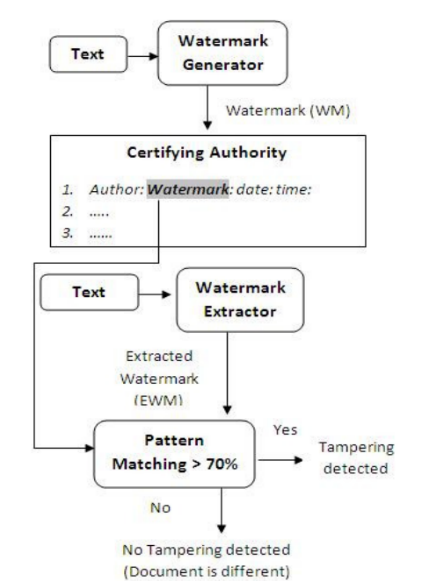


Figura 2 - Esquema de geração e extração da watermark

* + - 1. Vantagens

Tem uma entidade certificadora que vai possuir a *watermark* original, que vai ser a única pessoa a comparar os documentos, permitindo assim que o documento esteja só na posse de uma pessoa, diminuindo os ataques ou distribuição do documento. A composição da entidade certificadora é similar ao pretendido no código de barras.

* + - 1. Desvantagens

Tem testes realizados com ataques aleatórios, como inserir, apagar, reordenar e alterar, que para textos reais os resultados poderão ser diferenciados.

* + 1. Documentos texto baseados em Eigenvalues

A solução proposta pelos autores do artigo [3] envolve um algoritmo que guarda as posições de todas as palavras do documento numa matriz e o peso das mesmas em ASCII. Um documento normalmente é constituído por palavras, espaçamento, números e pontuações. Os autores consideram cada contagem para calcular o peso ASCII para gerar o esquema da *watermark* com base numa chave privada, que um utilizador autenticado vai verificar o documento recebido.

* + - 1. Testes

Os autores do artigo [3] fizeram testes com base num documento com 47 linhas. A figura 3 demonstra um excerto do texto antes de ter *watermark*. A figura 3 demonstra o texto depois da geração da *watermark.* Para simular um ataque foi alterada a palavra “OFF” para “ON” numa zona do texto e pelos resultados binários na figura 3, em primeiro está o original e de seguida o alterado, retira-se que os números são diferentes. Uma pequena alteração da palavra levou à mudança de 11 bits.

Os autores fizeram testes na mudança de vogais, consoantes, caracteres especiais, palavras, números pontuações, e alterações aleatórias no texto, obtiveram os resultados da figura 5.

Na coluna “tamper detection” da tabela 3 verifica-se que em todo o tipo de ataques, o algoritmo deteta a sua alteração sendo assim confiável.

Text

Description automatically generated

Figura 3 - Antes do Watermark

Text

Description automatically generated

Figura 4 - Depois do Watermark

Text

Description automatically generated

Figura 5 - Comparação Watermark

Tabela 3 - Resultados Eigenvalues

Table

Description automatically generated

* + - 1. Vantagens

Guarda a posição das palavras do texto numa matriz. Tem uma identidade certificadora que guarda o *watermark* e é a única pessoa que pode verificar se o documento é original.

* + - 1. Desvantagens

Difícil de compreensão, a execução do algoritmo leva muito tempo de execução e processamento para textos grandes.

* + 1. Line-Shift Coding

Os autores do artigo [4] expuseram outra forma de fazer watermarking que se denomina, *Line-Shift Coding*, que consiste em mover as linhas de texto de um documento para cima ou para baixo, enquanto as linhas adjacentes não são movidas. Na figura 6, demonstra-se um exemplo. Neste exemplo a linha do meio começada por “Effects…”, foi movida para baixo 1/300 inches, que equivale a 0.00846666667 cm. Esta mudança não é percetível ao olho humano, só através de um OCR, Optical Character Recognition, é que se conseguiria comparar ficheiros para verificar se existe ou não mudança nas linhas.

A screenshot of a computer

Description automatically generated with medium confidence

Figura 6 - Exemplo line-shift

2.2.4.1. Word-Shift Coding

Os autores do artigo [3] deram outra forma de fazer watermarking que se denomina *Word-Shift Coding*, que consiste em mover as palavras para a esquerda ou para a direita, enquanto as palavras adjacentes não são alteradas. A figura 7, tem-se um exemplo. A segunda linha, contém quatro palavras movidas com espaçamento de 1/150 inch, que equivale a 0.0169333333 cm, enquanto na primeira linha não se altera, a terceira é uma junção das duas anteriores. Como referido no ponto anterior esta mudança não percetível a olho humano, e para retificar a watermark recorrer a um OCR.

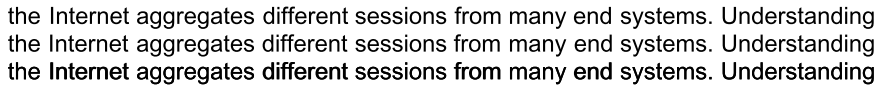


Figura 7 - Exemplo word-shift

2.2.4.2. Character Coding

Os autores do artigo [3] deram outra forma similar, referidas anteriormente, mas desta vez consistia em mover o caracter para cima ou baixo, enquanto adjacentes não se alteram. A figura 8, apresenta um exemplo. A primeira letra “e” da palavra “internet” foi movida para baixo 1/600 inches que equivale a 0.00423333333 cm. Como foi referido anteriormente noutros capítulos é necessário um OCR, para retificar as mudanças feitas no texto, isto é dispendioso em termos de tempo, e difícil de perceção.

A picture containing text

Description automatically generated

Figura 8 - Exemplo Character Coding

2.2.4.3. Vantagens dos métodos

Tem watermark invisível, ou seja, não é percetível para os olhos de um humano.

* + - 1. Desvantagens dos métodos

Uso de OCR para verificar documento, e ter acesso ao texto para inserir o espaçamento. Erros de impressão, por exemplo faltar letras no documento, pode levar à mal classificação de um documento, por exemplo ser falso, quando se tem a certeza de que o documento é original e não foi modificado.

* + 1. ECL zero-based watermarking

Os autores do artigo [5], propuseram um tipo de *watermarking*, que se denomina *ECL zero-based watermarking* que consiste num algoritmo que mantém o conteúdo original do texto do documento e constrói a *watermark* pela seleção de caracteres no documento. A watermark é guardada num sítio de confiança que se denomina *Certifying Authority*, para quando houver a necessidade de comparar o documento com a *watermark* para verificação da sua autenticidade, perceber que se o documento é original ou falsificado.

Os autores forneceram uma imagem (figura 9) em que se demonstra como funciona o algoritmo.

Diagram

Description automatically generated

Figura 9 - Funcionamento do algoritmo

2.2.5.1. Vantagens

Tem uma identidade certificadora que garante que o documento original esteja seguro com uma pessoa confiável na empresa.

2.2.5.2. Desvantagens

Falta de testes em cópias de documentos. Pouco efetivo com documentos que tenham um *effectiveness* *ratio* (ER) menor que 0.1. Effectiveness Ratio determina quais os caracteres para a geração da watermark.

# Qrcode

Atualmente vemos qrcode, nas faturas, em cartões de empresas e até mesmo na rua, maior parte das pessoas já viu, mas maior parte pode não saber o processo de leitura ou como surgiu o mesmo, para isso dediquei este capítulo ao qrcode.

O Qrcode (figura 10) surgiu em 1994 pela empresa Denso Wave, onde originalmente foi criado para categorizar peças de automóvel. Os Qrcode podem ter links para páginas web, texto, um endereço geográfico, uma imagem, um vídeo ou contacto telefónico.

Relativamente à estrutura do qrcode é constituído por 3 quadrados de deteção de posição (4.1. Figura 11) que permite a leitura em várias posições do scanner ou de um smartphone com câmara.

Por um padrão alinhamento (4.2. Figura 11) que corrige a distorção do qrcode em superfícies curvadas, o seu número varia consoante a informação contida.

Por padrões de temporização (4.3. Figura 11) que permite obter o tamanho da matriz de dados. Por informações da versão, que indica que versão do qrcode está a ser utilizada (1. Figura 11).

Por informações de formato, que contém informações sobre a tolerância de erros e o padrão da máscara de dados. Por códigos de dados e erros que podem ser do tipo L, M, Q, H, cujos estão apresentados na tabela 6.

Qr code

Description automatically generated

Figura 10 – Exemplo de um Qrcode

Graphical user interface, qr code

Description automatically generated

Figura 11 - Estrutura Qrcode

Tabela 4 - Classificação de erros do Qrcode

|  |  |
| --- | --- |
| **Nível de correção de erros** | **Percentagem de área danificada (%)** |
| L (Low) – Baixo | 7 |
| M (Medium) – Médio | 15 |
| Q (Quartile) - Quartil | 25 |
| H (High) – Alto | 30 |

1. Código de barras 128

O código de barras 128 (figura 12 – retirada do Wikipédia) é um tipo de códigos de barras linear que permite ter 128 caracteres da tabela ASCII (figura 13).

É constituído por zonas quietas 1- figura 12, por símbolo de início 2- figura 12, por dados que são colocados no qrcode 3- figura 12, por um símbolo de verificação que serve para verificar se o barcode foi bem gerado 4- figura 12, e por último um símbolo de stop, onde acaba o código de barras.

A largura do código de barras é dada pela quantidade de informação que este tem no ponto 3- figura 12.

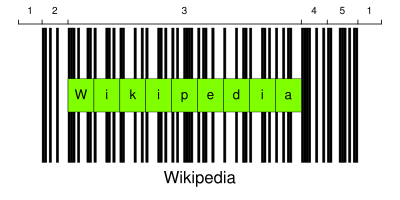


Figura 12 - Código de barra 128

Shape

Description automatically generated with medium confidence

Figura 13 - Tabela ASCII

# Entropia documento com watermark

O presente projeto de tese apresenta estudos e desenvolvimento de documentos com watermark, já que estes documentos são confidenciais, ou seja, não se pretende que qualquer pessoa que não seja autorizada os veja. Contudo no mundo atual existem pessoas que pretendem descobrir os documentos mais confidenciais e os divulgar, ou até mesmo, alterar para outros efeitos. Um programa de segurança deve evitar ao máximo que o algoritmo seja descoberto, ou seja, deve garantir que a entropia do algoritmo seja menor possível.

Entropia é definida como uma forma de medir o grau de incerteza de um sistema, ou seja, quanto mais um sistema for incerto menos probabilidade tem se de descobrir como um sistema ou algoritmo funciona. Para é necessário que o algoritmo desenvolvido seja mais discreto e mais aleatório possível.

# Métodos implementados

Antes de iniciar o desenvolvimento do programa e ter uma proposta final, foram colocadas em estudo várias técnicas.

Em primeiro o objetivo era usar o algoritmo de espaçamento de letras referido no estudo da arte no subcapítulo 2.2.4.1, contudo descartou-se porque era difícil de implementar e também a taxa de erros do espaçamento seria maior se o documento tivesse sido impresso torto ou com falta de letras, devido por exemplo à falta de tintores, ou a erros de impressão.

Em segundo pensou-se em alterar os caracteres do texto no documento, como foi colocado no subcapítulo 2.2.2., mas quando a empresa referiu que os documentos são imutáveis, ou seja, não permite a alteração de texto, descartou-se.

Em terceiro, pensou-se em ter um código de barras 128, que confirmaria o utilizador que assinou o documento, e um qrcode colocado no texto, que conteria as informações do documento, mas foi descartado devido ao espaço ocupado pelo código de barras no footer do documento, e porque a empresa queria que informação do documento estivesse toda agregada num só código de barras.

De seguida, implementou-se um qrcode, contendo como informação um índice de uma tabela da base de dados que contém as características do ficheiro, inserido numa posição aleatória do documento. Cujo irá ser abordado com mais detalhe no subcapítulo 6.1.

Entretanto surgiu a necessidade de criar funcionalidades de aprovação/rejeição de documentos gerados pelo programa, depois de se pensar na questão de o utilizador não querer que o qrcode esteja naquela posição documento, para isso adaptou-se o código e criou-se um programa novo que contém interface gráfica ao utilizador, levando a que o tempo de processamento demorasse mais devido à criação das instâncias visuais.

Como o processamento de imagem é lento surgiu a ideia de criar segmentos de retas para ligar posições de posicionamento do qrcode, abordado no subcapítulo 6.2.

* 1. Primeira implementação

Como referido anteriormente, a posição do qrcode vai ser aleatória no documento, podendo calhar em zonas que contém imagens, texto e zonas livres, que não tem texto nem imagens. Por exemplo na figura 14, que contém um documento de teste, o qrcode está inserido numa zona livre. Se o qrcode calhar numa imagem, é colocado noutra posição já que é impossível de ler qrcode com interferência de imagem. Se porventura calhar no texto, é possível ler, só que é necessário processar o documento, que obrigatoriamente tem de ser convertido para imagem, para permitir a alteração da cor dos pixéis.

Um leitor de qrcode só consegue ler se não existir interferência dentro do mesmo, para perceber melhor criou se a imagem 15, que demonstra as zonas críticas do qrcode num documento exemplar. Ou seja, quando aparece preto no cinza deve-se alterar para cinza, e quando aparece preto no branco, alterar para branco.

Para isso criou-se um algoritmo de processamento de imagem, que altera a cor do Qrcode no documento.

Este algoritmo consiste em processar o documento até encontrar a cor do qrcode, que é cinza-claro, se existir outro cinzento no documento a cor vai ser alterada também, que irá influenciar na leitura do qrcode se este sobrepor ou estar perto da cor cinza. Quando encontrar o qrcode, o algoritmo vai procurar os pixéis de cor preta, que simboliza o texto, de seguida lê a cor dos vizinhos – cima, baixo, esquerda, direita – mudando a cor do pixel consoante a tabela 5 e a cor pretendida como referido em cima, sendo o certo para mudar e errado para não mudar.

Para demonstrar os resultados do algoritmo, retirou-se um screenshot de dois ficheiros auxiliares o ficheiro com o qrcode (imagem 16) e o ficheiro processado com o algoritmo (imagem 17) com o tamanho do Qrcode de 75x75 pixels que equivale a aproximadamente 2cm. Pela análise da imagem 17 retira-se que o algoritmo fez a substituição da cor do qrcode com alguns restos, e que se passar um leitor de qrcode consegue-se ler, dependendo da qualidade telemóvel, é de salientar que o conteúdo do qrcode é exemplificativa. Contudo testar num documento com muito texto vai falhar devido ao tamanho do pixel do qrcode ser inferior ao caracter do texto, concluindo que o tamanho adequado é de 100x100.

A tabela 6 analisa a performance do algoritmo, avaliando o tempo de execução desde a criação do qrcode, inserção no documento, processamento de imagem à verificação da leitura. Na figura 19 tem-se o output do programa desenvolvido em c# em consola para efeitos de testes do documento, a parte vermelha e os outros documentos não se pode demonstrar devido à confidencialidade. Os prints representam o nome do ficheiro a processar, o resultado do id do ficheiro associado na base de dados e o tempo de execução do tipo “hh:mm:ss”.

A picture containing qr code

Description automatically generated

Figura 14 - Zona livre

Graphical user interface, text, application

Description automatically generated

Figura 15 - Zonas críticas

Tabela 5 - Mudança de cor

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **CIMA** | **BAIXO** | **ESQUERDA** | **DIREITA** |
| **CIMA** | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| **BAIXO** | Close with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| **ESQUERDA** | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill |
| **DIREITA** | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill |

Text

Description automatically generated

Figura 16 - Ficheiro exemplar

Diagram

Description automatically generated

Figura 17 - Ficheiro processado

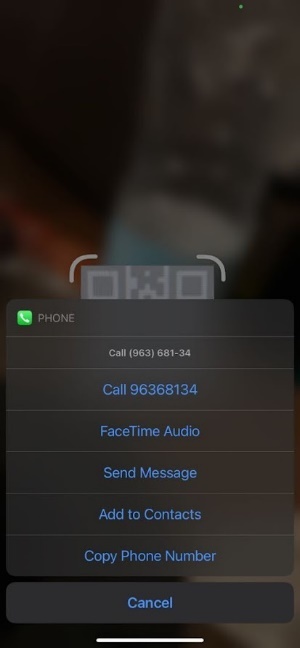


Figura 18 - Resultado Qrcode

Tabela 6 - Performance

|  |  |  |
| --- | --- | --- |
| **Tamanho Qrcode(pixels)** | **Ficheiro** | **Tempo (horas:minutos:segundos)** |
| 200x200 | NACIONAL\_1\_2022\_01000 | 00:00:41 |
| 100x100 | NACIONAL\_1\_2022\_01000 | 00:00:39 |
| 100x100 | NACIONAL\_19208\_2022\_29000 | 00:01:11 |
| 100x100 | NACIONAL\_23\_2022\_05000 | 00:01:10 |
| 100x100 | NATO\_190\_2021\_13000 | 00:00:34 |
| 75x75 | NACIONAL\_19208\_2022\_29000 | 6 min – falhou com 10 tentativas |

Text

Description automatically generated

Figura 19 - Resultado consola

* 1. Análise de quantidade de qrcode para análise forense

O objetivo deste capítulo é estudar o número ideal de qrcode para inserir no documento para guardar as letras com base no segmento de reta desenhados a partir dos quadrados de posição referidos no capítulo 3 figura 11. A figura 20 representa o que se pretende desenvolver para a análise forense.

Na tabela 7 está representado o número de estudo de Qrcode a usar no ficheiro, tem três colunas com as seguintes características “Nº de Qrcode”, “Nº Retas por quadrado de posição” e “Nº ligações”, as linhas representam os respetivos valores.

Em primeiro lugar pensou-se na possibilidade de haver só um segmento de reta a sair por quadrado de posição de um qrcode e retificar a interseção com a reta, mas como essa pode passar por várias letras, causa problemas de armazenamento e visualização difícil por utilizadores na análise forense. Para resolver esse problema e preencher todo o documento resolveu-se colocar 9 Qrcode numerados num novo documento PDF com base no documento PDF de entrada, e por cada quadrado de posição tem-se 3 segmentos de reta (figura 21), como decerto algumas retas se cruzam (círculos amarelos na figura 21) resolveu-se calcular o ponto de interseção e guardar o caracter que esteja inserido nesse ponto.

Diagram

Description automatically generated with medium confidence

Figura 20 - Análise forense pretendida

Tabela 7 – Estudo do número de qrcode a usar no ficheiro pdf

|  |  |  |
| --- | --- | --- |
| **Nº de Qrcode** | **Nº Retas por quadrado de posição** | **Nº ligações (Nº Retas que saem por quadrado ligadas ao quadrado de posicionamento)** |
| 2 | 1 | 9 |
| 2 | 3 | 9 |
| 4 | 1 | 18 |
| 6 | 1 | 27 |
| 8 | 1 | 36 |
| 9 | 3 | 81 |

Chart

Description automatically generated

Figura 21 - Segmentos de reta no ficheiro PDF

* 1. Análise forense com 9 qrcode com processamento de imagem

Antes do surgimento da ideia de usar os segmentos de reta no qrcode surgiu a proposta de realizar processamento de imagem no qrcode. O programa desenvolvido convertia o documento PDF numa imagem PNG (imagem 22), para alterar a cor dos pixéis, a conversão é feita com a utilização do package Freeware.Pdf2Png. De seguida calculava-se as posições dos qrcode que são as seguintes “2,25|8,25|12,25|2,32|8,32|12,32|2,41|8,41|12,41” cm, cuja ordem é do qrcode 1 para o 9.

Com base na posição de cada qrcode processava-se o documento na zona das posições (imagem 23), guardando a imagem do qrcode na posição para retificar se consegue ler a informação (conjunto de imagens com borda preta figura 24, cuja ordem da esquerda para direita e cima baixo é de qrcode 1 para 9). Através dos maus resultados e difícil perceção como também o tempo de execução da mesma descartou-se esta proposta.

Text, application

Description automatically generated

Figura 22 - Qrcode com processamento de imagem

Diagram

Description automatically generated

Figura 23 - Pós processamento

A picture containing room, scene, gallery

Description automatically generated

Figura 24 - Resultado processamento de imagem de 9 qrcodes

# Proposta de desenvolvimento

O desenvolvimento da solução para o problema consiste em duas partes, gerar o watermarking do ficheiro com base nos metadados, e verificar se o ficheiro recebido é original e não foi modificado.

Relativamente à verificação do ficheiro recebido como referido no capítulo 3.1 o tempo de execução do algoritmo de processamento de imagem aumenta devido à componente gráfica do Windows Form (C#) passa de 1minuto para 3 minutos. A solução criada envolve colocar um qrcode fixo no *header* do documento e outros numa posição aleatória, que irá ter uma distância mínima entre os mesmos, contendo a informação dos caracteres do texto através de segmentos de reta que irão ser desenhados e porventura escolher um caracter que apareça nesse segmento. Por último voltou-se a escolher o código de barras para ter a informação do documento e as posições dos qrcode inseridos no documento, isto para permitir o operador fazer uma análise rápida ao documento para retificar se o documento que recebeu é o mesmo que consta no código de barras, senão corresponder aí vai ser usado a análise forense que consiste em saber que zonas do documento foi alterado com ajuda dos qrcode.

Por uma questão de segurança o utilizador só tem acesso ao valor lido pelo leitor de barras, se estiver na base de dados do sistema, onde está a picar/ler o mesmo, isto descarta os mais curiosos caso queiram saber quem foi a pessoa que emitiu o documento fora do sistema.

1. Arquitetura da solução

Ao longo do desenvolvimento do programa de consola referido no capítulo 5.1 e necessidades da empresa, criou-se a seguinte arquitetura presente na figura 25.

O programa desenvolvido tem três camadas, a primeira é a camada do servidor destinada à base de dados que irá ter dados armazenados cerca dos ficheiros, códigos de barras, segmentos de reta traçados entre qrcode e posição dos caracteres no ficheiro de input. A segunda é a camada do programa onde acontece o desenvolvimento da criação do código de barras, do qrcode, do processamento do ficheiro para obter as posições dos caracteres no documento, da retificação do documento, e análise forense, esta também é responsável pelas conexões entre camadas, desencadeadas pelo utilizador. A última camada é onde acontece as ações que o utilizador pode fazer como por exemplo processar o documento ou retificar.

Diagram

Description automatically generated

Figura 25 - Arquitetura solução proposta

* 1. Estrutura do documento

O documento que vai dar entrada no sistema, é um documento PDF, que tem metadados - informações sobre o documento, como por exemplo, a classificação de segurança - introduzidos pelo utilizador no programa. Para realizar o processamento dos documentos estes têm de preencher dois requisitos o primeiro existirem na pasta predefinida, e o segundo existir metadados para o ficheiro no programa.

O programa trata dois estados de documentos.

* Documentos com watermarking para retificar a origem deles e se for necessário proceder a uma análise forense se o código de barras tiver informação diferente daquela que o documento apresenta, guardando os metadados do documento numa tabela da base de dados.
* Documentos sem watermarking que são utilizados para gerar a watermark, para que se consiga validar a autenticidade do documento. As watermark são colocadas na primeira página do documento, para a realização da tese.
  1. Base de dados

O intuito de usar base de dados no sistema é garantir segurança sobre as informações do código de barras e qrcode, como se referiu anteriormente, o código de barras vai ter um id aleatório representante das características do documento cujo vai ser inserido na base de dados aquando do seu processamento do documento para a criação da watermark.

A base de dados SQL criada localmente com o uso da ferramenta Microsoft SQL Server Management Studio 2018, tem um utilizador e base de dados dedicado apenas à consulta de informação pelo programa. As informações que se guardam são as características do documento, o código de barras, segmentos de reta traçados entre qrcode e posição dos caracteres no ficheiro de input.

* 1. Criação do código de barras e Qrcode

Para a criação do código de barras 128 e do qrcode foi utilizado um package para C# denominado Iron Barcode, que no anexo irá ser abordado mais em detalhe.

Como referido anteriormente o código de barras irá ter um id referenciador para o documento e para as posições dos Qrcode no documento. Em relação ao qrcode, irá ter a informação da versão do qrcode, caso se quiser mudar a estrutura futuramente e um identificador do documento para evitar perdas de informação caso o código de barras não seja visível.

* 1. Operações a desempenhar pelo utilizador

Um utilizador exerce 4 tipos de funções.

* Escolher entre
  + Processamento: que processa o documento sem a watermark para originar a watermark;
  + Retificar: que retifica o documento de input com a watermark para verificar se é o documento que se apresenta.
* Aceitar ou rejeitar o documento com watermark;
* Comparar a informação do documento com a watermark no submenu retificar, para determinar a sua autenticidade;
* Análise forense, para determinar a zona da alteração do documento, se por venturar o documento não ser autêntico.
  1. Processamento do documento

Com base no documento escolhido para processamento, gera-se um id aleatório identificador para a base de dados onde se vai inserir os metadados do documento inseridos previamente no programa, este programa só vai ter acesso um utilizador que seja fidedigno, ou seja, apenas pessoas que a empresa confie pode usar o código chave, fornecendo o programa, para inserir os metadados. Como a certo ponto pode existir colisões de criação do id, o programa verifica se já criou aquele id, se tiver gerado o mesmo número gera outro. De seguida cria-se os qrcode e o código de barras com a informação, referida no ponto 7.3. Seguidamente, coloca-se os qrcode numa posição fixa do documento e o código de barras no footer do mesmo.

Previamente lê-se as posições dos carateres do ficheiro para inserir na base de dados e traça-se retas invisíveis para calcular o ponto de interseção, caso o ponto calhe em alguma letra, a informação é guardada na base de dados para a análise forense.

Por fim o utilizador pré-visualiza o documento e pode aceitar ou rejeitar o documento com watermarking gerado, se aceitar o documento é guardado em formato PDF com o nome do documento inserido com adição “\_qrcode\_dd\_mm\_yy\_hh\_m\_ss”, sendo d:dia, mm:mês, yy:ano, hh:hora, m: min, ss:segundos, se rejeitar todo o processo é descartado, guardando na base de dados e de seguida o documento é processado novamente.

* 1. Retificar documento

Para retificar um documento é necessário que o documento escolhido para retificação, verificar se um documento é original, esteja inserido na base de dados. Caso esteja, o programa, lê o código de barras 128 e mostra a informação contida para o utilizador numa janela onde é possível ter a visualização do documento com a watermark escolhido e algumas informações sobre as características do documento (metadados). Cabe ao utilizador averiguar se o documento e as características demonstradas são diferentes ou iguais, se determinar que são diferentes o utilizador pode proceder à análise forense do documento, referido no ponto 6, determinando as zonas que foram alteradas no documento.

* 1. Análise Forense

Na análise forense resolveu-se por criar 3 segmentos de reta

# Casos teste

# Conclusões e futuro trabalho

# Referências

[1] M. B. Mohd, S. Mohd, R. Tanzila, and S. A. Rehman, “Replacement Attack: A New Zero Text Watermarking Attack,” *3D Res.*, vol. 8, 2017, doi: 10.1007/s13319-017-0118-y.

[2] Z. Jalil, A. M. Mirza, and H. Jabeen, “Word length based zero-watermarking algorithm for tamper detection in text documents,” *ICCET 2010 - 2010 Int. Conf. Comput. Eng. Technol. Proc.*, vol. 6, no. May, 2010, doi: 10.1109/ICCET.2010.5486185.

[3] T. Rethika, I. Prathap, R. Anitha, and S. V. Raghavan, “A novel approach to watermark text documents based on eigen values,” *2009 Int. Conf. Netw. Serv. Secur. N2S 2009*, no. c, pp. 1–5, 2009.

[4] J. T. Brassil, S. Low, and N. F. Maxemchuk, “Copyright protection for the electronic distribution of text documents,” *Proc. IEEE*, vol. 87, no. 7, pp. 1181–1196, 1999, doi: 10.1109/5.771071.

[5] T. Saba, M. Bashardoost, H. Kolivand, M. S. M. Rahim, A. Rehman, and M. A. Khan, “Enhancing fragility of zero-based text watermarking utilizing effective characters list,” *Multimed. Tools Appl.*, vol. 79, no. 1–2, pp. 341–354, Jan. 2020, doi: 10.1007/S11042-019-08084-0/TABLES/5.

[6] “About Us | IronSoftware.com.” https://ironsoftware.com/about-us/ (accessed Jul. 26, 2022).

[7] “C# .NET Barcode Quickstart Guide, Code Examples | Iron Barcode.” https://ironsoftware.com/csharp/barcode/examples/barcode-quickstart/ (accessed Jul. 26, 2022).

[8] “C# Tesseract OCR In 1 Line of Code | Iron OCR.” https://ironsoftware.com/csharp/ocr/examples/simple-csharp-ocr-tesseract/ (accessed Jul. 26, 2022).

Apêndice A

OCR

Entropia

Iron Software

Para desenvolver o programa foi necessário usar tecnologias de criação de Qrcode e leitura de Qrcode. Para isso usou-se um software open-Source, que tem a opção de comprar a licença para efeito de segurança e apoio, denominado Iron Software

[6]. O Iron software é usado por mais de 6.9 milhões de utilizadores tendo vários produtos como o Iron PDF, Iron OCR, Iron OCR, Iron Barcode, Iron XL, Iron Updates e Iron suite, listados na imagem abaixo. Sendo que neste projeto se usou o Iron OCR para leitura do qrcode e Iron barcode para gerar o qrcode.

Optou-se por usar o Iron Barcode [7] para gerar o qrcode devido a fornecer documentação e pelas propriedades de geração do qrcode, permitindo mudar o estilo do qrcode, como por exemplo a cor do qrcode, também permite colocar o qrcode no ficheiro PDF, mas eu optei por usar outro método, já que alguns documentos PDF são protegidos e neste pacote não tem opção de escrever neles sem saber a password.

Para ler/detetar o qrcode no ficheiro PDF optou-se por usar o Iron OCR [8], cujo procura num documento PDF o qrcode, e devolve o resultado.

