COPPE/UFRJ

ALGORITMOS DE ESTEGANOGRAFIA VIA RESTAURAÇÃO ESTATÍSTICA EM IMAGENS DIGITAIS

Gabriel Mayrink da Rocha Hospodar

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: José Gabriel Rodríguez Carneiro Gomes

Rio de Janeiro

Fevereiro de 2009

ALGORITMOS DE ESTEGANOGRAFIA VIA RESTAURAÇÃO ESTATÍSTICA EM IMAGENS DIGITAIS

Gabriel Mayrink da Rocha Hospodar

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. José Gabriel Rodríguez Carneiro Gomes, Ph.D.

Prof. Mariane Rembold Petraglia, Ph.D.

Prof. Murilo Bresciani de Carvalho, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

FEVEREIRO DE 2009

Hospodar, Gabriel Mayrink da Rocha

Algoritmos de Esteganografia via Restauração Estatística em Imagens Digitais/Gabriel Mayrink da Rocha Hospodar. – Rio de Janeiro: UFRJ/COPPE, 2009.

XI, 73 p.: il.; 29, 7cm.

Orientador: José Gabriel Rodríguez Carneiro Gomes

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2009.

Referências Bibliográficas: p. 57 – 60.

1. Esteganografia. 2. Casamento de Histogramas. 3. Marca d'Água. I. Gomes, José Gabriel Rodríguez Carneiro. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

À minha família.

Agradecimentos

Agradeço, primeiramente, a Deus.

Agradeço ao governo Brasileiro que, através do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), acreditou e investiu neste trabalho. Agradeço à COPPE/UFRJ, por ter me propiciado um ensino de excelência. Agradeço ao Laboratório de Processamento Analógico e Digital de Sinais (PADS) da COPPE/UFRJ pela excelente infra-estrutura que me foi disponibilizada para a realização deste projeto.

Agradeço ao brilhante Professor, orientador e amigo José Gabriel R. C. Gomes. Muito obrigado pelos ensinamentos passados sempre com muita clareza e simplicidade. Agradeço bastante, também, pelas incontáveis e preciosas horas semanais dedicadas às discussões relacionadas ao nosso trabalho. Foi uma honra trabalhar com uma pessoa extremamente inteligente, dedicada e acessível. Sou bastante grato pelos conselhos profissisonais e pelas diversas portas que este trabalho me abriu.

Agradeço aos ilustres Professores Mariane Rembold Petraglia e Murilo Bresciani de Carvalho, por terem aceitado o convite para participar da banca de defesa desta dissertação de mestrado. Muito obrigado pelas valiosas e acertadas sugestões de melhorias no texto, que enriqueceram bastante o trabalho.

Agradeço à minha família, de forma geral, por ter sido bastante paciente durante o meu período de estudos e por sempre ter me incentivado. Agradeço também aos amigos e colegas, que, de alguma forma, foram úteis para a realização do meu mestrado.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ALGORITMOS DE ESTEGANOGRAFIA VIA RESTAURAÇÃO ESTATÍSTICA EM IMAGENS DIGITAIS

Gabriel Mayrink da Rocha Hospodar

Fevereiro/2009

Orientador: José Gabriel Rodríguez Carneiro Gomes

Programa: Engenharia Elétrica

Sistemas de ocultação de marcas d'água em imagens digitais têm se mostrado cada vez mais importantes para aplicações como proteção de direitos autorais e segurança de informações militares. Este trabalho propõe três abordagens de sistemas de esteganografia via restauração estatística. Objetiva-se inserir o maior número de bits de informação possível nos coeficientes das DCTs dos blocos 8×8 de imagens digitais, sem que as estatísticas dessas imagens sejam alteradas. Neste trabalho, a estatística relevante é a função massa de probabilidade do conjunto de coeficientes das DCTs da imagem. O processo de escrita da marca d'água não deve introduzir distorções visualmente perceptíveis nas imagens hospedeiras. Os resultados obtidos apontam que uma das abordagens propostas oculta, de forma segura, uma quantidade expressiva de bits de informação por *pixel* da imagem hospedeira.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ALGORITHMS FOR DIGITAL IMAGE STEGANOGRAPHY VIA STATISTICAL RESTORATION

Gabriel Mayrink da Rocha Hospodar

February/2009

Advisor: José Gabriel Rodríguez Carneiro Gomes

Department: Electrical Engineering

Digital image watermarking systems have gained importance in applications such as copyright protection and security of military information. This work proposes three approaches for steganographic systems via statistical restoration. The main objective is to embed the maximum amount of bits of information into the coefficients of the 8×8 DCT blocks of digital images. The statistics of these images should not be changed. In this work, the probability mass function of the DCTs coefficients set is the relevant statistic. The watermark embedding process should not visually degrade the host images. Results show that one of the proposed approaches hides, in a secure way, an expressive amount of bits of information per host image pixel.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O rápido crescimento na demanda e no consumo de conteúdos em meios digitais tem mudado o estilo de vida da sociedade. As mudanças podem ser observadas na forma em que as pessoas se divertem, se comunicam, assimilam ou disseminam informação. Assim, diversas pesquisas têm focado em aplicações que envolvem: meios de garantir comunicações seguras, proteção de direitos autorais, autenticidade de dados digitais, etc.

Este trabalho propõe três abordagens de sistemas de esteganografia para imagens digitais. Objetiva-se ocultar, tanto visível quanto estatisticamente, a máxima quantidade de informação nos coeficientes do domínio da transformada discreta do cosseno (DCT) de imagens digitais. Uma parte desses coeficientes é destinada à escrita da mensagem, enquanto que o restante dos coeficientes é destinado à restauração ou compensação da estatística da imagem original. A compensação estatística faz com que a imagem marcada com a máxima quantidade de informação não apresente evidências de carregar uma mensagem escondida. A estatística considerada foi a função massa de probabilidade (PMF) dos coeficientes da DCT.

O Capítulo 2 apresenta um resumo sobre a área de Ocultação de Informação, que inclui a esteganografia. Um breve histórico da área é apresentado, assim como a terminologia utilizada no trabalho e um modelo atual para esteganografia.

O Capítulo 3 apresenta as Propostas A, B e C de sistemas de esteganografia via restauração estatística. Cada uma dessas propostas descreve o método de escrita e de compensação que elas utilizam. Este capítulo também descreve a abordagem utilizada por Sarkar e Manjunath em [2], que foi tomada como motivação para este

trabalho.

O Capítulo 4 apresenta a descrição dos testes feitos sobre as abordagens propostas. Dois sistemas esteganalistas baseados em redes neurais artificiais foram utilizados para testar as abordagens apresentadas no Capítulo 3. O principal objetivo desses testes é verificar a máxima quantidade de informação que pode ser ocultada nas imagens de forma segura. Os resultados obtidos a partir desses testes são apresentados no Capítulo 4.

O Capítulo 5 apresenta as conclusões deste trabalho, discutindo os resultados obtidos no Capítulo 4. Por fim, o Capítulo 5 apresenta uma breve proposta de trabalhos futuros.

Capítulo 2

Ocultação de Informação

Técnicas de ocultação de informação são importantes em diversas aplicações. Por exemplo, aplicações como as de sistemas militares de comunicações requerem a utilização de técnicas de segurança que, em vez de somente encriptar a mensagem, também ocultem o transmissor, o receptor ou a própria existência da mensagem. Imagens, áudio e vídeo digitais vêm sendo marcados de forma imperceptível com informações de direitos autorais, com números de série ou simplesmente com o objetivo de evitar cópias não autorizadas do conteúdo. Técnicas similares são usadas por alguns sistemas de telefonia celular e por esquemas propostos para eleições através de tecnologias digitais. Muitas das técnicas propostas no novo campo de ocultação de informação têm suas origens históricas na antiguidade. A principal referência utilizada neste capítulo foi o artigo *Information Hiding - A Survey* [3], de Fabien Petitcolas et al.

2.1 Introdução

Muitos consideram que uma comunicação se torna segura através apenas da utilização de criptografia. A prática tem mostrado que isto nem sempre é adequado. A ocultação de informação levanta menos suspeita do que a criptografia.

Por exemplo, tanto Æneas, o estrategista, quanto John Wilkins[1] preferiam a ocultação de informação à criptografia. Vale notar que as técnicas de criptografia

[1] John Wilkins (1614 - 1672) foi um clérigo anglicano e cientista. Chefiou faculdades na Universidade de Oxford e na Universidade de Cambridge. Foi um dos fundadores da *Royal Society*, em Londres.

moderna foram desenvolvidas durante o Renascimento[2]. Porém, a preferência pela utilização de técnicas de ocultação de informação persiste em diversos contextos operacionais até os dias de hoje. Por exemplo, uma mensagem digital criptografada trocada entre um traficante de drogas e alguém que ainda não esteja sob suspeita, pode ser facilmente interceptada. O trabalho policial fica reduzido a decriptar a mensagem.

Pode-se afirmar que o estudo de segurança nas comunicações inclui não somente criptografia, mas também segurança no tráfego de informação, cuja essência é a ocultação de informação [3]. Esta área inclui tecnologias como a de Espalhamento Espectral (*Spread Spectrum*), que é amplamente utilizada em sistemas militares de comunicação, para prevenir a localização dos transmissores e receptores.

Uma importante subárea de ocultação de informação é a esteganografia. A criptografia diz respeito à proteção do conteúdo das mensagens, enquanto a esteganografia trata de esconder a sua existência. A etimologia da palavra *esteganografia*, de raízes gregas, é composta por *estegano*, que significa ocultada, escondida ou coberta, e por *grafia*, que significa escrita. Logo, esteganografia significa "escrita escondida" e pode ser interpretada como a ocultação de uma informação sob outra informação.

Como imagens, áudio e vídeos são disponibilizados em meios digitais, a facilidade com que é possível fazer cópias ilegais, muitas vezes, leva à pirataria em larga escala. Isto significa um grande problema para a indústria, pois ela perde uma quantia enorme de dinheiro. Assim, a principal força que tem impulsionado a pesquisa na área de ocultação de informação é a proteção de direitos autorais. A primeira conferência acadêmica na área de ocultação de informação foi organizada em 1996, em Cambridge, Inglaterra [4].

Pesquisas recentes em marcas d'água digitais (*digital watermarking*) e em inserção de impressões digitais (*fingerprinting*) têm sido bastante significativas. Impressões digitais lidam com a ocultação de números de série, que podem ajudar a identificar violadores de direitos autorais. Marcas d'água digitais lidam com a ocultação de informações de direitos autorais, que servem para processar os violadores.

[2]Período da história da Europa entre fins do século XIII e meados do século XVII no qual diversas transformações ocorreram, assinalando o final da Idade Média e o início da Idade Moderna.

A indústria de DVD tem incentivado a busca de soluções para esquemas de inserção de informações de direitos autorais, para reforçar os sistemas de gerenciamento de cópias dessas mídias. A idéia é que aparelhos de DVD disponíveis para consumidores permitam copiar, ilimitadamente, vídeos caseiros e programas de TV, porém sem permitir os abusos da pirataria comercial. Assim, vídeos caseiros não devem ser marcados, programas de TV devem ser marcados de forma a permitir apenas uma cópia e vídeos comerciais devem ser marcados para jamais serem copiados [5].

Há uma infinidade de aplicações em diversas áreas que podem se beneficiar das técnicas de ocultação de informação. Profissionais de Direito e agências de inteligência têm se mostrado bastante interessados no entendimento dessas tecnologias e nas suas fraquezas, de tal forma a poderem detectar e rastrear mensagens escondidas. Tentativas recentes de alguns governos de limitar o direito de liberdade de expressão na Internet e de limitar o uso civil de técnicas de criptografia têm incentivado o desenvolvimento de técnicas para comunicação anônima através da rede, incluindo servidores de e-mail anônimos e de *Web proxies*. Esquemas para eleições digitais podem fazer uso de técnicas de comunicação anônima, para manter o sigilo do voto.

Além desses exemplos, podem-se citar agências de propaganda que utilizam técnicas como e-mails forjados para enviar inúmeras mensagens não solicitadas e indesejadas, evitando automaticamente respostas de usuários irritados.

Agências de segurança e órgãos militares precisam realizar, seguramente, comunicações sem obstruções. A detecção de um sinal num campo de batalha moderno, mesmo que ele esteja encriptado, pode levar rapidamente a um ataque do inimigo. Por esta razão, sistemas militares de comunicação usam técnicas como modulação de espalhamento espectral ou modos de propagação que exploram rastros ionizados de meteoros entrantes na atmosfera para estabelecer canais de comunicação secretos.

Criminosos também estão relacionados com o assunto. Suas tecnologias preferidas incluem telefones celulares pré-pagos, telefones celulares que tenham sido modificados para frequentemente mudarem suas identidades e ataques a roteadores, de forma que ligações ou conexões sejam redirecionadas.

As posições éticas na área de ocultação de informação não são tão claras quanto as da área de criptografia, em que bons usuários tentam manter suas comunicações

privadas enquanto que maus usuários tentam decifrar os conteúdos dessas comunicações. Por exemplo, usuários inofensivos da Internet podem precisar fazer uso de mecanismos de comunicações anônimas, para fazer denúncias ou para votarem secretamente em eleições *online*. Pelo outro lado, tais sistemas de comunicações anônimas não devem fornecer mecanismos que facilitem ataques por agentes maliciosos sobre tais infra-estruturas. Ferramentas para a indústria devem ser cuidadosamente projetadas, já que agentes mal-intencionados podem utilizá-las com o objetivo de passar informações escondidas em dados não suspeitos através de redes públicas.

Por último, aplicações não competitivas dessas tecnologias podem incluir marcação de faixas de áudio com informações de compra do CD ou DVD contendo a música de forma que uma pessoa ouvindo rádio em seu carro possa realizar essa compra simplesmente apertando um botão.

2.2 Terminologia

Como visto anteriormente, o interesse pelas áreas de *watermarking* e de *fingerprinting* por diferentes comunidades científicas e corporativas tem crescido consideravelmente. Como se trata de um novo campo de pesquisa, isso tem gerado confusão no que diz respeito à terminologia. A seguir será apresentada uma breve descrição sobre terminologia de acordo com a primeira conferência internacional de ocultação de informação [4].

O modelo genérico de ocultação de informação sob outra informação pode ser descrito da seguinte maneira. A informação inserida ou escrita é a mensagem que se deseja enviar secretamente. Geralmente, a informação é escondida numa mensagem inofensiva, que é chamada de objeto-camuflagem (*cover-text*) ou de imagem-camuflagem (*cover-image*), gerando o estego-objeto (*stego-text*) ou a estego-imagem (*stego-image*), no caso deste trabalho. Uma estego-chave é usada para controlar o processo de ocultação de forma a restringir a detecção ou a recuperação do dado escondido somente por agentes que a conheçam.

O propósito da esteganografia é realizar uma comunicação camuflada entre duas partes, cuja existência seja desconhecida a terceiros. Um ataque bem-sucedido sobre um sistema de esteganografia consiste em detectar a existência dessa comunicação.

Opostamente à esteganografia, a aplicação de inserção de informações de direitos autorais demanda o requisito adicional de robustez contra possíveis ataques. Nesse contexto, o conceito de robustez pode variar dependendo da aplicação.

Marcas de direitos autorais nem sempre precisam estar escondidas, já que alguns sistemas utilizam marcas d'água digitais visíveis. Marcas d'água visíveis estão fortemente ligadas às marcas d'água originais criadas no final do século XIII para diferenciar fabricantes de papéis da época. Exemplos de marcas d'água modernas visíveis incluem logotipos de empresas, símbolos de proteção de direitos autorais, ou outros padrões visuais sobre imagens digitais. Todavia, a maior parte da literatura tem focado em marcas d'água digitais invisíveis (transparentes), pois elas possuem mais aplicações práticas.

Na literatura sobre marcação digital, o estego-objeto é frequentemente referenciado como objeto marcado. Marcas d'água frágeis indicam se um dado foi alterado e providenciam informação do lugar da alteração. Tais marcas d'água não são úteis para aplicações de proteção, porém são de grande valia para aplicações como inserção de informações de pacientes em imagens médicas, etc. Essas marcas não precisam ser resistentes a técnicas de processamento ou a ataques intencionais, pois uma falha na leitura ou recuperação da marca d'água indica que o dado não é autêntico [6].

Outras aplicações requerem que a marca d'água esteja sempre presente no objeto marcado, independentemente de os ataques serem intencionais ou não. Tais marcas d'água são chamadas de robustas. O ideal é que as marcas robustas possuam características que tornem impossível a sua remoção de uma imagem sem destruí-la simultaneamente. Uma das formas de se agregar robustez a uma marca d'água é ocultá-la cuidadosamente nas componentes perceptualmente mais significantes do domínio disponível para inserção de informação [7].

Ataques não intencionais incluem aplicações envolvendo armazenamento e transmissão de dados, que frequentemente usam técnicas de compressão. Outras técnicas de processamento que degradam um sinal incluem filtragem, reamostragem, conversão analógico-digital e vice-versa, etc. Ataques intencionais são aqueles que visam remover a marca d'água do objeto (ou da imagem).

Alguns autores costumam fazer distinção entre os diversos tipos de marcas robustas. Impressões digitais (*fingerprints*) são frequentemente referenciadas como

números de série, que permitem que o dono da propriedade intelectual identifique qual consumidor violou o acordo de licença por repassar a propriedade intelectual para terceiros. Marcas d'água dizem respeito a poucos bits de informação, que indicam o proprietário do objeto.

Neste trabalho, a inserção de qualquer quantidade de informação em objetos, que no caso serão imagens digitais, será referenciada como inserção ou escrita de marca d'água. Isto será feito para simplificar a terminologia.

A Figura 2.1 ilustra o processo genérico de escrita. Dada uma imagem I, uma marca M e uma chave K, que normalmente é a semente de um gerador de números aleatórios, o processo de escrita ou ocultação pode ser definido como o mapeamento $I \times K \times M \rightarrow \widetilde{I}$, que é comum para todos os métodos de marcação.

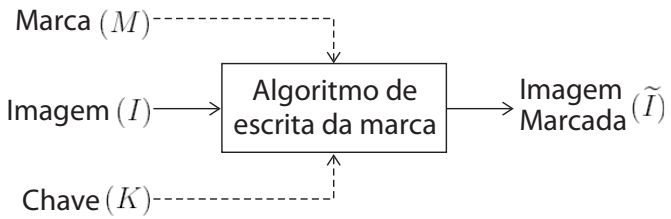


Figura 2.1: Processo genérico de escrita da marca d'água

O processo genérico de leitura ou detecção da marca (esteganálise) está ilustrado na Figura 2.2. A saída desse processo pode ser a marca recuperada M, ou algum tipo de medida, indicando a probabilidade de uma dada marca estar presente na imagem em análise $\widetilde{I'}$.

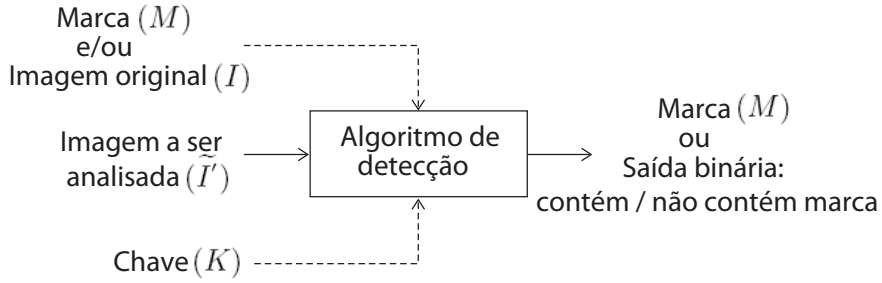


Figura 2.2: Processo genérico de leitura ou detecção da marca d'água

Há vários tipos de sistemas de marcação robusta. Eles são definidos através de suas entradas e saídas. Segue uma descrição do sistema de marcação privada, sis-

tema de marcação semi-privada, sistema de marcação pública e sistema de marcação assimétrica.

Sistemas de marcação privada requerem pelo menos a imagem original. Tais sistemas usam a imagem original para encontrar a localização da marca d'água em $\widetilde{I}'$. A saída de um sistema deste tipo pode ser a própria marca d'água M ou uma resposta binária (do tipo sim ou não) para a pergunta: $\widetilde{I}'$ contém a marca M? Esses sistemas podem ser definidos como $\widetilde{I}' \times I \times K \times M \rightarrow 0, 1$. Espera-se que esse tipo de marcação seja mais robusto do que os outros, já que ele precisa de pouca informação, e requer acesso à informação secreta.

A marcação semi-privada não usa a imagem original para a detecção ($\widetilde{I}' \times K \times M \rightarrow 0, 1$), porém responde à mesma pergunta acima. A principal aplicação das marcações privada e semi-privada é a possibilidade de provar propriedade e controle de cópia em mídias, como o DVD, em disputas judiciais.

Sistemas de marcação pública, ou marcação cega, representam o problema mais desafiador, pois eles não requerem nem a imagem original I nem a marca escondida M. De fato, tais sistemas realmente extraem a informação escondida da imagem marcada: $\widetilde{I}' \times K \rightarrow M$. Marcas cegas possuem mais aplicações do que as outras. Seus algoritmos de escrita podem ser usados comumente nas marcas privadas, aumentando-se a robustez ao mesmo tempo.

Por último, a marcação assimétrica, ou marcação de chave pública, deve ter a propriedade de qualquer usuário poder ler a marca d'água, sem conseguir removê-la.

2.3 Histórico

Esta seção apresentará técnicas utilizadas para inserção (ocultação) de informação. Muitas delas surgiram na antiguidade.

2.3.1 Segurança através de Obscuridade

A literatura de esteganografia começou a crescer, consideravelmente, durante os séculos XVI e XVII. Nessa época, surgiram novos métodos de ocultar informação. Gaspar Schott (1608 - 1666) explica, em seu livro de quatrocentas páginas *Schola Steganographica* [3], [8], como esconder mensagens em figuras musicais, fazendo com

que cada nota musical corresponda a uma letra. [9] menciona outro método, utilizado por Johann Sebastian Bach, que se baseia no número de ocorrências de notas musicais. Schott também explora o código "Ave Maria", proposto por Johannes Trithemius (1462-1516) em *Steganographiæ*, um dos primeiros livros dedicados à área. A expansão desse código usa quarenta tabelas. Cada uma delas possui vinte e quatro entradas (uma para cada letra do alfabeto naquela época) em quatro idiomas: Latim, Alemão, Italiano e Francês. Cada letra do texto original é substituída pela palavra ou frase que aparece na entrada correspondente da tabela e, finalmente, o estego-texto fica parecendo uma oração. Reeds [10] mostrou que essas tabelas podem ser decifradas através da redução a módulo vinte e cinco e, em seguida, usando-se um alfabeto reverso. Além disso, John Wilkins mostrou que dois músicos podem se comunicar tocando os seus instrumentos musicais. Ele também explica que é possível esconder uma mensagem em desenhos usando pontos, linhas ou triângulos de modo que suas diferentes combinações representem diferentes letras.

Outro exemplo interessante de ocultação de informação é o acróstico, que são formas textuais em que o conjunto formado pelas primeiras letras de cada palavra, verso ou outra sequência de palavras forma uma mensagem.

Prisioneiros de guerra frequentemente escondiam mensagens em cartas usando os pontos e traços das letras i, j, t e f para codificar um texto em código Morse. Esse tipo de semagrama[3] é eficiente para ocultar mensagens, porém pode se tornar problemático, já que é difícil criar um texto-camuflagem insuspeito que carregue a mensagem escondida. Censores interceptaram diversas tentativas de comunicação desse tipo durante as duas Grandes Guerras. Uma frase famosa interceptada na Primeira Guerra Mundial dizia "Father is dead". Suspeitando, o censor a modificou para "Father is deceased" e a enviou. A resposta obtida pelo censor foi "Is Father dead or deceased?", o que revelou claramente a tentativa de se comunicar uma mensagem oculta entre as duas partes. Durante a Segunda Guerra Mundial, censores norte-americanos impediram que um lote de relógios importados entrasse no país, pois se desconfiava que as posições dos ponteiros dos relógios pudessem conter algum tipo de informação.

[3] A etimologia da palavra *semagrama* vem do Grego, onde *sema* significa sinal e *grama* significa escrito ou desenhado.

Apesar de esteganografia ser diferente de criptografia [11], podem-se tomar emprestados algumas técnicas e conhecimentos de criptografia, pois há mais literatura sobre o assunto. Auguste Kerckhoffs, em 1883, enunciou o primeiro princípio de engenharia criptográfica, advertindo que se deve assumir que o método usado para encriptar dados é conhecido pelo oponente, de forma que a segurança fique baseada somente na escolha da chave [12]. Ou seja, não se deve assumir que o oponente permanecerá ignorante em relação ao sistema em uso. Desde então, a história da criptografia tem mostrado a insensatez da "Segurança através de Obscuridade".

Utilizando-se o princípio de Kerckhoffs como inspiração, obtém-se uma definição de segurança para sistemas de esteganografia. Diz-se que um sistema de esteganografia é seguro quando um oponente que entenda do sistema, mas que não conheça a estego-chave, não consiga obter evidências ou graus de desconfiança que uma comunicação esteja acontecendo. Ou seja, não há como se obter informação sobre a mensagem escondida a partir do objeto marcado (estego-imagem, por exemplo). Deve-se enfatizar que sistemas de esteganografia que objetivem ser amplamente utilizados devem ser publicados, da mesma forma que algoritmos e protocolos de criptografia.

2.3.2 Camuflagem

O sucesso do funcionamento de sistemas de segurança por obscuridade depende, basicamente, da sorte. A segurança em sistemas de esteganografia pode ser incrementada com o uso inteligente da camuflagem. Mesmo que um método seja conhecido a princípio, tornar a busca pela informação escondida muito custosa pode ser proveitoso. Desta forma, um oponente não conseguiria extrair a informação escondida, por falta de recursos (especialmente quando há um grande tráfego de objeto-camuflagem).

Desde os primórdios da Arquitetura, artistas têm notado que trabalhos como esculturas ou pinturas parecem diferentes quando observados a partir de diferentes ângulos. Isso os fez estabelecer regras de perspectiva e de anamorfose[4]. Durante os séculos XVI e XVII imagens anamórficas eram sinônimos de meios ideais para

[4]Anamorfose é uma projeção distorcida, requerendo que o observador use um dispositivo específico ou ocupe uma certa posição para reconstituir a imagem.

camuflar críticas políticas perigosas e idéias hereges. Uma obra-prima de anamorfose é o *Verxierbild*, criado em aproximadamente 1530 em Nürnberg por Shö, que foi um gravador aluno de Albrecht Dürer. Quando se olha a gravura normalmente, vê-se uma estranha paisagem, porém quando se olha a gravura pelo lado, retratos de reis famosos são revelados [3].

Um protocolo de segurança desenvolvido há séculos na China baseia-se no fato do transmissor e do receptor possuírem máscaras de papel iguais com buracos em posições específicas. O transmissor colocava a máscara sobre um papel em branco, escrevia a mensagem secreta nos buracos e, depois, terminava de preencher o papel com outros ideogramas. O receptor colocava uma versão igual da máscara sobre a carta recebida e extraia a informação escondida diretamente. Esse código foi reinventado no início do século XVI por Girolamo Cardan (1501 - 1576), um matemático italiano, que batizou o método de grelha de Cardan [3]. Em 1992 o método foi colocado em prática novamente por um banco inglês, que recomendou que seus clientes escondessem seus dados e senhas bancárias com um sistema análogo. Contudo, nesse caso uma implementação simples do método pode se tornar muito insegura [3]. [13] alerta que a maioria das fraudes em sistemas de segurança ocorrem devido a falhas ou erros nas implementações e não por ataques técnicos.

Uma variante dessa abordagem é marcar determidadas partes de um objeto através de erros ou de características de estilo. Um exemplo clássico foi a técnica utilizada por Francis Bacon (1561 - 1626) no seu alfabeto *biliteraire* (1587) [3], [14], que tem ligações com as controvérsias se ele teria sido o autor dos trabalhos de Shakespeare. Bacon codificava cada letra do alfabeto com grupos de cinco letras compostas por combinações das letras “A” e “B”. A letra “A” era codificada como “AAAAA”, a letra “B” era codificada como “AAAAB”, e assim sucessivamente. O interessante é que tal codificação casa precisamente com os primeiros vinte e quatro termos da escala binária, cuja invenção é atribuída a Gottfried Leibniz em 1671. Diversos exemplos têm origem em tabelas matemáticas. Publicadores de tabelas logarítmicas e de efemeridades astronômicas nos séculos XVII e XVIII costumavam abusar, deliberadamente, dos erros nos dígitos menos significativos [15].

Herodotus (486 - 425 A.C.) conta em seu livro *Histories* que, por volta de 440 A.C., Histiæus raspou a cabeça de seu escravo mais fiel e a tatuou com uma men-

sagem, que desapareceu depois do cabelo ter crescido novamente [16]. A mensagem visava instigar a revolta contra os Persas. Esse antigo método de ocultação de informação ainda foi usado por alguns espiões alemães no início do século XX.

Herodotus também relata como o rei Demaratus, um Grego na corte Persa, alertou Esparta sobre uma iminente invasão por Xerxes. Demaratus retirou a cera de uma tábua de escrita, escreveu a sua mensagem na madeira que estava por debaixo, e depois a cobriu novamente com cera. Assim a tábua não parecia conter mensagem. Até mesmo os destinatários em Esparta quase não perceberam que a tábua continha uma mensagem escondida.

Æneas, o estrategista, inventou diversas técnicas relacionadas à ocultação de informação. Exemplos delas incluem mensagens escritas nas solas dos pés de mensageiros, em brincos femininos e em notas transportadas por pombos. Sua técnica mais relevante constava de um esquema de passar linhas de costura através de cada um dos vinte e quatro buracos feitos em contas. Æneas também propôs esconder mensagens aumentando a altura de algumas letras num texto insuspeito. Outra proposta foi fazer furos muito pequenos acima ou abaixo de letras, de forma tal que os furos fossem disfarçados pelo contraste de cores entre as letras pretas e o papel branco. Essa técnica foi melhorada por John Wilkins, que utilizou tinta invisível para fazer pontos em vez de fazer furos, e foi usada durante o século XVII e nas duas Grandes Guerras por espiões alemães.

Originalmente as tintas invisíveis eram feitas de substâncias orgânicas, como leite ou urina, e eram reveladas através de calor. Apesar dos avanços na área de Química durante a Primeira Guerra Mundial terem incentivado a criação de formulações mais sofisticadas de tintas invisíveis e de reveladores, a invenção de reveladores universais fez a técnica cair em desuso. Os reveladores universais baseiam-se nos efeitos causados sobre as fibras do papel, para determinar em quais regiões houve adulterações com alguma substância. Atualmente, tintas especiais são utilizadas em notas de dinheiro, em cheques bancários e em diversos documentos para esconder informações que permitam comprovar originalidade. Um exemplo seria a tinta fluorescente, que, perante a exposição a um raio laser numa determinada frequência, revela uma informação única.

Durante a Guerra Franco-Prussiana (1870 - 1871) mensagens em microfilme

eram enviadas por pombos [17]. Na Guerra Russo-Japonesa de 1905, imagens microscópicas eram escondidas nas orelhas, narinas e sob as unhas dos militares e espiões. Na Primeira Guerra Mundial as mensagens trocadas por espiões foram reduzidas a micro-pontos, de forma que podiam ser inseridas em vírgulas de textos de uma revista, por exemplo.

No universo digital, muitas dessas técnicas de camuflagem são implementadas através de algoritmos de mascaramento. Assim como a maioria das técnicas de codificação, tais algoritmos levam em conta as propriedades do sistema perceptual humano.

Mascaramento de áudio, por exemplo, é um fenômeno no qual um som interfere na percepção de outro som. Esse mascaramento pode ocorrer na frequência ou no tempo. O mascaramento da frequência ocorre quando dois sons de frequências parecidas são tocados ao mesmo tempo com intensidades diferentes. O mascaramento no tempo ocorre quando um som com intensidade baixa é tocado imediatamente antes ou depois de um som com intensidade mais alta. É interessante citar que, depois de um som de intensidade alta ser tocado, o ouvido fica insensível a um som de intensidade mais baixa numa frequência parecida.

Algoritmos de mascaramento podem ser utilizados em sinais de vídeo que serão posteriormente processados por algoritmos de compressão de vídeo, tais como o MPEG [18]. Assim, estes sistemas inserem a informação escondida nas componentes mais perceptíveis do sinal hospedeiro, de forma que a informação escondida não seja destruída pela compressão.

2.4 Um Modelo para Esteganografia

Em 1983, G. S. Simmons [19] apresentou a versão moderna do problema de esteganografia - o problema do prisioneiro. Alice e Bob estão presos e querem combinar um plano de fuga. Entretanto, todas as comunicações passam pela carcereira Willie. Assim, a comunicação sobre o plano deve ser escondida de forma que Willie não desconfie. Logo, o desafio num sistema de esteganografia consiste em estabelecer uma comunicação com altas taxas sem que ela seja detectada via análise estatística ou perceptual. Willie faz o papel de estaganalista (*steganalyst*), sendo responsável

por determinar se um sinal é *cover* ou *stego*.

A esteganografia moderna tem se tornado um jogo cada vez mais sofisticado entre quem esconde a mensagem e o esteganalista, que tenta identificá-la. Muitas vezes, um sistema de esteganografia é proposto para enganar alguma técnica de esteganálise. Isso faz com que os métodos de estaganálise sejam melhorados. Percebe-se que é difícil definir quem é o vencedor desse jogo. Essas iterações entre quem esconde a mensagem e o estaganalista são benéficas para o desenvolvimento da área de esteganografia.

Shannon [20] propôs o conceito *One-Time Pad* (OTP) de comunicação perfeitamente segura[5] entre Alice e Bob na área de criptografia. O OTP não pode ser quebrado por um criptanalista mesmo que ele possua recursos computacionais ilimitados. O sistema de segurança é baseado no sigilo e na aleatoriedade da chave.

Um sistema equivalente ao OTP para esteganografia é aquele que atende a duas premissas. A primeira delas é que a carcereira Willie tenha conhecimento determinístico perfeito de todos os possíveis sinais hospedeiros insuspeitos (*covers*). Desta forma, considerando-se o caso de ocultação de informação sobre imagens, mesmo uma pequena mudança em apenas um pixel da imagem seria detectável. A segunda premissa é que Willie tenha recursos computacionais ilimitados, o que a permite realizar testes com todas as chaves secretas possíveis. Sob essas premissas, qualquer modificação feita por Alice no sinal hospedeiro (objeto-camuflagem) é suspeita e pode ser potencialmente identificada por Willie.

Há uma forma de Alice e Bob se comunicarem secretamente sob tais premissas idealizadas. No equivalente ao OTP em esteganografia se assume que Alice e Bob possuem a mesma chave secreta e o mesmo conjunto de imagens. O conjunto de imagens também é conhecido por Willie. Para comunicar uma mensagem a Bob, Alice envia uma imagem do conjunto disponível indexada pela mensagem, que por sua vez é embaralhada pela chave secreta. Esta idéia é equivalente à de Shannon exceto que, em vez de se enviar uma mensagem encriptada, Alice envia uma imagem, que é indexada pela mensagem encriptada.

Para se comunicar uma mensagem de n bits, Alice e Bob devem compartilhar um

[5]Refere-se aqui à segurança criptográfica, em que o significado da mensagem não é revelado. A segurança esteganográfica requer que a presença da comunicação não seja revelada.

conjunto de 2^n imagens. Como Willie também possui o mesmo conjunto de imagens, uma chave secreta de pelo menos n bits deve ser compartilhada entre o codificador (Alice) e o decodificador (Bob) para indexar as imagens. É interessante notar que um sistema perfeitamente seguro de esteganografia pode ser construído sem ocultar a informação.

Entretanto, deve-se observar que o sistema descrito possui limitações práticas. Analogamente ao OTP de Shannon, a chave secreta, que possui a mesma complexidade em número de bits que a mensagem, deve ser comunicada via um canal seguro alternativo. Além disso, uma chave pode ser usada somente uma vez. Um outro problema é que pode não ser viável o compartilhamento de um conjunto grande de imagens hospedeiras entre o codificador e o decodificador. Restringindo-se à situação na qual Alice e Bob não podem compartilhar o mesmo conjunto de imagens e na qual a comunicação deva ser feita apenas através de uma imagem, então a capacidade do sistema passa a ser $\log_2(1) = 0$. Em outras palavras, um sistema perfeitamente seguro de comunicação não seria praticável.

Num sistema prático, o estaganalista não possui conhecimento perfeito sobre as imagens hospedeiras (*cover-images*). Num sistema real, Willie tem, na melhor das hipóteses, somente um modelo estocástico em vez de determinístico para as imagens hospedeiras. Nesse caso, espera-se ter uma capacidade finita diferente de zero para se alcançar uma comunicação perfeitamente segura com uma ou poucas imagens. Sabendo que Willie não tem conhecimento perfeito de uma dada imagem, a Alice agora pode modificá-la de forma a esconder uma mensagem.

A capacidade do sistema, ou o número de bits que podem ser inseridos sem incitar a desconfiança de Willie, depende, estritamente, da acurácia do modelo estocástico dela. O entendimento de Willie sobre o que é uma imagem natural pode consistir de aspectos visuais e de observações estatísticas. Naturalmente, há a necessidade de que o processo de escrita da mensagem não cause distorção perceptível na imagem hospedeira. Diversas técnicas de esteganálise também empregam análises estatísticas para detectar a presença de mensagem inserida. Para que a comunicação não seja detectada, o processo de escrita deve satisfazer às duas condições a seguir:

- **Restrição perceptual:** A distorção perceptual entre a imagem original e a marcada não deve ser maior do que um certo valor D_1, para um dada medida

de distância de percepção.

- **Restrição estatística:** O processo de escrita não deve modificar as estatísticas da imagem hospedeira em mais do que um valor pequeno ϵ, para uma dada medida de distância estatística.

Capítulo 3

Esteganografia via Restauração Estatística

Como visto no Capítulo 2, a esteganografia é a arte e a ciência da comunicação de forma imperceptível a terceiros. Viu-se, também, que ela possui um histórico bastante rico.

O objetivo desse trabalho é inserir o máximo de informação possível em imagens, de forma que um esteganalista (Willie, conforme a Seção 2.4) não desconfie da existência dessa mensagem. A principal motivação é a abordagem feita por Sarkar e Manjunath [2] sobre capacidade esteganográfica usando o conceito de $\epsilon-$segurança em esteganografia proposto por Cachin [21]. Esse conceito é explicado na Seção 3.1. Fridrich et al. [22] definiram capacidade esteganográfica como o maior tamanho da mensagem que pode ser escrita num sinal hospedeiro, sem produzir distorções estatistica ou perceptualmente detectáveis.

A idéia para se encontrar um erro ϵ pequeno entre as distribuições do objeto *cover* e do objeto *stego* é empregar uma restauração estatística, gastando-se recursos para reparar o estrago feito na estatística da imagem original pelo processo de escrita da informação. Para assegurar que a restauração não interfira no processo de decodificação (ou leitura) da informação escondida, uma porcentagem fixa do sinal hospedeiro é separada para escrita, enquanto que o restante é usado para a restauração (ou compensação) da estatística. Assume-se que tanto o codificador quanto o decodificador possuam uma chave secreta que determina aonde a mensagem está escrita.

A Seção 3.3 apresenta a abordagem feita por Sarkar e Manjunath. Em [2], os autores buscam escrever o máximo de informação usando um método de escrita par/ímpar no domínio da transformada discreta do cosseno (DCT) quantizada. Após o processo de escrita, aplica-se o processo de compensação. A compensação objetiva igualar as funções densidade de probabilidade (PDFs) dos coeficientes da DCT da imagem original e da imagem marcada. Dessa forma, um sistema de esteganálise não consegue identificar que uma imagem possui mensagem escondida através da análise da PDF dos coeficientes da DCT da imagem. Sarkar e Manjunath utilizaram o método de compensação de histogramas proposto por Tzschoppe et al. [1].

A DCT é uma ferramenta amplamente utilizada em processamento de imagens, pelo fato de possuir diversas propriedades úteis para a representação ou análise de imagens naturais. A Seção 3.2 apresenta uma breve descrição dessa transformada, de suas propriedades e de suas aplicações na área de processamento de imagens.

As Seções 3.4, 3.5 e 3.6 apresentam três novas abordagens, que objetivam escrever, de forma indetectável, mais informação nas imagens do que o método utilizado por Sarkar e Manjunath. Essas abordagens são chamadas de Proposta A, Proposta B e Proposta C, respectivamente.

As análises práticas e os resultados obtidos por cada abordagem descrita neste trabalho serão apresentadas no Capítulo 4.

3.1 Conceito de ϵ-segurança de Cachin

O conceito de ϵ-segurança proposto por Cachin em [21] diz que a entropia relativa (divergência de Kullback-Leibler) entre as distribuições do *cover* e do *stego* deve ser menor ou igual a ϵ para que a comunicação seja segura.

A definição de segurança de Cachin assume que Willie permita que Alice envie qualquer imagem *cover* $\mathbf{c}$ para Bob (desde que o conjunto dos coeficientes das DCTs dos blocos desta imagem apresente uma distribuição de probabilidades P_{cover}). Esta distribuição de probablidades representa o conhecimento que Willie tem sobre quais tipos de transmissão entre Alice e Bob são legítimas. A probabilidade de se tomar uma imagem *cover* $\mathbf{c}$ desta distribuição é $P_{cover}(\mathbf{c})$.

Por exemplo, pode acontecer que a esteganalista Willie espere que Alice e Bob

compartilhem apenas fotos de paisagens. Assim, dado que uma foto $\mathbf{c}$ contém a imagem de uma paisagem, então há uma probablidade finita $P_{cover}(\mathbf{c})$ da imagem ter sido gerada a partir dessa distribuição. Caso a fotografia contenha uma imagem de caixotes num chão de fábrica, por exemplo, a probabilidade desta imagem ter sido gerada a partir da distribuição de probilidades P_{cover} é muito baixa ou zero. Consequentemente, há uma grande chance de Willie suspeitar desta imagem. A distribuição análoga à P_{cover} para imagens *stego* é a distribuição P_{stego}.

Cachin analisa a performance da esteganalista usando a teoria de teste de hipóteses. Dada uma imagem $\mathbf{c}$, Willie deve decidir entre duas hipóteses:

- H_0: hipótese de que a imagem $\mathbf{c}$ não contém informação escondida;
- H_1: hipótese de que a imagem $\mathbf{c}$ contém informação escondida.

Se a hipótese H_0 for verdadeira, então a imagem $\mathbf{c}$ foi gerada a partir da distribuição P_{cover}. Se a hipótese H_1 for verdadeira, então a imagem $\mathbf{c}$ foi gerada a partir da distribuição P_{stego}.

A comparação entre as distribuições P_{cover} e P_{stego} pode ser feita baseada na entropia relativa, também conhecida como divergência de Kullback-Leibler, definida por

$$D(P_{cover}||P_{stego}) = \sum_{\mathbf{c}\in cover} P_{cover}(\mathbf{c}) \log_2 \frac{P_{cover}(\mathbf{c})}{P_{stego}(\mathbf{c})}. \tag{3.1}$$

A entropia relativa é sempre não-negativa e vale zero se e somente se $P_{cover} = P_{stego}$. Quando isto ocorre, diz-se que o sistema de esteganografia de Alice é perfeitamente seguro, como definido por Cachin. É impossível para a esteganalista Willie distinguir entre imagens *cover* e *stego* quando $D(P_{cover}||P_{stego}) = 0$. Se $D(P_{cover}||P_{stego}) \leq \epsilon$, o sistema de esteganografia é definido como ϵ-seguro.

A resposta do detector de Willie é binária: $\mathbf{c}$ contém ou não contém mensagem escondida. Willie pode cometer dois tipos de erro [23]. O erro do Tipo I é um falso positivo, que ocorre quando Willie decide que uma mensagem escondida está presente, quando, de fato, ela não está. A probabilidade de se ocorrer um erro do Tipo I é α. O erro do Tipo II é um falso negativo, que ocorre quando Willie decide que não há mensagem escondida, quando, de fato, ela está presente. A probabilidade de se ocorrer um erro do Tipo II é β.

Assumindo que Willie receba apenas imagens tomadas da distribuição P_{cover}, a probabilidade de Willie apontar uma imagem como *cover* é $p_1 = 1 - \alpha$. A probabilidade de Willie apontar uma imagem como *stego* é $p_2 = \alpha$.

Assumindo que Willie receba apenas imagens tomadas da distribuição P_{stego}, a probabilidade de Willie apontar uma imagem como *cover* é $q_1 = \beta$. A probabilidade de Willie apontar uma imagem como *stego* é $q_2 = 1 - \beta$.

Assim, pode-se escrever

$$\begin{aligned} d(\alpha, \beta) &= D(P(\text{erro}|cover)||P(\text{erro}|stego)) \\ d(\alpha, \beta) &= (1-\alpha)\log_2 \tfrac{1-\alpha}{\beta} + \alpha \log_2 \tfrac{\alpha}{1-\beta}. \end{aligned} \tag{3.2}$$

Um resultado de teoria da informação declara que a entropia relativa de sinais processados nunca pode aumentar [24]. Sabe-se que a esteganalista Willie realiza um tipo de processamento (Figura 3.1). Assim, tem-se a Equação (3.3), que pode ser usada para se determinar um limite inferior da probabilidade de ocorrer um erro do Tipo II, β, dado um limite superior de probabilidade de ocorrer um erro do Tipo I, α.

Figura 3.1: Processamento

$$d(\alpha, \beta) \leq D(P_{cover}||P_{stego}) \tag{3.3}$$

Por exemplo, caso não se permita que Willie acuse Alice pela transmissão de uma mensagem escondida, quando, de fato, ela não existe ($\alpha = 0$), então a Equação (3.3) implica em

$$\log_2 \frac{1}{\beta} \leq D(P_{cover}||P_{stego}) \leq \epsilon. \tag{3.4}$$

Da Equação (3.4), tem-se que a probablidade β de um erro do Tipo II é dada pela Equação (3.5).

$$\beta \geq 2^{-\epsilon} \tag{3.5}$$

Quanto menor for o valor de ϵ, maior será a probabilidade de uma comunicação escondida não ser detectada.

3.2 Transformada Discreta do Cosseno (DCT)

A transformada discreta do cosseno (DCT, *Discrete Cosine Transform*) é uma ferramenta muito utilizada na área de Processamento de Imagens. Ela é especialmente útil para aplicações de compressão com perdas, pois esta transformada possui uma propriedade de forte compactação de energia. Assim, a maioria da informação do sinal tende a se concentrar em poucos componentes de baixas frequências da DCT.

Esta propriedade faz a DCT se aproximar da transformada Karhunen-Loève (KLT, *Karhunen-Loève Transform*) em alguns casos. A KLT é a transformada ótima no sentido de descorrelacionar os dados, maximizando a energia concentrada em um dado número de componentes da transformada.

A matriz $\mathbf{C} = \{c(k,n)\}$ da DCT $N \times N$ é definida pela Equação (3.6). As imagens base da DCT 8×8 são mostradas na Figura 3.2.

$$c(k,n) = \begin{cases} \frac{1}{\sqrt{N}}, \, k = 0, 0 \leq n \leq N-1 \\ \sqrt{\frac{2}{N}} \cos \frac{\pi(2n+1)k}{2N}, \, 1 \leq k \leq N-1, 0 \leq n \leq N-1 \end{cases} \tag{3.6}$$

A DCT apresenta algumas propriedades interessantes, como mostradas em [25]. Essas propriedades são enunciadas a seguir:

1) A DCT é real e ortogonal, ou seja,

$$\mathbf{C} = \mathbf{C}^* \Rightarrow \mathbf{C}^{-1} = \mathbf{C}^T; \tag{3.7}$$

2) A DCT não é a parte real da transformada de Fourier discreta unitária[1] (DFT);

3) A DCT é uma transformada rápida. A DCT de um vetor de N elementos pode ser calculada com complexidade $O(N \log_2 N)$ através de uma DFT de N pontos;

4) A DCT apresenta uma excelente compressão de energia para dados altamente correlatados;

[1]DFT unitária de uma imagem $\{u(m,n)\}$ $N \times N$: $v(k,l) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} u(m,n) W_N^{km} W_N^{ln}$, $0 \leq k,l \leq N-1$, onde $W_N = \exp^{\frac{-j2\pi}{N}}$.

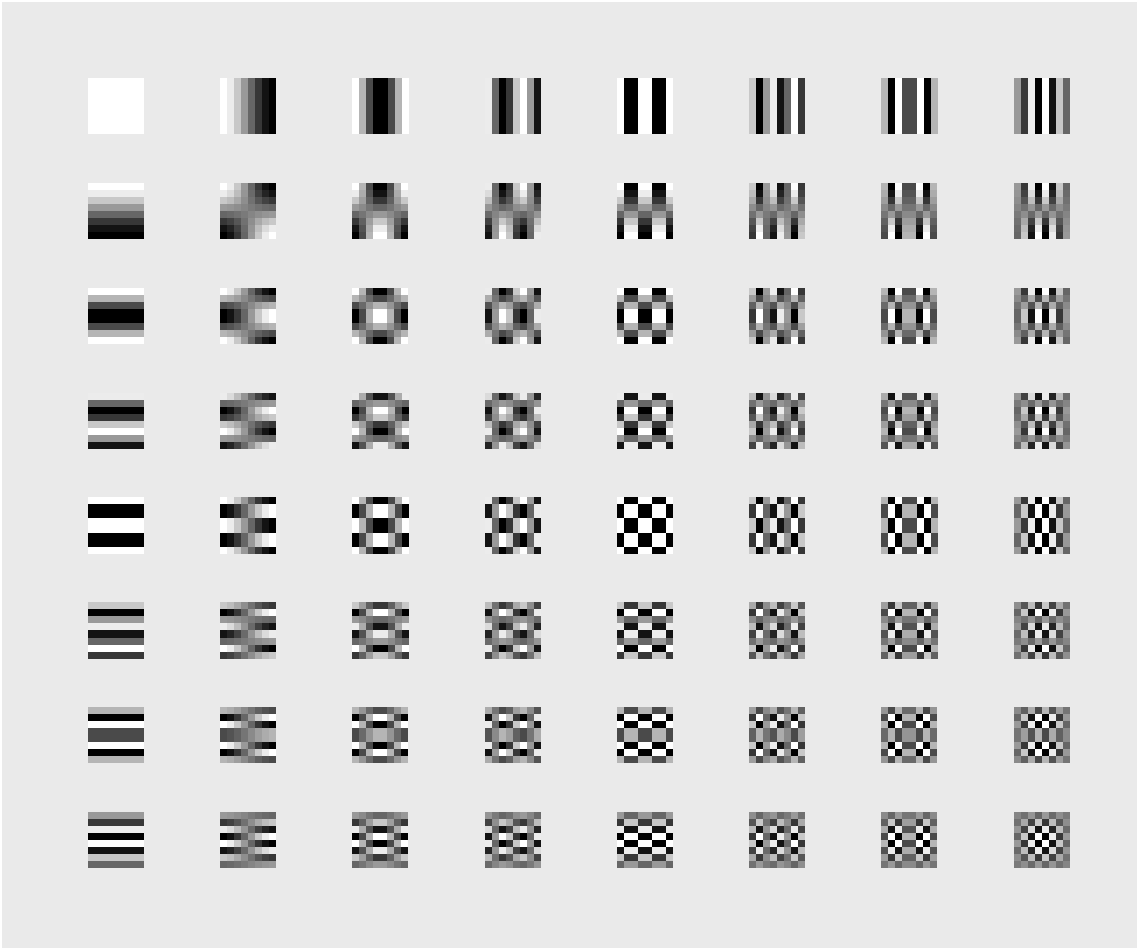


Figura 3.2: Imagens base da DCT 8×8

5) As linhas da matriz $\mathbf{C}$, que são os vetores base da DCT, são os autovetores da matriz simétrica $\mathbf{Q}_C$, definida conforme a Equação (3.8).

$$\mathbf{Q}_C = \begin{pmatrix} 1-\alpha & -\alpha & \dots & 0 & 0 \\ -\alpha & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -\alpha & 1-\alpha \end{pmatrix} \tag{3.8}$$

6) A DCT $N \times N$ é bastante parecida com a KLT de um processo de Markov estacionário de primeira ordem de tamanho N cuja matriz de covariância $\mathbf{R}$ é dada por

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & \dots & \dots & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{N-1} & \dots & \dots & \rho & 1 \end{pmatrix} \tag{3.9}$$

quando $\rho \to 1$.

Nota-se que a matriz $\mathbf{R}$ é do tipo Toeplitz, pois considera-se o processo estacionário. O motivo pelo qual, neste caso, a DCT se aproxima da KLT é que $\mathbf{R}^{-1}$ é uma matriz simétrica, que satisfaz a relação $\frac{(1-\rho^2)}{(1+\rho^2)}\mathbf{R}^{-1} \approx \mathbf{Q}_C$ para $\rho \to 1$. Assim, os autovetores de $\mathbf{R}$ e de $\mathbf{Q}_C$, que compõem a DCT, serão bastante parecidos.

Frequentemente, verifica-se que funções de covariância de *pixels* próximos em imagens naturais podem ser modeladas como separáveis e estacionárias, de acordo com a Equação (3.10). σ^2 representa a variância do processo aleatório bidimensional e $\rho_1 = \frac{r(1,0)}{\sigma^2}$ e $\rho_2 = \frac{r(0,1)}{\sigma^2}$ representam as autocovariâncias nas direções m e n, respectivamente. Logo, as DCTs de blocos pequenos, por exemplo 8×8, de imagens naturais aproximam muito bem a KLT [26].

$$r(m,n) = \sigma^2 \rho_1^{|m|} \rho_2^{|n|},\ |\rho_1| < 1,\ |\rho_2| < 1 \tag{3.10}$$

3.3 Abordagem de Sarkar e Manjunath

Esta seção apresenta a abordagem do sistema de esteganografia utilizado em [2].

A Seção 3.3.1 apresenta a formulação do problema e explica como o conjunto disponível para escrita e compensação deve ser dividido, de forma que a compensação perfeita da PDF desse conjunto seja realizável. Esta formulação também será válida para as Propostas A, B e C.

A Seção 3.3.2 apresenta o método de escrita par/ímpar, que foi utilizado por Sarkar e Manjunath.

A Seção 3.3.3, em conjunto com as noções apresentadas nas Seções 3.3.1 e 3.3.2, mostra como maximizar o tamanho da mensagem que pode ser inserida na imagem. Simultaneamente, garante-se o processo de compensação gastando-se menos recursos.

Por fim, a Seção 3.3.4 apresenta o método de compensação de histogramas de Tzschoppe [1], que foi utilizado por Sarkar e Manjunath.

3.3.1 Formulação do Problema

Seja um conjunto X de coeficientes da DCT disponíveis para escrita e compensação. Divide-se X em dois conjuntos disjuntos de acordo com a Equação (3.11), sendo o conjunto H para escrita e o conjunto C para compensação. O fator de escrita λ ($\lambda \in [0, 1]$) é dado pela relação entre as cardinalidades dos conjuntos H e X, de acordo com a Equação (3.12).

$$X = H \cup C, H \cap C = \emptyset \tag{3.11}$$

$$\lambda = \frac{|H|}{|X|} \tag{3.12}$$

Seja Y o conjunto obtido após os procedimentos de escrita e de compensação, como apresentado na Equação (3.13). Deve-se observar que o conjunto H, disponível para escrita, foi alterado para $\widehat{H}$, e que o conjunto C, disponível para compensação, foi alterado para $\widehat{C}$.

$$Y = \widehat{H} \cup \widehat{C}, \widehat{H} \cap \widehat{C} = \emptyset \tag{3.13}$$

Sabe-se que a função massa de probabilidade de uma distribuição é obtida através da normalização do histograma dessa distribuição pela sua quantidade de elementos. Sejam P_X e P_Y as funções massa de probablidade dos conjuntos X e Y, respectivamente.

O principal objetivo é encontrar o fator máximo de escrita λ_{opt}, que maximize $|H|$, com a restrição de que elementos suficientes sejam fornecidos para o processo de compensação tornar P_Y igual a P_X. Assim, o esteganalista não consegue identificar, estatisticamente, a presença da uma mensagem oculta na imagem.

Seja $B_K(i)$ a quantidade de termos iguais a i num conjunto K. Ou seja, $B_K(i)$ é a altura da torre do histograma de K centrada em i. Assim, podem-se escrever as Equações (3.14) e (3.15), pois tanto H e C quanto $\widehat{H}$ e $\widehat{C}$ são disjuntos.

$$B_X(i) = B_H(i) + B_C(i), \forall i \tag{3.14}$$

$$B_Y(i) = B_{\widehat{H}}(i) + B_{\widehat{C}}(i), \forall i \tag{3.15}$$

Para que $P_Y = P_X$, necessita-se que $B_Y(i) = B_X(i), \forall i$. Para um dado conjunto X, $B_X(i)$ é conhecido. Além disso, depois da escrita, o conjunto H é mudado para $\widehat{H}$, o que permite encontrar $B_{\widehat{H}}(i)$. Substituindo-se $B_Y(i)$ da Equação (3.15) por $B_X(i)$, obtém-se a Equação (3.16). Esta equação fornece a distribuição que os elementos formadores do conjunto de compensação devem apresentar após o processo de escrita, para que os histogramas de X e de Y fiquem idênticos.

$$B_{\widehat{C}}(i) = B_X(i) - B_{\widehat{H}}(i) \geq 0, \forall i \tag{3.16}$$

Deve-se notar que a restauração perfeita só é possível se $B_{\widehat{C}}(i)$ for não-negativo $\forall i$, pois histogramas não podem conter quantidades negativas. Em outras palavras, $B_{\widehat{C}}(i) < 0$ significa que nunca será possível compensar perfeitamente a quantidade de elementos i, pois não há recursos suficientes para isso no conjunto disponível para compensação. $B_{\widehat{C}}(i) \geq 0$ significa que, caso haja recursos disponíveis no conjunto de compensação, poder-se-á compensar perfeitamente a quantidade de elementos i.

O cálculo do fator máximo (fator ótimo) de escrita λ_{opt} é dado pela Equação (3.17). À medida que λ aumenta, a distância entre as duas funções massa

de probabilidade associadas a B_H e $B_{\widehat{H}}$ aumenta e restam menos termos disponíveis para a compensação.

$$\lambda_{opt} = \underset{\lambda=\frac{|H|}{|X|}}{\operatorname{argmax}} \{|H| = |\widehat{H}| : B_X(i) - B_{\widehat{H}}(i) \geq 0, \forall i\} \tag{3.17}$$

Vale ressaltar que toda a formulação explicada nessa seção é geral. X pode ser formado por qualquer conjunto de dados, não sendo restrito a coeficientes da DCT de uma imagem.

3.3.2 Método de Escrita Par/Ímpar

De uma forma indireta, a luminância da imagem é usada para ocultar a mensagem. O conjunto X é formado pelo seguinte procedimento: i) calcula-se a DCT de cada bloco 8×8 da imagem e, em seguida, divide-se ponto a ponto cada bloco resultante por uma matriz de quantização sujeita a um fator de qualidade; ii) uma banda de frequência é escolhida (por exemplo, do coeficiente 1 ao 19) e os coeficientes assim selecionados da DCT de cada bloco são arredondados para formar o conjunto X.

O método par/ímpar de escrita converte os termos do conjunto X para o inteiro par ou ímpar mais próximo dependendo do bit a ser escrito. Caso um termo de X seja par e o bit da marca d'água seja 1 (ímpar), então esse termo de X deve ser mapeado para o inteiro ímpar mais próximo, que pode ser obtido através da soma ou subtração da unidade. Caso o termo de X em análise seja par e o bit a ser inserido seja 0, nada precisa ser feito, pois o termo de X já está mapeado para o seu inteiro par mais próximo - ele mesmo.

Resumindo, caso as paridades do coeficiente de X e do bit da marca d'água sejam iguais, nada precisa ser feito. Ou seja, o coeficiente de X já estará carregando a informação referente a esse bit de informação. Caso as paridades do coeficiente de X e do bit da marca d'água sejam diferentes, deve-se somar 1 ou subtrair 1 de tal coeficiente. A probabilidade de se escolher a operação de soma ou de subtração é de 50% para cada uma delas.

O método par/ímpar de escrita é uma versão simplificada do método *Quantization Index Modulation* (QIM) [27].

Sendo c o coeficiente original da DCT quantizada, s o valor do mapeamento de c

em função do bit da marca d'água e Δ um valor sorteado aleatoriamente no intervalo $[-\frac{1}{2}, \frac{1}{2}]$, as Equações (3.18) e (3.19) definem os processos de escrita dos bits 0 e 1 da marca d'água, respectivamente. Se c é um número par (ímpar) e deseja-se inserir um bit 1 (0), c é mapeado para $c+1$ ou $c-1$ dependendo se Δ pertence ao intevalo $(0, \frac{1}{2}]$ ou $[-\frac{1}{2}, 0]$, respectivamente. A função $round(k)$ arredonda k para o seu inteiro mais próximo e a função $mod(m, n)$ retorna o resto da divisão de m por n.

$$s = round(c + 1 - mod(c + 1 - \Delta, 2)) \tag{3.18}$$

$$s = round(c + 1 - mod(c - \Delta, 2)) \tag{3.19}$$

Seja $X(i)$ o subconjunto de X tal que todos os seus elementos pertençam a X e sejam iguais a i. Seja $\widehat{H}(i)$ o subconjunto de $\widehat{H}$ definido analogamente a $X(i)$. Considera-se que λ seja o fator de escrita comum para todos os subconjuntos $X(i)$.

Assume-se que a mensagem a ser inserida possua um número igual de bits 0 e 1 que afetam os elementos em $X(i)$. Tal consideração é válida considerando-se que o tamanho da mensagem é grande o suficiente. Há uma probabilidade de 50% que cada termo de $X(i)$ seja alterado. Caso um termo de $X(i)$ seja alterado, ele pode ser mapeado para o maior ou para o menor inteiro mais próximo com probabilidades iguais. Deve-se notar que apenas uma fração λ ($\lambda \in [0, 1]$) de termos de cada subconjunto $X(i)$ será utilizada para escrita, pois a fração $(1 - \lambda)$ restante dos termos ficará reservada para o processo de compensação.

Assim, $\lambda.\frac{1}{2}$ termos de $X(i)$ permanecerão em $\widehat{H}(i)$. Frações iguais de $\lambda.\frac{1}{2}.\frac{1}{2}$ termos de $X(i)$ serão transferidos para $\widehat{H}(i-1)$ e para $\widehat{H}(i+1)$.

Baseando-se nessa análise, o número de termos em $\widehat{H}(i)$ é dado por

$$B_{\widehat{H}}(i) \approx \frac{\lambda.B_X(i)}{2} + \frac{\lambda.B_X(i-1)}{4} + \frac{\lambda.B_X(i+1)}{4}. \tag{3.20}$$

3.3.3 Fator de Escrita

Como visto na Seção 3.3.1, o fator de escrita λ é definido pela Equação (3.12). Ele representa a porcentagem de elementos do conjunto total de coeficientes da DCT disponíveis para alteração, X, que será usada para a ocultação da mensagem.

Viu-se que elementos de altas energias são pouco frequentes na transformada discreta do cosseno aplicada a blocos 8×8 de uma imagem. A distribuição dos coeficientes da DCT quantizada para imagens naturais possui valores muito elevados na região próxima de zero e decai, rapidamente, quando se afasta deste valor. Nem todos os coeficientes da transformada são disponibilizados para escrita e compensação. Elementos com magnitudes elevadas ocorrem poucas vezes e seriam muito improváveis de serem compensados, o que atrapalharia a segurança do sistema de esteganografia.

Os elementos considerados para computar o histograma dos coeficientes da DCT quantizada são aqueles, dentro da faixa de frequências escolhida previamente, do coeficiente 1 ao 19 por exemplo, cujos módulos sejam menores que um limiar T. Desta forma, garante-se uma maior probabilidade de compensação para todos os coeficientes. Para um dado T, há $2T + 1$ *bins* no intervalo $[-T, T]$. O processo de escrita ocorre em todos os *bins*, com exceção dos *bins* dos extremos ($-T$-ésimo *bin* e T-ésimo *bin*). Esses *bins* podem ser mais difíceis de serem perfeitamente compensados, pois eles só possuem uma vizinhança a partir da qual podem ser compensados.

Da Equação (3.16), deriva-se a Equação (3.21). Aplicando-se a Equação (3.20) na Equação (3.21), obtém-se a Equação (3.22), que fornece um limite superior para λ.

$$B_{\widehat{H}}(i) \leq B_X(i) \tag{3.21}$$

$$\lambda \leq \left\{ \frac{B_X(i)}{\frac{B_X(i-1)}{4} + \frac{B_X(i)}{2} + \frac{B_X(i+1)}{4}} \right\} \tag{3.22}$$

Como λ é o fator de escrita comum para todos os subconjuntos $X(i)$, a Equação (3.22) deve ser satisfeita para todos os coeficientes i do conjunto X. Define-se λ_i na Equação (3.23) de forma a se encontrar o menor deles — o fator de escrita efetivo λ^*. O λ_i pode ser visto como $\frac{B_X}{B_{\widehat{H}}}$, com $B_{\widehat{H}}$ calculado com $\lambda = 1$.

$$\lambda_i = \left\{ \frac{B_X(i)}{\frac{B_X(i-1)}{4} + \frac{B_X(i)}{2} + \frac{B_X(i+1)}{4}} \right\} \tag{3.23}$$

O fator de escrita efetivo $\lambda^*(T)$, para um dado limiar T, é dado pela

Equação (3.24).

$$\lambda^*(T) = \inf\{\lambda_i : \lambda_i > 0, i \in (-T, T)\} \tag{3.24}$$

A condição $\lambda_i > 0$ assegura que o fator de escrita não será reduzido para zero, no caso de haver *bins* sem elementos. Isto pode causar diferença entre as PMFs (*Probability Mass Functions*) dos coeficientes da DCT quantizada antes e após a escrita da marca d'água. Tal descasamento entre as PMFs é improvável de ser significativo estatisticamente, tornando-o desprezível para a detecção da informação oculta.

Uma vez calculadas as DCTs dos blocos da imagem, escolhe-se uma faixa de frequências para escrita e compensação. A fração máxima de termos que realmente pode ser usada para escrever uma mensagem em função de um certo limiar viabilizando a restauração estatística é chamada de taxa $R(T)$. Sendo P_X a PMF de X, a Equação (3.25) apresenta a fração de termos disponíveis para escrita, $G(T)$, sobre um limiar T. A taxa $R(T)$ é dada pela Equação (3.26).

$$G(T) = \sum_{-T<i<T} P_X(i) \tag{3.25}$$

$$R(T) = \lambda^*(T).G(T) \tag{3.26}$$

À medida que T aumenta, $G(T)$ aumenta enquanto que $\lambda^*(T)$ tende a diminuir. Isto acontece pois a probabilidade de se encontrar um λ_i menor num intervalo $[-T, T]$ maior aumenta, conforme a Equação (3.24).

Varia-se o limiar T e seleciona-se o limiar ótimo T_{opt} para o qual a taxa é maximizada, como apresentado na Equação (3.27).

$$T_{opt} = \underset{T}{\operatorname{argmax}}\, R(T) \tag{3.27}$$

A máxima taxa alcançável para o conjunto X, formado pelos coeficientes da DCT quantizada, usando-se o método de escrita par/ímpar e compensação é $R_{opt} = R(T_{opt})$.

A Seção 3.3.4 apresenta o método de compensação de histogramas utilizado em [2].

3.3.4 Método de Compensação

Pode-se considerar uma imagem como uma realização de um processo estocástico. Caso o esteganalista conheça esse processo estocástico, tudo o que ele precisa fazer para decidir se a imagem está inalterada ou não é usar o modelo conhecido. Contudo, na prática é impossível caracterizar tal processo estocástico. Isto leva o esteganalista a considerar estatísticas simplificadas para a sua análise, o que gera brechas para a inserção imperceptível de informação.

O sistema de escrita da mensagem tem a vantagem de conhecer a estatística da imagem original, pois ela está sob sua posse. Assim, o procedimento de escrita pode assegurar uma comunicação segura ao tornar a estatística da imagem marcada semelhante à da original. Uma forma de se alcançar essa segurança é reservar parte dos recursos alocados para inserção de informação para restaurar (compensar) a estatística original. Deve-se notar que uma estatística simplificada é considerada, e não o processo estocástico por completo. Neste trabalho, a estatística considerada é a PMF dos coeficientes de uma certa banda de frequência da DCT quantizada aplicada a cada bloco 8×8 da imagem. Consideram-se apenas os coeficientes cuja magnitude seja menor que um limiar.

Garante-se que os coeficientes alocados para escrita e para compensação sejam diferentes. Isto é feito para que o processo de restauração não interfira na decodificação ou leitura da informação da imagem marcada. O processo de compensação pode ser visto como o custo dispendido para aumentar a segurança, pois ele implica na diminuição do tamanho da mensagem que pode ser inserida na imagem.

Na Seção 3.3.1, o conjunto X foi dividido nos subconjuntos disjuntos H, para escrita, e C, para compensação (Equação (3.11)). Deve-se alterar o conjunto C de forma que sua distribuição passe a ser dada por $B_{\widehat{C}}$, como na Equação (3.16). Este procedimento não é tão direto quanto dizer que, caso o processo de escrita mude um coeficiente do conjunto de escrita de Y para Z, deve-se encontrar um coeficiente Z no conjunto de compensação e modificá-lo para Y. Se o processo de escrita modificasse outro símbolo hospedeiro de Z para Y, o procedimento de compensação anterior não seria necessário. Em [2], os autores usam um método de compensação com critério de mínimo erro médio quadrático (MMSE) [1] para mudar a distribuição de C e atingir um erro ϵ pequeno.

O problema de modificação de histograma com critério MMSE foi primeiramente considerado por Mese e Vaidyanathan [28]. Os autores propuseram a solução de um problema de programação linear para obter uma matriz de mapeamento entre os dados. Tzschoppe et al. [1] mostram que existe uma solução mais simples, que não requer a solução de uma problema de programação linear. A descrição dessa solução é apresentada a seguir.

Geralmente dados com valores iguais são mapeados para um mesmo valor. Desta forma, o histograma alvo $B_{\widehat{C}}$ pode ser somente aproximado, onde essa aproximação depende, fortemente, da natureza do histograma dos dados de entrada. Pode-se alcançar, perfeitamente, um histograma alvo, se o mapeamento dos dados permitir que porções dos dados com valores idênticos sejam mapeados para diferentes valores na saída.

Demanda-se que o mapeamento $C \rightarrow \widehat{C}$ introduza a mínima distorção possível entre os coeficientes originais e modificados. O erro médio quadrático (MSE) é utilizado como medida de distorção. Uma importante consequência do uso dessa medida de distorção é que o mapeamento $C \rightarrow \widehat{C}$ com mínimo MSE deve presevar as relações de ordem entre diferentes elementos. Ou seja, dados quaisquer dois coeficientes de entrada c_1 e c_2 tais que $c_1 \geq c_2$, os coeficientes mapeados correspondentes $\widehat{c}_1$ e $\widehat{c}_2$ devem satisfazer $\widehat{c}_1 \geq \widehat{c}_2$. Esta propriedade é mostrada no Apêndice A.

Para tornar a explicação mais clara, o texto a seguir utiliza a mesma notação proposta pelos autores de [1]. No final desta seção, a correspondência entre os símbolos utilizados por Tzschoppe et al. e os símbolos utilizadaos por Sarkar e Manjunath é indicada explicitamente na Tabela 3.1.

Seja f uma variável aleatória com o alfabeto finito χ de cardinalidade N_f apresentado na Equação (3.28). A PMF de f é P_f (histograma de entrada). Os dados de f são modificados para g, que possui PMF P_g (histograma alvo ou desejado).

$$\chi = \{f^{(1)}, f^{(2)}, \ldots, f^{(N_f)}\} \tag{3.28}$$

O processo de mapeamento dos dados $f \rightarrow g$ de forma que P_f fique igual P_g pode ser completamente caracterizado pelo quantizador escalar Q_t, que por sua vez é caracterizado pelo conjunto de níveis de decisão $\tau = \{t_1, t_2, \ldots, t_{N_f-1}\}$.

O algoritmo de mapeamento atua sobre os índices i dos possíveis símbolos de

entrada $f^{(i)}$. Seja $i_n \in \{1, 2, \ldots, N_f\}$ o índice de um símbolo de entrada f_n. Esse índice é mapeado, aleatoriamente, numa variável aleatória contínua t de acordo com a Equação (3.29), onde a_n é sorteado de uma variável aleatória uniforme a no intervalo $[0, 1)$.

$$t_n = i_n - a_n \tag{3.29}$$

Sendo as variáveis aleatórias i_n e a_n independentes, a PDF (*Probability Mass Function*) da variável aleatória t é dada pela Equação (3.30).

$$\widetilde{P}_f(t) = \sum_{i=1}^{N_f} P_f[i].rect\left(t + \frac{1}{2} - i\right), \text{ onde } rect(x) = \begin{cases} 1 & \text{, se } -\frac{1}{2} < x \leq \frac{1}{2} \\ 0 & \text{, caso contrário} \end{cases} \tag{3.30}$$

As PDFs associadas às funções massa $P_f[i]$ e $P_g[j]$ são dadas, respectivamente, por:

$$P_f(t) = \sum_{i=1}^{N_f} P_f[i].\delta(t - i) \tag{3.31}$$

$$P_g(t) = \sum_{j=1}^{N_f} P_g[j].\delta(t - j). \tag{3.32}$$

O quantizador $Q_t(t)$ é definido na Equação (3.33).

$$Q_t(t) = \begin{cases} f^{(1)} & \text{, se } t \leq t_1 \\ f^{(j)} & \text{, se } t_{j-1} < t \leq t_j, \forall j \in \{2, 3, \ldots, N_f - 1\} \\ f^{(N_f)} & \text{, se } t_{N_f - 1} < t \end{cases} \tag{3.33}$$

O mapeamento dos dados é representado pela Equação (3.34), onde o conjunto τ de níveis de decisão do quantizador deve obedecer às Equações (3.35) e (3.36).

$$g_n = Q_t(t_n) = Q_t(i_n - a_n) \tag{3.34}$$

$$\int_0^1 P_g(\tau)d\tau = \int_0^{t_1} \widetilde{P}_f(\tau)d\tau \tag{3.35}$$

$$\int_{j-1}^{j} P_g(\tau)d\tau = \int_{t_{j-1}}^{t_j} \widetilde{P}_f(\tau)d\tau, j \in \{2, 3, \ldots, N_f - 1\} \tag{3.36}$$

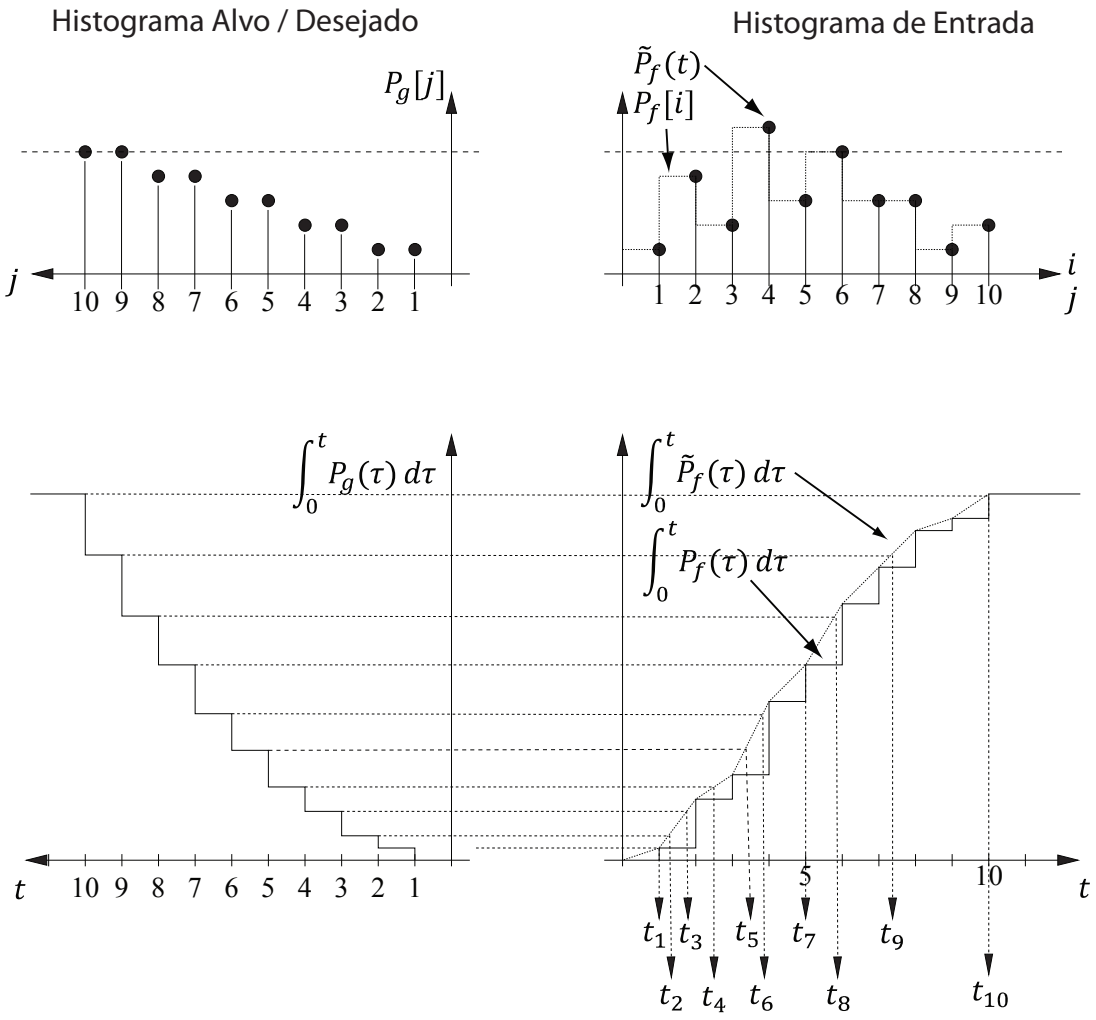


Figura 3.3: Exemplo de modificação de histograma conforme [1], com $N_f = 10$ usando as Equações (3.33) a (3.36)

A Figura 3.3 apresenta um exemplo com $N_f = 10$.

Na prática, Sarkar e Manjunath dividem o histograma dos coeficientes da DCT em dois outros histogramas. O primeiro deles é formado pelos *bins* de 0 até T_{opt}. O segundo deles é formado pelos *bins* de $-T_{opt}$ até -1. O procedimento de compensação é aplicado a cada um desses histogramas separadamente.

Tabela 3.1: Correspondência entre os símbolos utilizados por Tzschoppe et al. e os símbolos utilizados por Sarkar e Manjunath

Tzschoppe	Sarkar
f	C
$-T_{opt}, \ldots, T_{opt}$	$f^1, \ldots, f^{N_f}$
g	$\widehat{C}$
P_f	B_C
P_g	$B_{\widehat{C}}$

3.4 Proposta A

Esta seção apresenta a primeira proposta alternativa para se ocultar informação em imagens digitais. O objetivo desta, e das outras propostas alternativas à abordagem feita por Sarkar e Manjunath, é ocultar mais informação de forma segura.

O processo de escrita dos bits da marca d'água pode ser realizado de forma a ajudar automaticamente o processo de compensação. Uma forma de fazer isto é apresentada na Seção 3.4.1.

Na Seção 3.4.2, propõe-se um método de compensação de histogramas mais intuitivo do que aquele proposto por Tzschoppe et al.

Todas as considerações feitas na Seção 3.3.1 continuam valendo tanto para esta nova proposta quanto para as Propostas B e C, que serão apresentadas nas seções seguintes. Por simplicidade, os parâmetros T_{opt}, R_{opt} e λ_{opt} continuam sendo calculados de acordo com o que foi apresentado na Seção 3.3.3.

3.4.1 Método de Escrita

Este método de escrita visa melhorar o método de escrita par/ímpar. O processo de escrita continua a atuar apenas sobre o bit menos significativo de cada coeficiente do conjunto H.

O método de escrita par/ímpar da abordagem de Sarkar e Manjunath altera, caso necessário, o bit menos significativo de um coeficiente da DCT, para esconder um bit da marca d'água. Quando esta alteração é necessária, o algoritmo usado por eles sorteia, equiprovavelmente, se este coeficiente vai ser somado ou subtraído de 1.

Este sorteio não necessariamente facilita o processo de compensação. No caso em que se faz necessária a alteração da paridade de um coeficiente hospedeiro de um bit da marca d'água, infere-se que há casos em que se é mais interessante somar 1 a este coeficiente. Em outros casos, pode ser mais interessante subtrair 1 deste coeficiente. Uma escolha aleatória da operação a ser realizada sobre o coeficiente hospedeiro não se mostra ideal.

Propõe-se a criação de um histórico de alterações sofridas por cada valor de coeficiente da DCT disponível para escrita. Os valores possíveis desses coeficientes hospedeiros pertencem ao intevalo $[-T_{opt}, T_{opt}]$. Este histórico de alterações tem a função de ajudar o processo de escrita a decidir por somar ou subtrair 1 a um coeficiente da DCT, caso ele tenha que ser alterado. Este procedimento tenderá a fazer parte da compensação automaticamente durante a própria escrita da marca d'água.

O histórico de alterações pode ser representado por um vetor de tamanho $2T_{opt} + 1$. A primeira posição do vetor guarda o histórico de alterações dos coeficientes da DCT com valores iguais a $-T_{opt}$. A segunda posição do vetor guarda o histórico de alterações dos coeficientes da DCT com valores iguais a $-T_{opt}+1$. A $(2T_{opt}+1)$-ésima posição do vetor guarda o histórico de alterações dos coeficientes da DCT com valores iguais a T_{opt}.

Inicialmente, todas as posições do vetor histórico de alterações valem 0. As quantidades iniciais de cada coeficiente disponível para escrita são tomadas como referência. Caso um coeficiente que valha a seja alterado para b, o processo de escrita deverá subtrair 1 na posição referente aos elementos que valem a no vetor histórico de alterações, pois o *déficit* de elementos que valem a aumentará de 1.

Simultaneamente, o processo de escrita deverá somar 1 na posição referente aos elementos que valem b no vetor histórico de alterações, pois haverá mais um novo elemento valendo b.

Vale notar que $b = a - 1$ ou $b = a + 1$, pois apenas o bit menos significativo do coeficiente da DCT é alterado, se necessário, para carregar a informação de um bit da marca d'água. Para ajudar o processo de compensação, o processo de escrita deve alterar a para $a-1$, caso o *déficit* de elementos valendo $a-1$ seja maior que o *déficit* de elementos valendo $a+1$. Caso os valores do vetor histórico de alterações referentes aos coeficientes $a-1$ e $a+1$ sejam iguais, o coeficiente a deve ser alterado para $a-1$ ou para $a+1$ aleatoriamente.

Observa-se que o histórico de alterações fica mais rico na medida em que mais bits da marca d'água vão sendo inseridos sobre os coeficientes da DCT.

3.4.2 Método de Compensação

Propõe-se um método de compensação de histogramas mais intuitivo do que aquele proposto por Tzschoppe et al. O seu funcionamento é análogo ao preenchimento das sucessivas torres do histograma com um "líquido", como se este líquido "escorresse" das torres com excesso de probabilidade para as torres com *déficit* de probabilidade.

Dados o histograma alvo, $B_{\widehat{C}}$, e o histograma de entrada, B_C, deve-se "escorrer" as torres do histograma de entrada até que todas as suas torres fiquem iguais àquelas do histograma alvo. Isto é feito através do mapeamento dos dados originais do subconjunto C, geradores do histograma de entrada B_C, num novo conjunto de dados $\widehat{C}$.

Os *bins* de B_C e $B_{\widehat{C}}$ são analisados em ordem crescente. Assim, o algoritmo começa analisando os *bins* mais à esquerda de ambos histogramas. Por exemplo, caso o *bin* mais à esquerda do histograma de entrada possua uma torre menor do que o *bin* mais à esquerda do histograma alvo ($B_C(-T_{opt}) < B_{\widehat{C}}(-T_{opt})$), necessita-se aumentar a torre mais à esquerda do histograma de entrada. Isto é feito através do mapeamento do número necessário de elementos imediatamente maiores do que o *bin* em análise para o valor do *bin* que apresenta *déficit* em sua torre. Ou seja, procuram-se coeficientes iguais a $-T_{opt}+1$ no subconjunto C para serem mapeados para $-T_{opt}$, de forma a fazer com que $B_C(-T_{opt})$ fique igual a $B_{\widehat{C}}(-T_{opt})$. Caso não haja uma

quantidade suficiente de coeficientes iguais a $-T_{opt}+1$ para realizar a compensação da torre sobre o *bin* $-T_{opt}$, procuram-se coeficientes iguais a $-T_{opt}+2$ em C para serem mapeados para $-T_{opt}$. Este procedimento segue até que $B_C(-T_{opt}) = B_{\widehat{C}}(-T_{opt})$.

Analogamente, caso $B_C(i) > B_{\widehat{C}}(i)$, elementos iguais a i do subconjunto C são mapeados para $i+1$, de forma que se tenha $B_C(i) = B_{\widehat{C}}(i)$.

Resumindo, este processo compensa o histograma de entrada B_C, torre a torre, em função do histograma alvo $B_{\widehat{C}}$.

3.5 Proposta B

Esta seção apresenta a segunda proposta alternativa para ocultar marcas d'água grandes em histogramas de coeficientes da DCT de blocos de imagens digitais. Assim como a Proposta A, apresentada na Seção 3.4.1, esta proposta apresenta um novo procedimento de escrita da marca d'água, que objetiva diminuir a utilização do processo de compensação. A consequência direta das Propostas A e B é o aumento do tamanho da informação que pode ser inserida de forma que um esteganalista não perceba a presença da mensagem.

O novo método de escrita é apresentado na Seção 3.5.1. Optou-se por continuar utilizando o método de compensação do tipo "escorrimento d'água" nesta abordagem, pois este método é simples e apresenta bons resultados. Assim, esta seção não apresenta uma subseção dedicada a um método de compensação, pois o mesmo está descrito na Seção 3.4.2.

Assim como na seção anterior, as definições e considerações feitas nas Seções 3.3.1 e 3.3.3 continuam valendo para a Proposta B.

3.5.1 Método de Escrita

Este método de escrita também escreve a marca d'água sobre os bits menos significativos dos coeficientes do subconjunto H.

Divide-se o histograma B_H, formado pelos coeficientes do subconjunto H de coeficientes da DCT disponíveis para escrita, em dois novos histogramas. O primeiro deles contém os *bins* de $-T_{opt}$ até -1. O segundo deles contém os *bins* de 1 até T_{opt}. O processo de escrita é realizado separadamente em relação a cada um dos novos

histogramas. A escrita sobre coeficientes de H iguais a 0 é feita pelo método de escrita par/ímpar.

A idéia é que o processo de escrita comece pelos coeficientes menos frequentes de cada histograma novo. Por exemplo, os coeficientes menos frequentes no primeiro histograma novo são os de valor $-T_{opt}$. Esses coeficientes, quando precisarem ser alterados, só poderão ser mapeados para $-T_{opt}+1$. Em seguida, o processo de escrita deve escrever sobre os coeficientes $-T_{opt}+1$ do subconjunto H, que são os segundos coeficientes menos frequentes do primeiro histograma novo. Os coeficientes iguais a $-T_{opt}+1$ de H que precisarem ser alterados, devem ser primeiramente mapeados para $-T_{opt}$, de forma a compensar o *déficit* de coeficientes $-T_{opt}$ que a iteração anterior do processo de escrita causou. Os coeficientes $-T_{opt}+1$ que não forem mapeados para $-T_{opt}$, deverão ser mapeados, obrigatoriamente, para $-T_{opt}+2$. Isto causará um desbalanceamento sobre o *bin* $-T_{opt}+2$ do primeiro histograma novo. Este desbalanceamento tenderá a ser compensado pelo processo de escrita sobre os coeficientes iguais a $-T_{opt}+3$. O procedimento segue, sucessivamente, e é análogo para o segundo histograma novo, com *bins* positivos.

3.6 Proposta C

Como será visto no Capítulo 4, os resultados da Proposta A ficaram melhores do que os da Proposta B. Isto motivou a criação da Proposta C, que pode ser vista como uma mistura da Proposta A e da abordagem de Sarkar e Manjunath.

A Proposta C utiliza o método de escrita da Proposta A, apresentado na Seção 3.4.1, e o método de compensação usado por Sarkar e Manjunath, descrito na Seção 3.3.4.

3.7 Comentários Finais

Além dos métodos de escrita propostos nas Seções 3.4.1 e 3.5.1, foram consideradas, ainda, algumas abordagens alternativas, que não deram bons resultados.

Como exemplo, pode-se descrever o "método da escrita por solução de sistema linear sobre-determinado". Como explicado na Seção 3.3.2, assume-se que a marca d'água a ser ocultada tenha o mesmo número de bits 0 e 1 que afetam os elementos

em $X(i)$[2]. Assim, tenta-se prever como o método de escrita par/ímpar deve tomar a decisão de somar ou subtrair uma unidade de um coeficiente da DCT, caso este coeficiente tenha que ser alterado para carregar um bit da marca d'água.

Tal previsão pode ser representada por porcentagens de escolha entre cada uma das duas operações possíveis (soma ou subtração). Dada uma torre $B_H(i)$, tem-se que, após o processo de escrita da marca d'água, sua altura $B_{\widehat{H}}(i)$ passará a ser dada pela Equação (3.37). Para que o processo de escrita facilite o processo de compensação, deve-se fazer com que $B_{\widehat{H}}(i) = B_H(i)$ ainda no processo de escrita.

$$B_{\widehat{H}}(i) = \frac{1}{2}B_H(i) + \frac{1}{2}\gamma_{i-1,i}B_H(i-1) + \frac{1}{2}\gamma_{i+1,i}B_H(i+1) = B_H(i) \tag{3.37}$$

Na Equação (3.37), assume-se que metade dos coeficientes iguais a i permanecerão inalterados. Além disto, dos aproximadamente 50% de termos $i-1$ que podem ser mapeados para i, apenas uma porcentagem $\gamma_{i-1,i}$ sofrerá o mapeamento. Analogamente, dos aproximadamente 50% de termos $i+1$ que podem ser mapeados para i, apenas uma porcentagem $\gamma_{i+1,i}$ sofrerá o mapeamento. Deve-se notar que um termo i pode ser mapeado somente para $i-1$ ou $i+1$ (Equação (3.38)).

$$\gamma_{i,i-1} + \gamma_{i,i+1} = 1 \tag{3.38}$$

Desta forma, dado um histograma com $2T_{opt}+1$ *bins*, tem-se um sistema de $2(2T_{opt}+1)-2$ equações e $2(2T_{opt}+1)-4$ incógnitas, o que caracteriza um sistema linear sobre-determinado. Este sistema pode ser representado pela Equação (3.39), onde A é a matriz de tamanho $(2(2T_{opt}+1)-2) \times (2(2T_{opt}+1)-4)$ com os termos que multiplicam as incógnitas $\gamma_{m,n}$ nas Equações (3.37) e (3.38), Γ é o vetor coluna que contém as incógnitas $\gamma_{m,n}$ do sistema e b é o vetor coluna com os termos que aparecem nos lados direitos das Equações (3.37) e (3.38).

$$A\Gamma = b \tag{3.39}$$

A solução deste sistema é dada pela Equação (3.40), onde I é a matriz identidade

[2] $X(i)$ é o subconjunto de X tal que os seus elementos pertençam a X e sejam iguais a i.

de tamanho $(2(2T_{opt}+1)-2)\times(2(2T_{opt}+1)-2)$ e λ é um multiplicador de Lagrange.

$$\Gamma = (A^T A + \lambda I)^{-1} A^T b \tag{3.40}$$

Deve-se encontrar Γ que minimize a função custo J, apresentada na Equação (3.41). O multiplicador de Lagrange λ representa a restrição que $\gamma_{m,n} \leq 1$.

$$J = ||A\Gamma - b||^2 + \lambda||\Gamma||^2 \tag{3.41}$$

A restrição $0 \leq \gamma_{m,n} \leq 1$ se aplica pois $\gamma_{m,n}$ representa uma porcentagem. A solução final deste sistema não apresentou bons resultados na prática. Diversos valores foram atribuídos ao multiplicador de Lagrange λ, para buscar a melhor solução Γ. Porém, verificou-se que o sistema proposto frequentemente apresentava soluções como $\gamma_{i,i-1} = 1{,}4$ e $\gamma_{i,i+1} = -0{,}4$, o que é impossível na prática. Os resultados deste método não serão apresentados neste texto.

Por exemplo, seja $T_{opt} = 2$, $B_H(-2) = 0{,}10$, $B_H(-1) = 0{,}13$, $B_H(0) = 0{,}50$, $B_H(1) = 0{,}17$ e $B_H(2) = 0{,}10$. Assim, têm-se 6 incógnitas ($\gamma_{-1,-2}$, $\gamma_{-1,0}$, $\gamma_{0,-1}$, $\gamma_{0,1}$, $\gamma_{1,0}$, $\gamma_{1,2}$) e as 8 equações a seguir:

$$\begin{gathered} B_{\widehat{H}}(-2) = \tfrac{1}{2}B_H(-2) + \tfrac{1}{2}\gamma_{-1,-2}B_H(-1) = B_H(-2) \\ 0{,}13\gamma_{-1,-2} = 0{,}10 \end{gathered} \tag{3.42}$$

$$\begin{gathered} B_{\widehat{H}}(-1) = \tfrac{1}{2}B_H(-1) + \tfrac{1}{2}\gamma_{-2,-1}B_H(-2) + \tfrac{1}{2}\gamma_{0,-1}B_H(0) = B_H(-1) \\ 0{,}10 \times 0{,}5 + 0{,}50\gamma_{0,-1} = 0{,}13 \end{gathered} \tag{3.43}$$

$$\begin{gathered} B_{\widehat{H}}(0) = \tfrac{1}{2}B_H(0) + \tfrac{1}{2}\gamma_{-1,0}B_H(-1) + \tfrac{1}{2}\gamma_{1,0}B_H(1) = B_H(0) \\ 0{,}13\gamma_{-1,0} + 0{,}17\gamma_{1,0} = 0{,}50 \end{gathered} \tag{3.44}$$

$$\begin{gathered} B_{\widehat{H}}(1) = \tfrac{1}{2}B_H(1) + \tfrac{1}{2}\gamma_{0,1}B_H(0) + \tfrac{1}{2}\gamma_{2,1}B_H(2) = B_H(1) \\ 0{,}50\gamma_{0,1} + 0{,}10 \times 0{,}5 = 0{,}17 \end{gathered} \tag{3.45}$$

$$\begin{gathered} B_{\widehat{H}}(2) = \tfrac{1}{2}B_H(2)\tfrac{1}{2}\gamma_{1,2}B_H(1) = B_H(2) \\ 0{,}17\gamma_{1,2} = 0{,}1 \end{gathered} \tag{3.46}$$

$$\gamma_{-1,-2} + \gamma_{-1,0} = 1 \tag{3.47}$$

$$\gamma_{0,-1} + \gamma_{0,1} = 1 \tag{3.48}$$

$$\gamma_{1,0} + \gamma_{1,2} = 1 \tag{3.49}$$

Das equações acima, pode-se escrever:

$$A = \begin{bmatrix} 0{,}13 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}50 & 0 & 0 & 0 \\ 0 & 0{,}13 & 0 & 0 & 0{,}17 & 0 \\ 0 & 0 & 0 & 0{,}50 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}17 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \Gamma = \begin{bmatrix} \gamma_{-1,-2} \\ \gamma_{-1,0} \\ \gamma_{0,-1} \\ \gamma_{0,1} \\ \gamma_{1,0} \\ \gamma_{1,2} \end{bmatrix}, b = \begin{bmatrix} 0{,}10 \\ 0{,}13 - 0{,}10 \times 0{,}5 \\ 0{,}50 \\ 0{,}17 - 0{,}10 \times 0{,}5 \\ 0{,}10 \\ 1 \\ 1 \\ 1 \end{bmatrix}. \tag{3.50}$$

Utilizando a Equação (3.40) e fazendo $\lambda = 0{,}001$, por exemplo, tem-se

$$\Gamma = \begin{bmatrix} -0{,}21 \\ 1{,}22 \\ 0{,}43 \\ 0{,}51 \\ 1{,}20 \\ -0{,}17 \end{bmatrix}, \tag{3.51}$$

que possui dois valores negativos. Tais valores negativos tornam esta solução impraticável.

Capítulo 4

Resultados e Discussões

Dois sistemas esteganalistas foram implementados para testar as abordagens apresentadas no Capítulo 3. Deseja-se identificar qual abordagem é capaz de ocultar mais informação nas imagens. Os resultados obtidos são apresentados neste capítulo.

O banco de imagens utilizado nesse trabalho foi composto por 1200 imagens em preto e branco com 256 níveis de cinza no formato TIF [29]. As dimensões das imagens deste conjunto variam entre 256×256, 512×512 e 1024×1024, em $pixels \times pixels$.

4.1 Esteganálise

As versões do sistema esteganalista (Willie) foram implementadas através de redes neurais artificiais supervisionadas do tipo *perceptron* multi-camadas (*feed-forward*), treinadas por *error backpropagation*[1]. O Apêndice B apresenta um breve resumo sobre redes neurais e os parâmetros de treinamento utilizados. As redes neurais têm a função de classificar imagens como *cover* ou *stego*, ou seja, como imagem não marcada ou marcada.

O conjunto de treinamento de ambas redes neurais foi formado por histogramas dos coeficientes das DCTs dos blocos 8×8 de 800 imagens. 400 dessas imagens foram marcadas, utilizando-se o fator de escrita λ_{opt}, enquanto que as 400 imagens restantes foram preservadas. Todos os histogramas foram pré-processados por Análise de

[1]Por simplicidade, não foram considerados neste trabalho outros tipos de classificadores, como por exemplo o SVM (*Support Vector Machine*) que foi utilizado por Sarkar e Manjunath. O banco de imagens deles foi composto por 4500 imagens.

Componentes Principais (PCA) antes de serem apresentados às redes neurais. Após este pré-processamento, os vetores de entrada, que eram histogramas com 301 componentes, assim como em [2], passaram a ter apenas 3 componentes — mantendo 99 % da energia dos vetores de 301 componentes.

A PCA é uma transformada linear ortogonal bastante útil para reduzir a dimensão de conjuntos de dados. A PCA transforma os dados de forma que se possa descartar componentes de baixas energias, que carregam pouca informação. Isto é possível pois essa transformada descorrelaciona os dados originais. A PCA também é chamada de transformada de Karhunen-Loève [25].

Para um vetor $\mathbf{u}$ de cardinalidade N formado por números reais, os vetores base da PCA são dados pelos autovetores normalizados de sua matriz de covariância[2] $\mathbf{R}$, ou seja,

$$\mathbf{R}\phi_{\mathbf{k}} = \lambda_k \phi_{\mathbf{k}},\ 0 \leq k \leq N-1. \tag{4.1}$$

A PCA de $\mathbf{u}$ é definida pela Equação (4.2).

$$\mathbf{v} = \mathbf{\Phi}^{*\mathbf{T}}\mathbf{u} \tag{4.2}$$

$\phi_{\mathbf{k}}$ é a k-ésima coluna de $\mathbf{\Phi}$.

Para o treinamento da primeira rede neural ("especializada em Sarkar"), marcaram-se 400 dentre as 800 imagens do conjunto de treinamento pelo método de escrita par/ímpar. Para o treinamento da segunda rede neural ("especializada na Proposta A"), marcaram-se as mesmas 400 imagens dentre as 800 existentes no conjunto de treinamento pelo método de escrita da Proposta A (ou C). Ambos os conjuntos de treinamento foram apresentados juntos com um vetor alvo, que indicava cada histograma como *cover* ou *stego*, às redes neurais, para treiná-las. Utilizou-se o fator de escrita λ_{opt} para marcar as imagens.

As 400 imagens restantes, que não foram utilizadas para o treinamento das redes neurais, foram utilizadas para formar o conjunto de teste. Todas essas 400 imagens, utilizadas para teste, foram escritas e compensadas pelos métodos abordados. Em seguida, os histogramas de suas DCTs foram pré-processados por PCA, da mesma forma que o conjunto de treinamento, e foram apresentados para as redes neurais.

[2]Matriz de covariância: $\mathbf{R} = cov[\mathbf{u}] = E[(\mathbf{u}-\mu)(\mathbf{u}-\mu)^T]$, onde $\mu = E[\mathbf{u}]$.

Se a compensação funcionar bem, o esperado é que as redes neurais não identifiquem as imagens marcadas e compensadas como imagens marcadas, através da análise dos histogramas dos coeficientes das DCTs. Desta forma, consegue-se uma comunicação segura entre duas partes, pois o esteganalista não percebe a marca d'água oculta.

4.2 Testes

Para se verificar qual método consegue escrever mais informação de forma segura nas imagens, cada imagem de teste é escrita e compensada até o limite em que a rede neural não consiga identificar a presença da marca d'água. Assim, pode-se avaliar a quantidade máxima de informação que cada método consegue inserir sobre as imagens. Inicialmente, o fator de escrita utilizado para escrever em cada imagem é λ_{opt}, calculado na Equação (3.17). Enquanto a rede neural não identifica a imagem marcada e compensada como marcada, o fator de escrita vai sendo incrementado de 0.02. O tamanho da marca d'água aumenta de forma segura enquanto o custo dispendido pelo processo de compensação diminui.

Os parâmetros utilizados em todos os testes feitos nesse trabalho são os mesmos utilizados por Sarkar e Manjunath [2], para facilitar a avaliação dos resultados. O fator de qualidade utilizado foi 75, o que gerou a matriz de quantização M_{75} apresentada na Equação (4.3). Esta matriz é utilizada para normalizar as DCTs dos blocos 8×8 das imagens, como explicado no início da Seção 3.3.2.

$$M_{75} = \begin{pmatrix} 8 & 6 & 5 & 8 & 12 & 20 & 26 & 31 \\ 6 & 6 & 7 & 10 & 13 & 29 & 30 & 28 \\ 7 & 7 & 8 & 12 & 20 & 29 & 35 & 28 \\ 7 & 9 & 11 & 15 & 26 & 44 & 40 & 31 \\ 9 & 11 & 19 & 28 & 34 & 55 & 52 & 39 \\ 12 & 18 & 28 & 32 & 41 & 52 & 57 & 46 \\ 25 & 32 & 39 & 44 & 52 & 61 & 60 & 51 \\ 36 & 46 & 48 & 49 & 56 & 50 & 52 & 50 \end{pmatrix} \tag{4.3}$$

A banda de frequência escolhida sobre a DCT de cada bloco 8×8 de cada imagem para escrita e compensação é composta pelos primeiros 19 coeficientes AC tomados

em ziguezague, como mostrado na Figura 4.1 (coeficientes de 1 a 19).

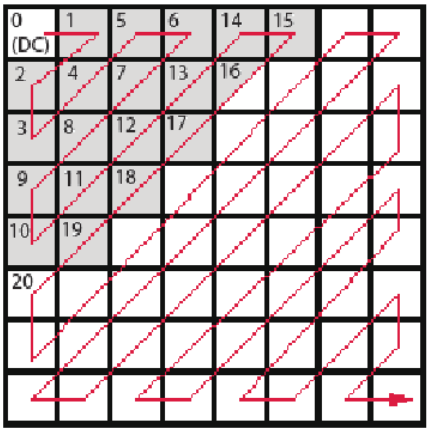


Figura 4.1: 19 coeficientes AC da DCT tomados em ziguezague (fundos em cinza)

Admite-se que T_{opt} valha no máximo 30 durante o processo de escrita da marca d'água. Isto é feito para garantir que o processo de escrita não escreva sobre coeficientes da DCT de magnitudes elevadas, pois eles são pouco frequentes e mais difíceis de serem compensados.

Para tornar o processo de leitura da marca d'água possível, tanto o codificador quanto o decodificador devem possuir uma chave secreta. Esta chave é a semente de um gerador de números pseudo-aleatórios que os dois possuem. A sequência pseudo-aleatória, gerada a partir desta semente, permite que o decodificador saiba em quais coeficientes da DCT a mensagem está inserida. Além disto, separam-se os 8 primeiros bits da marca d'água, para carregar a informação do limiar T_{opt}. Isto é feito pois o decodificador precisa saber se um coeficiente da DCT está carregando informação, o que pode acontecer quando seu módulo é menor ou igual a T_{opt}. Os 20 bits subsequentes carregam a informação do tamanho da própria marca d'água.

As figuras de mérito mais importantes para as análises dos resultados são o fator de escrita máximo alcançado em cada método, o máximo número de bits de informação inseridos por *pixel* da imagem (bit/*pixel*), a fração máxima R_{opt} de termos que realmente pode ser usada para escrever uma mensagem em função do limiar T_{opt} e a razão sinal-ruído de pico (PSNR, *Peak Signal-to-Noise Ratio*) entre as imagens original e marcada, e entre as imagens original e marcada e compensada. A PSNR em decibéis entre duas imagens em preto e branco com 256 níveis de cinza é

dada pela Equação (4.4). Quanto maior for a PSNR, maior será a similaridade entre as imagens. A grosso modo, imagens com PSNR em torno de 40 dB correspondem a uma boa qualidade na reprodução em relação à outra imagem.

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{255^2}{\text{MSE entre as imagens}} \right) \tag{4.4}$$

4.3 Resultados

As Figuras 4.2, 4.3, 4.4 e 4.5 apresentam resultados ilustrativos dos métodos do Sarkar e Manjunath e das Propostas A, B e C aplicados sobre a imagem LENA de tamanho 256×256. O fator de escrita foi $\lambda = 0,6237$ e o tamanho da marca d'água foi de 11.951 bits (0,185 bits/*pixel*).

Além dos testes ilustrativos aplicados sobre a imagem LENA, outro teste preliminar foi realizado. Todos os métodos foram aplicados a todas as 1200 imagens. Em seguida, avaliaram-se os valores médios de cada figura de mérito. Neste ponto, as redes neurais (esteganalistas) não foram aplicadas, pois aqui o interesse diz respeito somente aos valores médios das figuras de mérito calculadas com todos os parâmetros idênticos aos de Sarkar e Manjunath, incluindo o fator de escrita. Este experimento preliminar é útil para validar os resultados obtidos pela implementação do método de Sarkar e Manjunath realizada nesse trabalho.

A Tabela 4.1 apresenta as médias dos resultados preliminares tomados sobre as 1200 imagens. PSNR (H) representa a razão sinal-ruído em decibéis entre a imagem marcada e a imagem original. PSNR (H-C) representa a razão sinal-ruído em decibéis entre a imagem marcada e compensada e a imagem original. σ_{PSNR} (H) e σ_{PSNR} (H-C) representam os desvios-padrão associados às figuras de mérito PSNR (H) e PSNR (H-C), respectivamente. A coluna S & M representa os resultados referentes à implementação do método de Sarkar e Manjunath realizada neste trabalho. A coluna seguinte (S & M [2]) apresenta os resultados publicados pelos autores. Os autores não apresentam resultados referentes à PSNR.

Analisando-se a segunda e a terceira coluna da Tabela 4.1, verifica-se que os resultados encontrados pela implementação feita neste trabalho sobre o método de Sarkar e Manjunath estão corretos. Como neste teste preliminar não foram treinadas

Original

Marcada (PSNR = 40,9 dB)

Marcada e Compensada (PSNR = 36,0 dB)

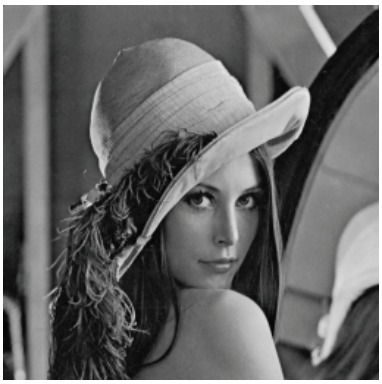
Figura 4.2: Método do Sarkar e Manjunath sobre a imagem LENA

Original

Marcada (PSNR = 41,0 dB)

Marcada e Compensada (PSNR = 39,6 dB)

Figura 4.3: Método da Proposta A sobre a imagem LENA



Original

Marcada (PSNR = 40,9 dB)

Marcada e Compensada (PSNR = 38,3 dB)

Figura 4.4: Método da Proposta B sobre a imagem LENA

Original

Marcada (PSNR = 41,0 dB)

Marcada e Compensada (PSNR = 37,7 dB)

Figura 4.5: Método da Proposta C sobre a imagem LENA

Tabela 4.1: Resultados preliminares sobre as 1200 imagens

Parâmetro	S & M	S & M [2]	Proposta A	Proposta B	Proposta C
T_{opt}	28	27	28	28	28
λ_{opt}	0,4642	0,4834	0,4642	0,4642	0,4642
R_{opt}	0,472	0,502	0,472	0,472	0,472
PSNR (H)	42,2	-	42,3	42,3	42,3
σ_{PSNR} (H)	1,6	-	1,5	1,6	1,5
PSNR (H-C)	37,2	-	40,4	39,9	38,2
σ_{PSNR} (H-C)	1,0	-	1,1	1,1	0,7
bits / *pixel*	0,136	0,141	0,136	0,136	0,136

redes neurais, o principal motivo para os resultados apresentarem uma pequena diferença entre si é a utilização de bancos de imagens diferentes.

Verifica-se que T_{opt}, λ_{opt}, R_{opt} e "bits/*pixel*" apresentam os mesmos resultados para todos os métodos, com exceção dos resultados da terceira coluna da tabela (extraídos de [2]). Isto ocorre pois os métodos de escrita implementados nesse trabalho calculam o fator de escrita λ_{opt} igualmente, como apresentado na Seção 3.3.3.

O método que, na média, distorce menos as imagens marcadas e compensadas é o da Proposta A. O método que causa a maior distorção nas imagens é o proposto por Sarkar e Manjunath. Da Tabela 4.1, tem-se $\text{PSNR}_{S\&M}$ (H-C) = 37,2 dB, enquanto que $\text{PSNR}_{PropostaA}$ (H-C) = 40,4 dB.

4.3.1 Resultados Finais

A primeira parte dos resultados finais, analisados sobre as 400 imagens do conjunto de teste usando como esteganalista a rede neural treinada sobre o método de escrita de Sarkar e Manjunath, são apresentados na Tabela 4.2.

Deve-se observar que agora se tem λ_{max} no lugar de λ_{opt}. Isto acontece porque esse teste final encontra o fator de escrita prático máximo λ_{max}, que permite escrever a maior marca d'água imperceptível tanto visual quanto estatisticamente.

Os resultados da Tabela 4.2 mostram que a Proposta C é a melhor abordagem. Observa-se que todos os resultados referentes à Proposta C são superiores em todas

Tabela 4.2: Resultados sobre as 400 imagens de teste com rede neural treinada sobre o método de escrita de Sarkar e Manjunath

Parâmetro	S & M	Proposta A	Proposta B	Proposta C
λ_{max}	0,5477	0,5429	0,5251	0,5759
R_{opt}	0,557	0,552	0,534	0,586
PSNR (H)	41,4	41,6	41,7	41,4
σ_{PSNR} (H)	1,2	1,3	1,3	1,1
PSNR (H-C)	36,8	37,2	37,2	37,9
σ_{PSNR} (H-C)	1,0	1,4	1,4	0,7
bits / *pixel*	0,161	0,159	0,154	0,169

as figuras de mérito.

A Prosposta C consegue inserir a maior marca d'água numa imagem digital sem que a rede neural treinada sobre o método de escrita de Sarkar e Manjunath a identifique. Sob o ponto de vista deste sistema esteganalista, a Proposta C insere, em média, $0,169$ bits de informação por *pixel* da imagem hospedeira. O método de Sarkar e Manjunath, que apresenta o segundo melhor desempenho, insere, em média, $0,161$ bits/*pixel*.

Os valores das PSNRs das imagens escritas e compensadas (PSNR (H-C)) ficaram no limite do aceitável para todas as abordagens. Uma imagem processada que apresente PSNR em torno de 37 dB fica no limite de apresentar ruído visualmente perceptível.

A segunda parte dos resultados finais diz respeito ao experimento com o esteganalista implementado pela rede neural treinada sobre o método de escrita usado nas Propostas A e C. Os resultados estão na Tabela 4.3.

Os resultados da Tabela 4.3 também mostram que a Proposta C é a melhor abordagem, inclusive quando o esteganalista é a rede neural treinada sobre o método de escrita das Propostas A e C.

A Proposta C insere 0,182 bits/*pixel*, enquanto que a abordagem de Sarkar e Manjunath, que aqui também apresenta o segundo melhor desempenho, insere 0,168

Tabela 4.3: Resultados sobre as 400 imagens de teste com rede neural treinada sobre o método de escrita das Propostas A e C

Parâmetro	S & M	Proposta A	Proposta B	Proposta C
λ_{max}	0,5716	0,5578	0,5430	0,6181
R_{opt}	0,582	0,568	0,553	0,629
PSNR (H)	41,3	41,5	41,6	41,1
σ_{PSNR} (H)	1,2	1,4	1,4	1,2
PSNR (H-C)	36,8	37,0	36,9	37,9
σ_{PSNR} (H-C)	1,0	1,2	1,3	0,8
bits / *pixel*	0,168	0,164	0,160	0,182

bits/*pixel*.

A Proposta C ocultou menos informação nas imagens quando ela foi submetida ao sistema esteganalista especializado na abordagem de Sarkar e Manjunath do que quando ela foi submetida ao sistema esteganalista especializado no método de escrita da própria Proposta C. Era esperado, intuitivamente, que a Proposta C conseguisse inserir mais informação nas imagens quando esta abordagem fosse submetida ao sistema esteganalista especializado na abordagem de Sarkar e Manjunath, pois a rede neural não teria conhecimento do método de escrita da Proposta C. Infere-se que isso tenha ocorrido devido ao processo de treinamento das redes neurais: o conjunto de treino gerado pela escrita da Proposta C parece estar levando a um treinamento mais difícil em termos do reconhecimento da própria escrita da Proposta C.

Os resultados das Tabelas 4.2 e 4.3 permitem afirmar que a Proposta C apresentou o melhor desempenho. O segundo melhor desempenho foi alcançado pela abordagem de Sarkar e Manjunath. Em seguida, colocam-se os resultados das Propostas A e B.

Capítulo 5

Conclusões

Neste trabalho, três abordagens de sistemas de esteganografia para imagens digitais foram propostas. O objetivo principal do trabalho foi inserir a maior marca d'água possível no domínio da transformada discreta do cosseno (DCT) de imagens digitais. As imagens marcadas não poderiam apresentar distorções visuais severas. Além disto, sistemas de esteganálise que investigassem a PMF dos coeficientes das DCTs dos blocos 8×8 das imagens marcadas não podiam identificar a presença da informação oculta.

Os resultados obtidos pelas Propostas A, B e C foram comparados com aqueles da implementação da abordagem de Sarkar e Manjunath [2]. As Propostas A e B apresentaram duas abordagens de escrita da marca d'água alternativas ao método de escrita par/ímpar, utilizado por Sarkar e Manjunath. Além disso, tanto a Proposta A quanto a Proposta B utilizaram uma abordagem nova para realizar o processo de compensação, que visava reparar o estrago feito na estatística da imagem original causado pelo processo de escrita. A Proposta C utilizou o método de escrita da Proposta A e o método de compensação de Tzschoppe et al. [1], utilizado por Sarkar e Manjunath.

Dois sistemas esteganalistas baseados em redes neurais artificiais supervisionadas do tipo *perceptron* multi-camadas (*feed-forward*) foram implementados para classificar as imagens escritas e compensadas como *cover* ou *stego*. Desta forma, verificou-se o máximo de informação que cada abordagem apresentada conseguiu inserir seguramente.

Os resultados apresentados no Capítulo 4 mostram que a Proposta C é a abor-

dagem que consegue ocultar o maior número de bits de informação por *pixel* da imagem. Quando o sistema esteganalista é especializado no método de escrita par/ímpar, a Proposta C oculta $0,169$ bits/*pixel*, enquanto que a implementação feita neste trabalho da abordagem de Sarkar e Manjunath oculta $0,161$ bits/*pixel*. Quando o sistema esteganalista é especializado no método de escrita utilizado nas Propostas A e C, a Proposta C oculta $0,182$ bits/*pixel*, enquanto que a implementação feita neste trabalho da abordagem de Sarkar e Manjunath oculta $0,168$ bits/*pixel*. Quanto à distorção visual, os resultados de todas abordagens são praticamente iguais. As imagens marcadas e compensadas ficaram no limite de apresentar ruído visível.

A área de esteganografia é relativamente nova, o que implica em inúmeras possibilidades de propostas para trabalhos futuros. Pesquisas e aplicações na área de segurança podem ser vistas como um jogo, onde métodos cada vez mais sofisticados de defesa e de ataque são criados. Sempre haverá demanda por melhorias em sistemas de segurança.

Uma proposta direta de trabalho futuro seria a abordagem em que a esteganalista, Willie, pode atacar a imagem que vai ser enviada do codificador, Alice, para o decodificador, Bob. Desta forma, faz-se necessária a incorporação de técnicas que tornem a marca d'água, além de imperceptível, robusta a ataques.

Também seria interessante estudar, a nível teórico, os algoritmos de escrita considerados neste trabalho, desenvolvendo expressões teóricas que permitam a previsão da capacidade de escrita para uma dada imagem.

Uma última idéia de trabalho futuro seria investigar os motivos pelos quais um esteganalista treinado com uma proposta de escrita acaba mais sensível à escrita da outra proposta.

Referências Bibliográficas

[1] TZSCHOPPE, R., BAUML, R., EGGERS, J., *Histogram modifications with minimum MSE distortion*, Relatório técnico, Telecommunication Laboratory, University of Erlangen-Nuremberg, Dezembro de 2001.

[2] SARKAR, A., MANJUNATH, B. S., "Estimating steganographic capacity for odd-even based embedding and its use in individual compensation". Em: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, v. 1, pp. 409–412, San Antonio, TX, EUA, Setembro de 2007.

[3] PETITCOLAS, F. A. P., ANDERSON, R. J., KUHN, M. G., "Information hiding - a survey", *Proceedings of the IEEE*, v. 87, no. 7, pp. 1062–1078, Julho de 1999.

[4] ANDERSON, R. J. (Ed.), *Information Hiding, First International Workshop, Cambridge, U.K., May 30 - June 1, 1996, Proceedings*, v. 1174, *Lecture Notes in Computer Science*. Ed. Springer, 1996.

[5] MILLER, M., COX, I., BLOOM, J., "Watermarking in the real world: an application to DVD". Em: *Proceedings of the Workshop at ACM Multimedia '98*, v. 41, pp. 71–76, Bristol, Reino Unido, Fevereiro de 1998.

[6] LANGELAAR, G. C., SETYAWAN, I., LAGENDIJK, R. L., "Watermarking digital image and video data: a state-of-the-art overview", *IEEE Signal Processing Magazine*, v. 17, no. 5, pp. 20–46, Septembro de 2000.

[7] COX, I. J., MILLER, M. L., "A review of watermarking and the importance of perceptual modeling". Em: *Proceedings of the SPIE*, v. 3016, no. 8, pp. 92–99, Fevereiro de 1997.

[8] SCHOTT, G., *Schola steganographica: in classes octo distributa....* [Cortesia do Museu Whipple de História da Ciência, Cambridge, Reino Unido], 1680.

[9] BAUER, F. L., *Decrypted Secrets: Methods and Maxims of Cryptology*. Ed. Springer-Verlag New York, Secaucus, NJ, EUA, 2000.

[10] REEDS, J., "Solved: the ciphers in book III of Trithemius' steganographia", *Cryptologia*, v. 22, no. 4, pp. 291–317, Outubro de 1998.

[11] COX, I. J., DOËRR, G., FURON, T., "Watermarking is not cryptography". Em: *Proceedings of the IWDW 2006*, v. 4283, pp. 1–15, Ilha Jeju, Coréia do Sul, Novembro de 2006.

[12] KERCKHOFFS, A., "La cryptographie militaire", *Journal des Sciences Militaires*, v. 9, pp. 5–83, Janeiro de 1883.

[13] ANDERSON, R., "Why cryptosystems fail", *Communications of the ACM*, v. 37, pp. 32–40, Novembro de 1994.

[14] BACON, F., *Of the advancement and proficiencie of learning or the partitions of sciences*. Cópia de Sir Geofrey Keynes, 1640, cortesia da Seção de Livros Raros da Livraria da Universidade de Cambridge, Cambridge, Reino Unido.

[15] WAGNER, N. R., "Fingerprinting". Em: *Proceedings of the 1983 IEEE Symposium on Security and Privacy (SP 83)*, pp. 18–22, IEEE Computer Society, Washington, DC, EUA, Abril de 1983.

[16] HERODOTUS, *The Histories*. Ed. Penguin Classics, Abril de 2003.

[17] HAYHURST, J. D., *The pigeon post into Paris 1870 - 1871*. Ed. J. D. Hayhurst, 1970.

[18] BRITISH STANDARD, BSI, *Information technology - generic coding of moving pictures and associated audio information - Part 3: audio*, Implementação da ISO/IEC 13818-3:1995, Londres, Reino Unido, Outubro de 1995.

[19] SIMMONS, G. J., "The prisoner's problem and the subliminal channel". Em: *Advances in Cryptology: Proceedings of the CRYPTO '83, Plenum Press*, pp. 51–67, 1984.

[20] SHANNON, C. E., "Communication Theory of Secrecy Systems", *Bell System Technical Journal*, v. 28, 1949.

[21] CACHIN, C., "An information-theoretic model for steganography". Em: *Information and Computation*, v. 192, no. 1, pp. 41–56, Ed. Academic, Duluth, MN, EUA, Julho de 2004.

[22] FRIDRICH, J., GOLJAN, M., HOGEA, D., et al., "Quantitative steganalysis of digital images: estimating the secret message length", *Multimedia Systems Journal, Special issue on Multimedia Security*, v. 9, no. 3, pp. 288–302, Setembro de 2003.

[23] COX, I. J., MILLER, M. L., BLOOM, J. A., et al., *Digital Watermarking and Steganography*. Ed. Morgan Kaufmann, MA, EUA, Novembro de 2007.

[24] COVER, T. M., THOMAS, J. A., *Elements of Information Theory*. Ed. Wiley-Interscience, NY, EUA, Agosto de 1991.

[25] JAIN, A. K., *Fundamentals of digital image processing*. Ed. Prentice-Hall, NJ, EUA, Outubro de 1989.

[26] SILVA, E. A. B., *Disciplina COE784 - processamento digital de imagens*, Notas de aula, COPPE - Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Junho a Setembro de 2008.

[27] CHEN, B., WORNELL, G. W., "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding", *IEEE Transactions on Information Theory*, v. 47, no. 4, pp. 1423–1443, Maio de 2001.

[28] MESE, M., VAIDYANATHAN, P. P., "Optimal histogram modification with MSE metric". Em: *Proceedings of the IEEE International Conference on*

Acoustics, Speech, and Signal Processing (ICASSP), v. 3, pp. 1665–1668, Salt Lake City, UT, EUA, Maio de 2001.

[29] TIFF REVISION 6.0, http://partners.adobe.com/public/developer/en/tiff/tiff6.pdf, Junho de 1992.

[30] WASSERMAN, P. D., *Neural computing: theory and practice*. Ed. Van Nostrand Reinhold, NY, EUA, Junho de 1989.

[31] CARVALHO, Y. U., PETRAGLIA, M. R., *Algoritmo de extensão em frequência baseado em redes neurais e filtragem ótima*, Dissertação de mestrado, COPPE - Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Junho de 2006.

[32] GOMES, J. G. R. C., *Fundamentos de redes neurais*, Minicurso, Congresso Brasileiro de Redes Neurais (CBRN), SC, Brasil, Outubro de 2007.

Apêndice A

Mínimo Erro Médio Quadrático (MMSE) entre Vetores com Componentes em Ordem Crescente

Como visto na Seção 3.3.4, deseja-se que o mapeamento do conjunto de dados C para $\widehat{C}$, de forma que a distribuição de C fique igual à distribuição de $\widehat{C}$, possua a mínima distorção possível. A medida de distorção considerada é o erro médio quadrático (MSE) entre os histogramas C e $\widehat{C}$.

Em [1], Tzschoppe et al. dizem que o mapeamento $C \rightarrow \widehat{C}$ com MSE mínimo deve preservar as relações de ordem crescente entre diferentes elementos. Eles declaram isso mais precisamente da seguinte forma.

Sejam os vetores $\mathbf{x}$ e $\mathbf{y}$ de tamanho L_x com seus elementos distribuídos em ordem crescente ($x_1 \leq x_2 \leq \ldots \leq x_{L_x}$ e $y_1 \leq y_2 \leq \ldots \leq y_{L_x}$). A distorção $D = \frac{1}{L_x}\sum_{n=1}^{L_x}(x_n - y_n)^2$ nunca é maior do que a distorção $D_\pi = \frac{1}{L_x}\sum_{n=1}^{L_x}(x_n - y_{\pi(n)})^2$, onde $\pi(n)$ denota uma permutação arbitrária dos índices dos elementos. Segue a demonstração por indução matemática:

1) Primeiramente, mostra-se a validade para $L_x = 2$. Só há uma permutação

possível, com $\pi(1) = 2$ e $\pi(2) = 1$. Para esta permutação, tem-se:

$$\begin{aligned} D_\pi - D &= \tfrac{1}{2}\left((x_1 - y_2)^2 + (x_2 - y_1)^2 - (x_1 - y_2)^2 - (x_2 - y_2)^2\right) \\ D_\pi - D &= -x_1 y_2 - x_2 y_1 + x_1 y_1 + x_2 y_2 \\ D_\pi - D &= (x_2 - x_1)(y_2 - y_1) \geq 0, \end{aligned} \tag{A.1}$$

implicando em $D_\pi \geq D$.

2) Agora, mostra-se a validade para $L_x = N + 1$ supondo validade para $L_x = N$. Assuma validade para o subvetor $\mathbf{x}^{(N)} = [x_1, \ldots, x_N]^T$, de tamanho N, de forma que qualquer subvetor $\mathbf{y}^{(N)}$, de tamanho N, que origine distorção mínima $D^N = \frac{1}{N}||\mathbf{x}^{(N)} - \mathbf{y}^{(N)}||^2$ esteja distribuído em ordem crescente. Assim, restam somente N diferentes permutações do vetor $\mathbf{y}^{(N+1)}$ que, basicamente, movem o termo y_k, com $k \in 1, 2, \ldots, N$, para o fim do vetor permutado $\mathbf{y}_\pi^{(N+1)}$. A distorção resultante D_π é

$$D_\pi = \frac{1}{L_x}\left(\sum_{n=1}^{k-1}(x_n - y_n)^2 + \sum_{n=1}^{k-1}(x_n - y_{n+1})^2 + (x_{N+1} - y_k)^2\right). \tag{A.2}$$

Logo,

$$\begin{aligned} D_\pi - D &= -\tfrac{2}{L_x}\left(\textstyle\sum_{n=1}^{k-1} x_n y_n + \sum_{n=k}^{N} x_n y_{n+1} + x_{N+1} y_k - \sum_{n=1}^{N+1} x_n y_n\right) \\ D_\pi - D &= -\tfrac{2}{L_x}\left(\textstyle\sum_{n=k}^{N} x_n (y_{n+1} - y_n) + x_{N+1}(y_k - y_{N+1})\right) \\ D_\pi - D &\geq -\tfrac{2}{L_x}\left(x_{N+1}\textstyle\sum_{n=k}^{N}(y_{n+1} - y_n) + x_{N+1}(y_k - y_{N+1})\right) \\ D_\pi - D &\geq -\tfrac{2}{L_x} x_{N+1}(y_{N+1} - y_k + y_k - y_{N+1}) \\ D_\pi - D &\geq 0. \end{aligned} \tag{A.3}$$

Apêndice B

Redes Neurais

Este apêndice apresenta um resumo sobre redes neurais artificiais e os parâmetros utilizados para a implementação dos sistemas esteganalistas utilizados neste trabalho.

Redes neurais artificiais são estruturas computacionais destinadas a processar dados inspiradas no funcionamento cerebral. Elas são formadas por camadas de neurônios artificiais que, assim como os conjuntos de neurônios reais, aos serem treinados, possuem capacidade de processamento e aprendizado. A idéia dos neurônios artificiais é apresentada na Seção B.1. A Seção B.3 explica como é feito o treinamento de redes neurais do tipo *feed-forward* de múltiplas camadas com o algoritmo *error backpropagation*.

Na prática, as redes neurais artificiais não emulam o funcionamento cerebral. Elas são utilizadas para implementar mecanismos mais simples. O projeto de redes neurais artificiais depende do objetivo do processamento. Suas aplicações incluem reconhecimento de padrões, aproximações de funções, sistemas de classificação não-lineares etc.

B.1 Neurônios Artificiais

Os neurônios são células cerebrais que possuem capacidade de receber, processar e transmitir sinais eletroquímicos [31]. Uma ilustração de neurônios pode ser vista na Figura B.1. Os axônios são responsáveis por enviar sinais, enquanto que os dendritos os recebem. O ponto de conexão entre um axônio e um dendrito é chamado sinapse.

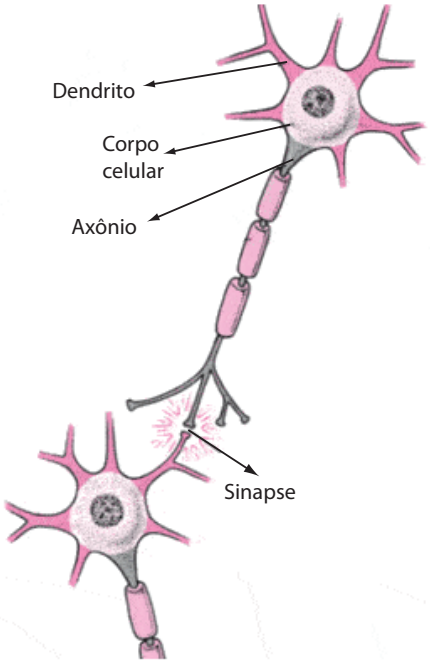


Figura B.1: Neurônios

Resumidamente, os sinais recebidos por um neurônio são somados. Alguns sinais tendem a estimular a célula e outros tendem a inibí-la. Quando a acumulação resultante dessas excitações atingem um limite, o neurônio dispara um sinal através do axônio.

O neurônio artificial é uma estrutura elementar inspirada no neurônio cerebral. Ele visa simular a característica de primeira ordem do neurônio biológico [30]. A Figura B.2 ilustra o neurônio artificial.

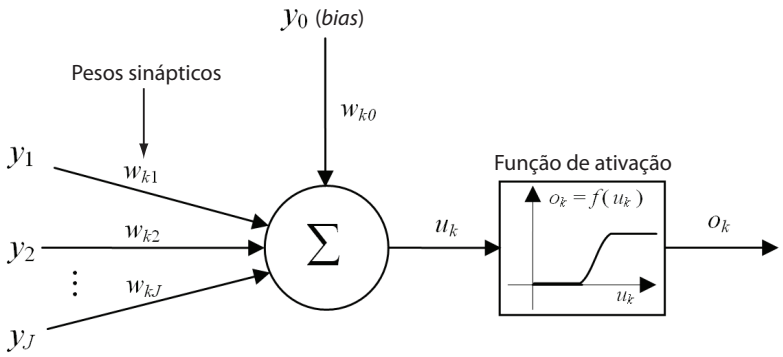


Figura B.2: Neurônio artificial

As entradas de um neurônio artificial podem ser tanto os dados de entrada da

rede neural, quanto as saídas de outros neurônios. Cada uma dessas entradas é multiplicada por um peso sináptico w_{kj}. Os resultados dessas multiplicações são somados. Por fim, aplica-se uma função f, chamada de função de ativação, no resultado das operações anteriores, para gerar a saída do neurônio artificial. A Equação (B.1) apresenta um modelo matemático para o neurônio artificial.

$$o_k = f(u_k) = f\left(\sum_{j=0}^{J} w_{kj} y_j\right) \tag{B.1}$$

Um neurônio pode ser classificado como linear ou não-linear, dependendo da função de ativação. Para se realizar o treinamento da rede neural através do algoritmo *error backpropagation*, faz-se necessário que a função de ativação seja diferenciável. Geralmente, a função de ativação utilizada é a tangente hiperbólica, que é não-linear.

A função tangente hiperbólica, apresentada na Equação (B.2) e mostrada na Figura B.3, é monotônica e apresenta um comportamento aproximadamente linear para valores intermediários. Caso esses valores cresçam ou decresçam, as saídas desta função se aproximam, assintoticamente, de um limite superior ou inferior, respectivamente.

$$tanh(\alpha) = \frac{1 - \exp^{-2\alpha}}{1 + \exp^{-2\alpha}} \tag{B.2}$$

Algumas características dos neurônios biológicos, como atrasos que afetam a dinâmica do sistema, efeitos de sincronismo e modulação em frequência, não são levadas em consideração pela modelagem de neurônio artificial apresentada. Contudo, essa modelagem de neurônio artificial apresenta bons resultados na prática.

A próxima seção trata de conjuntos de neurônios artificiais dipostos em camadas.

B.2 Organização em Camadas

As redes neurais artificiais *feed-forward* são compostas por uma ou mais camadas de neurônios e não possuem realimentação. Cada camada é composta pelos pesos sinápticos e pelos neurônios subsequentes, que possuem *bias* e função de ativação. A Figura B.4 mostra uma rede neural *feed-forward* de duas camadas. A primeira camada é chamada de camada intermediária ou escondida. Ela processa os dados de

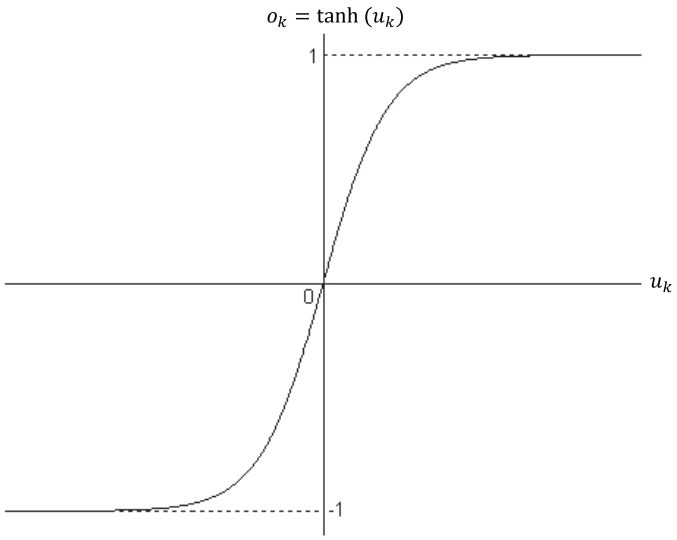


Figura B.3: Função tangente hiperbólica

entrada da rede neural. A segunda camada é chamada camada de saída, que gera as saídas da rede neural.

As redes neurais são ditas totalmente conectadas se todos os neurônios de cada camada possuírem ligações sinápticas com as saídas de todos os neurônios da camada anterior. No caso da primeira camada, todos os neurônios devem possuir conexões sinápticas com todos os elementos do vetor de entrada da rede neural.

B.3 Treinamento da Rede Neural

O treinamento da rede neural serve para atualizar seus pesos sinápticos. Assim, a rede neural aprende a realizar o processamento ao qual ela foi projetada.

O treinamento de uma rede neural pode ser supervisionado ou não-supervisionado. O treinamento supervisionado requer que se apresente para a rede neural os dados de entrada e as suas respostas esperadas (pares entrada-saída) após o processamento, para que o processo de aprendizagem ocorra. No treinamento não-supervisionado, não se faz necessária a apresentação para a rede neural da saída desejada. Neste trabalho, utilizou-se o treinamento supervisionado.

Considerando-se o treinamento supervisionado, sejam os pares entrada-saída definidos pelos vetores $\mathbf{x}_p$ e $\mathbf{d}_p$, respectivamente, apresentados nas Equações (B.3)

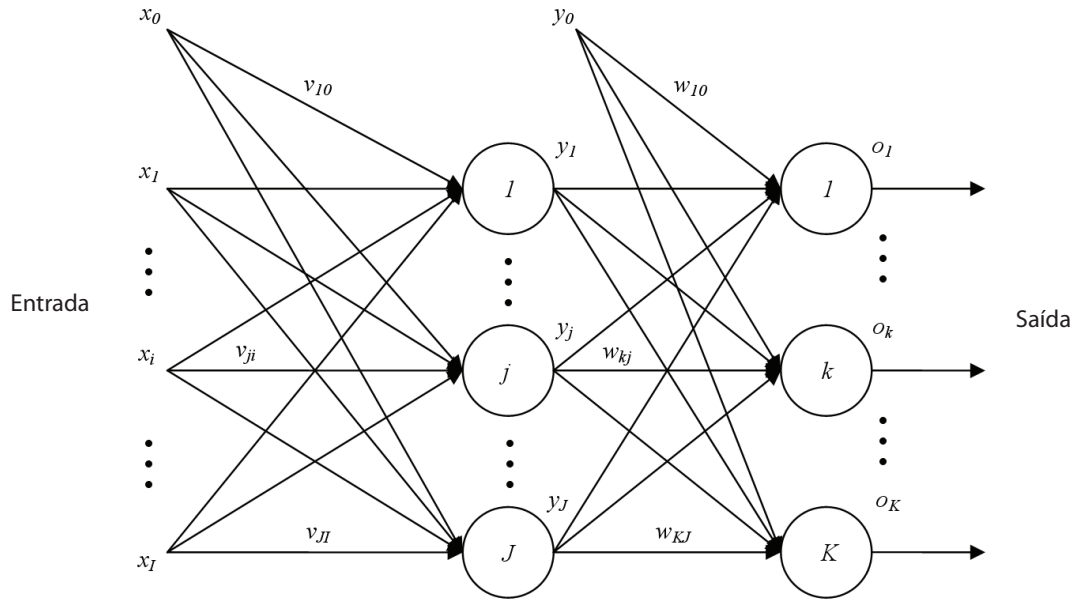


Figura B.4: Rede neural *feed-forward* de duas camadas

e (B.4). O subíndice p representa o número do par entrada-saída.

$$\mathbf{x}_p = [x_1, \ldots, x_i, \ldots, x_I]_p^T \tag{B.3}$$

$$\mathbf{d}_p = [d_1, \ldots, d_k, \ldots, d_K]_p^T \tag{B.4}$$

O objetivo do treinamento da rede é fazer com que o conjunto de saídas da rede, $\mathbf{o}_p = [o_1, \ldots, o_k, \ldots, o_K]_p^T$, convirja para o conjunto de saídas desejadas $\mathbf{d}_p$. Cada par entrada-saída é apresentado à rede neural, que gera um sinal de saída. Comparando-se a saída obtida com a saída desejada, obtém-se um sinal de erro. O algoritmo de treinamento atualiza os pesos sinápticos em função deste sinal de erro, de forma a tentar minimizá-lo. Este procedimento é realizado em todos os pares entrada-saída, para se atingir um erro aceitável.

Os pesos sinápticos influenciam diretamente no desempenho da rede neural. Tal desempenho é avaliado matematicamente em termos de uma função custo. Um exemplo de função custo é o erro médio quadrático (MSE) entre a saída obtida e a saída desejada.

Geralmente, a função custo apresenta uma relação não-linear com os pesos

sinápticos da rede neural. Assim, a superfície de erro pode não ser convexa, com a existência garantida de um valor mínimo. Logo, a busca pelo melhor conjunto de pesos sinápticos, que implica no valor mínimo da função custo, é um problema de otimização não-linear.

Um dos métodos mais conhecidos e utilizados para a resolução de problemas de otimização não-linear é o do gradiente descendente. Este método não apresenta garantias de encontrar o mínimo global de uma função, porém, normalmente, apresenta bons resultados para encontrar mínimos locais.

O vetor gradiente de uma função, no ponto em que é calculado, aponta na direção do máximo crescimento desta função. Então, o sentido oposto a ele aponta para o mínimo. O método do gradiente descendente busca soluções sempre na direção oposta àquela do vetor gradiente calculado num ponto da função custo.

No contexto de redes neurais, o gradiente da função custo F em relação aos pesos sinápticos é utilizado para se buscar o valor mínimo dela. A regra de aprendizagem dos pesos sinápticos dos neurônios da camada de saída da rede neural apresentada na Figura B.4 é dada pela Equação (B.5), onde n é o número da iteração e μ é a taxa de aprendizagem.

$$w_{kj}(n+1) = w_{kj}(n) + \Delta w_{kj}(n) = w_{kj}(n) - \mu \frac{\partial F(n)}{\partial w_{kj}} \tag{B.5}$$

Deve-se observar que uma taxa de aprendizagem μ muito grande dificulta o método do gradiente descendente a encontrar o mínimo da função custo, pois os pesos sinápticos são bastante alterados a cada iteração. O processo pode ficar instável. Uma taxa de aprendizagem muito pequena pode tornar o processo de convergência do algoritmo muito lento. Uma boa escolha de μ depende do problema em questão.

B.3.1 Algoritmo *Error Backpropagation*

O algoritmo *error backpropagation* é um dos métodos mais utilizados para treinamento de redes neurais. Ele é um método de otimização de primeira ordem que utiliza o método do gradiente descendente para atualizar os pesos sinápticos. O algoritmo *error backpropagation* retropropaga o sinal de erro, que é gerado ao se comparar o resultado obtido com o resultado esperado quando os pares entrada-saída são apresentados para a rede neural.

Considerando-se a rede neural da Figura B.4, pode-se escrever a função custo F de acordo com a Equação (B.6), onde K é o número de neurônios da camada de saída.

$$F = \sum_{k=1}^{K}(d_k - o_k)^2 = \sum_{k=1}^{K}(e_k)^2 \tag{B.6}$$

Há duas formas de atualizar os pesos sinápticos da rede neural. A forma de atualização dos pesos sinápticos da rede neural que atua a cada iteração, ou seja, à medida que cada par de dados entrada-saída é apresentado à rede, é chamada de Regra Delta[1]. A outra forma de atualização dos pesos sinápticos é chamada de Batelada. Nela, a atualização ocorre após o cálculo do gradiente com base em diversos pares entrada-saída.

A Equação (B.5) apresenta a lei de aprendizagem, ou de atualização, dos pesos sinápticos da camada de saída da rede. Além disso, sabe-se que cada componente do gradiente da função custo pode ser escrita de acordo com a Equação (B.7).

$$\nabla_{kj} F = \frac{\partial F}{\partial w_{kj}} \tag{B.7}$$

Aplicando-se a regra da cadeia na Equação (B.7), obtém-se a Equação (B.8).

$$\frac{\partial F}{\partial w_{kj}} = \frac{\partial F}{\partial e_k}\frac{\partial e_k}{\partial o_k}\frac{\partial o_k}{\partial u_k}\frac{\partial u_k}{\partial w_{kj}} \tag{B.8}$$

Além disto, tem-se:

$$\frac{\partial F}{\partial e_k} = 2e_k, \tag{B.9}$$

$$\frac{\partial e_k}{\partial o_k} = -1, \tag{B.10}$$

$$\frac{\partial o_k}{\partial u_{kj}} = f^{'}(u_k), \tag{B.11}$$

$$\frac{\partial u_k}{\partial w_{kj}} = y_j. \tag{B.12}$$

[1]A Regra Delta foi criada por Widrow e Hoff. Ela também é conhecida como método *Least Mean Square* (LMS).

Então, aplicando-se as Equações (B.9) a (B.12) na Equação (B.8), tem-se:

$$\frac{\partial F}{\partial w_{kj}} = -2e_k f^{'}(u_k) y_j. \tag{B.13}$$

Finalmente, substituindo-se a Equação (B.13) na Equação (B.5), tem-se a Equação (B.14), que representa a lei de aprendizagem dos neurônios da camada de saída da rede neural apresentada na Figura B.4.

$$w_{kj}(n+1) = w_{kj}(n) + 2\mu e_k f^{'}(u_k) y_j \tag{B.14}$$

A dedução da equação de aprendizagem da camada intermediária da rede neural da Figura B.4 é análoga à dedução da Equação (B.14). Sejam t_j e y_j, $1 \leq j \leq J$, definidos conforme as Equações (B.15) e (B.16), respectivamente. O termo correspondente à atualização dos pesos sinápticos dos neurônios da camada intermediária é apresentado na Equação (B.17).

$$t_j = \sum_{i=0}^{I} v_{ji} x_i \tag{B.15}$$

$$y_j = f(t_j) \tag{B.16}$$

$$\Delta v_{ji} = -\mu \frac{\partial F}{\partial v_{ji}} = \mu \frac{\partial F}{\partial y_j} \frac{\partial y_j}{\partial t_j} \frac{\partial t_j}{\partial v_{ji}} \tag{B.17}$$

Além disto, tem-se:

$$\frac{\partial t_j}{\partial v_{ji}} = x_i \tag{B.18}$$

$$\frac{\partial y_j}{\partial t_j} = f^{'}(t_j) \tag{B.19}$$

$$\begin{array}{l} \frac{\partial F}{\partial y_j} = \frac{\partial}{\partial y_j} \left\{ \sum_{k=1}^{K} [d_k - f(u_k(\mathbf{y}))]^2 \right\}, \text{ com } \mathbf{y} = [y_1, \ldots, y_j, \ldots, y_j]^T \\ \frac{\partial F}{\partial y_j} = -2 \sum_{k=1}^{K} [d_k - o_k] \frac{\partial}{\partial y_j} \{ f(u_k(\mathbf{y})) \} \end{array} \tag{B.20}$$

Utilizando-se a regra da cadeia na derivada parcial da Equação (B.20), tem-se:

$$\frac{\partial}{\partial y_j} f(u_k(\mathbf{y})) = \frac{\partial}{\partial u_k} f(u_k(\mathbf{y})) \frac{\partial u_k}{\partial y_j} = f^{'}(u_k) w_{kj} \tag{B.21}$$

Substituindo-se a Equação (B.21) na Equação (B.20), tem-se:

$$\frac{\partial F}{\partial y_j} = -2 \sum_{k=1}^{K} e_k f^{'}(u_k) w_{kj} \tag{B.22}$$

Finalmente, aplicando-se as Equações (B.18), (B.19) e (B.22) na Equação (B.17), pode-se escrever a equação de atualização dos pesos sinápticos dos neurônios da camada intermediária da rede neural apresentada na Figura B.4, conforme a Equação (B.23).

$$v_{ji}(n+1) = v_{ji}(n) + 2\mu x_i f^{'}(u_j) \sum_{k=1}^{K} e_k f^{'}(u_k) w_{kj} \tag{B.23}$$

Considerando-se a utilização da função tangente hiperbólica como função de ativação dos neurônios, a sua derivada é dada pela Equação (B.24).

$$\begin{aligned} f^{'}(u) &= 1 - o^2, \\ f^{'}(t) &= 1 - y^2 \end{aligned} \tag{B.24}$$

B.4 Constante de Momento

A constante de momento η, $0 \leq \eta \leq 1$, serve para repetir parcialmente os passos de treinamento executados em iterações anteriores, de modo que o processo de treinamento fique menos susceptível ao aprisionamento em mínimos locais da função custo. Isto é feito substituindo-se, respectivamente, Δw_{kj} e Δv_{ji} das equações de aprendizagem por

$$\Delta w_{kj}(n) = \eta \Delta w_{kj}(n-1) - \mu \frac{\partial F(n)}{\partial w_{kj}} \tag{B.25}$$

$$\Delta v_{ji}(n) = \eta \Delta v_{ji}(n-1) - \mu \frac{\partial F(n)}{\partial v_{ji}}. \tag{B.26}$$

As novas equações de aprendizagem são chamadas de Regra Delta Generalizada [32]. A inclusão da constante de momento representa uma pequena modificação na atualização dos pesos sinápticos da rede neural. Isto tende a trazer efeitos benéficos no comportamento do algoritmo de aprendizagem.

B.5 Sistemas Esteganalistas

Os dois sistemas esteganalistas, implementados neste trabalho, foram modelados através de redes neurais *feed-forward* treinadas com o algoritmo *error backpropagation*. A topologia e os parâmetros de treinamento das duas redes são iguais. A diferença entre elas é que conjuntos de treinamento diferentes foram apresentados a cada uma das redes, como descrito no Capítulo 4. As redes neurais foram treinadas no modo batelada.

A Tabela B.1 apresenta os parâmetros de treinamento utilizados. Apesar de os histogramas terem 301 componentes, a rede neural tem apenas três entradas. Para obter estas três entradas, a média dos dados foi feita igual a zero e as três direções principais foram calculadas. Estas três componentes foram normalizadas para terem variância igual a $1,0$. L representa o número de camadas da rede, J representa o número de neurônios da camada intermediária, K representa o número de neurônios da camada de saída, Δ representa o critério de parada (valor de F aceitável), E representa o número de iterações por época, α representa o fator de decaimento da taxa de aprendizagem, η representa a constante de momento e μ representa a taxa de aprendizagem inicial.

Tabela B.1: Parâmetros de treinamento das duas redes neurais

Parâmetro	Valor
Número de Entradas	3
L	2
J	20
K	1
Δ	1e-9
E	800[1]
α	0,9999
η	0,95
μ	0,001

[1]Todos os dados de entrada disponíveis no conjunto de treino.