2
Preliminares

2.1
Toolkit

Um *toolkit* de software é um conjunto de algoritmos desenvolvido para um tipo específico de trabalho, mantido de forma unitária e com capacidade de ser expandido de acordo com necessidades específicas, se fornecido com seu código fonte. Atualmente, existem toolkits para praticamente tudo relacionado à computação, desde aqueles específicos para operações complexas no campo da matemática até aqueles desenvolvidos especialmente para criar componentes gráficos em uma aplicação.

O objetivo dos toolkits é facilitar o desenvolvimento de software através do fornecimento de uma série de funções e serviços prontos. Os toolkits são normalmente implementados na forma de biblioteca de rotinas ou plataforma para aplicações. No caso das bibliotecas de rotinas, suas funções e serviços são compilados em separado e linkados ao software que a utiliza. Se implementado como plataforma, seus módulos ou pacotes passam a fazer parte do software em desenvolvimento.

Muitos toolkits de código aberto disponíveis hoje em dia são extensíveis, o que significa que podem receber novas funcionalidades, aumentando assim a gama de serviços oferecidos.

Dentre alguns toolkits muito populares podemos citar o AWT – Abstract Window Toolkit, implementado como uma biblioteca de funções com o objetivo de prover serviços de criação de janelas, botões e outros recursos gráficos para softwares desenvolvidos em Java.

PUC-Rio - Certificação Digital Nº 0511033/CA

2.2
Tecnologias de sensores

Com o avanço tecnológico, os sensores estão evoluindo rapidamente e sendo aplicados a um número cada vez maior de situações. A queda nos preços aliada a maior autonomia e confiabilidade tem contribuído sensivelmente para a popularização dos sensores. Existem atualmente diversos tipos, dentre eles os sensores ópticos, de pressão diferencial, orgânicos, eletro-eletrônicos e de vibração.

Independente da área de estudo à qual as tecnologias de sensores estão sendo empregadas, a maneira mais popular de aplicação leva em conta o posicionamento estratégico de sensores em diversos pontos onde ocorrem os fenômenos físicos a serem monitorados, formando o que conhecemos como "redes de sensores". As redes de sensores são formadas a partir de pequenos dispositivos que se comunicam através de um meio sem fio. Estes dispositivos são capazes de medir, processar e transmitir os dados monitorados sendo que, a capacidade de processamento efetiva nas redes de sensores é obtida pelo trabalho em conjunto de cada dispositivo. Cada um deles processa os dados coletados localmente para extrair as características de interesse, além de armazenar a informação momentaneamente, fazer o roteamento e usar a rede sem fio para transmitir a informação para os seus vizinhos. Estas redes possuem características bem particulares que incluem a limitação dos recursos computacionais, requerimento de vida longa, acoplamento com o mundo físico e condições ambientais dinâmicas (e muitas vezes hostis).

Os sistemas que rodam nestes sensores devem ser os mais otimizados possíveis em relação ao consumo de bateria, visto que em algumas situações a substituição de um dispositivo pode ter um custo extremamente elevado, como é o caso de sensores de pressão de fundo em poços de petróleo.

Dentro dos poços, as empresas de petróleo normalmente optam por sensores eletro-eletrônicos com fios (*wired sensors*), pois podem ser energizados continuamente. Em caso de falha, o procedimento de troca de um destes sensores demanda uma intervenção do poço com parada de produção, o que ocasiona custos enormes. Em alguns casos estes custos são tão altos que os sensores defeituosos são simplesmente abandonados. Estima-se que na indústria de petróleo apenas um

PUC-Rio - Certificação Digital Nº 0511033/CA

percentual variando de 60% a 90% dos sensores instalados esteja gerando dados até o receptor final.

Embora a confiabilidade dos sensores esteja aumentando gradativamente, ainda é impensável considerá-los seguros, principalmente no que tange a comunicação via rádio. Em levantamento recente, a Petrobras S.A. concluiu que, em aproximadamente 80% dos casos onde houve erros em dados adquiridos por sensores, eles foram causados por algum tipo de falha de comunicação.

Desta forma, ao tratarmos de dados adquiridos por sensores devemos levar em consideração o ambiente muitas vezes hostil e a complexidade das transmissões via rádio, que demandam sincronismo extremo. Estas características tornam tais dados extremamente sujeitos à falhas e distorções.

Quando um sensor falha, normalmente ele pára de funcionar ou funciona de forma totalmente inconsistente. Em caso de falhas temporárias, as séries históricas passam a conter buracos. No entanto, quando distorções são percebidas no dado gerado pelo sensor, comumente elas aparecem durante pequenas falhas na transmissão ou por alterações no ambiente de transmissão. Estas distorções são conhecidas como *ruídos*.

Os ruídos podem ser definidos como flutuações causadas por fatores externos em um sinal durante a transmissão. Estas flutuações indesejadas podem ser tanto interferências aleatórias como interferências que seguem um padrão específico (*patterned interference*). Apenas a última pode ser removida com sucesso em sistemas analógicos. Já os sistemas digitais são normalmente construídos de forma que seu sinal quantizado pode ser perfeitamente reconstruído, desde que os ruídos não ultrapassem determinado patamar pré-estabelecido.

Embora possa parecer impensável o uso dos sensores analógicos nos dias atuais, eles ainda são uma realidade em muitos segmentos da indústria, incluindo o segmento de gás e energia, onde ambientes de difícil acesso demandam custos altíssimos de substituição. Logo, dados contaminados por ruídos fazem parte da realidade destas indústrias.

Claude Shannon contribuiu enormemente no tratamento de ruído durante a transmissão de sinais. Shannon conseguiu definir um modelo matemático para calcular a capacidade de um determinado canal de se comunicar sem ruído. Po-

PUC-Rio - Certificação Digital Nº 0511033/CA

rém, este assunto não está no escopo deste trabalho, focaremos nos dados que chegam ao receptor e em como lidar com estes dados.

2.3 Processamento de sinais

2.3.1 Introdução

A natureza física dos sinais e dos sistemas utilizados nos mais diversos segmentos da indústria pode ser drasticamente diferente. No entanto, possuem duas características muito básicas em comum. Sinais são funções de uma ou mais variáveis independentes e que, tipicamente, contém informações a respeito do comportamento ou natureza de algum fenômeno, enquanto os sistemas respondem a sinais particulares produzindo outros sinais. Voltagem e corrente em função do tempo, tratando-se de circuitos elétricos, são exemplos de sinais, enquanto um circuito propriamente dito é um exemplo de sistema. Um programa de computador para o diagnóstico automatizado de eletrocardiogramas pode ser considerado um sistema que tem como entrada os dados digitalizados do eletrocardiograma e produz parâmetros como a taxa de batimentos cardíacos. Uma câmera fotográfica é um sistema que recebe luz de diferentes fontes e produz uma fotografia.

Nos mais diferentes contextos há uma variedade de problemas e questões a serem consideradas. Algumas vezes somos apresentados a um sistema específico e devemos analisar seu comportamento em relação a uma série de entradas diferentes. Uma aplicação deste tipo pode estudar, por exemplo, a percepção dos sentidos humanos.

Em outros casos, ao invés de analisarmos sistemas existentes, teremos que desenvolver novos com a finalidade de processar sinais específicos. Por exemplo, o desenvolvimento de um sistema para tentar prever o comportamento de uma determinada ação na bolsa de valores com base em seu histórico.

Uma outra aplicação muito popular é a restauração de sinais que tenham sido degradados de alguma forma. Em sistemas de telefonia, por exemplo, poderíamos remover os ruídos do ambiente, fazendo com que a voz do transmissor chegasse limpa ao receptor.

PUC-Rio - Certificação Digital Nº 0511033/CA

Os exemplos acima são apenas algumas das milhares de aplicações para os conceitos de sinais e sistemas. Em alguns dos exemplos os sinais variam continuamente no tempo e, em outros, sua evolução é descrita em alguns pontos discretos no tempo. Por exemplo, uma música é um sinal sonoro contínuo no tempo. Já no caso dos sensores, como eles são configurados para adquirir os dados em intervalos de tempo específicos, o sinal gerado não é contínuo e sim discreto. O melhor gráfico representativo para um sinal sonoro é uma curva contínua, enquanto um gráfico de um sinal adquirido por sensor pode facilmente ser representado como um conjunto de números associados a instantes de tempo discretos.

Embora relacionadas de alguma forma, a análise de sinais possui duas vertentes: uma relacionada aos eventos e fenômenos que ocorrem em **tempo discreto** e outra relacionada aos que ocorrem em **tempo contínuo**. Embora conceitualmente estejam estritamente relacionadas, no passado suas aplicações eram tratadas de maneiras bem distintas e, por essa forma, o desenvolvimento das duas vertentes se deu de forma bastante paralela. Porém, nas últimas décadas, devido ao avanço dramático nas tecnologias de circuitos e telecomunicações, estas duas vertentes têm se aproximado cada vez mais. Muitas vezes os sinais de tempo contínuo são analisados por métodos discretos através de amostragem.

Devido à natureza discreta dos sinais adquiridos por sensores, iremos focar exclusivamente na análise de sinais em tempo discreto.

2.3.2 Ferramentas para análise de sinais

Sistemas de análise de sinais basicamente transformam um sinal original em um outro sinal de acordo com necessidades específicas. Por este motivo encontraremos na literatura uma série de ferramentas matemáticas desenvolvidas para transformar sinais, dentre elas a **transformada de Fourier** e a **transformada Wavelet**.

Sem dúvidas a transformada de Fourier é a técnica mais popular no campo de processamento de sinais. No entanto, nem sempre é a mais indicada para determinados casos. Por exemplo, dado um sinal sonoro ou uma imagem, a análise de Fourier pode facilmente calcular as freqüências e amplitudes das componentes do sinal, fornecendo uma visão geral de suas características. No entanto, mesmo

PUC-Rio - Certificação Digital Nº 0511033/CA

sendo possível aplicar a inversa de Fourier em algumas circunstâncias, seus métodos nem sempre são boas alternativas para recapturar o sinal, principalmente se o mesmo for pouco suavizado. Ele necessita de muita informação para reconstruir um determinado trecho do sinal. Nestes casos, a análise Wavelet é mais eficiente, pois fornece uma abordagem simples para tratar aspectos localizados de um sinal. A análise Wavelet também nos fornece novos métodos de remoção de ruídos e compressão.

Devido ao fato de que sensores normalmente medem propriedades físico-químicas do mundo real, os sinais gerados por eles, na maioria das vezes, não são muito suaves. Como o objetivo deste trabalho está associado à compressão e tratamento básico de sinais, a análise Wavelet foi selecionada, e por este motivo, não iremos nos aprofundar em teoremas matemáticos que demonstram o funcionamento da análise de Fourier [4].

2.3.3 Análise Wavelet

As Wavelets são ferramentas matemáticas para decomposição hierárquica de funções. Elas permitem que uma função seja descrita em termos de seu comportamento padrão, adicionando detalhes em diferentes níveis. Desta maneira, estas transformadas oferecem uma forma elegante de representar os níveis de detalhe de um sinal, seja ele gerado pela variação de cores de uma imagem, uma superfície 3D ou uma curva gerada por dados adquiridos por sensores. A transformada de Wavelet se baseia na transformada de Fourier. A idéia central da transformada de Wavelet consiste em decompor um sinal em diferentes níveis de resolução, processo conhecido como multi-resolução. A representação em multi-resolução fornece uma moldura hierárquica simples para interpretação de informação da curva. Em diferentes resoluções, os detalhes de uma curva geralmente caracterizam diferentes estruturas físicas da mesma. Se olharmos uma curva através de uma escala grande, estes detalhes geralmente caracterizam as grandes estruturas que fornecem o contexto ou o comportamento geral da curva. Com o aumento da resolução, obtemos detalhes mais finos. O procedimento de análise por Wavelet adota uma função Wavelet base, chamada Wavelet mãe. A análise tem-

PUC-Rio - Certificação Digital Nº 0511033/CA

poral é executada com uma contração de base Wavelet de alta freqüência, enquanto análise de freqüência é executada por uma dilatação da base Wavelet de baixa freqüência. Como o sinal original ou a função podem ser representados em termos de uma expansão de Wavelet (usando coeficientes em uma combinação linear das funções Wavelet), as operações com os dados podem ser realizadas apenas com o uso dos coeficientes de Wavelet correspondentes. Podemos representar a informação de maneira esparsa, se escolhermos a base Wavelet gerando coeficientes próximos de zero. Podemos, também, truncar a informação abaixo de um limiar (*t-hreshold*). Estes são procedimentos relativamente importantes na tarefa de limpar os dados de natureza ruidosa como os adquiridos por sensores.

Um sistema físico natural é quase sempre não-estacionário, o que significa dizer que, no mundo real dos dados coletados por sensores, qualquer segmento temporal que tomarmos para análise irá apresentar momentos estatísticos, média e variâncias diferentes, dificultando enormemente a descoberta de padrões de comportamento dos dados. Com o intuito de resolver este problema, diversos esforços na área da matemática e física têm sido feitos, um desses esforços é a transformada Wavelet. A primeira menção a Wavelet apareceu em um apêndice da tese do matemático alemão Alfred Haar em 1909. Uma das propriedades da Wavelet de Haar é ter suporte compacto, significando que, fora do intervalo da sua definição até infinito, a função tem valor zero. Infelizmente a Wavelet de Haar não é continuamente diferenciável, o que limita seu campo de aplicação.

Nos anos trinta, surgiram vários grupos que trabalhavam independentemente pesquisando a representação de funções através de funções base variando em escala.

No início da década de oitenta, Grossman e Morlet, um físico e outro engenheiro, respectivamente, definiram completamente as Wavelets no contexto da física quântica. Estes dois pesquisadores criaram um modo para Wavelets baseado na intuição física. Em 1985, S. Mallat deu as Wavelets um impulso adicional com o seu trabalho em processamento digital de sinal. Ele descobriu algumas relações entre filtros de quadratura conjugada (QMF- Quadrature Mirror Filters), algoritmos de pirâmide, e base Wavelet ortonormal. Inspirado em parte por estes resultados, Yves Meyer construiu a primeira Wavelet não trivial. Diferente da Wavelet de Haar, as Wavelets de Meyer são continuamente diferenciáveis; porém elas não

PUC-Rio - Certificação Digital Nº 0511033/CA

têm suporte compacto. Anos mais tarde, Ingrid Daubechies inspirada no trabalho de Mallat construiu um conjunto de funções Wavelets com base ortonormal. Essas funções são talvez as mais elegantes, e se tornaram um divisor de águas nas aplicações de Wavelet nos dias de hoje.

Como dissemos anteriormente, as transformadas de Wavelet foram desenvolvidas a partir da transformada de Fourier, na verdade, ela foi desenvolvida como uma alternativa para a STFT (Short Time Fourier Transform), a qual foi à primeira modificação da transformada de Fourier, para permitir a análise de dados não estacionários. A STFT não obteve muito sucesso, pois analisava o sinal inteiro na mesma janela de tempo. A transformada de Wavelet é a única transformação linear que pode analisar sinais não-estacionários a taxas de resolução diferentes.

Uma Wavelet é uma onda com duração limitada cujo valor médio é zero. Assim, como a análise de Fourier consiste na quebra de um sinal em ondas senoidais de várias freqüências, a análise de Wavelet é a quebra de um sinal em versões deslocadas e em diferentes escalas de uma Wavelet original (conhecida como Wavelet mãe).

Aplicando-se a transformada Wavelet é aplicada ao conjunto original de dados, gera-se uma lista de coeficientes é gerada, na proporção de um para um. Esta lista contém coeficientes com diferentes escalas em diferentes seções do sinal. A correspondência entre escala e freqüência se dá da seguinte forma:

Fator de escala pequeno
Wavelet comprimida
Detalhes que mudam rápido
Freqüência alta

Fator de escala grande
Wavelet expandida
Pequenas variações
Freqüência baixa

Tabela 1 : Tabela de correspondência entre escalas e freqüências

O cálculo dos coeficientes Wavelet em todo o conjunto de escalas possível gera uma quantidade de dados desnecessária. A escolha de apenas um subconjunto de escalas torna a análise mais eficiente sem perda de precisão. O resultado desta análise é conhecido como *Wavelet discreta*. Neste tipo de análise, são gerados coeficientes relacionados à aproximação e relacionados a detalhes. As apro-

PUC-Rio - Certificação Digital Nº 0511033/CA

ximações são os componentes de baixa freqüência e representam o comportamento geral do sinal. Já os detalhes são compostos por componentes com fator de escala pequeno, de alta freqüência.

Para demonstrarmos melhor como os coeficientes são gerados, tomemos como exemplo a mais simples das funções base, a wavelet de Haar, mostrada abaixo:

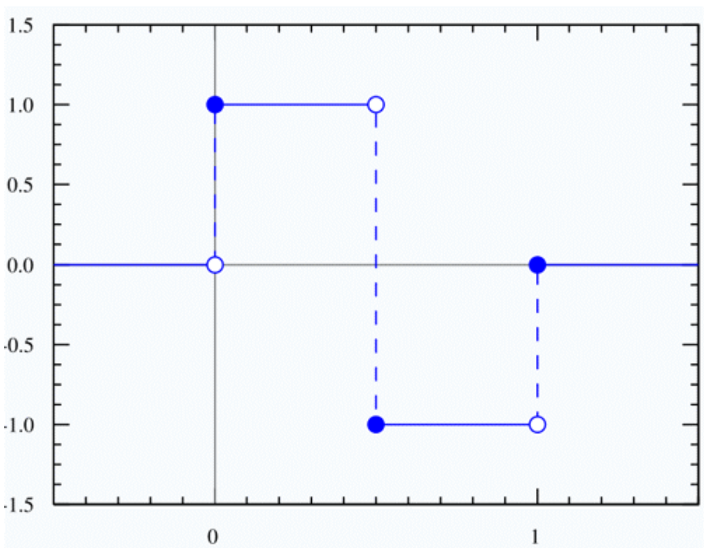

Figura 1: Função base de Haar

Dado a seguinte lista de valores originais:

[9 7 3 5]

Podemos representar este conjunto na base Haar começando por agrupar os valores pela média dos pares:

(9 + 7) / 2 = 8
(3 + 5) / 2 = 4

[8 4]

De maneira clara, podemos perceber que seria impossível restaurar a seqüência inicial de quatro valores sem perda de precisão. Desta forma, devemos

PUC-Rio - Certificação Digital Nº 0511033/CA

armazenar valores que nos ajudem a recuperar os detalhes da série original. Chamaremos estes valores de "coeficientes de detalhe". No nosso exemplo, o coeficiente de detalhe para restaurar os dois primeiros componentes é 1 visto que, 8 se relaciona com 9 e 7 através de soma e subtração de 1:

$8 + 1 = 9$

$8 - 1 = 7$

Já o elemento 4 se relaciona com 3 e 5 através da soma e subtração de -1**:**

$4 + (-1) = 3 – (-1) = 5$

Desta forma, ficaríamos com:

Aproximação	**Detalhe**
[8 4]	**[1 -1]**

Sucessivamente podemos repetir o processo agrupando 8 e 4 pela média do par:

$(8 + 4) / 2 = 6$

Novamente, precisamos de coeficientes de detalhe para chegar ao 8 e 4 a partir da média 6. Neste caso precisamos de apenas um coeficiente de detalhe, o número 2 já que:

$6 + 2 = 8$

$6 – 2 = 4$

Então:

Aproximação	**Detalhe**
[6]	**[2]**

PUC-Rio - Certificação Digital Nº 0511033/CA

Em resumo, teremos os seguintes conjuntos:

Resolução	Aproximação	Detalhes
4	[9 7 3 5]	
2	[8 4]	[1 -1]
1	[6]	[2]

Finalmente, podemos definir a transformada Wavelet (ou decomposição wavelet) dos valores originais como sendo um simples coeficiente representando a média geral dos seus valores seguida por uma série de coeficientes de detalhe que servirão para aumentar gradativamente a resolução. Logo, para a base Haar de uma dimensão a transformada Wavelet é dada por:

[6 2 1 -1]

Podemos perceber que a lista de coeficientes inicia com o valor representativo de toda a curva (6), logo depois vem o coeficiente de detalhe para a resolução mais baixa (2), seguido pelos coeficientes de detalhe da resolução mais alta (1,-1). Devida à pequena quantidade de dados originais no exemplo, talvez não seja muito evidente que, quanto mais para o final um coeficiente de detalhe estiver, maior a resolução que ele ajuda a restaurar, conseqüentemente maior o nível de detalhe associado a ele.

A maneira como calculamos esta transformada Wavelet através de somas, subtrações e médias dos coeficientes é conhecida como *filter bank* e este processo pode ser generalizado para outras funções base além da de Haar. Como podemos perceber, não houve ganho ou perda em termos de precisão ou espaço já que o resultado da transformada contém a mesma quantidade de informação do dado original e através dela podemos restabelecer os valores originais sem nenhuma perda.

PUC-Rio - Certificação Digital Nº 0511033/CA

2.4
Técnicas para compressão de dados

2.4.1 Introdução

Compressão de dados consiste em representar informação de uma maneira compacta. Criamos estas representações compactas identificando e utilizando as estruturas existentes no próprio dado. Os dados podem ser caracteres em um arquivo de texto, números amostrados de uma onda sonora ou séries históricas adquiridas por sensores.

Muitas das técnicas de compressão utilizam a estrutura estatística do conjunto de dados, porém esta não é a única estrutura que pode ser explorada para compactação. Quando os dados a serem compactados são analisados por humanos, podemos nos beneficiar das limitações de percepção do corpo humano. Por exemplo, nós não conseguimos escutar as componentes sonoras de altíssima freqüência que os cães podem ouvir, logo, ao efetuarmos a gravação de uma onda sonora, não se faz necessário manter estas componentes no sinal. Da mesma forma, mesmo que o corpo humano possa perceber a informação, dependendo da natureza da análise, ela pode não contribuir em nada. Por exemplo, quando temos que analisar um conjunto de dados relativamente grande, que forma uma longa série histórica, muitas vezes não estamos prestando atenção em detalhes muito minuciosos, mas estamos bastante interessados no comportamento geral da curva.

Quando falamos de técnicas de compressão, implicitamente estamos falando do desenvolvimento de dois algoritmos: um algoritmo que recebe como entrada o conjunto X original, codificando-o no conjunto X' que necessita de um número menor de bits; e um algoritmo de reconstrução que recebe X' transformando-o no conjunto Y, que pode ou não ser exatamente igual a X. Por convenção, quando nos referirmos a um algoritmo de compressão estaremos falando de ambos.

Existem basicamente dois grandes grupos de técnicas de compressão. Quando há perda de informação durante o processo, temos um esquema de **Compressão com perdas** e quando um conjunto de dados pode ser perfeitamente reconstruído, temos um esquema de **Compressão sem perdas**.

PUC-Rio - Certificação Digital Nº 0511033/CA

2.4.2 Compressão sem perdas

O esquema de compressão sem perdas não envolve perda de informações. Um conjunto de dados codificado pode ser perfeitamente decodificado a partir do conjunto compacto.

A compressão de texto é uma área importante onde à técnica sem perdas é utilizada. Neste caso o requisito básico é que o texto original possa ser preservado visto que, a substituição de palavras ao longo do texto pode torná-lo ilegível ou desprovido de significado.

Outro campo onde compressão sem perdas é bastante utilizado é o campo de imagens médicas. Imagine se uma imagem digitalizada de uma radiografia for comprimida com um esquema que introduza perda de informação. Esta perda pode levar o médico a um diagnóstico completamente equivocado. Neste caso, é um requisito básico que a imagem original seja preservada.

Existem muitos campos onde a compressão sem perdas deve ser utilizada mas, existem muitos outros onde este requisito pode ser relaxado para atingir uma taxa de compressão mais agressiva. Para estas situações existem as técnicas de compressão com perdas.

2.4.3 Compressão com perdas

Neste esquema de compressão as técnicas utilizadas envolvem algum tipo de perda de informação. Uma vez que o conjunto original é transformado no conjunto compactado, a decodificação não conseguirá transformar os dados compactados novamente no conjunto original. Como benefício por aceitarmos esta perda, podemos geralmente obter taxas de compressão muito maiores do que aquelas obtidas pelos esquemas sem perda.

Em muitas aplicações a perda introduzida pelo processo não é um problema. Por exemplo, para armazenar e transmitir voz não é necessário o valor exato de cada amostra do sinal. Dependendo do requisito de qualidade do sistema que usará a informação, podemos variar a quantidade de informação perdida. No caso de uma rede de telefonia, a qualidade da voz de um usuário deve ser suficientemente boa para que ambos possam manter uma conversação de maneira clara. Porém, se

PUC-Rio - Certificação Digital Nº 0511033/CA

a voz fosse de um locutor de história infantil armazenada em um CD-R, a quantidade de perda tolerada seria bem menor.

Tanto as imagens que assistimos nos canais de televisão como àquelas vistas em filmes de DVD possuem algum esquema de compressão com perdas. Sinais de vídeo utilizam uma grande variedade de métodos deste tipo. O fato de a reconstrução ser diferente da original não influencia na visualização do mesmo, desde que a perda de informação não resulte em artefatos incômodos para experiência do visualizador.

As fotografias digitais também são alvo dos algoritmos de compressão com perdas. Um dos esquemas mais populares do mundo é o JPG2000. Ele foi criado no ano 2000 pelo *Joint Photographic Experts Group* com a intenção de substituir o esquema até então utilizado, baseado na transformada de cossenos discretos, criado em 1991, pelo esquema baseado em transformadas Wavelet.

O processo utilizado no padrão JPG2000 inicia-se com uma transformação no espaço de cores de RGB para YCbCr (Irreversível, introduzindo perda) ou YUV (reversível, sem perda). Após esta transformação, os três componentes básicos Vermelho (R), Verde (G) e Azul (B) podem ser tratados individualmente. O passo posterior é dividir toda a imagem em pequenos quadrados de tamanho fixo. A partir deste ponto, uma transformada Wavelet é aplicada em cada um dos quadrados fornecendo uma lista de coeficientes que posteriormente são quantizados de forma escalar com o objetivo de reduzir a quantidade de bits necessários para representá-los. A saída é um conjunto de números inteiros que devem ser codificados bit a bit. A quantidade de perda introduzida neste processo é definida pelo nível da quantização que pode ser selecionado. Como resultado da quantização, temos uma série de grupos de coeficientes que devem agora ser codificados seguindo uma lógica específica que não convém entrarmos em detalhes. Esta codificação utiliza entre outros algoritmos, um conhecido como *codificação aritmética*.

2.4.4 Codificação de Huffman

A codificação de Huffman é um algoritmo de codificação entrópico utilizado para comprimir um conjunto de dados sem que haja perda de informação. Chamamos de codificação o processo de atribuir uma seqüência binária para cada

PUC-Rio - Certificação Digital Nº 0511033/CA

símbolo em um alfabeto. Um alfabeto por sua vez é um conjunto de símbolos (letras). Por exemplo, o alfabeto usado para escrever a maioria dos livros é composto por 26 letras minúsculas, 26 maiúsculas e uma variedade de caracteres de pontuação. Neste contexto, uma vírgula é uma letra. O código ASCII para a letra A é 1000011, portanto dizemos que a letra "**A**" é codificada como 1000011. Já uma vírgula **","** é codificada como 0011010. Podemos notar que a tabela ASCII codifica todas as letras com o mesmo número de bits (8 bits ou 1 byte). Este esquema de codificação é conhecido como *código de comprimento fixo*.

Para reduzirmos a quantidade de bits necessária para representar uma mensagem, devemos usar um número diferente de bits para cada símbolo. Se usarmos um número menor de bits para representar os símbolos que ocorrem em maior freqüência, na média usaremos menos bits por símbolo. Este esquema de codificação é conhecido como *código de comprimento variável.*

A idéia geral por trás da codificação de Huffman é substituir cada símbolo do conjunto original pelo seu código binário correspondente. Vejamos um exemplo. Supondo que o alfabeto seja composto pelas letras A, B, C e D com as seguintes probabilidades P(A) = 1/2 ; P(B) = 1/4 ; P(C) = 1/8 ; P(D) = 1/8 vamos analisar os seguintes códigos para os símbolos:

Letras	Código 1	Código 2
A	0	0
B	1	10
C	00	110
D	11	111
Tamanho médio (bits/simbolo)	**1.25**	**1.75**

Tabela 2: Exemplo de codificação de um alfabeto

Codificar a sequência "BAA" resultaria em:

Usando **Código 1**: 100

Usando **Código 2**: 1000

PUC-Rio - Certificação Digital Nº 0511033/CA

PUC-Rio - Certificação Digital Nº 0511033/CA

Como podemos perceber, usando o código 1 alcançamos uma taxa de compressão maior pois representamos o mesmo conjunto com menos informação. No entanto, vejamos o que acontece quando temos que decodificar o conjunto:

Código 1

100 ► 1: B ; 0:A ; 0:A ► BAA

ou

100► 1:B ; 00:C ► BC

Uma vez a mensagem codificada com o **código 1** ela nunca mais conseguirá ser decodificada com certeza e isto não é uma propriedade desejada para um código. Precisamos que o código seja decodificável de forma única. Vejamos o que acontece quando decodificamos com o **código 2**:

Código 2

1000 ► 10:B ; 0:A ; 0:A ► BAA

Podemos notar que só existe uma possibilidade neste caso, logo, o **código 2** é considerado unicamente decodificável. Além disso, ele possui uma propriedade importante que torna fácil identificar a consideração: nenhuma seqüência binária atribuída aos símbolos no **código 2** é prefixo de nenhuma outra. Códigos deste tipo são conhecidos como *códigos livres de prefixo* e são a chave para codificação de Huffman.

Os códigos livres de prefixo gerados através da técnica desenvolvida por Huffman são conhecidos como *códigos de Huffman*. O procedimento é baseado em duas observações a respeito de códigos livres de prefixo ótimos:

1- Em um código ótimo, símbolos que ocorrem mais freqüentemente (possuindo maior probabilidade de ocorrência) terão seqüências binárias menores do que aqueles que ocorrerem com menor freqüência.

2- Em um código ótimo, os dois símbolos que ocorrem com menos freqüência terão seqüências binárias correspondentes de mesmo comprimento, variando apenas o último bit.

PUC-Rio - Certificação Digital Nº 0511033/CA