**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# REDUÇÃO DE DIMENSIONALIDADE USANDO AGRUPAMENTO E DISCRETIZAÇÃO PONDERADA PARA A RECUPERAÇÃO DE IMAGENS POR CONTEÚDO.

**FRANCISCO ROCHA PIROLLA**

**ORIENTADORA: PROFª. DRª. MARCELA XAVIER RIBEIRO**

São Carlos - SP
Outubro/2012

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# REDUÇÃO DE DIMENSIONALIDADE USANDO AGRUPAMENTO E DISCRETIZAÇÃO PONDERADA PARA A RECUPERAÇÃO DE IMAGENS POR CONTEÚDO.

**FRANCISCO ROCHA PIROLLA**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software.
Orientadora: Dra. Marcela Xavier Ribeiro.

São Carlos - SP
Outubro/2012

**Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar**

P671rd

Pirolla, Francisco Rocha.
Redução de dimensionalidade usando agrupamento e discretização ponderada para a recuperação de imagens por conteúdo / Francisco Rocha Pirolla. -- São Carlos : UFSCar, 2012.
74 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2012.

1. Processamento de imagens. 2. Discretização. 3. Pré-processamento. 4. CBIR. I. Título.

CDD: 006.42 (20ª)

# Universidade Federal de São Carlos

## Centro de Ciências Exatas e de Tecnologia

### Programa de Pós-Graduação em Ciência da Computação

# "Redução de Dimensionalidade usando Agrupamento e Discretização Ponderada para a Recuperação de Imagens por Conteúdo"

Francisco Rocha Pirolla

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Marcela Xavier Ribeiro

Profa. Dra. Marcela Xavier Ribeiro
(Orientadora - DC/UFSCar)

Marilde Santos

Profa. Dra. Marilde Terezinha Prado Santos
(DC/UFSCar)

L.Romani

Profa. Dra. Luciana Alvim Santos Romani
(EMBRAPA)

São Carlos
Novembro/2012

# AGRADECIMENTOS

A Deus, por sempre estar ao meu lado, à minha orientadora Marcela, a todos os meus professores, à minha família, ao CNPQ e à FAPESP pelo apoio financeiro.

# RESUMO

Neste trabalho, propomos diminuir o *gap* semântico e os problemas de maldição de dimensionalidade apresentando duas técnicas de pré-processamento do vetor de características com o objetivo de melhorar a recuperação de imagens baseada em conteúdo e sistemas de classificação de imagens: um método de redução de dimensionalidade do vetor de características original, baseado no algoritmo *k-means,* chamado FTK (*Feature Transformation based on K-means*) e um método de discretização ponderada de características que privilegia as faixas de características mais importantes para distinguir imagens, chamado WFD (*Weighted Feature Discretization*). Os métodos propostos foram utilizados para pré-processar os vetores de características nas abordagens CBIR e classificação, comparando o pré-processamento executado pelo método PCA e os resultados dos vetores de características originais. O algoritmo FTK promoveu uma redução no tamanho do vetor de características com uma melhoria na precisão da consulta e na precisão de classificação. O algoritmo WFD melhorou a precisão da consulta e classificação; a combinação de dos dois algoritmos propostos também melhorou a precisão da consulta e classificação. Estes resultados são muito importantes, especialmente quando comparados com os resultados do método PCA, que também leva a uma redução no tamanho do vetor de características, a um menor aumento na precisão da consulta e a menor aumento na precisão da classificação. Além disso, as técnicas propostas têm custo computacional linear, enquanto o PCA tem um custo computacional cúbico. Os resultados indicam que os métodos propostos são abordagens adequadas para realizar pré-processamento dos vetores de características de imagens em sistemas CBIR e em sistemas de classificação.

**Palavras-chave:** transformação de características, discretização, pré-processamento, vetor de características, agrupamento, CBIR, classificação.

# ABSTRACT

This work proposes two new techniques of feature vector pre-processing to improve CBIR and image classification systems: a method of feature transformation based on the *k-means clustering* approach (*Feature Transformation based on K-means* - FTK) and a method of *Weighted Feature Discretization* - WFD. The FTK method employs the *clustering* principle of *k*-means to compact the feature vector space. The WFD method performs a weighted feature discretization, privileging the most important feature ranges to distinguish images. The proposed methods were employed to pre-process the feature vector in CBIR and in classification approaches, comparing the results with the pre-processing performed by PCA (a well known feature transformation method) and the original feature vector: FTK produced a reduction in the feature vector size with an improving in the query precision and a improvement in the classification accuracy; WFD improved the query precision up to and a improvement in the classification accuracy; the combination of WFD and FTK improved also the query precision and a improvement in the classification accuracy. These are very important results, especially when compared with PCA results, which leads to a minor reduction in the feature vector size, a minor increase in the query precision and a minor increase in the classification accuracy. Also the proposed approaches have linear computational cost where PCA has a cubic computational cost. The results indicate that the proposed approaches are well-suited to perform image feature vector pre-processing improving the overall quality of CBIR and classification systems.

**Keywords:** feature transformation, discretization, pre-processing, feature vector, *clustering*, CBIR, classification.

# LISTA DE FIGURAS

# SUMÁRIO

# Capítulo 1

# INTRODUÇÃO

## 1.1 Considerações Iniciais

A medicina é uma das áreas que mais tem se beneficiado do aperfeiçoamento dos equipamentos eletrônicos e dos sistemas computacionais. O volume de dados médicos armazenados, que incluem exames, diagnósticos e procedimentos de tratamento, tem uma tendência de crescimento exponencial. Esse grande volume de dados históricos é uma valiosa fonte de conhecimento, que pode ser usada para o auxílio ao diagnóstico médico e para o ensino da medicina (RIBEIRO, 2009).

Em virtude da complexidade da análise e tratamento dos dados que incluem imagens, os profissionais da área médica ainda não se beneficiam de grande parte dessa fonte de conhecimento. As imagens são representadas computacionalmente por meio de vetores de características. No entanto, existe um problema na definição de um vetor de características adequado para a representação das imagens: existe um grande número de características que podem ser extraídas das imagens e usadas para sua representação, mas são desconhecidas quais delas são as mais importantes para cada tipo de aplicação.

Para a representação das imagens, é importante o uso de um conjunto compacto e representativo de características, pois o uso de um grande número de características pode levar ao problema da “maldição da alta dimensionalidade” (JEONG, 2007), que degrada a precisão e o tempo de busca.

Devido a tais desafios, as técnicas de recuperação de imagens por conteúdo têm sido bastante pesquisadas nos últimos anos. Neste trabalho foram utilizadas técnicas de redução de dimensionalidade e discretização para pré-processar vetores de características visando melhorar a precisão das buscas por conteúdo.

## 1.2 Motivação

As técnicas de recuperação de imagens por conteúdo (*CBIR, do Inglês – Content-Based Image Retrieval*) ainda não atendem, plenamente, às necessidades de consulta dos usuários, retornando frequentemente resultados não relevantes ou imprecisos. As limitações de CBIR se devem principalmente ao problema de representação da informação subjetiva, ou seja, da interpretação da imagem, utilizando descritores matemáticos de baixo nível (vetores de características), resultando no problema conhecido como "*gap* semântico" (DESERNO, 2008). O *gap* semântico pode ser tratado usando retrorealimentação de relevância, técnicas de seleção de características e/ou consultas considerando, além da similaridade visual, descrições textuais de alto-nível das imagens.

Existem técnicas na literatura que propõem retrorealimentação de relevância para tentar capturar a intenção e a percepção do usuário (como demonstra Liu, 2007). A desvantagem destes métodos é que o usuário é obrigado a passar por várias rodadas (ciclos de retrorealimentação de relevância), e que não é possível prever ou otimizar a qualidade da recuperação de antemão.

## 1.3 Objetivos

Este projeto se direciona ao desenvolvimento de técnicas de pré-processamento de vetores de características para a recuperação de imagens por conteúdo visando aumentar a precisão de consultas relacionadas à área médica. As técnicas propostas devem permitir a realização de consultas por conteúdo tendo como dados de entrada vetores de características de uma imagem de consulta ou a

própria imagem de consulta. O objetivo principal é utilizar o pré-processamento de imagens (transformação de características - *feature transformation* e discretização ponderada de características – *weighted feature discretization*) para aumentar a precisão da busca por conteúdo e diminuir o *gap* semântico, melhorando o poder de representação do vetor de características.

## 1.4 Organização do trabalho

Esta dissertação está organizada em 6 capítulos. O capítulo 1 discute-se a introdução, a motivação e os objetivos desta dissertação; O capítulo 2 apresenta os conceitos envolvidos num sistema de recuperação de imagem por conteúdo (CBIR), suas etapas. No capítulo 3 são apresentadas as técnicas de exploração de dados, tais como redução de dimensionalidade, agrupamento (*clustering),* classificação e discretização de dados. No capítulo 4 são explanados os trabalhos correlatos de redução de dimensionalidade, agrupamento e classificação aplicados ao processamento de imagens.

O capítulo 5 apresenta os métodos FTK (*Feature Transformation Based on K-means* - transformação de características baseada no algoritmo *k-means*) e WFD (*Weight Feature Discretizer* – discretizador de características com ponderação). Também são descritos os experimentos realizados com estes algoritmos e outros existentes na literatura no capítulo 6 e finalmente a monografia é finalizada com as conclusões.

# Capítulo 2

# CONSULTA POR CONTEÚDO (CBIR)

## 2.1 Considerações Iniciais

Os **sistemas de recuperação de imagens por conteúdo** (CBIR) abordam tecnologias e métodos voltados para organização de grandes repositórios de imagens digitais por meio do conteúdo visual das mesmas. Várias áreas da computação contribuem para o desenvolvimento dos sistemas CBIR, dentre elas destacam-se: base de dados, processamento de imagens, visão computacional e interação humano-computador (RIBEIRO, 2009). As imagens são representadas através de vetores de características tais como: cor, forma e textura. Ao se trabalhar com imagens, existe um grande desafio: encontrar a melhor representação numérica que sintetize a essência da imagem por meio deste vetor de características, bem como indexar imagens retornadas e avaliar a eficiência de métodos empregados. Este capítulo apresenta um detalhamento dos sistemas CBIR necessário para o entendimento deste trabalho de mestrado.

## 2.2 Etapas de um Sistema CBIR

Inicialmente, uma imagem de consulta é fornecida como entrada para o sistema. O sistema CBIR se encarrega primeiramente de processar a imagem (filtrar, equalizar, segmentar, etc.). Após esta etapa, ocorre a extração de vetores de características da imagem, tais como cor (histogramas), forma (momentos invariantes), textura (descritores de HARALICK (1973)), etc. O próximo passo é utilizar uma estrutura de indexação (*slim-tree* (TRAINA, 2000), *R-tree* (GUTTMAN, 1984), *quad-tree* (SAMET, 1984), etc.) para recuperar imagens no banco de dados de imagem que contém um grande volume de dados.

A indexação é associada a um método de busca por similaridade, isto é, busca em que se considera quão “próximos” (similares) dois dados são entre si. Considerando $O_i$ a imagem para consulta e $O_j$ uma imagem do banco de dados de imagem, a similaridade entre os dados é definida por meio de uma função de distância $d(O_i, O_j)$, que retorna zero se ambos os objetos $O_i$ e $O_j$ forem idênticos, e um valor positivo que aumenta quanto maior for a distância (ou dissimilaridade) entre os objetos. Finalmente, é retornado um conjunto de imagens recuperadas similares à imagem de consulta. As etapas envolvidas num CBIR estão ilustradas na Figura 2.1.

**Figura 2.1: Visão geral de um sistema CBIR. Adaptado de (FILARDI, 2008).**

## 2.3 Representação das Imagens

Os sistemas CBIR representam as imagens por meio de suas características relacionadas a descritores (vetores de características) que buscam aproximar-se da percepção humana de cor, formato e textura. A inconsistência entre a informação de baixo nível automaticamente extraída das imagens e a interpretação humana de alto nível é chamada de "*gap* semântico" (DESERNO, 2008).

Em geral, as ferramentas de recuperação de imagens por conteúdo possuem dois obstáculos: conseguir uma recuperação rápida de imagens armazenadas em grandes bases de dados e reduzir o "*gap* semântico". Nesse sentido, o uso da mineração de regras de associação pode ajudar a efetivamente reduzir o "*gap* semântico" entre as características de baixo nível da imagem (representação numérica) e a interpretação humana (descrição semântica) (FELIPE, 2005).

## 2.4 Tipos de consulta

Uma consulta por similaridade pode ser classificada em *consulta por abrangência* ou *consulta por vizinhança*, de acordo com a sua abrangência ou cobertura.

Na consulta por raio de abrangência (do inglês, *range query* - RQ), é fornecido um objeto de referência *O* e um raio de cobertura *r*. O conjunto resposta $R_{rq}$ inclui todos os elementos *S* da base que se encontram a uma distância menor ou igual a *r* do elemento *O*, ou seja:

$$R_{rq} = \{S | d(S, O) \leq r\} \qquad (2.1)$$

A Figura 2.2 ilustra um exemplo de consulta por raio de abrangência no domínio bidimensional, onde o conjunto de resposta contém 8 elementos (incluindo o objeto de consulta) ou 7 não considerando o objeto de consulta como parte da base. A função de distância utilizada neste caso é a função euclidiana ($L_2$):

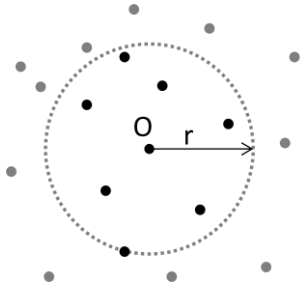



**Figura 2.2: Exemplo de uma consulta por raio de abrangência onde o conjunto de resposta contém 8 elementos (incluindo o objeto de consulta).**

Na consulta aos *k*-vizinhos mais próximos (do inglês, *k-nearest neighboor query* - $k_{NN}$) são fornecidos um objeto de referência *O* e um número inteiro *k* referente ao número de elementos mais próximos do elemento *O* que se deseja obter como conjunto de resposta $R_{kNN}$. Formalmente tem-se:

$$R_{kNN} = \{S | \forall\, P \in \{\Omega - R_{kNN}\}, d(O,S) \leq d(O,P), |R_{kNN}| = K\} \quad (2.2)$$

onde Ω representa o conjunto de todos os elementos.

A Figura 2.3 ilustra um exemplo de consulta do tipo $k_{NN}$ no domínio bidimensional, onde o conjunto de resposta contém cinco elementos.

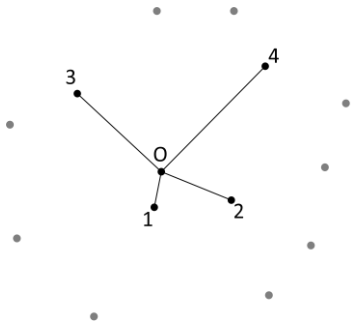



**Figura 2.3: Exemplo de uma consulta do tipo $k_{NN}$ onde *k*=5.**

## 2.5 Características para representar imagens

Assinaturas de forma

Assinaturas de formas podem ser obtidas a partir da representação de forma baseada em regiões e contornos. Em geral, as assinaturas baseadas em contorno

são criadas a partir de um ponto inicial do contorno da forma e percorrendo-a, em sentido horário ou anti-horário. Para ilustrar este conceito, representa-se graficamente a distância entre cada ponto de contorno e o centróide da forma em termos da sequencia dos pontos de contorno, que atuam como parâmetros (Figura 2.4) (COSTA, 2001).

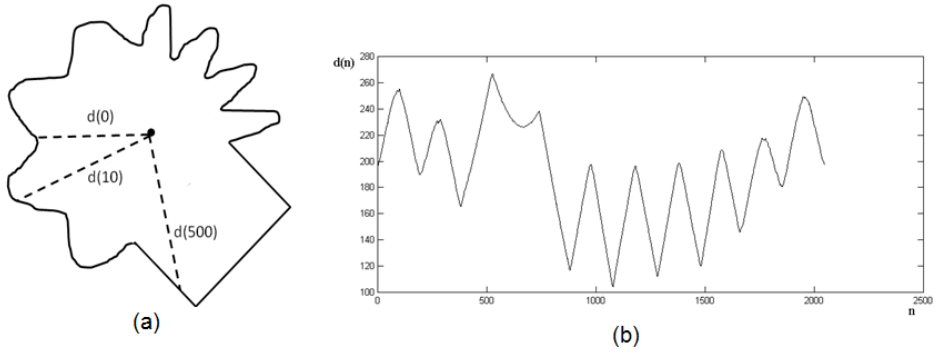



**Figura 2.4: (a) Exemplo de contorno e (b) sua assinatura *d*(*n*) (COSTA, 2001).**

As linhas que unem o centróide e alguns pontos de amostragem ao longo do contorno são mostrados, juntamente com a indicação do respectivo parâmetro.

Histograma

Entre as técnicas mais simples e mais úteis para o tratamento de imagens digitais estão aquelas baseadas em histogramas. Um histograma é um mapeamento que atribui um valor numérico *h(n)* para cada *n* nível de cinza da imagem (histograma de brilho) ou, para imagens coloridas, o nível de ocorrência de cada cor. Esses valores correspondem ao número de vezes que cada *n* nível de cinza ou cor específica ocorre na imagem (**COSTA**, 2001). A Figura 2.5 ilustra uma imagem e seu histograma.

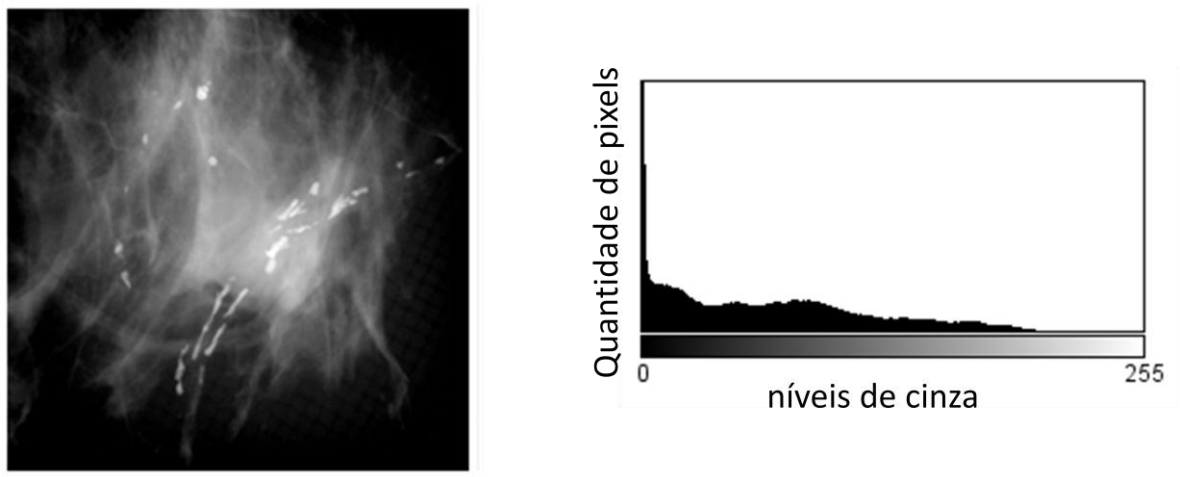



**Figura 2.5: Exemplo de imagem e seu histograma.**

Histogramas de cores são amplamente utilizados para obtenção de resultados em sistemas de recuperação de imagens por conteúdo, pois são muito eficientes em matéria de computabilidade e oferecem insensibilidade para pequenas mudanças em relação à posição do objeto (translação e rotação), e objetos distintos frequentemente possuem histogramas diferentes. Porém, quando utilizados para se realizar comparações e buscas por similaridade, apresentam uma importante limitação: sua caracterização grosseira de uma imagem que pode resultar em histogramas semelhantes ou até mesmo iguais de imagens diferentes (Figura 2.6).

Além disso, os métodos tradicionais de comparação entre histogramas tendem a aumentar esse volume de falsos positivos, devido à sua rigidez de tratamento dos diferentes *bins* (níveis de cinza ou cor) dos histogramas, e também por não levarem em consideração as características globais dos mesmos (FELIPE, 2005).

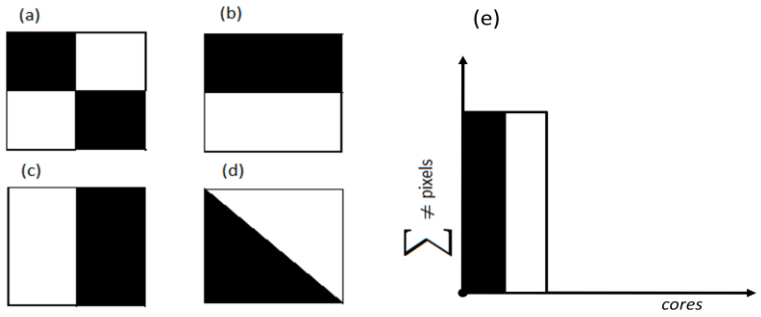
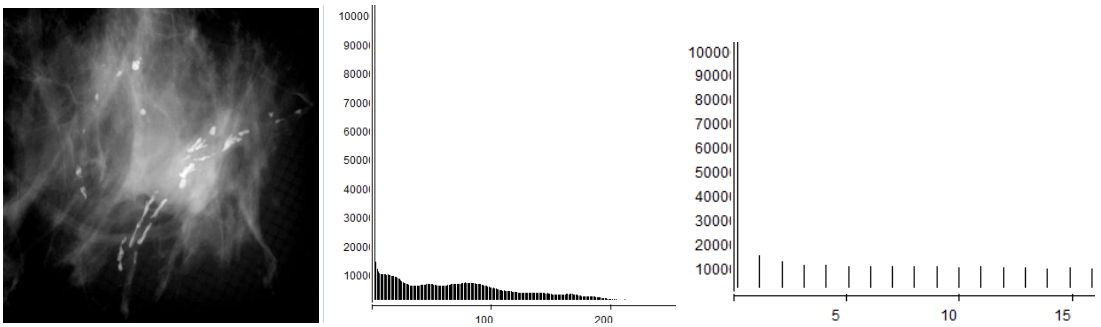



**Figura 2.6: Exemplo de imagens (a - d) com o mesmo histograma (e).**

Para aumentar a eficiência do processamento, as cores da imagem ou tons de cinza podem ser requantizadas, com o objetivo de diminuir o número de cores (ou níveis de cinza) possíveis e facilitar o tratamento das mesmas. O processo de requantização do histograma consiste em comprimir um intervalo de valores de intensidade em um único valor, chamado *quantum*. A Figura 2.7 apresenta uma imagem com histograma de 256 e 16 níveis de cinza.




**(a)** **(b)** **(c)**

**Figura 2.7: (a) imagem original, (b) histograma de 256 níveis e (c) histograma de 16 níveis.**

Textura

Uma importante abordagem para a descrição de regiões é a quantificação de seu conteúdo de textura. Não existe nenhuma definição formal de textura, porém esse descritor fornece características tais como: suavidade, rugosidade e regularidade. Em processamento de imagens são usadas as seguintes abordagens para se descrever textura: a estatística, a estrutural e a espectral. As abordagens estatísticas levam em consideração características tais como suave, áspera, granular, etc. As técnicas estruturais são indicadas quando são analisadas imagens com padrões fixos e repetitivos de textura. As técnicas espectrais são utilizadas basicamente na detecção de periodicidade global em uma imagem através de picos de alta-energia no espectro (GONZALEZ, 2000).

Haralick (1973) propõe 14 medidas que podem ser extraídas de uma matriz de co-ocorrência.

## 2.6 Cálculo de Distância

A tarefa de avaliação de similaridade entre elementos de um repositório de imagens é realizada mediante a escolha adequada de uma função de distância. A função de distância indica o nível de semelhança entre duas imagens quando é aplicada a vetores de características obtidos a partir do processamento das mesmas. Como existem várias funções de distâncias na literatura, trabalhos têm sido desenvolvidos para determinar qual é a melhor combinação de características e funções de distância que devem ser utilizadas para cada tipo de imagem [Silva, 2008]. A escolha do método e da função de distância a serem utilizados depende de uma série de fatores, tais como o contexto, as propriedades das imagens e as características requeridas.

Para uma função de distância ser considerada uma "função de distância métrica" é preciso que ela obedeça quatro propriedades. Considerando $p_1$, $p_2$ e $p_3$ como pontos em um espaço de características, as propriedades que tal função deve obedecer são as seguintes:

1. Simetria: $d(p_1, p_2) = d(p_2, p_1)$
2. Não-negatividade: $d(p_1, p_2) \geq 0$
3. Identidade: $d(p_1, p_2) = 0 \Leftrightarrow p_1 = p_2$
4. Desigualdade triangular: $d(p_1, p_2) \leq d(p_1, p_3) + d(p_3, p_2)$

Um sistema que efetue a armazenagem e a manipulação de imagens possui três componentes fundamentais: um conjunto de extratores de características (descritores) de imagens, uma estrutura de indexação e uma família de distâncias. A família de distâncias proposta por Minkowski é comumente conhecida como família de distâncias $L_p$ e assume a seguinte forma:

$$L_p(R,S) = \sqrt[\mathrm{p}]{\sum_{i=1}^{N-1} |R_i - S_i|^{\mathrm{p}}} \qquad (2.3)$$

onde $R = [R_0, R_1, \ldots, R_{N-1}]$ e $S = [S_0, S_1, \ldots, S_{N-1}]$ são vetores de $N$ características.

Existem três funções de distância conhecidas que fazem parte da família $L_p$. A distância $L_1$ corresponde à distância Manhattan, ou "*city block*":

$$L_1(R,S) = \sum_{i=1}^{N-1} |R_i - S_i| \qquad (2.4)$$

Se p=2, tem-se a distância euclidiana $L_2$, que é uma das funções de distância mais conhecidas e utilizadas.

$$L_2(R,S) = \sqrt[2]{\sum_{i=1}^{N-1} (R_i - S_i)^2} \qquad (2.5)$$

Fazendo $p \to \infty$ tem-se a distância $L_\infty$ ou $L_{infinity}$ definida como:

$$L_\infty(R,S) = \max_{0 \le i < N} [|R_i - S_i|] \qquad (2.6)$$

A Figura 2.8 ilustra as configurações de um conjunto de pontos equidistantes considerando as distâncias $L_1$, $L_2$ e $L_{infinity}$ num espaço bidimensional.

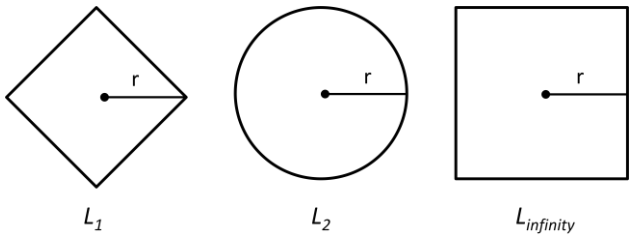



**Figura 2.8: configurações de um conjunto de pontos equidistantes considerando as distâncias $L_1$, $L_2$ e $L_{infinity}$ num espaço bi-dimensional.**

### Distância quadrática e distância de Mahalanobis

A função de distância quadrática leva em consideração a correspondência (similaridade) existente entre as características de índices iguais ou diferentes no vetor. A distância quadrática é definida como:

$$d_{quad}(R,S) = \sqrt{(R-S)^T A (R-S)} \qquad (2.7)$$

onde *A* é uma matriz de dimensão *NxN* de elementos $a_{ij}$ que correspondem ao coeficiente de afinidade entre os elementos de índices *i* e *j*. O valor de $a_{ij}$ é dado por:

$$a_{ij} = 1 - \frac{d_{ij}}{max[d_{ij}]} \quad onde \quad d_{ij} = |R_i - S_i| \qquad (2.8)$$

A distância Mahalanobis é um caso especial da distância quadrática onde a matriz de transformação *A* corresponde à matriz inversa da matriz de covariância, obtida a partir de um conjunto de treinamento de vetores de características. Para se alcançar a definição desta função é preciso considerar um vetor de variáveis aleatórias $X = [X_0, X_1, ..., X_{N-1}]$ que assume os valores de características dos vetores que constituem um conjunto de treinamento estabelecido. A matriz de covariância *V* é dada por $V = \left[\sigma_{ij}^2\right] onde \ \sigma_{ij}^2 = E\left[X_i, X_j\right] - E[X_i]E[X_j]$ e $E[X]$ é o valor esperado da variável aleatória. Assim, a distância Mahalanobis entre dois vetores R e S é definida como:

$$d_{mah} = \sqrt{(R-S)^T V^{-1} (R-S)} \qquad (2.9)$$

No caso especial onde as variáveis $X_0, X_1, ..., X_{N-1}$ são estatisticamente independentes, a matriz de covariância *V* é a matriz diagonal:

$$V = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}$$

Neste caso a distância Mahalanobis fica reduzida à seguinte forma:

$$d_{mah-i}(R,S) = \sum_{i=0}^{N-1} \frac{(R_i - S_i)^2}{\sigma_i^2} \qquad (2.10)$$

Esta última corresponde a uma função de distância $L_2$ ponderada pela variância. Para as características que possuem pequena variância dentro do conjunto de treinamento é atribuído um peso maior, enquanto que para características de maior variância o peso é menor.

**Funções de distância típicas para vetores de características específicos**

Os descritores de uma imagem podem ser classificados como genéricos ou específicos do domínio, de acordo com as características visuais (FELIPE, 2005). Os descritores genéricos contemplam características tais como cor, textura e forma, enquanto os descritores específicos do domínio são dependentes do contexto, como faces humanas, impressões digitais ou nódulos de tumor.

Devido à subjetividade da percepção humana, não existe uma única representação ótima para um dado descritor, mas sim múltiplas opções de representação que caracterizam o descritor a partir de diferentes perspectivas. Ainda assim, as pesquisas têm indicado a adequação de certas funções de distância para serem usadas com diferentes descritores. As utilizações mais frequentes das funções de distância estão associadas a descritores específicos. Segundo Hand (2001), são as seguintes:

- histogramas de cores: distância Mahalanobis ou interseção de histogramas;
- textura: distância Euclidiana Ponderada ou distância Mahalanobis;
- forma: distância Euclidiana simples (algoritmo *k-means*).

O método comparativo (mais restritivo, ou menos restritivo, por exemplo) deve ter uma função de distância condizente associada. Assim, cada caso de aplicação exige um estudo específico.

## 2.7 Considerações Finais

Este capítulo apresentou uma visão geral de um sistema de recuperação de imagens por conteúdo. Um dos maiores desafios em CBIR está diretamente ligado à tarefa de extração de características das imagens e sua representação matemática para que seja possível medir o nível de semelhança visual entre elas, bem como definir as características mais adequadas para um determinado domínio de imagens. Nos capítulos seguintes, serão discutidas as técnicas de exploração de dados, técnicas de processamento de imagens, bem como a discretização e redução de dimensionalidade, aplicadas ao processamento de imagem.

# Capítulo 3

# TÉCNICAS DE EXPLORAÇÃO DE DADOS

## 3.1 Considerações Iniciais

Neste capítulo serão descritos conceitos de redução de dimensionalidade, agrupamento (*clustering),* classificação e discretização de dados, conceitos necessários para o entendimento deste trabalho de mestrado, cujo foco principal é o desenvolvimento de técnicas de pré-processamento (redução de dimensionalidade e discretização ponderada) para o aumento do poder de representação do vetor de características.

## 3.2 Redução de Dimensionalidade

Redução de dimensionalidade é o processo de redução do número de características empregadas para representar um conjunto de dados determinado. Um conjunto limitado de características relevantes simplifica a representação do padrão e, consequentemente, o processo de comparação será mais rápido e usará menos memória (JAIN, 2000). Técnicas de redução de dimensionalidade podem ser classificadas em técnicas de seleção de características e técnicas de transformação de características. Diferentes abordagens para redução de dimensionalidade podem

ser descritas de acordo com a hierarquia mostrada na Figura 3.1, destacando o método FTK, proposto nesta dissertação e detalhado em capítulo posterior.

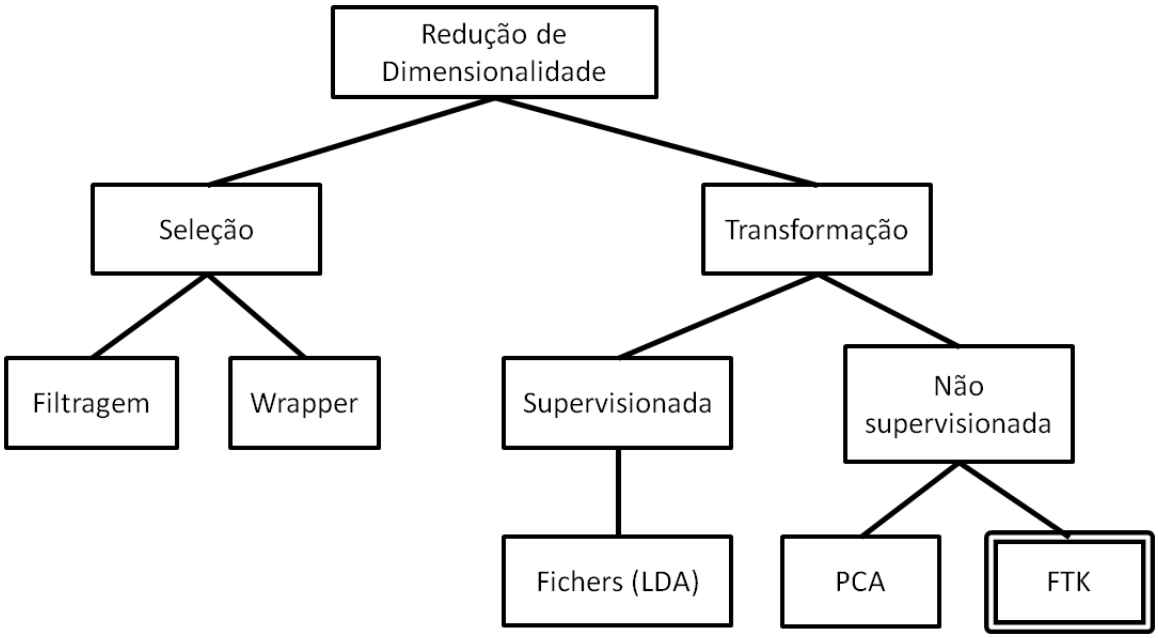



**Figura 3.1: Taxonomia das abordagens de redução de dimensionalidade.**

### 3.2.1. Técnicas de Seleção de Características

As técnicas de seleção de características não modificam o espaço de característica original, produzindo subconjuntos dos vetores de características originais. Um processo de seleção de características consiste em quatro etapas básicas: geração, avaliação, critério de parada e de validação de resultado de cada subconjunto. A etapa de geração de subconjuntos consiste em uma pesquisa que produz subconjuntos de características candidatos para avaliação com base em uma estratégia de pesquisa específica e cada subconjunto candidato é avaliado e comparado com os subconjuntos melhores anteriores, de acordo com um critério de avaliação determinado. Se novos subconjuntos virem a ser melhores, eles substituirão os anteriores. A geração de subconjuntos e avaliação são repetidas até que um determinado critério de parada seja satisfeito. Então, o melhor subconjunto selecionado geralmente precisa ser validado utilizando um conjunto de dados de teste. Quando um algoritmo indutivo é utilizado para avaliação do subconjunto de características, a seleção de características é classificada em dois grupos: métodos de filtragem e *wrapper*. Métodos de filtragem avaliam a "adequabilidade"

(*properness*) do subconjunto de características, usando a característica intrínseca dos dados. Alguns critérios comumente empregados são medidas de distância, medidas de informação, medidas de dependência e consistência (LIU, 2005).

Uma inconsistência ocorre se duas instâncias têm a mesma característica, mas pertencem a classes diferentes. Os métodos de filtragem não são computacionalmente caros, uma vez que não envolvem algoritmos de indução. No entanto, eles também têm o risco de selecionar subconjuntos de características que podem não funcionar bem com o algoritmo aplicado no aplicativo do usuário. Os métodos de *wrapper*, por outro lado, utilizam o algoritmo de mineração final (agrupamento, classificação e etc.) próprios para avaliar os subconjuntos de características candidatos. Eles geralmente selecionam as características mais adequadas para a tarefa de mineração do que os métodos de filtragem, mas geralmente apresentam maior custo computacional (SILVA, 2011).

O custo computacional e especificidade são as principais desvantagens do método *wrapper*, quando comparado com os métodos de filtragem. No entanto, o que efetivamente conta para o usuário é a busca, uma vez que a seleção de características é, normalmente, empregada na etapa de pré-processamento (SILVA, 2011).

### 3.2.2. Técnicas de Transformação de Características

As técnicas de transformação de características, também chamadas de técnicas de extração ou de redução de características, alteram o espaço de características original, produzindo um conjunto completamente novo de características para representar os dados.

Uma estratégia natural para redução de dimensionalidade é extrair componentes importantes a partir de dados originais, que podem contribuir para a divisão de clusters. Análise de componentes principais (PCA - *Principle Component Analysis*), ou transformação *Karhunen-Loéve,* é uma das abordagens clássicas, relacionadas com a construção de uma combinação linear de um conjunto de vetores que podem descrever melhor a variância dos dados (XU, 2005).

Análise de componentes principais é um método não supervisionado de análise de dados que fornece uma sequência das melhores aproximações lineares

para um determinado conjunto de dados de alta dimensão. É uma das técnicas mais populares para redução de dimensionalidade (BARSHAN, 2011).

Dada a matriz *NxD* de padrões de entrada $X = [x_1, ..., x_j, ..., x_N]^T$, o mapeamento linear $Y = XV$ projeta cada *feature* $x_i$ $(1<=i<=N)$ em um subespaço de menor dimensão $L$ $(L<D)$, em que $Y$ é a matriz resultante $NxL$ e $V$ é a matriz de projeção, cujas colunas são autovetores que correspondem aos $L$ maiores valores da matriz de covariância $\Sigma_{DxD}$, calculada a partir do conjunto de dados (por isso os vetores das colunas são ortonormais) (XU, 2005).

PCA é uma transformação linear que transforma os dados para um novo sistema de coordenadas de tal modo que o novo conjunto de variáveis, os componentes principais, são funções lineares das variáveis originais, não estão correlacionados, e a maior variância, por qualquer projeção dos dados, estende-se sobre a primeira coordenada, a segunda maior variância sobre a segunda coordenada, e assim por diante. Para tanto, é calculada a matriz de covariância para o conjunto de dados completo. Em seguida, os autovetores e autovalores da matriz de covariância são computados e classificados de acordo com o autovalor (JOLIFE, 1986).

A Análise de Discriminantes Lineares ou Análise Discriminante de Fisher (FDA - *Fisher Discriminant Analysis* ou LDA - *Linear Discriminant Analysis*), (FISHER, 1936) é um método supervisionado tradicional de redução de dimensionalidade. Para um problema com $n$ classes, a LDA mapeia os dados para um espaço com dimensão $n-1$ tal que a distância entre as classes é maximizada enquanto a variância dentro da classe é minimizada.

## 3.3 Agrupamento (*clustering*)

O conceito de agrupamento é utilizado para o desenvolvimento de uma técnica de transformação de características, descrita com detalhes em capítulo posterior. O problema de *clustering*, também conhecido como agrupamento ou classificação não supervisionada, consiste em agrupar uma determinada coleção de padrões não-rotulados em agrupamentos significativos.

De um modo geral, *clustering,* ou agrupamento, é um método de organizar objetos em grupos cujos membros são similares entre si. Mais especificamente, um problema de agrupamento consiste em se agrupar os objetos (elementos) de um dado conjunto *X* (base de dados), tal que os elementos mais similares fiquem no mesmo *cluster* (grupo) e os elementos menos similares sejam colocados em *clusters* diferentes.

Para Jain (1999), o agrupamento é a classificação não supervisionada de padrões (observações, itens de dados ou vetores de características) em grupos (*clusters*). Trata-se de agrupar uma determinada coleção de padrões não rotulados em agrupamentos significativos. Os rótulos podem ser associados com os *clusters* também, mas esses rótulos de categoria são orientados a dados, ou seja, são obtidos somente a partir dos dados.

O problema de agrupamento tem sido abordado em diversos contextos e por pesquisadores em muitas disciplinas, o que reflete o seu grande apelo e utilidade na análise exploratória de dados. Algoritmos de agrupamento também podem ser aplicados às áreas de segmentação de imagens, reconhecimento de objetos, e recuperação da informação. A Figura 3.2 ilustra o processo de agrupamento, mediante os parâmetros de entrada.

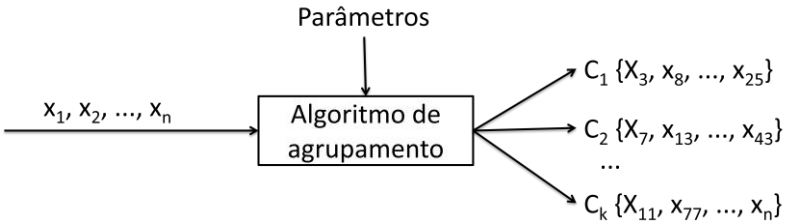



**Figura 3.2: O processo de agrupamento (*clusterização*).**

Problemas de agrupamento podem ser definidos formalmente da seguinte maneira (DIAS, 2004): dado um conjunto com *n* elementos $X = \{x_1,x_2,...,x_n\}$, o problema de agrupamento consiste na obtenção de um conjunto $C=\{c_1,c_2,...,c_n\}$ de *k clusters* ($k \leq n$), tal que os objetos contidos em um *cluster* $C_i$ possuam uma maior similaridade entre si do que com os objetos de qualquer um dos demais *clusters* do conjunto *C*. O conjunto *C* é considerado um agrupamento com *k clusters* caso as seguintes condições ocorram:

$$1)\ X = \bigcup_{i=1}^{k} C_i \qquad (3.1)$$

2) $C_i \neq \emptyset,\ \ 1 \leq i \leq k$ (3.2)

3) $C_i \cap C_j = \emptyset,\ \ 1 \leq i, j \leq k\ \ e\ \ i \neq j$ (3.3)

O valor de *k* pode ser conhecido ou não. Caso o valor de *k* seja fornecido como parâmetro para a solução, o problema é referenciado na literatura como "*problema de k-agrupamento"*. Caso contrário, isto é, caso o *k* seja desconhecido, o problema é referenciado como "*problema de agrupamento automático*" e a obtenção do valor de *k* faz parte do processo de solução do problema.

### 3.3.1 Técnica baseada em Centróide : O método *k-means*

O método *K-means* é simples e pode ser usado para diferentes tipos de dados. É bastante utilizado no Sensoriamento Remoto para executar procedimentos de classificação não supervisionada de imagens de satélite (SCHOWENGERDT, 1997).

Trata-se de um método de agrupamento não hierárquico que busca minimizar a distância dos elementos de forma iterativa. Primeiro, escolhemos K centróides iniciais, onde K é um parâmetro especificado pelo usuário que indica o número de grupos desejados. Cada objeto é atribuído ao centróide mais próximo, e cada conjunto de objetos atribuído a um centróide constitui um grupo. O centróide de cada grupo é então atualizado baseado nos objetos atribuídos ao grupo. Repetem-se os passos de atribuição e atualização até que nenhum objeto mude de grupo ou até que os centróides permaneçam os mesmos (TAN, 2006).

O objetivo do agrupamento geralmente é expresso por uma função objetivo que depende da proximidade dos pontos entre si ou dos centróides do grupo, isto é, minimiza a distância quadrada de cada ponto com seu centróide mais próximo. Esta função objetivo, que mede a qualidade do agrupamento, é conhecida como *soma do erro quadrado* (*SSE - Sum of the Squared Error).* Formalmente, temos:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |c_i - x|^2 \qquad (3.4)$$

onde:

*x*: elemento do *cluster* $C_i$

$C_i$: i-ésimo cluster i=(1,...,K)

$c_i$: centróide do cluster $C_i$

$c$: centróide de todos os pontos

$m_i$: número de objetos no i-ésimo cluster

$m$: número de objetos do conjunto

$K$: número de clusters

$|c_i - x|^2$: distância Euclidiana ($L_2$) entre dois objetos.

O centróide para o algoritmo *k-means* pode ser derivado matematicamente quando a função de proximidade for a distância Euclidiana e o objetivo seja minimizar a função SSE. É possível atualizar um centróide de um grupo de forma que a SSE do grupo seja minimizada resolvendo para o k-ésimo centróide $c_k$, o qual minimiza a equação **Erro! Fonte de referência não encontrada.**, diferenciando a função SSE e igualando-a a zero:

$$\frac{\partial}{\partial c_k} SSE = \frac{\partial}{\partial c_k} \sum_{i=1}^{k} \sum_{x \in C_i} (c_i - x)^2 = \sum_{i=1}^{k} \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2 = \sum_{x \in C_k} 2 * (c_k - x_k) = 0$$

$$\sum_{x \in C_k} 2 * (c_k - x_k) = 0 \;\Rightarrow\; m_k c_k = \sum_{x \in C_k} x_k \;\Rightarrow\; c_k = \frac{1}{m_k} \sum\nolimits_{x \in C_k} x_k \qquad (3.5)$$

Assim, o i-ésimo *cluster* $C_i$ pode ser identificado por seu centróide $c_i$:

$$c_i = \frac{1}{m_i} \sum\nolimits_{x \in C_i} x \qquad (3.6)$$

Por exemplo, o centróide de um *cluster* que contém os três pontos em duas dimensões, (1,1), (2,5) e (6,3), é $\left(\left(\frac{1+2+6}{3}\right), \left(\frac{1+5+3}{3}\right)\right) = (3,3)$. Na Figura 3.3 é ilustrado o centróide marcado como "o" deste exemplo.

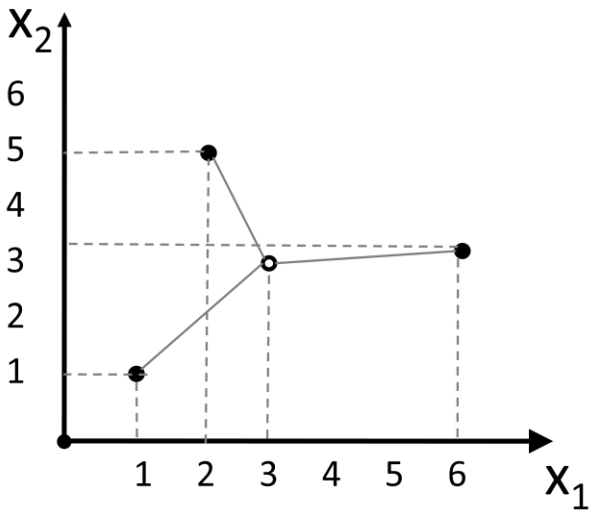



**Figura 3.3: Centróide localizado com o método *k-means* marcado com "o".**

O algoritmo de *k-means* tem o parâmetro de entrada, *k*, e particiona um conjunto de *m* objetos em *k clusters* de modo que a semelhança *intracluster* resultante é alta (objetos tão homogêneos quanto possível), mas a semelhança *intercluster* é baixa. A similaridade do *cluster* é medida em relação ao valor médio dos objetos em um *cluster*. A Figura 3.4 ilustra o algoritmo.

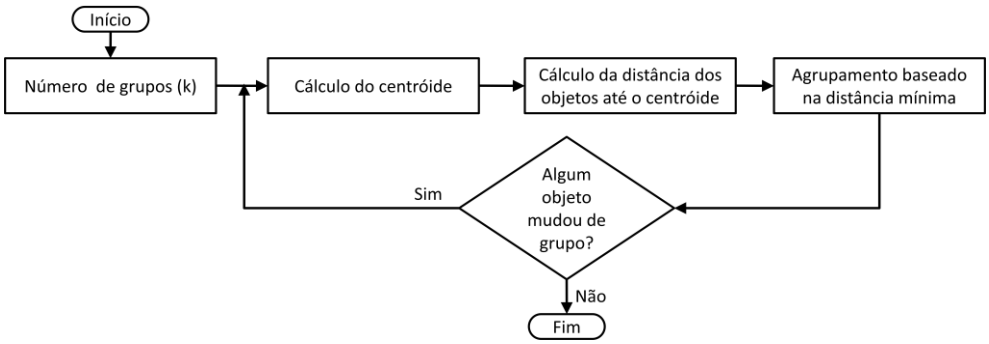



**Figura 3.4: Fluxograma do algoritmo *k-means*.**

Os passos do algoritmo *k-means* são:

1. Selecione k pontos como centróide s iniciais
2. **Repita**
    a. Forme k grupos atribuindo cada ponto ao seu centróide mais próximo (com base na distância mínima).
    b. Recalcule o centróide de cada grupo

3. **Até que** os centróide s não mudem ou nenhum objeto mude de grupo

Este algoritmo é veloz, geralmente convergindo em poucas iterações para uma configuração estável, na qual nenhum elemento está designado para um *cluster* cujo centro não lhe seja o mais próximo.

## 3.4 Classificação de dados

A classificação de dados permite agrupar dados em uma hierarquia de classes, de acordo com os valores dos seus atributos. Os registros agrupados em classes são formados por um atributo de classe, que determina a classe do registro, e um conjunto de atributos de predição. O objetivo é descobrir as relações existentes entre os atributos de predição e o atributo alvo, utilizando registros cuja classificação é conhecida.

Um classificador é uma função que possui como entrada padrões desconhecidos e como saída, rótulos que identificam a que classe tais padrões pertencem.

Formalmente, temos: dado um padrão desconhecido x, pertencente ao conjunto padrões de teste X em um espaço de características, e o conjunto $\Omega$ de todas as classes existentes, um classificador é uma função $f: X \rightarrow \Omega$, tal que $f(x) = \omega_i$, em que $\omega_i$ é uma i-ésima classe de $\Omega$.

O classificador OneR (WITTEN, 2005) cria uma regra para cada atributo dos dados de treino determinando a classe mais frequente (classe que mais vezes aparece) para cada atributo. Em seguida, é selecionada a regra com menor porcentagem de erro, a chamada regra única.

Uma regra é um conjunto de valores de atributos limitados pela sua classe majoritária. A porcentagem de erro de uma regra é o número de instâncias de treino na qual a classe de um valor de atributo não é concordante com a classificação desse atributo na regra. Se duas ou mais regras possuírem a mesma porcentagem de erro, a regra é escolhida ao acaso. Esse classificador é utilizado nos experimentos desta dissertação para a validação dos métodos desenvolvidos. Ele foi

escolhido para ser usado nos experimentos como base de comparação com outros algoritmos, devido à sua simplicidade e necessidade de apenas um atributo.

O algoritmo J48 (QUINLAN, 1993) permite a criação de modelos de decisão em árvore. O modelo de árvore de decisão é construído pela análise dos dados de treino e o modelo utilizado para classificar dados ainda não classificados. Este algoritmo gera árvores de decisão, em que cada nó da árvore avalia a existência ou significância de cada atributo individual, escolhendo o atributo mais apropriado para cada situação. Uma vez escolhido o atributo, os dados de treino são divididos em subgrupos, referentes aos diferentes valores dos atributos. O processo é repetido para cada subgrupo até que uma grande parte dos atributos em cada subgrupo pertença a uma única classe. Esse classificador também é utilizado nos experimentos desta dissertação para a validação dos métodos desenvolvidos por ser um dos mais utilizados da literatura.

Um classificador bayesiano é um classificador estatístico baseado no teorema de Bayes (JOHN e LANGLEY, 1995). O modelo construído por este algoritmo é um conjunto de probabilidades. O classificador simples bayesiano (Naive Bayes) assume que o valor de um atributo é independente do valor dos demais atributos (independência condicional). As probabilidades são estimadas pela contagem da frequência de cada valor de característica para as instâncias dos dados de treino. Dada uma nova instância, o classificador estima a probabilidade dessa instância pertencer a uma classe específica, baseada no produto das probabilidades condicionais independentes para os valores específicos da instância. Esse classificador também é utilizado nos experimentos desta dissertação para a validação dos métodos desenvolvidos por também ser um dos mais utilizados da literatura.

## 3.5 Discretização

Métodos de discretização de características têm sido propostos com o objetivo de obter representações da base de dados que são mais adequadas para a mineração de dados. O uso dessas técnicas pode levar a uma menor utilização de memória para a representação dos dados. Alguns trabalhos da literatura indicam

que os métodos de discretização geralmente levam a um aumento na precisão e a uma diminuição no tempo de treinamento, em comparação ao uso das características originais (FERREIRA e FIGUEIREDO, 2011].

Na mineração de dados, os tipos mais comuns de atributos (características) são os atributos contínuos, discretos e nominais. Conjuntos contínuos possuem valores intermediários entre dois quaisquer elementos. Um exemplo de conjunto contínuo é o conjunto dos números reais.

Conjuntos discretos são aqueles que, uma vez estabelecida a ordem de seus elementos, entre dois elementos sucessivos não existe elemento. Um exemplo de conjunto discreto é o conjunto dos números naturais.

Os atributos nominais ou categóricos exprimem atributos de qualidade. Não existe qualquer relação de ordem entre elas. Esses atributos frequentemente assumem um número limitado de valores. Um exemplo de atributo categórico é a condição socioeconômica de uma família: alta, média e baixa.

O processo de mapeamento do domínio de um atributo contínuo para discreto é chamado discretização. A discretização de dados é um processo de dividir um conjunto composto de valores contínuos em intervalos. O objetivo de um algoritmo de discretização é determinar qual é o melhor conjunto de pontos de corte para discretizar os dados. Um ponto de corte é um limite de um intervalo de valores reais (RIBEIRO, 2009).

Para discretizar dados, podemos utilizar a distribuição normal de probabilidade. A distribuição normal padronizada ou *z-score* facilita os cálculos de probabilidade, projetando qualquer análise mediante utilização de escores *z*. A distribuição normal padronizada é uma distribuição normal de probabilidade que tem a média $\mu=0$ e desvio-padrão $\sigma=1$.

$$z = \frac{x - \mu}{\sigma} \qquad \textbf{(3.7)}$$

Onde:

*x* é o escore bruto a ser normalizado,

$\mu$ é a média aritmética da população e

$\sigma$ é o desvio padrão da população.

Neste trabalho, foi utilizado o *z-score* para o desenvolvimento de uma técnica de discretização ponderada.

A discretização é uma etapa opcional executada no pré-processamento dos dados, quando é necessário reduzir o número de valores dos atributos. Muitos algoritmos de mineração exigem discretização prévia dos dados de entrada, provenientes de aplicações reais.

A desvantagem da discretização de dados é a perda de informação, que pode levar o algoritmo de mineração a ter resultados distorcidos. Apesar da perda de informação inerente ao processo de discretização, muitos trabalhos relatam um aumento significativo na precisão e na velocidade dos algoritmos de mineração ao utilizar uma técnica de discretização adequada no pré-processamento (ABRAHAM, 2006; LIU, 2002; KURGAN E CIOS, 2004).

## 3.6 Considerações Finais

Neste capítulo, foram discutidos os conceitos de redução de dimensionalidade (técnicas de seleção e transformação de características), agrupamento (método *k-means),* classificação e discretização de dados. Tais conceitos foram explanados com o objetivo de fundamentar teoricamente os métodos desenvolvidos neste trabalho de pesquisa. Observa-se que o pré-processamento de dados, quando bem empregado, pode melhorar a qualidade da mineração de dados. Nessa direção, ele também pode ser utilizado para melhorar a qualidade de sistemas de CBIR, como discutido no capítulo 5.

# Capítulo 4

# TRABALHOS CORRELATOS

## 4.1 Considerações Iniciais

Neste capítulo, são comentados os trabalhos existentes na literatura relacionados aos conceitos de redução de dimensionalidade, agrupamento e classificação aplicados à representação de imagens por conteúdo, que é o assunto deste trabalho de mestrado. Não foram encontrados trabalhos relacionados que aplicam discretização ponderada na representação de imagens.

## 4.2 Redução de dimensionalidade aplicada ao processamento de imagens

Durante os últimos anos, alguns trabalhos de pesquisa tiveram como objetivo tratar dos problemas da transformação de características e redução de dimensionalidade. A maioria deles empregou estratégias *wrapper* com o objetivo de melhorar os resultados da mineração de dados. Neste sentido, Jin propôs a transformação espaço de características, empregando padrões frequentes processados por um algoritmo de mineração que utiliza regras de associação (2010). Os resultados experimentais mostraram que a transformação do espaço de características obtém uma precisão mais elevada, quando comparado ao uso somente de padrões frequentes.

Barshan propôs uma técnica supervisionada de redução de dimensionalidade chamada ‘‘*supervised principal component analysis (supervised PCA)*’’, uma generalização do PCA (2011). Este método visa estimar uma sequência de componentes principais que possuem dependência máxima sobre a variável de resposta. A principal desvantagem desta técnica é sua complexidade cúbica, pois trata-se de um método baseado no PCA.

Segundo Poonguzhali, um método baseado em PCA foi utilizado para reduzir o vetor de características (de textura) para representar imagens de fígado (2007). Em seguida, as imagens foram agrupadas utilizando o método *k-means*, onde os erros de agrupamento, baseados na informação das classes normal, cisto, benigno e maligno, foram reduzidos. Esta técnica também possui complexidade cúbica, sendo esta sua principal desvantagem.

O método proposto por Loog, é derivado da análise de discriminante linear clássica (LDA), estendendo-se esta técnica para os casos em que existe dependência entre as variáveis de saída, isto é, os rótulos de classe, e não apenas entre as variáveis de entrada (2005). Diferentemente do método LDA, que leva em conta um único rótulo de classe para cada vetor de características, essa técnica também considera rótulos de classe da sua vizinhança na análise. A principal desvantagem deste método é a redução de dimensionalidade não linear.

## 4.3 Agrupamento aplicado ao processamento de imagens

A análise de agrupamento (*cluster*) é a organização de uma coleção de padrões (geralmente representado como um vetor de medições, ou um ponto em um espaço multidimensional) em grupos com base na similaridade (JAIN, 1999).

O método proposto por An também é baseado no refinamento de histogramas de imagens (2008). Este método refina ainda mais o histograma, dividindo os pixels em uma determinada partição em várias classes com base em vetores de coerência por cor (pixels coerentes e incoerentes, por exemplo). Várias características são calculadas para cada um dos clusters e classificadas usando o método *K-means*. Não só as características de cor da imagem são utilizadas, mas também a informação espacial é incorporada para refinar o histograma.

O método proposto por An é composto das seguintes etapas: na primeira etapa, a imagem colorida (RGB) é processada e convertida em tons de cinza (uma matriz de 2 dimensões contendo valores 0-255.) e em seguida realiza-se a quantização para reduzir o número de níveis da imagem (2008). São reduzidos de 256 níveis (*bins)* para 4 níveis (*bins)* na imagem quantizada usando quantização uniforme. Após esta etapa, são classificados os pixels coerentes e incoerentes. Um pixel é coerente se é uma parte de alguma região similar colorida considerada, caso contrário, é incoerente. Assim, os pixels são classificados como coerentes ou incoerentes dentro de cada segmento de cor. Se um pixel é parte de um grande grupo de pixels da mesma cor que formam, pelo menos, cinco por cento da imagem, então, um pixel é coerente e o grupo é chamado grupo coerente ou cluster. Caso contrário, o pixel é incoerente e o grupo é incoerente. Em seguida, mais duas propriedades são calculadas para cada bin. Primeiro, o número de *clusters* é encontrado para cada caso, ou seja, coerente e incoerente em cada um dos bins. Em segundo lugar, a média de cada grupo é computada. Assim, para cada posição, há seis valores: um para cada percentual de pixels coerente e incoerente, o número de *clusters* coerentes e incoerentes e a média de *clusters* coerentes e incoerentes. Em seguida, é calculada a distância entre todos os vetores, isto é, encontrada a similaridade entre cada par de objetos no conjunto de dados. O resultado é uma matriz de distâncias. Para cada *cluster* considerado coerente, são calculados: o tamanho do maior cluster, o tamanho médio dos clusters, o tamanho do menor cluster e a variância. A principal desvantagem desta técnica é a imprecisão na classificação de pixels coerentes e incoerentes. A Figura 4.1 ilustra o processo.

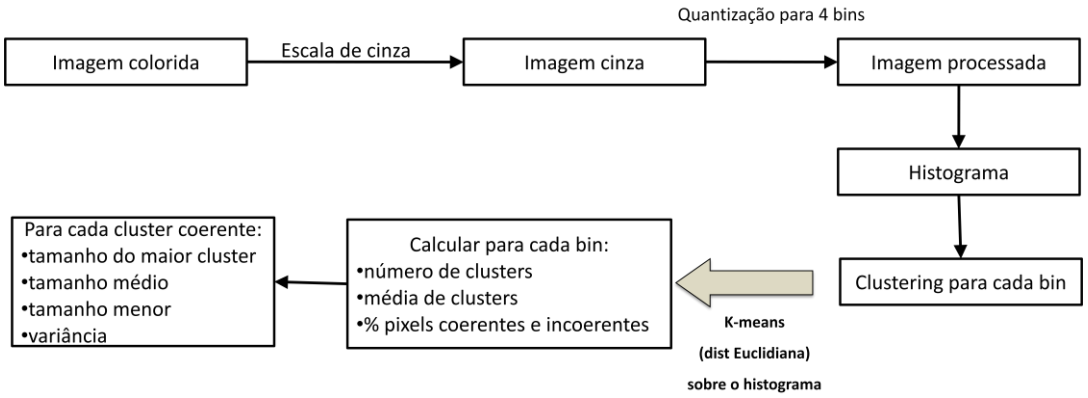



**Figura 4.1: O método proposto por An utiliza *k-means* no processamento de imagens (2008).**

Em um sistema de recuperação de imagem baseada em conteúdo, as imagens-alvo são classificadas por semelhanças de características. É possível a utilização do agrupamento *k-means* para a classificação do conjunto de características obtido a partir do histograma que fornece um conjunto de características para CBIR. Devido à sua eficiência e insensibilidade relativas a pequenas alterações, os histogramas são amplamente utilizados para recuperação de imagens por conteúdo. Mas, a principal desvantagem do emprego de histogramas é que muitas imagens de diferentes aparências podem ter histogramas semelhantes.

## 4.4 Considerações Finais

Neste capítulo foram comentados trabalhos correlatos ao desenvolvido neste trabalho de mestrado. Observa-se que os trabalhos discutidos envolvem, após a redução de dimensionalidade, processamentos extras que tornam essas técnicas custosas computacionalmente. Visando aplicar a idéia de *cluster* e minimizar o custo computacional, foi desenvolvido o método FTK de redução de dimensionalidade detalhado no próximo capítulo. No próximo capítulo também é apresentada a técnica WFD de discretização ponderada de vetores de características desenvolvida neste trabalho de pesquisa. No entanto, não foram encontrados trabalhos relacionados que aplicam discretização ponderada na representação de imagens.

# Capítulo 5

# MÉTODOS DESENVOLVIDOS

## 5.1 Considerações iniciais

Neste capítulo apresentamos duas técnicas de pré-processamento do vetor de características com o objetivo de melhorar a recuperação de imagens por conteúdo e sistemas de classificação de imagens: um método de redução de dimensionalidade baseado no algoritmo *k-means* chamado FTK (*Feature Transformation based on K-means*) e um método de discretização ponderada de características chamado WFD (*Weighted Feature Discretization*). Os métodos de pré-processamento aplicados aos vetores de características visam aumentar a precisão das consultas e a acurácia dos classificadores. O algoritmo FTK propõe a transformação de características do vetor original, agrupando o vetor de características por meio do algoritmo *k-means,* com o objetivo de compactar o vetor de características. O método WFD realiza uma discretização ponderada de características, privilegiando as faixas de características mais importantes para distinguir imagens. A utilização do método FTK em um sistema CBIR é descrita na Figura 5.1.

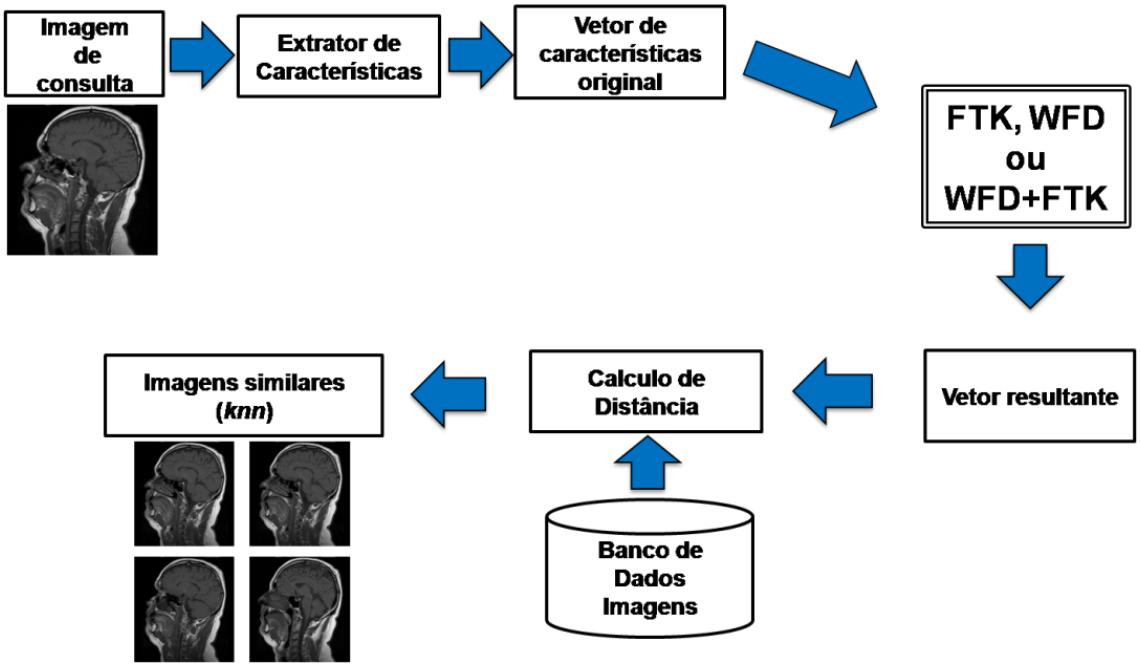



**Figura 5.1: Utilização dos métodos desenvolvidos em um sistema CBIR**

Os métodos propostos foram utilizados para pré-processar os vetores de características nas abordagens CBIR e classificação, comparando os resultados com o pré-processamento executado pelo método PCA (um método de transformação de características bem conhecido) e com os vetores de características originais. Os resultados dessas comparações são apresentados no próximo capítulo.

## 5.2 Algoritmo FTK

O método desenvolvido FTK (*Feature Transformation Based on K-means* - Transformação de Características Baseada no Algoritmo *K-means*) é utilizado com o objetivo de reduzir a dimensionalidade do vetor de características de uma imagem por meio da transformação de características. A principal motivação é o uso de uma técnica de agrupamento amplamente presente na literatura (método *k-means*) a partir de histogramas (vetores de características originais, extraídos das imagens) aliada a cálculos de complexidade linear, com o objetivo de simplificar as consultas por conteúdo por meio da redução de dimensionalidade do vetor original.

O método FTK é um método de filtragem que emprega uma técnica de agrupamento baseada no algoritmo *k-means* para processar os vetores de características, substituindo os valores das características pelos valores dos centróide s no processo de agrupamento. Seu objetivo é promover a redução de dimensionalidade do vetor de características, a fim de torná-lo mais compacto e discriminativo. Não foram encontrados métodos na literatura que atuam desta forma com esses objetivos.

Técnicas de agrupamento geralmente empregam o banco de dados particionado horizontalmente, onde cada objeto analisado é composto de um conjunto de características. Em relação a imagens, a técnica de agrupamento as agrupa de acordo com os valores dos seus vetores de características. Neste sentido, o método FTK a abordagem de agrupamento é utilizada sobre as características e sobre objetos.

Considere o vetor de características de uma imagem $V_i$ composto de um conjunto de características $f_{ij}$, isto é, $V_i = \{f_{i1}, f_{i2}, \ldots f_{in}\}$, onde $0 < i \leq m$ e $0 < j \leq n$; $m$ é o tamanho do banco de dados; $n$ é o tamanho do vetor de características. O conjunto $V_i$ de características $f_{ij}$ é organizado em *clusters*. O centróide $c_{ip}$ é obtido em cada *cluster* gerado, tal que $0 < p \leq k$, $k<n$, onde $k$ é o número de *clusters*. No método FTK, o vetor de características $V_i$ é substituído por um "vetor de centróides" $V_i' = \{c_{i0}, c_{i1,\ldots}, c_{ik}\}$, onde $k < n$. A saída do método é $R = \{\cup V_i' \mid 0 < i \leq m\}$, composto de vetores de características transformados, para cada imagem $i$ do banco de dados. As definições do método são apresentadas na Tabela 5.1.

**Tabela 5.1: Definições utilizadas no método FTK.**

| | |
|---|---|
| $I$ | Imagem |
| $J$ | Característica |
| $V_i$ | Vetor de características da imagem $i$ |
| $f_{ij}$ | Valor da característica $j$ da imagem $i$ |
| $n$ | Número de características |
| $m$ | Tamanho do banco de dados (número de imagens) |
| $c_{ip}$ | Centróide do cluster $p$ da imagem $i$ |
| $p$ | Cluster |
| $k$ | Número de clusters |
| $V_i'$ | Vetor de características transformado |
| $C_{ip}$ | Cluster $p$ da imagem $i$ |

| | |
|---|---|
| $C_i$ | Conjunto de clusters da imagem *i* |
| *S* | Número de iterações do método |

O método FTK pode ser dividido em três etapas: (1) determinar os centróides iniciais; (2) criação de *clusters*; (3) substituição dos valores do vetor de características pelos centróides. O método FTK (descrito no algoritmo 1) é detalhado a seguir.

**Passo 1**. Determinando os centróide s iniciais (linhas 1-3 do algoritmo 1): inicialmente, o vetor de entrada $V_i = \{f_{i1}, f_{i2}, \ldots f_{in}\}$ é dividido em *k clusters* iniciais $C_{ip}$, $1<p\leq k$, onde *k* é o número de clusters dado como entrada: $C_{ip} = \{f_{i((n/k)*(p-1)+1)}, .. f_{i((n/k)*p)}\}$, onde o conjunto inicial de *clusters* da a imagem *i* é dado por $C_i = \{\cup C_{ip} \mid 0 < p \leq k\}$.

Para a imagem *i*, a média aritmética $c_{ip}$ do cluster inicial $C_{ip}$ é calculada, de acordo com a equação 5.1. A equação 5.1 define o centróide inicial $c_{ip}$ do grupo *p* ($0<p\leq k$) do valor das características da imagem *i*, onde *n* é o tamanho original do vetor e *k* é o número de clusters pré-determinado.

$$c_{ip} = \frac{1}{n/k} \sum_{l=(n/k)*(p-1)+1}^{((n/k)*p)} f_{il} \qquad (5.1)$$

A média aritmética de cada *cluster p* da imagem *i* é usada como centróide inicial $c_{ip}$.

**Passo 2**. Criando os *clusters* (linhas 4-8 do algoritmo 1): dados os centróide s iniciais calculados no passo 1, os *clusters* são processados utilizando um procedimento iterativo baseado no algoritmo *k-means*. Neste passo, o algoritmo atribui o valor $f_{il}$ (para cada imagem *i*) ao *cluster* $C_{ip}$ que está mais próximo do centróide $c_{ip}$ (usando a distância $L_1$) e recalcula os valores dos centróides de cada grupo. Os centróides são recalculados por meio da equação 5.2. Este procedimento é executado iterativamente. Um valor *S* de entrada que limita o número de iterações é utilizado.

$$c_{ip} = \frac{1}{n/k} \sum_{f_{il} \in C_{ip}} f_{il} \qquad (5.2)$$

**Passo 3.** Criação do vetor de características transformado (linhas 9-10, algoritmo 1): Neste passo, o vetor de características de cada imagem é substituído pelo valor dos centróides (*vetor de centróides*).

**Algoritmo 1:** método FTK.

**Entrada:** conjunto de vetores de características do banco de dados de imagens: $V = \{\cup V_i \mid 0 < i < m\}$, onde *m* é o tamanho do banco de dados de imagens e $V_i$ é o vetor de características da imagem *i*; *k*: número de clusters (tamanho do vetor de características transformado); *S*: número de iterações executadas pelo algoritmo.

**Saída:** conjunto de vetores de características transformados: $R = \{\cup V_i' \mid 0 < i < k\}$

1. **Passo 1.** para cada imagem *i*
2. para cada grupo *p* ($0 < p \leq k$)
3. calcular $c_{ip}$ (Equação 5.1)
4. **Passo 2.** para *s=0* até *S*
5. para cada imagem *i*
6. para cada característica *j* ($0 < j \leq n$)
7. atribuir objeto $f_{ij}$ para o centróide mais próximo (usando distância $L_1$)
8. recalcular o centróide de cada cluster $C_{ip}$ (Equação 5.2)
9. **Passo 3.** para cada imagem *i*
10. criar o vetor de características transformado $V_i' = \{ \cup c_{ip}, 0 < p \leq k \}$
11. Retornar o conjunto de vetores de características transformados $R = \{\cup V_i' \mid 0 < i < n\}$

O algoritmo é escalável e eficiente no processamento de grandes conjuntos de dados, uma vez que a complexidade computacional do algoritmo é *O (nmkS)*, onde *n* é o número de características, *m* é o número total de objetos, *k* é o número de *cluster*s a serem criados e *S* é o número de iterações (geralmente, $s \leq k \leq n \leq m$).

## 5.3 Algoritmo WFD

O método desenvolvido WFD (*Weighted Feature Discretization*) é um novo discretizador estatístico e não-supervisionado que faz o pré-processamento de características de imagens, ponderando-as de acordo com seus valores. À medida que o valor da característica se distancia de sua média, seu peso é aumentado de acordo com uma potência, cuja base é fornecida pelo usuário. O objetivo é privilegiar os valores de características mais importantes para distinguir imagens de diferentes classes. Quando os valores de características estão próximos da média, têm maior probabilidade de apresentar um comportamento uniforme ao longo de classes de

imagem diferentes. No entanto, quando o valor da função é distante da média, é mais provável que eles representam um comportamento de uma classe de imagem específica.

A principal motivação é o uso de uma técnica estatística amplamente presente na literatura - *z-score* (LARSON E FARBER, 2010) aliada a cálculos de complexidade lineares, com o objetivo de simplificar as consultas por conteúdo, por meio da discretização e ponderação das características do vetor original.

O algoritmo WFD utiliza uma discretização flexível que leva em conta os valores de média e desvio padrão e a desigualdade de Chebyshev. A desigualdade de Chebyshev estabelece que os valores de uma amostra de dados, em qualquer distribuição, raramente podem estar mais distantes da média do que alguns desvios padrão. Generalizando tem-se que:

pelo menos, $\left(\frac{1}{k^2}.100\%\right)$ da amostra são valores entre a média e os *k* desvios padrão, para $k > 1$.

O método WFD discretiza os dados usando a média e o desvio-padrão dos valores, criando intervalos que são ponderados de acordo com a sua distância da média, beneficiando as faixas de características mais importantes para distinguir as classes de imagens. O método é não-supervisionado e tem como objetivo amplificar os valores extremos de características, ao mesmo tempo que reduz o número de valores diferentes que uma característica pode assumir. O método WFD (descrito no algoritmo 2) é detalhado a seguir. As definições apresentadas na tabela 5.2 são utilizadas no detalhamento do método.

**Tabela 5.2: Definições utilizadas no método WFD.**

| | |
|---|---|
| $I$ | Imagem |
| $j$ | Característica |
| $V_i$ | Vetor de características da imagem *i* |
| $f_{ij}$ | Valor da característica *j* da imagem *i* |
| $n$ | Número de características |
| $m$ | Tamanho do banco de dados (número de imagens) |
| $\mu_j$ | Média aritmética da característica *j* |
| $\sigma_j$ | Desvio padrão da característica *j* |
| $V_i'$ | Vetor de características transformadas |
| $z_{ij}$ | z-score da característica *j* da imagem *i* |
| δ, ω | Funções de mapeamento |

| | |
|---|---|
| $\beta$ | Base da função ω usada para ponderar os intervalos de características |

**Passo1:** O algoritmo WFD calcula o desvio padrão e a média aritmética de cada característica das imagens (linhas 1-7 do algoritmo 2). Em seguida, as características são normalizadas calculando o *z-score* de cada característica (equação 5.3).

$$z_{ij} = \frac{f_{ij} - \mu_j}{\sigma_j} \qquad (5.3)$$

Onde:

$f_{ij}$ valor da característica *j* da imagem *i*

$\mu_j$ média aritmética da característica *j*

$\sigma_j$ Desvio padrão da característica *j*

Em seguida, atribuímos o valor da característica *j* da imagem *i* ($f_{ij}$) ao valor de z-score ($z_{ij}$) correspondente: $f_{ij} = z_{ij}$.

**Passo 2:** Aplicação da discretização ponderada das características (linhas 8-13 do algoritmo 2). Neste passo, para cada imagem *i* e característica *j*, calcula-se as funções de mapeamento $\delta(f_{ij})$ e $\omega(f_{ij})$ de acordo com as equações 5.4 e 5.5. As funções $\delta$ e $\omega$ são funções de mapeamento que têm por objetivo atribuir valores discretos (dentro de intervalos) para cada $f_{ij}$ processada no passo anterior.

$$\delta(f_{ij}) = \begin{cases} f_{ij} \leq & -1.00, & -4 \\ -1.00 < f_{ij} \leq & -0.75, & -3 \\ -0.75 < f_{ij} \leq & -0.50, & -2 \\ -0.50 < f_{ij} \leq & -0.25, & -1 \\ -0.25 < f_{ij} \leq & 0.25, & 0 \\ 0.25 < f_{ij} \leq & 0.50, & 1 \\ 0.50 < f_{ij} \leq & 0.75, & 2 \\ 0.75 < f_{ij} \leq & 1.00, & 3 \\ f_{ij} > & 1.00, & 4 \end{cases} \qquad (5.4)$$

$$\omega(f_{ij}) = \begin{cases} & f_{ij} \leq & -1.00, & 4 \\ -1.00 & < f_{ij} \leq & -0.75, & 3 \\ -0.75 & < f_{ij} \leq & -0.50, & 2 \\ -0.50 & < f_{ij} \leq & -0.25, & 1 \\ -0.25 & < f_{ij} \leq & 0.25, & 0 \\ 0.25 & < f_{ij} \leq & 0.50, & 1 \\ 0.50 & < f_{ij} \leq & 0.75, & 2 \\ 0.75 & < f_{ij} \leq & 1.00, & 3 \\ & f_{ij} > & 1.00, & 4 \end{cases} \quad (5.5)$$

Em seguida, dada a base $\beta$ (fornecida pelo usuário), é realizada a ponderação dos intervalos de características. Neste passo, o vetor de características de cada imagem é substituído pelo valor ponderado e discretizado, calculado pela equação 5.6.

$$\mathrm{f_{ij}} = \delta(\mathrm{f_{ij}}) * \beta^{\omega(\mathrm{f_{ij}})} \qquad (5.6)$$

É retornado o conjunto de vetores de características discretizados e ponderados $\mathrm{R} = \{\cup \mathrm{V_i'} \mid 0 < i < m\}$, onde $V_i' = \{\cup f_{ij} \mid 0 < j < n\}$.

A Figura 5.2 ilustra uma ponderação e discretização dos intervalos realizada pelo método WFD, considerando $\beta = 2$.

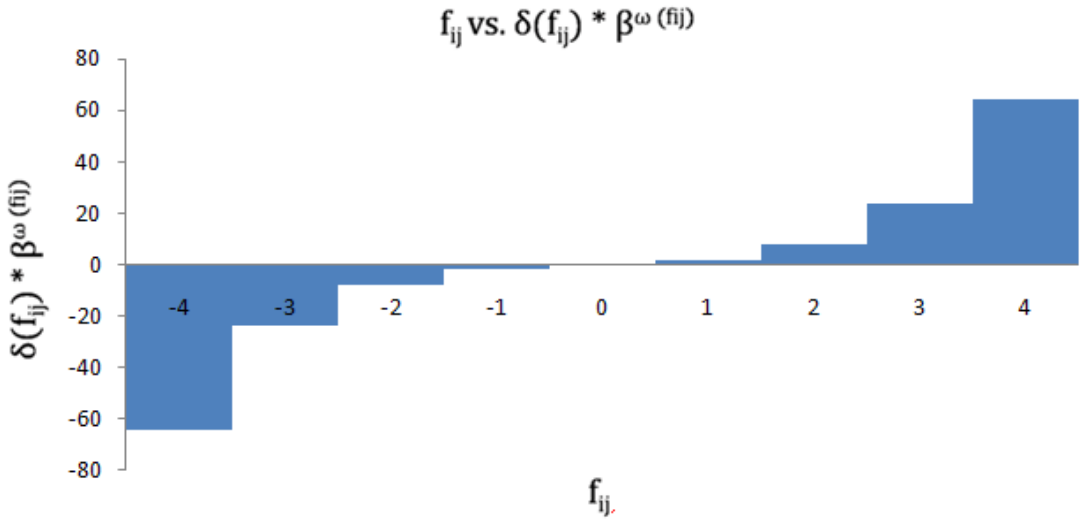



**Figura 5.2: Exemplo de ponderação e discretização dos intervalos realizada pelo método WFD considerando *β* =2.**

O algoritmo 2 implementa o método WFD segundo os passos descritos anteriormente.

**Algoritmo 2:** método WFD.

**Entrada:** conjunto de vetores de características do banco de dados de imagens: $V = \{\cup V_i \mid 0 < i < m\}$, onde *m* é o tamanho do banco de dados de imagens e $V_i$ é o vetor de características da imagem *i*, composta de *n* características; *β:* base da função utilizada para ponderar os intervalos de características.

**Saída:** conjunto dos vetores discretizados e ponderados $R = \{\cup R_i \mid 0 < i < m\}$ , onde $R_i$ é composto de *n* características;

1. **Passo1:** Normalizar os dados usando *z-score* (equação 5.3):
2. Para cada característica *j*, calcular o desvio padrão e a média aritmética (*0 < j < n*)
3. Para cada imagem *i (0 < i < m)*
4. Para cada característica *j* (*0 < j < n*)
5. Calcular $z_{ij}$ (equação 5.3)
6. Atribuir à característica $f_{ij}$ o valor $z_{ij}$ correspondente:
7. $f_{ij} = z_{ij}$
8. **Passo 2:** Aplicando a discretização e ponderação das características.
9. Para cada imagem *i (0 < i < m)*
10. Para cada característica *j* (*0 < j < n*)
11. Calcular as funções de mapeamento $\delta(f_{ij})$ e $\omega(f_{ij})$ (equações 5.4 e 5.5).
12. Atribuir à característica $f_{ij}$ os novos valores discretizados e ponderados
13. $f_{ij} = \delta(f_{ij}) * \beta^{\omega(f_{ij})}$

14. Retornar o conjunto de vetores de características discretizados e ponderados $R = \{\cup V_i^{'} \mid 0 < i < m\}$, onde $V_i^{'} = \{\cup f_{ij} \mid 0 < j < n\}$

O algoritmo é escalável e eficiente no processamento de grandes conjuntos de dados, uma vez que a complexidade computacional do algoritmo é *O (2mn)*, onde *m* é o número total de objetos e *n* é o número de características (geralmente *n* < *m)*. O WFD pode ser usado para o pré-processamento de algoritmos de mineração de dados que necessitam de dados discretos para funcionar, como o algoritmo *Apriori* (AGRAWAL E SRIKANT, 1994) de mineração de regras de associação.

## 5.4 Considerações finais

Neste capítulo foram apresentados os métodos desenvolvidos durante o mestrado – FTK e WFD, suas características, aplicações e motivação. No próximo capítulo serão discutidos os experimentos realizados com tais técnicas, as comparações com métodos existentes, analisando a eficiência e validação dos métodos propostos.

# Capítulo 6

# EXPERIMENTOS

## 6.1 Considerações Iniciais

No capítulo anterior, foram descritos os métodos desenvolvidos neste trabalho de mestrado. Neste capítulo, os métodos propostos foram utilizados para pré-processar os vetores de características nas abordagens CBIR e classificação, comparando os resultados com o pré-processamento executado pelo método PCA e com os vetores de características originais. Os resultados dessas comparações, conforme descrito neste capítulo, indicam que os métodos desenvolvidos são adequados para o pré-processamento de vetores de características de imagens, aumentando a qualidade da resposta de sistemas CBIR e dos algoritmos de classificação.

## 6.2 Avaliação de Consultas em CBIR

Para avaliar a eficiência dos sistemas CBIR é necessário utilizar métodos adequados de avaliação de desempenho de uma operação de consulta. Nesta seção de experimentos utilizamos uma das estratégias mais utilizadas para avaliar a eficácia dos sistemas de busca por conteúdo, que é a análise dos gráficos de

precisão e revocação (*precision vs. recall - P&R*) [Baeza-Yates, 1999]. A seguir são apresentados detalhes da construção de tais gráficos nestes experimentos.

Para uma dada consulta, considere que $R_e$ seja o número total de itens relevantes existentes na base, $R_{eR}$ o número total de itens relevantes recuperados e $I_R$ o total de itens recuperados.

*Revocação* é a fração do conjunto de elementos relevantes ($R_e$) que foram recuperados na consulta:

$$Revocação = \frac{R_{eR}}{R_e} \qquad (6.1)$$

*Precisão* é a fração do conjunto de elementos recuperados ($I_R$) que são relevantes:

$$Precisão = \frac{R_{eR}}{I_R} \qquad (6.2)$$

A Figura 6.1 ilustra os conjuntos de elementos relevantes e recuperados.

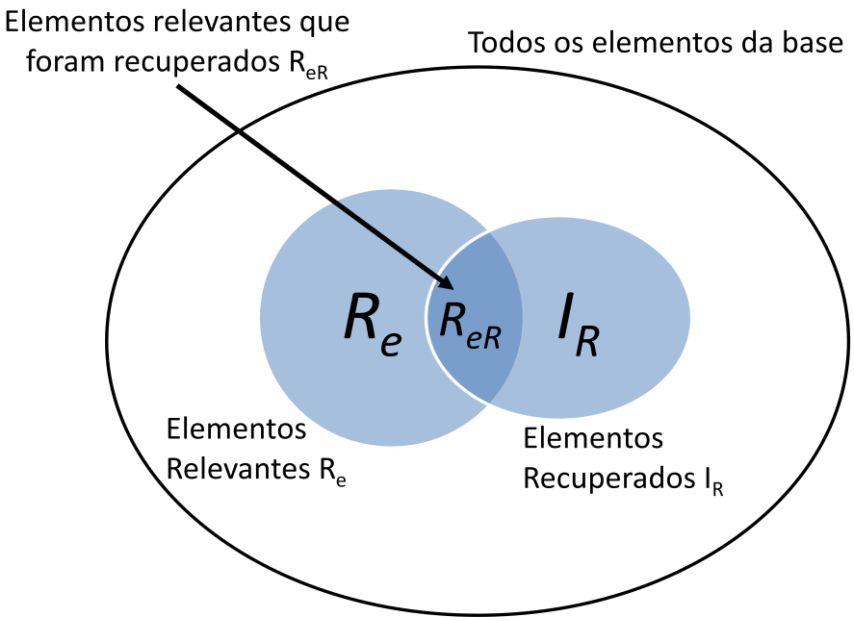



**Figura 6.1: Conjuntos referentes as medidas de precisão e revocação para uma determinada operação de busca. Adaptado de Balan (2007).**

Diversas operações de consultas devem ser realizadas e consideradas para uma avaliação confiável dos resultados obtidos por um determinado sistema de recuperação. Para tanto, foi construída uma curva de precisão e revocação que representa a média dos desempenhos das diversas consultas realizadas. Esta curva

foi construída, calculando-se valores de precisão para escalas determinadas de revocação (a cada intervalo de 5% de revocação).

A avaliação do desempenho é realizada observando-se as curvas obtidas. Quanto mais próxima a curva está do topo do gráfico, melhor é o resultado da operação de consulta. Sendo assim, a curva ideal para uma consulta apresenta 100% de precisão para todos os valores de revocação.

## 6.3 Validação

Foi utilizada a distância Euclidiana como medida de similaridade para comparar vetores de características. A distância euclidiana foi escolhida porque é considerada uma base para validar algoritmos CBIR.

Foram utilizados gráficos de precisão e revocação para analisar a qualidade do ganho da redução de dimensionalidade obtido pelo método FTK e discretização ponderada de características obtida pelo método WFD, quando comparado com o vetor de característica original e o vetor de características obtidos pelo método PCA, empregados para representar as imagens.

Além disso, foram construídos gráficos que comparam a acurácia dos métodos propostos por meio de classificadores OneR (WITTEN, 2005), J48 (QUINLAN, 1993) e Naïve Bayes (JOHN E LANGLEY, 1995) com o objetivo de avaliar a qualidade do pré-processamento dos métodos propostos para a mineração de dados pois estes classificadores são considerados base para a comparação de diversas técnicas. Foram construídos gráficos que comparam o tamanho da árvore de aprendizado gerado pelo classificador J48 a partir dos vetores resultantes dos métodos propostos. O J48 gera uma árvore de decisão: o número de nós representa o tamanho do modelo, no entanto, quanto menor o modelo, mais rápido é o processo de classificação e também maior a interpretabilidade do modelo.

Foi usada a acurácia para comparar os classificadores. A acurácia é também utilizada como uma medida estatística de quão bem um teste de classificação binária identifica corretamente ou exclui uma condição, ou seja, a acurácia é a proporção de resultados verdadeiros (verdadeiros positivos e verdadeiros negativos)

na população corretamente identificados pelo classificador. A acurácia pode ser calculada de acordo com a equação 6.3.

$$acurácia = \frac{n° \ de \ verdadeiros \ positivos + n° \ de \ verdadeiros \ negativos}{n° \ de \ verdadeiros \ positivos + falsos \ positivos + falsos \ negativos + verdadeiros \ negativos} \quad (6.3)$$

As bases de imagens foram submetidas aos métodos FTK, WFD e a combinação dos métodos, chamada "WFD+FTK". Neste último, utilizamos o algoritmo FTK para redução de dimensionalidade dos vetores de características pré-processados por WFD.

Entre várias bases de dados avaliadas, são apresentados aqui os resultados das seis mais significativas. Como parâmetro de entrada do método FTK, foram definidas 4 iterações (*S=4*), pois maximiza o desempenho do método. Para o método WFD, utilizamos $\beta = 2$, pelo mesmo motivo.

Nos experimentos, os métodos FTK, WFD e WFD+FTK são comparados com o método PCA (JOLLIFE, 1986), um método de transformação de características clássico, utilizado como base para a comparação.

## 6.4 Experimento 1

Os métodos propostos foram aplicados a um banco de dados chamado RMI220, que consiste de 220 imagens de ressonância magnética (RMI) obtidas do Hospital da Universidade de São Paulo de Ribeirão Preto, SP, Brasil. As imagens do banco de dados são classificadas em seis categorias, conforme a Tabela 6.1.

**Tabela 6.1: Distribuição das imagens do banco de dados RMI220.**

| Categoria da imagem | Número de imagens |
|---|---|
| Angiograma | 21 |
| Pélvis Axial | 33 |
| Cabeça Axial | 50 |
| Abdômen Coronal | 34 |
| Cabeça Sagital | 38 |
| Espinha Sagital | 44 |

Foram utilizados histogramas de níveis de cinza para representar as imagens do banco de dados RMI. Histogramas são amplamente utilizados em sistemas CBIR porque eles têm baixo custo computacional e são invariantes para as operações de translação e rotação de imagens. A fim de tornar o histograma também invariante para as operações de escala, o histograma foi normalizado. A normalização do histograma foi realizada por divisão de cada *bin* do histograma (característica) pelo valor do maior *bin* atingido para cada imagem, gerando histogramas em que o valor máximo de cada *bin* é 1.

O vetor de características originais (histograma padrão), composto de 256 características, foi submetido ao método FTK, usando o parâmetro de entrada $k = 8$ (número de *clusters* a serem gerados). O método FTK produziu um novo vetor (transformado) de 8 posições (características).

Foram realizadas várias consultas do tipo $k_{NN}$, empregando o vetor de característica original para representar as imagens do banco de dados RMI e o vetor de características produzido pelo método FTK. Os gráficos de precisão e revocação (P&R) comparando os métodos são mostrados na Figura 6.2.

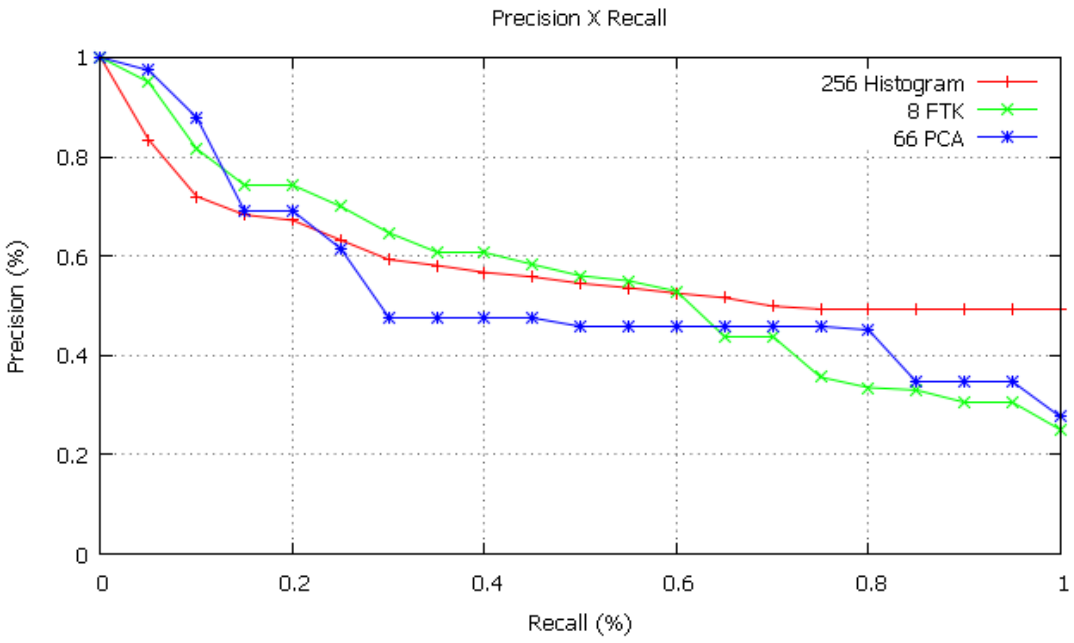



**Figura 6.2: gráfico de P&R comparando os resultados do uso do vetor de característica original (histograma - 256 características,-+-, vermelho), o vetor de característica resultante do método FTK (8 características, -x-, verde) e o vetor de características obtido pelo método PCA (66 características, -*-, azul), para representar as imagens do banco de dados RMI220.**

O gráfico da Figura 6.2 mostra que o vetor resultante do método FTK produz valores mais precisos quando comparado ao vetor de características original, aumentando a precisão em 8% e aumentando em 35% quando comparado ao método PCA, para uma revocação de 30%. Além disso, o método FTK reduziu o tamanho do vetor de característica original em aproximadamente 97%, acelerando o processo de recuperação.

A Tabela 6.2 e a Tabela 6.3 mostram exemplo de consultas $k_{NN}$ geradas a partir da base RMI usando a imagem de consulta ilustrada na Figura 6.3. A Tabela 6.2 mostra o resultado obtido utilizando o vetor característica original (256 características) e a Tabela 6.3 mostra o resultado obtido empregando o vetor característico resultante do método FTK (8 características).

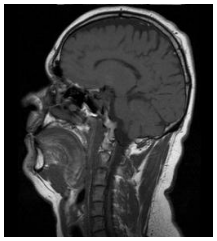


**Figura 6.3. Imagem de consulta utilizada no primeiro experimento.**

**Tabela 6.2: Exemplo dos resultados de uma consulta $K_{NN}$ = 5 a partir do vetor de características original (histograma) representando as imagens da base RMI. A distância descrita é a partir da imagem de consulta da Figura 6.3.**

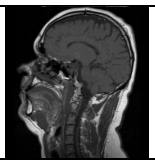
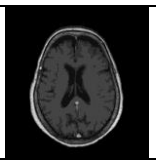
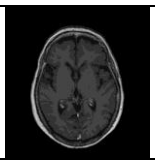
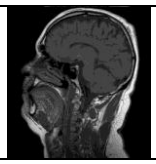
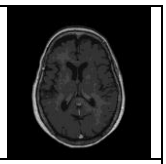
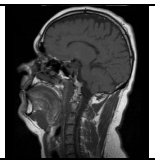
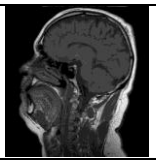
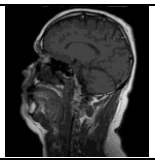
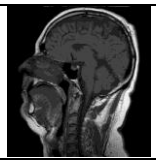
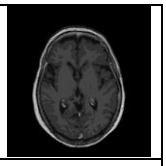
| Imagem |  |  |  |  |  |
|---|---|---|---|---|---|
| Distância | 0,0 | 2,1413 | 2,1904 | 2,2944 | 2,3597 |

**Tabela 6.3: Exemplo dos resultados de uma consulta $K_{NN}$ = 5 a partir do vetor de características resultante do método FTK para representar as imagens da base RMI. A distância descrita é a partir da imagem de consulta da Figura 6.3.**

| Imagem |  |  |  |  |  |
|---|---|---|---|---|---|
| Distância | 0,0 | 0,0732 | 0,0891 | 0,1706 | 0,1773 |

Os resultados da Tabela 6.2 (utilizando o vetor de características original) e da Tabela 6.3 (usando o vetor característico resultante do método FTK) mostram uma melhoria importante em termos de precisão usando o método FTK. Além disso, há uma redução no intervalo de valores de distância, utilizando o método proposto. O método FTK tende a eliminar os valores de ruído que interferem negativamente no cálculo da distância. Esta remoção de ruído aproxima imagens de mesma classe.

## 6.5 Experimento 2

Os métodos propostos foram aplicados à base de dados MIAS (*Mammographic Image Analysis Society*), do Departamento de Física do Hospital Royal Marsden (SUCKLING, 1994). A base de dados é composta de 322 mamografias. As imagens da base de dados são classificadas em sete categorias (calcificação, bem definidas / massas circunscritas, massas espiculadas; outros / massas mal definidas, distorção arquitetural, assimetria e normal).

Utilizamos o histograma de níveis de cinza para representar as imagens do banco de dados. Para tornar o histograma também invariante para as operações de escala, o mesmo foi normalizado. O vetor de características original (o histograma padrão), composto de 256 características, foi submetido ao método FTK, usando o parâmetro de entrada *k = 4* (número de *clusters* a serem gerados). O método FTK produziu um vetor transformado composto de 4 características. O vetor original também foi submetido ao método PCA, que produziu um vetor transformado composto de 24 características. O vetor original também foi submetido ao método WFD, que produziu um vetor composto de 256 características. Os gráficos de precisão *vs.* revocação comparando o vetor original e os vetores características resultantes são mostrados na Figura 6.4.

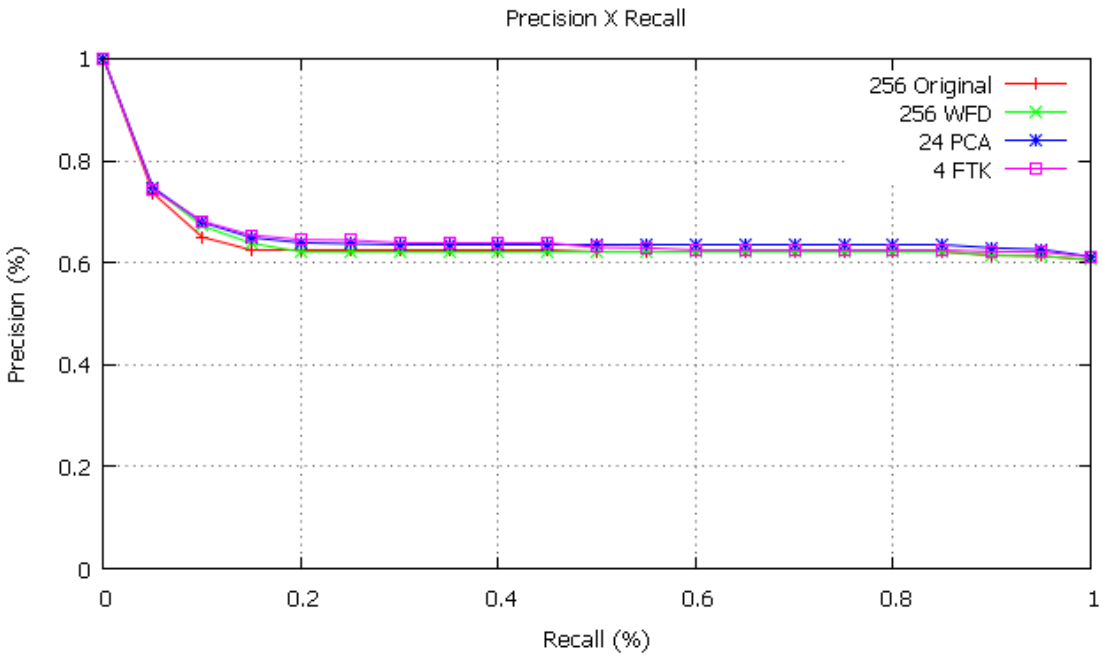



**Figura 6.4: gráficos de P&R comparando os resultados do uso do vetor de característica original (histograma - 256 características,-x-, vermelho), o vetor de característica resultante do método WFD (256 características, -x-, verde), o vetor de característica resultante do método FTK (4 características, -o-, rosa) e o vetor de características obtido pelo método PCA (24 características, -*-, azul) para representar as imagens do banco de dados MIAS.**

O gráfico da Figura 6.4 mostra que o vetor de características resultante do método FTK aumenta ligeiramente os valores de precisão para a região entre 10% e 15%. Além disso, o vetor de características do método WFD aumenta ligeiramente os valores de precisão para a região de 10%. Este é um resultado importante, uma vez que esse ganho foi conseguido em regiões de baixo valor de *recall* que são as mais utilizadas de fato em uma consulta, pois o usuário em cada consulta geralmente recupera um pequeno número de imagens da base de dados. Além disso, o método FTK reduziu o tamanho do vetor de características original em cerca de 98%, promovendo uma redução significativa da dimensionalidade do vetor de características. O gráfico Precisão vs. Classificadores comparando o vetor de características original e os vetores de características transformados são ilutrados na Figura 6.5.

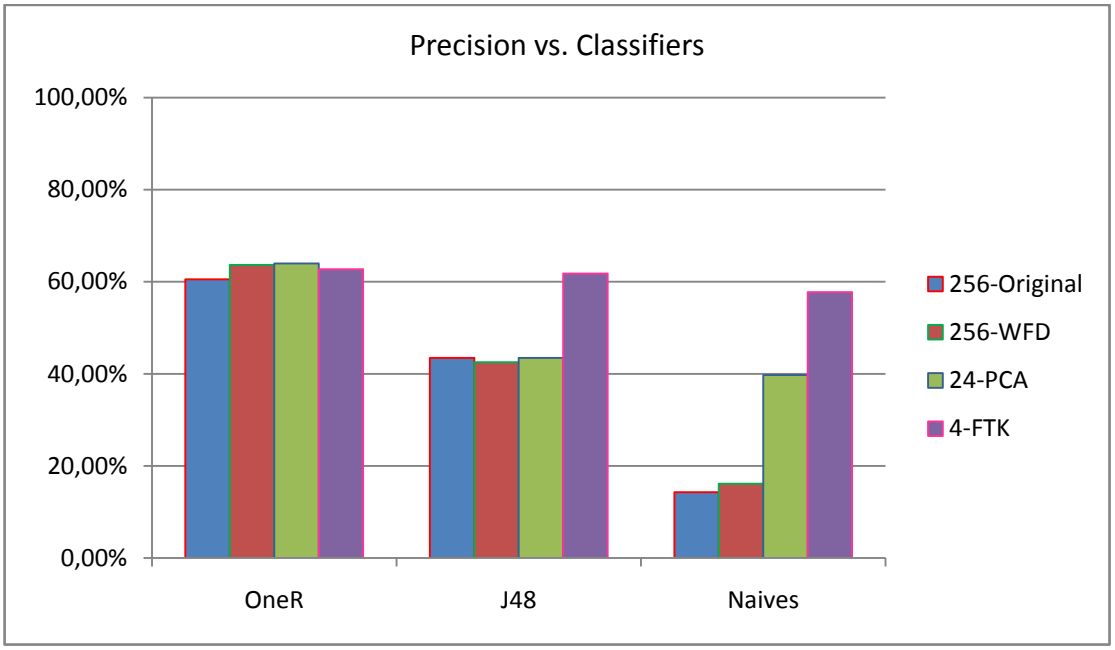



**Figura 6.5 gráficos de Precisão vs. Classificadores comparando os resultados do uso do vetor de característica original (histograma - 256 características - azul), o vetor de característica resultante do método WFD (256 características - vermelho), o vetor de característica resultante do método FTK (4 características - roxo) e o vetor de características obtido pelo método PCA (24 características - verde) para representar as imagens do banco de dados MIAS.**

Os gráficos da Figura 6.5 mostram que o vetor de características resultante do método FTK produz valores mais elevados de precisão em comparação com o vetor de características originais e vetor de características do método PCA, aumentando a precisão até 37% para o classificador J48. Além disso, o vetor de características FTK produz valores mais elevados de precisão em comparação com o vetor de características original, aumentando a precisão em até 200%; e, em até 45%, quando comparado com o método PCA, para o classificador Naives.

## 6.6 Experimento 3

O método proposto foi aplicado à base dados LungCancer obtida do repósitório *UCI Machine Learning Repository* (FRANK E ASUNCION, 2010). A base de dados é composta por 32 imagens de três tipos de câncer de pulmão.

O vetor de características original é composto de 56 características. Este vetor foi submetido ao método FTK, usando o parâmetro de entrada $k = 32$ (número

de *clusters* a serem gerados). O método FTK produziu um vetor transformado composto de 32 características. O vetor original também foi submetido ao método PCA, que produziu um vetor transformado composto de 21 características. O vetor original também foi submetido ao método WFD, que produziu um vetor composto de 56 características. O vetor original também foi submetido aos métodos combinados WFD+FTK, que produziu um vetor composto de 32 características. Os gráficos de precisão *vs.* revocação comparando o vetor original e os vetores transformados são mostrados na Figura 6.6.

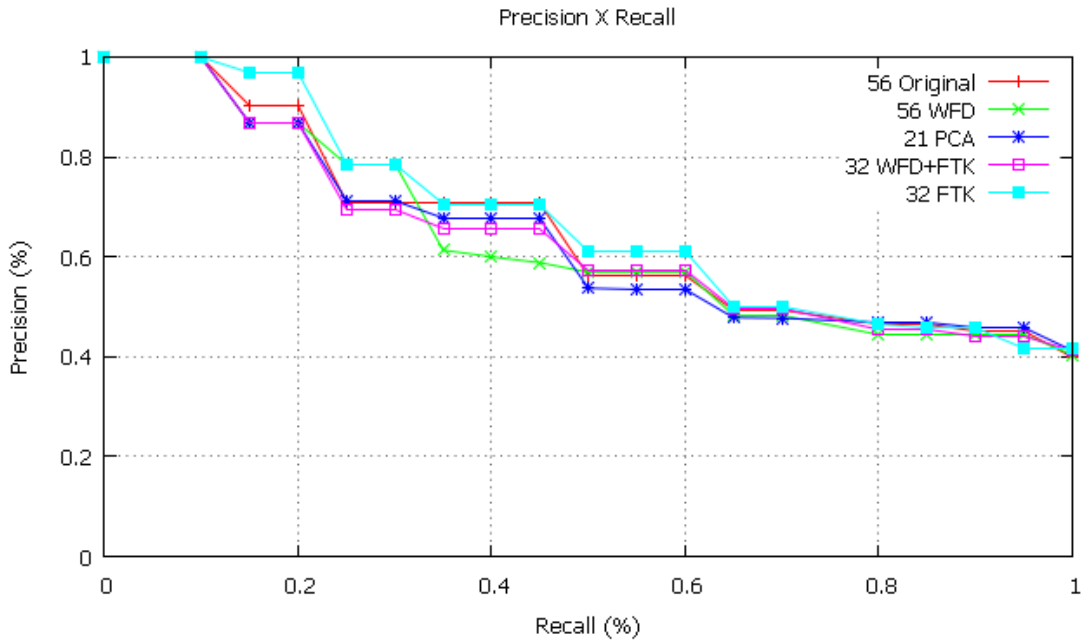



**Figura 6.6: gráficos de P&R comparando os resultados do uso do vetor de característica original (56 características,-+-, vermelho), o vetor de característica resultante do método WFD (32 características, -x-, verde), o vetor de característica resultante do método PCA (21 características, -*-, azul), o vetor de característica resultante dos métodos WFD e FTK combinados (32 características, -o-, rosa) e o vetor de característica resultante do método FTK (32 características, -o-, azul claro) para representar as imagens do banco de dados LungCancer.**

O gráfico da Figura 6.6 mostra que o vetor resultante dos métodos FTK e WFD aumenta a precisão da consulta em 9% na região de revocação de 30%, quando comparado ao vetor original e ao vetor resultante do método PCA. Além disso, os métodos FTK e WFD+FTK reduziram o tamanho do vetor de características original em aproximadamente 43%, acelerando o processo de recuperação. O gráfico Precisão vs. Classificadores, comparando o vetor de características original e os vetores de características transformados, é ilutrados na Figura 6.7.

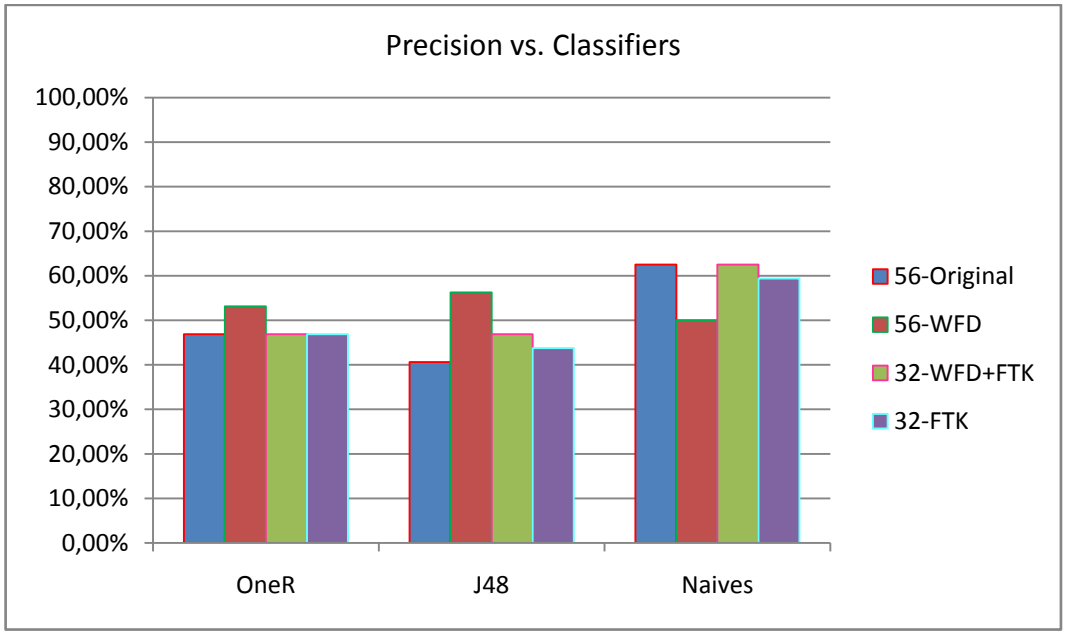



**Figura 6.7: gráficos de Precisão vs. classificadores comparando os resultados do uso do vetor de característica original (56 características, azul), o vetor de característica resultante do método WFD (32 características, vermelho), o vetor de característica resultante dos métodos WFD e FTK combinados (32 características, - verde) e o vetor de característica resultante do método FTK (32 características, - roxo) para representar as imagens do banco de dados LungCancer.**

Os gráficos da Figura 6.7 mostram que o vetor de características resultante do método WFD produz valores mais elevados de precisão em comparação com o vetor de características original, aumentando a precisão até 37% para o classificador J48. Além disso, o vetor de características WFD produz valores mais elevados de precisão em comparação com o vetor de características original, aumentando a precisão em mais de 15%, para o classificador Naives.

O vetor de características resultante do método WFD+FTK produz valores mais elevados de precisão em comparação com o vetor de características original, para os classificadores J48 e Naives. O gráfico da Figura 6.8 compara o tamanho da árvore do aprendizado gerado pelo classificador J48 a partir dos vetores resultantes dos métodos propostos.

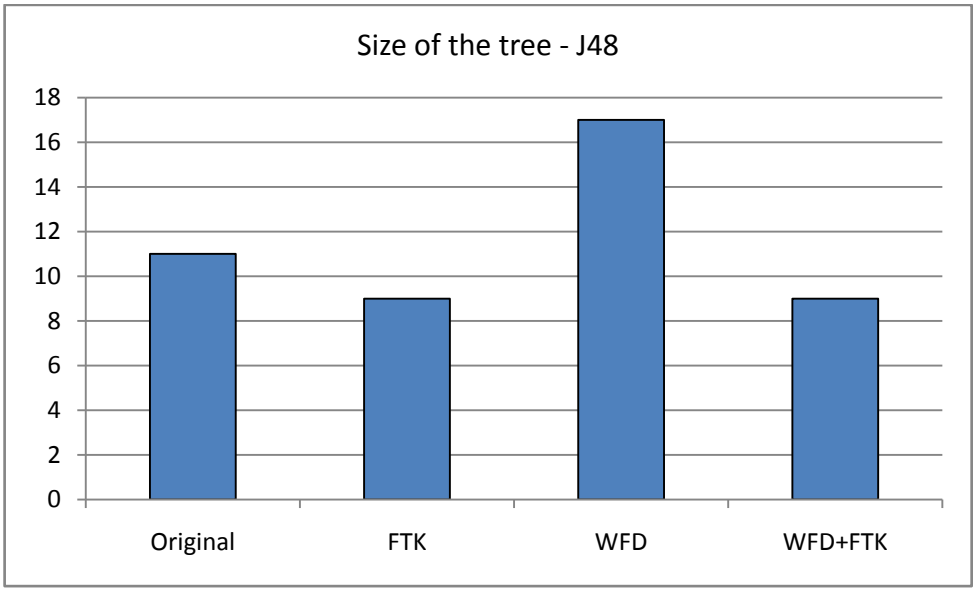



**Figura 6.8: Gráficos de precisão vs. classificadores comparando o tamanho da árvore de aprendizagem gerado pelo classificador J48 a partir dos vetores original, FTK, WFD e WFD+FTK para representar as imagens de banco de dados Lung Cancer.**

O gráfico da Figura 6.8 mostra que a dimensão da árvore de aprendizagem gerada pelo classificador J48 reduz de forma significativa, quando comparado com o vetor de características original e o vetor de características gerado pelo FTK. Comparando este gráfico com o gráfico P&R, concluímos que podemos diminuir o modelo de aprendizagem (tamanho da árvore), aumentando a precisão de busca de imagens similares.

## 6.7 Experimento 4

O método proposto foi aplicado a 240 imagens coloridas da base de dados *Amsterdam Library of Object Images* (ALOI). A base de dados é composta de imagens de objetos divididas em dez categorias, variando o ângulo de visão, o ângulo de iluminação e a cor de iluminação para cada objeto. Detalhes podem ser obtidos em (GEUSEBROEK, 2005). Foi utilizado o histograma de níveis de cinza para representar as imagens do banco de dados.

O vetor de características original (histograma composto de 256 características – níveis) foi submetido ao método FTK, usando o parâmetro de entrada *k = 128* (número de *clusters* a serem gerados). O método FTK produziu um

vetor transformado composto de 128 características. O vetor original também foi submetido ao método WFD, que produziu um vetor transformado composto de 256 características. O vetor original também foi submetido aos métodos WFD e FTK combinados, que produziu um vetor composto de 128 características. Os gráficos de precisão *vs.* revocação comparando o vetor original e os vetores transformados são mostrados na Figura 6.9.

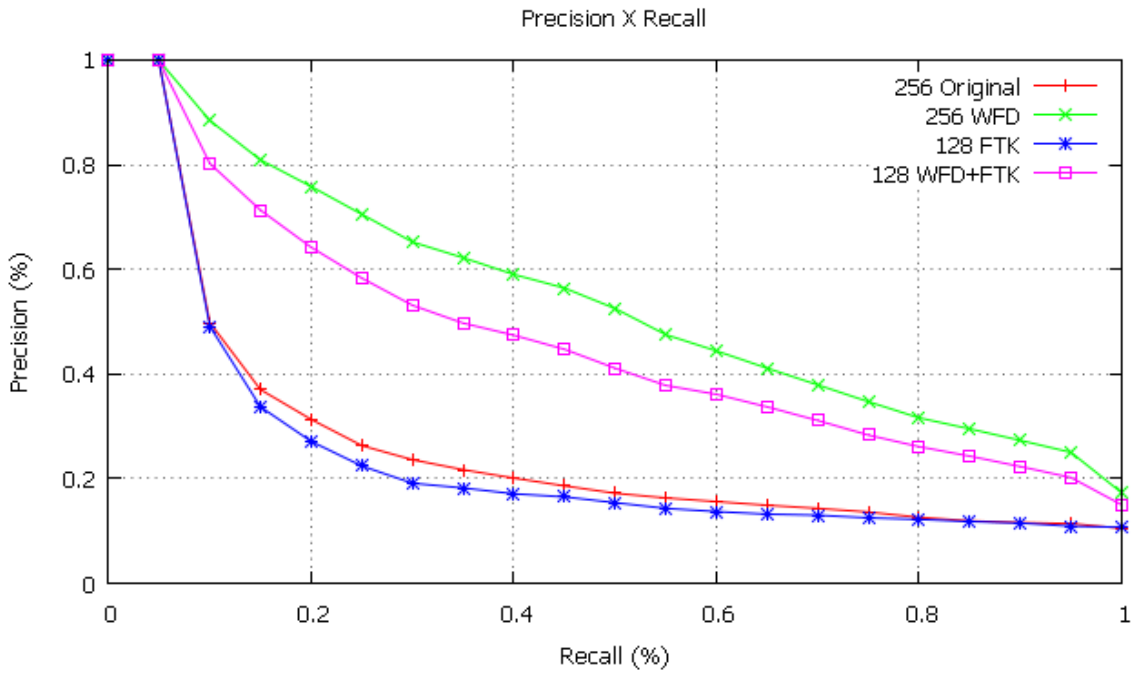



**Figura 6.9: gráficos de P&R comparando os resultados do uso do vetor de característica original (256 características,-+-, vermelho), o vetor de característica resultante do método WFD (256 características, -x-, verde), o vetor de característica resultante do método FTK (128 características, -*-, azul) e o vetor de característica resultante dos métodos WFD e FTK combinados (128 características, -o-, rosa) para representar as imagens da base de dados ALOI.**

O gráfico da Figura 6.9 mostra que o vetor resultante do método WFD aumenta a precisão da consulta em 140% na região de revocação de 20%, quando comparados ao vetor original. Além disso, os métodos WFD e FTK combinados aumentaram a precisão da consulta em 100% na mesma região de revocação (20%), quando comparados ao vetor original e reduziram o tamanho do vetor de característica original em aproximadamente 50%, acelerando o processo de recuperação.

## 6.8 Experimento 5

Os métodos propostos foram aplicados à base de dados chamada *Wisconsin Breast Cancer Database (WBCD)*, que consiste de 569 imagens médicas obtidas do *University of Wisconsin Hospitals, Madison, Wisconsin, USA*. As imagens do banco de dados são classificadas em duas classes: benignas (357 imagens) ou malignas (212 imagens). Pode-se obter mais detalhes por (WOLBERG E MANGASARIAN, 1990).

O vetor de características original composto de 30 características, foi submetido ao método FTK, usando o parâmetro de entrada *k = 4* (número de *clusters* a serem gerados). O método FTK produziu um vetor transformado composto de 4 características. Os gráficos de precisão *vs.* revocação comparando o vetor original e os vetores transformados são mostrados na Figura 6.10.

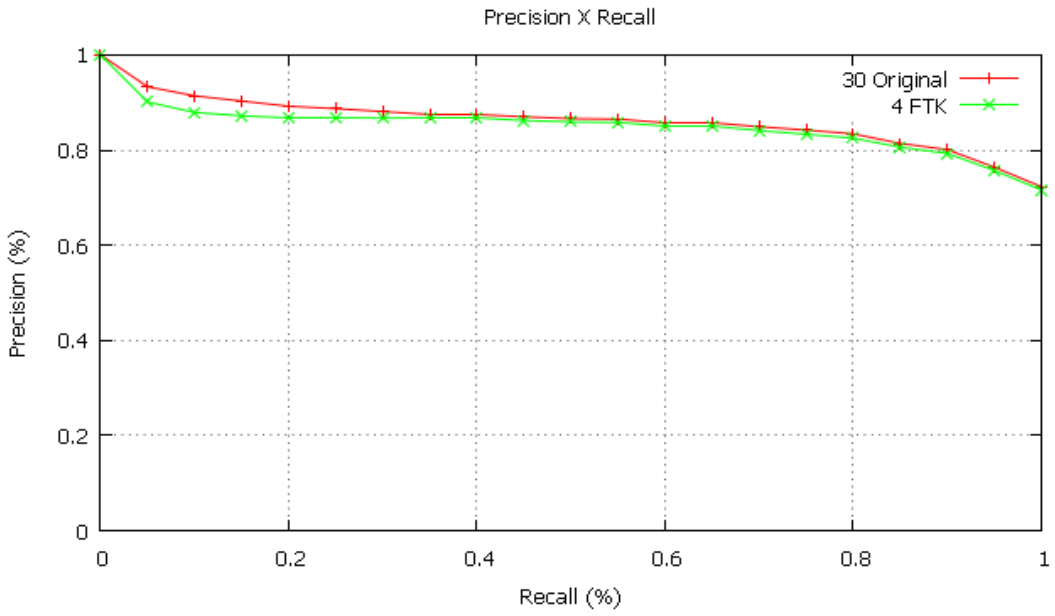



**Figura 6.10: gráficos de P&R comparando os resultados do uso do vetor de característica original (30 características,-+-, vermelho) e o vetor de características resultante do método FTK (4 características, -x-, verde) para representar as imagens do banco de dados WBCD.**

O gráfico da Figura 6.10 mostra que método FTK reduziu o tamanho do vetor de característica original em aproximadamente 87%, promovendo uma considerável redução de dimensionalidade do vetor de características, não afetando em muito a precisão das consultas.

## 6.9 Experimento 6

Os métodos propostos foram aplicados a um banco de dados chamado RMI704, que consiste de 704 imagens de ressonância magnética (RMI), obtidas do Hospital da Universidade de São Paulo de Ribeirão Preto, SP, Brasil. As imagens do banco de dados são classificadas em oito categorias, conforme a Tabela 6.4.

**Tabela 6.4: Distribuição das imagens do banco de dados RMI704.**

| Categoria | Número de imagens |
|---|---|
| Angiograma | 36 |
| Pélvis Axial | 86 |
| Cabeça Axial | 155 |
| Cabeça Sagital | 258 |
| Abdômen Coronal | 23 |
| Espinha Sagital | 59 |
| Abdômen Axial | 51 |
| Cabeça Coronal | 36 |

A base de dados foi previamente segmentada pelo método proposto por Balan (2007). A segmentação da base foi realizada considerando cinco regiões de segmentação, que é a configuração que produz os maiores valores de precisão para esta base. Para cada região segmentada, seis características foram extraídas: a massa *m,* ou tamanho; o centro de massa, ou centróide $(x_c;y_c)$; o nível de cinza médio *a*; a dimensão fractal (dimensão de correlação) $D_2$ e o coeficiente linear *b* usado para estimar $D_2$. Assim, quando a imagem é segmentada em 5 classes, o vetor de características tem 5*6 = 30 elementos. Os gráficos de precisão *vs.* Revocação, comparando o vetor original e os vetores transformados, são mostrados na Figura 6.11.

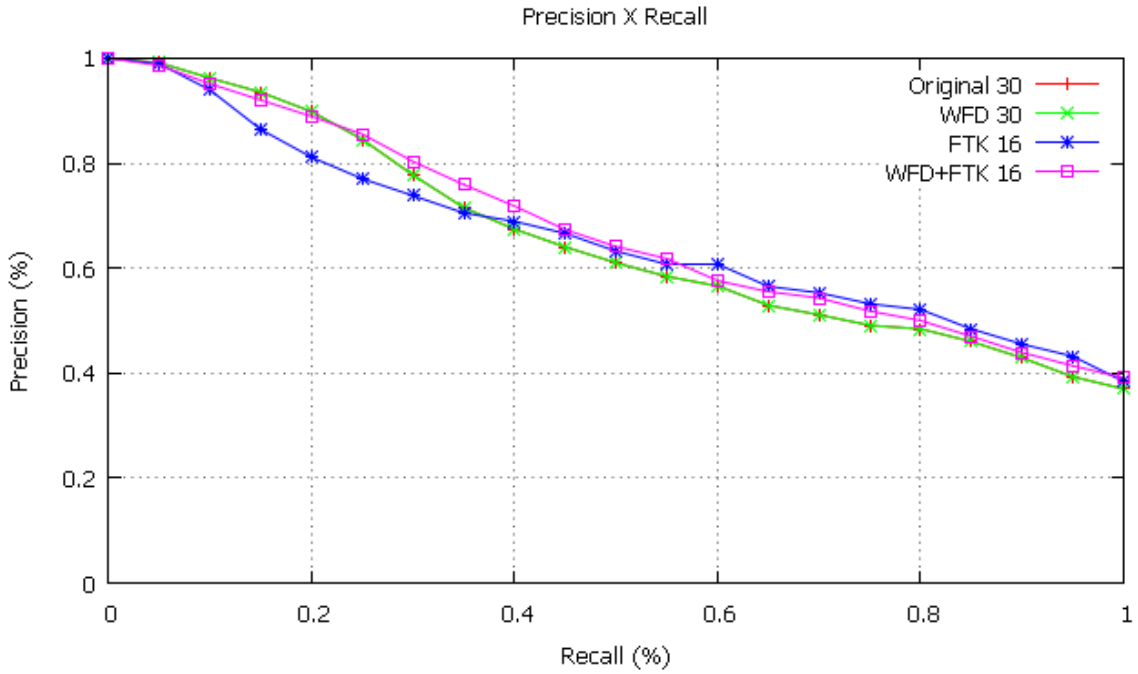



**Figura 6.11: gráficos de P&R comparando os resultados do uso do vetor de característica original (30 características,-+-, vermelho), o vetor de característica resultante do método WFD (30 características, -x-, verde), o vetor de característica resultante do método FTK (16 características, -*-, azul) e o vetor de característica resultante dos métodos WFD e FTK combinados (16 características, -o-, rosa) para representar as imagens do banco de dados RMI704.**

O gráfico da Figura 6.11 mostra que o vetor de características resultante da combinação dos métodos WFD+FTK aumenta ligeiramente os valores de precisão para a região entre 30% e 35%. Além disso, os métodos FTK e WFD+FTK reduziram o tamanho do vetor de características original em cerca de 47%, promovendo uma redução significativa da dimensionalidade do vetor de características.

## 6.10 Considerações finais

Neste capítulo, apresentamos os experimentos realizados com os algoritmos propostos neste trabalho. Comparamos o tamanho da árvore de aprendizado para verificar o efeito das técnicas propostas no modelo de aprendizado, e verificou-se que elas produzem um modelo de aprendizado mais compacto, concluindo-se, desta forma, que o vetor de características transformado (resultante dos métodos

propostos) é compacto e representativo, o que leva à interpretação de mais alto nível dos dados.

De um modo geral, o algoritmo FTK promoveu uma redução considerável no tamanho do vetor de características e uma melhoria na precisão da consulta e na acurácia de classificação. Em diversos experimentos, o algoritmo WFD também promoveu uma melhora na precisão da consulta e na acurácia da classificação; a combinação dos dois algoritmos também promoveu a melhoria de qualidade das consultas e classificação. Tais resultados são muito importantes, especialmente quando comparados com os resultados do método PCA, o que leva a uma menor redução no tamanho do vetor de características, a um menor aumento na precisão da consulta e a um menor aumento na acurácia na classificação. Além disso, as técnicas propostas têm custo computacional linear enquanto PCA tem um custo computacional cúbico. Os resultados indicam que os métodos propostos são abordagens adequadas para se realizar pré-processamento dos vetores de características de imagens em sistemas CBIR e sistemas de classificação.

# CONCLUSÃO

O método FTK emprega análise de *cluster* em valores de características para promover a transformação de características. O processo de análise de agrupamento altera os valores das características de imagem por valores centróide s, tendendo a eliminar o ruído que interfere negativamente no processo de consulta. A transformação de características, realizada pela análise de agrupamento, funciona como um filtro que reduz as diferenças entre imagens semelhantes. O método WFD realiza uma discretização ponderada de características, privilegiando as faixas de características mais importantes para distinguir imagens.

De um modo geral, no melhor caso, FTK produziu uma redução no tamanho do vetor de características, com um aperfeiçoamento da precisão de consulta e uma melhoria da acurácia da classificação; WFD melhorou a precisão da consulta e a acurácia da classificação. Já a combinação de WFD e FTK melhorou a precisão da consulta e a acurácia da classificação. Estes resultados são muito importantes, especialmente quando comparados com os resultados do método PCA, que, em geral, leva a menores ganhos. Além disso, as abordagens propostas têm custo computacional linear, enquanto o PCA tem um custo computacional cúbico.

As experiências mostram que os métodos FTK e WFD aumentam a precisão de consultas CBIR, ao mesmo tempo em que promovem uma redução significativa no tamanho do vetor de características, melhorando a qualidade e eficiência dos sistemas CBIR. Os resultados indicam que as abordagens propostas são bastante apropriadas para realizar pré-processamento do vetor de características de imagens, melhorando a qualidade dos sistemas CBIR e dos sistemas de classificação.

Os principais direcionamentos no futuro para este trabalho de pesquisa são realizar testes com outras bases, comparar com outras técnicas, usar o WFD para pré-processamento de algoritmos de mineração de dados (associação, regressão, etc.).

# REFERÊNCIAS


ABRAHAM, R. (et al). A comparative analysis of discretization methods for medical datamining with naive bayesian classifier. In: 9TH INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY. 2006. p. 235–6.*

AGRAWAL, R. E SRIKANT, R. Fast algorithms for mining association rules. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES (VLDB). 1994. Santiago de Chile, 1994. p. 487–499.

AN, Y. (et al). Classification of Feature Set using K-means Clustering from Histogram Refinement Method. In: FOURTH INTERNATIONAL CONFERENCE ON NETWORKED COMPUTING AND ADVANCED INFORMATION MANAGEMENT. 2008. p.320-4.

BAEZA-YATES, R. A. E RIBEIRO-NETO, B. A. **Modern Information Retrieval**. New York: ACM press, 1999.

BALAN, A. G. R. **Métodos adaptativos de segmentação aplicados à recuperação de imagens por conteúdo**. São Carlos, 2007, p. 183. Originalmente apresentada como teste de doutorado, Universidade de São Paulo, 2007.

Barshan, E. (et al). Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. p.1357-1371. v. 44, 2011. **Pattern Recognition Society - Elsevier Science Ltd.**

COSTA, L. D. F. Shape Analysis and Classification - Theory and Pratice. Boca Raton - Florida: CRC Press LLC, 2001.

DESERNO, T. M. (et al). Ontology of Gaps in Content-Based Image Retrieval. **Journal of Digital Imaging**. p.1-14, 2008.

DIAS, C. R. E OCHI, L.S. Desenvolvimento e Análise Experimental de Algoritmos Evolutivos para o problema de clusterização automática em grafos orientados. Niterói, 2004.

FAYYAD, U. M., (et al). Advances in Knowledge Discovery and Data Mining. Cambridge: MIT Press, 1996.

FELIPE, J. C. **Desenvolvimento de métodos para extração, comparação e análise de características intrínsecas de imagens médicas, visando à recuperação perceptual por conteúdo.** São Carlos, 2005, p. 176. Originalmente apresentada como teste de doutorado, Universidade de São Paulo, 2005.

FERREIRA, A. J. E FIGUEIREDO, M. A. T. An unsupervised approach to feature discretization and selection. **Pattern Recognition Society - Elsevier Science**

**Ltd**, v.45, 2011, p.3048-3060.

FILARDI, A. L. E TRAINA, A. J. M. Avaliação de Interface de Sistemas de Recuperação de Imagens Médicas por Conteúdo. In: IV WORKSHOP DE VISÃO COMPUTACIONAL - UNESP – Bauru. 2008. p. 6.

FISHER, R. A. **The use of multiple measurements in taxonomic problems, Annals of Eugenics**, v. 7. 1936. p. 179–188.

FRANK, A. E ASUNCION, A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2010.

FRAWLEY, W. J. (et al). Knowledge discovery in databases: an overview. AAAI/MIT Press. 1991. p.1-27.

GONZALEZ, R. C. E WOODS, R. E. **Processamento de Imagens Digitais**. São Paulo: Blucher, 2000.

GEUSEBROEK, J. M. (et al). The Amsterdam library of object images (ALOI). **International Journal of Computer Vision**, v.61, 2005. p.103-112.

GUTTMAN, A. R-trees: a dynamic index structure for spatial searching. **Nacional Science Foundation**. ACM, 1984.

HAND, D. (et al) **Principles of Data Mining**. Cambridge, Massachusetts London, England: Massachusetts Institute of Technology Press. 2001.

HARALICK, R. M. (et al). Textural features for image classification. **IEEE Transactions on Systems, Man and Cybernetics**, v.3, 1973. p.610-621.

JAIN, A. K., MURTY, M. N. (et al). Data Clustering: A Review. **ACM Computing Surveys**, v.31, 1999. p.60.

JAIN, A. K., DUIN, R. P. W. (et al). Statistical Pattern Recognition: A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v.22, 2000. p. 4-37.

JIN, C. H. (et al). Classification of Closed Frequent Patterns Improved by Feature Space Transformation. **IEEE International Conference on Computer and Information Technology**, v.10, 2010. p.1306-1311.

JEONG, S. (et al). Dimensionality reduction in high-dimensional space for multimedia information retrieval. In: 18TH INTERNATIONAL CONFERENCE ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, 2007. p. 404-413.

JOHN, G. H. E LANGLEY, P. **Estimating continuous distributions in bayesian classifiers.** San Mateo: Morgan Kaufmann. 1995. p. 338–345.

JOLLIFFE, I. T. **Principal Component Analysis**. New York: Springer - Verlag, 1986.

KURGAN, L. A. E CIOS, K. J. Caim discretization algorithm. **IEEE Transactions on Knowledge and Data Engineering (TKDE)**, v. 16. 2004. p. 145–153.

LARSON, R. E FARBER, B. **Estatística aplicada.** São Paulo: Pearson, 2010. 4ª ed.

LIU, H., HUSSAIN, F. (et al). Discretization: an enabling technique. **Data Mining and Knowledge Discovery**, v. 6, 2002. p. 393–423.

LIU, H. E YU, L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. **IEEE Transactions on Knowledge and Data Engineering (TKDE)**, v.17, 2005. p.491-502.

LIU, Y., ZHANG, D. (et al). A survey of content-based image retrieval with high-level semantics. **Elsevier - Pattern Recognition**, v.40, 2007. p.262 – 282.

LOOG, M., GINNEKEN, B. V. (et al). Dimensionality reduction of image features using the canonical contextual correlation projection. **Elsevier - Pattern Recognition**, v.38, 2005. p.2409 – 2418.

MARQUES FILHO, O. E VIEIRA NETO, H. **Processamento Digital de Imagens**. Rio de Janeiro: Brasport Livros e Multimidia LTDA. 1999. p. 406.

POONGUZHALI, S. (et al). Optimal feature selection and automatic classification of abnormal masses in ultrasound liver images. In: INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING, COMMUNICATIONS AND NETWORKING, 2007. p.503–6.

QUINLAN, J. R. **C4.5: programs for machine learning.** San Francisco, CA: Morgan Kaufmann Publishers Inc., 1993.

RIBEIRO, M. X. **Suporte a sistemas de auxílio ao diagnóstico e de recuperação de imagens por conteúdo usando mineração de regras de associação.** São Carlos, 2009. p. 202. Originalmente apresentada como teste de doutorado, Universidade de São Paulo, 2009.

SAMET, H. The Quadtree and Related Hierarchical Data Structures. **ACM Computing Surveys**, v.16, 1984. p.74.

SCHOWENGERDT, R. A. **Spectral Transforms in Remote Sensing: Models and Methods**. London: Academic Press. 1997. p. 522.

SILVA, M. P. D. **Processamento de Consultas por Similaridade em Imagens Médicas Visando à Recuperação Perceptual Guiada pelo Usuário**. São Carlos, 2008, p. 55. Originalmente apresentada como dissertação de mestrado, Universidade de São Paulo, 2008.

SILVA, S. F. D., RIBEIRO, M. X. (et al). Improving the ranking quality of medical image retrieval using a genetic feature selection method. **Elsevier - Decision Support Systems**, v.51, 2011. p.810-820.

SUCKLING, J. (et al). The Mammographic Image Analysis Society Digital Mammogram Database. **Excerta Medica.** INTERNATIONAL CONGRESS SERIES, v.1069, 1994. p.375-8. [http://peipa.essex.ac.uk/ipa/pix/mias/]

TAN, P. N. (et al). **Introduction to Data Mining**. Boston: Pearson Addison-Wesley. 2006. p. 487-569.

TRAINA JR., C. (et al). **Slim-trees: High Performance Metric Trees Minimizing Overlap Between Nodes**. In: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY (EDBT), 2000. p.27-31.

WOLBERG, W. H. E MANGASARIAN, O. L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. **Proceedings of the National Academy of Sciences**. U.S.A., v. 87, December 1990. p 9193-6.

WITTEN, I. H. E FRANK, E. **Data mining: practical machine learning tools and techniques.** San Francisco: Morgan Kaufmann, 2005. p. 85-90.

XU, R. E Wunsch-II, D. Survey of Clustering Algorithms. **IEEE Transactions on Neural Networks**, v.16, 2005. p.645-678.