

**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA  
COMPUTAÇÃO**

**Antonio da Luz Júnior**

**UMA METODOLOGIA EFICIENTE PARA  
RECUPERAÇÃO DE EXAMES MÉDICOS DICOM  
POR SIMILARIDADE DE CARACTERÍSTICAS  
VISUAIS**

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação

**Prof. Dr. rer. nat. Eros Comunello**

Florianópolis, Março de 2007.

**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA  
COMPUTAÇÃO**

**Antonio da Luz Júnior**

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Ciências da Computação, área de concentração Sistemas de Computação, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

---

Prof. Dr. Rogério Cid Bastos  
Coordenador do Curso

Banca examinadora

---

Prof. Dr. rer. nat. Eros Comunello  
Orientador

---

Prof. Dr. rer.nat. Aldo von Wangenheim  
Co-Orientador

---

Prof. Dr. Luiz Felipe Nobre, MD

---

Profa. Dra. Agma Juci Machado Traina

## **Agradecimentos**

Agradeço a Deus e a Nossa Senhora Aparecida pelas bênçãos que sempre têm me proporcionado.

A meus pais pelo apoio e amor dedicados incondicionalmente.

Aos parentes e amigos pelo incentivo que sempre me deram.

Aos profs. Eros Comunello e Aldo von Wangenheim pela oportunidade e colaboração em mais esta etapa de minha formação.

Ao Daniel Duarte Abdala, grande mestre e amigo, por todos os ensinamentos: culinária, matemática, processamento de imagens, português, religião, entre outros temas que surgiram durante as discussões de todas as horas.

Aos grandes amigos Levi Ferreira e Rodrigo Copetti pelo carinho e compreensão que sempre tiveram comigo.

Aos profs. Eduardo Leal e Fabiano Fagundes que me incentivaram e sempre acreditaram que eu iria conseguir.

Aos amigos e companheiros membros do projeto Cyclops por dividirem comigo o peso desta caminhada.

# SUMÁRIO

Capítulo 1	1
Introdução	1
1.1 Definição do problema	3
1.2 Motivação	5
1.3 Objetivos	5
1.3.1 Objetivo Geral	5
1.3.2 Objetivos Específicos	5
1.4 Hipótese	6
1.5 Restrições	6
1.6 Material	7
1.7 Contribuições	8
1.8 Estado da Arte	8
1.8.1 SIM/HUG	8
1.8.2 RWTH/Aachen	9
1.8.3 RVL/Purdue	10
1.8.4 GBDI/USP	11
1.8.5 Outros	12
1.9 Estrutura do Trabalho	12
Capítulo 2	14
Similaridade: Características Visuais e Métricas de Distância	14
2.1 Cor	15
2.1.1 Histograma	15
2.1.2 Histograma Normalizado	16
2.2 Textura	17
2.2.1 Transformada de Fourier	17
2.2.2 Matriz de Co-ocorrência	18
2.3 Forma	20
2.3.1 Limites	20
2.3.1.1 Detector de Bordas de Sobel	21
2.3.1.2 Momentos	21

2.3.1.3	Descritores de Fourier	22
2.3.2	Regiões	22
2.3.2.1	Momentos Invariantes	22
2.3.2.2	Momentos de Zernike	23
2.4	Métricas de Distância	24
2.5	Considerações	25
Capítulo 3		26
Recuperação de Informação por Similaridade		26
3.1	Raciocínio Baseado em Casos	26
3.2	Técnicas de Classificação de Dados	28
3.2.1	Não-Supervisionadas	28
3.2.1.1	<i>k</i> -Means	29
3.2.2	Supervisionada	30
3.2.2.1	<i>k</i> -Nearest Neighbor	30
3.3	Técnicas para Seleção de Características	31
3.3.1	StARMiner	31
3.3.2	Customized-Queries	33
3.4	Métricas para Avaliação de Resultados	33
3.4.1	<i>Precision &amp; Recall</i>	34
3.4.2	Proposta do SIM/HUG-Viper	35
3.5	Considerações	36
Capítulo 4		37
Recuperação de Exames Médicos		37
4.1	Conjuntos de Informações em Exames Médicos	38
4.2	Recuperação Baseada na Descrição do Procedimento	40
4.2.1	Em Exames DICOM	41
4.3	Recuperação Baseada no Diagnóstico	43
4.4	Recuperação Baseada no Conteúdo de Imagens	45
4.5	CBMIR x RBC	45
4.6	Considerações	47
Capítulo 5		48

Metodologia para Recuperação de Exames Médicos DICOM por Similaridade de Características Visuais	48
5.1 Metodologia Proposta	48
5.2 Módulo de Especificação	49
5.3 Extração de Características	50
5.3.1 Características Pseudo-Semânticas	50
5.3.2 Categorização Nível 1	52
5.3.3 Características Visuais	52
5.3.4 Categorização Nível 2	53
5.4 Mineração	54
5.5 Similaridade	54
5.6 Adaptação de Casos	55
5.7 Reutilização	56
5.8 Construção e Manutenção da Base de Casos	56
5.9 Considerações	59
Capítulo 6	60
Resultados e Discussões	60
6.1 Experimentos e Resultados	60
6.1.1 Experimento #1	62
6.1.2 Experimento #2	64
6.1.3 Experimento #3	66
6.1.4 Experimento #4	68
6.1.5 Experimento #5	70
6.1.6 Experimento #6	72
6.2 Considerações	74
Capítulo 7	75
Conclusões	75
7.1 Objetivos Alcançados	75
7.2 Trabalhos Futuros	76
Referências	78

## Índice de Figuras

<b>Figura 1:</b> Exemplos de Imagens existentes na base de testes	7
<b>Figura 2:</b> Esquema de Integração de dados suportados pelo <i>Casim@ge</i>	9
<b>Figura 3:</b> Representação de Imagem através de Características Visuais	15
<b>Figura 4:</b> Imagem de CT (a) e seu Histograma (b)	16
<b>Figura 5:</b> Histograma Normalizado para a imagem 4a	17
<b>Figura 6:</b> Exemplo de Construção de Matriz de Co-ocorrência. Direção 0° Grau e Distância 2, para uma imagem de valores de pixel 0 e 1	18
<b>Figura 7:</b> Máscaras do Operador de Sobel	21
<b>Figura 8:</b> Ciclo Básico de Sistema de RBC	28
<b>Figura 9:</b> Exemplo de Agrupamento <i>k-Means</i>	29
<b>Figura 10:</b> Exemplo de Classificação <i>k-NN</i>	30
<b>Figura 11:</b> Algoritmo StARMiner	32
<b>Figura 12:</b> Representação de Dados Relevantes e Irrelevantes no processo de Recuperação de Informação	34
<b>Figura 13:</b> Exemplos de laudos médicos: manuscrito (esq.); ditado utilizando gravador de mão (centro) e; Interface para elaboração de laudo estruturado (dir.)	39
<b>Figura 14:</b> Modelo Simplificado para Recuperação de Exames	40
<b>Figura 15:</b> Camadas e Componentes da Metodologia para Recuperação de Exames Proposta	49
<b>Figura 16:</b> Hierarquia de Classificação, Nível 1	52
<b>Figura 17:</b> Hierarquia de Classificação, Nível 2	54
<b>Figura 18:</b> Estrutura para Construção e Manutenção da Base de Casos	57
<b>Figura 19:</b> Imagem de Consulta utilizada nos Experimentos	62
<b>Figura 20:</b> 20 primeiros resultados – Experimento #1	63
<b>Figura 21:</b> 20 primeiros resultados – Experimento #2	65
<b>Figura 22:</b> 20 primeiros resultados – Experimento #3	67
<b>Figura 23:</b> 20 primeiros resultados – Experimento #4	69
<b>Figura 24:</b> 20 primeiros resultados – Experimento #5	71
<b>Figura 25:</b> 20 primeiros resultados – Experimento #6	73

## Índice de Tabelas

<b>Tabela 1:</b> Atributos Descritores definidos por Haralick	19
<b>Tabela 2:</b> Momentos de Zernike de Ordem 0 a 10	24
<b>Tabela 3:</b> Medidas de Similaridade	25
<b>Tabela 4:</b> Valores críticos para $z$ , para $z=0,9$ , $z=0,95$ e $z=0,99$	32
<b>Tabela 5:</b> Exemplos de <i>Data Elements</i>	42
<b>Tabela 6:</b> Informações DICOM consideradas no processo de Recuperação de Exames	51
<b>Tabela 7:</b> <i>Precision, Recall</i> e Tempo de Execução – Experimento #1	63
<b>Tabela 8:</b> <i>Precision, Recall</i> e Tempo de Execução – Experimento #2	65
<b>Tabela 9:</b> <i>Precision, Recall</i> e Tempo de Execução – Experimento #3	67
<b>Tabela 10:</b> <i>Precision, Recall</i> e Tempo de Execução – Experimento #4	70
<b>Tabela 11:</b> <i>Precision, Recall</i> e Tempo de Execução – Experimento #5	71
<b>Tabela 12:</b> <i>Precision, Recall</i> e Tempo de Execução – Experimento #6	73

## Índice de Fórmulas

<b>Fórmula 1-2:</b> Conversão para Espectro 2D em 1D – direção e raio	18
<b>Fórmula 3:</b> Energia	19
<b>Fórmula 4:</b> Contraste	19
<b>Fórmula 5:</b> Correlação	19
<b>Fórmula 6:</b> Variância	19
<b>Fórmula 7:</b> Variância Inversa	19
<b>Fórmula 8:</b> Soma Média	19
<b>Fórmula 9:</b> Soma Variância	19
<b>Fórmula 10:</b> Soma Entropia	19
<b>Fórmula 11:</b> Entropia	19
<b>Fórmula 12:</b> Diferença da Variância	19
<b>Fórmula 13:</b> Diferença da Entropia	19
<b>Fórmula 14:</b> Medida de Correlação (1)	20
<b>Fórmula 15:</b> Medida de Correlação (2)	20
<b>Fórmula 16:</b> Máximo Coeficiente de Correlação	20
<b>Fórmula 17:</b> Magnitude do Gradiente	21
<b>Fórmula 18:</b> Direção de Borda	21
<b>Fórmulas 19-20:</b> Momento de Ordem $n$	22
<b>Fórmula 21:</b> Transformada Discreta de Fourier	22
<b>Fórmula 22:</b> Momentos	22
<b>Fórmula 23:</b> Momentos Centrais	23
<b>Fórmula 24:</b> Momentos Centrais Normalizados	23
<b>Fórmulas 25-31:</b> Momentos Invariantes	23
<b>Fórmulas 32-34:</b> Momentos de Zernike	23
<b>Fórmula 35:</b> Distância de Hamming	25
<b>Fórmula 36:</b> Distância de Manhattan	25
<b>Fórmula 37:</b> Distância Euclidiana	25
<b>Fórmula 38:</b> Entropia Relativa	25
<b>Fórmula 39:</b> Coeficiente de Discriminação $F$	29
<b>Fórmula 40:</b> Valor Médio de um Atributo	32

<b>Fórmula 41:</b> Desvio Padrão de um Atributo	32
<b>Fórmula 42:</b> Coeficiente Z	32
<b>Fórmula 43:</b> <i>Precision</i>	34
<b>Fórmula 44:</b> <i>Recall</i>	35
<b>Fórmula 45:</b> <i>Rank</i> Normalizado	35

## Resumo

À medida que se iniciou o processo de popularização de exames médicos em formato digital, surgiu à necessidade de se desenvolver técnicas capazes de facilitar o processo de tomada de decisão médica. Nesse contexto, técnicas de Recuperação de Imagens Médicas Baseada no Conteúdo têm sido empregadas.

Uma possível abordagem a ser utilizada em pesquisas por características visuais em bases de dados de imagens é a Consulta por Exemplo. O modelo de recuperação proposto por este trabalho manipula grandes bases heterogêneas de imagens, porém a extração das características visuais (cor, textura, forma e estrutura) se dá com métodos específicos para cada tipo de imagem. Considerando as particularidades de cada grupo de imagens, pretende-se aumentar a precisão do cálculo de similaridade e o grau de relevância dos resultados durante o processo de recuperação. O modelo apresentado analisa características léxico-sintáticas (ex.: histogramas de cores, descritores de textura e descritores de forma) e pseudo-semânticas (ex.: região do corpo analisada, modalidade de exame e dados do paciente) extraídas dos exames médicos analisados. Foi definida uma abordagem a priori para cálculo de similaridade entre exames. O processo de recuperação foi projetado para ser executado sobre as informações de um banco de casos contendo exames médicos no padrão DICOM.

O crescimento da utilização da tecnologia digital na realização de exames médicos por imagem tem gerado um grande volume de dados que por falta de métodos para acesso eficiente, deixam de ser utilizados no auxílio ao diagnóstico. A metodologia desenvolvida neste trabalho veio de encontro a esta necessidade. Os resultados apresentados comprovam o ganho de desempenho na execução de uma solicitação de recuperação de exames por similaridade do conteúdo visual. A redução de tempo necessário para execução de solicitações de recuperação chega próxima a 95%.

**Palavras-chave:** CBIR, CBMIR, DICOM, Mineração de Características, RBC.

## Abstract

As soon the healthcare institutions become to be more and more sustained by digital media, a new necessity becomes preeminent: the development of techniques capable the medical decision making process. Within this context, techniques known as Content Based Medical Image Retrieval where developed and since some time, are being applied with acceptable success.

One of the possible approaches for searching based on visual features on large image database is the Query by Example. The retrieval model presented by this dissertation, works over large heterogeneous medical image databases. However the visual features retrieval (color, texture, shape and structure) takes place using specific image processing methods. They take in account specific characteristics of each image type, in order to maximize the accuracy of similarity computation as well as the relevance grad by the retrieval process in large. The presented model analyzes the lexical-syntactical features (color histograms, texture descriptors and shape descriptors) and pseudo-semantics features (body part examined, modality and patients data) extracted from medical exams. It was defined also, an approach to calculate in advantage the images similarities, at constructions of case base time. The retrieval process was developed to be executed with information extracted from a case base composed only by DICOM medical exams.

The growing in usage of digital technology in medical image exams generated in a sort time a large volume of data. For lack of efficient methods to access those data, they where unfortunately misused to further aid the medical diagnosis using chronological data. The methodology developed in this work come to fulfill this lack in the already existing technological support in image retrieval. The obtained results prove the performance enhancement in retrieval by visual similarity execution process. The reduction in time necessity to this task is near 95%.

**Keywords:** CBIR, CBMIR, DICOM, Features Miner, CBR.

# **Capítulo 1**

## **Introdução**

À medida que se iniciou o processo de popularização de exames médicos em formato digital, surgiu à necessidade de se desenvolver técnicas capazes de facilitar o processo de tomada de decisão médica. Esse auxílio pode ser realizado via ferramentas que possibilitam a análise de imagens, como *Workstations Radiológicas*<sup>1</sup>, ou através de técnicas que permitam a comparação do exame em estudo, com exames anteriormente realizados e diagnosticados, existentes na base de dados, buscando soluções registradas em casos similares, para que sirvam de embasamento na elaboração do diagnóstico para o caso atual.

Um Exame Médico pode ser decomposto em macro conjuntos distintos de informações: a) Descrição do Procedimento; b) Laudo ou Diagnóstico, e; c) Imagens, Sinais ou dados laboratoriais. Esses dados, embora dotados de características distintas, referenciam um mesmo procedimento clínico realizado para um determinado paciente. A análise de dados de exames anteriormente realizados pode fornecer as informações necessárias para que o especialista médico determine a melhor alternativa a ser adotada no novo caso.

Nesse contexto, técnicas de Recuperação de Imagens Médicas Baseada no Conteúdo – *Content-Based Medical Image Retrieval (CBMIR)* [MULLER, 2004a] têm sido empregadas. As propostas encontradas atualmente nessa área têm se apresentado como uma alternativa viável e necessária, para a otimização eficiente do fluxo de trabalho de radiologistas e clínicos em geral. No entanto, pesquisas em bases de imagens médicas ainda dependem do desenvolvimento de técnicas adequadas para maximizar a qualidade dos resultados provenientes do processo de recuperação. Os modelos de pesquisa adotados por essas bases de imagens, em geral, decorrentes do modelo relacional ou orientado a objetos [SILBERSCHATZ, 1999], se demonstram inadequados

---

<sup>1</sup> Workstation Radiológica é a denominação dada a softwares que auxiliam o médico durante o processo de elaboração de diagnóstico, proporcionando ferramentas para a manipulação e análise de imagens médicas.

no contexto do processo de recuperação baseada em características visuais da imagem. Em bases de dados em que a informação contida é representada por números e textos, a pesquisa usando álgebra relacional fundamentada pela lógica de primeira ordem mostrou-se correntemente viável, funcional e eficiente para a recuperação dos dados utilizada a algumas décadas em inúmeros contextos e sistemas computacionais, tais como sistemas de controle de estoque e gerência de informações. No entanto, em bases de dados em que a informação a ser recuperada é, em sua grande maioria, composta de características extraídas de imagens bi ou tridimensionais, tal lógica apresenta-se insuficientemente expressiva para descrever os atributos e fornecer base comparativa entre os dados.

Uma possível abordagem a ser utilizada em pesquisas por características visuais em bases de dados de imagens é a Consulta por Exemplos – *Query by Example (QBE)* [SANTINI, 1998], um modelo de pesquisa onde uma imagem, chamada *imagem de consulta - IC*, é utilizada como exemplo de pesquisa. Alguns atributos da IC são extraídos e comparados com os mesmos atributos obtidos das imagens contidas na base de dados, e os exemplares mais similares a IC, dentro de um limite pré-definido, são então retornados como resultado do processo de recuperação. Tal modelo prevê a análise de aspectos léxicos da imagem (intensidade de sinal dos *pixels*) e sintáticos (localização espacial, bordas e geometria de estruturas) extraídos via algoritmos de PDI (Processamento Digital de Imagens) e convertidos para seus relativos numéricos.

Encontra-se ainda na literatura propostas para se executar a recuperação de imagens médicas, como podem ser observadas em [LEHMANN, 2004a], [MULLER, 2004b], [WILLY, 2004], [ANTANI, 2004], [MARQUES, 2002], [TRAINA, 1997]. Essas propostas podem ser classificadas em dois diferentes grupos, de acordo com a sua abordagem: a) tratam o problema de recuperação para um grupo específico de imagens - as propostas desse grupo destinam-se a auxiliar diretamente o processo de diagnóstico médico. Como um exemplo, podemos citar a identificação de possíveis anomalias em mamografias [MARQUES, 2002]; b) tratam todas as imagens de maneira única, não analisando as características particulares das diferentes classes de imagens médicas [LEHMANN, 2004a]. As propostas desse segundo grupo se destinam à recuperação de imagens em grandes Bases de Imagens.

O modelo de recuperação proposto por este trabalho atua de maneira híbrida. De maneira similar à da abordagem “*b*”, trabalha-se com grandes bases heterogêneas, porém a extração das características visuais (cor, textura, forma e estrutura) se dá de maneira similar à da abordagem “*a*”, com métodos específicos para cada tipo de imagem. Considerando as particularidades de cada grupo de imagens, pretende-se aumentar a precisão do cálculo de similaridade e o grau de relevância dos resultados durante o processo de recuperação.

Este trabalho apresenta um modelo para CBMIR. O modelo apresentado analisa características léxico-sintáticas (ex.: histogramas de cores, descritores de textura e descritores de forma) e pseudo-semânticas (ex.: região do corpo analisada, modalidade de exame e dados do paciente) extraídas dos exames médicos analisados. O processo de recuperação foi projetado para ser executado sobre as informações de um banco de casos contendo exames médicos no padrão DICOM [DICOM, 2006a] – *Digital Imaging and Communications in Medicine* (Imaginografia Digital e Comunicação na Medicina).

O modelo desenvolvido prevê a utilização de técnicas que permitem a identificação de quais características melhor descrevem cada uma das categorias de exames existentes na base de casos considerados neste trabalho. Com isso pretende-se definir um modelo para construção de ferramentas de CBMIR que possua um processo eficiente para a recuperação de exames médicos.

## **1.1 Definição do problema**

Com o início da realização de exames médicos digitais, como: Tomografia Computadorizada – CT, Ressonância Magnética – MR, Raio-X Computadorizada – CR, Angiografia – XA, Ultra-sonografia – US e Cintilografia - NM, tornou-se necessário o desenvolvimento de abordagens para acesso às informações contidas nesses exames, além de novas formas de comparação entre exames diferentes com o objetivo de auxiliar o profissional médico no instante do diagnóstico. O advento da radiologia digital introduziu uma nova ferramenta para diagnóstico de doenças, e criou uma nova área de pesquisa e aplicação, a manipulação e análise de imagens médicas. Ocorreu então uma alteração nos processos clínicos e cirúrgicos dentro da medicina, com ganhos reais para pacientes e médicos.

Anteriormente à utilização de exames digitais, quando se fazia necessária a comparação de diagnósticos emitidos para casos passados, para auxílio na elaboração de um laudo para um problema atual, era preciso utilizar um procedimento “manual”. Inicialmente, localizar exames do mesmo tipo do exame de referência numa sala de arquivos da sua clínica, ou hospital. Um problema encontrado nesse processo é que dada a grande quantidade de exames, o espaço de busca do especialista, geralmente, se restringia aos casos dos quais ele possuía conhecimento prévio de que poderiam ser similares ao problema em estudo. Com isso, nem sempre as informações sobre todos os casos anteriormente realizados eram compartilhadas e acessadas por todos os médicos. Aqui ainda podem ser adicionados os problemas de arquivamento e identificação dos exames, que é problema recorrente em instituições que não utilizam o modelo de arquivamento digital.

Porém, após conseguir localizar um conjunto de exames semelhantes ao caso em estudo, surge o problema de pesquisa “fina” dos casos. Colocar os exames lado-a-lado em um negatoscópio e a olho nu, ou com ajuda de uma lupa, procurar manualmente por similaridades entre os exames, até conseguir identificar os casos que realmente se demonstram semelhantes, para só então tentar verificar se os diagnósticos atribuídos àqueles exames verificados se aplicam ao novo caso.

Com a utilização de exames digitais, geralmente, não se têm mais os exames armazenados em forma de filme, inviabilizando o processo de busca manual por exames similares. Com isso, são necessárias técnicas computacionais de recuperação, capazes de atender às necessidades de comparação entre exames para uma possível reutilização de diagnósticos. Ou seja, definição de métodos capazes de analisarem as características dos exames armazenados na base de dados e classificá-los de acordo como seu grau de similaridade com um dado caso em estudo pelo especialista.

Esse problema, que inicialmente parece ser de fácil solução, se agrava ao constatar-se que durante o processo de recuperação manual o médico se utiliza de características tanto sintáticas quanto semânticas dos filmes para realizar o processo de avaliação de similaridade entre os exames. Ou seja, os métodos computacionais que se propuserem a resolver o problema de recuperação de exames baseado na similaridade de seus conteúdos, além de analisarem as características visuais dos exames, como cor, forma e textura, deverão ser capazes de identificar em qual contexto aquelas

características estão inseridas. Uma possível solução para o problema de recuperação automática de exames médicos é a utilização de técnicas de Inteligência Artificial, como Raciocínio Baseado em Casos, unida a métodos de Processamento de Imagens Digitais, para extração e identificação de características contidas nos exames, e a conceitos de Recuperação de Informação, como CBMIR.

## **1.2 Motivação**

Este trabalho se motiva frente à crescente dificuldade observada no que tange o acesso e recuperação a informações arquivadas em bases de dados de imagens médicas. Esta dificuldade pode ser considerada como um efeito colateral da informatização e digitalização dos dados de exames médicos.

Neste contexto, a disponibilização de métodos para fornecer a possibilidade de uma recuperação acurada de experiências passadas de especialistas em problemas que tem seu espectro definido desde os mais triviais até casos de mais alta complexidade, possibilitando o acesso rápido e facilitando o acesso às informações valiosas que de outra forma estariam perdidas, se apresentou como fator motivacional para que o presente trabalho tomasse forma.

## **1.3 Objetivos**

### **1.3.1 Objetivo Geral**

Elaborar uma metodologia híbrida para recuperação de exames DICOM baseada no conteúdo de imagens médicas, visando obter um método de recuperação ágil e eficiente, que possa servir de modelo para a elaboração de ferramentas de recuperação de exames independente de sua categoria.

### **1.3.2 Objetivos Específicos**

1. Definir uma estrutura modular que permita uma recuperação ágil e eficiente;
2. Definir quais das informações existentes no exame são úteis para o processo de recuperação de imagens;

3. Definir um método de filtragem de relevância dos exames a serem comparados, de forma a reduzir a quantidade de dados a serem verificados pelos processos mais custosos em termos computacionais;
4. Utilizar um método de seleção de características visuais do exame, que adapte os atributos analisados a cada uma das classes de exames identificadas neste trabalho;
5. Definir uma estratégia de pré-processamento para extração de características, visando minimizar a sobrecarga de processamento na estação de armazenamento de dados;
6. Definir uma abordagem eficiente para cálculo de similaridade entre os casos, permitindo uma execução ágil do processo de recuperação.

## **1.4 Hipótese**

Utilizando um conjunto de informações pseudo-semânticas, obtidas com a análise da descrição do procedimento adotado para a realização de exames DICOM, e léxico-sintáticas, extraídas do conteúdo das imagens médicas pertencentes a esses exames, é possível a definição de um modelo de recuperação de exames médicos digitais que consiga realizar um processo otimizado de recuperação em grandes volumes de dados, de forma a retornar todas as ocorrências similares a um dado caso em estudo, de modo a auxiliar no processo de elaboração do diagnóstico para o problema atual.

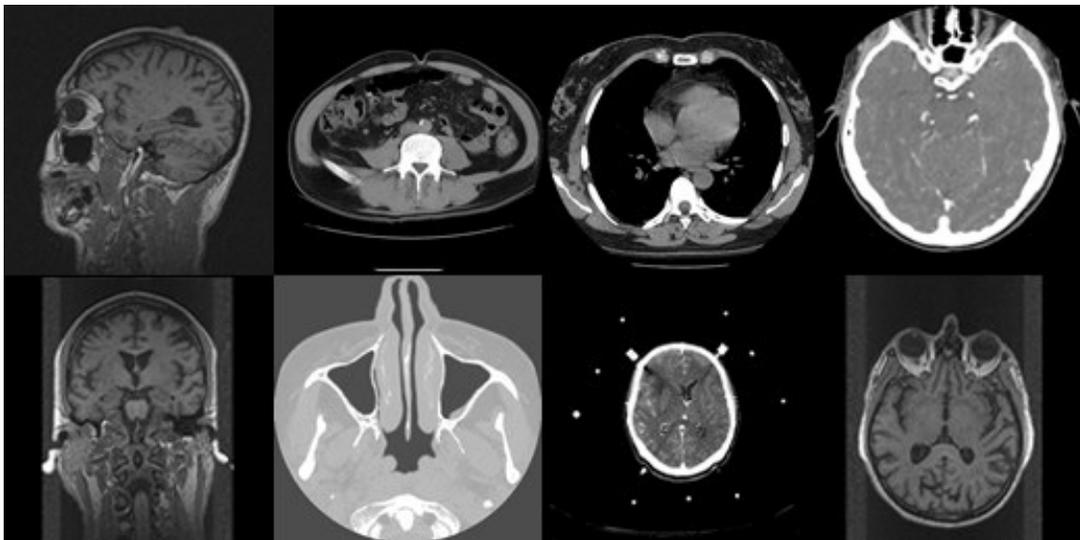
## **1.5 Restrições**

O modelo de recuperação de exames baseada no conteúdo apresentado neste trabalho foi concebido para suportar a tarefa de recuperação em todas as modalidades de exames médicos previsto no padrão DICOM. Todavia, as validações realizadas neste trabalho contemplaram as modalidades de CT – *Computed Tomography* (Tomografia Computadorizada) e MR – *Magnetic Resonancy* (Ressonância Magnética). A modalidade de CT foi utilizada por ter sido considerada comum considerando o volume de exames realizados e os exames de MR por ter sido observada maior complexidade nos casos de estudo, ou seja, os exames de MR estão relacionados a problemas mais complexos. Com isso se pretende melhor adequar o tempo disposto para realização deste trabalho com a necessidade mais eminente de seus beneficiários finais, os profissionais

médicos e pacientes. No entanto, como o modelo proposto foi desenvolvido para abordar o problema de recuperação de exames de uma maneira genérica, o modelo apresentado pode ser facilmente estendido para contemplar o suporte as outras modalidades de exames.

## 1.6 Material

Este trabalho fez uso de uma base de dados de exames médicos composta por 23.871 imagens, organizadas em 312 pacientes, 322 estudos e 627 séries<sup>2</sup>. As imagens utilizadas compõem um subconjunto de exames extraído da base de dados do Laboratório de Telemedicina da Universidade Federal de Santa Catarina. Foram utilizados dois tipos de modalidade de exames: CT e MR. Sendo que, cada uma dessas modalidades pode conter exames de diferentes subtipos, como: *MR-T1*, *MR-T2*, *MR-DTI*, e, também, diferentes partes do corpo, tal como definido pelo padrão DICOM [DICOM, 2006b], como: *Abdomen*, *Head*, *Chest*, e *Extremity*. A Figura 1 apresenta alguns exemplos de imagens de exames existentes no banco de dados de imagens utilizado neste trabalho.



**Figura 1:** Exemplos de Imagens existentes na base de testes.

---

<sup>2</sup> A organização em pacientes, estudos e séries deriva diretamente do padrão DICOM. Uma forma comum de referenciar as imagens médicas. Mais informações [DICOM, 2006b]

## **1.7 Contribuições**

A principal contribuição deste trabalho é a definição de um modelo híbrido para o desenvolvimento de ferramentas para recuperação de exames médicos de acordo com a similaridade de suas características visuais e pseudo-semânticas. Com isso possibilitando a utilização prática deste tipo de abordagem de maneira eficiente no dia-a-dia de profissionais da área da saúde.

A metodologia utilizada para elaboração do modelo proposto neste trabalho integra as vantagens dos conceitos apresentados nos trabalhos de referência na área de recuperação de imagens médicas de modo a aproveitar em um único modelo as contribuições relevantes de modelos anteriormente propostos, contribuindo para o aprimoramento do estado da arte desta área de aplicação com a elaboração de um modelo homogêneo.

## **1.8 Estado da Arte**

Nesta seção é apresentada uma visão geral do cenário atual da área de recuperação de imagens por conteúdo, aplicada no contexto médico. A seção é dividida em cinco tópicos, onde são expostas as contribuições dos principais grupos de pesquisa, além de alguns outros resultados interessantes obtidos por outros pesquisadores.

### **1.8.1 SIM/HUG**

O SIM/HUG [SIM/HUG, 2007] – *Service d’Informatique Médicale des Hôpitaux Universitaires de Genève* (Serviço de Informática Médica do Hospital Universitário de Genebra), na Suíça, possui importantes trabalhos, [MULLER, 2004a; 2005a; 2006a], que afiguram como um bom ponto de partida para pesquisas na área de Recuperação de Imagens Médicas por Conteúdo.

Em conjunto com outros parceiros os esforços têm se focado na elaboração de métodos eficientes para a avaliação da qualidade de metodologias de CBMIR [CLOUGH, 2004; 2005] e [MULLER, 2005b]. Além disso, mantêm dois outros projetos: *Casim@ge* [CASIMAGE, 2007] e [MULLER, 2004d] – *Case of Imaging* (Imagens de

Casos) e *medGIFT* [FLOREA, 2006], [MULLER, 2005c], [MULLER, 2006b] – *GNU Medical Image Finding Tool* (Ferramenta GNU para Busca de Imagens Médicas).

O *Casim@ge* tem por objetivo ser um repositório de imagens médicas, fornecendo conteúdo para pesquisas na área de informática médica, além de servir de base comum para realização de testes de qualidade de metodologias de recuperação de imagens médicas. A Figura 2 mostra o modelo de integração entre banco de dados, geradores de imagens e fontes de acesso ao conteúdo armazenado na base do *Casim@ge*.



**Figura 2:** Esquema de integração de dados suportado pelo *Casim@ge* [CASIMAGE, 2007].

Com o objetivo de desenvolver uma ferramenta de código-livre que ofereça suporte ao trabalho de especialistas médicos, além de servir como modelo para o desenvolvimento de outras ferramentas de CBMIR, o SIM-HUG criou o *medGIFT*, adaptado a partir de outras aplicações para CBIR [MULLER, 2004e].

### 1.8.2 RWTH/Aachen

O grupo de Informática Médica do RWTH (website)– *Rheinisch-Westfälische Technischen Hochschule* (Universidade de Tecnologia) de Aachen, na Alemanha, é o responsável pelo desenvolvimento de um dos mais importantes projetos na área de Recuperação de Imagens Médicas por Conteúdo, o IRMA [IRMA, 2007] – *Image Retrieval in Medical Applications* (Recuperação de Imagens em Aplicações Médicas).

Propõe uma metodologia de recuperação modular [LEHMANN, 2001; 2004a] que visa atender a tarefa de recuperação de imagens médicas sem a necessidade de se restringir um domínio específico, como se trabalhavam na grande maioria das propostas existentes. Para realizar esse objetivo, o enfoque principal dos trabalhos desenvolvidos foi a classificação automática de imagens médicas [GÜLD, 2004], [LEHMANN, 2005a], [KEYSERS, 2004], [DAHMEN, 2000], [LEHMANN, 2004b], procurando identificar melhores formas de agrupamento e características a serem utilizadas para realizar o processo de agrupamento (*clustering*) das imagens.

Para validar a metodologia de recuperação em desenvolvimento, foi elaborada uma ferramenta, [DASERLAERS, 2004] e [GÜLD, 2003; 2005], que implementa o modelo proposto. Em [THIES, 2004] e [LEHMANN, 2003], respectivamente, são apresentados relatos da utilização dessa ferramenta integrada a um PACS e, também, uma tentativa de integração com o projeto *Casim@ge* (descrito na subseção anterior).

### 1.8.3 RVL/Purdue

O RVL/Purdue [RVL/PURDUE, 2007] – *Robot Vision Lab. of Purdue University* (Laboratório de Visão Robótica da Universidade de Purdue), nos Estados Unidos, desenvolve um projeto de destaque na área de CBMIR. O objetivo do projeto é o de desenvolver técnicas para auxílio ao diagnóstico em exames de Tomografia Computadorizada. Os testes, [AISEN, 2003], [SHYU, 1999; 2000] e [KAK, 2000], com o sistema de recuperação desenvolvido, ASSERT – *Automated Search and Selection Engine with Retrieval Tools* (Motor de Seleção e Busca Automática com Ferramentas de Recuperação), são realizados em um conjunto de dados de domínio restrito. Tal fato dificulta a avaliação ou comparação com outras ferramentas.

Contudo, foi a partir desse projeto que surgiu a técnica de *customized-queries* (consultas customizadas) e, também, algoritmos que realizam a seleção automática de características simultaneamente ao agrupamento das imagens [DY, 1999a; 1999b; 2003]. Essas propostas auxiliam no aumento da eficiência de métodos de recuperação de imagens por conteúdo também em domínio aberto.

#### 1.8.4 GBDI/USP

No GBDI/USP [GBDI/USP, 2007] – Grupo de Bases de Dados e Imagens da Universidade de São Paulo – tem sido dedicado bastante esforço a pesquisa na área de Recuperação de Imagens Médicas por Conteúdo, o que pode ser comprovado pela existência de trabalhos relevantes em diferentes frentes dentro dessa área:

- No desenvolvimento de novas técnicas para Recuperação e proposta de novas características a serem avaliadas, [BALAN, 2004a; 2004b], [BUENO, 2002], [CASTAÑÓN, 2002; 2003a; 2003b], [WATANABE, 2006], [CAMPO, 2003] e [BORTOLOTTI, 2004], com destaque para [BUENO, 2002], onde é proposta uma nova técnica para análise de características de imagens, denominado Histograma Métrico. Essa técnica realiza uma redução da quantidade de *bins* (pode ser interpretado como uma posição num Histograma, representando um determinado nível de intensidade) a serem analisados, através de um esquema de agrupamento em *buckets* (conjuntos de *bins*);
- Nas tentativas de aplicação prática de ferramentas de CBMIR no dia-a-dia do profissional médico, [ARAUJO, 2002a], [CHINO, 2002], [TRAINA, 2002] e [ROSA, 2002]. Vale ser ressaltado [TRAINA, 2002], que aborda o desenvolvimento de um novo conceito de PACS com o recurso de Recuperação de Imagens Baseada no Conteúdo;
- Com propostas que visam reduzir o número de características a serem avaliadas, utilizando técnicas estatísticas e de mineração de dados, [TRAINA, 2004], [RODRIGUES JR, 2003], [RIBEIRO, 2005] e [RIBEIRO, 2006]. Nessa linha, em [RIBEIRO, 2005] é apresentado o método StARMiner – *Statistical Association Rule Miner* (Minerador Estatístico de Regras de Associação), que tem por objetivo determinar o grupo de características determinantes para identificação de uma categoria distinta de imagens; e,
- Propondo a integração de métodos de CBMIR às ferramentas de Gerenciamento de Bancos de Dados e extensão da SQL – *Structured Query Language* (Linguagem Estrutura de Consulta) [TRANA, 1997a; 1997b; 1998], [ARAUJO, 2002b] e [BARIONI, 2006] para suportar

solicitações de Recuperação por Conteúdo. Em [ARAUJO, 2002b] é apresentado um interpretador de comandos que realiza a tradução de solicitações de consultas por conteúdo visual para consultas SQL padrão, atuando como uma proposta de extensão a linguagem de consultas tradicional para SGBDR – Sistemas Gerenciadores de Bancos de Dados Relacional.

### **1.8.5 Outros**

Além dos trabalhos na área de CBMIR desenvolvidos pelos grupos anteriormente citados, existem outras propostas que devem ser consideradas no contexto do desenvolvimento desse trabalho. Em [XIONG, 2005] é apresentado o conceito de QTDIF – *Query Topic Dependent Image Features* (Consulta dependente das características do tipo da imagem), a proposta não é exaustivamente detalhada. Até onde se pôde analisar, é semelhante ao apresentado em [DY, 1999a]. Mas, em momento algum é feita referência a outra proposta no texto.

Em [MOSHFEGUI, 2004] é descrita uma nova métrica para cálculo de distância entre imagens médicas. A proposta sugere a utilização da entropia relativa entre duas imagens. Os resultados iniciais têm apresentado bons resultados na tarefa de recuperação, além de não requerer uma fase de pré-processamento incluindo segmentação de imagens.

## **1.9 Estrutura do Trabalho**

Este capítulo apresentou uma introdução ao tema abordado neste trabalho, bem como os objetivos e contribuições da proposta que será apresentada a seguir. O restante deste trabalho está organizado da seguinte forma:

### **Capítulo 2 – Similaridade: Características Visuais e Métricas de Distância**

Descreve o conceito de similaridade entre objetos. Apresenta algumas das características visuais possíveis de serem utilizadas como parâmetros de comparação entre imagens. E, ainda, exemplos de métricas de distância que podem ser adotadas para mensurar o devido grau de similaridade.

### **Capítulo 3 – Recuperação de Informação por Similaridade**

Apresenta a técnica de Raciocínio Baseado em Casos, modelo sobre o qual se estrutura a metodologia proposta neste trabalho. Também são descritas técnicas de classificação de dados. Essas técnicas são utilizadas para determinar a qual categoria pertence um determinado objeto.

### **Capítulo 4 – Recuperação de Exames Médicos**

Descreve as diferentes formas de estruturação de um exame médico: Manuscrito, Ditado e Estruturado. Apresenta as possíveis formas de recuperação de exames de acordo com cada um de seus componentes: Descrição do Procedimento, Diagnóstico e Imagem. Faz, ainda, uma comparação entre os conceitos de RBC e CBMIR na recuperação de exames médicos.

### **Capítulo 5 – Metodologia para Recuperação de Exames Médicos DICOM por Similaridade de Características Visuais**

Apresenta a metodologia desenvolvida neste trabalho. Descreve a arquitetura definida para processo de recuperação, seus módulos e componentes. Também mostra o esquema definido para a tarefa de construção da Base de Casos.

### **Capítulo 6 – Experimentos e Resultados**

Descreve alguns dos experimentos realizados durante a elaboração deste trabalho e os resultados obtidos. São apresentados os primeiros casos resultantes dos processos de recuperação e realizada uma comparação entre esses resultados.

### **Capítulo 7 – Conclusões**

Apresenta considerações gerais sobre a metodologia proposta. Descreve um conjunto de pesquisas que podem ser originadas a partir dos resultados obtidos neste trabalho.

## **Capítulo 2**

### ***Similaridade: Características Visuais e Métricas de Distância***

Para realizar a recuperação de imagens baseada na similaridade de conteúdos é necessário que seja realizado um processo de identificação de características da imagem para que sejam definidos pontos de comparação entre a imagem de consulta e as imagens contidas na base de casos onde será realizada a busca. A literatura pertinente apresenta diferentes características que podem ser utilizadas pela tarefa de representação do conteúdo léxico-sintático de imagens, como: distribuição de cores, atributos de textura e forma dos objetos existentes na imagem. Essas características são utilizadas para simbolizar o conteúdo intrínseco na imagem analisada. As informações extraídas das imagens são organizadas em um vetor multidimensional que é utilizado como assinatura da imagem, ou seja, representa e identifica a imagem. Esse vetor é conhecido como *Vetor de Características*.

De modo semelhante são encontradas diferentes formas de se realizar a avaliação do grau de similaridade entre os vetores de características que representa o conjunto de atributos extraídos de uma imagem, como: Distância Euclidiana, Distância de Hamming e Modelo de Contraste (TVERSKY, 1977). Tradicionalmente, os atributos extraídos de uma imagem são classificados em dois tipos: Locais, que identificam uma região ou segmento da imagem, e Globais, que representam a imagem como um todo.

Neste capítulo serão apresentadas algumas das características passíveis de serem extraídas em imagens médicas. Essas características estão divididas em grupos, de acordo com o tipo de característica visual que representa. São eles: a) Cor; b) Textura, ou; c) Forma. Outros conjuntos de características possíveis para se analisar as informações existentes em uma imagem são a Geométrica e a Estrutural, que considera as relações topológicas entre os objetos. A Figura 3 mostra um esquema de representação de uma imagem através de suas características visuais. São descritas algumas propostas para medir a distância entre objetos, utilizadas para definir o grau de semelhança entre eles.



**Figura 3:** Representação de Imagem através de Características Visuais.

## 2.1 Cor

As características pertencentes a este grupo de informações, são obtidas através da análise da distribuição dos valores de *pixels* existentes na imagem. Os métodos que trabalham neste domínio são dependentes do espaço de cores utilizado<sup>3</sup>: Tons de Cinza, RGB, dentre outros. Em certas situações, antes que a extração de características possa ser realizada, uma mudança no espaço de cores da imagem, no intuito de obter uma maior agilidade no processo de extração e posterior comparação, deve ser executada. No entanto, a dinâmica do funcionamento geral dos métodos é semelhante para todos os espaços de cores existentes.

### 2.1.1 Histograma

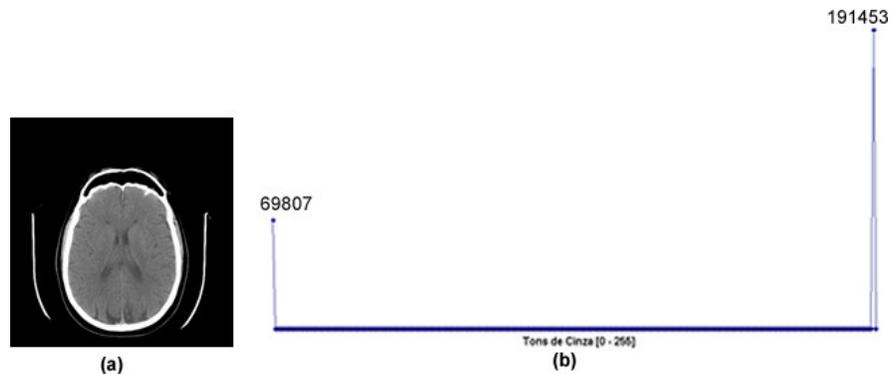
Um Histograma é uma técnica de agregação de informações correlatas proveniente da estatística. Ele descreve uma imagem através da contagem da quantidade de cada valor de intensidade de cor encontrada nessa imagem. Esse método é computacionalmente simples<sup>4</sup> e por isso, bastante utilizado em aplicações de recuperação de imagens por conteúdo.

Resultados positivos da utilização isolada dessa característica têm sido relatados em aplicações de domínios abertos. Em aplicações específicas como a área médica, onde

<sup>3</sup> Um espaço de cor é um modelo matemático abstrato utilizado para descrever a maneira com a qual as cores podem ser representadas, normalmente na forma de tuplas com 3 ou 4 valores de componentes de cor. O valor associado a cada valor da tupla corresponde a uma coordenada dentro do espaço de cores.

<sup>4</sup> Sempre  $O(N)$ , sendo  $N$  o número de *pixels* que compõem a imagem.

a restrição ao contexto das imagens utilizadas é imposta e, geralmente as imagens estão em tons de cinza, a análise isolada da distribuição dos valores de intensidade dos *pixels*, não tem gerado resultados satisfatórios até então. Por outro lado, essa característica analisada em conjunto com outras, de preferência as que levam em consideração o posicionamento dos *pixels* na imagem, demonstra-se como lugar comum em aplicações de CBMIR que relatam bons resultados. A Figura 4b apresenta uma representação gráfica de um Histograma gerado para a imagem de CT, 4a.



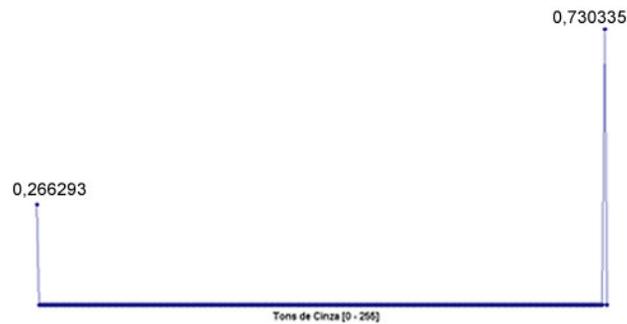
**Figura 4:** Imagem de CT (a) e seu Histograma (b).

### 2.1.2 Histograma Normalizado

O Histograma Normalizado [GONZALEZ, 1992] é obtido com a transformação dos valores de um histograma para uma faixa de valores conhecida. Na maioria das aplicações a faixa de valores adotada é  $[0,1]$ . Histogramas Normalizados são úteis em aplicações para a área médica, principalmente quando se trabalha com imagens padrão DICOM, onde os valores dos *pixels* não se encontram em um espaço de cor tradicional. Nesse tipo de imagem, os *pixels* representam valores em HU - *Unidades de HounsField* [DICOM, 2006b], conhecido como intensidade radiológica. A faixa de valores é muito ampla e varia de acordo com a modalidade de exame. Em exames de tomografia, por exemplo, a faixa de valores possíveis está entre  $-4000$  e  $+3999$ <sup>5</sup>. A operação de normalização é realizada através do cálculo de *regra de três simples*. O Histograma Normalizado possui o mesmo número de *bins*, posição no vetor ou número de cores avaliadas, que o histograma original, a única diferença entre eles está nos valores de cada

<sup>5</sup> Esses valores valem apenas para Tomografia. Em ressonância não existe uma limitação de faixa, e US normalmente é RGB GS direto ou 3 canais R, G, B.

uma dessas posições. Na Figura 5b é apresentado um exemplo de Histograma Normalizado para a imagem de CT, 5a.



**Figura 5:** Histograma Normalizado para a imagem 4a.

## 2.2 Textura

Na literatura da área de PDI são encontradas diferentes propostas para a análise da característica de textura de imagens, como: Transformada de Fourier [COOLEY, 1965], Wavelets [LEE, 1996], Matriz de Co-ocorrência [HARILACK, 1973], entre outras. Nesta seção são apresentados dois destes métodos, Transformada de Fourier e Matriz de Co-ocorrência, e alguns dos atributos que podem ser extraídos após a aplicação desses métodos para a representação da informação de textura da imagem.

### 2.2.1 Transformada de Fourier

A Transformada de Fourier converte o conteúdo do domínio do valor para o da frequência. Para que se possa realizar análise de conteúdo em imagens considerando o domínio da frequência dos valores de *pixel*, é necessário que ocorra uma conversão entre domínios, de valor para frequência.

Este é um método computacionalmente intenso,  $O(n^2)$ , o que implica em uma implementação ineficiente do método original. Existem diferentes formas de se realizar essa transformação, o método mais eficiente conhecido é o FFT - *Fast Fourier Transform* (Transformada Rápida de Fourier), complexidade  $O(n \cdot \log(n))$ . Existem diferentes propostas para o cálculo do FFT, [COOLEY, 1965] e [BRUUN, 1978], que tradicionalmente é definida para dados unidimensionais. Esses métodos podem ser

estendidos para funcionar em espaços bi e tridimensionais. Uma abordagem é calcular o FFT unidimensional para todas as linhas e, em seguida, para todas as colunas.

O resultado do FFT é dado em números complexos, compostos de uma parte real e uma parte imaginária. Ao final da execução da transformada são obtidas duas matrizes de tamanhos iguais ao tamanho original da imagem, uma contendo a parte real e outra a parte imaginária resultante da execução do FFT.

Após realizar a transformada, podem ser extraídos diferentes atributos que representam a característica de textura de uma imagem. Em [GONZALEZ, 1992] é encontrado um conjunto desses atributos, como: posição do maior valor, média, variância e distância entre a média e o maior valor. Para o cálculo desses atributos o espectro 2D obtido na transformada é convertido para espectro 1D. A conversão é realizada considerando as variações de direção ( $\theta$ ) e raio ( $r$ ). Em (1) e (2) é apresentado o modelo para a conversão.

$$S(r) = \sum_{\theta=0}^{\pi} S_{\theta}(r) \quad (1)$$

$$S(\theta) = \sum_{r=1}^R S_r(\theta) \quad (2)$$

### 2.2.2 Matriz de Co-ocorrência

A análise feita a partir desse tipo de matriz foi inicialmente introduzida por [HARALICK, 1973]. A técnica se baseia na análise da vizinhança dos *pixels*. São considerados os vizinhos nas direções 0°, 45°, 90° e 135°. A distância de análise da vizinhança é definida por um parâmetro  $d$ . Os valores armazenados na matriz são interpretados da seguinte maneira: Um valor  $V(x,y)$  qualquer pertencente a matriz, na linha  $x$  e coluna  $y$ , representa a quantidade de valores  $i$  que possuem vizinhos de valor  $j$  a uma distância  $d$ , a esquerda ou a direita. A Figura 6 apresenta um exemplo de construção de uma matriz de co-ocorrência para um conjunto de valores de *pixel*.

0	0	0	0
1	0	0	0
1	1	1	1
1	1	0	1
1	0	0	0

0	8	3
1	3	6
	0	1

**Figura 6:** Exemplo de Construção de Matriz de Co-ocorrência. Direção 0° e Distância 2, para uma imagem de valores de *pixel* 0 e 1.

Em [HARALICK, 1973] são apresentados um conjunto de atributos que podem ser extraídos de uma imagem a partir da utilização de técnicas de análise estatística de 1ª ou 2ª ordem. Esses atributos foram originalmente definidos para serem calculados sobre o resultado obtido da aplicação do método de matriz de co-ocorrência. Os resultados gerados com a utilização desses descritores estão relacionados à característica visual de textura de uma imagem. A Tabela 1 apresenta uma relação com esses descritores e suas respectivas formulações matemáticas.

**Tabela 1:** Atributos Descritores definidos por [HARALICK, 1973].

Descritor	Formulação	
Energia	$\sum_{i,j} p(i, j)^2$	(3)
Contraste	$\sum_n n^2 \sum_{i,j i-j =n} p(i, j)$	(4)
Correlação	$\frac{\sum_{i,j} ij p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ <p>onde, <math>\mu_x, \mu_y</math> e <math>\sigma_x, \sigma_y</math> são a média e o desvio padrão de <math>p_x(i) = \sum_k p(i, k)</math> e <math>p_y(j) = \sum_k p(k, j)</math>.</p>	(5)
Variância	$\sum_{i,j} (i - \mu)^2 p(i, j)$ onde, $\mu$ é a média da função de densidade $p(i, j)$ .	(6)
Variância Inversa	$\sum_{i,j} \frac{1}{1 + (i - j)^2} p(i, j)$	(7)
Soma Média	$\sum_{i=2}^{2\omega} i p_{x+y}(i)$ onde, $p_{x+y}(i) = \sum_{j,k} p(j, k)$ ; $j + k = i$ .	(8)
Soma Variância	$\sum_{i=2}^{2\omega} \left( i + \sum_{j=2}^{\omega} p_{x+y}(j) \log(p_{x+y}(j)) \right)^2 p_{x+y}(i)$	(9)
Soma Entropia	$-\sum_{i=2}^{\omega} p_{x+y}(i) \log(p_{x+y}(i))$	(10)
Entropia	$-\sum_{i,j} p(i, j) \log(p(i, j))$	(11)
Diferença da Variância	variância de $p_{x-y}$	(12)
Diferença da Entropia	$-\sum_{i=0}^{\omega-1} p_{x-y}(i) \log(p_{x-y}(i))$	(13)

Medida de Correlação (1)	$\frac{HXY - HXY1}{\max(HX, HY)}$ <p>onde, <math>HX</math> e <math>HY</math> são entropias de <math>p_x</math> e <math>p_y</math>,</p> $HXY = -\sum_{i,j} p(i, j)\log(p(i, j)),$ $HXY1 = -\sum_{i,j} p(i, j)\log(p_x(i)p_y(j)),$ $HXY2 = -\sum_{i,j} p_x(i)p_y(j)\log(p_x(i)p_y(j)).$	(14)
Medida de Correlação (2)	$\sqrt{1 - e^{-2(HXY2 - HXY)}}$	(15)
Máximo Coeficiente de Correlação	$\sqrt{\text{o segundo maior eigenvalue de } Q \text{ onde,}}$ $Q(i, j) = \sum_k \frac{p(i, k)p(j, k)}{p_x(i)p_y(k)}.$	(16)

## 2.3 Forma

A extração de características de forma de uma imagem, compreende a análise dos valores de intensidade de *pixel* encontrados na imagem, levando em consideração a sua disposição no plano bidimensional e, também, analisando a sua vizinhança. São empregadas duas abordagens para a representação do conteúdo das imagens em relação a sua forma. Uma que considera apenas os *pixels* de bordas da imagem, limites de objetos, e outra que analisa a imagem como um todo.

### 2.3.1 Limites

Nesta abordagem são utilizados apenas o conjunto de *pixels* identificados como pertencentes a bordas de objetos dentro da imagem para extração de atributos que caracterizam as forma presentes na imagem. Para isso é necessária a utilização de um método para detecção de bordas. Na literatura são encontradas diferentes propostas: Canny [CANNY, 1986], Sobel, Roberts e Robinson [ZIOU, 1998]. É utilizado o detector de Sobel por ser ágil e oferecer uma qualidade de resultado aceitável para o processo de detecção de bordas. Após a detecção são empregadas as técnicas de Momentos e Descritores de Fourier [GONZALEZ, 1992] para extração de características que representem a forma das bordas identificadas.

### 2.3.1.1 Detector de Bordas de Sobel

O processo de análise, ou detecção, de bordas em imagens compreende da identificação de regiões de mudança abrupta nos valores de intensidade de *pixel* na imagem. Esse processo é semelhante ao reconhecimento de limites de objetos realizado pelo cérebro humano, que identifica áreas através do processo de observação de variações nas tonalidades de cores das superfícies.

O método de Detecção de Bordas de Sobel é também conhecido como um Operador de Convolução. Isso por aplicar sobre a imagem, máscaras ortogonais para determinar o valor de magnitude de gradiente da imagem,  $|G|$ . A Figura 7 apresenta as máscaras utilizadas pelo operador de Sobel para definição dos valores de gradiente nas direções horizontal,  $G_x$ , e vertical,  $G_y$ .

$$G_x = \begin{vmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{vmatrix} \quad G_y = \begin{vmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{vmatrix}$$

**Figura 7:** Máscaras do Operador de Sobel.

A formulação para o cálculo da magnitude do gradiente é definida por (17). A direção da borda pode ser definida através do cálculo de (18). Se a magnitude calculada é maior do que um valor limite, conhecido por parâmetro, o *pixel* é considerado ser parte de uma região de borda.

$$M(x, y) = \sqrt{G_x^2 + G_y^2} \quad (17)$$

$$\alpha = \arctan\left(\frac{G_y}{G_x}\right) \quad (18)$$

### 2.3.1.2 Momentos

Este método fornece uma representação numérica para a forma dos *pixels* de borda detectados na imagem. O cálculo é computacionalmente simples, o que se torna uma vantagem para o uso dessa abordagem. O segundo momento ( $n = 2$ ) mede a propagação da curva sobre o valor médio de  $r$  e o terceiro momento ( $n = 3$ ) representa a simetria com relação a média [GONZALEZ, 1992]. Geralmente esses valores são utilizados para descrever a característica de forma dos seguimentos de bordas. Em (19) é

apresentada a formulação para o cálculo do Momento de ordem  $n$  de um seguimento de borda.

$$\mu_n(r) = \sum_{i=1}^L (r_i - m)^n g(r_i) \quad \text{onde,} \quad (19)$$

$$m = \sum_{i=1}^L r_i g(r_i) \quad (20)$$

### 2.3.1.3 Descritores de Fourier

Esta técnica consiste na utilização de um conjunto  $s$  de  $N$  elementos de bordas da imagem para prover uma descrição através da aplicação da Transformada Discreta de Fourier [GONZALEZ, 1992]. Ao final do processo é obtido um conjunto  $a$ , com  $N$  descritores que representam a forma do seguimento de borda da imagem. Em (21) é apresentada a fórmula para o cálculo da Transformada Discreta de Fourier.

$$a(u) = \frac{1}{N} \sum_{k=0}^{N-1} s(k) e^{\frac{-j2\pi uk}{N}} \quad \text{onde, } u = k = 0, 1, 2, \dots, N - 1. \quad (21)$$

## 2.3.2 Regiões

Os métodos utilizados nesta abordagem analisam todos os *pixels* presentes na imagem e suas relações de vizinhança provendo um atributo numérico que representa a característica de forma da imagem como um todo. As técnicas utilizadas são as de Momentos Invariantes [HU, 1962] e Momentos de Zernike [KHOTANZAD, 1990]. Ambas fornecem uma representação invariante à rotação, translação e escala.

### 2.3.2.1 Momentos Invariantes

A utilização do cálculo de Momentos Centrais para representar a forma de um objeto não é capaz de manter a relação entre dois objetos que possuem a mesma forma, porém, estão em escala diferentes, ou ainda, passaram por processo de rotação ou translação. Em (22), (23) e (24) são apresentadas as fórmulas para o cálculo de Momentos, Momentos Centrais e Momentos Centrais Normalizados, respectivamente.

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \quad (22)$$

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \text{ onde, } \bar{x} = \frac{m_{10}}{m_{00}} \text{ e } \bar{y} = \frac{m_{01}}{m_{00}}. \quad (23)$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}, \text{ onde } \gamma = \frac{p+q}{2} + 1 \quad (24)$$

Em [HU, 1962] foi apresentado um conjunto com 7 seqüências de momentos capazes de representar as formas contidas na imagem independente de modificações de rotação, translação e escala. Esse conjunto é conhecido como *Momentos Invariantes* e é apresentado de (25) a (31) [HU, 1962].

$$\Phi_1 = \eta_{20} + \eta_{02} \quad (25)$$

$$\Phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \quad (26)$$

$$\Phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (27)$$

$$\Phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (28)$$

$$\Phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (29)$$

$$\Phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \quad (30)$$

$$\Phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03}) [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (31)$$

### 2.3.2.2 Momentos de Zernike

Este método consiste em um conjunto de polinômios complexos proposto por [ZERNIKE, 1934]. Na proposta original era apenas invariante a rotação, porém realizando uma etapa de normalização é possível obter a invariância a escala e translação [KHOTANZAD, 1990]. Em (32) é apresentada a fórmula para o cálculo dos Momentos de Zernike. A Tabela 2 apresenta os Momentos de Zernike até a 10ª ordem.

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x, y) V(\rho, \theta), \quad x^2 + y^2 \leq 1. \quad (32)$$

$$V_{nm}(x, y) = V_{nm}(\rho, \theta) = R_{nm}(\rho) \exp(jm\theta), \text{ onde } j = \sqrt{-1} \quad (33)$$

$$R_{nm}(\rho) = \sum_{s=0}^{n-|m|/2} (-1)^s \frac{(n-s)!}{s! \left(\frac{n+|m|}{2} - s\right)! \left(\frac{n-|m|}{2} - s\right)!} \rho^{n-2s} \quad (34)$$

**Tabela 2:** Momentos de Zernike de Ordem 0 a 10.

Ordem	Momentos	Num. Momentos.
0	$A_{00}$	1
1	$A_{11}$	1
2	$A_{20}, A_{22}$	2
3	$A_{31}, A_{33}$	2
4	$A_{40}, A_{42}, A_{44}$	3
5	$A_{51}, A_{53}, A_{55}$	3
6	$A_{60}, A_{62}, A_{64}, A_{66}$	4
7	$A_{71}, A_{73}, A_{75}, A_{77}$	4
8	$A_{80}, A_{82}, A_{84}, A_{86}, A_{88}$	5
9	$A_{91}, A_{93}, A_{95}, A_{97}, A_{99}$	5
10	$A_{100}, A_{102}, A_{104}, A_{106}, A_{108}, A_{1010}$	6

## 2.4 Métricas de Distância

O termo “similar” é empregado quando se deseja expressar que dois, ou mais, conjuntos, com as mesmas propriedades, possuem elementos próximos da igualdade, porém não o são completamente iguais. Para se realizar a determinação de quão similares são esses conjuntos é necessária a utilização de uma métrica de distância. Essa métrica irá encontrar o grau de similaridade entre os conjuntos analisados.

Os atributos extraídos de uma imagem, representando as diferentes características visuais: Cor, Textura e Forma; são utilizados como uma assinatura que representa a imagem. Porém, em um sistema de recuperação de imagens baseado no conteúdo, apenas representar o conteúdo das imagens não é suficiente para se determinar quais são as imagens mais similares a uma imagem de consulta qualquer. Para atender esse objetivo é necessário utilizar métricas que atuem como base de comparação.

Na literatura são encontradas diferentes propostas para realização da medida de similaridade entre características de imagens, ou casos, com: Modelo de Contraste, Distância Euclidiana, Distância de Manhattan, Entropia Relativa e Intersecção de Histogramas. Em [SANTINI, 1999] e [VELTKAMP, 2001] são discutidas diferentes abordagens para medidas de similaridade. Essas técnicas podem ser utilizadas em dois contextos diferentes: a) para realizar o cálculo da similaridade local, considerando apenas a relação existente entre cada um dos atributos de forma isolada; ou b) similaridade global, calculando a similaridade entre casos considerando um conjunto de

características definidas para representar esses objetos ou o conjunto de valores de Similaridade Local encontrados para cada uma das características que representam esses casos.

Uma medida, ou métrica, de distância é uma base utilizada para definir o quão diferentes são os conteúdos de dois, ou mais, objetos. A partir desse valor obtido pela análise da medida de distância, é definido um grau de similaridade entre os objetos avaliados. Na Tabela 3 são apresentadas algumas métricas existentes, que dependendo de suas características, podem ser utilizadas tanto para similaridade local quanto para medida global.

**Tabela 3:** Medidas de Similaridade.

<b>Medida</b>	<b>Formulação</b>	
Distância de Hamming	$d(x, y) = \sum_{i=0}^n \delta_i$ , onde $\delta_i = \begin{cases} 1 & \text{se } x_i \neq y_i \\ 0 & \text{outros casos.} \end{cases}$	(35)
Distância de Manhattan	$d(x, y) = \sum_{i=0}^n  x_i - y_i $	(36)
Distância Euclidiana	$d(x, y) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2}$	(37)
Entropia Relativa	$d(x, y) = \sum_{i=0}^n x_i \log_2 \left( \frac{x_i}{y_i} \right)$	(38)

## 2.5 Considerações

O processo de recuperação de imagens baseado em conteúdo, é possível apenas com o uso extensivo de técnicas de processamento digital de imagens. Durante as três últimas décadas, técnicas isoladas foram desenvolvidas para diferentes finalidades. Essas técnicas resultaram num conjunto disperso, de técnicas que podem ser utilizadas para a execução do processo de CBIR. A transformação de imagens médicas em atributos alfanuméricos tratáveis por computador, representa parte fundamental do processo de CBMIR e a correta escolha dos métodos e atributos a serem utilizados, atua diretamente na probabilidade de sucesso ou fracasso de uma metodologia que se destina a realizar esse tipo de tarefa.

## **Capítulo 3**

### **Recuperação de Informação por Similaridade**

Neste capítulo são apresentados alguns dos componentes de um processo de recuperação de informação. O processo descrito é baseado no modelo tradicional para um sistema de Raciocínio Baseado em Casos - RBC [AAMODT, 1994]. São vistas técnicas de classificação de dados, utilizadas na fase de agrupamento dos dados, para identificação das categorias de exames. Serão apresentadas ainda, técnicas para seleção de características que melhor descrevem cada uma das categorias de imagens, visando agilizar o processo de recuperação. São descritas técnicas tradicionais para a avaliação de resultados obtidos com a execução de processos de recuperação de informação e uma proposta para comparação de eficiência entre métodos de recuperação de imagens baseada em conteúdo.

#### **3.1 Raciocínio Baseado em Casos**

O RBC é uma técnica de Inteligência Artificial (IA) Simbólica. Têm por objetivo a resolução de problemas através da recuperação e adaptação de experiências passadas, chamadas de caso, armazenadas em uma base de casos. Uma nova solução é proposta a partir de um processo de adaptação de problemas classificados como similares ao problema em estudo, que já possuem uma solução conhecida.

Um sistema de RBC é composto basicamente por quatro elementos [WANGENHEIM, 2003]:

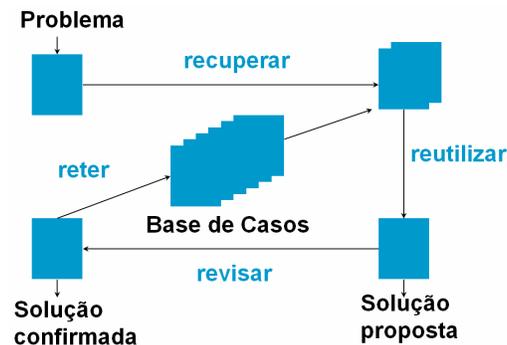
- **Representação do Conhecimento:** consiste na modelagem do conhecimento, ou seja, as informações que serão utilizadas pelo sistema durante o processo de identificação de casos similares e adaptação para geração de uma nova solução. São encontrados quatro diferentes tipos de conhecimento que necessitam ser representados: **vocabulário** que descreve o conhecimento geral do domínio e é utilizado para fazer a identificação de, ou correlação entre, termos; **casos**, são fatos concretos

obtidos a partir de experiências passadas e que são estruturados e armazenados em uma base de casos para que fiquem disponíveis para servirem de exemplo para formulação de novas soluções; **similaridade**, consiste nos elementos que devem ser avaliados para que se tenha condições de estabelecer um grau de semelhança entre os casos existentes na base de casos e um novo caso avaliado, e; **adaptação**, que se caracteriza pelo conjunto de regras ou técnicas que devem ser utilizadas para transformar soluções conhecidas de casos similares, em uma nova solução para um determinado caso avaliado;

- **Medida de Similaridade:** ao se planejar um elemento responsável pelo cálculo de similaridade entre casos em um sistema de RBC, deve ser considerado que é necessário que seja algo capaz de identificar quais, dentre os casos armazenados na base de casos, são os casos mais similares a um novo problema apresentado e, ainda, esse grau de similaridade a ser definido deve estar diretamente ligado ao grau de relevância que esse caso terá na elaboração de uma solução para o novo caso;
- **Adaptação:** a tarefa de se adaptar casos anteriores para auxiliar na solução de novos problemas deve considerar as diferenças existentes entre os casos avaliados e ser capaz de identificar quais as alterações necessárias para composição da nova solução, observando o critério de relevância do resultado do processo de RBC;
- **Aprendizado:** esse elemento é importante para garantir autonomia de um sistema de RBC, por possibilitar que a base de casos seja alimentada sempre que um novo caso for avaliado e uma nova solução for proposta. Em algumas situações pode se optar por avaliar se o resultado gerado pelo sistema foi ou não satisfatório antes de se reter o novo conhecimento.

Na Figura 8 pode-se observar o Ciclo Básico de funcionamento de sistemas de RBC. Esse ciclo é composto pelas fases de *Recuperação*, que identifica os casos similares ao novo problema apresentado; *Reutilização*, na qual é realizado o processo de adaptação das soluções encontradas, de forma a atender o novo caso avaliado; *Revisão*, onde se faz uma análise da relevância do resultado do processo de adaptação, ou seja, é

identificado se a solução proposta é útil para solução do problema apresentado, e; *Retenção*, que realiza o aprendizado novo conhecimento gerado.



**Figura 8:** Ciclo Básico de Sistema de RBC. [WANGENHEIM, 2006].

### 3.2 Técnicas de Classificação de Dados

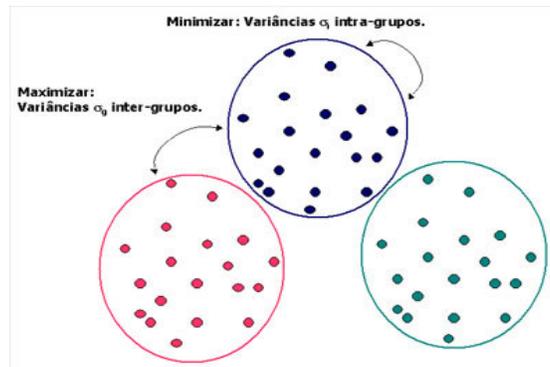
A classificação de dados ocorre quando se deseja identificar a qual classe um determinado objeto, ou caso, analisado pertence. Ainda é necessário para que se possa observar quais as categorias naturais existentes dentro de um conjunto de dados. Para a realização dessas tarefas, são necessárias diferentes técnicas de classificação de dados. Essas técnicas podem ser divididas em: a) *supervisionadas* e b) *não-supervisionadas*. Nesta seção são descritas técnicas pertencentes a esses grupos utilizadas posteriormente na fase de testes deste trabalho.

#### 3.2.1 Não-Supervisionadas

Uma técnica de classificação é considerada como não-supervisionada, ou exploratória, quando trabalha com um conjunto de dados dos quais não se conhecem as informações de categorias existentes, ou seja, compõem apenas um conjunto de observações de um determinado fenômeno. Nessa situação essa técnica é empregada visando efetuar a identificação das classes naturais para os dados desse conjunto e realizar a classificação dos dados de acordo com a sua categoria. Existem diferentes abordagens para classificação exploratória, como: *k-Means* [MACQUEEN, 1965] e *Expectation-Maximization* [DEMPSTER, 1977].

### 3.2.1.1 *k*-Means

O método de agrupamento por *k*-Means (k-Médias) é uma abordagem estatística para o particionamento de conjuntos de casos em classes de similaridade. A idéia central dessa técnica é a maximização da distância entre agrupamentos, de mesmo modo que minimiza as distâncias entre os grupos. A precisão da classificação pode ser avaliada comparando-se a variabilidade intra-grupo (que é pequena se a classificação é boa) e a variabilidade inter-grupos. Os valores obtidos devem ser inversamente proporcionais, ou seja, uma classificação é considerada boa quando a variação intra-grupos for pequena e a inter-grupos grande. Na Figura 9 pode-se observar um exemplo de agrupamento através do método *k*-Médias.



**Figura 9:** Exemplo de Agrupamento *k*-Means [WANGENHEIM, 2006].

Esse algoritmo sempre retornará os melhores agrupamentos para os casos, dado um número *k* de classes. Para isso, um valor *k* representando o número de classes, ou agrupamentos, em que os dados serão divididos deve ser conhecido. Como nem sempre é possível se conhecer a priori essa informação, pode se utilizar um teste estatístico para determinar se a quantidade de agrupamentos formada é o melhor número para representar a população observada. Uma medida de discriminação entre grupos pode ser dada pelo cálculo das variâncias inter e intra-grupos e pelo coeficiente de discriminação *F*. Pode-se utilizar *F* como medida da qualidade de um determinado conjunto de classes como subdivisão natural de um conjunto de dados, para se determinar o número de classes mais adequando para um conjunto de dados.

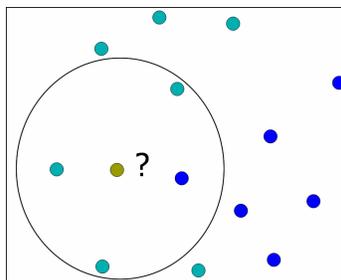
$$F = \frac{\text{variância entre grupos}}{\text{variância dentro dos grupos}} \quad (39)$$

### 3.2.2 Supervisionada

O processo de classificação supervisionado é caracterizado pela existência de um conjunto de dados do qual se conhece, a priori, as categorias em que estes estão divididos e em qual dessas, pelo menos um dos padrões está inserido. A tarefa de classificação consiste em: com base nesse conjunto de dados realizar a categorização de um novo elemento, utilizando para isso a técnica de medida de similaridade para determinar quais são os casos que mais se assemelham ao caso estudado. Uma classe para o novo dado é atribuída de acordo com a classe dos casos selecionados no processo de análise de similaridade. O *k-Nearest Neighbor* [DASARATHY, 1991] é uma das mais tradicionais para esse tipo de tarefa.

#### 3.2.2.1 k-Nearest Neighbor

Esta técnica de classificação é uma variação de *Nearest Neighbor* (Vizinho mais próximo), que ao invés de considerar apenas o vizinho imediato ao caso analisado, utiliza os  $k$  vizinhos mais próximos, durante o processo de classificação. É utilizada uma medida de similaridade pré-definida para determinar o grau de similaridade entre o caso em estudo e os demais elementos conhecidos. Baseado no resultado obtido com esse cálculo é determinada a relação de vizinhança entre os casos. O caso “desconhecido” é classificado de acordo com a classe mais recorrente dentre as classes de seus  $k$  vizinhos. Caso exista igualdade entre duas ou mais categorias deve ser definida um critério para desempate. A Figura 10 apresenta um exemplo de classificação utilizando  $k$ -NN com  $k=4$ . Nesta situação o novo caso será classificado como pertencendo à mesma classe dos casos que obtiveram maior frequência, 03 (três) ocorrências.



**Figura 10:** Exemplo de Classificação  $k$ -NN.  $k=4$  [WANGENHEIM, 2006].

### 3.3 Técnicas para Seleção de Características

A Seleção, ou Mineração, de Características é um processo utilizado para identificar quais, dentre todos os atributos que foram definidos para representar uma determinada classe de casos, possuem maior relevância durante a fase de comparação entre dois elementos desta mesma categoria. Existem diferentes abordagens aplicadas no domínio de imagens médicas que se propõem tratar este problema. Essas alternativas fundamentam-se tanto em métodos estatísticos [QUINLAN, 1993] e [RIBEIRO, 2005], como em técnicas de IA Conexionista, [WERBOS, 1974] e [KOHONEN, 1987]. Independente de sua linha de fundamentação, o objetivo é comum, reduzir a dimensionalidade do vetor de características durante a fase de medida de similaridade entre casos em tarefas de recuperação de informação. Nesta seção serão apresentadas algumas dessas abordagens, ambas estruturadas sob modelos estatísticos.

#### 3.3.1 StARMiner

O método StARMiner – *Statistical Association Rule Miner* (Minerador Estatístico de Regras de Associação), têm por objetivo identificar regras envolvendo as características que melhor descrevem um conjunto de imagens em classes de similaridade [RIBEIRO, 2006]. Usando conceitos de regras de associação estatísticas para identificar os atributos mais relevantes para o processo de medida de similaridade entre casos de uma mesma categoria.

Na Figura 11 é apresentado um algoritmo para o método StARMiner [RIBEIRO, 2005]. Nas linhas 2 a 5 é calculado um valor médio para cada atributo armazenado na base. Nas linhas 6 a 10 mede-se o desvio padrão para cada atributo e um valor  $Z$  que é utilizado no teste de hipótese. Nas linhas 11 e 12 são analisadas regras de restrições. Uma regra só é aceita caso os valores encontrados estejam dentro dos limites ( $\Delta\mu_{min}$ ,  $\sigma_{max}$  e  $\gamma_{min}$ ), informados nos parâmetros de entrada. Onde  $\Delta\mu_{min}$  indica o valor mínimo aceitável para a diferença entre a média de um atributo  $A_i$  em casos de uma categoria  $x$  e a média de  $A_i$  nos demais casos da base;  $\sigma_{max}$  representa o maior valor permitido para o desvio padrão de um atributo  $A_i$  em casos de uma categoria  $x$ , e;  $\gamma_{min}$  indica o valor de confiança mínimo para rejeitar uma hipótese  $H_0$ .

**Algoritmo StARMiner**

**Input:** as instâncias no banco de dados devem estar estruturadas como  $\{x, A_1, A_2, \dots, A_N\}$ , onde  $x$  é a categoria da imagem e  $A_i$  são os atributos que representam a imagem; parâmetros  $\Delta\mu_{\min}$ ,  $\sigma_{\max}$  e  $\gamma_{\min}$ .

**Output:** regras mineradas.

1. **Início**
2. **varre** banco de dados
3. **para** cada atributo  $A_i$
4.     **para** cada categoria  $x$  existente
5.         **calcule**  $\mu_{A_i}(T_x)$  e  $\mu_{A_i}(T - T_x)$
6. **varre** banco de dados
7. **para** cada atributo  $A_i$
8.     **para** cada categoria  $x$  existente
9.         **calcule**  $\sigma_{A_i}(T_x)$
10.         **calcule** o coeficiente  $z$
11.         **if**  $|\mu_{A_i}(T_x) - \mu_{A_i}(T - T_x)| \geq \Delta\mu_{\min}$  e
12.              $|\sigma_{A_i}(T_x)| \leq \sigma_{\max}$  e ( $z < z1$  ou  $z > z2$ )
13.             escreva( $x + \rightarrow + A_i$ )
14. **Fim**

**Figura 11:** Algoritmo StARMiner. Traduzido de [RIBEIRO, 2005].

Na Figura 11  $T$  representa o conjunto de casos existente na base de dados e  $T_x$  o conjunto de casos de uma dada categoria  $x$ . Os cálculos de  $\mu_{A_i}(V)$  e  $\sigma_{A_i}(V)$ , onde  $V$  é um conjunto de imagens, e do coeficiente  $z$ , são realizados através das equações (40), (41) e (42), respectivamente. A Tabela 4 apresenta alguns valores críticos,  $z1$  e  $z2$ , para  $z$ .

$$\mu_{A_i}(V) = \frac{\sum_{j \in V} (A_i)_j}{|V|} \quad (40)$$

$$\sigma_{A_i}(V) = \sqrt{\frac{\sum_{j \in V} ((A_i)_j - \mu_{A_i}(V))^2}{|V|}} \quad (41)$$

$$z = \frac{\mu_{A_i}(T_x) - \mu_{A_i}(T - T_x)}{\frac{\sigma_{A_i}(T_x)}{\sqrt{|T_x|}}} \quad (42)$$

**Tabela 4:** Valores críticos para  $z$ , para  $z = 0,9$ ,  $z = 0,95$  e  $z = 0,99$  [RIBEIRO, 2005].

	<b>0,9</b>	<b>0,95</b>	<b>0,99</b>
<b>z1</b>	1,64	1,96	2,58
<b>z2</b>	-1,64	-1,96	-2,58

### 3.3.2 Customized-Queries

A técnica de Consultas Customizadas (*Customized-Queries*) aborda o problema da seleção de características definindo uma hierarquia para organização dos casos armazenados. Em [DY, 1999a; 1999b] são apresentados resultados com a utilização de três níveis de hierarquia (raiz, Nível 1 e Nível 2), porém a idéia pode ser expandida para suportar níveis de pesquisa mais extensos.

Inicialmente é definido um nó-raiz que contém todos os casos conhecidos. A partir deste nó são definidos outros que compõem o Nível 1, separados de acordo com a classe a qual estes pertencem. A classificação desses casos é previamente conhecida. É aplicado a todos os nós pertencentes ao Nível 1, o algoritmo de agrupamento MLC++ [DY, 1999a]. Esse algoritmo é capaz de determinar as características que melhor discriminam cada um dos subconjuntos de casos nos quais foi aplicado. Em seguida é definido o Nível 2 realizando um processo de agrupamento de casos em cada um dos nós do nível anterior, de forma a gerar diferentes sub-agrupamentos. Após a realização do agrupamento, é executado o algoritmo MLC++ em cada um dos nós gerados no Nível 2 para definir quais os atributos de maior relevância entre eles.

### 3.4 Métricas para Avaliação de Resultados

Quando se realiza um processo de Recuperação de Informação, independente do domínio da aplicação em que se está trabalhando, é importante avaliar o resultado obtido após a execução dessa tarefa. Isso é necessário para se verificar o grau de relevância dos dados recuperados. Geralmente, técnicas para avaliação de resultados são empregadas para determinar a eficiência de uma aplicação de recuperação de informação.

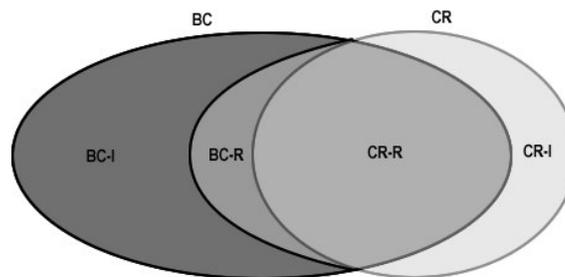
Existem algumas abordagens tradicionais para esse tipo de verificação, como: *Precision & Recall*, *F-Measure* [RIJSBERGEN, 1979] e Cobertura; essas técnicas podem ser aplicadas em qualquer domínio de aplicação. Porém, além de se verificar a qualidade da recuperação em uma determinada metodologia, existe a necessidade de se comparar a eficiência de abordagens diferentes, para isso é necessário que se tenham tarefas de recuperação e bases de casos definidas como “padrão” para esse tipo de avaliação. Nesse sentido é encontrada a proposta descrita em [MULLER, 2001] para a avaliação de métodos de recuperação de imagens, incluindo a área de Recuperação de

Imagens Médicas. Mais a diante serão apresentadas as características dessas abordagens. Durante a fase de validação deste trabalho tais tipos de técnicas são utilizados para verificar a relevância do resultado do processo de recuperação com base no modelo proposto, além de permitir a comparação deste com outros modelos existentes.

### 3.4.1 *Precision & Recall*

Os conceitos de *Precision* e *Recall* têm sido amplamente descritos e adotados na literatura pertinente, em especial quando se deseja determinar o grau de relevância dos resultados de um processo de recuperação, assim como avaliar o método em si.

Na Figura 12 é apresentado um cenário de um processo de recuperação genérico. Esse cenário será utilizado para ilustrar o que representa cada uma dos conceitos de avaliação de resultados. No cenário, *BC* – Base de Casos, contém *BC-I* – Casos Irrelevantes na Base de Casos; e, *BC-R* – Casos Relevantes na Base de Casos. *RC* – Casos Recuperados está dividido em: *CR-R* – Casos Recuperados Relevantes; e, *CR-I* – Casos Recuperados Irrelevantes.



**Figura 12:** Representação de Dados Relevantes e Irrelevantes no processo de Recuperação de Informação.

*Precision* representa a relação existente entre a quantidade de casos relevantes recuperados – *CRR*, no processo de recuperação, e o número total de registros recuperados – *CR*, considerando-se os casos Relevantes (*CRR*) e Irrelevantes (*CRI*). Essa relação pode ser expressa em termos percentuais através da formulação:

$$P = \frac{CRR}{CR} \text{ sendo, que } CR = CRR + CRI. \quad (43)$$

*Recall* representa a relação existente entre a quantidade de casos relevantes recuperados – *CRR*, no processo de recuperação, e o número total de registros relevantes

existentes na base de casos – *BCR*. Essa relação pode ser expressa em termos percentuais através da formulação:

$$R = \frac{CRR}{BCR} \text{ sendo, que } BCR \subset CRR. \quad (44)$$

### 3.4.2 Proposta do SIM/HUG-Viper

Os grupos do SIM/HUG e Viper [VIPER, 2007] propuseram [MULLER, 2001] a utilização de um conjunto de atributos a serem utilizados como parâmetros de comparação entre métodos de recuperação de imagens por conteúdo, incluindo a recuperação de imagens médicas. Postula-se então a necessidade da utilização de uma base de casos comum para todos os métodos a serem comparados e, também, os atributos a serem analisados:

- *Rank<sub>l</sub>* e *Rank Normalizado*: *Rank<sub>l</sub>* representa a posição em que o caso mais relevante é recuperado. O *Rank Normalizado* é a média normalizada das posições das imagens mais relevantes, que pode ser formulado como:

$$\overline{Rank} = \frac{1}{N * N_R} \left( \sum_{i=1}^{N_R} R_i - \frac{N_R(N_R - 1)}{2} \right) \quad (45)$$

onde,  $R_i$  é a posição em que aparece a  $i$ -ésima imagem mais relevante recuperada,  $N$  é o número de casos existentes na base e  $N_R$  o número de casos relevantes recuperados;

- $P(20)$ ,  $P(50)$  e  $P(N_R)$ : esses atributos se referem ao valor de *Precision* após a recuperação da 20-ésima, 50-ésima e  $N_R$ -ésima imagem mais relevante, respectivamente;
- $R_P(0,5)$  e  $R(100)$ : representando o valor de *Recall* ao se atingir o valor 0,5 para o *Precision* e ao se recuperar a 100-ésima imagem mais relevante;
- *PR-Graph* – Gráfico de *Precision* e *Recall*: avaliação do gráfico que cruza os valores obtidos para os atributos de *Precision* e *Recall* durante todo o processo de recuperação.

### **3.5 Considerações**

O processo de recuperação de imagens baseado nas características, ou no conteúdo, das imagens apresenta variações de possibilidade de estruturação, pois o processo de recuperação pode ser baseado em diferentes características visuais, como: cor, forma, textura e identidades dos objetos, dentre outras.

Neste capítulo foram revistas várias técnicas de classificação e seleção de similaridade. O raciocínio baseado em casos, atua neste contexto como um modelo consolidado para realização de recuperação de informação por similaridade de conteúdo. São conhecidas diferentes técnicas para descrição e classificação dos padrões criados a partir da etapa de extração de características descrita no capítulo 2. A seleção das técnicas mais eficientes para a avaliação da similaridade, todavia, implica em validação empírica no domínio de aplicação.

## **Capítulo 4**

### ***Recuperação de Exames Médicos***

*A recuperação de exames médicos é o ato de se buscar por exames anteriormente realizados visando sua reutilização de alguma forma. Essa recuperação pode gerar diferentes resultados, de acordo com o tipo de parâmetro utilizado na realização dessa tarefa. De acordo com cada um dos conjuntos de informações atribuídos a um exame: Descrição do Procedimento, Laudo ou Diagnóstico e Imagens, pode-se definir uma metodologia de recuperação que contemple as particularidades das características desse conjunto. Cada uma dessas abordagens, portanto, é capaz de medir o grau de similaridade entre exames em um nível diferenciado. Ao se utilizar apenas informações que descrevem o procedimento realizado, todavia, não é possível analisar características como, por exemplo, uma anomalia relatada no laudo, tampouco, características referentes às imagens do respectivo exame. Com a avaliação do diagnóstico emitido para um exame, dependendo de como o seu conteúdo foi construído, não é possível traçar correlação com as informações das imagens e, também, da descrição do exame. De forma similar, quando se emprega uma metodologia de recuperação que se utiliza do conteúdo existente nas imagens ou sinais gerados durante a realização do exame, dependendo da organização empregada no armazenamento dos dados do exame, não se tem acesso às informações da descrição do procedimento realizado ou diagnóstico emitido.*

Para que um método de recuperação consiga responder perguntas como: *“Quais os exames já realizados que possuem características semelhantes a um exame atual, de forma a ser capaz de auxiliar na elaboração do diagnóstico para esse caso?”*, é necessário que se tenha acesso ao procedimento utilizado para realizar todos os exames avaliados e, também, aos dados gerados em cada um desses exames. Não é, obrigatoriamente, necessário que se consiga interpretar o conteúdo dos laudos anteriormente elaborados, mas é requisito que exista relação entre um exame e seu respectivo diagnóstico. Uma metodologia de recuperação de exames médicos para atender a todos os possíveis tipos de objetivos com a realização dessa tarefa, deve ser

capaz de receber como entrada, e manipular, qualquer um dos conjuntos de informações relacionados ao exame.

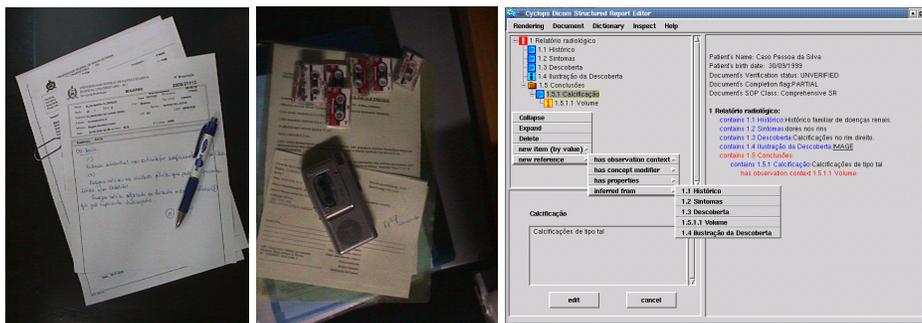
#### **4.1 Conjuntos de Informações em Exames Médicos**

Os diferentes grupos de informação que são encontrados no exame médico são responsáveis por descrever suas características implícitas. Essas características são utilizadas, dentre outras formas, como parâmetros de comparação entre exames durante o processo de recuperação de exames por similaridade. Um Banco de Dados responsável pelo armazenamento de exames médicos, deve ser estruturado de forma a suportar os diferentes conjuntos de informações gerados para representar um exame. Essa estrutura pode variar entre bancos de dados devido às necessidades particulares das modalidades de exames suportadas.

O exame médico em formato DICOM possui informações utilizadas para descrever o procedimento utilizado na sua realização, como: modalidade de exame, região do corpo analisada, orientação do paciente, uso ou não de contraste, informações do equipamento de aquisição, informações do paciente, dentre muitas outras. Essas informações podem ser agrupadas, no que nesse trabalho é definido como *Descrição do Procedimento*. Essas características, embora capazes de identificar exames apenas em um nível de abstração mais elevado, são importantes por gerarem uma descrição pseudo-semântica do exame.

As informações presentes nos Laudos ou Diagnósticos médicos possuem características capazes de fornecer uma descrição acerca da conclusão resultante da análise de um determinado exame. Dentre outras informações, pode-se encontrar a descrição da anomalia encontrada e sugestões de tratamentos posteriores. Essas informações não são em absoluto, suficientes para determinar as características das imagens ou sinais gerados pelo exame e, dependendo da forma em que estejam estruturadas, não são capazes de descrever o procedimento realizado para aquisição da imagem. Porém, a partir dessas informações, o especialista médico consegue extrair dados que lhe auxiliam durante a elaboração do diagnóstico de outro exame similar. As informações extraídas do laudo podem ser agrupadas em um conjunto de características descritivas do exame, denominado neste trabalho como *Descrição do Diagnóstico*.

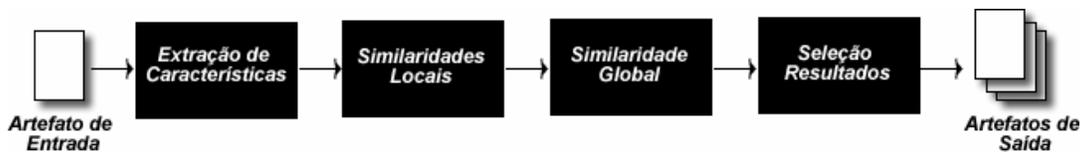
Na Figura 13, são apresentados exemplos de laudos médicos. À esquerda vê-se um exemplo típico de laudo manuscrito. No centro, um exemplo de laudo ditado e gravado em fitas magnéticas. À direita é apresentado um trecho de um arquivo contendo um laudo estruturado.



**Figura 13:** Exemplos de laudos médicos: manuscrito (*esq.*); ditado utilizando gravador de mão (*centro*) e; Interface para elaboração de laudo estruturado [BORTOLUZZI, 2003](*dir.*).

Um outro conjunto importante de informações extraídas a partir de um exame são as características relativas ao *Conteúdo de Imagens ou Sinais* médicos gerados durante a realização do procedimento clínico. Na Figura 1, foram apresentados exemplos de imagens contidas em exames. A partir dessas imagens, são obtidas diferentes informações responsáveis pela descrição léxico-sintática do exame gerado, que podem ser descritas como padrões representativos das imagens. Uma operação de recuperação de exames com base nesse tipo de características não é capaz de identificar completamente o procedimento adotado para realização do exame, tampouco, sugerir qual o seu diagnóstico. Porém, com essas informações é possível criar uma base de informações para posteriormente realizar recuperação de imagens ou sinais similares aos do exame em estudo.

Na Figura 14, é apresentado um modelo simplificado de recuperação de exames que pode ser utilizado independente do artefato de entrada, *Descrição de Procedimento*, *Diagnóstico*, *Imagem ou Sinal*. Esse modelo é similar aos modelos padrão para recuperação de imagens por conteúdo, encontrados na literatura.



**Figura 14:** Modelo Simplificado para Recuperação de Exames.

Seguindo o modelo simplificado apresentado, inicialmente realiza-se a extração de características do artefato passado como parâmetro. Essas características são então organizadas em um Vetor de Características; e em seguida, são calculadas as similaridades locais para cada um dos atributos, utilizando para tal, uma medida de similaridade pré-definida. Na seqüência, é realizada uma medida de similaridade global para determinar o grau de dissimilaridade entre o objeto de consulta e os objetos armazenados na base de dados, e; são selecionados os que obtiverem um resultado dentro do grau da taxa de similaridade definida como aceitável.

## **4.2 Recuperação Baseada na Descrição do Procedimento**

Para realizar uma operação de recuperação utilizando as informações de descrição do procedimento adotado na geração do exame médico, devem ser selecionadas quais dentre as características serão utilizadas no processo de medida de similaridade. A abordagem a ser adotada para esse acesso dependerá da forma como as características estão armazenadas. Tal como no modelo simplificado apresentado na Figura 14, são aplicadas a seguir as medidas de similaridades locais que melhor se adaptam a cada um dos tipos de valor de atributos. Um grau de similaridade global é gerado entre cada exame analisado e o exame em estudo. Os exames que possuem similaridade dentro do limite pré-estabelecido são recuperados para posterior apresentação ao usuário.

O resultado obtido com um método de recuperação de exames que utilize apenas características do grupo de *descrição do procedimento*, certamente pode classificar como semelhantes exames que possuem atributos, imagens, sinais ou laudos, visivelmente dissimilares. Porém, de acordo com os atributos considerados, possuem estes um grau de similaridade aceitável. Esse processo realiza um tipo de agrupamento pseudo-semântico

dos exames e, pode ser utilizado como técnica de redução do espaço de busca, a ser utilizado em adição aos demais modelos de recuperação que serão apresentados a seguir.

O processo de Recuperação descrito é indiferente à existência ou não de um padrão na estruturação do exame, como nos exames padrão DICOM. A diferença a ser observada em processos nos quais são utilizados exames não padronizados, com os que possuem padronização, reside no momento da extração das características pseudo-sintáticas. A partir do momento em que se têm acesso às características que serão utilizadas para a medida de similaridade, a aplicação do processo de medida de similaridade e seleção de resultados ocorre de maneira similar.

Um ponto a ser observado nesse tipo de processo para exames não padronizados, é a importância em se conseguir manter uma relação entre as características classificadas como de *descrição do procedimento* e os demais artefatos gerados durante a realização do exame, imagens ou sinais e laudo. A capacidade de se manter essa relação é essencial para o processo de recuperação de exames, caso contrário, de pouco adianta analisar um conjunto de informações descritivas acerca de procedimentos de realização de exames e não ser capaz de recuperar as imagens e laudos gerados em cada um desses exames.

Uma observação importante a ser feita quando se trata do processo de Recuperação Baseada na Descrição do Procedimento se refere ao conteúdo dos atributos analisados, informações de procedimentos e pacientes, é de caráter textual e são inseridos manualmente. Essa característica faz com que seja necessário dispensar especial atenção ao se trabalhar com esse conteúdo devido a possibilidade de ocorrência de erros durante a digitação dos dados ou, ainda, com a utilização de termos sinônimos para se referenciar a uma mesma característica. A utilização de um dicionário de dados pode ser utilizado para auxiliar a comparação entre termos sinônimos.

#### **4.2.1 Em Exames DICOM**

Em exames realizados em equipamentos que exportam os dados de acordo com o padrão DICOM, as informações que descrevem o paciente, o procedimento e o equipamento utilizados para a realização dos exames são incluídas juntamente com cada um dos arquivos de imagem, ou sinal, gerados. Essas informações são o que é conhecido como *header* (cabeçalho) de um arquivo DICOM. Além de descrever o procedimento utilizado na realização do exame, essas informações são utilizadas para caracterizar um

determinado arquivo e identificá-lo dentro de um conjunto de arquivos onde este esteja armazenado.

Essas informações são organizadas dentro dos arquivos de exames em grupos conhecidos como *DE – Data Elements* (Elementos de Dados). A Tabela 5 apresenta exemplos de *DE's* existentes num arquivo DICOM e que são passíveis de serem utilizados para identificação ou comparação de similaridade entre exames. Para maior detalhamento dos *DE's* consulte [DICOM, 2006c].

**Tabela 5:** Exemplos de *Data Elements*.

<b>Tag</b>	<b>Nome</b>	<b>Descrição</b>
(0008,0008)	<i>Image Type</i>	Identifica o Tipo de Imagem
(0008,0016)	<i>Sop Class UID</i>	Identifica a Modalidade da Imagem
(0008,0020)	<i>Study Date</i>	Identifica a Data do Estudo
(0008,0021)	<i>Series Date</i>	Identifica a Data da Série
(0008,0022)	<i>Acquisition Date</i>	Identifica a Data de Aquisição da Imagem
(0008,0090)	<i>Referring Physicians Name</i>	Identifica o Médico Solicitante
(0008,1030)	<i>Study Description</i>	Descrição do Estudo
(0008,103e)	<i>Series Description</i>	Descrição da Série
(0010,0010)	<i>Patients Name</i>	Identifica o nome do paciente
(0010,0020)	<i>Patients ID</i>	Identifica o código do paciente
(0018,0015)	<i>Body Part Examined</i>	Identifica a região examinada
(0020,0010)	<i>Study ID</i>	Identifica o estudo
(0020,0011)	<i>Series Number</i>	Identifica a serie
(0020,0032)	<i>Image Position Patient</i>	Posicionamento da Imagem em Relação ao Paciente

Para que se tenha maior agilidade na execução de tarefas de recuperação que se utilizem das informações disponíveis no cabeçalho DICOM é importante que seja previsto em tempo de armazenamento dos exames quais as características serão analisadas durante o processo. As informações selecionadas devem ser armazenadas de forma independente do arquivo. Um exame DICOM pode ser entendido como uma seqüência (*stream*) de bytes que armazena em um único arquivo, todos os dados pertencentes ao exame. Caso não se tenha esse tipo de atenção com o armazenamento das características, sempre que for necessário acessar algum atributo de um exame deverá ser realizada pesquisa na *stream* em busca da informação desejada.

### **4.3 Recuperação Baseada no Diagnóstico**

A recuperação de exames baseada no diagnóstico utiliza como referência para a comparação entre exames o conteúdo dos laudos emitidos para os exames. O funcionamento, geralmente, é localizar exames que contenham diagnósticos semelhantes observando-se os conjuntos de palavras contidos no texto do laudo.

Um documento que contém um diagnóstico médico, caso não siga um padrão estruturado, consiste em conjuntos de informações expressas em linguagem natural, o que dificulta o processo de identificação de padrões de informações neles existentes. Para solucionar essa dificuldade é necessário utilizar dicionários de termos globalmente aceitos que são utilizados para elaboração dos laudos. Quando se necessita referenciar algum tipo de informação, ou característica da informação, que pode possuir diferentes interpretações dentro da mesma área de estudo, faz-se uso desse tipo de dicionário, possibilitando uma correta interpretação das informações contidas no documento. Os estudos realizados em [ALEXANDRINI, 2005] tentam encontrar soluções para este tipo de recuperação.

Mesmo com a utilização de dicionários de termos ainda se faz necessário a definição de um dicionário de sinônimos. Através desse tipo de técnica, é possível se identificar a relação existente entre as possíveis variações de termos médicos que são utilizados para descrever uma mesma característica ou informação existente num diagnóstico elaborado para um exame.

Existem diferentes formas de elaboração do diagnóstico por parte do especialista médico:

- **Laudos Manuscritos:** nessa modalidade o especialista elabora o texto do laudo de próprio punho. Posteriormente uma outra pessoa realiza a digitação do laudo. Deve ser registrado que nesse tipo de processo existe uma recorrente presença de erros de digitação, geralmente ocasionados pela dificuldade de interpretação do conteúdo do laudo;
- **Laudos Ditados:** em laudos elaborados com essa abordagem, antes de se executar o processo de análise de linguagem natural, têm-se a necessidade de se converter o áudio gerado durante a realização do laudo para o formato de dados textuais. Porém, nesse modelo é eliminada a

possibilidade de erro de interpretação do conteúdo do laudo, existente no modelo anterior;

- **Laudos Estruturados:** o modelo de laudo estruturado independe da forma como o diagnóstico foi efetuado pelo especialista, manuscrito ou ditado. Essa abordagem se refere à estruturação do conteúdo do laudo, após ter sido digitalizado em forma textual. Com isso se possibilita um maior nível de identificação do conteúdo existente no laudo, além de facilitar a tarefa de traçar correlação entre conteúdo de laudos de diferentes especialistas. A partir da existência de documentos de diagnósticos médicos estruturados de acordo com um padrão conhecido, podem ser realizadas consultas ao conteúdo desses documentos através das características das informações contidas nos documentos.

O padrão DICOM estabelece um formato, ou melhor, um padrão para estruturação de documentos médicos, incluindo laudos, denominado DICOM SR – *Structured Report* (Relatório Estruturado). Com a utilização do DICOM SR é possível obter laudos médicos que possuem uma mesma estrutura de formatação, o que possibilita o processo de identificação de características do conteúdo existente em diferentes documentos. Uma descrição detalhada do padrão DICOM SR pode ser obtida em [CLUNIE, 2000].

O processo de recuperação de exames com base no conteúdo da informação descrita no diagnóstico médico possui uma maior complexidade de concepção do que a recuperação baseada apenas na *Descrição do Procedimento*. Esse fato se dá devido à necessidade de interpretação semântica do conteúdo do laudo emitido para o exame. Essa etapa deve ser realizada antes da fase de extração de características e conseqüente mensuração de similaridades locais e globais. Isso serve para que sejam evitados erros durante a análise de termos sinônimos ou, mesmo, a interpretação de falsos positivos, quando da existência de termos negados.

#### **4.4 Recuperação Baseada no Conteúdo de Imagens**

O processo de recuperação de exames baseada no conteúdo das imagens apresenta variações de possibilidade de estruturação, pois o processo de recuperação pode ser fundamentado com base em diferentes características léxico-sintáticas, como: cor, forma, estrutura, textura e identidades dos objetos, entre outras. A idéia central da recuperação baseada nesse tipo de características é poder localizar um conjunto de imagens similares à imagem passada por parâmetro. Esse parâmetro é identificado durante o processo como Imagem de Consulta - IC.

Após analisadas e convertidas em valores alfanuméricos, as características extraídas são organizadas em uma estrutura de informações, que passa a representar a imagem descrita. A vetorização de características facilita o processo de elaboração de consultas e comparações necessárias na tarefa de recuperação de imagens por conteúdo. Sistemas que suportam esse modelo sem considerar as particularidades existentes entre grupos distintos de imagens, ou sinais, apresentam discrepância entre o resultado obtido e resultados esperados de uma consulta. Essa diferença no resultado, encarada como taxa de erro da operação de recuperação, é indesejada em todos os tipos de sistemas de recuperação de imagem por conteúdo, mas, em se tratando de uma aplicação na área médica, a existência de um grau elevado na taxa de erro se apresenta como um complicador difícil de ser superado no instante da implantação desse tipo de sistema no dia-a-dia médico, despertando desconfiança e rejeição por parte dos especialistas. A exemplo dos demais possíveis processos de recuperação para exames médicos, o modelo básico para recuperação baseada no conteúdo de imagens pertencentes ao exame, possui uma seqüência de passos que pode ser definida como apresentado na Figura 14.

#### **4.5 CBMIR x RBC**

As metodologias de CBMIR são abordadas na literatura como sendo aplicações de CBIR com o objetivo de recuperar imagens obtidas a partir de exames médicos de diagnóstico por imagem. CBIR é o resultado da união de técnicas de Recuperação de Informação com métodos de Reconhecimento de Padrões em Imagens. Ou seja, é uma seleção de técnicas que com base em uma determinada imagem, tida como IC, pesquisa

em uma base de dados contendo diferentes tipos de imagens e recupera um subconjunto de imagens mais similares à imagem de referência. É desejável que uma metodologia de CBIR seja capaz de selecionar os atributos a serem extraídos de acordo com a classe de imagem analisada. Para realizar o processo de cálculo de similaridade, são empregadas técnicas de medida de similaridade, que realizam a tarefa de comparação entre os objetos tomando por base os valores das características léxico-sintáticas extraídas das imagens analisadas.

Esse tipo de abordagem não considera o objetivo da recuperação, ou seja, para que se procedeu com uma análise por imagens mais similares a uma determinada imagem. Ao contrário de aplicações de CBIR em domínio aberto, quando se trabalha em aplicações médicas se tem um objetivo muito claro quando se utiliza deste tipo de ferramenta: encontrar diagnósticos anteriormente atribuídos a um determinado tipo de exame, de forma a contribuir na elaboração do laudo para um caso em estudo para o qual não se tem plena certeza do diagnóstico. Quando este cenário é compreendido, pode-se perceber que a utilização de uma metodologia que se preocupe apenas com a tarefa de recuperar imagens mais similares, não considerando sua finalidade, terá dificuldade em atender às necessidades do usuário médico.

A RBC é uma metodologia que permite a reutilização de observações de ocorrências de um dado fenômeno em um domínio de aplicação, conhecidas como casos, armazenados em uma BC, de forma a auxiliar no processo de tomada de decisão em novas situações. Um caso pode ser entendido como um par  $(p,s)$  onde:  $p$  é um problema e  $s$  é uma solução. A solução pode conter informações adicionais. No caso de exames médicos, temos:  $p$  sendo o conjunto de características que descrevem o exame,  $s$  como sendo o conjunto de artefatos que compõem o exame, imagens e laudos.

Com isso, a tarefa de recuperação de exames médicos pode ser entendida como uma forma de aplicação de RBC integrada a funções de CBIR e com a capacidade de retornar múltiplos artefatos a partir de padrões com diferentes atributos que representam o modelo do domínio, de acordo com a sua classe, de forma a auxiliar no processo de tomada de decisão em novas situações. É importante ressaltar que para que se possa recuperar as informações de diagnósticos ou quaisquer outras informações relacionadas a um exame, essas informações precisam estar disponíveis para acesso. Além disso, durante a fase de modelagem dos casos, ou representação do conhecimento, devem ser

definidas as relações existentes entre todos os artefatos pertencentes a um exame: imagens, laudos, etc.

## **4.6 Considerações**

A abordagem de recuperação de exames a ser adotada por uma ferramenta que se proponha a realizar esse tipo de tarefa é dependente do objetivo final do processo, ou seja, o que se pretende com o resultado obtido na recuperação. Quando se deseja auxiliar na elaboração de um novo diagnóstico, deve-se garantir que o resultado da recuperação fornecerá os artefatos necessários para contribuir com esse objetivo. Isso implica que a metodologia de recuperação deve avaliar o conteúdo do exame e o contexto no qual esse conjunto de informações está inserido. Ainda é desejável que se mantenham as relações entre os diferentes componentes de um exame: descrição do procedimento utilizado, imagens e laudo. Uma proposta de um modelo para recuperação possuindo essas características é apresentado no Capítulo 5.

## ***Capítulo 5***

### ***Metodologia para Recuperação de Exames Médicos DICOM por Similaridade de Características Visuais***

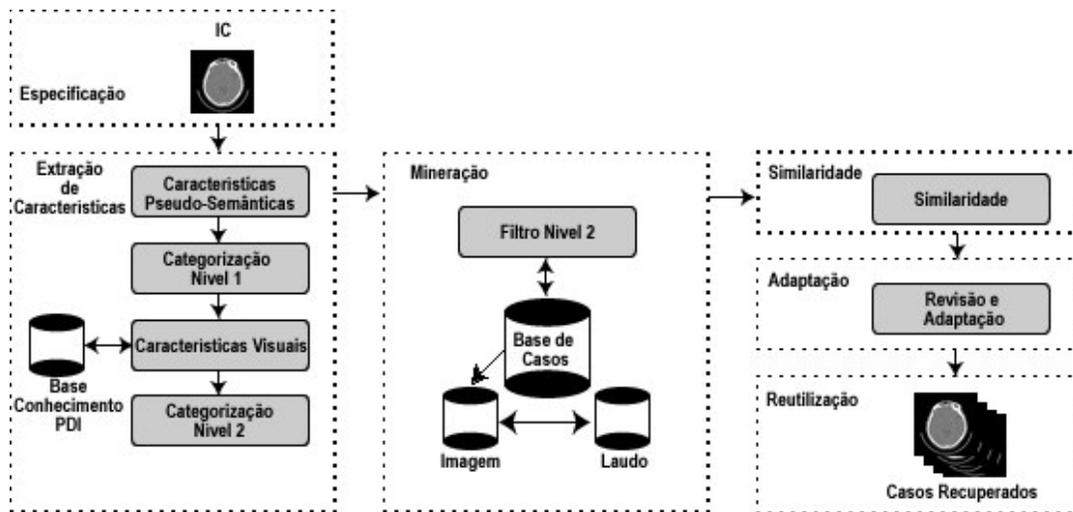
Neste Capítulo é apresentada uma metodologia para Recuperação de Exames Médicos que utiliza características extraídas de imagens contidas no exame como parâmetros para comparação entre eles. A metodologia proposta é composta por diferentes camadas. Cada uma dessas camadas contém módulos responsáveis pela realização de uma tarefa específica. As camadas e seus módulos são descritos de forma a facilitar a compreensão da interação existente entre eles e da metodologia como um todo.

#### ***5.1 Metodologia Proposta***

As abordagens de recuperação de imagens médicas baseadas no conteúdo encontradas na literatura procuram trabalhar os problemas da área médica sem muitas vezes considerar as particularidades do objeto que está sendo analisado ou não se preocupando com a necessidade de reutilização do resultado desse processo no apoio a tomada de decisões médicas. Algumas das propostas verificadas possuem características importantes e que devem ser consideradas em uma proposta que tenha por objetivo servir de modelo para a tarefa de recuperação de exames.

Para atender às necessidades da área médica é proposta uma metodologia para recuperação de exames que contempla as características comumente encontradas num exame médico de diagnóstico por imagem: estruturação de dados seguindo o padrão DICOM, tratamento diferenciado para cada tipo de modalidade de exame e região do corpo observada e as relações existentes entre os diferentes artefatos que compõem o exame; integrando ou adaptando características consideradas relevantes encontradas em outros trabalhos acadêmicos de referência nessa área, com o objetivo de construir uma metodologia que sirva como referência para recuperação de exames médicos.

A metodologia de recuperação de exames proposta utiliza a abordagem de Consultas por Exemplos. A partir de uma imagem recebida como parâmetro, IC, é extraído um conjunto de características, módulo de Extração de Características, que irão representar essa imagem durante o processo de recuperação. Utilizando-se dos atributos pseudo-semânticos extraídos da IC é realizada uma consulta à base de dados, módulo Recuperação, que contém todos os exames disponíveis armazenados na forma de casos, ou seja, representados por um conjunto de características extraídas de seu conteúdo. Os casos selecionados nesse processo de consulta são comparados com a IC, módulo de Similaridade. Com base no grau de similaridade encontrado os casos que satisfizerem o limiar de relevância aceitável definido para o processo de recuperação são selecionados. Esses casos podem passar por um processo de adaptação antes de serem retornados como resultados do processo de recuperação, módulo de Adaptação, ou reutilizados em sua forma original. A Figura 15 apresenta uma representação das camadas e componentes pertencentes à metodologia de recuperação proposta neste trabalho.



**Figura 15:** Camadas e Componentes da Metodologia para Recuperação de Exames Proposta.

## 5.2 Módulo de Especificação

O Módulo de Especificação é o ponto inicial da metodologia de recuperação. É através desse módulo que são interpretados os arquivos de exames padrão DICOM. Para realizar essa tarefa é necessário utilizar-se dos conceitos descritos na documentação do

padrão, para que seja possível se separar o conteúdo pseudo-semântico dos dados da imagem gerada durante a realização do exame. Segundo o DICOM um arquivo de exame contém ambos os tipos de informação.

Os dados interpretados a partir do exame são armazenados em uma estrutura de dados capaz de representar tanto as informações descritivas quanto os *pixels* da imagem, ou pontos de intensidade do sinal médico. É essa estrutura que é repassada ao Módulo de Extração de Características para realizar a análise e extração das características relevantes ao processo de recuperação.

### **5.3 Extração de Características**

A estrutura contendo as informações obtidas a partir da leitura de um exame médico é o parâmetro de entrada para o módulo de Análise de Características. Neste módulo são realizadas as operações de extração de características pseudo-semânticas, categorização de exames baseada nas informações obtidas do cabeçalho DICOM, extração de características visuais e categorização de exames de acordo com as características extraídas da imagem, nessa ordem, cada uma em um componente específico.

#### **5.3.1 Características Pseudo-Semânticas**

Utilizando-se de um processo de seleção de características rígido, ou seja, o conjunto de informações a ser analisado é fixado anteriormente, o componente de extração de características pseudo-semânticas verifica os dados obtidos a partir da leitura do conteúdo descritivo do exame, ou cabeçalho DICOM. Essas informações, se analisadas em conjunto, provêm um grau de descrição acerca do conteúdo semântico do exame. Com isso é possível se identificar dados relevantes sobre o objeto analisado, mas não fornece referência completa do contexto em que está inserido.

As informações consideradas são apresentadas na Tabela 6. A seleção dessas características foi feita através da observação de resultados de testes em laboratório e, também, relatos descritos em outros trabalhos da área, como [GÜLD, 2002], e complementada com o conhecimento obtido com a análise da documentação oficial do padrão DICOM [DICOM, 2006b].

**Tabela 6:** Informações DICOM consideradas no processo de Recuperação de Exames.

Tag	Nome	Descrição
(0008,0060)	<i>Modality</i>	Identifica a Modalidade da Imagem. Possui valores pré-definidos.
(0008,0018)	<i>Sop Instance UID</i>	Identificação Única da Imagem
(0020,000d)	<i>Study Instance UID</i>	Identificação Única do Estudo
(0020,000e)	<i>Series Instance UID</i>	Identificação Única da Série
(0018,0015)	<i>Body Part Examined</i>	Identifica a região do corpo examinada
(0028,0004)	<i>Photometric Interpretation</i>	Identifica a forma de interpretação dos <i>pixels</i> da imagem. Possui valores pré-definidos.

O atributo de Modalidade do Exame – *Modality* – é útil para eliminar a ocorrência de comparação desnecessária entre casos de modalidades diferentes. Os valores possíveis para este atributo são definidos no padrão DICOM [DICOM, 2006b]. Para os tipos de exames considerados neste trabalho, os valores esperados para este atributo são: *CT*, *MR* e *US*.

De acordo com o conhecimento obtido acerca do domínio, outros atributos foram extraídos do exame com o mesmo objetivo de se evitar comparação entre instâncias incompatíveis: *Body Part Examined* – Região do Corpo Examinada, e *Photometric Interpretation* – Interpretação Fotométrica. Estes atributos se referem, respectivamente, à região do corpo que está sendo analisada durante a realização do exame e a forma de interpretação dos *pixels* da imagem. Ambos os atributos possuem valores pré-definidos na documentação do padrão. Para a Região do Corpo Examinada é esperado um dentre os seguintes valores: *SKULL*, *CSPINE*, *TSPINE*, *LSPINE*, *SSPINE*, *COCCYX*, *CHEST*, *CLAVICLE*, *BREAST*, *ABDOMEN*, *PELVIS*, *HIP*, *SHOULDER*, *ELBOW*, *KNEE*, *ANKLE*, *HAND*, *FOOT*, *EXTREMITY*, *HEAD*, *HEART*, *NECK*, *LEG*, *ARM* e *JAW*. Os valores possíveis para a Interpretação Fotométrica são: *MONOCHROME1*, *MONOCHROME2*, *PALETTE COLOR*, *RGB*, *YBR\_FULL*, *YBR\_FULL\_422*, *YBR\_PARTIAL\_422*, *YBR\_PARTIAL\_420*, *YBR\_ICT* e *YBR\_RCT*.

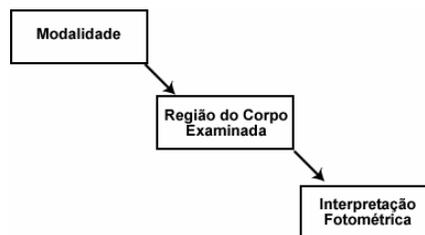
Os atributos de Identificação Única de Estudo, Série e Imagem – *Study Instance UID*, *Series Instance UID* e *SOP Instance UID* respectivamente são utilizados para que se possa manter as relações existentes entre imagens, séries e estudos definidas no DICOM. Isso serve para que não sejam comparadas imagens pertencentes a um mesmo estudo, pois a probabilidade de que estas sejam selecionadas como mais similares é alta e, nesta situação as chances de contribuírem na elaboração de um novo diagnóstico são

consideradas baixas. Além disso, as identificações de séries de imagens são úteis para se estabelecer o relacionamento com o laudo atribuído a um exame, que é composto por uma série de imagens, e, que em caso de laudo estruturado, pode conter referência a uma ou mais imagens dentre as de sua série.

A informação de Plano de Corte da imagem é uma das características que permitem identificar similaridade entre imagens. Porém para a análise dessa informação é necessário que se considerem particularidades em cada um dos tipos de exames. Devem, ainda, ser observadas diferenças entre processos de aquisição das imagens. Ou seja, como o operador do equipamento realizou o ajuste, quais foram os pontos de referência para centralizar as imagens. Um aprofundamento nesta linha foge ao escopo deste trabalho.

### 5.3.2 Categorização Nível 1

A fase de categorização de exames é executada para classificar a IC como pertencente a um dos grupos formados a partir da análise dos diferentes valores possíveis, conforme descrito no DICOM, para os dados pseudo-semânticos considerados. O objetivo desta classificação é organizar os casos em *clusters* de ocorrências similares. Esse agrupamento serve para identificar quais as características que devem ser extraídas da IC. Também auxilia na organização dos casos armazenados na base de casos. Durante a execução da tarefa de recuperação, a categorização da IC, neste nível, é realizada através da identificação dos dados obtidos na extração de características pseudo-semânticas, adotando-se a hierarquia apresentada na Figura 16.



**Figura 16:** Hierarquia para Classificação, Nível 1.

### 5.3.3 Características Visuais

Os atributos visuais a serem extraídos do exame representam as características léxicas, sintáticas e morfológicas da imagem. Com o auxílio de técnicas de mineração de regras de associação e a construção de uma base de conhecimentos, adotou-se uma

abordagem de especialização de características a serem extraídas de acordo com a categoria da imagem analisada. Com isso se pretende reduzir a dimensionalidade do vetor de características que representam um dado caso. Essa redução agiliza o processo de medida de similaridade entre casos. Além disso, aumenta o grau de relevância dos resultados do processo de recuperação.

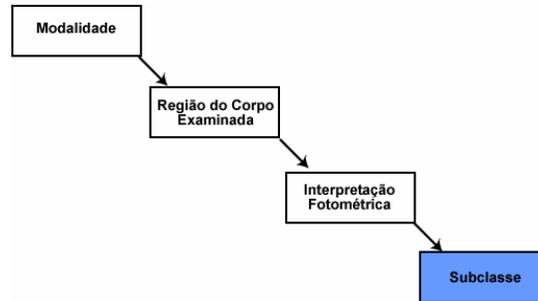
Uma vez que se sabe qual o conjunto de características que melhor discriminam uma determinada classe de imagens, todas as imagens pertencentes a essa classe serão representadas através desse grupo de atributos. A informação de quais os métodos e atributos a serem utilizados para a extração das características visuais de cada um dos tipos de exames DICOM existentes é repassada pela Base de Conhecimento. Este conhecimento deve ser adquirido anterior a construção da Base de Casos.

Após a extração das características, é construído um vetor multidimensional que contém todas essas informações, denominado *Vetor de Características*. Um vetor de características extraídas da IC é que será utilizado para realização da medida de similaridade entre vetores de características gerados para os casos armazenados na base de casos.

#### **5.3.4 Categorização Nível 2**

Na segunda etapa de classificação de exames é utilizado como parâmetro os atributos visuais extraídos da imagem. O Vetor de Características do exame é submetido ao método de classificação. Um exemplo deste tipo de método é o  $k$ -NN. Durante o processo de recuperação, são comparados os atributos do exame com os valores médios de atributos obtidos para cada um dos subgrupos gerados, durante a fase de construção da Base de Casos, para a Classe, nível 1, ao qual esse exame pertence.

O objetivo deste segundo nível é reduzir o espaço de busca por casos similares ao exame analisado. Depois de realizado esse processo de classificação, a hierarquia de busca é representada pelo caminho apresentado na Figura 17.



**Figura 17:** Hierarquia para Classificação, Nível 2.

## **5.4 Mineração**

O Módulo de Mineração é o responsável por realizar o acesso a Base de Casos e recuperar o conjunto de casos que serão utilizados na fase de medida de similaridade. Para isso aplica um processo de filtragem de conteúdo na base de casos. Este filtro tem por objetivo selecionar apenas os casos pertencentes a mesma categoria da IC e, ainda, filtrar casos pertencentes ao mesmo Estudo. Para verificar se os casos pertencem a mesma classe é adotada a hierarquia de classificação de exames, Figura 17, recuperando apenas casos que pertençam à subclasse desejada. Essa tarefa é executada pelo componente de Filtro de Categorias.

Na figura esquemática da metodologia é possível observar que os casos armazenados na base mantêm uma relação com as imagens do exame, possivelmente armazenadas em uma base de imagens, e que estas imagens podem estar relacionadas a um diagnóstico. A existência deste relacionamento é importante para se garantir a recuperação de todos os artefatos que compõem o exame: dados descritivos, imagens e laudos. São estes artefatos que serão apresentados ao especialista médico para auxiliá-lo no processo de tomada de decisão. Antes disso, podem passar por um processo de adaptação, o módulo de Adaptação é descrito na Seção 5.6.

## **5.5 Similaridade**

No módulo de Similaridade é realizada a mensuração do grau de similaridade entre o vetor de características da IC e os vetores construídos para os casos selecionados

no módulo de Mineração. O resultado final deste módulo é o conjunto de casos recuperados da BC com seus respectivos graus de similaridade com a IC calculados.

A abordagem adotada nesta proposta visa agilizar o processo de recuperação, minimizando o número de cálculos de similaridade. Para isto, em tempo de construção da BC é gerado um valor médio para cada um dos atributos pertencentes a uma determinada classe. O conjunto de valores médios representa um caso-médio dessa classe. Em seguida, todos os casos classificados como pertencentes a essa classe tem sua similaridade calculada com esses valores médios.

Durante o processo de recuperação, é calculada a distância entre a IC e os valores médios para os atributos da classe à qual ela pertence. A similaridade entre os casos e a média da sua classe é definida através do valor de Entropia Relativa de seus atributos. A distância entre a IC e os casos recuperados no módulo de Mineração é medida utilizando a Distância de Manhattan entre o valor de distância da IC para o caso-médio e os valores obtidos para cada um dos casos selecionados e a média.

## **5.6 Adaptação de Casos**

Este módulo é utilizado para possibilitar a transformação do conteúdo encontrado nos laudos atribuídos aos casos recuperados como mais similares a IC e que passaram pela etapa da filtragem de relevância. Para isso é necessário o auxílio de um especialista no domínio, um profissional médico com conhecimento na área do exame analisado, que fornece o conhecimento acerca das informações que podem ser reaproveitadas e indicar o que deve ser modificado para facilitar a reutilização dos laudos recuperados para a elaboração de um novo diagnóstico.

Para realização do processo de adaptação, ainda faz-se necessário a utilização de técnicas computacionais, como Processamento de Linguagem Natural, para a identificação automática do conteúdo dos laudos médicos. A aplicação deste tipo de tecnologia não é o foco deste trabalho.

É importante destacar que a não inclusão deste módulo em uma ferramenta de recuperação de exames não elimina a possibilidade da utilização dos laudos dos casos retornados como similares ao exame analisado. O processo de reutilização pode ser

realizado de forma manual pelo especialista, avaliando e aproveitando cada um dos diagnósticos que sejam julgados como relevantes.

## **5.7 Reutilização**

Este módulo possui um Filtro de Relevância que tem o objetivo de excluir do conjunto de resultados do processo de recuperação os casos que mesmo pertencendo à mesma classe da IC possuem um grau de similaridade global abaixo de um limite de tolerância definido como aceitável. Com isso se aumenta o nível de relevância dos resultados da recuperação e evita que sejam retornados casos que sabidamente não terão utilidade no processo de construção do diagnóstico para o novo caso. Este Filtro pode ser implementado com a utilização de um *threshold* para o valor de similaridade global entre os casos analisados.

O resultado do processo de filtragem de relevância, consistindo de um conjunto de par  $(c,s)$ , onde  $c$  é um caso armazenado na BC e  $s$  o grau de similaridade deste com a IC, é organizado para ser retornado ao especialista médico. O conjunto de funcionalidades incluídas neste módulo e suas características é dependente da ferramenta que implementa a metodologia de recuperação. Em sistemas mais elaborados, pode sugerir automaticamente, com base nos resultados, qual o diagnóstico a ser atribuído para o caso em estudo. Ou, mais simplesmente, retornar o conjunto de laudos atribuídos a cada um dos casos em ordem de similaridade e deixar a tarefa de verificação do seu conteúdo a encargo do próprio especialista.

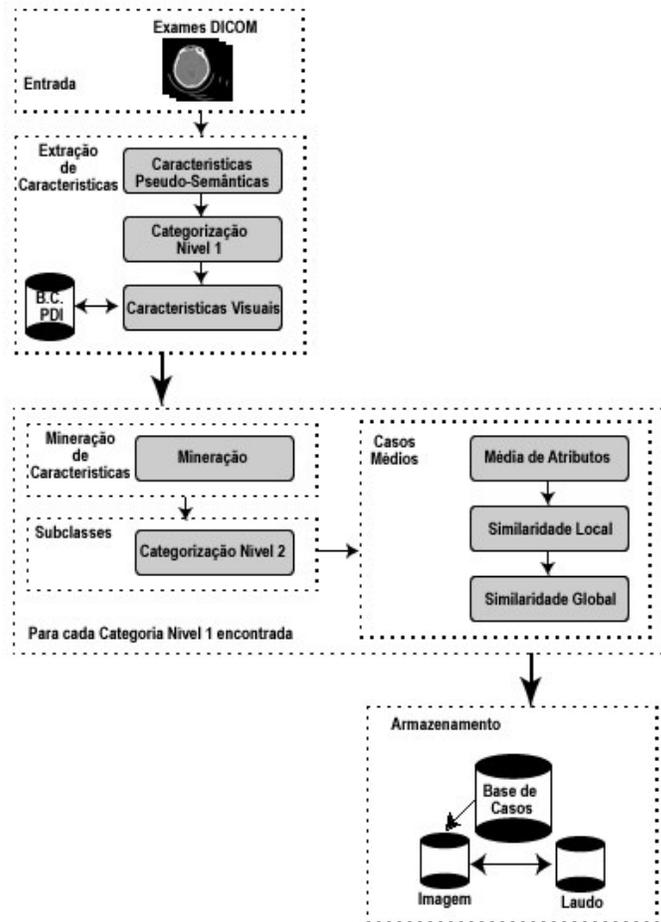
## **5.8 Construção e Manutenção da Base de Casos**

Em uma metodologia em que se utiliza de um conjunto de observações de fenômenos, casos, de um determinado domínio para realizar um processo de recuperação de instâncias similares a uma nova observação, com qualquer que seja o objetivo final, é necessário que se conheça a priori as características desse domínio. Essas características irão permitir a construção de uma Base capaz de armazenar todos os casos observados.

No caso de exames médicos é desejado que a Base de Casos seja capaz de representar tanto as características que descrevem o procedimento utilizado para

realização de um exame, como, também, um conjunto de atributos que consigam identificar o conteúdo presente nas imagens que compõem esse exame. Além disso, é esperado que se mantenha as relações existentes entre os exames e seus respectivos diagnósticos.

Para agilizar o processo de recuperação de exames a BC deve ser construída a priori. Os exames devem ser categorizados, em ambos os níveis, e armazenados de forma a facilitar recuperar os casos de uma mesma categoria. Na Figura 18 é apresentada a estrutura definida para a construção e manutenção de uma Base de Casos.



**Figura 18:** Estrutura para Construção e Manutenção da Base de Casos.

O módulo de entrada recebe um conjunto de exames médicos, os quais irão compor a base de casos. Cada um desses exames é repassado ao módulo de Extração de Características, onde tem suas informações pseudo-semânticas analisadas e em seguida são submetidos a um processo de categorização de acordo com esses dados. Ainda neste

módulo são extraídas as características visuais da imagem contida no exame. É extraído o mesmo conjunto de atributos para todos os exames. A definição dos algoritmos e características a serem extraídas foge ao escopo da proposta deste trabalho. Estas informações devem fazer parte da Base de Conhecimento em PDI. Com o auxílio de especialistas ou, ainda, de testes com diferentes abordagens a informação dos melhores métodos e atributos deve ser adquirida e estruturada na forma de uma Base de Conhecimento. Com isto se consegue um processo de recuperação automático e adaptável.

Após serem extraídas as características de todos os exames, estes são submetidos aos próximos passos do processo de construção da BC. Onde, para cada categoria identificada na etapa de categorização nível 1 é procedida a definição dos atributos mais significativos para representar essa dada classe. Para determinar o conjunto de atributos que melhor representa cada uma das categorias de imagens existentes na base de casos, uma das diferentes abordagens encontradas na literatura, exemplos de métodos com essas características são encontrados em [RIBEIRO, 2005] e [DY, 1999a].

Em seguida, utilizando-se esse conjunto de características selecionadas, é executado um processo de agrupamento dos casos que compõem a dada categoria, gerando subclasses de similaridade. Os casos são identificados como pertencente a uma única subcategoria. Após, é gerado um caso-médio para cada uma das subclasses, a partir do cálculo da média matemática dos valores dos atributos de todos os casos que a compõe.

As Similaridades entre cada caso e o caso-médio de sua categoria são definidas através de seus valores de Entropia Relativa. O valor da similaridade encontrada para cada caso é armazenado. Todos os casos gerados, cada um identificando um exame DICOM repassado por parâmetro, são então armazenados na BC.

A estrutura definida para a BC é capaz de representar a relação existente entre Casos, Imagens e Laudos, permitindo que a partir da seleção de um determinado caso se consiga acessar a imagem relativa a esta observação, armazenada em uma base de dados para esse fim, e, ainda, recuperar o laudo médico atribuído ao exame ao qual essa imagem pertence.

## **5.9 Considerações**

Para uma efetiva utilização da metodologia apresentada neste capítulo faz-se necessário que sejam tomadas algumas decisões, como: atributos visuais a serem analisados, métodos que realizarão a extração dessas características, abordagem para mineração de características, técnica de agrupamento de casos, entre outras. Porém, tais opções devem levar em consideração o conjunto de imagens com o qual se está trabalhando. Isto por que ainda não se têm definido quais as melhores configurações para cada tipo de exame.

Contudo, independente dos diferentes caminhos a serem seguidos, a metodologia apresentada, se adotada na íntegra, provê uma execução mais rápida e com similar qualidade no resultado da recuperação que demais abordagens anteriores. Isto devido ao seu esquema de classificação em níveis, especialização de características visuais para cada grupo de exames e cálculo prévio da similaridade entre os casos conhecidos. A comprovação destes resultados pode ser observada no próximo capítulo, onde são apresentados alguns dos experimentos realizados durante a elaboração deste trabalho.

## **Capítulo 6**

### **Resultados e Discussões**

Neste capítulo serão apresentados os experimentos realizados com o objetivo de validar a metodologia de recuperação de exames apresentada no capítulo anterior. Para viabilizar a realização dos testes, a proposta apresentada neste trabalho foi implementada. Os métodos para extração de características, bem como os atributos avaliados e os algoritmos de agrupamento de casos e seleção de características, foram selecionados com base em informações levantadas através de pesquisa na literatura da área.

Foram realizados diferentes testes efetuando a tarefa de recuperação de exames, cada um adotando uma característica da metodologia proposta neste trabalho. Partindo de um modelo tradicional de recuperação de imagens por conteúdo, foram sendo adicionados os módulos, ou componentes, a este modelo e verificando o impacto desta inclusão no resultado do processo de recuperação.

A apresentação dos experimentos está dividida em duas partes: a) descrição geral dos experimentos realizados, quais os métodos e algoritmos utilizados e; b) o objetivo de cada um dos casos de testes e a avaliação dos Resultados, onde são discutidos os resultados obtidos e analisados os impactos da inclusão dos diferentes módulos da metodologia.

#### **6.1 Experimentos e Resultados**

Nos testes realizados foi extraído um conjunto de 1223 características de cada um dos exames analisados, sendo: 08 pseudo-semânticas e 1215 visuais. Essas características foram utilizadas para representar cada um dos casos. As informações pseudo-semânticas analisadas foram as mesmas descritas na Tabela 6. Foram considerados diferentes atributos visuais: cor, textura e forma. Para identificação dessas características, os seguintes métodos foram adotados:

- Histograma Normalizado (cor);

- Matriz de Co-ocorrência, distâncias: 1, 2 e 3, com atributos propostos por [HARALICK, 1973] (textura);
- Transformada de Fourier, com verificação de atributos descritos em [GONZALEZ, 1992] (textura);
- 7 Momentos Invariantes (forma - região);
- 50 primeiros Momentos de Zernike (forma - região);
- Descritor de Borda de Sobel;
- 2º e 3º Momentos das Bordas da Imagem (forma - borda);
- Descritores de Fourier, considerando os 1024 primeiros *pixels* de Borda da Imagem (forma - borda).

Para a implementação das abordagens de agrupamento de casos e seleção de características relevantes, foram utilizados os algoritmos de *k-means* e *StARMiner*, respectivamente. O *k-means* foi adotado por ser uma abordagem bastante difundida e fornecer bons resultados no processo de classificação exploratória de dados. O *StARMiner* foi selecionado por ter demonstrado em [RIBEIRO, 2005] resultado satisfatório e eficiência superior a outros algoritmos tradicionais para mineração de regras de associação.

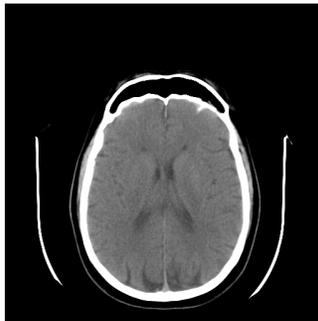
Nos experimentos foram utilizados 23.222 exames médicos divididos em duas modalidades, CT e MR. Essas modalidades continham diferentes regiões do corpo analisadas: HEAD, ABDOMEN, CSPINE, LSPINE, CHEST, PELVIS, ORBITS, EXTREMITY e SKULL. As informações extraídas dos exames foram organizadas em uma estrutura de casos, onde cada caso representa um exame específico. Os casos gerados foram armazenados em um Banco de Casos construído utilizando o PostgreSQL versão 8.1 [POSTGRESQL, 2007]. O equipamento onde foram executados os testes possui como hardware: Processador 1.81Ghz, 1GB de memória RAM e disco rígido SATA<sup>6</sup>.

Para cada um dos casos de testes desenvolvidos são apresentados os respectivos valores de *Precision e Recall* após a recuperação do 100-ésimo exame e o Tempo de Execução Total do processo e, ainda, imagem contendo os 20 primeiros resultados.

---

<sup>6</sup> SATA – *Serial Advanced Technology Attachment* é uma tecnologia para transmissão de dados utilizada em discos rígidos.

Quando o número de resultados é inferior a esse valor, todos os exames retornados são apresentados. A relevância dos exames retornados, observada durante o cálculo dos índices de qualidade do processo de recuperação, é determinada de acordo com a classe, ou subclasse, a qual o exame recuperado pertence. Ou seja, um exame é considerado relevante quando pertencer à mesma classe da imagem de consulta. A exceção desta regra se aplica apenas nos casos onde houve a participação de um especialista médico realizando uma classificação fina das imagens de mesma classe. Nesse caso, consideram-se como relevantes os exames classificados como significantes pelo especialista. A Figura 19 apresenta a imagem de consulta utilizada nos testes. De acordo com as informações pseudo-semânticas extraídas da IC, esta foi classificada como: *CT* → *HEAD* → *MONOCRHOME2*.

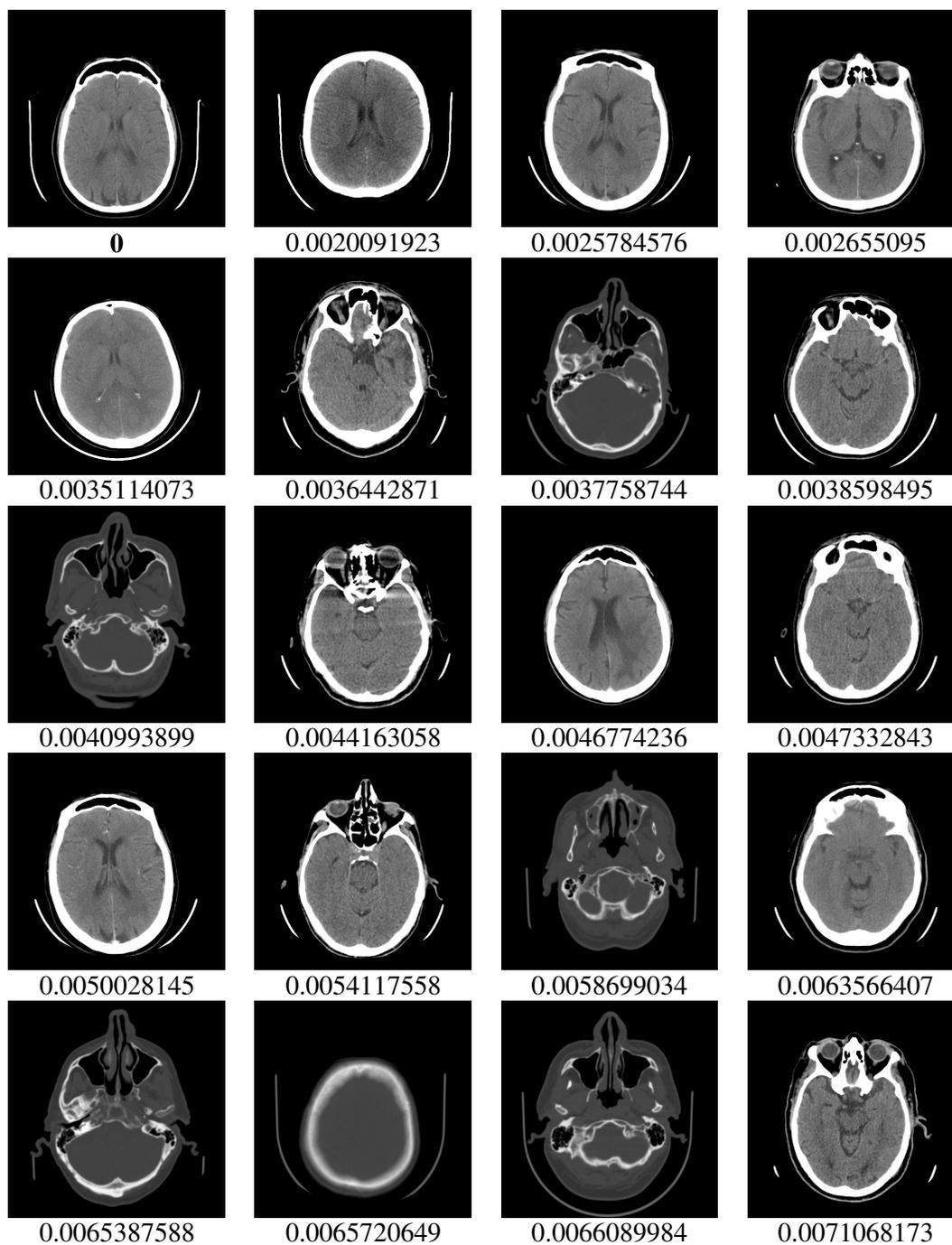


**Figura 19:** Imagem de Consulta utilizada nos Experimentos.

### 6.1.1 Experimento #1

O primeiro teste foi realizado utilizando um modelo convencional para sistemas de recuperação de imagens por conteúdo. Neste tipo de modelo, são extraídas as características visuais da IC (para todos os tipos de imagens de consulta são extraídas as mesmas características). É realizado o acesso a BC para levantar todos os casos armazenados. Após isso, calcula-se o grau de similaridade entre a IC e os casos retornados da BC, utilizando uma métrica de distância qualquer. Os casos são ordenados em ordem crescente de similaridade e retornados como resultado do processo de recuperação.

Utilizando a IC apresentada na Figura 19, foi realizada a tarefa de recuperação seguindo o modelo tradicional descrito. A Figura 20 mostra os primeiros resultados retornados e a Tabela 7 apresenta os valores obtidos para *Precision*, *Recall* e Tempo de Execução.



**Figura 20:** 20 primeiros resultados – Experimento #1.

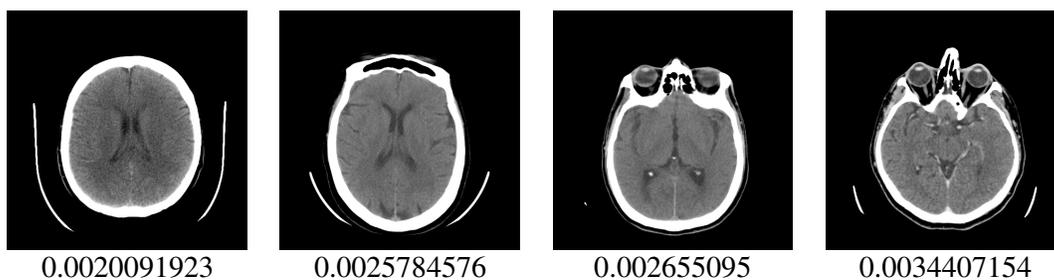
**Tabela 7:** *Precision*, *Recall* e Tempo de Execução – Experimento #1.

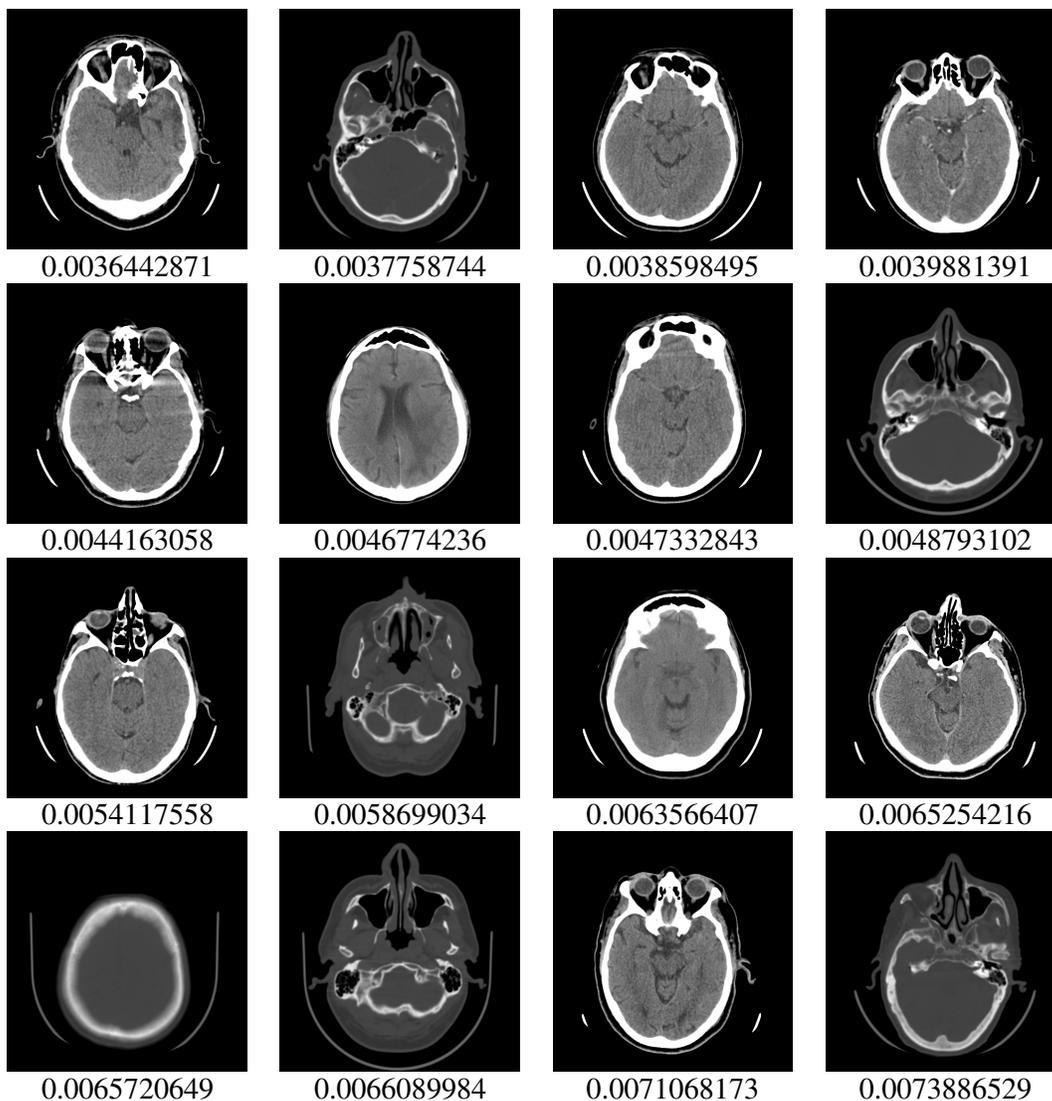
<i>Precision</i>	<i>Recall</i>	Tempo de Execução (s)
100 %	2.75255 %	1.437,431

No modelo utilizado neste experimento todos os casos armazenados na BC são analisados e tem o seu grau de similaridade com a IC calculado. Este fato implica em um grande consumo de processamento e tempo necessário para a execução da tarefa de recuperação. No teste apresentado as 100 primeiras imagens pertencem a mesma classe de exames que a IC. Porém, como essa informação não é considerada durante o processo de recuperação, não se tem a garantia de que essa Precisão será mantida em qualquer outro caso. Cabe observar que a imagem retornada como mais similar é a mesma IC. O baixo índice de *Recall* se deve ao fato de se existir um número maior de exames relevantes na BC do que a quantidade de exames considerada. Isto faz com que a relação entre *Casos Relevantes Recuperados (CRR) / Casos Relevantes Existentes (CRE) na BC* seja baixa. Neste caso,  $CRR = 100$  e  $CRE = 3633$ .

### 6.1.2 Experimento #2

Para a realização do experimento #2 foram adicionados ao modelo tradicional os componentes de extração de características pseudo-semânticas e o filtro de casos a serem selecionados na BC. As informações pseudo-semânticas são obtidas analisando o conteúdo do exame DICOM. São elas que permitem a classificação dos exames de acordo com sua modalidade, região do corpo examinada e interpretação fotométrica. O filtro de casos atua fazendo com que sejam selecionadas na BC apenas os casos pertencentes à mesma classe que a IC. Este filtro também, permite que sejam recuperadas apenas exames que pertençam a Série DICOM diferente a da IC. Este filtro de Série visa aumentar a relevância dos resultados. A Figura 21 mostra os primeiros resultados retornados por este experimento e a Tabela 8 apresenta os valores obtidos para *Precision*, *Recall* e Tempo de Execução.





**Figura 21:** 20 primeiros resultados – Experimento #2.

**Tabela 8:** *Precision*, *Recall* e Tempo de Execução – Experimento #2.

<i>Precision</i>	<i>Recall</i>	Tempo de Execução (s)
100 %	2.75255 %	258,468

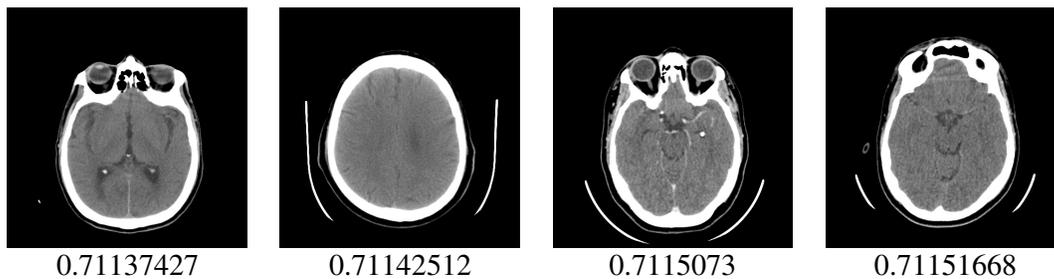
Neste modelo, com a adoção do filtro por categoria, apenas os casos que possuem a mesma classificação pseudo-semântica que a IC são selecionados na BC. Com isto se reduz consideravelmente o custo de tempo e processamento necessários para a realização da recuperação por similaridade. Assim como no resultado obtido neste teste, onde todas as 100 primeiras imagens pertencem a mesma classe de exames que a IC, com a

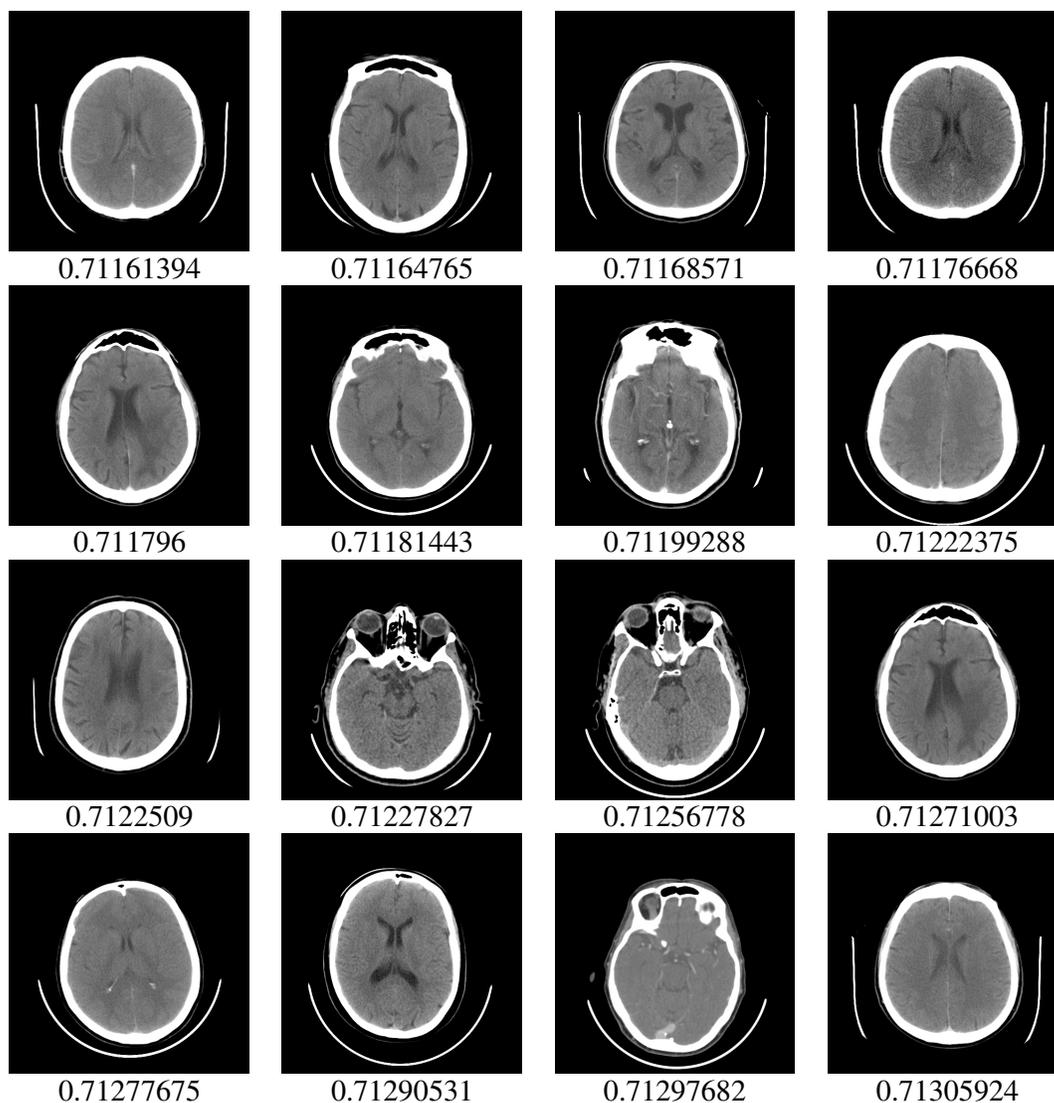
utilização deste modelo pode-se afirmar que todos os casos resultantes do processo pertencerão a mesma classe. Com isso se tem a garantia de que essa Precisão será mantida em qualquer outro caso. Assim como no experimento anterior, o baixo índice de *Recall* se deve ao fato de se existir um número maior de exames relevantes na BC do que a quantidade de exames considerada. Também neste caso,  $CRR = 100$  e  $CRE = 3633$ .

### 6.1.3 Experimento #3

Seguindo com o processo de incremento do modelo de recuperação utilizado, neste experimento as características visuais extraídas das imagens são selecionadas de acordo com o grupo ao qual o exame pertence. Para realizar o processo de mineração de atributos foi utilizado o algoritmo *StARMiner*. Devido à grande diferença entre os valores dos atributos analisados, estes foram agrupados em conjuntos com valores homogêneos antes de serem processados. Para cada um dos grupos de características, foram utilizados três diferentes valores para a diferença mínima entre médias ( $\Delta\mu_{\min}$ ) e outros três para desvio padrão máximo ( $\sigma_{\max}$ ), o grau de confiança ( $\gamma_{\min}$ ) utilizado foi o de 95%. Com isso, foram realizadas, para cada conjunto de atributos, nove execuções do algoritmo.

O resultado final da seleção de características foi dado pelo conjunto de atributos que foram selecionados por mais de duas execuções. Esta abordagem foi adotada pela dificuldade de se identificar os melhores parâmetros para seleção de cada grupo de características. Com isso, pretendeu-se eliminar possíveis discrepâncias do processo de mineração. A característica de cor, Histograma Normalizado, não foi submetida ao processo de mineração. Porém, esta característica também foi utilizada nesta tarefa de recuperação. A Figura 22 mostra os primeiros resultados retornados e a Tabela 9 apresenta os valores para *Precision*, *Recall* e Tempo de Execução.





**Figura 22:** 20 primeiros resultados – Experimento #3.

**Tabela 9:** *Precision*, *Recall* e Tempo de Execução – Experimento #3.

<i>Precision</i>	<i>Recall</i>	Tempo de Execução (s)
100 %	2.75255 %	191,937

Com a utilização de uma abordagem de seleção de características, o processo de recuperação ganha em agilidade. Isto se deve ao fato de se reduzir o número de atributos a serem analisados durante o cálculo de similaridade entre os casos. Dependendo de como esteja a estrutura da BC e de como são selecionados os casos a serem comparados à IC, também pode ser reduzido o tempo para seleção dos casos. Assim como nos testes

anteriores, o valor de  $CRR = 100$  e  $CRE = 3633$ . O tempo de execução total do processo é reduzido em relação aos demais. O grau dessa redução está diretamente relacionado à quantidade de características mineradas. Quanto menor o número de atributos mais ágil se torna o processo. Deve-se considerar, porém, que a redução das características deve ser realizada sem prejudicar a qualidade dos resultados do processo de recuperação. Interessante avaliar diferentes métodos para a seleção de atributos e parâmetros utilizados, para verificar em qual par (método, parâmetro) se consegue os melhores custo/benefício de acordo com o conjunto de elementos na BC.

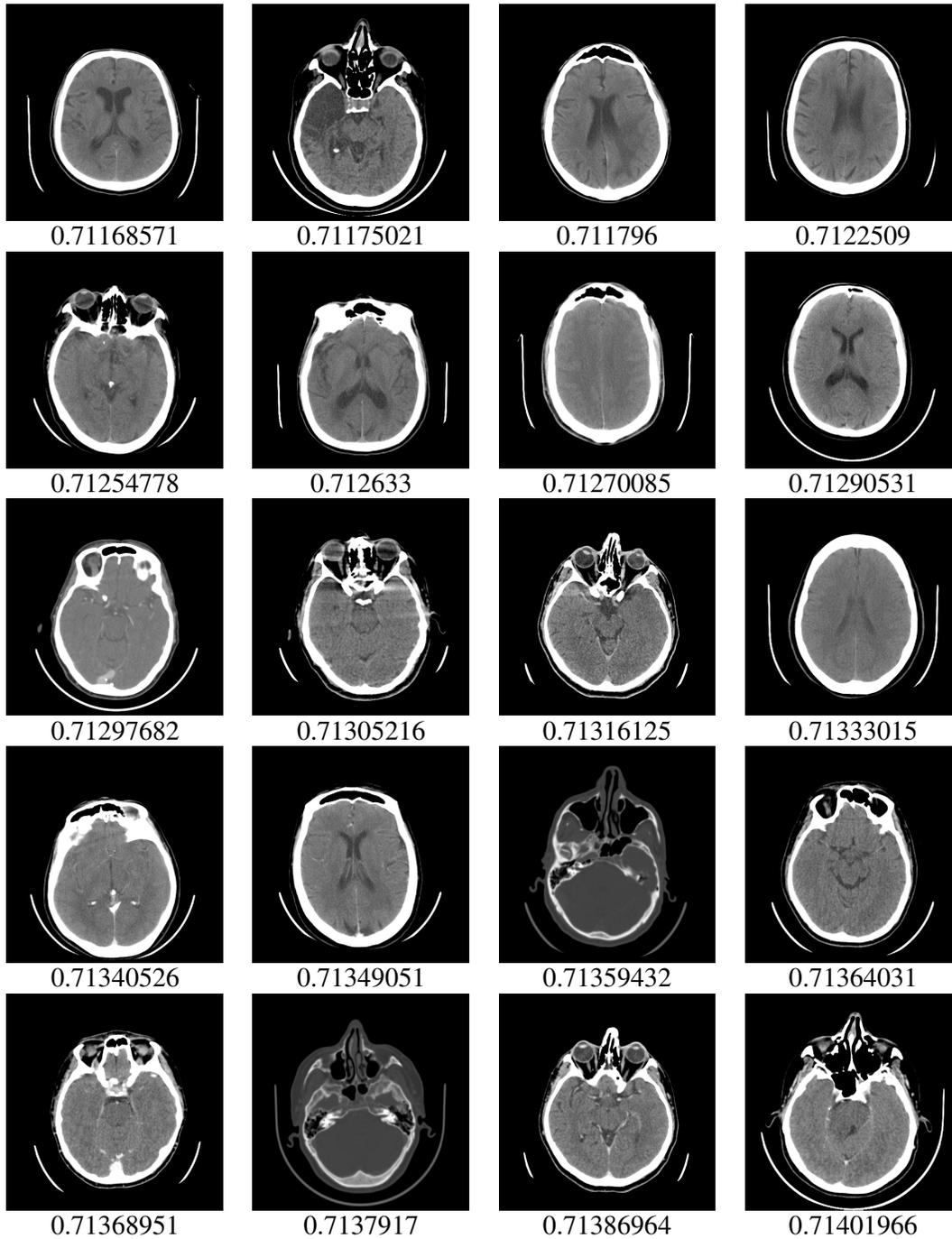
#### 6.1.4 Experimento #4

Com base nas classes de imagens geradas com a análise das informações pseudo-semânticas e utilizando as características mineradas para cada um desses grupos de exames, no experimento #4 é definido um segundo nível de classificação de exames. Essa nova classificação é baseada na similaridade do conteúdo visual. Ou seja, procurou-se agrupar as imagens em subclasses de similaridade visual com o objetivo de reduzir o espaço de busca por casos similares, ampliando a relevância dos resultados e diminuindo o tempo necessário para execução da tarefa de recuperação de exames.

Neste teste, após serem extraídas as características visuais da IC esta é subclassificada. Para isto é calculada a distância entre a IC com o centro de cada uma das subclasses geradas para a classe de exames a qual ela pertence. É utilizado o  $k$ -NN, com  $k=1$ , para determinar qual a subclasse mais próxima. Além do novo nível de classificação, a outra modificação no modelo para este experimento é que o Filtro de Casos, que até então era realizado apenas com as informações pseudo-semânticas, agora considera também a subclasse da IC.

Os agrupamentos de casos foram gerados em tempo de construção da BC. Para tal, foi utilizado o algoritmo  $k$ -Means. Este é um algoritmo tradicional para este tipo de tarefa. Foram gerados agrupamentos com 2, 3, 4 e 5 subclasses para cada uma das classes de exames identificadas. Foi observado que não existe um parâmetro ótimo para o número de subclasses geradas que possa ser utilizado por todos os tipos de exames. Os melhores resultados observados foram onde se conseguiu uma distribuição homogênea dos casos entre as diferentes subclasses. Neste experimento não se utilizou nenhuma medida estatística para determinar a qualidade dos agrupamentos, porém, tal abordagem

pode ser adotada para esta verificação. A Figura 23 mostra os primeiros resultados retornados para a situação onde foram geradas 4 classes. Esta apresentou melhores resultados que os demais agrupamentos para este tipo de exame, considerando os atributos analisados. Na Tabela 10 são apresentados os valores de *Precision*, *Recall* e Tempo de Execução.



**Figura 23:** 20 primeiros resultados – Experimento #4.

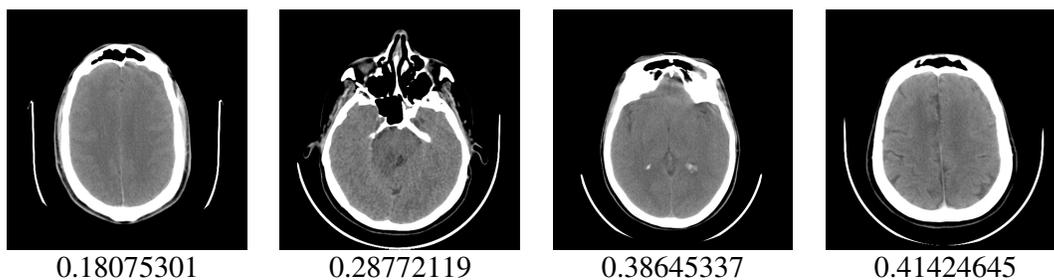
**Tabela 10:** *Precision*, *Recall* e Tempo de Execução – Experimento #4.

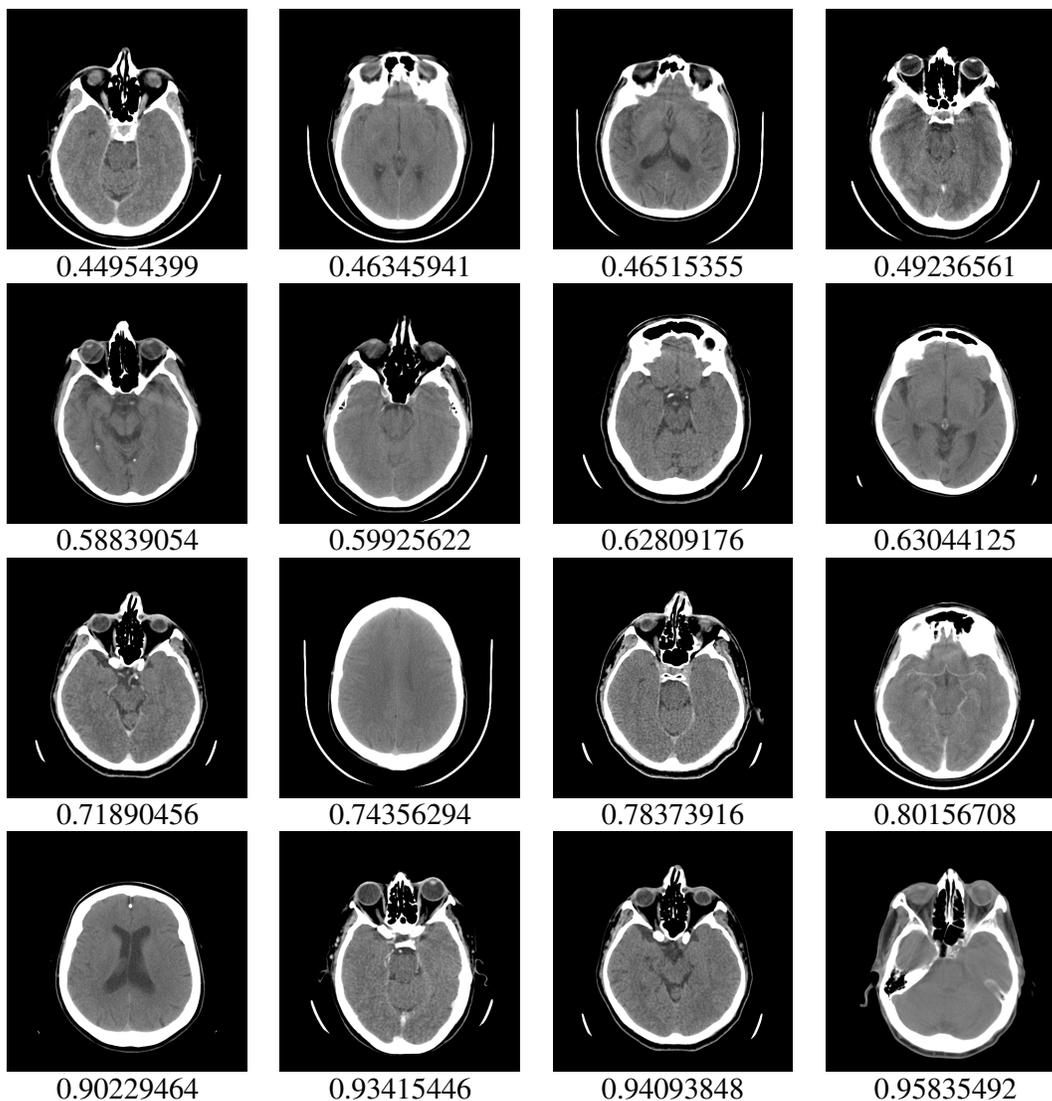
<i>Precision</i>	<i>Recall</i>	Tempo de Execução (s)
100 %	9.46073 %	99,562

Os resultados obtidos demonstram uma redução considerável no tempo de execução. Isto se deve à menor quantidade de casos comparados com a IC. Com o agrupamento em subclasses de similaridade de acordo com os valores de atributos visuais, reduz-se o espaço de busca na BC. O filtro, na fase de seleção de casos, retorna apenas casos pertencentes a mesma subclasse que a IC. O valor de *Recall* aumenta pois a quantidade de casos considerados no processo de recuperação diminui. Neste caso, CRR = 100 e CRE = 1057.

### 6.1.5 Experimento #5

Este experimento apresenta a implementação na íntegra da metodologia proposta neste trabalho. Em tempo de construção da BC é calculado um valor médio para cada um dos atributos de cada uma das subclasses encontradas durante os processos de classificação de exames. Esses conjuntos de atributos médios servem para identificar o ponto central de cada uma das subclasses. Todos os casos têm sua distância para o centro de sua subclasse calculada e armazenada na BC. Durante o processo de recuperação no instante de se determinar o grau de similaridade entre os casos, é calculada a distância entre a IC e o centro de sua subclasse, em seguida é calculada a distância entre esse valor e o valor obtido para cada um dos casos retornados da BC. Este é adotado como o grau de similaridade entre a IC e os demais elementos de sua subclasse. A Figura 24 mostra os 20 primeiros resultados. A Tabela 11 mostra os valores de *Precision*, *Recall* e Tempo de Execução.





**Figura 24:** 20 primeiros resultados – Experimento #5.

**Tabela 11:** *Precision*, *Recall* e Tempo de Execução – Experimento #5.

<i>Precision</i>	<i>Recall</i>	Tempo de Execução (s)
100 %	9.46073 %	64,732

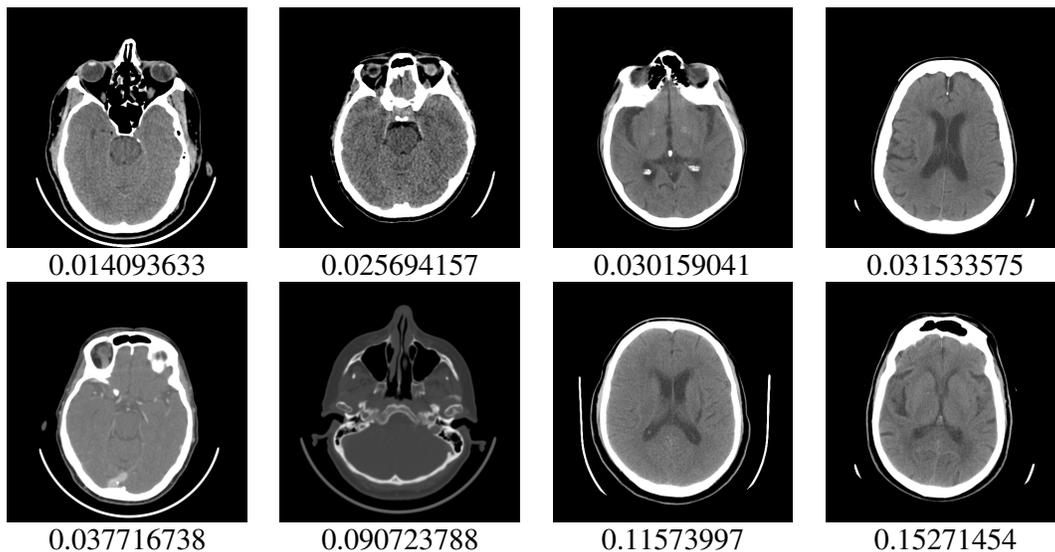
A existência dos valores de similaridade pré-calculados entre os casos contidos na BC, sem dúvida reduz a exigência de processamento para o atendimento de uma solicitação de recuperação de exames por conteúdo. De igual modo, o tempo gasto com a execução dessa tarefa é reduzido consideravelmente. A redução do tempo está diretamente relacionada à quantidade de casos pertencentes a mesma classe da IC,

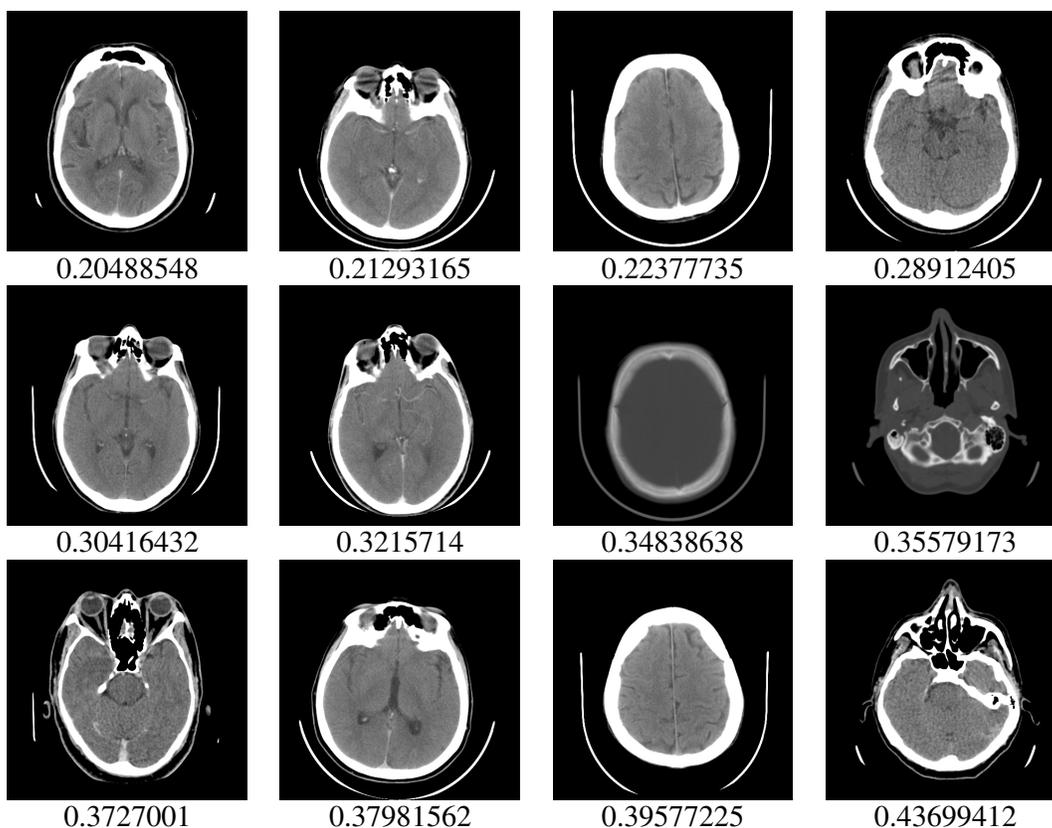
quanto maior o número de casos mais visível será a diferença de tempos entre o modelo deste experimento em relação aos dos demais. A precisão dos resultados continua assegurada devido ao filtro de casos no instante da seleção na BC. A relevância dos resultados, de acordo com a similaridade visual, é dependente dos métodos atributos e analisados e não é influenciada pela abordagem de cálculo de similaridade adotada neste modelo.

### 6.1.6 Experimento #6

Neste caso de teste foi utilizado um conjunto de 57 casos selecionados aleatoriamente da BC. Esta quantidade representa aproximadamente 30% do total de casos possíveis de serem retornados pelo processo de recuperação, visto que encontramos 171 Séries de exames da mesma classe que a IC. Esses casos foram ordenados, previamente, de acordo com sua similaridade com a IC por um especialista. Esta classificação foi adotada como parâmetro para determinar a relevância dos exames recuperados.

Foi executada a tarefa de recuperação utilizando o mesmo modelo do Experimento #5. O objetivo deste experimento foi o de avaliar o grau de relevância “fina” dos resultados. Esta relevância é dependente do conjunto de atributos extraídos do exame e dos métodos utilizados para este fim. Isso porque são eles que determinam quais os aspectos que devem ser julgados para determinar a similaridade entre os casos. Na Figura 25 os primeiros resultados são apresentados. A Tabela 12 mostra os valores de *Precision*, *Recall* e Tempo de Execução.





**Figura 23:** 20 primeiros resultados – Experimento #6.

**Tabela 15:** *Precision*, *Recall* e Tempo de Execução – Experimento #6.

<i>Precision</i>	<i>Recall</i>	Tempo de Execução (s)
15%	75%	32,047

Foram utilizados 32,047s para execução da tarefa de recuperação. Destes, 31,579s foram gastos com a extração de características e categorização da imagem. Os valores de *Precision* e *Recall* foram obtidos considerando os 20 primeiros resultados. Do conjunto de amostras repassadas ao especialista, foram selecionadas apenas 4 imagens como sendo similares a IC. Com esse parâmetro o valor máximo para precisão seria de 20%. Deste total o experimento alcançou 15% por ter conseguido recuperar apenas 3 das imagens selecionadas. Por outro lado, o valor de *Recall* é alto por terem sido recuperadas  $\frac{3}{4}$  da imagens possível.

O baixo valor de precisão se justifica no fato de o especialista ter considerado como parâmetro para definição de similaridade entre imagens o plano de corte no qual

cada imagem pertence. Este atributo não é levado em conta por nenhum dos métodos de extração de características ou medidas de similaridade adotadas durante a realização dos testes. Contudo, em conversas com o especialista se pode concluir que tal parâmetro deve ser considerado caso se pretenda alcançar resultados realmente relevantes na prática. A dificuldade em verificar o plano de corte de uma imagem é a de que este atributo é variável de acordo com o grupo de imagens em que se está trabalhando. Ou seja, não existe um parâmetro ótimo para determinar dissimilaridade entre planos de corte que possa ser levado em consideração para CT\_HEAD e CT ABDOMEN, por exemplo.

## **6.2 Considerações**

Durante a elaboração deste trabalho, diferentes experimentos com outras classes de imagens foram realizados. Em cada um deles foram utilizadas mais de uma IC. Todos apresentaram resultados similares aos descritos neste capítulo. A adoção do modelo proposto no capítulo 5 permite uma ágil execução do processo de recuperação de exames por conteúdo visual sem embutir perda de relevância nos resultados.

Com a construção de uma Base de Conhecimento de PDI capaz de identificar os melhores métodos, parâmetros e atributos a serem adotados na recuperação de cada um dos tipos de exames deverá contribuir positivamente para a obtenção de resultados ainda mais relevantes. Esta afirmação pode ser feita com base na observação dos resultados obtidos com a especialização das características analisadas, através do método de mineração de atributos e com os resultados obtidos no experimento #6. Os resultados deste último teste validam a idéia de se definir estratégias específicas para extração de características e cálculo de similaridade em cada uma das categorias de exames existentes.

## **Capítulo 7**

### **Conclusões**

A falta de ferramentas ágeis para a execução da tarefa de recuperação de exames médicos digitais é um dos grandes motivos para a não implantação efetiva deste tipo de benefício no dia-a-dia de profissionais que lidam com esse tipo de informação. O crescimento da utilização da tecnologia digital na realização de exames médicos por imagem tem gerado um grande volume de dados que por falta de métodos para acesso eficiente, deixam de ser utilizados no auxílio ao diagnóstico. Verificando experiências anteriores, médicos teriam mais um parâmetro para considerar no instante da elaboração de seus laudos.

A metodologia desenvolvida neste trabalho veio de encontro a esta necessidade. Os resultados apresentados comprovam o ganho de desempenho na execução de uma solicitação de recuperação de exames por similaridade do conteúdo visual. É importante observar que a relevância dos resultados obtidos se mantém próxima às obtidas com métodos mais tradicionais. A qualidade do processo de recuperação está diretamente associada aos métodos e atributos visuais considerados no cálculo da similaridade. A adoção de abordagens específicas para cada tipo de exame acarretará um ganho na relevância dos resultados. Isso devido proporcionarem uma representação mais aproximada de cada categoria.

#### **7.1 Objetivos Alcançados**

Neste trabalho definiu-se uma arquitetura capaz de realizar a recuperação de exames médicos por similaridade de conteúdo visual de forma ágil e eficiente. A estrutura apresentada é dividida em módulos, cada um sendo encarregado de uma função específica durante o processo. Esse tipo de divisão permite um maior controle do processo.

Foram definidas informações que podem ser extraídas diretamente do conteúdo de um exame DICOM e que auxiliam consideravelmente na execução da tarefa de recuperação. Estas informações atuam como uma descrição pseudo-semântica do exame.

Um filtro de casos armazenados na base de casos foi definido com o objetivo de reduzir o espaço de busca. Este filtro se baseia no conteúdo pseudo-semântico e classificação visual do exame.

Durante os testes foi utilizado o algoritmo *StARMiner* para realizar a seleção de características em cada uma das classes de exames. Contudo, este método pode ser substituído por outro semelhante sem acarretar em prejuízo para a metodologia. Com a experiência da realização deste trabalho concluiu-se que a utilização direta de um método para mineração de atributos não surtirá os efeitos desejados se o conjunto de atributos no qual será executado não forem suficientemente relevantes para representar as características visuais da imagem. Sendo necessário, portanto, a definição de um conjunto representativo de atributos para cada uma das categorias de exames.

A alimentação de dados na base de casos deve ser um processo contínuo. Na medida em que forem gerados novos exames estes devem passar a constarem como experiências, sendo armazenados na base de casos. O processo de extração de características de exames exige bastante processamento em uma estação de trabalho<sup>7</sup>. A tarefa de manutenção da base de casos tipicamente se apresenta como um processo de alta-disponibilidade e deve ser executada à parte das solicitações de recuperação de exames. Por isso, neste trabalho, definiu-se um esquema para realização desta tarefa.

Foi definida uma abordagem a priori para cálculo de similaridade entre exames. Com este modelo conseguiu-se reduzir em 95,5%<sup>8</sup> do tempo necessário para execução da recuperação de casos similares. Além deste ganho, deve-se destacar a redução das necessidades de processamento para a realização desta tarefa. Outro ponto importante é a obtenção de relevância dos resultados similar a observada nos demais modelos.

## **7.2 Trabalhos Futuros**

Com a elaboração deste trabalho observou-se a necessidade de aprofundamento de pesquisa em alguns temas relacionados. O desenvolvimento destes novos projetos certamente contribuirá positivamente para o processo de recuperação como um todo.

---

<sup>7</sup> Conclusão baseada no equipamento utilizado durante a realização dos testes.

<sup>8</sup> Com base nos testes realizados. Comparando a abordagem proposta com o tempo do modelo tradicional.

- Identificação das características que melhor descrevem cada uma das categorias de exames definidas a partir da análise dos possíveis valores para as informações pseudo-semânticas. As classes podem ser identificadas com a observação de dados fornecidos em [DICOM, 2006b];
- Identificação de métodos para extração de características e melhores configurações para obtenção dos atributos visuais do exame. Com base nos resultados de um projeto com este foco, pode-se construir uma Base de Conhecimento em PDI;
- Definir estratégias para identificação de planos de corte de imagem relevantes para serem comparados entre si. Para isso deve-se considerar cada uma das particularidades dos tipos de exames possíveis;
- Para a verificação da qualidade da recuperação é necessário que alguém com experiência na área de domínio da imagem, faça uma classificação manual que sirva de base para certificação do resultado do processo informatizado. Para tal, faz necessária a construção de ferramenta para facilitar o processo de validação de resultados por parte de especialistas;
- Realização de testes comparativos da abordagem proposta para o cálculo de similaridade utilizando outros modelos computacionais que não os adotados neste trabalho. Isto pode levar a um maior embasamento de que a utilização de Entropia Relativa é a melhor solução para este problema. Ou, ainda, que existem outras abordagens que podem tratar de forma mais eficaz esta necessidade;
- Construção de uma interface amigável para a realização de solicitação de recuperação de exames. Devem ser levadas em consideração as necessidades do usuário final. Um levantamento de requisitos e validação com um grupo de especialistas pode ser necessário;
- Definição de técnicas para adaptação e reaproveitamento de diagnósticos médicos. Isto contribuirá para a demonstração da utilidade do uso de ferramentas de recuperação de exames. Além de facilitar o trabalho de especialistas quando da necessidade de reuso de diagnósticos.

## Referências

- [AAMODT, 1994] AAMODT, Agnar. PLAZA, Enric. *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*. Artificial Intelligence Communications 7, no. 1, 39-52, 1994.
- [AISEN, 2003] AISEN, A. M., BRODERICK, L. S., WINER-MURAM, H., BRODLEY, C.E., KAK A.C., PAVLOPOULOU, C., DY, J., SHYU, C.R. MARCHIORI, A.. *Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment*, Radiology, Vol. 228, pp. 265-270, 2003.
- [ALEXANDRINI, 2005] ALEXANDRINI, Fábio, BORTOLUZZI, Mariana Kessler, WANGENHEIM, Aldo v. *Improving Content Based Recovery on a Radiological Reports Database*. Wissensmanagement: 309-315, 2005.
- [ANTANI, 2004] ANTANI, S., LONG, L. R., THOMA, G. R.. *Content-Based Image Retrieval for Large Biomedical Image Archives*. MEDINFO, 2004.
- [ARAUJO, 2002a] ARAUJO, Myrian Renata Barros; TRAINA JR., Caetano; TRAINA, Agma Juci Machado; MARQUES, Paulo M. d. A.. *"Recuperação de Exames em Sistemas de Informação Hospitalar com Suporte a Busca de Imagens Baseada em Conteúdo"*. VIII Congresso Brasileiro de Informática em Saúde- CBIS'2002, realizado Natal - RN, de 29 de setembro a 02 de outubro de 2002.
- [ARAUJO, 2002b] ARAUJO, Myrian Renata Barros; TRAINA JR., Caetano; TRAINA, Agma Juci Machado; BUENO, Josiane M.; RAZENTE, Humberto. *"Extending Relational Databases to Support Content-based Retrieval of Medical Images"*. in IEEE International Conference on Computer Based Medical Systems - CBMS, Maribor, Slovenia, 2002, pp. 303-308 (Best Paper Award).
- [BALAN, 2004a] BALAN, Andre G. R.; BORTOLOTTI, Luis M.; TRAINA Jr., Caetano. *"Content-based Image Retrieval Using Approximate Shape of Objects"*. in 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04), Bethesda, Maryland, 2004, pp. 91-96.
- [BALAN, 2004b] BALAN, Andre G. R.; TRAINA, Agma Juci Machado; TRAINA JR., Caetano; MARQUES, Paulo M. d. A.. *"Integrando Textura e Forma para a Recuperação de Imagens por Conteúdo"*. IX Congresso Brasileiro de Informática em Saúde - CBIS'2004, Ribeirão Preto, SP, Brazil, 2004, pp. 6 p.
- [BARIONI, 2006] BARIONI, Maria Camila N.; TRAINA, Agma Juci Machado; TRAINA JR., Caetano. *"Recuperação de Imagens Médicas por Conteúdo em um Sistema de Gerenciamento de Banco de Dados de Código Livre"*. X Congresso Brasileiro de Informática em Saúde (CBIS), Florianópolis (SC), p. 1561-1566, 2006.

[BORTOLOTTI, 2004] BORTOLOTTI, Luis M.; TRAINA, Agma Juci Machado. "*Recuperação de Imagens por Conteúdo Através de Regiões Determinadas Automaticamente pela Energia*". IX Congresso Brasileiro de Informática em Saúde - CBIS'2004, realizado Ribeirão Preto, SP, de 07-10 de novembro de 2004.

[BORTOLUZZI, 2003] BORTOLUZZI, Mariana Kessler. Desenvolvimento e Implementação de um editor de documentos estruturados no padrão DICOM Structured Report. Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2003.

[BUENO, 2002] BUENO, Josiane M. *Suporte a Recuperação de Imagens Médicas Baseada em Conteúdo através de Histogramas Métricos*. Tese Doutorado. Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional do ICMC/USP. São Carlos-SP, 2002.

[BRUUN, 1978] BRUUN, Georg. *z-Transform DFT filters and FFTs*. IEEE Trans. on Acoustics, Speech and Signal Processing (ASSP) 26 (1), 56-63, 1978.

[CAMPO, 2003] CAMPO, C. Y. ; TRAINA, Agma Juci Machado . *Uma Abordagem Eficiente para Recuperação de Imagens Médicas*. In: WIM - Workshop de Informática Médica, 2003, Fortaleza - CE. Uma Abordagem Eficiente para Recuperação de Imagens Médicas, 2003.

[CANNY, 1986] CANNY, J.. *A Computational Approach To Edge Detection*. IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986.

[CASIMAGE, 2007] Casim@ge Project. Disponível em: [http://www.sim.hcuge.ch/casimage/01\\_CASIMAGE\\_Presentation\\_EN.htm](http://www.sim.hcuge.ch/casimage/01_CASIMAGE_Presentation_EN.htm). Acessado em, Fevereiro de 2007.

[CASTAÑÓN, 2002] CASTAÑÓN, Cesar A. B.; TRAINA, Agma Juci Machado. "*Extração de Características de Imagens Médicas Baseadas na Distribuição de Cor dos Espaços de Wavelets*". Anais do 2º Workshop de Informática Médica – WIM'2002, junto ao Simpósio Brasileiro de Engenharia de Software (SBES) da Sociedade Brasileira de Computação. Gramado, RS, 17 de outubro de 2002.

[CASTAÑÓN, 2003a] CASTAÑÓN, Cesar A. B.; TRAINA, Agma Juci Machado. "*Caracterização de Imagens através de Análise Multiresolução por Wavelets para Sistemas CBIR*". XXIX Conferência LatinoAmericana de Informática - CLEI'2003 29 de setembro a 2 de outubro de 2003, pp. 12 pags.

[CASTAÑÓN, 2003b] CASTAÑÓN, César A. *Recuperação de imagens por conteúdo através de análise multi-resolução por Wavelets*. Dissertação de Mestrado. Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional do ICMC/USP. São Carlos-SP, 2003.

[CHINO, 2002] CHINO, Fabio J. T.; TRAINA, Agma Juci Machado; TRAINA JR., Caetano; MARQUES, Paulo M. d. A.. "*How to Add Content-based Image Retrieval Capability in a PACS*". IEEE International Conference on Computer Based Medical Systems - CBMS, Maribor, Slovenia, 2002, pp. 321-326.

- [CLOUGH, 2004] CLOUGH, Paul, SANDERSON, Mark, MÜLLER, Henning, *A proposal for the CLEF Cross Language Image Retrieval Task 2004*, Proceedings of the 2004 CIVR conference, Springer Lecture Notes in Computer Science, Dublin, Ireland, 2004.
- [CLOUGH, 2005] CLOUGH, Paul, MÜLLER, Henning, SANDERSON, Mark, *The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004*, CLEF Proceedings - Springer Lecture Notes in Computer Science, LNCS 3491, pages 597-613, 2005.
- [COOLEY, 1965] COOLEY, James W., TUKEY, John W.. *An algorithm for the machine calculation of complex Fourier series. Math. Comput.* 19, 297–301, 1965.
- [CLUNIE, 2000] CLUNIE, David. *DICOM Structured Reporting*. PixelMed Publishing, 2000.
- [DAHMEN, 2000] DAHMEN, J., THEINER, T., KEYSERS, D., NEY, H., LEHMANN, T.M., WEIN, B.B.. *Classification of radiographs in the image retrieval in medical application system IRMA*. Proceedings 6th International RIAO Conference on Content-Based Multimedia Information Access, Paris, France, 2000; 551-566.
- [DASARATHY, 1991] DASARATHY, B.V., *Nearest-Neighbor Classification Techniques*. IEEE Computer Society Press, Los Alamos, CA, 1991.
- [DEMPSTER, 1977] DEMPSTER, Arthur, LAIRD, Nan, RUBIN, Donald. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977.
- [DESELAERS, 2004] DESELAERS, T., KEYSERS, D., NEY, H.. *Fire-flexible image retrieval engine: ImageCLEF 2004 evaluation*. Proceedings of Cross Language Evaluation Forum (CLEF) Workshop 2004, Bath, UK, 2004.
- [DICOM, 2006a] DICOM – Parte 1.
- [DICOM, 2006b] DICOM – Parte 3.
- [DICOM, 2006c] DICOM – Parte 6.
- [DY, 1999a] DY, J. G., BRODLEY, C. E., KAK, A., SHYU, C. R., BRODERICK, L. *The customized-queries approach to CBIR*. SPIE Storage and Retrieval for Image and Video Databases VII, 3656:22–32, Janeiro, 1999.
- [DY, 1999b] DY, J. G., BRODLEY, C. E., KAK, A., SHYU, C. R., BRODERICK, L. *The customized-queries approach to CBIR using EM*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 400-406, Fort Collins, CO, Junho, 1999.
- [DY, 2003] DY, J. G., BRODLEY, C. E., KAK, A., BRODERICK, L. S., AISEN, A. M. *Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images*. IEEE Trans. Pattern Anal. Mach. Intell. 25, 3, 373-378., Março, 2003.

- [FLOREA, 2006] FLOREA, F., MULLER, H., ROGOZAN, A., GEISSBUHLER, A., DARMONI, S.. *Medical Image Categorization with MedIC and MedGIFT*. 20th International Congress of the European Federation for Medical Informatics (MIE 2006). Maastricht, Netherlands. 2006.
- [GBDI/USP, 2007] Grupo de Bases de Dados e Imagens da Universidade de São Paulo. Disponível em: <http://gbdi.icmc.usp.br/> . Acessado em: Fevereiro, 2007.
- [GONZALEZ, 1992] GONZALEZ, R., WOODS, Richard. *Digital Image Processing*. Addison Wesley, 1992.
- [GÜLD, 2002] GÜLD, M.O., KOHNEN, M., KEYSERS, D., SCHUBERT, H., WEIN, BB., BREDNO, J., LEHMANN, T.M. *Quality of DICOM header information for image categorization*. Proceedings SPIE; 4685: 280-287, 2002.
- [GÜLD, 2003] GÜLD, M.O, THIES, C., FISCHER, B., KEYSERS, D., WEIN, B.B., LEHMANN, T.M.. *A platform for distributed image processing and image retrieval*. Proceedings SPIE 2003; 5150:1109-1120.
- [GÜLD, 2004] GÜLD, M.O., KEYSERS, D., DESELAERS, T., LEISTEN, M., SCHUBERT, H., NEY, H., LEHMANN, T.M.. *Comparison of global features for categorization of medical images*. Proceedings SPIE 2004; 5371: 211-222.
- [GÜLD, 2005] GÜLD, M.O., THIES, C., FISCHER, B., LEHMANN, T.M.. *A generic concept for the implementation of medical image retrieval systems*. Connecting medical informatics and bio-informatics. Proceedings of MIE 2005. IOS Press, Amsterdam, 2005; 459-464.
- [HARALICK, 1973] HARALICK, R. M. et al. *Textural Features for Image Classification*. IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, n. 6, Novembro, 1973.
- [HU, 1962] HU, M.-K. *Visual Pattern Recognition by Moments Invariants*. IRE Transactions on Inform. Theory, Vol. IT-8, pp. 179-187, Feb. 1962.
- [IRMA, 2007] *Image Retrieval in Medical Applications*. Disponível em: <http://phobos.imib.rwth-aachen.de/irma/index.php> . Acessado em: Fevereiro, 2007.
- [KAK, 2000] KAK, A. C., PAVLOPOULOU, C.. *Computer vision techniques for content-based image retrieval from large medical databases*. 7th Workshop on Machine Vision Applications, 2000.
- [KEYSERS, 2004] KEYSERS, D., GOLLAN, C., NEY, H.. *Classification of medical images using non-linear distortion models*. Bildverarbeitung für die Medizin 2004, Springer-Verlag, Berlin, 2004; 366-370.
- [KHOTANZAD, 1990 ] KHOTANZAD, A., HONG, Y. H. *Invariant Image Recognition by Zernike Moments*. IEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, N. 5, pp. 489-497, 1990.

- [KOHONEN, 1987] KOHONEN, Teuvo. *An Introduction to Neural Computing*. Finland: Helsinki University of Technology, 1987.
- [LEE, 1996] LEE, T. S. *Image Representation Using 2D Gabor Wavelets*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 10, 1996.
- [LEHMANN, 2001] LEHMANN, T.M., WEIN, B., KEYSERS, D., BREDNO, J., GÜLD, M.O., SCHUBERT, H., KOHNEN, M., *Image retrieval in medical applications: The IRMA-Approach*. Proc. VISIM Workshop on In-formation Retrieval and Exploration from Large Medical Image Collections (MICCAI 2001 Satellite), Utrecht, The Netherlands, 2001.
- [LEHMANN, 2003] LEHMANN, T.M., WEIN, B.B., GREENSPAN, H. *Integration of content-based Image retrieval to picture archiving and communication systems*. Proceedings Medical Informatics Europe (MIE 2003), IOS Press, Amsterdam, ISBN 1 58603 347 6 (CD-ROM only).
- [LEHMANN, 2004a] LEHMANN, T.M., GÜLD, M.O., THIES, C., FISCHER, B., SPITZER, K., KEYSERS, D., NEY, H., KOHNEN, M., SCHUBERT, H., WEIN, B.B.. *Content-based image retrieval in medical applications*. Methods of Information in Medicine; 43(4): 354-361, 2004.
- [LEHMANN, 2004b] LEHMANN, T.M., GÜLD, M.O., KEYSERS, D., DESELAERS, T., SCHUBERT, H., WEIN, B.B., SPITZER, K.. *Similarity of medical images computed from global feature vectors for content-based retrieval*. Lecture Notes in Artificial Intelligence 2004; 3214: 989-995.
- [LEHMANN, 2005a] LEHMANN, T.M., GÜLD, M.O., DESELAERS, T., KEYSERS, D., SCHUBERT, H., SPITZER, K., NEY, H., WEIN, B.B.. *Automatic categorization of medical images for content-based retrieval and data mining*. Computerized Medical Imaging and Graphics 2005; 29(2): 143-155.
- [MACQUEEN, 1965] MACQUEEN, J. B.. *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1965.
- [MARQUES, 2002] MARQUES, Paulo M. d. A.; HONDA, Marcelo Ossamu; RODRIGUES, José Antonio H.; SANTOS, Rildo Ribeiro dos; TRAINA, Agma Juci Machado; TRAINA JR., Caetano. "Recuperação de Imagens Baseada em Conteúdo: Uso de Atributos de Textura para Caracterização de Microcalcificações Mamográficas". Revista Brasileira de Radiologia, vol. 35, pp. 93-98, 2002.
- [MOSHEFEGHI, 2004] MOSHEFEGHI, Mehran, SAIZ, Craig, YU, Hu: *Content-based Retrieval of Medical Images with Relative Entropy*. In: Medical Imaging 2004: PACS and Imaging Informatics. Proceedings of the SPIE, Volume 5371, pp. 259-267, 2004.
- [MÜLLER, 2001] MÜLLER, Henning, MULLER, Wolfgang, SQUIRE, David, MARCHAND-MAILLET, Stephane, PUN, THIERRY. *Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals*. Patter Recognition Letters, Vol. 22, n. 5, pp. 593-601, 2001.

- [MÜLLER, 2004a] MÜLLER, Henning, MICHOUX, Nicolas, BANDON, David, GEISSBUHLER , Antoine. *A review of content-based image retrieval systems in medicine - clinical benefits and future directions*. International Journal of Medical Informatics, 73:1--23, 2004.
- [MÜLLER, 2004b] MÜLLER, Henning, ROSSET, Antoine, VALLÉE, Jean-Paul, GEISSBUHLER, Antoine, *Comparing feature sets for content-based medical information retrieval*, SPIE Medical Imaging , San Diego, CA, USA, February 2004.
- [MÜLLER, 2004d] MÜLLER, Henning, ROSSET, Antoine, VALLÉE, Jean-Paul, TERRIER, François, GEISSBUHLER, Antoine, *A reference data set for the evaluation of medical image retrieval systems*, Journal on Computerized Medical Imaging and Graphics, volume 28, pages 65-77, 2004.
- [MÜLLER, 2004e] MÜLLER, Henning, RUCH, Patrick, SQUIRE, David McG., LOVIS, Christian, GEISSBUHLER, Antoine, *Using open source software for medical image retrieval*, Medinfo 2004, San Francisco, CA, USA, september 2004.
- [MÜLLER, 2005a] MÜLLER, Henning, RUCH, Patrick, GEISSBUHLER, Antoine, *Enriching content-based image retrieval with multi-lingual search terms*, Swiss Medical Informatics, volume 54, pages 6-11, 2005.
- [MÜLLER, 2005b] MÜLLER, Henning, GEISSBUHLER, Antoine, RUCH, Patrick, *Report on the CLEF Experiment: Combining image and multi-lingual search for medical image retrieval*, CLEF Proceedings - Springer Lecture Notes in Computer Science, LNCS 3491, pages 718-727, 2005.
- [MÜLLER, 2005c] MÜLLER, Henning, GEISSBUHLER, Antoine, *Medical Image Retrieval - the medGIFT project*, Medical Imaging and Telemedicine (MIT 2005), WuyiShan, China, August 2005.
- [MÜLLER, 2006a] MÜLLER, H., DESPONT-GROSA, C., HERSH, W., JENSEN, J., LOVIS, C., GEISSBUHLER, A., *Health care professionals' image use and search behaviour*, Proceedings of Medical Informatics Europe 2006.
- [MÜLLER, 2006b] MÜLLER, Henning, GEISSBUHLER, Antoine, MARTY, Johan, LOVIS, Christian, RUCH, Patrick, *The use of medGIFT and easyIR for ImageCLEF 2005*, Springer Lecture Notes in Computer Science, 2006 - to appear.
- [QUINLAN, 1993] QUINLAN, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
- [RIBEIRO, 2005] RIBEIRO, Marcela; BALAN, Andre G. R.; FELIPE, Joaquim C.; TRAINA, Agma Juci Machado; TRAINA JR., Caetano. *"Mining Statistical Association Rules to Select the Most Relevant Medical Image Features"*. First International Workshop on Mining Complex Data (IEEE MCD'05), Houston, USA, November 2005, pp.91-98.

- [RIBEIRO, 2006] RIBEIRO, Marcela; MARQUES, Joselene; TRAINA, Agma Juci Machado; TRAINA JR., Caetano. "Statistical Association Rules and Relevance Feedback: Powerful Allies to Improve the Retrieval of Medical Images". Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems, Salt Lake City, Utah, USA, June 2006, pp. 887 – 892.
- [RIJSBERGEN, 1979] RIJSBERGEN, C. J. v. *Information Retrieval*. 2nd edition, London, Butterworths, 1979.
- [RODRIGUES JR, 2003] RODRIGUES JR, José F.; CASTAÑÓN, Cesar A. B.; TRAINA, Agma Juci Machado; TRAINA JR., Caetano. "Using Efficient Visual Exploration Techniques to Evaluate Features for Content-based Image Retrieval". Proc. 16th Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAP'03), São Carlos - SP, October 12-15, 2003, pp. 183-190.
- [ROSA, 2002] ROSA, Natália Abdala *Uma abordagem pratica e eficiente de consultas por similaridade para suporte a diagnostico por imagens*. Dissertação de Mestrado. Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional do ICMC/USP. São Carlos-SP, 2002.
- [RVL/PURDUE, 2007] *Robot Vision Lab of Purdue University*. Disponível em: <http://cobweb.ecn.purdue.edu/RVL/> . Acessado em: Fevereiro, 2007.
- [SANTINI, 1998] SANTINI, S., JAIN, R., "Beyond Query by Example". Proceedings of the IEEE 1998.
- [SANTINI, 1999] SANTINI, S., JAIN, Ramesh. *Similarity Measures*. IEEE Transaction on Pattern Recognition and Machine Intelligence. Vol. 21, n. 9, pp. 871-883, Setembro, 1999.
- [SHYU, 1999] SHYU, C. R., BRODLEY, C. E., KAK, A. C., KOSAKA, A., AISEN, A. M., BRODERICK, L. S., "ASSERT: A Physician-in-the-loop Content-Based Image Retrieval System for HRCT Image Databases," Computer Vision and Image Understanding (Special Issue on Content-Based Retrieval from Image Databases), pp. 111-131, 1999.
- [SHYU, 2000] SHYU, C.R.; KAK, A.C.; BRODLEY, C.E.; PAVLOPOULOU, C.; CHYAN, M.F.; BRODERICK, L.S., *A Web-based CBIR-assisted learning tool for radiology education-anytime and anyplace*. IEEE International Conference on Multimedia and Exp., vol.1, no.pp.23-26 vol.1, 2000.
- [SILBERSCHATZ, 1999] SILBERSCHATZ, Abraham; KORTH, Henry F. e SUDARSHAN, S. *Sistema de Banco de Dados*. 3ª ed. – São Paulo: MAKRON Books, 1999.
- [SIM/HUG, 2007] *Service d'Informatique Médicale des Hôpitaux Universitaires de Genève*. Disponível em: <http://www.sim.hcuge.ch/>. Acessado em: Fevereiro, 2007.
- [THIES, 2004] THIES, C., GÜLD, M.O., FISCHER, B., LEHMANN, T.M.. *Content-based queries on the casimage database with the IRMA framework: A field report*. Results of the CLEF 2004 Cross-Language System Evaluation Campaign - Working Notes for the CLEF 2004 Workshop. ISTI-CNR, Pisa, Italy, 2004; 1: 611-20.

- [TRAINA, 1997a] TRAINA, Agma Juci Machado; SANTOS, Rildo Ribeiro dos; SENZAKO, Edna Y.. "A Support System for Content-Based Medical Image Retrieval in Object-Oriented Databases". Journal of Medical Systems, Plenum Press, vol 21, n. 6, 1997, pp 339-352.
- [TRAINA, 1997b] TRAINA, Agma Juci Machado; RIBEIRO, R. S.; SENZAKO, Edna Y.. "Content-based Medical Images Retrieval in Object Oriented Database". Proceedings of 10th IEEE Symposium on Computer-Based Medical System - Part II, Maribor- Slovenia June 1997, 67-72.
- [TRAINA, 1998] TRAINA, Agma Juci Machado; RIBEIRO, R. S.; SENZAKO, Edna Y.. "Support to Content-Based Image Query in Object-Oriented Databases". Proceedings of 1998 ACM Symposium on Applied Computing (SAC'98) - ACM Press, Atlanta, USA, 28/02 to 1/03/1998, pp.241-248.
- [TRAINA, 2002] TRAINA, Agma Juci Machado; TRAINA JR., Caetano; MARQUES, Paulo M. d. A. "cbPACS: PACS com Suporte à Recuperação de Imagens Médicas Baseada em Conteúdo". VIII Congresso Brasileiro de Informática em Saúde- CBIS'2002, Natal, RN-Brazil, 2002.
- [TRAINA, 2004] TRAINA, Agma Juci Machado; FIGUEIREDO, Josiel Maimoni de. "Including Conditional Operators in Content-Based Image Retrieval in Large Sets of Medical Exams". 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04), Bethesda, Maryland, 2004, pp. 85-90.
- [TVERSKY, 1977] Tversky, A. *Features of similarity*. Psychological Review 84(4), 327–352, 1977.
- [XIONG, 2005] XIONG, W., QIU, B., TIAN, Q., MULLER, H., XU, C. *A novel content-based medical image retrieval method based on query topic dependent image features (QTDIF)*. Proceedings of SPIE, 5748:123--133, 2005.
- [WANGENHEIM, 2003] WANGENHEIM, Christiane G. v., WANGENHEIM, Aldo v.. *Raciocínio Baseado em Casos*. 1ª ed. Manole, 2003.
- [WANGENHEIM, 2006] WANGENHEIM, Aldo v.. *Notas de Aula - Reconhecimento de Padrões I e II*. Universidade Federal de Santa Catarina, Curso de Ciência da Computação. 2006.
- [WATANABE, 2006] WATANABE, Carolina Y. V.; TRAINA, Agma Juci Machado. "Recuperação de Imagens Médicas por Conteúdo utilizando Wavelets e PCA". X Congresso Brasileiro de Informática em Saúde, Florianópolis- SC, Outubro 2006, p. 1-6.
- [WERBOS, 1974] WERBOS, P. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. PhD Dissertation. Committee on Applied Math., Harvard University, 1974.
- [WILLY, 2004] WILLY, P.M., KUFER, K.H.. *Content-based Medical Image Retrieval (CBMIR): An Intelligent Retrieval System for Handling Multiple Organs of Interest*. 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04), 24-25 June, Bethesda, MD, USA. p. 103.
- [VELTKAMP, 2001] VELTKAMP, Remco C. *Shape Matching: Similarity Measures and Algorithms*. Relatório Técnico, Dept. Ciência da Computação, Universidade de Utrecht – Holanda, Janeiro, 2001.

[VIPER, 2007] Group on Multimedia Information Retrieval and Management. Disponível em: <http://vipер.unige.ch/>. Acessado em: Fevereiro, 2007.

[ZERNIKE, 1934] F. ZERNIKE. . *Physica*, vol. 1, pp. 689, 1934.

[ZIOU, 1998] ZIOU, D., TABBONE, S. *Edge Detection Techniques – An Overview*. *International Journal of Pattern Recognition and Image Analysis*. Vol. 8, pp. 537--559, 1998.