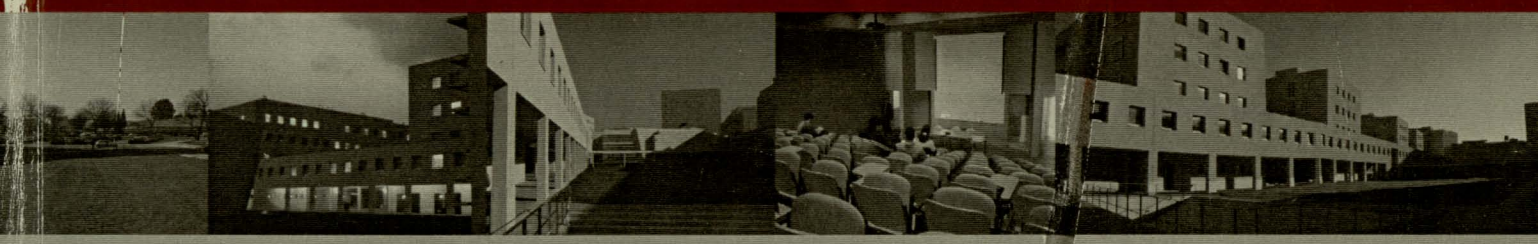
Universidade do Porto
Faculdade de Engenharia
FEUP



Ricardo Pereira Moura

Pesquisa em Imagens Combinando Informação Visual e Informação Textual

MGI
orto, 2004

Faculdade de Engenharia da Universidade do Porto

Mestrado em Gestão de Informação

Pesquisa em Imagens Combinando Informação Visual e Informação Textual

Dissertação do MGI – 2002/2003

Ricardo Pereira Moura

Orientadora: Prof. Maria Cristina Ribeiro

Julho de 2004

À minha família

Resumo

A quantidade de informação visual produzida, modificada e consumida aumentou exponencialmente na última década, devido ao aparecimento de novas vias de comunicação, à melhoria das tecnologias de comunicação existentes e à disponibilização de novas ferramentas de produção de conteúdos multimédia ao público em geral. Este aumento criou a necessidade de novos métodos nas áreas de pesquisa e recuperação de imagens, menos dependentes da intervenção humana. Neste documento são apresentados os principais desenvolvimentos ocorridos nestas áreas nos últimos anos.

Para além da necessidade de automatização, coloca-se o problema de ultrapassar a diferença entre as características de alto nível semântico, tipicamente extraídas de textos, e as características de baixo nível semântico, tipicamente extraídas de imagens. Actualmente, estas últimas características não podem fornecer por si só informação significativa para o utilizador que quer comunicar com o sistema de recuperação de informação através de uma linguagem de nível semântico mais elevado.

Na grande maioria das soluções propostas até ao momento para este problema foram utilizadas estruturas hierárquicas que procuram representar os diferentes níveis semânticos, ou foram adaptadas técnicas utilizadas previamente para pesquisa apenas em texto ou apenas em imagens. Pretende-se mostrar nesta dissertação que a pesquisa por conteúdos em imagens pode beneficiar da informação extraída destas por ferramentas de análise automática, tipicamente de baixo nível semântico, se esta for combinada com a informação textual proveniente de anotações relativas a essas imagens.

A grande quantidade de fontes de informação fez surgir a necessidade de uma estrutura de meta-informação adaptável mas com regras bem definidas, que possibilite a integração das várias estruturas de informação produzidas. O MPEG-7 visa responder a esta necessidade.

Propõe-se, nesta dissertação, testar uma resposta ao problema enunciado e ao mesmo tempo avaliar as características de baixo nível do módulo XM do MPEG-7. Com estes objectivos foi utilizada a técnica *Latent Semantic Indexing* (LSI), que permitiu agrupar as ocorrências de palavras e as características de imagens numa só estrutura de informação, servindo esta de base à pesquisa.

O protótipo construído para testar as soluções propostas utiliza como dados de teste ex-votos – objectos, neste caso, quadros, oferecidos a uma divindade em agradecimento de um favor – digitalizados e acompanhados de descrições textuais. Foram obtidos bons resultados nos testes sobre este protótipo, principalmente tendo em conta a ausência de coerência estilística nos dados utilizados.

Agradecimentos

À Prof. Maria Cristina Ribeiro pela atenção e paciência, ao Lucian e ao Catalin pela ajuda prestada na utilização do módulo de extracção de informação das imagens, ao Prof. Agostinho Araújo pelo empréstimo prolongado de literatura votiva.

E a todos os amigos e familiares que nunca pararam de perguntar por que é que a dissertação ainda não estava acabada.

Índice de Conteúdos

1 Introdução 1

1.1 A importância da anotação e pesquisa de imagens 1

1.2 O problema da interpretação de conteúdos e proposta de solução 1

1.3 As imagens de ex-votos 1

1.4 Estrutura da dissertação 2

2 Estado da arte na área da anotação e pesquisa de imagens 4

2.1 Pesquisa de texto 4

2.2 Anotação de conteúdos visuais 9

2.3 Utilização de características de baixo nível semântico 11

2.4 Pesquisa de imagens 13

2.5 Avaliação de resultados para sistemas de pesquisa em imagem 14

3 Combinação de características de diferentes níveis semânticos 16

3.1 A distância entre texto e imagem 16

3.2 Propostas anteriores 17

3.3 Abordagem ao problema 18

3.4 O módulo XM do MPEG-7 21

4 O protótipo 23

4.1 Descrição geral 23

4.2 Estrutura de objectos 25

4.3 *Software* utilizado 32

4.4 Caracterização dos dados utilizados 33

5 Resultados 35

5.1 Configuração e parâmetros do sistema 35

5.2 Resultados globais 35

5.3 Resultados de interrogações individuais com um só conceito 41

5.4 Resultados de interrogações individuais com mais do que um conceito 45

5.5 Comparação do desempenho dos descritores do MPEG-7 46

6 Conclusões e experiências futuras 52

Referências e Bibliografia 54

Software utilizado 57

1 Introdução

1.1 A importância da anotação e pesquisa de imagens

É indiscutível que uma das características da sociedade moderna, mais do que qualquer outra que a tenha precedido, é a enorme quantidade de informação produzida e consumida:

- Por todo o mundo, arquivos de informação visual em suporte físico conservaram ou pretendem conservar a sua informação em formato digital, de forma a evitar a deterioração da sua qualidade.
- Os equipamentos de produção de imagem e vídeo digital encontram-se cada vez mais acessíveis ao público em geral.
- Milhares de estações de televisão transmitem informação 24 horas por dia, 365 dias por ano.
- Até recentemente, apenas uma pequena parte das centenas de milhões de páginas na *World Wide Web* era indexada pelos motores de pesquisa [25].

O aumento do nível de produção e consumo acarretou a necessidade de sistemas que permitam pesquisar bases de dados de grandes dimensões. Inicialmente, a esmagadora maioria da investigação sobre técnicas de pesquisa de informação focou-se na utilização de texto [22], mas hoje em dia são necessárias técnicas específicas para a recuperação de imagens e vídeo, assim como esquemas de anotação que facilitem a interacção entre bases de dados distintas, como o MPEG-7.

1.2 O problema da interpretação de conteúdos e proposta de solução

Desde os primórdios do estudo de técnicas automáticas de recuperação de informação os investigadores têm-se deparado com o problema da subjectividade da informação retornada por qualquer sistema de recuperação que não distinga qual o utilizador que lhe está a aceder, podendo um mesmo documento ser considerado relevante ou irrelevante conforme a pessoa que o visualiza. No caso específico da recuperação de imagens, coloca-se outro problema, o de relacionar a informação passível de ser extraída da imagem, tipicamente de baixo nível semântico, com a interrogação introduzida pelo utilizador, tipicamente de alto nível semântico, o chamado *semantic gap* [44].

Propõe-se, nesta dissertação, a utilização de um método de pesquisa de imagens com o objectivo de minimizar este problema, integrando características de texto e características de baixo nível retiradas da imagem, tendo sido utilizados o módulo XM do MPEG-7 para a extracção destas últimas.

1.3 As imagens de ex-votos

O dicionário da língua portuguesa da Porto Editora define um ex-voto como "um objecto, quase sempre de índole piedosa, que se oferece a Deus ou a um Santo, em cumprimento de um voto."

Os ex-votos existem desde o nascimento das primeiras religiões. Os ex-votos cristãos mais antigos encontrados datam do século IV D.C. [39]. A maioria destes objectos são quadros em

madeira, mas muitos tipos de ex-votos foram oferecidos ao longo dos séculos, desde esculturas representando partes anatómicas (mãos, braços, etc.) até velas de cera e sacos de trigo. Henrique Rema [39] considera, inclusivamente, que até a Basílica e o Convento de Mafra são ex-votos, já que D. João V prometeu a sua construção se Santo António lhe desse herdeiros. Quando acompanhados por legendas, a sua linguagem apresenta características únicas, devidas não só à época em que foram produzidos mas também à sua natureza popular e religiosa. Nestas legendas, inscritas na maior parte das vezes nos próprios, é referido o milagre ocorrido e a entidade a quem se destina o agradecimento pela ocorrência deste: Cristo, Nossa Senhora ou um santo.

Segundo José Luís Porfírio [12], um ex-voto deve ser entendido como um agradecimento por uma ajuda divina, como uma memória, normalmente narrada em legenda, tanto de um acontecimento como das pessoas que o viveram ou presenciaram e como um testemunho, na forma da repetição de uma imagem cujo original muitas vezes estava exposto num santuário. Por exemplo, um mesmo santo aparece representado em vários ex-votos com postura e simbologia semelhante. Estas repetições tornam essas imagens em ícones, sendo mais facilmente identificáveis. Joaquim Oliveira Caetano também refere este aspecto dos ex-votos: "muitas vezes a pintura votiva resulta numa espécie de meta-imagem, no sentido em que frequentemente comporta dentro de si a imagem responsável pela acção miraculosa (...) [1]".

As obras pintadas ou esculpidas são na maioria toscas e, embora a maior parte dos autores não se identifique, nalguns casos foi possível relacionar o mesmo autor com várias obras espalhadas em zonas relativamente próximas. Não há um estilo distinto que percorra todos os ex-votos, apenas podem ser estabelecidas associações entre obras dedicadas à mesma entidade divina ou pertencentes ao mesmo santuário. Também não podem ser entendidos como parte de uma corrente artística: são simplesmente ingénuos [12].

A importância histórica dos ex-votos, como narrações pessoais de épocas e meios dos quais muitas vezes são fontes únicas de informação, é inestimável.

Neste trabalho é usada uma colecção de pinturas votivas digitalizadas do livro "Do Gesto à Memória – Ex-votos" [12], provenientes originalmente dos distritos de Viseu, Guarda e Castelo Branco. Estas pinturas datam maioritariamente dos sécs. XVII e XVIII.

1.4 Estrutura da dissertação

A dissertação encontra-se estruturada em seis capítulos, incluindo o presente.

No segundo capítulo são revistas as experiências mais interessantes dos últimos anos nas áreas de pesquisa e anotação de texto e imagem.

No terceiro capítulo é apresentado o problema do *semantic gap* e uma proposta para um protótipo de pesquisa de imagens que utiliza características de alto e baixo nível semântico, com o objectivo de minimizá-lo. São também apresentadas propostas anteriores com este mesmo objectivo e é introduzido o módulo XM do MPEG-7, que compreende as características de baixo nível utilizadas.

No quarto capítulo o protótipo é delineado, sendo apresentada a sua estrutura geral e as ferramentas externas utilizadas. Também se fornece aqui uma descrição mais técnica dos dados de teste utilizados, sendo focados os pormenores com mais interesse para a sua utilização no protótipo.

No quinto capítulo apresentam-se os resultados dos testes no protótipo, sendo apreciados os resultados quando são utilizadas todas as características de baixo nível simultaneamente e quando são utilizadas individualmente, de forma a comparar o desempenho dos vários descritores do MPEG-7. São também analisadas interrogações específicas, de forma a detectar possíveis anomalias no funcionamento do protótipo.

Por fim, no sexto capítulo são apresentadas as conclusões deste trabalho e referidas possíveis experiências futuras.

2 Estado da arte na área da anotação e pesquisa de imagens

As áreas da anotação e pesquisa de conteúdos multimédia têm sofrido uma enorme evolução nos últimos anos, propulsionada pelo aumento vertiginoso da quantidade de informação multimédia disponível e pelo interesse dos produtores de conteúdos e dos seus utilizadores. Nesta secção são revistos os principais progressos nestas áreas, não sendo referidos aspectos específicos relativos à anotação e pesquisa de vídeo que não possam ser também utilizados com imagens estáticas.

2.1 Pesquisa de texto

A investigação na pesquisa em texto começou o seu desenvolvimento muito antes da pesquisa em imagem, sendo natural que as técnicas aplicadas nesta última área descendam de técnicas utilizadas previamente na pesquisa em texto. Ao mesmo tempo, muitos dos problemas e limitações encontrados no campo da pesquisa em texto surgem também no campo da pesquisa em imagem, não se podendo portanto ignorar os desenvolvimentos essenciais da pesquisa em texto no estudo da pesquisa em imagem.

Para a pesquisa em texto existem três modelos tradicionais: o lógico booleano, o vectorial e o probabilístico. Para além da orientação segundo um destes modelos ou segundo modelos alternativos, pode também ser considerada num sistema de recuperação de informação a utilização de técnicas baseadas no estudo linguístico como, por exemplo, o uso de ontologias pré-existentes (como a *WordNet* [28]) ou a análise da sintaxe de um documento.

Modelo lógico

No modelo lógico, as interrogações são constituídas por termos relacionados por operadores booleanos (essencialmente AND, OR e NOT), sendo retornados ao utilizador documentos cujos termos satisfaçam a expressão lógica formada pela interrogação. O modelo lógico clássico tem caído em desuso, já que tem vindo a apresentar resultados inferiores aos modelos vectorial e probabilístico. Ao mesmo tempo, os testes realizados em sistemas com base neste modelo nas últimas conferências TREC têm mostrado as suas limitações [22]. A ineficácia dos sistemas que implementam este modelo foi atribuída a muitos factores, tendo Ed Greengrass destacado o problema das expressões booleanas que compõe cada interrogação poderem ser apenas verdadeiras ou falsas [17]. Esta característica origina a não devolução ao utilizador de qualquer documento ou a devolução de um número de documentos demasiado grande para o utilizador processar. Pela mesma razão é difícil nestes sistemas determinar a relevância dos resultados obtidos. De forma a minimizar estes problemas foram introduzidas várias extensões ao modelo tradicional que utilizam operadores booleanos modificados, permitindo a obtenção de valores de verdade para os resultados de expressões booleanas que não apenas verdadeiro ou falso.

Uma das modificações mais importantes é a utilização dos conjuntos difusos, que usam uma extensão da teoria de conjuntos tradicionais. Enquanto que nesta a função de pertença de um elemento a um conjunto só pode tomar os valores de 1 (verdadeiro, o elemento pertence ao conjunto) ou 0 (falso, o elemento não pertence ao conjunto), num conjunto difuso o elemento pode ter um grau de pertença ao conjunto em causa entre 0 e 1. Procura-se assim representar a incerteza que existe na vida real em praticamente todas as suposições, não podendo esta

incerteza ser modelada com a lógica booleana convencional. Na recuperação de informação, os conjuntos difusos permitem estabelecer relações incertas entre tópicos e documentos [5].

Modelo vectorial

No modelo vectorial, cada termo constituinte da interrogação é representado por uma coordenada num espaço de informação multidimensional, sendo este espaço conservado numa matriz de termos por documentos. Quando o utilizador apresenta uma interrogação ao sistema, é construído um vector com os termos desta, sendo o vector comparado com os vectores representativos de cada documento da base de dados pesquisada – estes vectores formam as colunas da matriz referida. A investigação deste modelo tem sido concentrada no cálculo dos pesos de cada termo na representação de cada documento e nas medidas de semelhança da interrogação com os documentos.

No cálculo do peso de termos no modelo vectorial, têm sido considerados essencialmente três métodos [23]:

Cálculo de peso com base na *Inverse Document Frequency* (IDF) [21] – É um método muito discriminante, já que a IDF reflecte a importância de um termo não só dentro de um documento mas dentro da colecção inteira de dados. Normalmente a IDF é multiplicada pela frequência do termo no documento em causa.

Rácio sinal-ruído [23] – É calculado através de dois indicadores: sinal e ruído. O ruído indica se o termo está concentrado num número pequeno ou grande de documentos, sendo maior quanto mais distribuído o termo se encontra pelos documentos que compõem a colecção de dados. O sinal é calculado pela subtracção do ruído ao número de documentos que contêm o termo em causa. Uma vez que o peso de um termo aumenta de acordo com a sua raridade, é semelhante à IDF.

Normalização através do tamanho do documento [40] – Esta normalização é utilizada em medidas de semelhança de vectores documentais como, por exemplo, no cálculo do coseno do ângulo entre dois vectores, para minimizar o efeito do tamanho de um documento (no que diz respeito ao número de termos nele contidos) na sua classificação. Da mesma forma, esta normalização é utilizada no cálculo de pesos para relativizar a importância do tamanho de um documento no cálculo da importância local e global de um termo.

Uma alternativa a estes métodos de cálculo de pesos, que também pode ser aplicada em conjunção com eles, é a utilização de métodos probabilísticos, em que um termo obtém um peso mais elevado de acordo com a sua presença em documentos relevantes para uma determinada interrogação realizada previamente por um utilizador.

Após o cálculo do peso de cada termo, independentemente do método utilizado, é possível a utilização de técnicas estatísticas para tentar encontrar co-ocorrências de valores na matriz de termos por documentos. Uma destas técnicas é o LSI, que é usado no modelo proposto (ver secção 3.3).

As medidas de semelhança podem ser categorizadas naquelas que utilizam constantes nas fórmulas, de forma a permitir ajustar a performance do sistema para diferentes contextos, e nas que se baseiam no cálculo do produto interno dos pares de vectores interrogação-documento. Jung *et al* consideram que a segunda classe de medidas é preferível [23], já que a primeira implica a utilização de diferentes constantes para optimizar o desempenho consoante o conjunto de dados. A medida mais utilizada no cálculo da semelhança, pertencente à

segunda classe, é o coseno entre o ângulo formado pelo vector de interrogação e cada vector de documento. Como foi já referido, esta medida normaliza o produto interno dos vectores, minimizando os efeitos da presença de documentos com tamanhos muito díspares. Outras medidas pertencentes à mesma classe modificam essencialmente o método de normalização.

Modelo probabilístico

No modelo probabilístico, para cada interrogação assume-se a existência de um conjunto de documentos relevantes R e de um conjunto de documentos irrelevantes NR, sendo retornados ao utilizador documentos ordenados pela probabilidade de pertencerem a R. A forma como esta probabilidade é calculada distingue as variantes deste modelo. O modelo probabilístico convencional tem a desvantagem de necessitar que os seus utilizadores indiquem a relevância dos documentos retornados para as suas interrogações, de forma a estimar as probabilidades iniciais; esta restrição pode, no entanto, ser levantada facilmente.

A utilização de redes Bayesianas, que podem ser vistas como alternativas a este modelo, obteve um impacto significativo na recuperação de informação e na representação de conhecimento, tendo sido introduzidas variantes à sua arquitectura tradicional. Estas redes baseiam-se na teoria de probabilidades bayesiana convencional, em que é relacionada a probabilidade *a priori* da relevância de um qualquer documento com a probabilidade *a posteriori* de um documento analisado ser relevante, de acordo com as suas características [17]. Uma rede Bayesiana é um grafo dirigido acíclico em que cada nó representa uma variável aleatória e cada aresta representa as relações entre essas variáveis. Na estrutura mais básica, existem dois níveis de nós: no primeiro, cada nó representa a probabilidade de um determinado documento ser relevante para uma determinada interrogação. No segundo nível, cada nó representa a probabilidade de um determinado documento possuir um determinado termo. As arestas são estabelecidas entre cada nó do primeiro nível e um ou mais nós do segundo nível. Esta estrutura presume a determinação de um conjunto de probabilidades representativas das variáveis aos dois níveis. Para os nós do primeiro nível, é necessário determinar a probabilidade do documento representado ser relevante para a interrogação em causa. Para os nós do segundo nível, é necessário determinar a probabilidade de uma determinado termo estar presente no documento representado, sendo esta probabilidade condicionada pela relevância deste documento para as interrogações nas quais o termo é encontrado. A tarefa de recuperação de informação neste contexto resume-se ao cálculo da probabilidade *a posteriori* de um documento ser relevante para uma determinada interrogação de acordo com a presença ou ausência dos vários termos associados a essa interrogação no documento em causa, podendo esta probabilidade posterior ser calculada através dos dois conjuntos de probabilidades previamente referidos [15].

Desenvolvimentos e resultados recentes

As *Text REtrieval Conferences* (TRECs) têm desempenhado um papel essencial há mais de uma década, não só por terem propiciado avaliações independentes e extensas dos processos de indexação e pesquisa de sistemas de recuperação de texto mas também por terem estimulado novos desenvolvimentos nesta área [22]. Todas as TRECs se têm focado em duas tarefas de pesquisa, a *ad hoc* e a *routing*. Em tarefas *ad hoc*, assume-se que estão a ser utilizadas novas interrogações – as necessidades de informação do utilizador estão em mudança constante – sobre um conjunto de dados estático, sendo dada grande importância à velocidade de recuperação de informação. Em tarefas *routing*, assume-se que são sempre

utilizadas as mesmas interrogações sobre um conjunto de dados dinâmico, ao qual novos dados podem ser adicionados. Neste contexto, o objectivo é melhorar a interrogação do utilizador, aproximando-a o mais possível das necessidades de informação deste [18][15].

Spärck Jones referiu em [22] algumas tendências e resultados, tanto positivos como negativos, da utilização de diferentes técnicas na recuperação de texto que as TRECs permitiram observar:

- Em tarefas de pesquisa *ad hoc* generalizada, as classificações construídas explicitamente e as ontologias, quer construídas de forma automática quer de forma manual, não acrescentaram um valor substancial aos resultados obtidos sem a sua utilização;
- A análise sofisticada de linguagem natural na recuperação não melhorou significativamente os resultados na recuperação de informação;
- A utilização de técnicas estatísticas produziu desempenhos razoáveis a partir de uma interrogação inicial pobre, principalmente quando foi utilizada *query relevance feedback*, o que demonstra a validade destas técnicas mesmo nos casos em que o empenho do utilizador não é garantido;
- Em termos gerais, os modelos baseados em procedimentos estatísticos, como o modelo vectorial e o probabilístico, obtiveram resultados bons e semelhantes;

Comparando os três modelos clássicos, a investigação na área da recuperação de informação tem vindo a obter melhores resultados com a utilização dos modelos vectorial e probabilístico do que com a utilização do modelo lógico, à custa de uma maior utilização de recursos computacionais [15]. Um quarto modelo com bases probabilísticas, o modelo de linguagem, foi considerado por Spärck Jones o desenvolvimento mais interessante dos últimos tempos [22].

No modelo de linguagem, a noção principal é que, ao contrário do modelo probabilístico convencional, a interrogação é gerada pelo documento. Cada documento é equacionado como parte duma linguagem, sendo estimadas as probabilidades da presença de termos individuais em cada documento. Dado o conjunto de termos de uma interrogação, são computadas as probabilidades de geração destes termos de acordo com a estrutura de cada documento. A classificação dos documentos é feita através da multiplicação das probabilidades de geração de cada termo da interrogação, sendo que um documento é mais relevante quanto maior for a probabilidade deste gerar a interrogação [45].

Pesquisa na WWW

A explosão da WWW incentivou a investigação na área da pesquisa de informação e introduziu novas possibilidades e desafios que não se colocam na pesquisa da maioria das bases de dados. O gigantesco volume de informação presente na WWW inviabiliza a pesquisa em tempo real sem utilização de indexação. A constante mutação desta informação torna necessária a actualização constante dos índices dos motores de pesquisa, de forma a que estes se mantenham actualizados. Outros motivos contribuem para a impossibilidade da indexação total da WWW por parte de um motor de pesquisa [47][36]:

- A existência de páginas para as quais não existem hiperligações que apontem para elas, de acesso restrito e cujo conteúdo é gerado automaticamente de acordo com a interacção dos utilizadores;

- As limitações existentes na capacidade de armazenamento de dados e na velocidade de processamento destes;
- O tamanho estimado da WWW, que até recentemente tinha vindo a crescer mais depressa do que o número de páginas armazenadas pelos motores de pesquisa.

Estes dois últimos factores têm vindo a ser contrariados, o primeiro deles devido à evolução tecnológica, o segundo devido à estagnação da WWW que, inclusivamente, decresceu de tamanho entre 2001 e 2002, em contraste com o seu crescimento exponencial em anos anteriores [36]. Ao mesmo tempo, o motor de pesquisa *Google*[1], neste momento com o maior índice da WWW, passou de 500 mil páginas indexadas em Junho de 2000 para 3.2 biliões em Setembro de 2003 [48]. Relativamente a estes valores, é importante referir que eles foram reclamados pelo próprio *Google*, que uma percentagem significativa dos documentos apenas é indexada parcialmente, por diversas razões (páginas nunca processadas pelo *Google*, páginas desaparecidas, etc.) [30] e que apenas são indexados os primeiros 101 KB de cada página [31]. Outra dificuldade encontrada no desenvolvimento de motores de pesquisa para a WWW prende-se com a inexistência de controlo de qualidade e a heterogeneidade do seu conteúdo [38].

Apesar dos problemas referidos, a estrutura da WWW possibilita o desenvolvimento de técnicas únicas para a recuperação de informação através do aproveitamento das características dos ficheiros HTML que a compõem. Em particular, a avaliação da importância de cada página fornecida pelas hiperligações tem despoletado grande interesse, sendo esta tecnologia utilizada em muitos motores com sucesso, sendo o exemplo mais conhecido o *Google*. O *PageRank*, um dos algoritmos deste motor de pesquisa, ainda hoje é relevante e foi inovador na altura da sua criação, já que anteriormente apenas tinha sido utilizada a contagem de hiperligações que apontavam para uma determinada página para avalizar a sua importância. O *PageRank* classifica uma página A de acordo com a soma das classificações das páginas que apontam para A, sendo assim previstos tanto os casos em que uma página tem muitos apontadores para si própria como os casos em que apenas poucas páginas com uma classificação alta para ela apontam [38].

Outra técnica muito influente que utiliza hiperligações na WWW e que gerou várias adaptações foi introduzida por Kleinberg [24], baseada nos conceitos de páginas *authority* e páginas *hub*. Páginas *authority* são aquelas que, dentro do conjunto de páginas relevantes para uma determinada interrogação, são as essenciais, as mais representativas. Páginas *hub* são aquelas que possuem várias hiperligações para páginas *authority* interrelacionadas. O algoritmo desenvolvido pelo mesmo autor explora estes dois conceitos, identificando durante a indexação as páginas que mais se enquadram nestes dois tipos e classificando estas páginas como as mais relevantes para os tópicos que representam.

A utilização de informação relativa a hiperligações resolve dois problemas encontrados na WWW de difícil, se não impossível, resolução com a utilização de técnicas puramente textuais. O primeiro é o problema da abundância, como o denominou Kleinberg [24]: "o número de páginas que poderia ser considerado relevante para uma determinada interrogação é demasiado grande para o seu consumo por parte de um utilizador". Esta situação ocorre com grande frequência na WWW, tornando-se assim importante definir quais as páginas *authority* para interrogações que devolvam um grande número de documentos relevantes. O segundo

[1] http://www.google.com/

problema refere-se ao facto de que muitas páginas na WWW não contêm no seu texto os termos necessários para a sua descrição ou, embora sejam as páginas mais relevantes para uma determinada interrogação, não são as que usam os termos dessa interrogação com mais frequência.

A utilização de hiperligações acarreta também alguns problemas próprios: foi necessário introduzir medidas no algoritmo *PageRank* para impedir que entrasse em ciclo, devido a situações em que é possível seguir várias hiperligações de um ponto inicial até se regressar ao mesmo ponto [38], um problema enfrentado por todas as técnicas que sigam este rumo. Ao mesmo tempo, muitas das hiperligações são criadas apenas para navegação ou para fins publicitários. Outro problema referido por Kleinberg [24] na versão mais simples do seu algoritmo prende-se com a necessidade de distinguir páginas populares (com muitos apontadores para elas) de páginas relevantes, já que uma página muito popular teria sempre uma classificação elevada para uma grande parte das interrogações mas não seria relevante para a maioria destas.

É interessante referir a observação de Spärck Jones [22] de que, no que diz respeito aos resultados observados ao longo das TRECs, a utilização da informação relativa a hiperligações na pesquisa generalizada da *World Wide Web* (WWW) não melhorou os resultados obtidos com a utilização de termos relativos apenas ao conteúdo, o que contrasta com o aparente sucesso encontrado por vários motores de pesquisa com esta direcção.

A existência de inúmeros motores de pesquisa na WWW incentivou a criação de meta-motores de pesquisa, como o são o *MetaCrawler*[2] e o *Copernic*[3]. Um meta-motor apresenta interrogações em paralelo a um conjunto de motores de pesquisa pré-existentes, integra os resultados obtidos e retorna-os ao utilizador. Um sistema deste género procura melhorar os resultados dos motores de pesquisa utilizados, retirando destes as páginas que considerar irrelevantes para a interrogação do utilizador [19]. Mais recentemente, meta-motores de pesquisa como o *Vivisimo*[4] utilizaram *clustering* nos resultados obtidos, organizando hierarquias de classes que são apresentadas ao utilizador.

Estes sistemas têm como vantagens o uso de motores de pesquisa que podem não ser conhecidos pelo utilizador [19] e a possibilidade de apresentar mais resultados relevantes do que aqueles retornados por um só motor de pesquisa. Podem também aplicar algoritmos de classificação próprios que melhoram os resultados retirados dos motores de pesquisa nos quais se apoiam.

2.2 Anotação de conteúdos visuais

A exploração da informação visual é uma área em que tem requerido uma continuada reflexão sobre a semântica das imagens e a forma de a captar. Obter uma representação de uma imagem com significado para um utilizador, ou seja, de alto nível semântico, é uma tarefa executada facilmente por pessoas, mas a sua realização sem interferência humana implica sempre ultrapassar o *semantic gap*, já que é necessário relacionar as características de baixo nível da imagem, fáceis de obter automaticamente, com os objectos e conceitos representados.

---

[2] http://www.metacrawler.com/

[3] http://www.copernic.com/

[4] http://www.vivisimo.com/

O texto que eventualmente acompanha a imagem pode ser-lhe associado, com ou sem o uso de uma taxionomia de termos pré-definida, como a *WordNet* [28], que estabelece relações semânticas entre os termos (por exemplo, relações de sinonímia e de generalização). Para além do problema do *semantic gap*, na maior parte dos casos as características de alto e baixo nível semântico são extraídas usando técnicas e ferramentas diferentes [4]. A alternativa é a anotação manual, sendo associados termos a cada imagem numa base de dados, mas esta apresenta várias desvantagens [10]:

- É demasiado custosa em bases de dados de grande tamanho;
- As anotações são subjectivas, já que na maior parte dos casos o anotador e o utilizador não são a mesma pessoa; para o próprio anotador não se pode garantir consistência ao longo do tempo e capacidade de análise semelhante em domínios diversos;
- A utilização de termos não permite normalmente que seja apresentada uma imagem como interrogação.

De forma a transpor estes problemas, foram feitas propostas que conjugam numa só estrutura de dados características de baixo nível da imagem e informação de alto nível semântico.

Colombo *et al* propuseram um sistema para a recuperação de imagens em que é feita uma anotação prévia, sendo esta baseada em estudos anteriores sobre a relação entre as características de baixo nível da imagem e as emoções e as impressões que estas causam no observador [10]. As imagens são categorizadas em emoções, sendo as principais características da imagem de interrogação analisadas para verificar em que categorias esta se enquadra e sendo retornadas imagens que se supõe proporcionarem emoções semelhantes. A interrogação é feita através de características de baixo nível da imagem, como o contraste de saturação e o contraste de calor.

Barnard *et al* propuseram uma estrutura hierárquica de termos de diferentes níveis semânticos [3]. Cada imagem utilizada neste sistema é segmentada, sendo utilizados atributos de cor, textura e forma para caracterizar cada segmento. São utilizados métodos estatísticos para analisar correlações de termos e características de imagem, sendo estimadas as probabilidades de um dado segmento ou termo pertencer a um determinado nível semântico ou a um determinado nó da estrutura hierárquica. Os autores prevêem que a recuperação de imagens possa ser feita através de interrogações textuais, tendo, no entanto, sido criado apenas o esquema de anotação em si e não um motor de pesquisa para o mesmo.

Benitez e Chang propuseram novos métodos para a integração e síntese de informação de diferentes níveis semânticos, desenvolvidos dentro do sistema *Intelligent Multimedia Knowledge Application* (IMKA) [4]. Este sistema visa a construção de estruturas de conhecimento a partir de informação multimédia e a possibilidade de desenvolvimento de aplicações que utilizem essas estruturas. O IMKA utiliza a estrutura de representação de conhecimento *MediaNet*, que integra conceitos perceptuais, como histogramas de cor e padrões de textura, e conceitos semânticos, como palavras, estabelecendo relações entre estes elementos. As relações estabelecidas tanto podem ser relações semânticas tradicionais, como as existentes no *WordNet* [28], ou relações de semelhança ou de restrição entre as características de baixo nível semântico. Para construir estas relações é em primeiro lugar estimada a presença de conceitos numa imagem ou nas descrições associadas através de algoritmos de classificação executados sobre os descritores de alto e baixo nível semântico, sendo em seguida construída uma rede Bayesiana para encontrar relações estatísticas entre os conceitos apreendidos. Por fim, a informação apreendida através das duas técnicas é integrada

numa única estrutura. A síntese desta estrutura é realizada através do agrupamento de conceitos num único conceito e da consequente redução de relações entre eles.

A norma ISO MPEG-7 foi desenvolvida para promover a integração de anotações provenientes de fontes de dados diversas num único conjunto de estruturas de anotação e facilitar a pesquisa e recuperação eficientes de informação através da utilização destas estruturas, numa época em que cada vez mais imagens digitais e vídeos são fornecidos ao público por produtores diversificados [43]. O MPEG-7 abrange apenas a descrição de conteúdos, não procura especificar como deve ser feita a produção ou consumo destes.

Os conceitos de base do MPEG-7 são os de Esquemas de Descrição (*Description Schemes* – DSs), Descritores (Descriptors – Ds) e Linguagem de Definição de Descritores (*Description Definition Language* – DDL). Os DSs são estruturas de meta-informação escritas em XML, que descrevem a estrutura e a semântica de conteúdos multimédia. No contexto destes conteúdos, os DSs foram concebidos para representar objectos de alto nível semântico como, por exemplo, regiões de uma imagem ou texto, enquanto que os Ds pretendem representar características de baixo nível semântico, como por exemplo cor ou textura. Um DS pode representar estruturas mais complexas, podendo ser constituído por vários DSs e Ds. Os DSs são definidos utilizando a DDL, baseada na XML *Schema Language* (criada para definir a estrutura de documentos XML). O MPEG-7 prevê um conjunto de DSs e Ds que satisfaçam as necessidades relativas à criação e utilização de meta-informação da grande maioria das aplicações multimédia. A DDL também permite a criação e modificação de DSs e Ds para os restantes casos [27]. As estruturas de dados do MPEG-7 concebidas para a descrição e anotação de conteúdos audiovisuais são os Esquemas de descrição Multimédia (Multimedia Description Tools – MDS), sendo estes compostos por um conjunto de DSs e Ds.

2.3 Utilização de características de baixo nível semântico

Nesta secção são revistas as principais características de baixo nível extraídas automaticamente de imagens e as suas possíveis aplicações, não sendo analisados os algoritmos de extracção destas ou a eficiência computacional desta extracção. Estas características são na maior parte dos casos relativas a pormenores observáveis na imagem [6], podendo ser relativas à cor, à textura ou à forma, ou ser formadas por uma combinação destas categorias.

Na extracção de uma destas características coloca-se o problema de manter a sua robustez quando a imagem é deslocada, rodada ou quando o seu tamanho é modificado – exceptuando quando estas modificações não introduzem alterações no descritor devido à natureza do próprio, como é o caso do histograma de cor simples. Quanto mais robusta uma característica for, mais fiável é a sua utilização na comparação de duas ou mais imagens.

Um passo que normalmente precede a extracção de características de baixo nível é a segmentação, que procura dividir uma imagem nos segmentos que a compõem. Um exemplo desta técnica é utilizado no sistema Blobworld [7], em que a informação de cor e textura contribui para a detecção de zonas coerentes.

Características de cor

Relativamente à cor, as características mais vulgares são os histogramas, que indicam o número de ocorrências de cada cor. Na computação destas características, é escolhido em primeiro lugar um espaço de cor (por exemplo, RGB, HSV, LUV, etc.) e em seguida as

ocorrências de cada cor na imagem são distribuídas por intervalos pertencentes a cada canal (por exemplo, o espaço de cor RGB tem os canais R, G e B). A escolha do número de intervalos é importante, já que quanto menor este número é, mais eficiente é a computação e menos exacta é a representação da imagem. Outra técnica, os *color moments*, caracterizam a distribuição da cor nos seus vários canais [9]. O principal defeito destas duas técnicas é que não fornecem qualquer informação relativa à posição da cor na imagem, o que reconhecidamente torna difícil a distinção de objectos numa imagem sem recorrer a outras características de textura e de forma [46]. Mais recentemente foram introduzidas outras características que tentam colmatar esta lacuna, nomeadamente os correlogramas, introduzidos por Huang *et al* [20], que relacionam as cores dos pixéis com as distâncias entre eles, os autocorrelogramas, que são um subconjunto dos correlogramas, o *Color Structure Descriptor* e o *Color Layout Descriptor* do MPEG-7 (ver secção 3.4).

Ojala *et al* [35] utilizaram autocorrelogramas com o espaço HSV em vez do espaço RGB usado previamente em experiências com correlogramas, tendo obtido resultados ligeiramente superiores e bons resultados na distinção de categorias semânticas em que as imagens haviam sido enquadradas manualmente. O espaço HSV supostamente fornece uma correspondência mais próxima da percepção humana da diferença entre cores do que outros espaços, nomeadamente o RGB.

Chiang *et al* [9] propuseram uma técnica que adiciona informação sobre a região a que um pixel pertence aos histogramas de cor e aos *color moments*, sendo as regiões determinadas com uma segmentação inicial da imagem. Foram obtidos bons resultados comparativamente às mesmas características sem utilização de segmentação. Tao e Grosky introduziram os anglogramas de cor [49], que são histogramas dos ângulos dos triângulos formados pelas posições de cores semelhantes numa imagem. Esta técnica tem a vantagem de ser resistente às três modificações já referidas. Todas estas características podem ser utilizadas em comparação vectorial, já que são discretas e de dimensão constante, isto é, o número de atributos numéricos que compõe cada característica é sempre o mesmo, independentemente da imagem à qual a característica se refere.

Características de textura

As características relativas à textura são normalmente extraídas de imagens em escala de cinza, que podem ser obtidas convertendo qualquer imagem a cores para esta escala. A robustez destas características é avaliada não só pela estabilidade quando são efectuadas translações, rotações e mudanças de tamanho mas também quando as propriedades da escala de cinza são alteradas. As técnicas de extracção relativas à textura tentam normalmente gerar descritores que não apresentam variações significativas quando uma ou, no máximo, duas destas quatro modificações ocorrem. Os três paradigmas principais na extracção deste tipo de características são: w*avelets*, *Gauss-Markov Random Field* (GMRF) e filtragem de Gabor [32], esta última utilizada no *Homogeneous Texture Descriptor* do MPEG-7 (ver secção 3.4).

Características de forma

Dois exemplos de características de baixo nível relativas a formas são o *Edge Histogram Descriptor* (EHD) do MPEG-7 (ver secção 3.4) e o anglograma, podendo-se considerar que este último também caracteriza a cor de uma imagem. O EHD é bastante semelhante na concepção ao anglograma, já que também cria um histograma dos ângulos das arestas

encontrados nas imagens, mas classifica-os apenas em cinco categorias: verticais, horizontais, orientados a 45°, orientados a 135° e isotrópicos [34].

2.4 Pesquisa de imagens

Nos últimos anos o interesse na pesquisa em imagens tem aumentado, já que até recentemente a maioria do esforço de pesquisa se havia concentrado na recuperação de texto. Assim, não é surpreendente verificar que muitas das soluções apresentadas para a recuperação de imagens foram adaptadas de soluções criadas previamente para a recuperação textual.

Inicialmente as técnicas de recuperação de imagens baseavam-se apenas no texto existente no documento a elas associado ou apenas nas características de baixo nível da imagem. A recuperação de informação utilizando estas últimas técnicas é denominada *Content Based Image Retrieval* – CBIR. Ultimamente têm surgido técnicas híbridas, que tentam conjugar características de alto e baixo nível semântico, sendo estas que se discutem em mais pormenor na secção 3.2.

Sistemas QBIC

Nos sistemas *Query-By-Image Content* (QBIC) a interrogação introduzida é uma imagem ou uma característica de baixo nível de imagem. O sistema QBIC desenvolvido pela IBM foi dos primeiros deste género, tendo-lhe inclusivamente dado o nome [14]. Este sistema permitia ao utilizador pesquisar uma base de dados de imagens através de exemplos de imagens, esboços desenhados pelo utilizador e padrões de cor e textura. A alternativa aos sistemas QBIC são sistemas em que a interrogação é introduzida sob a forma de texto, quer numa linguagem estrita (como por exemplo SQL) quer em linguagem natural.

Sclaroff et al [41] expandiram a ideia da utilização de um conjunto de robots, como já é feito em muitos dos motores de pesquisa de texto da WWW, como o *Google* ou o *HotBot*[5], para pesquisarem e tratarem imagens em simultâneo. Nesta experiência é apresentado ao utilizador um conjunto de imagens aleatórias, indicando este quais considera relevantes ou pedindo outro conjunto de imagens. A *query relevance feedback* é realizada neste sistema através de um algoritmo próprio que define as métricas de distância das imagens seleccionadas às restantes em tempo real.

Shahabi e Chen propuseram um sistema QBIC de pesquisa com interrogações *soft*, baseado em conjuntos difusos, em que uma imagem pode ser enquadrada em mais do que uma categoria ao mesmo tempo [42]. Esta classificação é baseada em *query relevance feedback*, tanto na comparação de características de baixo nível das imagens seleccionadas como na comparação de outras propriedades previamente associadas à imagem, podendo estas últimas ser pesquisadas através de interrogações SQL. Nesta abordagem existe um perfil único para cada utilizador, ou seja, podem ser retornados resultados diferentes conforme a pessoa que está a aceder ao sistema. Cada utilizador pode ainda indicar outros utilizadores em cujas opiniões confie, sendo estas também utilizadas na classificação de resultados.

Outro sistema QBIC é o Blobworld [7], que emprega segmentação na imagem introduzida pelo utilizador para lhe permitir seleccionar os segmentos que considera relevantes, sendo depois utilizada separadamente a distribuição de cor e textura em cada região na recuperação

[5] http://www.hotbot.com/

de informação. Este método permitiu aumentar a precisão relativamente aos resultados obtidos sem a utilização de segmentação. Ozcanli e Vural [37] construíram um sistema em que todo o processo de segmentação é conduzido pelo utilizador. A interrogação é introduzida através de uma imagem que é dividida em rectângulos e na qual o próprio utilizador selecciona os rectângulos que contêm as características em que está interessado. O utilizador escolhe também as características de baixo nível, nomeadamente de cor e textura, que são relevantes para representar cada um destes segmentos. O conjunto de rectângulos seleccionado é deslizado sobre as restantes imagens da base de dados, de forma a que os rectângulos sejam comparados nas mesmas posições relativas da imagem de interrogação. Os resultados obtidos mostraram melhorias relativamente a técnicas que usam características globais da imagem, mas o principal defeito do sistema é a sua sensibilidade à modificação de tamanho e rotação das imagens.

Sistemas com interrogações textuais

Um exemplo de um sistema de recuperação de imagens em que a interrogação é introduzida sob a forma de texto e em que apenas é utilizado texto na indexação e recuperação de conteúdos é o European Visual Archive (EVA) [11], sendo este método também utilizado na pesquisa de imagens do *Google* e na maioria dos outros sistemas existentes na WWW. O EVA permite aceder de momento a uma colecção de fotografias dos arquivos das cidades de Antuérpia e de Londres, podendo o utilizador realizar a sua pesquisa em várias línguas, neste momento inglês, alemão e holandês.

2.5 Avaliação de resultados para sistemas de pesquisa em imagem

A maior parte dos métodos de avaliação dos resultados de pesquisa em texto aplicam-se também à pesquisa de imagens. As medidas mais comuns são a recuperação e precisão. A recuperação é definida pelo número de documentos relevantes recuperados sobre o número total de documentos relevantes presentes na base de dados pesquisada. A precisão é definida pelo número de documentos relevantes obtidos sobre o total de documentos recuperados. A partir destas duas medidas podem ser criadas curvas de precisão/recuperação, que são a métrica *de facto* na análise de desempenho na área de recuperação de informação, permitindo comparar graficamente, de forma intuitiva, diferentes algoritmos de recuperação [2].

Outra métrica utilizada é a precisão-R, que se calcula da seguinte forma: se R é o número total de documentos relevantes para uma determinada interrogação, a precisão-R é a razão entre o número de documentos relevantes encontrados nos primeiros R documentos retornados e R. Podem ser obtidos histogramas desta medida, sendo esta útil para observar o comportamento de um algoritmo numa interrogação individual [2].

Quando são usadas interrogações *soft*, as métricas tradicionais referidas não podem ser utilizadas, sendo aplicadas medidas de semelhança entre os resultados de uma interrogação e os resultados esperados pelo utilizador.

A comparação do desempenho de diferentes propostas para a recuperação de imagens é um problema difícil, já que estas utilizam diferentes conjuntos de imagens de teste. Aliás, a maior parte das experiências utiliza conjuntos propositadamente favoráveis, em que existem grupos de imagens muito diferentes entre si, mas em que dentro de cada grupo as imagens são muito semelhantes [29]. Müller *et al* apresentaram uma proposta inovadora para um método automático e quantitativo de comparação do desempenho de sistemas CBIR [29], baseado no

Multimedia Retrieval Mark-up Language (MRML[6]), desenvolvido pelos mesmos autores. O MRML é um protocolo de comunicação baseado em XML que visa unificar o acesso a sistemas de recuperação multimédia, separando a interface de pesquisa do motor de pesquisa em si, sendo destinado em particular a sistemas CBIR. O sistema desenvolvido prevê um servidor e uma base de dados associada a este. O servidor comunica com um qualquer sistema de recuperação de informação, desde que este utilize o MRML, e fornece informações sobre a base de dados para que o sistema de recuperação possa indexá-la. Em seguida, este último é interrogado e, após receber os resultados de cada interrogação, o servidor calcula e devolve automaticamente medidas de precisão, recuperação, precisão-R e gráficos de precisão/recuperação.

[6] http://mrml.net/

3 Combinação de características de diferentes níveis semânticos

Neste capítulo é discutida a problemática do estabelecimento automático de relações com significado entre texto e imagem e possíveis utilizações conjuntas destes dois tipos de informação. Em seguida são apresentados sistemas já construídos que procurem solucionar o problema do *semantic gap* e ao mesmo tempo fornecer aos seus utilizadores uma interface que lhes permita introduzir interrogações de alto nível. Por fim, é apresentada uma nova proposta para uma abordagem a este problema.

3.1 A distância entre texto e imagem

O texto e a imagem, quando ocorrem em conjunto num documento, tendem a fornecer informação de formas distintas mas, muitas vezes, complementam-se. A imagem pode conter informação semântica indispensável para a compreensão total do texto e vice-versa. Por outro lado, texto e imagem podem, por vezes, ser compreendidos de forma independente, quando a imagem representa conceitos independentes deste ou é meramente ilustrativa, não contendo qualquer relação observável com o texto.

A grande maioria das soluções propostas até ao momento para a recuperação automática de imagens explora apenas a semântica do texto, utilizando técnicas já existentes para a recuperação de informação textual, ou apenas a semântica da imagem, utilizando características de baixo nível da imagem, como cor, textura ou forma. O principal problema do primeiro tipo de técnicas prende-se com as situações em que o texto não contém todos os conceitos necessários para representar uma imagem, não podendo portanto satisfazer todas as interrogações relevantes para esta. O principal problema do segundo tipo de técnicas está no já referido *semantic gap*.

As técnicas híbridas, que exploram tanto a informação contida na imagem como no texto que a rodeia, têm sido alvo de muito interesse e de desenvolvimentos recentes. Estas técnicas procuram conciliar a informação de baixo nível semântico extraída das imagens com a informação de alto nível semântico encontrada no texto. Uma hipótese para a descoberta de relações entre texto e imagem é o tratamento na indexação de vários documentos em conjunto, podendo ser utilizadas técnicas estatísticas para a análise de co-ocorrências de características em documentos semelhantes. Assim, um termo que aparece muitas vezes em conjunto com uma determinada característica das imagens pode ser associado a outros documentos que contenham essa característica mas não o termo em causa. A importância de um termo na caracterização de um determinado documento também pode diminuir se este termo não for encontrado em documentos com características semelhantes. Outra hipótese é a análise das hiperligações entre os documentos, como é comum em motores de pesquisa da WWW como, por exemplo, o *Google*.

Outra questão que se coloca para muitas das soluções propostas é a sua adaptabilidade. Na pesquisa de imagens pertencentes a domínios específicos, as características das imagens podem ser previstas e subsequentemente podem ser utilizados modelos exclusivos para o domínio em causa, utilizando essas características. No entanto, para domínios mais alargados são necessários modelos cuja utilização não dependa da existência de características próprias nas imagens, ou seja, cujos conceitos subjacentes se possam generalizar para utilização em pesquisa na maioria das bases de dados de imagens existentes.

3.2 Propostas anteriores

Nesta secção são apresentadas propostas recentes que utilizam características de alto e baixo nível semântico conjuntamente na procura de uma solução para o *semantic gap* e que permitem ao utilizador interagir com o sistema através de interrogações de alto nível.

Os sistemas aqui apresentados podem ser agrupados em duas categorias. Na primeira, as relações entre características de diferentes níveis semânticos são estabelecidas apenas implicitamente, ou seja, as relações semânticas não são exploradas abertamente. Na segunda categoria são estabelecidas relações entre as características de alto e baixo nível semântico numa mesma estrutura de dados, ou seja, é construído explicitamente um modelo dos vários níveis semânticos encontrados na informação indexada.

Dentro da primeira categoria, a técnica LSI foi usada em diversos sistemas. O LSI analisa ocorrências de elementos num conjunto de documentos, agrupando aqueles que co-ocorrem com mais frequência e excluindo os elementos mais comuns. Zhao e Grosky propuseram utilizar o LSI para conciliar texto e características de baixo nível da imagem num sistema único [50]. Até ao surgimento desta proposta o LSI havia sido utilizado apenas para a análise de documentos apenas com texto ou apenas com características de baixo nível, sendo possível a utilização para estes dois tipos de informação em conjunto. Nesta experiência foram utilizados histogramas de cor e anglogramas (ver secção 2.2) como características de baixo nível, tendo sido obtidos resultados marginalmente melhores com a utilização de histogramas de cor e resultados muito superiores com a utilização de anglogramas comparativamente à utilização do LSI apenas com texto e à utilização da comparação vectorial directa. Também van Gemert [16] utilizou o LSI desta forma, utilizando como característica de baixo nível a estrutura da cor da imagem.

Outra proposta que utiliza o LSI, mas na WWW, é a de La Cascia *et al* [8]. O LSI é utilizado aqui apenas no texto, sendo criado um vector para cada documento com os pesos de cada termo. Em seguida são concatenadas ao vector as características de baixo nível. As interrogações do utilizador são feitas inicialmente sob a forma de texto, sendo utilizada em seguida uma técnica de *query relevance feedback* para permitir ao utilizador seleccionar as imagens retornadas mais próximas das desejadas. Após este passo o sistema já consegue utilizar todos os componentes do vector para comparar as imagens seleccionadas com as imagens restantes da base de dados. Como características de baixo nível foram utilizados os histogramas de cor e os histogramas de orientação dominante.

Relativamente à segunda categoria, Colombo *et al* propuseram um sistema [10] que relaciona as emoções transmitidas por uma imagem com as suas características de baixo nível, nomeadamente de cor e de linhas (por exemplo, a calma é associada à cor azul). As interrogações são realizadas através de imagens ou através da indicação de categorias emocionais pré-definidas em que as imagens a retornar se devem inserir. Neste sistema cada imagem é dividida em regiões de cor uniforme, sendo associadas emoções às características de baixo nível semântico extraídas de cada região, de acordo com uma correspondência pré-estabelecida. Foram obtidos resultados interessantes na comparação das imagens retornadas pelo sistema, para várias categorias, com imagens especificadas como relevantes por especialistas.

Outra possibilidade nesta categoria é a utilização de estruturas de conhecimento, como as apresentadas em [3] e [4]. Uma vez que nestas são relacionadas características de baixo nível

extraídas de imagens com termos textuais, é possível permitir ao utilizador fazer interrogações de alto nível sobre estas estruturas.

3.3 Abordagem ao problema

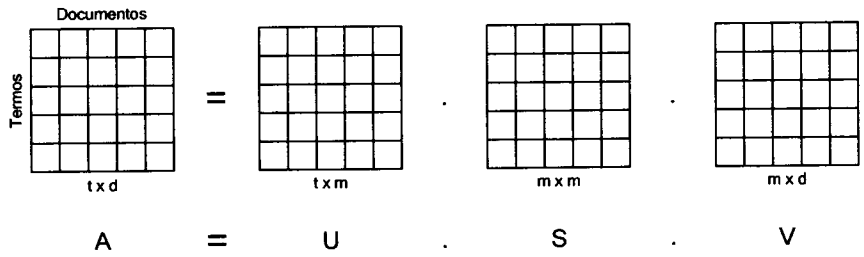
Na proposta que se apresenta para um protótipo de recuperação automática de imagens, procurou-se uma solução que tirasse partido da informação oferecida pelo texto e pela imagem, minimizando a distância semântica entre os dois, e que não fosse dependente nem das características de baixo nível das imagens utilizadas nem da homogeneidade do conjunto destas imagens. Para atingir estes objectivos, escolheu-se basear a solução proposta na utilização do LSI, à semelhança da experiência referida em [50], mas com as características de baixo nível de imagens definidas no módulo XM do MPEG-7 e procurando atingir os seguintes objectivos:

- Comparar os resultados obtidos com a utilização de características de baixo nível com a utilização exclusiva de texto;
- Verificar a prestação do LSI quando este é utilizado com um conjunto de dados de teste desfavorável, como é o caso dos ex-votos utilizados;
- Testar a eficácia de cada uma das características de baixo nível do MPEG-7 na recuperação automática de imagens;

O LSI foi introduzido por Deerwester, Dumais e Harshman [13], baseando-se na ideia de que existe uma estrutura semântica latente num conjunto de documentos relacionados e tendo sido pensado inicialmente apenas para a recuperação de texto. Esta estrutura é encontrada através do agrupamento numa mesma dimensão de termos que co-ocorrem em muitos documentos e da redução da importância de termos que ocorrem na maioria dos documentos e não são portanto importantes para discriminar documentos específicos.

Esta estrutura não pode ser captada por técnicas que apenas comparam os termos da interrogação com os termos existentes em cada documento, já que estas não evitam os problemas relacionados com a sinonímia e polissemia. A sinonímia refere-se à capacidade de várias palavras exprimirem o mesmo significado, o que pode fazer com que não sejam retornados documentos que exprimem os conceitos pretendidos mas que não contêm exactamente os termos pesquisados. A polissemia refere-se ao facto de uma palavra poder ter vários significados conforme o contexto, o que pode dar azo a que num dado sistema sejam retornados ao utilizador documentos que contêm os termos pesquisados mas exprimem conceitos não relevantes para este.

A implementação original do LSI baseia-se numa técnica de álgebra vectorial, a *Singular Value Decomposition* (SVD), que, como o nome indica, decompõe uma matriz inicial em várias matrizes. No nosso caso, a matriz inicial **A** é de termos por documentos, contendo cada posição $\mathbf{T_{ij}}$ na matriz o peso que o termo **i** tem no documento **j**. Os pesos podem ser calculados de várias maneiras, já que o SVD é uma técnica puramente aritmética, não dependendo o seu funcionamento da existência de valores específicos na matriz. O esquema da página seguinte ilustra a decomposição da matriz **A** [13].



Em que:
A é uma matriz de termos por documentos
As colunas de **U** e **V** são ortogonais ($U^TU = V^TV = I$)
S é uma matriz diagonal com os valores singulares de **A**
t é o número de linhas de **A**
d é o número de colunas de **A**
m é menor ou igual ao mínimo entre **t** e **d**.

Figura 1 – *Singular Value Decomposition*

Esta transformação permite obter uma nova aproximação $\mathbf{A_k}$ da matriz **A** inicial. Para esse fim, são mantidos apenas os **k** maiores valores singulares da matriz **S**, ficando estes ordenados por ordem decrescente na matriz $\mathbf{S_k}$. A matriz aproximada $\mathbf{A_k}$ é, garantidamente, a melhor aproximação para cada valor de **k** da matriz original **A**.

O valor de **k** determina a redução de dimensões. Um valor pequeno produz um menor número de dimensões, ou seja, são agrupados mais termos em cada dimensão e são eliminadas as dimensões menos importantes. Um valor grande produz um grande número de dimensões, ficando a matriz $\mathbf{A_k}$ mais próxima da matriz inicial **A**. Não há consenso acerca da obtenção do valor ideal de k, devendo este ser experimentado e ser escolhido aquele que permitir obter os melhores resultados na recuperação. O seguinte esquema ilustra a obtenção de $\mathbf{A_k}$:

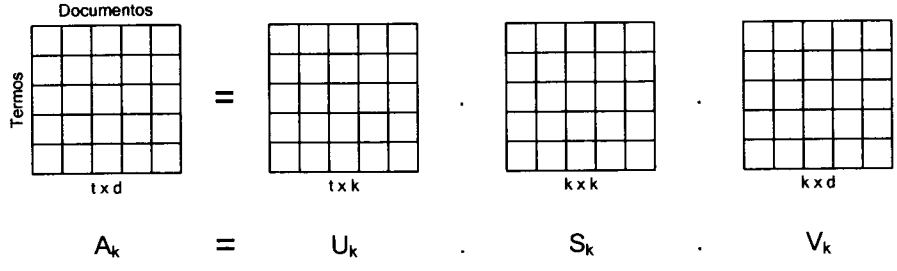


Figura 2 – Obtenção de uma matriz aproximada através da utilização do *Singular Value Decomposition*

Antes da multiplicação das matrizes, o número de colunas de **U** e o número de linhas de **V** são reduzidos para o valor **k**. Depois de obtida a matriz $\mathbf{A_k}$ podem desta ser eliminados valores abaixo de um certo limite que se considere que não vão ter influência na recuperação de documentos. Esta é já a matriz final que será utilizada na recuperação para comparar o vector de pesquisa de um utilizador com os vectores representativos de cada documento.

Para integrar no sistema as características de baixo nível de cada imagem, o peso de cada uma na representação de cada documento é incluído na matriz inicial **A**, passando esta a ser uma matriz de termos e características de baixo nível por documentos. Como já foi referido, para a realização do SVD é indiferente quais são os valores presentes na matriz inicial, podendo ser integrada nesta qualquer informação relativa aos vários documentos que possa ser reduzida a um valor numérico. Uma das vantagens deste sistema é que permite recuperar documentos que contenham apenas imagens e não texto, já que o SVD pode associar características de uma imagem que não tenha legenda a características semelhantes de outras imagens que contenham termos associados. Os termos que co-ocorrem com as características semelhantes às da imagem sem legenda são, por consequência, associados a esta.

A descrição do processo de indexação focou-se até ao momento nas alterações sofridas pela matriz de indexação, mas antes deste processo é necessário proceder ao cálculo dos pesos que compõem a matriz inicial **A**, como pode ser visto no seguinte esquema:

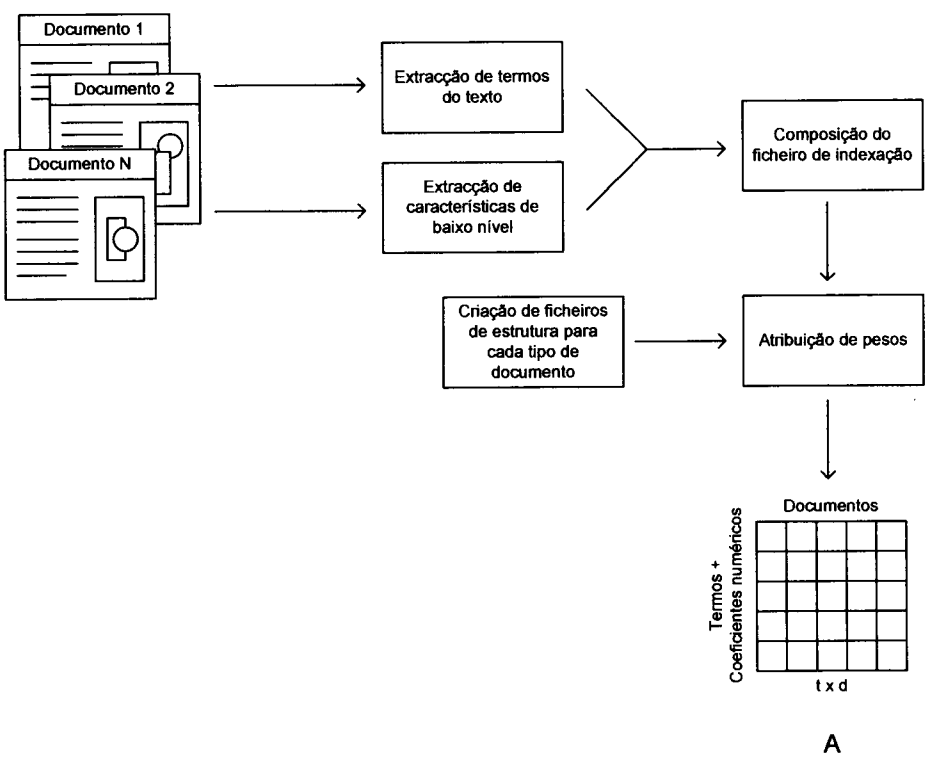


Figura 3 – Processo de criação da matriz de indexação inicial

Inicialmente são extraídos da colecção de documentos escolhidos para indexação os termos do texto e as características de baixo nível das imagens, sendo estes dois tipos de informação integrados no ficheiro de indexação, que reúne os dados necessários à indexação de todos os documentos.

Pretende-se que o protótipo suporte a utilização de diferentes estruturas de documentos em simultâneo, sendo para isso necessário criar manualmente ficheiros de estrutura para descrever cada tipo de documento. Os ficheiros de estrutura contêm as secções de cada documento, podendo cada secção ser textual ou numérica (todas as características de baixo nível das imagens são representadas por secções numéricas). A cada secção é atribuída uma

importância, podendo esta ser utilizada no cálculo de pesos. Desta forma, é possível atribuir importâncias diferentes a secções correspondentes a blocos de texto e secções correspondentes a características de baixo nível das imagens, o que pode ser útil para comparações de desempenho do protótipo.

3.4 O módulo XM do MPEG-7

O módulo XM (eXperimentation Model) do MPEG-7 é a plataforma de testes dos descritores do MPEG-7, sendo o seu objectivo a especificação e implementação de algoritmos de extracção, codificação e descodificação de características de imagem, vídeo e som [43]. A selecção dos descritores do XM foi feita com base em critérios como a eficiência, complexidade e aplicabilidade num grande número de situações diferentes [26], estando prevista a introdução de novos descritores cuja utilização tenha vantagens em relação aos existentes [27].

Os descritores visuais de baixo nível do MPEG-7 foram classificados como sendo relativos à cor ou à textura. Os espaços de cor utilizados no MPEG-7 incluem o monocromático, o RGB, o HSV, o YCrCb e o HMMD, o mais recente. Relativamente à cor foram considerados até ao momento quatro descritores principais [26]:

Scalable Color Descriptor (SCD) – Um histograma de cor, no espaço HSV, quantificado em 256 coeficientes. Pode ser utilizada uma transformada de Haar para obter histogramas de 128, 64 e 32 coeficientes. Quanto menor o número de coeficientes, menor é o espaço ocupado em memória e menor é a informação contida.

Color Structure Descriptor (CSD) – Expressa a estrutura local de cor, contendo portanto informação não só relativa às cores presentes na imagem mas à sua localização relativa. Tal como o SCD, pode ser quantificado entre 256 a 32 coeficientes. Utiliza o espaço de cor HMMD.

Dominant Color Descriptor (DCD) – Define o conjunto das cores dominantes na imagem ou num segmento desta, consistindo nas percentagens das cores mais representativas, na sua coerência espacial e nas variações de cor para cada cor principal. A diferença entre este descritor e o SCD é que aqui as cores são computadas para cada imagem, em vez de serem enquadradas em intervalos fixos, sendo, portanto, esta representação mais exacta e compacta. No entanto, este descritor usa uma medida de semelhança própria, não se prestando assim à comparação vectorial.

Color Layout Descriptor (CLD) – Expressa a distribuição espacial da cor numa imagem ou num segmento cuja forma pode ser especificada, sendo destinado particularmente a aplicações que envolvem segmentação. É um descritor compacto, com um número recomendado de coeficientes de 63, mas é também eficiente na recuperação de imagens. Utiliza o espaço de cor YCrCb.

Relativamente à textura o MPEG-7 oferece três descritores [26]:

Texture Browsing Descriptor (TBD) – Analisa atributos relativos à percepção, como a direcção, a regularidade e a granularidade de uma textura. É compacto, sendo destinado principalmente às aplicações de navegação em bases de dados de imagens. Não se presta à comparação vectorial.

Homogeneous Texture Descriptor (HTD) – Fornece uma perspectiva quantitativa em relação à textura, sendo destinado a ser utilizado na comparação de imagens. É um descritor robusto

em relação a rotações e mudanças de tamanho da imagem. É recomendado que, antes da comparação vectorial, cada valor seja normalizado utilizando o desvio padrão do conjunto de valores da base de dados em que está a ser efectuada a pesquisa.

Edge Histogram Descriptor (EHD) – Descreve a distribuição espacial de arestas, sendo útil na comparação de imagens em que a textura não é homogénea. É composto por 80 coeficientes.

Timo Ojala *et al* avaliaram, relativamente às taxas de precisão e recuperação dos resultados obtidos com a sua utilização, os descritores de cor [33] e de textura [34] do módulo XM. Na avaliação dos descritores de cor nos primeiros 20% dos documentos recuperados, e utilizando o *trade-off* das taxas de precisão e de recuperação como medida, os descritores ficaram assim classificados: 1. *Color Structure*, 2. *Scalable Color*, 3. *Color Layout*, 4. *Dominant Color*.

Na avaliação dos descritores de texturas os mesmos autores obtiveram taxas de recuperação, nos primeiros 20 resultados, entre 23.8% (com quantificação num menor número de coeficientes) e 34.4% (sem quantificação) para o *Edge Histogram Descriptor*; entre 47.4% (sem quantificação) e 47.0% (com quantificação) para o *Homogeneous Texture Descriptor*.

4 O protótipo

Descrevem-se agora os pormenores da implementação de um protótipo de sistema de recuperação de imagens de acordo com a proposta apresentada no Capítulo anterior.

A linguagem escolhida para a sua programação foi Java, procurando-se tirar partido da possibilidade da utilização do programa concebido em várias plataformas e da existência de diversas bibliotecas para uso público de grande utilidade e qualidade reconhecidas. A linguagem Java pertence à família das linguagens de programação orientadas por objectos.

4.1 Descrição geral

O diagrama apresentado a seguir representa o funcionamento do sistema de recuperação de imagens proposto (adaptado do diagrama de sistemas de recuperação de texto de [2]):

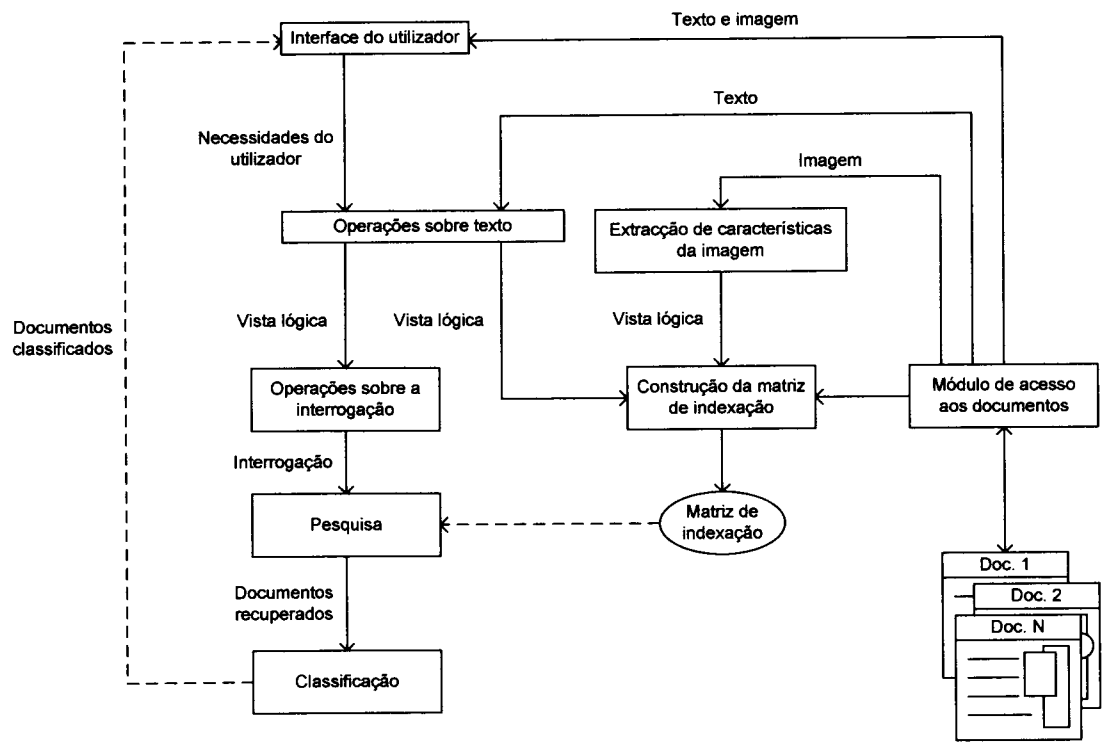


Figura 4 – Diagrama do funcionamento geral do protótipo

A recuperação de informação baseia-se numa matriz de indexação que tem de ser criada antes do sistema poder ser utilizado. A matriz fornece meta-informação dos documentos a indexar, indicando o peso dos vários termos nela contidos na representação de cada documento. As várias partes do sistema interagem com os documentos através de um módulo de acesso (ver secção 4.2. – Processamento dos documentos).

De forma a serem eliminadas diferentes grafias dos vários termos encontrados na base de dados, estes são uniformizados antes de serem indexados (ver secção 4.2. – Uniformização de termos). A extracção de características de baixo nível das imagens é também efectuada antes

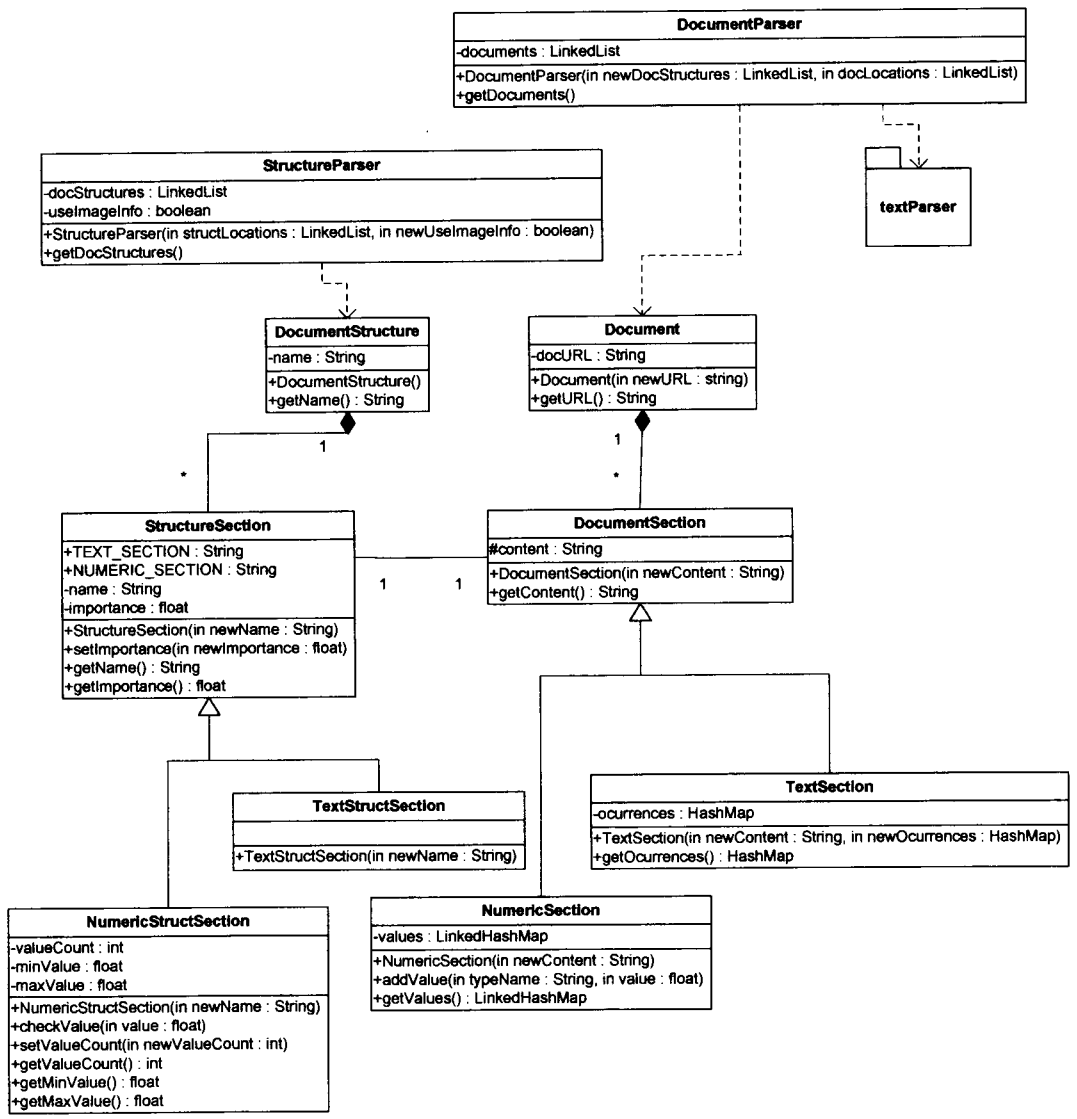


Figura 6 – Diagrama de classes do pacote "documentHandler"

Estas estruturas de dados visam conservar os documentos e respectivas estruturas, tendo em vista a utilização dos dados nelas contidos para a construção da matriz de indexação e também na apresentação de resultados ao utilizador, já que a matriz de indexação não conserva todos os aspectos de cada documento.

Cada documento é representado por uma instância da classe "Document" e contém uma ou mais secções, representadas por instâncias da classe "DocumentSection". Cada secção pode ser de tipo numérico ("NumericSection") ou de tipo textual ("TextSection"). A cada documento corresponde uma estrutura, representada por uma instância da classe "DocumentStructure". Da mesma forma, cada secção de um documento tem uma estrutura ("StructureSection"), e é a este nível que os documentos e as suas estruturas se encontram unidos nas estruturas de dados conservadas. A estrutura de um documento determina o número de secções que este irá conter e o tipo e importância relativa de cada secção. Cada tipo de secção de um documento (numérica ou textual) tem uma estrutura própria.

Os ficheiros de estruturas são lidos primeiramente, através do construtor da classe "StructureParser", que lê os ficheiros de estruturas, conservados em formato XML. Estes

ficheiros são construídos manualmente para os vários tipos de documento a utilizar. A seguinte figura é um exemplo do conteúdo de um ficheiro de estrutura:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<DocumentStructure Name = "ExVoto">
  <Description Type = "Text">1.0</Description>
  <ColorLayout Type = "Numeric" ValueCount = "12">0.5</ColorLayout>
</DocumentStructure>
```

Figura 7 – Ficheiro de estrutura

Os ficheiros de documentos são lidos de um ficheiro em formato XML, através do construtor da classe "DocumentParser". Este construtor recebe uma lista das estruturas de documentos processadas previamente, já que necessita de ter acesso a esta lista para relacionar cada secção de um documento com a secção de estrutura correspondente.

Um exemplo do conteúdo de um ficheiro de documentos pode ser visto na figura abaixo. Neste exemplo podem ser observados dois documentos, contendo, cada um, duas secções, "Description" e "ColorLayout", sendo a primeira de tipo textual e a segunda de tipo numérico. Cada documento tem obrigatoriamente o URL da imagem a que corresponde.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<Documents>
  <Document Type = "ExVoto" URL = "21.png">
    <Description>Milagre que fez Nosso Senhor da Fraga a Ludovina de Jesus, do
      lugar de Vide, que estando doente e sem esperança de vida chamou-se ao
      divino Senhor e logo melhorou. No ano de 1869.</Description>
    <ColorLayout>30 26 35 15 14 19 15 16 20 18 13 13</ColorLayout>
  </Document>
  <Document Type = "ExVoto" URL = "24.png">
    <Description>Milagre que fez Nosso Senhor dos Passos a Josefina Joaquina,
      da Vila da Ponte. Estando seu marido em perigo de vida, recorreu ao
      divino Senhor logo teve melhoras. Ano de 1894.</Description>
    <ColorLayout>36 12 47 17 20 16 18 14 15 11 17 20</ColorLayout>
  </Document>
</Documents>
```

Figura 8 – Ficheiro de documentos

Os ficheiros apresentados nos exemplos acima, assim como os utilizados na realização de experiências sobre o protótipo, foram criados manualmente, tendo sido incorporadas as legendas digitalizadas do livro "Do gesto à memória – Ex-votos" [12] e as características de baixo nível das imagens que foram digitalizadas a partir do mesmo livro. Estas características foram extraídas através da utilização do *software* XM [S1].

A cada secção é atribuída uma importância, que pode variar entre 1.0 e 0.0 e é utilizada mais tarde no processo de indexação para calcular os pesos dos elementos de cada secção. Neste exemplo, foi atribuída uma importância de 1.0 à secção textual "Description" e uma importância de 0.5 à secção numérica "ColorLayout" (ver secção 3.3).

Podem ser usadas diferentes estruturas e diferentes tipos de documentos correspondentes no mesmo processo de indexação, sendo para isso especificados vários ficheiros de estrutura. Cada ficheiro de documentos pode conter documentos com estruturas diferentes, sendo a relação feita entre o nome da estrutura e o tipo do documento (os documentos da fig. 8

contêm a estrutura da fig. 7, podendo-se observar que a propriedade "Name" no ficheiro de estrutura e a propriedade "Type" nos dois documentos têm o mesmo valor de "ExVoto").

Em cada secção de estrutura é conservado o nome e importância. Nas secções de estrutura numéricas é conservado também o número de coeficientes numéricos da secção e os valores mínimo e máximo de cada coeficiente para todos os documentos. Estes dados serão utilizados na indexação para o cálculo dos pesos dos coeficientes numéricos.

Nas secções dos documentos é conservado o conteúdo da secção no formato de texto, sendo que nas secções numéricas é também conservada a lista de valores numéricos pertencentes a cada secção. Nas secções textuais são conservados os termos encontrados no texto da secção; durante a leitura dos ficheiros de documentos é utilizada a função "applyRules" do pacote "textParser" (ver secção "Uniformização de termos") para uniformizar cada um destes termos.

Indexação

O processo de indexação é conduzido pelos objectos contidos no pacote "indexer", cuja estrutura se encontra detalhada na seguinte figura:

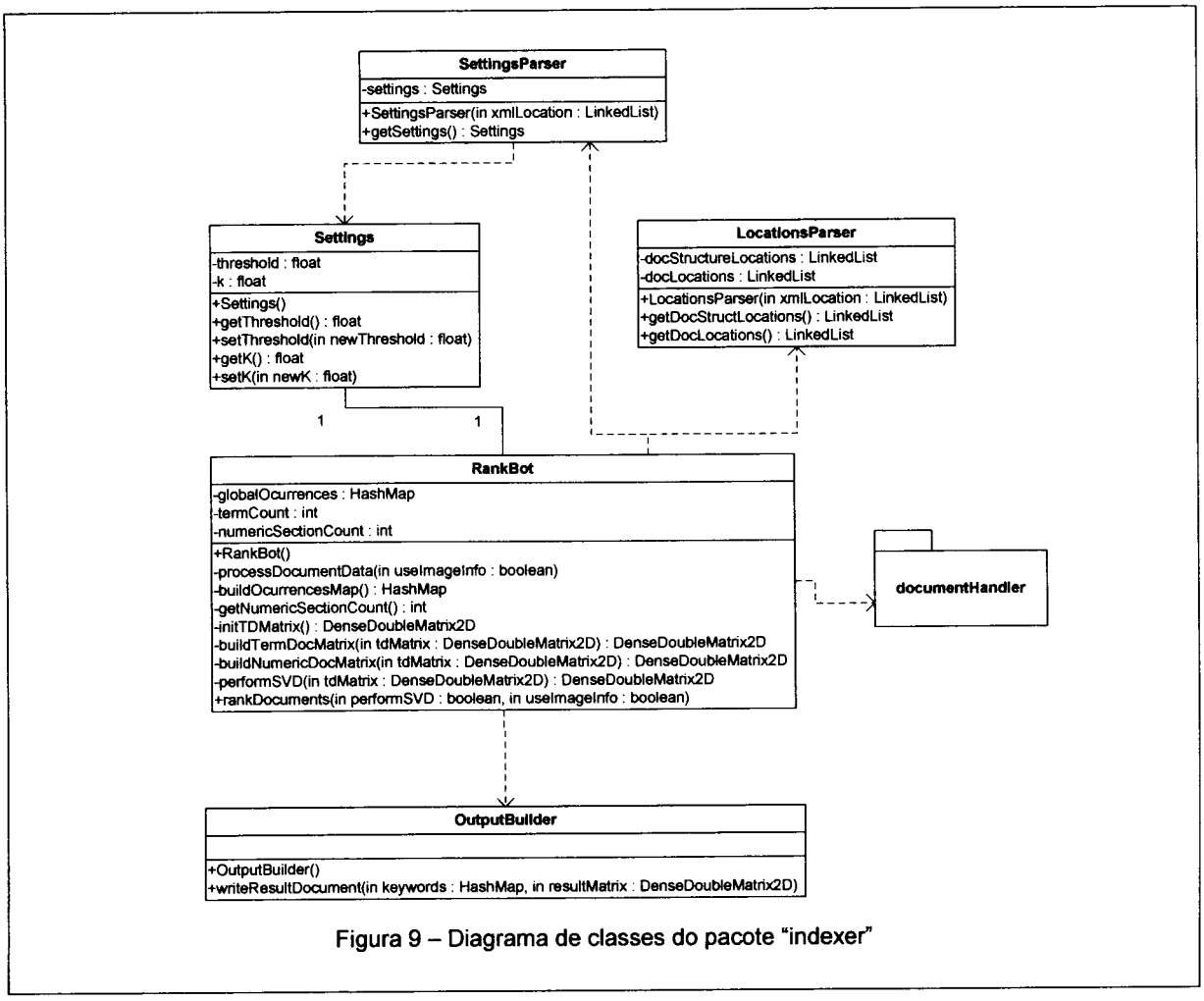


Figura 9 – Diagrama de classes do pacote "indexer"

A classe principal é a "RankBot", sendo criada uma instância desta para ser utilizada ao longo da indexação. Inicialmente é lido o conteúdo do ficheiro de parâmetros do sistema para a classe "Settings", através do construtor da classe "SettingsParser". Os dois parâmetros contidos neste ficheiro são "**k**", que indica a percentagem de dimensões retidas após aplicação

do SVD, e "threshold", cuja utilização visa reduzir os valores insignificantes da matriz de indexação final através da eliminação dos pesos inferiores ao valor numérico especificado. Em seguida é utilizado o construtor da classe "LocationsParser" para ler o ficheiro que contém as localizações dos vários ficheiros de documentos e de estruturas de documentos. Nesta altura são empregues instâncias das classes "StructureParser" e "DocumentParser" para traduzir para estruturas de dados os ficheiros de estruturas de documentos e os documentos em si, respectivamente. Após este passo, os dados contidos nestas estruturas são passados para a matriz de indexação. A seguinte figura representa as transformações pelas quais esta matriz passa:

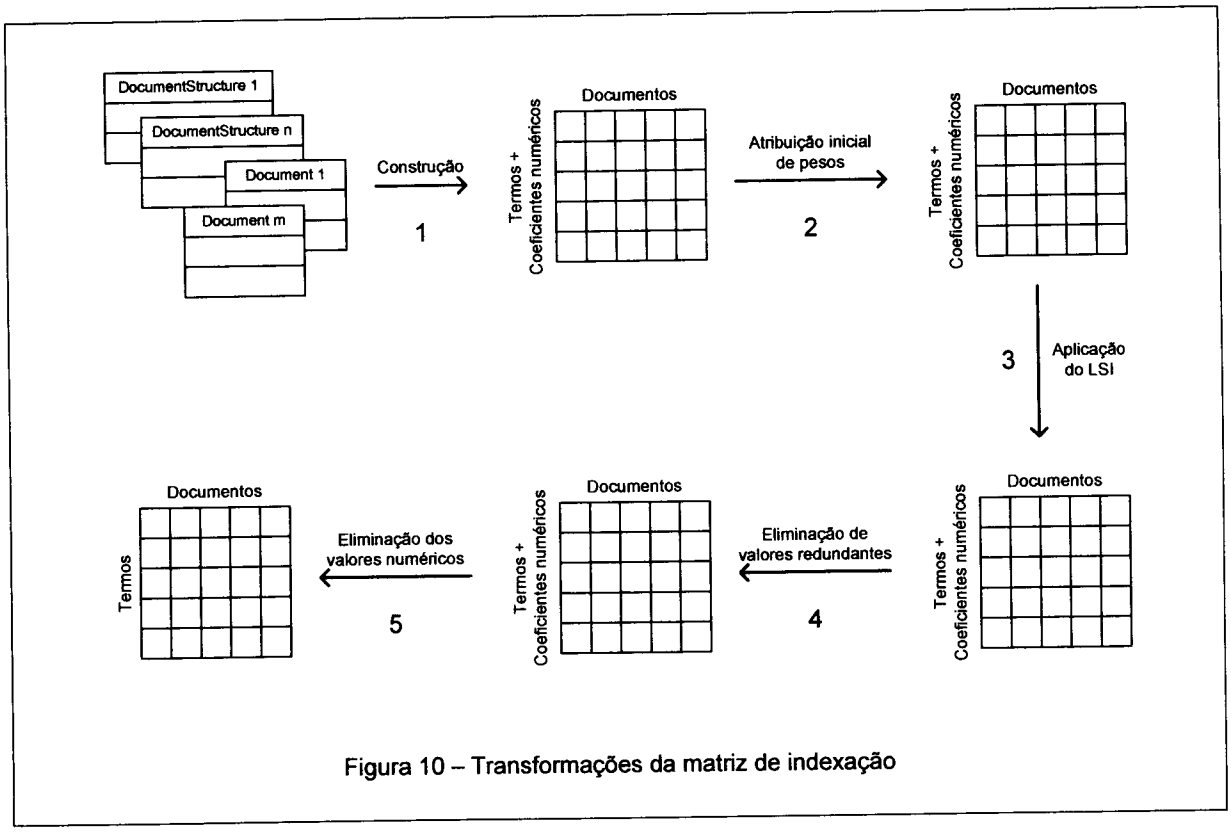


Figura 10 – Transformações da matriz de indexação

Passo 1 – a matriz é criada sem ser preenchida, sendo definido o seu número de linhas e colunas: o número de colunas corresponde ao número total de documentos, enquanto que o número de linhas corresponde à soma do número de termos diferentes encontrados nas secções textuais de todos os documentos com a soma do número de coeficientes de cada secção numérica. Para o exemplo das figs. 7 e 8, existem 42 termos diferentes e apenas uma secção numérica com 12 coeficientes, logo a matriz teria 54 linhas. Na classe "RankBot" este passo é realizado pela função "initTDMatrix".

Passo 2 – São percorridas as estruturas de dados que conservam os documentos e as estruturas respectivas, sendo atribuídos pesos que são calculados de forma diferente para os termos do texto e os coeficientes numéricos. A fórmula de cálculo do peso de um termo para um documento é:

$$p_{tdoc} = \frac{\sum_{s=1}^{n} (o_{ts} i_s)}{\sum_{d=1}^{k} o_{td}}$$

em que o_{ts} é o número de ocorrências do termo em cada secção, i_s a importância dessa secção, n o número de secções do documento, o_{td} o número total de ocorrências do termo em cada documento e k o número total de documentos. O valor mínimo deste peso é 0.0 e o máximo 1.0. Esta fórmula é uma adaptação da fórmula básica de cálculo de pesos, que divide as ocorrências de um termo num determinado documento pela soma das ocorrências desse termo em todos os documentos. A importância da secção em que o termo foi encontrado nesta fórmula visa introduzir a possibilidade de serem distinguidas secções num documento relativamente à relevância do seu conteúdo. Na classe "RankBot" estes pesos são atribuídos pela função "buildTermDocMatrix".

A fórmula de cálculo do peso de um coeficiente numérico para um documento é:

$$p_{cdoc} = \left(\frac{v_{cd} - \min_c}{\max_c - \min_c}\right) i_s$$

Em que v_{cd} é o valor do coeficiente para o documento em causa, min_c e max_c são o mínimo e máximo do coeficiente em todos os documentos e i_s é a importância da secção numérica em que o coeficiente se encontra. Os mínimos e máximos do coeficiente são utilizados para que, tal como na obtenção do peso de um termo, o valor mínimo seja 0.0 e o máximo 1.0. Desta forma termos textuais e coeficientes numéricos são comparados equitativamente. Também nesta fórmula a importância da secção é introduzida para ser possível indicar a sua influência em relação às restantes secções do documento. Na classe "RankBot" estes pesos são atribuídos pela função "buildNumericDocMatrix".

Passo 3 – O SVD é realizado sobre a matriz, sendo utilizado o parâmetro "**k**" referido como medida do número de dimensões retidas. Os pesos dos coeficientes numéricos afectam nesta altura os pesos dos termos, de acordo com as co-ocorrências observadas entre os dois tipos de dados.

Passo 4 – São eliminados da matriz os valores inferiores ao parâmetro "threshold" referido anteriormente, para que não seja retida informação desnecessária.

Passo 5 – São eliminadas da matriz as linhas correspondentes aos coeficientes numéricos, uma vez que os pesos destes já exerceram efeito sobre os pesos dos termos e a pesquisa é feita apenas através dos termos.

Os três passos anteriores são realizados na classe "RankBot" pela função "performSVD".

É importante referir que o protótipo pode também construir uma matriz final sem utilizar o LSI ou utilizando o LSI apenas com texto. No primeiro caso, não são realizados os passos 3, 4 e 5. No segundo não são incluídos na matriz inicial os coeficientes numéricos das características de baixo nível. A função "rankDocuments" da classe "RankBot" recebe dois argumentos booleanos que permitem especificar qual a variante do método de indexação a realizar. Esta função invoca todas as funções necessárias para a realização dos vários passos referidos.

Uma vez obtida a matriz final, a classe "RankBot" invoca a classe "OutputBuilder", que constrói uma representação da matriz em formato XML.

Pesquisa

As pesquisas são realizadas através do pacote "searchVector", cujas classes se encontram explanadas na seguinte figura:

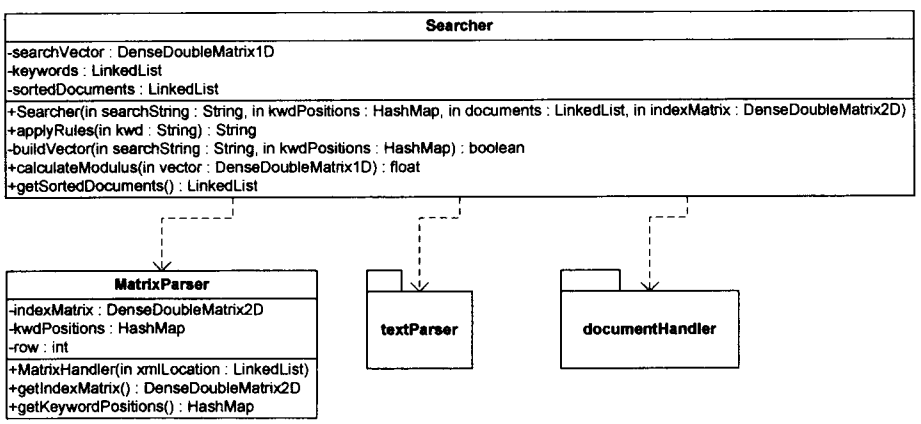


Figura 11 – Diagrama de classes do pacote "searchVector"

A classe principal é a "Searcher", cuja função essencial é comparar o vector da frase de pesquisa fornecida pelo utilizador com os vectores dos documentos indexados. Inicialmente, é necessário utilizar a classe "MatrixParser" para ler o ficheiro da matriz de indexação gerado pelo pacote "indexer" e fornecer à classe "Searcher" esta matriz. O ficheiro contém também a lista dos termos correspondentes a cada linha. A classe principal depende também do pacote "documentHandler" para lhe fornecer a lista de documentos existentes, já que após a comparação de vectores constrói uma lista de documentos ordenados pela relevância para a pesquisa em causa. Após receber a frase de pesquisa, a classe "Searcher" utiliza a classe "textParser" para aplicar aos termos contidos na frase as mesmas regras utilizadas pelo pacote "documentHandler" na leitura dos termos existentes em cada documento.

Na comparação vectorial, inicialmente é construído o vector de pesquisa, em que a cada termo existente na frase de pesquisa é atribuída uma importância de 1.0. Os termos que não existam nos documentos indexados são excluídos automaticamente do vector de pesquisa. Em seguida, este vector é comparado com os vectores dos documentos. Um vector de um documento é constituído pelos pesos dos termos para esse documento presentes na matriz de indexação. A comparação entre vectores é feita através do coseno do ângulo entre dois vectores, sendo um documento mais relevante quanto menor for este valor. Após a

comparação com todos os vectores dos documentos são retornados os documentos ordenados pela relevância.

Uniformização de termos

Como foi referido nas secções anteriores, os termos encontrados nos documentos a indexar e nas frases de pesquisa introduzidas pelo utilizador são uniformizados pelo pacote "textParser", que é composto por uma única classe:

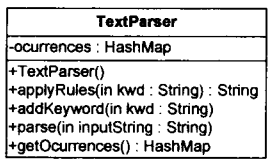


Figura 12 – Diagrama de classes do pacote "textParser"

A classe "TextParser" processa um texto, separando os termos nele contidos por espaços e sinais de pontuação (ponto de exclamação, parêntesis, vírgula, etc.). Após a obtenção de uma lista de termos, cada termo é sujeito às seguintes intervenções:

1. Todas as letras são convertidas para maiúsculas;
2. As siglas são unidas (por exemplo, a sigla O.N.U. seria convertida para ONU).

Não foram eliminadas palavras menos significantes, como artigos definidos e conjunções, nem foi realizado *stemming* sobre cada palavra, já que o LSI retira importância a palavras que ocorrem com grande frequência e os tempos verbais nas descrições utilizadas são muito constantes.

4.3 *Software* utilizado

Foram utilizadas várias bibliotecas de código externas ao projecto. A leitura da informação dos ficheiros de documentos, de estruturas de documentos e da matriz de indexação foi efectuada através da *Simple API for XML* (SAX) [S3], que disponibiliza *parsers* de fácil utilização, orientados por eventos, para a leitura de ficheiros XML. A escrita do ficheiro que contém a matriz resultante do processo de indexação foi realizada através de *parsers Document Object Model* (DOM) [S4], que permitem a construção de uma estrutura de objectos que pode depois ser transposta automaticamente para um ficheiro XML. Os dois tipos de *parsers* foram utilizados através das implementações disponibilizadas no SDK Java 2, Standard Edition (J2EE) [S5].

A realização do SVD sobre a matriz de termos por documentos foi efectuada através da biblioteca *Open Source* Colt [S2], destinada a fornecer uma infra-estrutura para computação nas áreas de análise de dados, álgebra linear, *arrays* multi-dimensionais e outras aplicações. Foram utilizadas as classes que permitem fazer o SVD (*SingularValueDecomposition*), as classes que representam matrizes (*DenseDoubleMatrix2D*) para conservação das matrizes nos vários passos da indexação e as classes que representam vectores (*DenseDoubleMatrix1D*) para a comparação dos vectores de pesquisa com os vectores representativos de cada documento.

A extracção das características de baixo nível utilizadas no protótipo foi realizada com o *software* XM [S1], tendo este sido concebido de acordo com as especificações do MPEG-7. Este *software* recebe um conjunto de ficheiros de imagens e fabrica um ficheiro XML com as características de baixo nível, sendo esta informação facilmente integrada nos ficheiros dos documentos.

Para a criação e modificação de código foi utilizada a IDE *Eclipse* [S6], uma ferramenta *Open Source*, multiplataforma e multilingue.

4.4 Caracterização dos dados utilizados

Os ex-votos da colecção de teste são quadros que foram reunidos em livro ("Do gesto à memória – Ex-votos"[12]) pelos museus da Guarda, de Grão Vasco (Viseu) e de Lamego, encontrando-se originalmente em santuários da região delimitada a norte pelo Douro e a sul pelo curso médio do Mondego, a oeste pelas cristas do Caramulo e da Gralheira (que dividem a Beira Alta do litoral) e a este pela fronteira com Espanha. Estas delimitações enquadram as dioceses de Guarda, Viseu e Lamego, no contexto das quais os ex-votos devem ser compreendidos.

Foram escolhidos 89 quadros, tendo sido utilizada como critério de selecção a distinção entre os elementos que compõem a peça, para que a utilização das características da imagem seja o mais útil possível. Os quadros escolhidos cobrem um largo período temporal: o mais antigo data de 1635, enquanto que o mais recente é do século passado (1917).

Na esmagadora maioria dos casos, os ex-votos escolhidos são constituídos por dois elementos: a representação do suposto milagre e a representação da entidade à qual o agradecimento é destinado. Estes elementos encontram-se distintamente separados na imagem, com a entidade protectora a fazer-se acompanhar por uma nuvem que sublinha a sua divindade. As cenas representadas dividem-se em exteriores, onde normalmente são representados acidentes, e interiores, em que, na maior parte dos casos, é representada uma enfermidade: a pessoa, ainda doente, aparece deitada numa cama, habitualmente acompanhada por familiares [12].

As legendas que acompanham as imagens são geralmente breves. São enunciados o beneficiado pela intervenção divina e a entidade à qual é atribuída a responsabilidade da intervenção. Estes destinatários dos agradecimentos aparecem com várias denominações. Cristo é referido como Senhor da Aflição, dos Aflitos, do Calvário, etc. Nossa Senhora aparece como Nossa Senhora das Dores, dos Remédios, da Saúde, etc. Também é referida por vezes pelo nome de uma localidade à qual tenha ficado associada, por exemplo, Nossa Senhora da Lapa ou Nossa Senhora do Viso. Os santos são chamados apenas pelo seu nome, por exemplo, Santa Quitéria ou S. Brás. São por vezes também recordadas outras testemunhas ou intervenientes, e até familiares, já que existe uma crença implícita de que estes também serão, de alguma forma, agraciados pela divindade. Fernando Paulo Baptista caracteriza o tipo da escrita nestas legendas como "«popular» que não decorre de uma concepção universal de escola nem de uma acção pedagógica globalmente concertada, planeada e generalizada, mas que deflui tão somente da monádica e singular competência de comunicação (...) oral-escrita (...) de cada um dos artesãos-narradores [12]". A brevidade das legendas pode introduzir algumas dificuldades ao LSI, já que é difícil para este analisar co-ocorrências com pouco texto disponível.

Apesar da incoerência estilística na pintura dos ex-votos, há símbolos que são associados às várias denominações de cada entidade divina, sendo estes símbolos coerentes na maioria das representações da entidade. Por exemplo, a Senhora das Necessidades aparece carregando um rosário que segura com as duas mãos, enquanto que Santa Eufémia segura numa das mãos um ramo de uma planta. Estas semelhanças podem ser importantes para a detecção de estruturas de cor ou formas. À falta de consistência no estilo corresponde uma dispersão da informação no ex-voto, não se encontrando os elementos que o compõem localizados relativamente uns aos outros de forma previsível.

Os dados utilizados no protótipo são as características de baixo nível extraídas dos quadros digitalizados e os textos das traduções das legendas que os acompanham. Estas traduções foram retiradas de "Do Gesto à Memória – Ex-votos" [12], tendo sido desenvolvidas as abreviaturas e actualizada a ortografia das legendas originais.

5 Resultados

Nesta secção são apresentados os testes realizados sobre o protótipo delineado no capítulo anterior e os resultados destes.

5.1 Configuração e parâmetros do sistema

Os resultados aqui demonstrados foram obtidos através da variação de três parâmetros essenciais ao sistema:

- Percentagem de dimensões retidas na matriz após o processamento do LSI ($\mathbf{k_p}$);
- Número de coeficientes de cada característica de baixo nível do MPEG-7;
- Importância relativa da informação textual e da informação de baixo nível proveniente das imagens.

Este último parâmetro representa a possibilidade de, na aplicação do LSI, fazer com que os termos ou coeficientes numéricos sejam considerados num grau variável. Nos testes que se seguem a importância relativa destes foi variada, sendo esta quantificada em percentagem. É de salientar que as percentagens consideradas para descrições textuais e características de baixo nível podem ser ajustadas independentemente, como se verá a seguir.

De forma a avaliar os resultados das alterações nestes parâmetros foram calculadas as médias de recuperação e precisão, nos primeiros vinte resultados, das catorze interrogações presentes na seguinte tabela:

Nosso Senhor
Santa Eufémia
Jesus
1706
enfermo
malina
moléstia
vaca
família
caiu
Senhor da Livração
Padre
filhinhos
Virgem

Tabela 1 – Interrogações com apenas um conceito utilizadas na avaliação de resultados

As imagens relevantes para cada interrogação foram seleccionadas através da análise dos elementos que as compõem e das descrições associadas, não tendo intervindo especialistas no processo.

5.2 Resultados globais

Na comparação vectorial directa de termos, em que nenhum dos parâmetros referidos é relevante, obtiveram-se médias de recuperação e precisão de 38.4% e 92.5%, respectivamente. A grande taxa de precisão apresentada por este método deve-se ao facto de, por definição, apenas serem retornados documentos em que as palavras contidas na

interrogação apareçam. No entanto, esta taxa não é de 100% porque nalguns documentos as palavras especificadas são referidas com significados diferentes dos pretendidos.

Na comparação vectorial, após utilização do LSI com as descrições textuais, apenas $\mathbf{k_p}$ introduz variações nos resultados, patentes na seguinte figura:

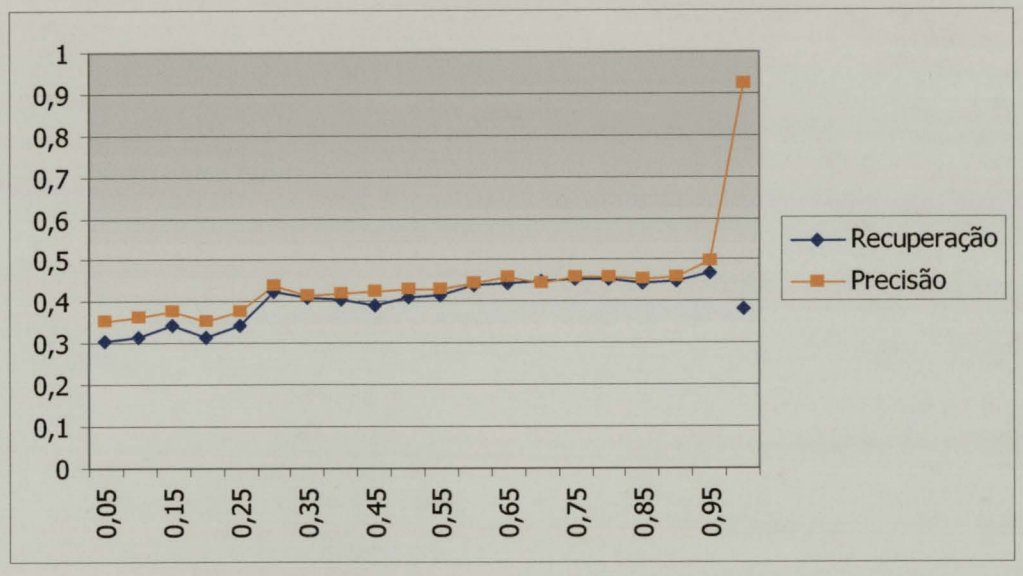


Figura 13 – Variação das taxas de recuperação e precisão com $\mathbf{k_p}$

A média das taxas de recuperação e de precisão para os vários valores de $\mathbf{k_p}$ é, respectivamente, de 40.5% e 42.4%. Pode-se observar que a taxa de recuperação aumenta com o número de dimensões retidas, registando-se um valor máximo de 47.0% quando $\mathbf{k_p}$ é de 0.95, muito superior ao obtido com a comparação vectorial simples. A taxa de precisão segue uma curva semelhante, atingindo um valor de 49.9% para o mesmo valor de $\mathbf{k_p}$, ou seja, conclui-se que, utilizando apenas as descrições textuais dos ex-votos, se obtêm os melhores resultados quando o LSI diminui apenas ligeiramente as dimensões da matriz inicial.

Quando $\mathbf{k_p}$ é igual a 1.0, a matriz inicial mantém-se inalterada e as taxas de recuperação e precisão são idênticas às obtidas com a comparação vectorial simples.

É importante notar que, utilizando o LSI, nunca são retornados menos de vinte documentos ao utilizador, ou seja, a taxa de precisão é sempre calculada com um divisor de 20. Como já foi referido, a comparação vectorial simples retorna apenas os documentos em que os termos pesquisados se encontram directamente referidos, ou seja, a taxa de precisão é quase sempre calculada com divisores inferiores a 20 e é por isso muito superior às taxas registadas com a utilização do LSI. Caso fosse sempre utilizado um divisor de 20 no cálculo da taxa de precisão esta seria de 36.1% para a comparação vectorial simples.

Utilizando o LSI com as descrições textuais e todas as características de baixo nível, todos os parâmetros referidos previamente produzem variações nos resultados. Inicialmente escolheu-se variar $\mathbf{k_p}$ e as importâncias relativas do texto, tendo as características de baixo nível da imagem ficado com um número de coeficientes fixos:

- *Scalable Color Descriptor* (SCD) – 64 coeficientes.
- *Color Structure Descriptor* (CSD) – 64 coeficientes.
- *Color Layout Descriptor* (CLD) – 12 coeficientes.
- *Homogeneous Texture Descriptor* (HTD) – 62 coeficientes.
- *Edge Histogram Descriptor* (EHD) – 80 coeficientes.

Procurou-se que as várias características fossem compostas por um número de coeficientes o mais próximo possível, de forma que todas pudessem exercer uma influência semelhante nos resultados finais. No entanto, o EHD encontra-se limitado por definição a 80 coeficientes e o CLD estava limitado no *software* utilizado na extracção a 12 coeficientes.

Na realização dos testes, $\mathbf{k_p}$ fez-se variar entre 0.05 e 0.95, e em conjunto com cada um destes valores foram testadas oito hipóteses diferentes para as importâncias relativas do texto e das características de baixo nível. Na primeira hipótese foi atribuída a mesma importância às descrições textuais e às características de baixo nível. Na segunda, terceira e quarta hipóteses foi diminuída a importância das características de baixo nível, enquanto que nas quatro últimas hipóteses foi reduzida a importância das descrições textuais. Obtiveram-se as seguintes médias para as taxas de recuperação e precisão:

Importância das descrições textuais	**Importância das características de baixo nível da imagem**	**Média de recuperação**	**Média de precisão**
100%	100%	44.1%	44.1%
100%	75%	44.0%	43.9%
100%	50%	43.5%	44.2%
100%	25%	42.5%	43.4%
75%	100%	44.5%	43.9%
50%	100%	44.9%	43.9%
25%	100%	45.9%	44.0%
10%	100%	46.2%	44.0%

Tabela 2 – Variação das médias de recuperação e precisão com as importâncias relativas do texto e da imagem

Como se pode observar na tabela, obtém-se a melhor média da taxa de recuperação com importâncias relativas de 10/100 (10% de importância para o texto e 100% de importância para as características de baixo nível), enquanto que a taxa de precisão neste caso é praticamente idêntica à melhor obtida, o que demonstra que para os dados de teste utilizados, as características de baixo nível têm maior capacidade de discriminação quando lhes é dada mais influência que o texto.

Os melhores resultados individuais verificam-se para um $\mathbf{k_p}$ de 0.85 e importâncias relativas de 25/100 – taxa de recuperação de 50,9% e taxa de precisão de 47.5% – e para um $\mathbf{k_p}$ de 0.9 e importâncias relativas de 10/100 – taxa de recuperação de 51,1% e taxa de precisão de 48.2%. Neste último caso, a precisão registada é apenas ligeiramente inferior à maior obtida com a utilização de características de baixo nível – 48.9%.

Pode-se assim comprovar que através da utilização das características de baixo nível em conjunção com o texto foi recuperado um maior número de documentos relevantes do que

com a utilização de apenas texto – 51,1% contra 47.0%. Relativamente à taxa de precisão para este valor de recuperação, ela é inferior em apenas 1.7% à máxima registada com a utilização do LSI exclusivamente com texto.

O gráfico seguinte permite comparar, para a gama de valores de $\mathbf{k_p}$, a taxa de recuperação observada utilizando o LSI apenas nas descrições textuais com a mesma taxa quando são também utilizadas as características de baixo nível, tendo esta segunda taxa sido calculada fazendo a média dos resultados para as diferentes importâncias testadas para o texto e as características de baixo nível, não reflectindo portanto os máximos obtidos para as várias combinações testadas dos valores de $\mathbf{k_p}$ e das importâncias referidas.

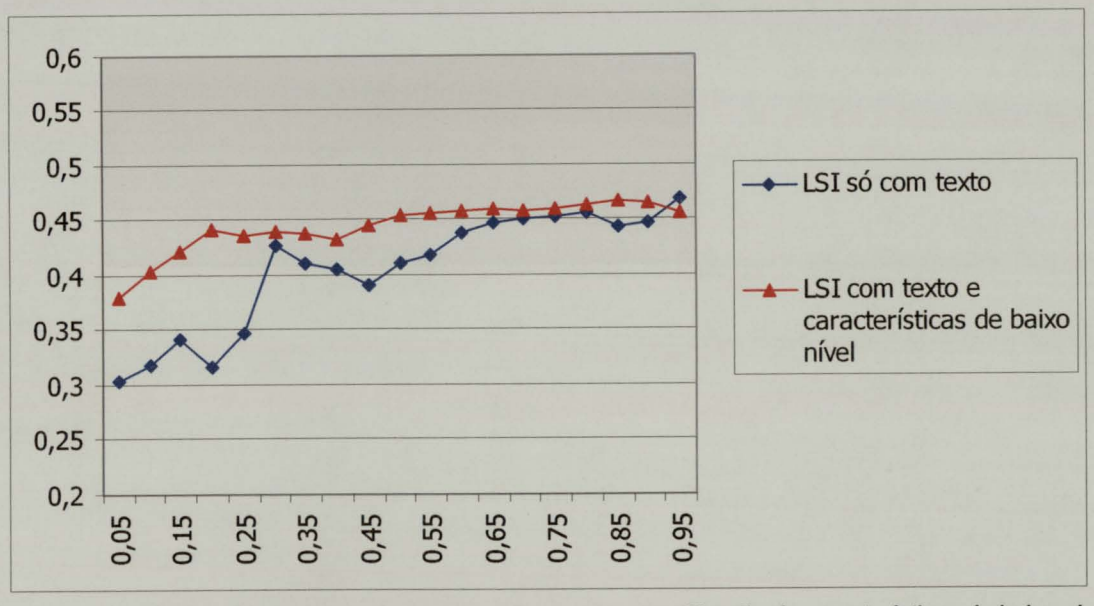


Figura 14 – Comparação das taxas de recuperação com e sem utilização de características de baixo nível semântico para diferentes valores de $\mathbf{k_p}$

Pode-se observar que os resultados obtidos utilizando as características das imagens são consistentemente melhores do que os observados com a utilização de apenas texto, já que só quando $\mathbf{k_p}$ é 0.95 o LSI apenas com texto regista melhor taxa de recuperação – o máximo referido de 47.0%. É também visível que para valores baixos de $\mathbf{k_p}$, ou seja, quando a informação é reunida num menor número de dimensões, a utilização do LSI com texto e características de baixo nível apresenta resultados muito superiores aos obtidos com a utilização de texto em exclusivo.

O gráfico abaixo permite efectuar a mesma comparação para a taxa de precisão:

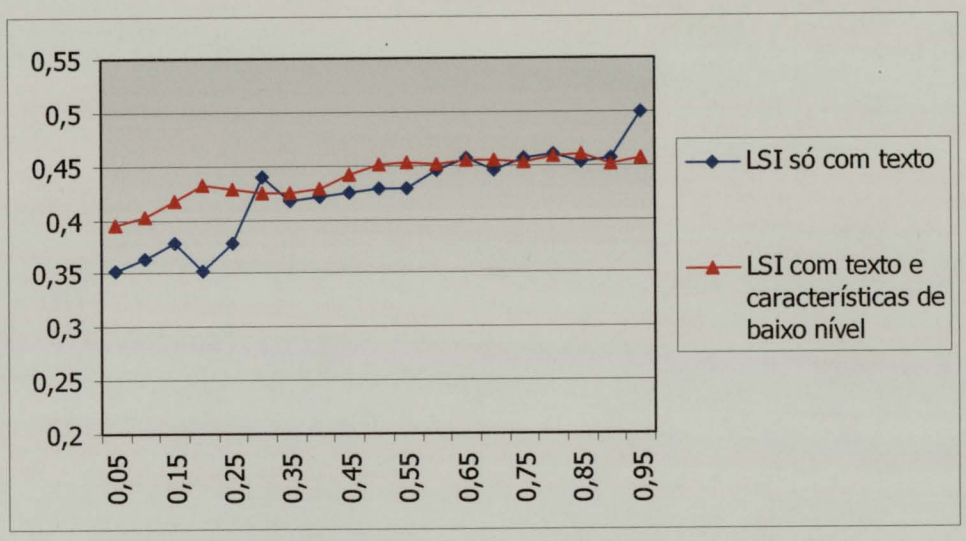


Figura 15 – Comparação das taxas de precisão com e sem utilização de características de baixo nível semântico para diferentes valores de k_p

Relativamente à taxa de precisão pode-se notar que os valores registados para as duas experiências em questão são mais próximos entre si do que os valores observados para a taxa de recuperação. No entanto, os valores registados com a utilização das características de baixo nível são mais homogéneos para os vários valores de k_p e, tal como acontece para a taxa de recuperação, são bastante superiores aos valores observados para o LSI apenas com texto quando o valor de k_p é baixo. É também visível que, para um valor de k_p de 0.95, a utilização do LSI apenas com texto apresenta um valor muito superior à média dos valores obtidos com a utilização do LSI com características de baixo nível, sendo este valor também o máximo referido de 49.9%.

Até ao momento apenas foram apresentados resultados globais das taxas de recuperação e precisão para os primeiros vinte documentos retornados pelo protótipo, não tendo sido possível estimar a consistência dos resultados à medida que são retornados mais documentos. Para este fim serão utilizadas curvas de precisão/recuperação, que permitem avaliar a evolução da taxa de precisão à medida que a taxa de recuperação aumenta. No gráfico da página seguinte encontram-se comparadas estas curvas para a utilização de comparação vectorial simples, LSI, utilizando apenas descrições textuais, e LSI utilizando descrições textuais e características das imagens. Para estes dois últimos foram utilizados os parâmetros que produziram os melhores resultados globais para as taxas de recuperação e precisão, ou seja, um k_p de 0.95 para o LSI apenas com texto e um k_p de 0.9 e importâncias relativas de 10/100 para o LSI com texto e características de baixo nível.

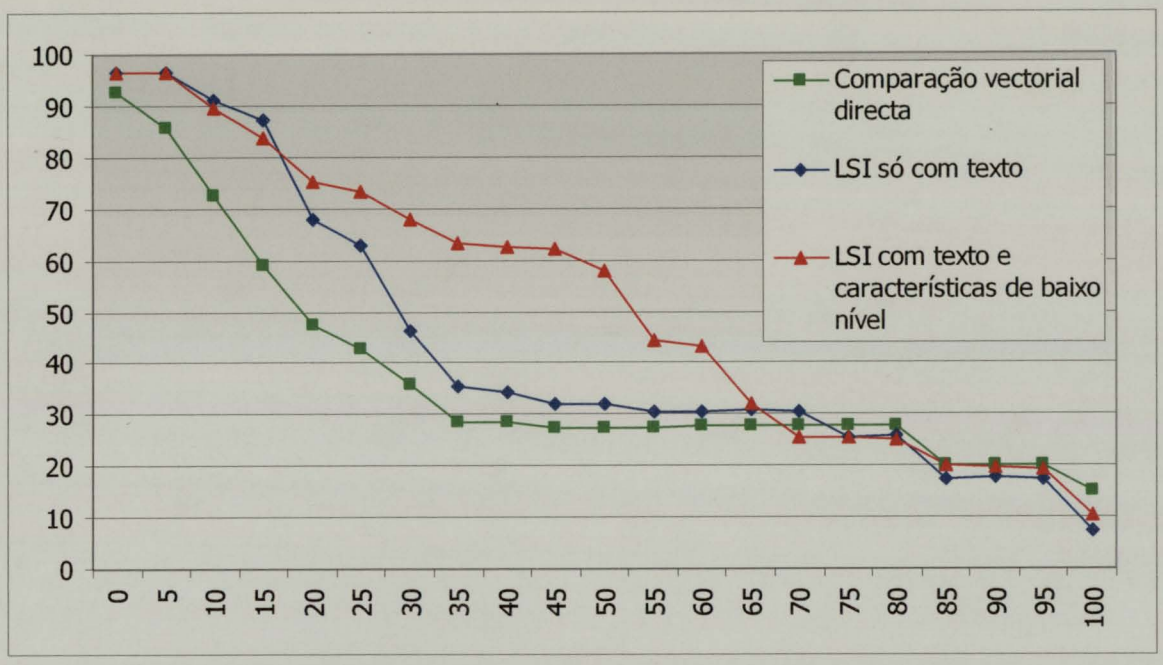


Figura 16 – Curvas de precisão/recuperação para diferentes modelos de indexação

Pode-se observar que a experiência com utilização de texto e características de baixo nível é a que apresenta uma degradação da precisão menos acentuada à medida que a taxa de recuperação aumenta, apresentando resultados muito superiores para taxas de recuperação de 20% a 65%. Dos 0% aos 15%, a mesma experiência apresenta resultados apenas ligeiramente inferiores aos obtidos com a utilização do LSI com apenas texto. Dos 75% aos 100% de taxa de recuperação, a comparação vectorial directa obtém as taxas de precisão mais elevadas, sendo, no entanto, notório que no cômputo global apresenta os piores resultados das três experiências em causa.

De forma a comprovar a robustez dos resultados observados para a utilização do LSI com texto e características de baixo nível foram também comparadas as curvas de precisão/recuperação para os parâmetros do protótipo que permitiram obter as melhores taxas de recuperação e precisão, como se pode ver no gráfico da página seguinte.

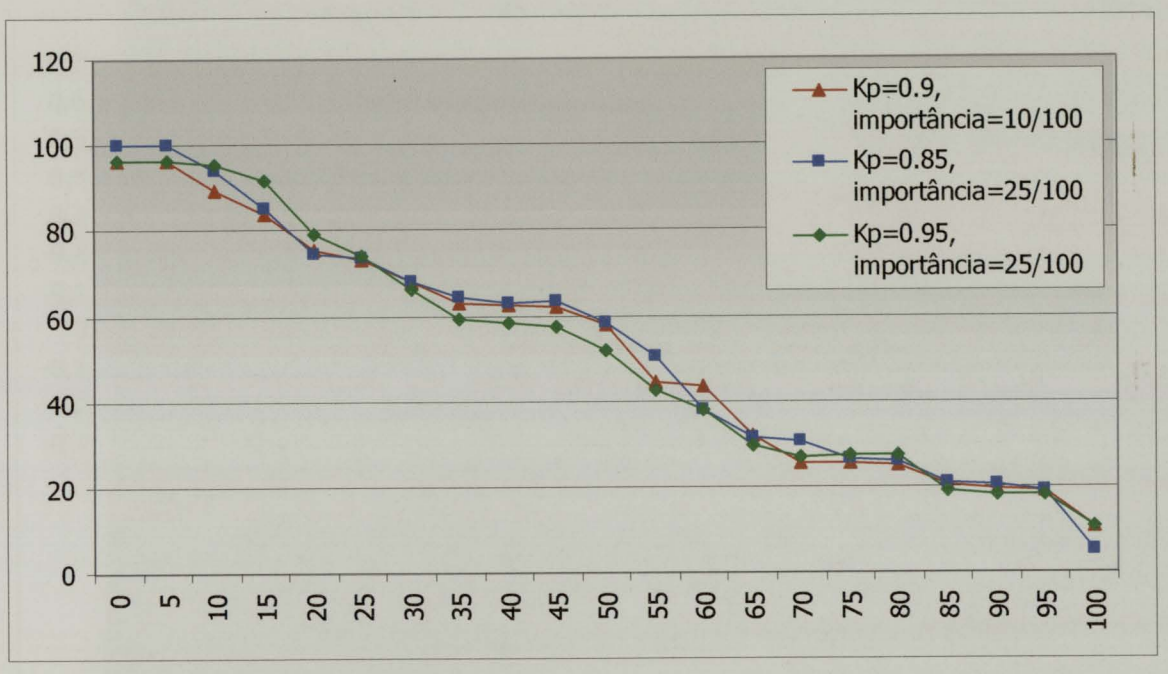


Figura 17 – Curvas de precisão/recuperação utilizando LSI com texto e características de baixo nível para diferentes configurações do sistema

É visível que as curvas para as diferentes configurações do protótipo são bastante aproximadas, o que comprova, juntamente com as curvas observadas no gráfico da figura 12, que a utilização do LSI com texto e características da imagem permite obter resultados consistentes para diferentes quantidades de documentos recuperadas.

5.3 Resultados de interrogações individuais com um só conceito

Na avaliação de resultados globais, através de taxas de recuperação, taxas de precisão e das curvas de precisão/recuperação, podem não ser verificadas anomalias patentes em interrogações individuais e, ao mesmo tempo, não é possível detectar em que interrogações cada algoritmo obtém melhores ou piores resultados [2]. Assim, é importante utilizar medidas que permitam avaliar o resultado das interrogações uma a uma, tendo sido escolhida para este fim a medida precisão-R, que se calcula através da divisão do número de documentos relevantes retornados para uma interrogação pelo total de documentos relevantes para essa interrogação. A precisão-R traduz a sensibilidade de uma determinada configuração do protótipo ao número de documentos recuperados.

Foram criados histogramas desta medida para a comparação, dois a dois, dos três paradigmas utilizados na indexação. Em cada histograma é subtraído o valor da precisão-R de um paradigma ao outro, observando-se, portanto, um valor positivo, de zero ou negativo para uma dada interrogação, conforme o primeiro paradigma apresente um valor superior, idêntico ou inferior ao da precisão-R do segundo paradigma.

No gráfico da página seguinte encontram-se comparadas a utilização do LSI apenas com texto e a comparação vectorial directa.

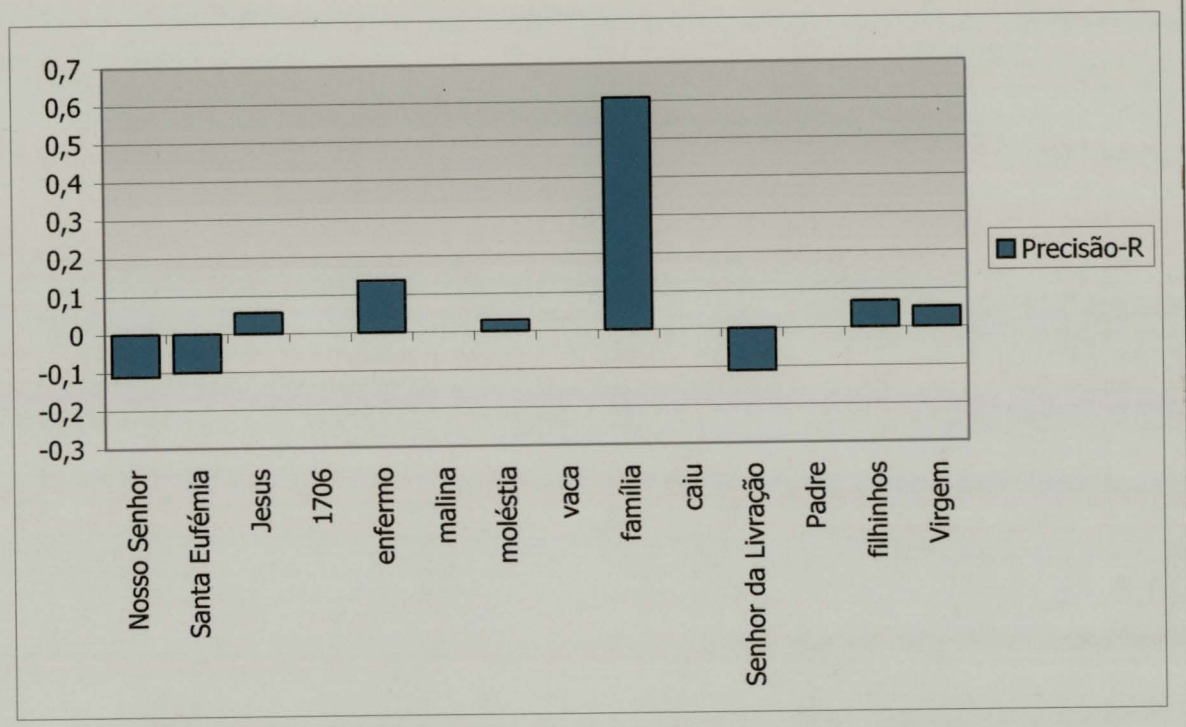


Figura 18 – Histograma de precisão-R para LSI apenas com texto/comparação vectorial directa

Pode-se verificar que em seis das catorze interrogações o LSI apenas com texto produz melhores resultados que a comparação vectorial directa, sendo o ganho no caso da interrogação "família" muito elevado. Uma explicação possível para este ganho é que o LSI tenha relacionado palavras como "irmão" ou "pais" a "família" e tenha assim retornado documentos que embora não contivessem a palavra "família" representavam famílias.

A comparação vectorial directa apresenta melhores resultados nas interrogações "Nosso Senhor", "Santa Eufémia" e "Senhor da Livração". Possivelmente o LSI terá estabelecido relações erróneas entre estes termos e outros presentes nas descrições textuais, sendo assim considerados relevantes documentos sem interesse para estas interrogações.

Milagre que fez Nossa Senhora das Necessidades a Francisco Marques de Vila Meã por uma moléstia grave e a Senhora lhe deu saúde. Ano de 1856.

Manuel Rodrigues de Loureiro, de Travassos de Baixo, próximo de Viseu, estando gravemente enfermo recorreu à Nossa Senhora da Lapa e ela lhe deu saúde. 1892.

Tabela 3 – Ex-votos e transcrições das descrições respectivas para português moderno

No gráfico da página seguinte é relacionada a utilização do LSI com texto e características de baixo nível com a comparação vectorial directa.

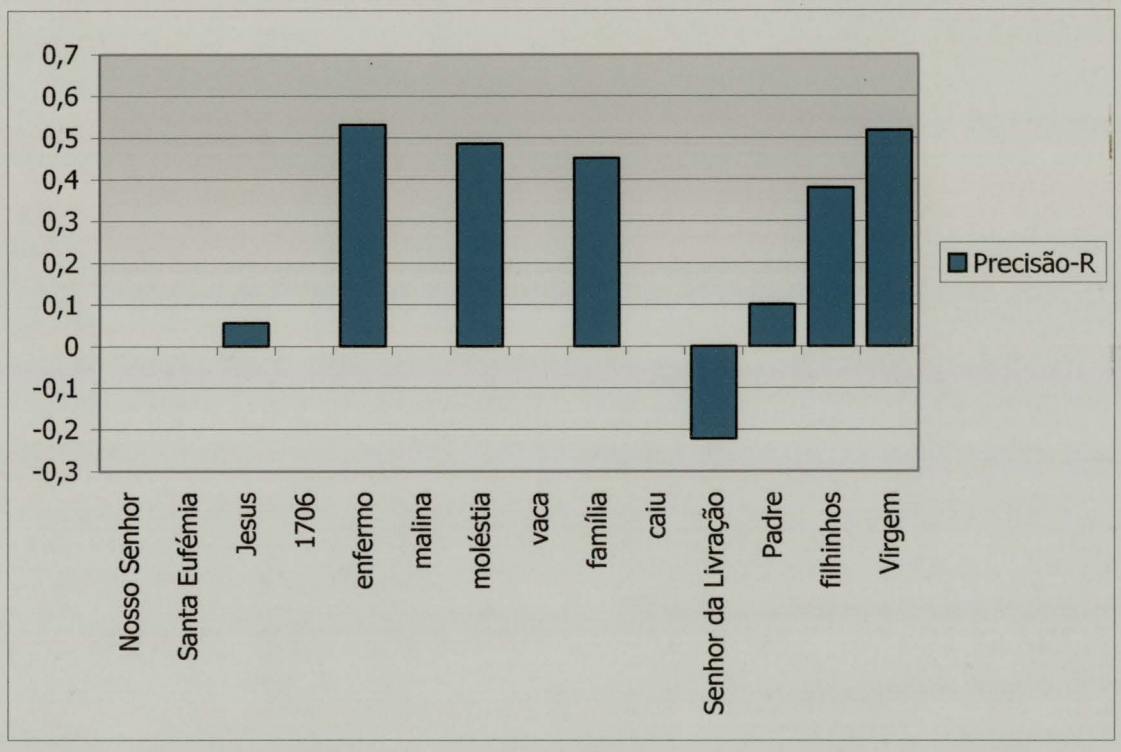


Figura 19 – Histograma de precisão-R para LSI com texto e características de baixo nível/comparação vectorial directa

Como se pode observar, a utilização de texto e características de baixo nível produz melhores resultados que a comparação vectorial simples em sete das catorze interrogações e resultados idênticos em seis das restantes. A interrogação "Senhor da Livração" é a única em que a comparação vectorial directa produz melhores resultados, tendo provavelmente o LSI associado características da imagem às palavras da interrogação que não se traduziram em relações realmente significativas. É também visível que, para as interrogações "enfermo", "moléstia", "família", "filhinhos" e "Virgem", os resultados foram substancialmente melhores, o que se pode explicar pela detecção por parte do LSI da existência de características textuais e visuais em comum na maioria dos documentos relevantes para cada uma destas. Nas duas primeiras interrogações, a presença do doente na cama, de elementos de decoração de interiores (cortinas, móveis, etc.) e da família do doente ajoelhada em prece na maioria dos ex-votos aos quais as palavras "enfermo" e "moléstia" se encontram associados (ver tabela 2) pode ter levado o LSI a agrupar estes elementos, sendo também retornadas imagens relevantes que não contêm as palavras da interrogação, apenas os elementos visuais referidos. Relativamente às interrogações "filhinhos" e "virgem", a presença de vários filhos prostrados, em muitos dos ex-votos em que a primeira palavra é referida, e de Nossa Senhora, em muitos dos ex-votos em que a segunda palavra é mencionada, pode também ter levado o LSI a estabelecer correspondências entre estes elementos textuais e visuais.

No seguinte gráfico é comparado o LSI utilizando texto e características de baixo nível com o LSI utilizando apenas texto.

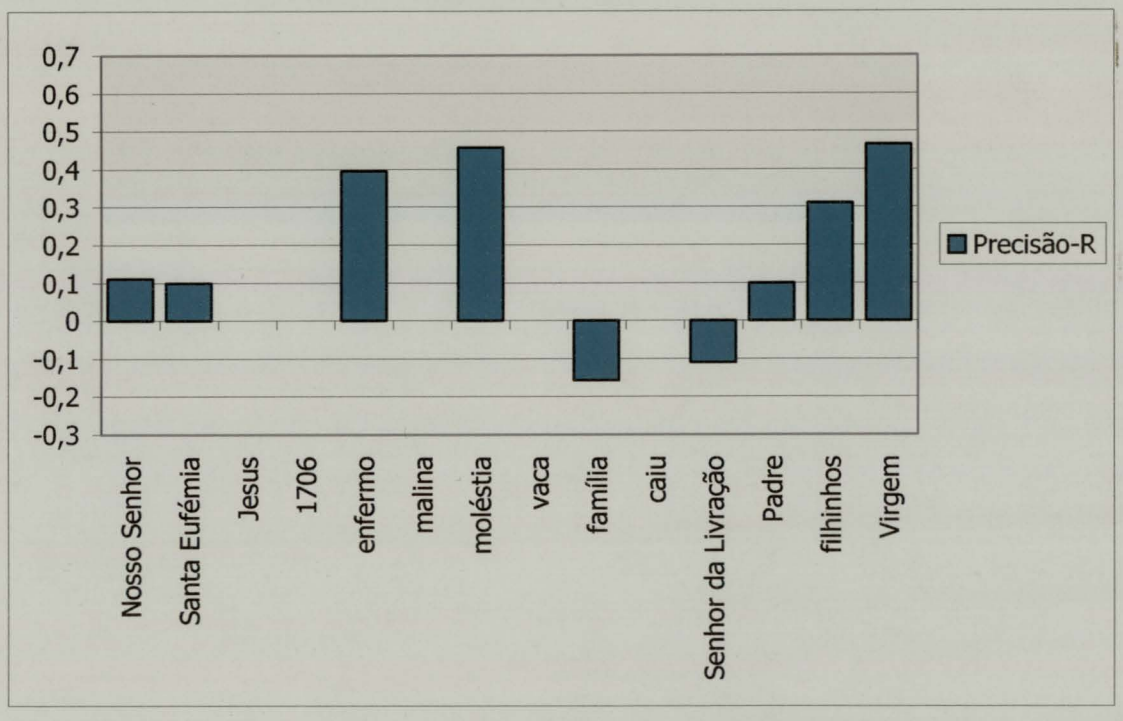


Figura 20 – Histograma de precisão-R para LSI com texto e características de baixo nível/LSI apenas com texto

É visível que os ganhos do LSI com texto e características de baixo nível em relação ao LSI apenas com texto são menores do que em relação à comparação vectorial directa, apresentando, no entanto, melhores resultados em sete das catorze interrogações e resultados idênticos em cinco das restantes. Nas interrogações "Nosso Senhor" e "Santa Eufémia", a utilização de características visuais parece corrigir as perdas introduzidas pelo LSI apenas com texto, enquanto que nas interrogações "enfermo", "moléstia", "filhinhos" e "Virgem" as razões referidas para os ganhos em relação à comparação vectorial directa também se aplicam aqui.

5.4 Resultados de interrogações individuais com mais do que um conceito

Até ao momento, as interrogações utilizadas procuravam obter imagens que traduzem apenas um conceito, mas é possível que, mesmo num conjunto de dados de temática reduzida como é o aqui tratado, que os utilizadores procurem imagens que exprimam vários conceitos. Ao mesmo tempo, também é possível que o utilizador procure obter mais resultados sobre um determinado conceito através do uso de sinónimos. Assim, foram realizados testes sobre as interrogações presentes na seguinte tabela:

Nosso Senhor enfermo
Santa Eufémia moléstia
Santa Eufémia enfermo
Senhor da Livração filhinhos
Nossa Senhora família
Virgem padre
enfermo filhinhos
Nossa Senhora Virgem
Nosso Senhor Jesus

Tabela 4 – Interrogações com mais do que um conceito, utilizadas na avaliação de resultados

No seguinte gráfico encontram-se comparadas as curvas de precisão/recuperação para a utilização de comparação vectorial directa, LSI apenas com texto e LSI com texto e características de baixo nível, tendo sido utilizados para estes dois últimos os parâmetros que melhores resultados produziram para o primeiro conjunto de interrogações – $\mathbf{k_p}$ de 0.95 para o LSI apenas com texto, $\mathbf{k_p}$ de 0.9 e importâncias relativas de 10/100 para o LSI com texto e características da imagem.

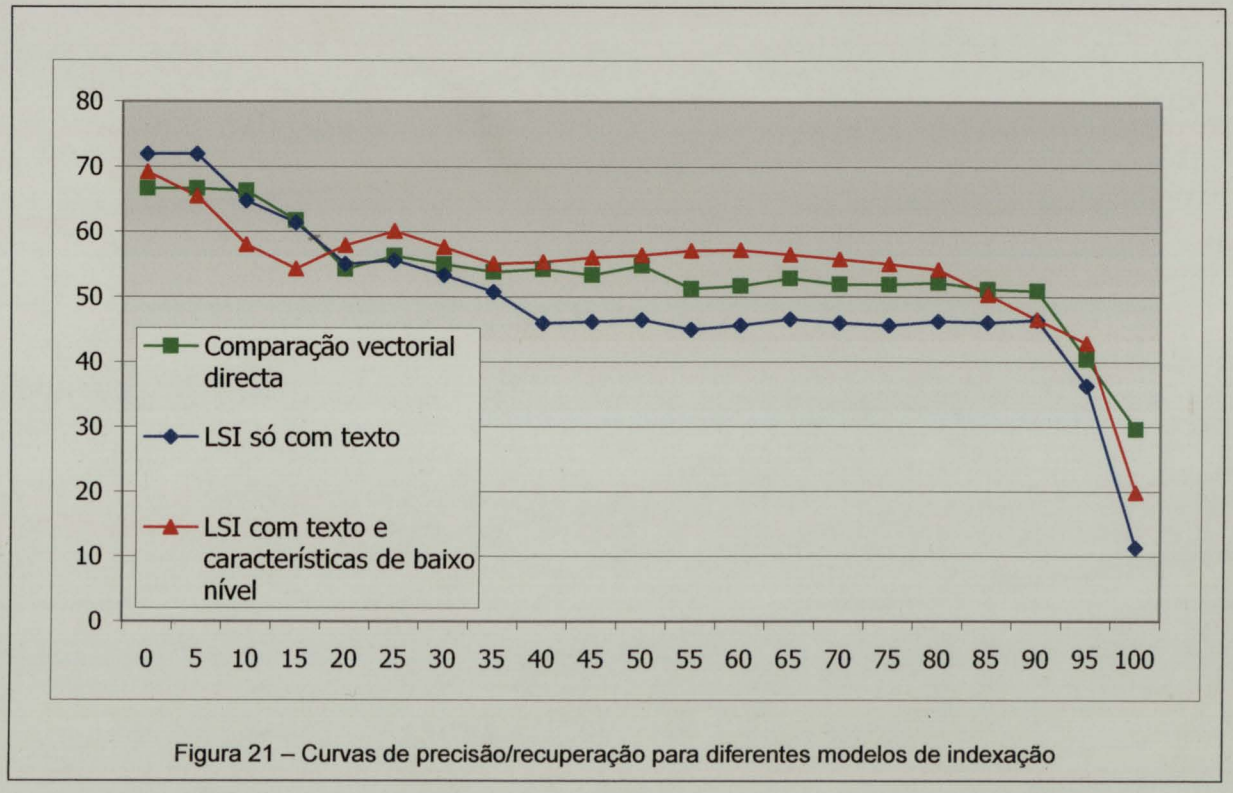


Figura 21 – Curvas de precisão/recuperação para diferentes modelos de indexação

É visível que quando é utilizado o LSI com texto e características da imagem a taxa de precisão é superior para a maior parte das taxas de recuperação (entre 20% a 80%). No entanto, este método apresenta os piores resultados para taxas de recuperação entre 0% a

15%. Para os três métodos pode-se concluir que a deterioração da taxa de precisão é menos acentuada para este conjunto de interrogações do que para o primeiro utilizado (ver figura 16), o que pode ser explicado pelo maior número de documentos relevantes para este segundo conjunto. É também de salientar que a comparação vectorial simples apresenta melhores resultados para a maioria das taxas de recuperação do que o LSI apenas com descrições textuais.

De forma a comprovar a consistência dos resultados observados com a utilização do LSI com texto e características de baixo nível para este conjunto de interrogações, foram também comparadas as curvas de precisão/recuperação para as configurações do sistema que permitiram obter os melhores resultados para o conjunto de interrogações anterior. O resultado desta comparação pode ser observado no seguinte gráfico:

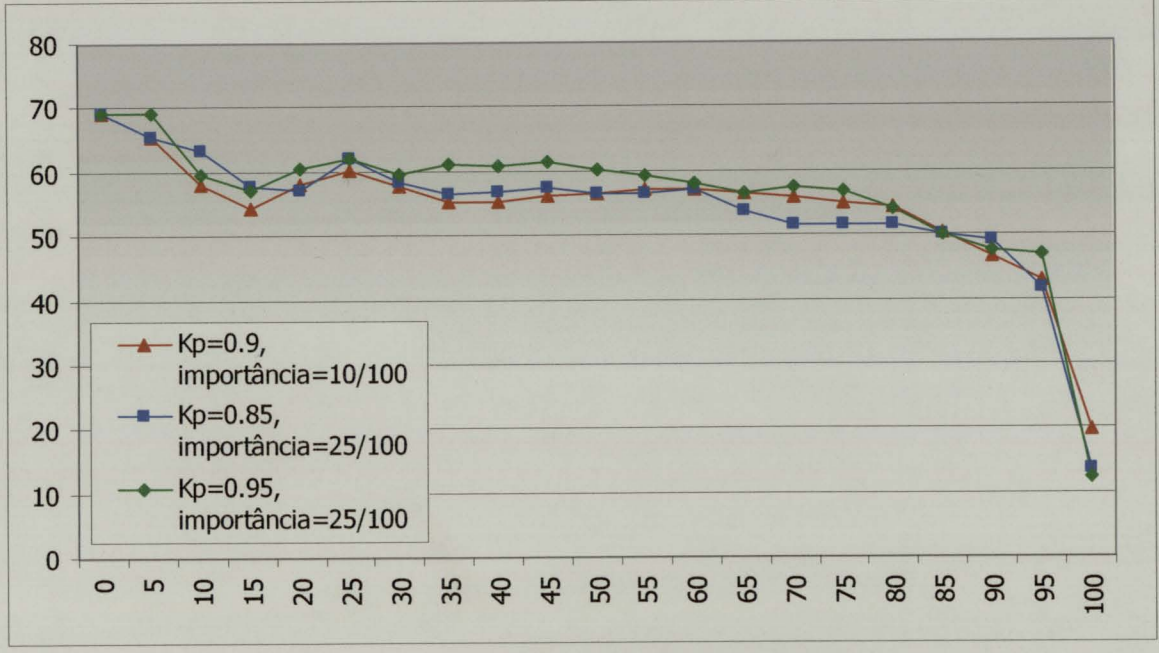


Figura 22 – Curvas de precisão/recuperação utilizando LSI com texto e características de baixo nível para diferentes configurações do sistema

Pode-se concluir, uma vez que as curvas para as diferentes configurações são bastante aproximadas entre si, que a utilização do LSI com descrições textuais e características da imagem tem pouca sensibilidade no que diz respeito ao número de documentos recuperados e à variação dos parâmetros utilizados no sistema. Apesar desta proximidade, pode-se observar que a configuração com um $\mathbf{k_p}$ de 0.95 e importâncias relativas de 25/100 é a que obtém melhores resultados para quase toda a gama de taxas de recuperação.

5.5 Comparação do desempenho dos descritores do MPEG-7

À semelhança da avaliação dos resultados utilizando os descritores do MPEG-7 em conjunto, a avaliação destes foi efectuada com a variação de dois parâmetros do sistema, a percentagem de dimensões retidas na matriz após o processamento do LSI ($\mathbf{k_p}$) e a importância relativa da informação textual e da informação de baixo nível proveniente das imagens. O número de coeficientes utilizado para cada descritor foi sempre o máximo possível, ou seja, foram utilizados 256 coeficientes para o SCD e para o CSD, tendo os restantes retido o mesmo

número. Cada descritor foi testado individualmente, tendo sido utilizadas as mesmas catorze interrogações apresentadas previamente.

Na seguinte tabela encontra-se, para cada um dos descritores, a variação da taxa de recuperação com diferentes importâncias relativas para o texto e características da imagem:

Importância das descrições textuais	Importância das características de baixo nível da imagem	Média de recuperação				
		SCD	CSD	CLD	HTD	EHD
100%	100%	41.0%	41.8%	42.0%	40.0%	45.2%
100%	75%	40.7%	41.3%	40.5%	40.2%	44.4%
100%	50%	41.0%	40.6%	39.5%	40.9%	42.8%
100%	25%	40.4%	40.0%	39.2%	40.3%	40.7%
75%	100%	40.6%	42.3%	42.7%	40.2%	45.3%
50%	100%	41.1%	42.9%	42.6%	40.5%	45.4%
25%	100%	41.9%	43.8%	42.5%	40.9%	46.3%
10%	100%	41.9%	44.9%	42.9%	41.0%	46.4%

Tabela 5 – Variação das médias de recuperação com as importâncias relativas do texto e da imagem para cada descritor do MPEG-7

Pode-se observar que o EHD apresenta os melhores resultados para as várias importâncias relativas, apresentando também o valor máximo absoluto (46.4%), para importâncias relativas de 10/100. Dos restantes descritores, o CSD apresenta os melhores resultados para importâncias reduzidas do texto. O HTD é o que apresenta os piores resultados. Pode-se observar que, quando a importância das características de baixo nível é diminuída, os resultados obtidos pelos vários descritores se aproximam entre si: os resultados dependem quase só das descrições textuais.

Na tabela abaixo pode ser observada uma comparação idêntica para a taxa de precisão:

Importância das descrições textuais	Importância das características de baixo nível da imagem	Média de precisão				
		SCD	CSD	CLD	HTD	EHD
100%	100%	41.4%	42.6%	43.0%	41.4%	44.3%
100%	75%	41.5%	42.6%	41.3%	41.5%	43.9%
100%	50%	42.0%	42.0%	40.7%	41.4%	43.3%
100%	25%	41.4%	40.5%	40.1%	40.8%	41.5%
75%	100%	41.2%	42.9%	43.8%	41.7%	44.2%
50%	100%	41.2%	43.6%	44.4%	42.0%	43.8%
25%	100%	42.0%	43.8%	43.8%	41.9%	43.8%
10%	100%	42.2%	44.2%	43.9%	42.4%	43.7%

Tabela 6 – Variação das médias de precisão com as importâncias relativas do texto e da imagem para cada descritor do MPEG-7

O EHD apresenta também aqui os melhores resultados, encontrando-se o CSD mais próximo. Os valores mais elevados são obtidos pelo CLD para importâncias relativas de 50/100 (44.4%), pelo EHD para importâncias relativas de 100/100 (44.3%) e 75/100 (44.2%) e pelo CSD para importâncias relativas de 10/100 (44.2%). Os piores resultados são também apresentados pelo HTD, embora esteja mais perto dos outros descritores do que no caso da taxa de recuperação.

O gráfico da página seguinte compara, para valores de $\mathbf{k_p}$ entre 0.05 e 0.95, a taxa de recuperação observada na utilização do LSI com cada descritor, tendo sido calculada a média dos valores obtidos com cada configuração das importâncias relativas para cada valor de $\mathbf{k_p}$.

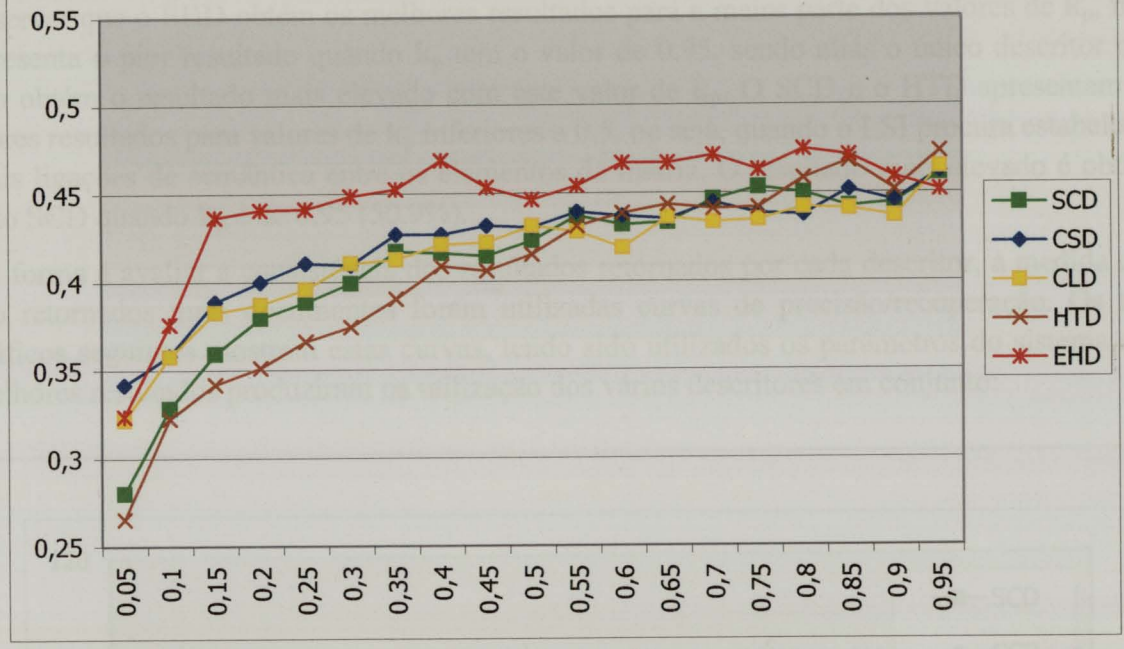


Figura 23 – Comparação das taxas de recuperação das várias características de baixo nível semântico para diferentes valores de $\mathbf{k_p}$

É visível que o EHD apresenta os melhores resultados para a maior parte dos valores de $\mathbf{k_p}$, embora obtenha o pior resultado quando $\mathbf{k_p}$ tem um valor de 0.95. O HTD apresenta os piores resultados para valores de $\mathbf{k_p}$ inferiores a 0.55. Os resultados mais elevados são obtidos pelo EHD para um $\mathbf{k_p}$ de 0.8 (47.5%) e pelo HTD para um $\mathbf{k_p}$ de 0.95 (47.3%).

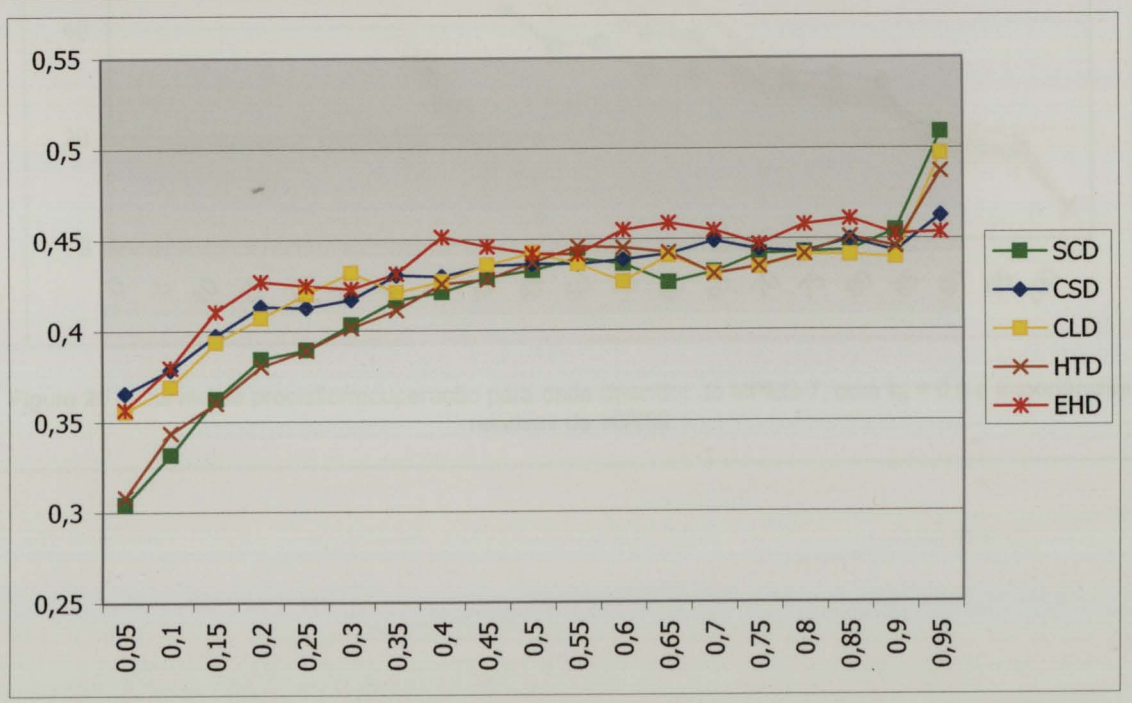


Figura 24 – Comparação das taxas de precisão das várias características de baixo nível semântico para diferentes valores de $\mathbf{k_p}$

O gráfico da página anterior apresenta a mesma comparação para a taxa de precisão. Pode-se observar que o EHD obtém os melhores resultados para a maior parte dos valores de $\mathbf{k_p}$, mas apresenta o pior resultado quando $\mathbf{k_p}$ tem o valor de 0.95, sendo aliás o único descritor que não obtém o resultado mais elevado com este valor de $\mathbf{k_p}$. O SCD e o HTD apresentam os piores resultados para valores de $\mathbf{k_p}$ inferiores a 0.5, ou seja, quando o LSI procura estabelecer mais ligações de semântica entre os elementos da matriz. O resultado mais elevado é obtido pelo SCD quando $\mathbf{k_p}$ é de 0.95 (50.9%).

De forma a avaliar a consistência dos resultados retornados por cada descritor, à medida que são retornados mais documentos foram utilizadas curvas de precisão/recuperação. Os três gráficos seguintes mostram estas curvas, tendo sido utilizados os parâmetros do sistema que melhores resultados produziram na utilização dos vários descritores em conjunto:

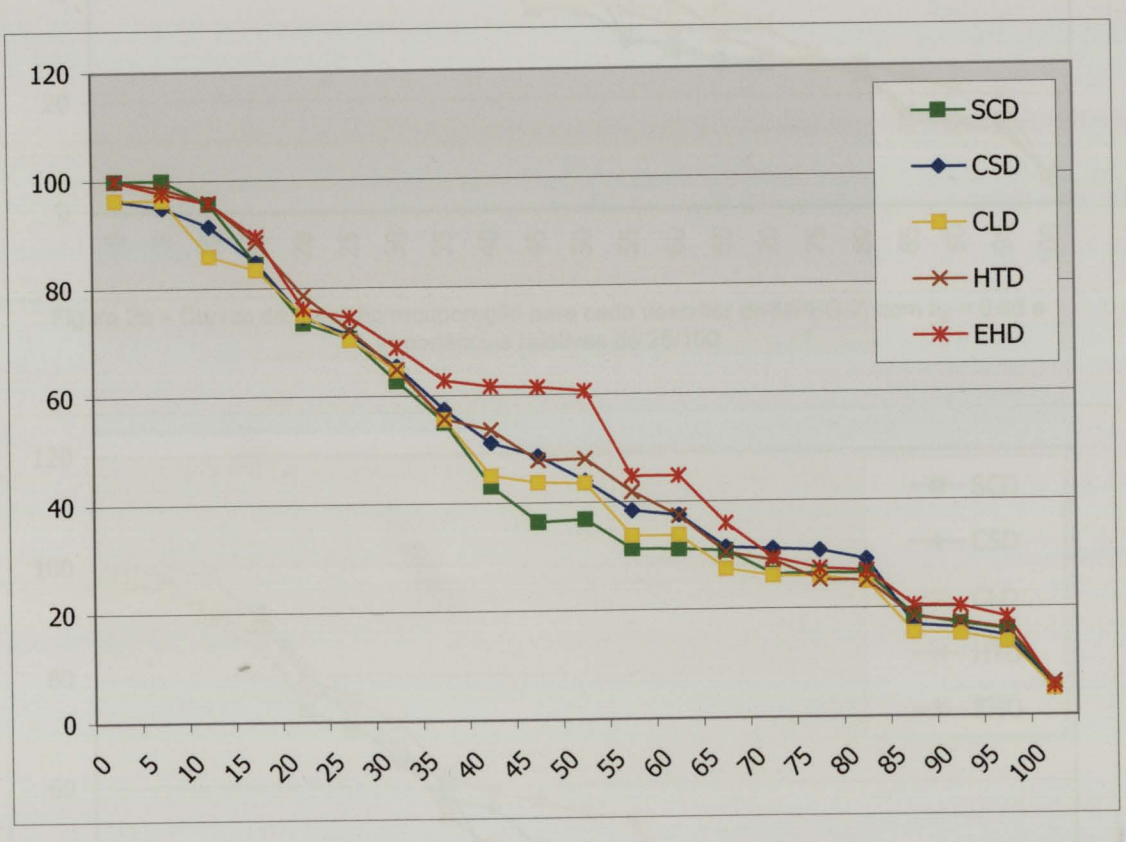


Figura 25 – Curvas de precisão/recuperação para cada descritor do MPEG-7, com $\mathbf{k_p}$ = 0.9 e importâncias relativas de 10/100

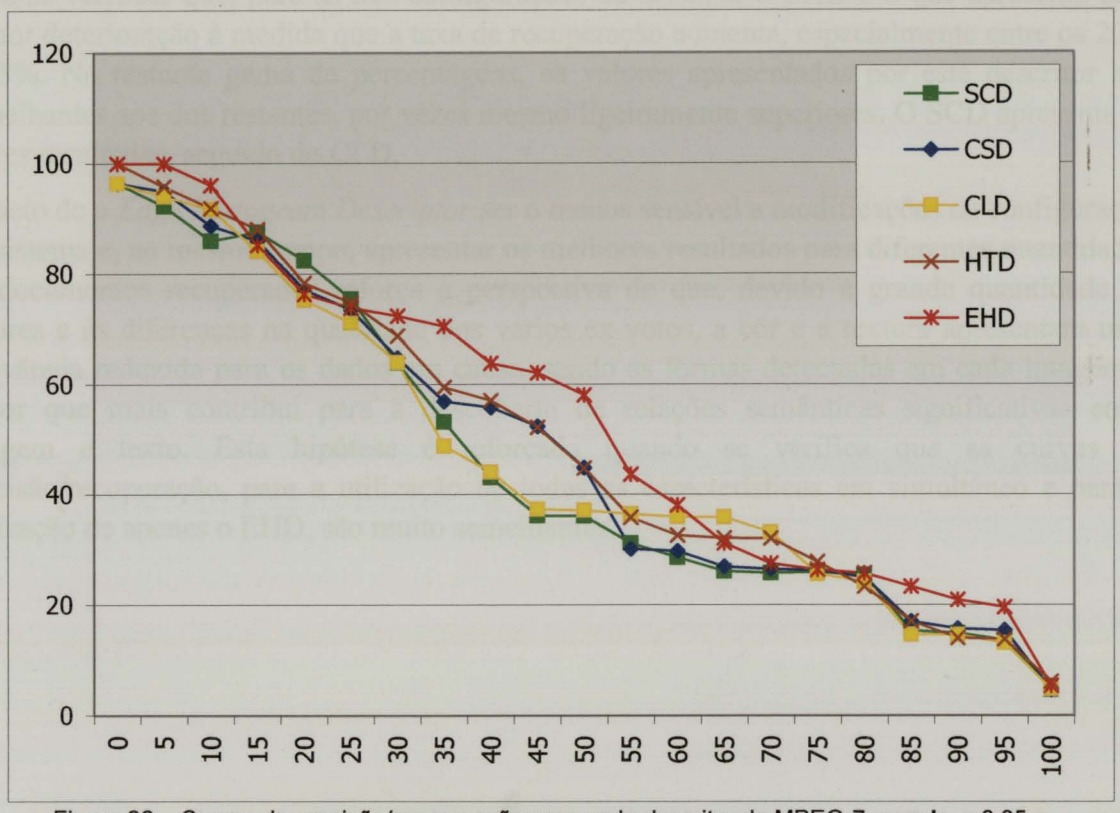


Figura 26 – Curvas de precisão/recuperação para cada descritor do MPEG-7, com k_p = 0.85 e importâncias relativas de 25/100

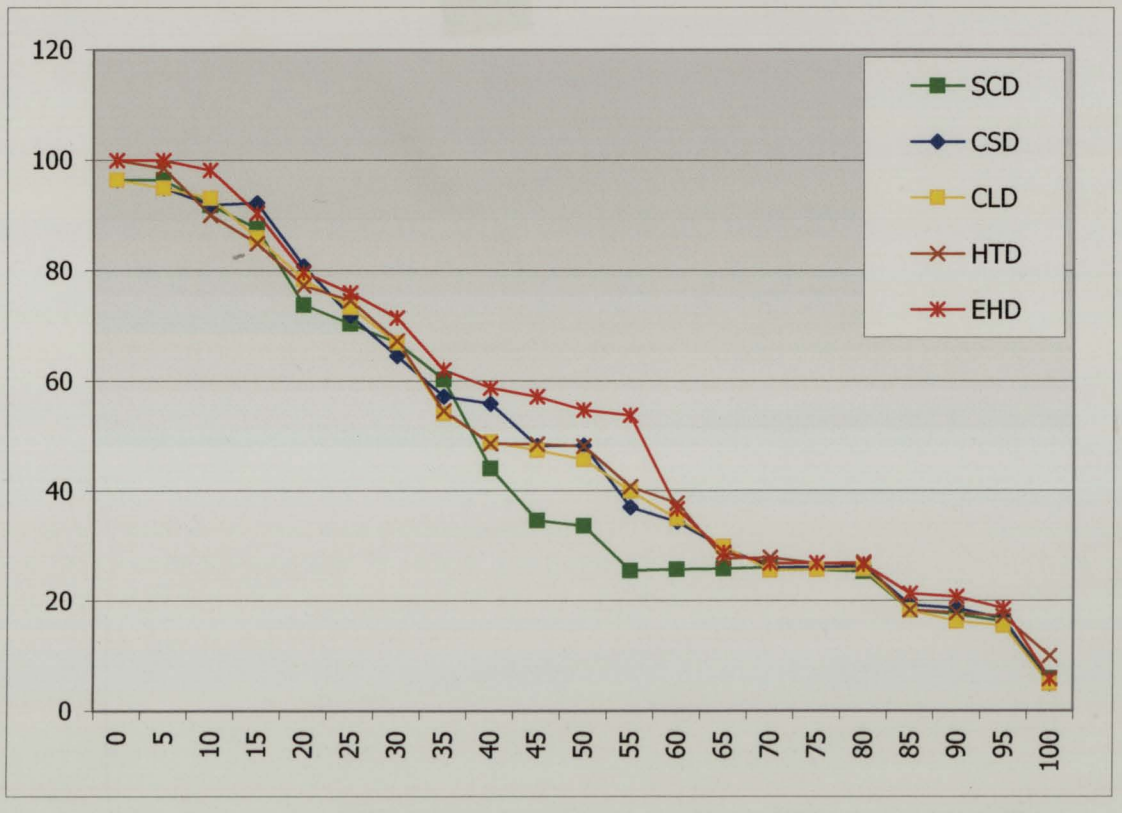


Figura 27 – Curvas de precisão/recuperação para cada descritor do MPEG-7, com k_p = 0.95 e importâncias relativas de 25/100

Pode-se verificar que, para as três configurações do sistema, o EHD é o que apresenta uma menor deterioração à medida que a taxa de recuperação aumenta, especialmente entre os 20% a 65%. Na restante gama de percentagens, os valores apresentados por este descritor são semelhantes aos dos restantes, por vezes mesmo ligeiramente superiores. O SCD apresenta os piores resultados, seguido do CLD.

O facto de o *Edge Histogram Descriptor* ser o menos sensível a modificações na configuração do sistema e, ao mesmo tempo, apresentar os melhores resultados para diferentes quantidades de documentos recuperados reforça a perspectiva de que, devido à grande quantidade de autores e às diferenças na qualidade dos vários ex-votos, a cor e a textura apresentam uma relevância reduzida para os dados em causa, sendo as formas detectadas em cada imagem o factor que mais contribui para a descoberta de relações semânticas significativas entre imagem e texto. Esta hipótese é reforçada quando se verifica que as curvas de precisão/recuperação, para a utilização de todas as características em simultâneo e para a utilização de apenas o EHD, são muito semelhantes.

6 Conclusões e experiências futuras

Foi apresentada uma proposta para a pesquisa automática de imagens com interrogações textuais, sendo utilizado o LSI para a integração de características de baixo nível semântico retiradas da imagem e características de alto nível semântico retiradas do texto. Esta proposta visa facilitar a utilização simultânea de documentos com várias estruturas, provenientes de diferentes bases de dados.

O protótipo concebido a partir desta proposta produziu bons resultados com a utilização de texto e de características da imagem em simultâneo, melhorando os resultados obtidos, quer com a utilização de comparação vectorial, quer com a utilização do LSI apenas com texto, para uma grande gama de documentos recuperados. Os resultados obtidos mostraram-se resistentes a variações dos parâmetros do sistema. Estes resultados são particularmente favoráveis se forem tidos em conta o vocabulário inconstante das descrições textuais e as variações de estilo nas imagens utilizadas.

Foram também comparados os desempenhos individuais de cinco dos descritores do MPEG-7 usados, tendo o *Edge Histogram Descriptor* obtido os melhores resultados. Este facto justifica-se com as enormes variações nas cores e texturas presentes nas imagens de teste utilizadas, o que torna difícil a utilização eficiente dos restantes descritores, e com a presença de ícones com características definidas, nomeadamente Jesus, Nossa Senhora e vários santos, e de outras formas diferentes, como camas e pessoas ajoelhadas, na maioria das imagens.

Relativamente a experiências futuras, seria interessante testar a utilização do protótipo com outros conjuntos de dados em que a cor e a textura apresentassem um papel mais relevante.

Seria interessante definir um algoritmo que determinasse automaticamente as importâncias relativas do texto e das características de baixo nível da imagem, conforme o número relativo dos termos e dos coeficientes das características, já que é possível que a obtenção de melhores resultados com importâncias reduzidas do texto se possa dever ao maior espaço ocupado na matriz pelos termos deste, relativamente ao pouco espaço ocupado pelas características de baixo nível. Aliás, o espaço ocupado pelos termos relativamente a estas características aumentaria caso fosse utilizada uma base de dados de maior dimensão, já que, independentemente do número de imagens nesta, o número de coeficientes das características de baixo nível é constante. Este problema já foi identificado por van Gemert [16], que sugere equilibrar o tamanho dos vectores de texto e de características de baixo nível de cada documento através da sua normalização.

Outra experiência com potencial passaria por utilizar esta técnica com uma base de dados que conjugasse anotações textuais e vídeo, até porque seria possível integrar características relativas às relações entre as "frames" e até o som, da mesma forma que os descritores de imagem foram integrados.

De forma a melhorar os resultados obtidos, poderia ainda ser utilizada segmentação. Uma hipótese seria utilizá-la colocando cada segmento como um objecto separado na matriz de indexação, com características de baixo nível relativas apenas a ele e não à totalidade da imagem, sendo necessária uma referência para o documento a que pertence caso seja necessário retorná-lo ao utilizador. Era nesta altura necessária a criação de uma nova operação na recuperação de documentos, já que seria possível serem retornados segmentos pertencentes ao mesmo documento mas com classificações de interesse para o utilizador bastante

diferentes. Outra hipótese de segmentação seria utilizar um algoritmo de correlação para, após a segmentação e o cálculo das características de baixo nível separadamente para cada segmento, integrar novamente as características como se pertencessem a um único documento, sendo mantidos apenas os valores mais constantes.

A *query relevance feedback* seria outra técnica que poderia melhorar a qualidade da interacção do utilizador com o sistema. Após a introdução de uma interrogação inicial na forma de texto, o utilizador poderia seleccionar uma ou mais imagens que considerasse relevante(s). Em seguida, poderiam ser utilizados três sistemas diferentes. No primeiro, seriam pesquisadas imediatamente imagens cujas características visuais fossem semelhantes às da imagem seleccionada, já que agora a interrogação seria uma imagem e não texto. No segundo, os pesos dos termos associados à imagem ou imagens seriam incrementados, sendo mais provável a sua recuperação em pesquisas futuras. No terceiro, que só seria funcional caso mais do que uma imagem fosse escolhida pelo utilizador, seria utilizado o LSI apenas sobre os documentos correspondentes às imagens seleccionadas, o que permitiria observar correlações entre as características dos documentos nos quais o utilizador havia manifestado interesse.

Também na fase de comparação do vector de pesquisa com os vectores dos documentos, poderiam ser testadas outras métricas de distância para além do coseno do ângulo entre os vectores como, por exemplo, a norma L1, utilizada em muitas experiências.

Referências e Bibliografia

[1] Araújo, A., D. Carneiro, et al. (1998). Estórias de Dor Esperança e Festa - O Brasil em Ex-Votos Portugueses (Séculos XVII-XIX), Comissão Nacional para as Comemorações dos Descobrimentos Portugueses.
[2] Baeza-Yates, R. and B. Ribeiro-Neto (1999). Modern Information Retrieval, Addison-Wesley.
[3] Barnard, K., P. Duygulu, et al. (2003). "Matching Words and Pictures." Journal of Machine Learning Research **3**: 1107–1135.
[4] Benitez, A. B. and S.-F. Chang (2002). "Multimedia Knowledge Integration, Summarization and Evaluation."
[5] Bezdek, J. C. (1993). "Fuzzy Models - What Are They, and Why?" IEEE Transactions on Fuzzy Systems **1**(1): 1-6.
[6] Bimbo, A. D. (1999). Visual Information Retrieval, Morgan Kaufmann.
[7] Carson, C., S. Belongie, et al. (2002). "Blobworld: Image segmentation using Expectation-Maximization and its application to image querying." IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(8): 1026–1038.
[8] Cascia, M. L., S. Sethi, et al. (1998). Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. IEEE Workshop on Content-based Access of Image and Video Libraries, Santa Barbara, California.
[9] Chiang, C.-C., D.-W. Fu, et al. (2003). On Extracting Color-Size Features for Image Classification. 16th IPPR Conference on Computer Vision, Graphics and Image Processing.
[10] Colombo, C., A. D. Bimbo, et al. (1999). "Semantics in visual information retrieval." IEEE Multimedia **6**(3): 38-53.
[11] Consortium, E. (2000). Functional and technical design, European Visual Archive. URL: http://www.eva-eu.org.
[12] Correia, A., J. L. Porfíriro, et al (1998). Do Gesto à Memória, Instituto Português de Museus – Ministério da Cultura.
[13] Deerwester, S., S. T. Dumais, et al. (1990). "Indexing by latent semantic analysis." Journal of the American Society of Information Science **41**(6): 391-407.
[14] Flickner, M., H. Sawhney, et al. (1995). "Query by image and video content: The QBIC system." IEEE Computer **28**(9): 23-32.
[15] Fung, R. and B. D. Favero (1995). "Applying Bayesian Networks to Information Retrieval." Communications of the ACM **38**(3): 42-48.
[16] Gemert, J. C. van Gemert (2003). Retrieving Images as Text. Amsterdão, Universidade de Amsterdão.
[17] Greengrass, E. (2000). Information Retrieval: A Survey. Baltimore, University of Maryland, Baltimore County Center for Architectures for Data-Driven Information Processing: 15.
[18] Harman, D. (1998). The Text REtrieval Conferences (TRECs): Providing a Test-Bed for Information Retrieval Systems.
URL: http://www.asis.org/Bulletin/Apr-98/harman.html
[19] Howe, A. E. and D. Dreilinger (1997). SavvySearch: A meta-search engine that learns which search engines to query. AI Magazine. **18**.
[20] Huang, J., S. R. Kumar, et al. (1997). Image Indexing Using Color Correlograms. IEEE Comp. Soc. Conf. Comp. Vis. and Patt. Rec.
[21] Jones, K. S. (1972). "A statistical interpretation of term specificity and its application in retrieval." Journal of Documentation **28**(1): 11-20.

[22] Jones, K. S. (2003) Document Retrieval: Shallow Data, Deep Theories; Historical Reflections, Potential Directions. 25th European Conference on Information Retrieval (ECIR-03), (Ed. F. Sebastiani), Lecture Notes in Computer Science 2633, Berlin: Springer, 1-11.
[23] Jung, Y., H. Park, et al. (2000). An Effective Term-Weighting Scheme for Information Retrieval. Minneapolis, University of Minnesota.
[24] Kleinberg, J. M. (1999). "Authoritative Sources in a Hyperlinked Environment." Journal of the ACM **46**(5): 604–632.
[25] Lawrence, S. and C. L. Giles (1998). Searching the World Wide Web. Science. **280:** 98-100.
[26] Manjunath, B. S., J.-R. Ohm, et al. (2001). "Color and Texture Descriptors." Transactions on Circuits and Systems for Video Technology **11**(6): 703-715.
[27] Martinez, J.M. (2001). Overview of the MPEG-7 standard (version 5.0). ISO/IEC JTC1/SC29/WG11 N4031.
URL: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
[28] Miller, G. A., R. Beckwith, et al. (1993). Introduction to WordNet: An On-line Lexical Database.
[29] Müller, H., W. Müller, et al. (2001). Automated Benchmarking in Content-based Image Retrieval. Geneva, Computer Vision Group, Universidade de Geneva.
[30] Notess, G. R. (2002). Google Special Report: Google's Unindexed URLs.
URL: http://www.searchengineshowdown.com/features/google/unindexed.shtml
[31] Notess, G. R. (2004). Review of Google.
URL: http://www.searchengineshowdown.com/features/google/review.html
[32] Ojala, T., M. Pietikäinen, et al. (2002). "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns." IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7): 971-987.
[33] Ojala, T., M. Aittola, et al. (2002). "Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories."
[34] Ojala, T., T. Mäenpää, et al. (2002). "Empirical Evaluation of MPEG-7 Texture Descriptors with A Large-Scale Experiment."
[35] Ojala, T., M. Rautiainen, et al. (2001). Semantic image retrieval with HSV correlograms. 12th Scandinavian Conference on Image Analysis, Bergen, Noruega.
[36] O'Neill, E. T., B. F. Lavoie, et al. (2003). Trends in the Evolution of the Public Web. D-Lib Magazine. **9**.
URL: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html
[37] Ozcanli, O. C. and F. T. Y. Vural (2002). A Content Based Image Retrieval System Based on Localization of the Query. Ankara, Turkey, Middle East Technical University.
[38] Page, L., S. Brin, et al. (1998). The Pagerank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project.
[39] Rema, H. P. (2003). Santo António De Lisboa - Ex-Votos, Quetzal Editores/Bertrand Editora.
[40] Salton, G. and C. Buckey (1988). "Term Weighting Approaches in Automatic Text Retrieval." Information Processing and Management **24**(5): 513-523.
[41] Sclaroff, S., L. Taycher, et al. (1997). Imagerover: A Content-based Image browser for the World Wide Web. IEEE Workshop on Content-based Access of Image and Video Libraries.
[42] Shahabi, C. and Y.-S. Chen (2000). Soft Query in Image Retrieval Systems. SPIE internet imaging (EI14), San Jose, California.
[43] Sikora, T. (2001). "The MPEG-7 Visual Standard for Content Description—An Overview." IEEE Transactions on Circuits and Systems for Video Technology **11**(6): 696-702.
[44] Smeulders, A. W. M., M. Worring, et al. (2000). "Content-Based Image Retrieval

at the End of the Early Years." IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12): 1349-1380.
[45] Song, F. and W. B. Croft (1999). A General Language Model for Information Retrieval. Eight International Conference on Information and Knowledge Management.
[46] Stricker, M. and M. Orengo (1995). "Similarity of Color Images." SPIE Storage and Retrieval for Image and Video Databases III **2420**: 381-392.
[47] Sullivan, D. (1997). How Big Are Search Engines?
URL: http://searchenginewatch.com/sereport/article.php/2165301
[48] Sullivan, D. (2003). Search Engine Sizes.
URL: http://searchenginewatch.com/reports/article.php/2156481
[49] Tao, Y. and W. I. Grosky (2001). "Spatial Color Indexing Using Rotation, Translation, and Scale Invariant Anglograms." Multimedia Tools and Applications **15**(3): 247–268.
[50] Zhao, R. and W. I. Grosky (2002). "Narrowing the Semantic Gap - Improved Text-Based Web Document Retrieval Using Visual Features." IEEE Transactions on Multimedia **4**(2): 189-200.

Ex-voto à Senhora da Lapa (s.d.)

Ex-voto à Senhora das Necessidades (s. d.)

Ex-voto ao Senhor dos Aflitos (1898)

Ex-voto ao Senhor dos Passos (1888)

Ex-voto à Senhora da Lapa (1886)

Ex-voto à Senhora das Necessidades (1909)