FACULDADE DE ENGENHARIA DA
UNIVERSIDADE DO PORTO

FEUP

Recuperação de Informação Multimédia em Larga Escala: Aplicação na Ilustração de Conteúdos Textuais

Filipe Coelho

Programa Doutoral em Engenharia Informática

Orientador: Cristina Ribeiro (Professora)

15 de Março de 2014

Recuperação de Informação Multimédia em Larga Escala: Aplicação na Ilustração de Conteúdos Textuais

Filipe Coelho

Programa Doutoral em Engenharia Informática

15 de Março de 2014

Informação de Contacto

Filipe Coelho
Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto

Rua Dr. Roberto Frias, s/n
4200-465 Porto
PORTUGAL

Email: filipe.coelho@fe.up.pt
URL: http://sites.google.com/site/filcoelhosoftmir/

Filipe Coelho
"Recuperação de Informação Multimédia em Larga Escala: Aplicação na Ilustração de Conteúdos Textuais"

"Here on earth you will have many trials and sorrows.
But take heart, because I have overcome the world."

John 16:33

Resumo

Vivemos atualmente na era da "sobrecarga de informação". Profissionais e consumidores desfrutam do acesso a dispositivos e formas de armazenamento capazes de capturar grandes quantidades de informação relevante. No entanto, estas características introduzem novas questões sobre a utilidade das coleções massivas de imagens, áudio, vídeo e texto face às capacidade de análise dos motores de pesquisa atuais.

Esta "explosão" de conteúdo multimédia constantemente produzido em áreas como o entretenimento, jornalismo, medicina e uso pessoal, entre outras, requer meios adequados de pesquisa e recuperação. Torna-se assim necessário analisar avanços alcançados em áreas distintas de recuperação de informação textual, visual e processamento em grande escala para determinar a viabilidade da aplicação e combinação dos mesmos na descoberta e recomendação de conteúdo.

A ilustração de textos representou a área escolhida de aplicação de conceitos de pesquisa multimédia a tarefas reais do quotidiano. Editores jornalísticos que preparam artigos noticiosos, "bloggers" que publicam conteúdos nos seus sites pessoais, ou criadores de histórias educativas para crianças, são exemplos de utilizadores que podem beneficiar de técnicas de ilustração de texto disponibilizadas por sistemas de recuperação multimédia. Sendo uma tarefa de interatividade acentuada, a sua execução está ligada à obtenção de resultados de pesquisas em tempo aceitável.

A combinação de descritores avançados de análise de conteúdo com a análise dos metadados disponíveis, e o uso de algoritmos de pesquisa aproximada tornaram possível tirar partido de coleções de dados com milhões de itens multimédia, exigindo poucos recursos de hardware. A utilização de metadados para determinar o contexto e a ordenação de imagens por conteúdo permitiu encontrar rapidamente grupos de fotos similares representativas do mesmo evento, ou identificar imagens visualmente distintas para destacar peças jornalísticas ou entradas em blogs.

A aplicação de filtros de abstração visual reduziu substancialmente as necessidades de armazenamento e a quantidade de informação processada, removendo detalhe visual redundante face aos descritores adotados. A análise de conteúdo permitiu também a pesquisa exploratória através do uso de fotos-exemplo, de forma a descobrir em tempo real imagens semelhantes espalhadas por coleções extensas, ou encontrar "duplicados aproximados", isto é, fotos do mesmo evento captadas de ângulos diferentes, por exemplo.

A abordagem foi posteriormente adaptada ao domínio da pesquisa de informação musical, nas tarefas de geração de *playlists* e exploração de coleções musicais de grande escala. O impacto dos filtros de abstração visual foi também analisado no processo de reconhecimento facial de personalidades.

A elaboração de testes seguindo as perspetivas técnica e de satisfação dos utilizadores permitiu uma avaliação realista e um refinamento das metodologias de recuperação propostas. Os sistemas podem ser avaliados considerando a sua capacidade de análise de grandes quantidades de informação, e a qualidade dos resultados produzidos por essas mesmas pesquisas. Enquanto que o primeiro ponto pode ser determinado de forma objetiva, tendo obtido assim maior ênfase neste trabalho, o segundo ponto depende de necessidades de informação específicas e expetativas relativamente aos resultados, que variam entre utilizadores.

Um estudo suportado por *crowdsourcing* demonstrou que uma percentagem considerável considerou que a reordenação dos resultados apresentados com base no conteúdo visual permitiu obter fotos mais adequadas do que as pesquisas tradicionais baseadas apenas em metadados. Testes efetuados na coleção IRMA-2007, com 10 mil radiografias, na coleção MIRFlickr-25k, com 25 mil fotos recolhidas do Flickr, e na coleção de fotos jornalísticas SAPO-Lusa, contendo mais de 1 milhão de fotos e legendas associadas, demonstraram o bom desempenho das estratégias adotadas.

Os resultados obtidos na coleção Million Song Dataset, com informação contextual e de conteúdo referente a 1 milhão de músicas, demonstrou a transversalidade da abordagem apresentada obtendo resultados válidos em tempo aceitável, mesmo perante a dimensão dos dados analisados. A utilização dos filtros de abstração visual no reconhecimento de personalidades reduziu significativamente o espaço necessário

para o armazenamento das fotos, sem afetar negativamente o desempenho já obtido pelos algoritmos de reconhecimento facial analisados.

Abstract

We currently live in the era of "information overload". Professionals and consumers have access to storage and media devices capable of capturing huge amounts of relevant information. But these developments present new questions about the usefulness of massive collections of pictures, audio, video and text with current search engines capabilities.

Hence, this "explosion" of multimedia content constantly being produced in several areas, such as entertainment, journalism, medicine, personal use, among others, requires proper means of search and retrieval. It then becomes necessary to analyze advances in key areas of multimedia information retrieval and large-scale processing, in order to determine the viability of applying and combining several of them for content recommendation and discovery .

The illustration of textual content was the chosen area for applying multimedia retrieval concepts to real-world user tasks. Journalists preparing news articles, bloggers reporting stories on their sites, users improving travel or holiday stories with photos taken during their trips and educational stories for children are some examples that can benefit from automatic text illustration techniques in multimedia retrieval systems. Being a user-interactive task, its execution is connected to obtaining search results in an acceptable timeframe.

The combination of state of the art content-based visual descriptors with available metadata, and the use of approximate search algorithms allows us to take advantage of multimedia datasets with millions of resources with just a minimum of hardware resources. Using metadata to determine context and ranking images by content enables users to quickly find groups of similar photos belonging to the same event, or identify visually distinguishing pictures to highlight news stories or blog posts.

Applying visual abstraction filters substantially reduced storage needs and the amount of data to be processed, by removing redundant visual detail for the adopted descriptors. Content-based analysis can also support exploratory search by using example photos to discover similar images across entire collections in near-realtime, to detect near-duplicates such as photos of the same event from different angles, for example.

This approach was later demonstrated in the music information retrieval area, for generating playlists and exploring large collections. The impact of visual abstraction filters was also analyzed in the facial recognition of public figures.

Testing the system from technical and user satisfaction perspectives allows for a more realistic evaluation and better refinement of the proposed retrieval methodology. Systems can be evaluated considering the ability to quickly analyze large amounts of information and the quality of its results. While the former can be objectively evaluated, and therefore the focus of the evaluation process, the latter depends on specific information needs and results expectancies that may change from person to person, and are deeply associated with the intended retrieval tasks.

A user study supported by crowdsourcing shown that a considerable percentage of users found visually reranked results to better illustrate text than traditional metadata-based searches. Tests performed on the IRMA-2007 collection, with 10 thousand medical images, the MIRFlickr-25k collection, with 25 thousand photos crawled from Flickr, and the SAPO-Lusa news photo collection, containing more than 1 million photos with captions, demonstrated the good performance of the presented approaches.

The results obtained in The Million Song Dataset, with context and content information about 1 million songs, demonstrated the versatility of the presented approach, obtaining valid results in an acceptable timeframe, even considering the scale of the analyzed data. Using the visual abstraction filters in the facial recognition process significantly reduced the photo storage requirements, without interfering with the results obtained with the adopted algorithms.

Agradecimentos

Primeiramente, gostaria de agradecer à minha orientadora, Cristina Ribeiro, pela supervisão prestada ao longo de todo o processo, como a concretização das tarefas de investigação, o foco na área de ilustração de textos e a obtenção das coleções necessárias à avaliação e demonstração do trabalho efetuado.

Gostaria também de agradecer aos membros da comissão de acompanhamento, Eduarda Rodrigues e Nuno Vasconcelos, pela supervisão da elaboração da tese, sugestões e críticas construtivas.

Aos meus colegas do Laboratório SAPO / U.Porto, José Devezas e Pedro Pontes, obrigado pelo apoio no desenvolvimento dos sistemas de recomendação de fotos e exploração visual, ramificações nas áreas de pesquisa de informação musical e reconhecimento facial, com as respetivas demonstrações.

Aos meus colegas do INESC Porto, Igor Amaral e Jaime Cardoso, obrigado pelo acompanhamento na fase inicial de investigação efetuada na área de recuperação de informação visual médica.

Gostaria de destacar as instituições ligadas ao trabalho desenvolvido, nomeadamente o INESC Porto como instituição de acolhimento, e o Laboratório SAPO / U.Porto pelo acesso à coleção SAPO-Lusa e financiamento dos projetos Dpikt, Juggle e Visage.

Finalmente, gostaria sobretudo de agradecer à minha família, em especial à minha esposa Joana, aos meus pais Jaime e Fernanda, ao meu irmão Pedro e à minha avó Fernanda por todo o amor, carinho e compreensão nesta jornada. Que todo e qualquer reconhecimento prestado ao trabalho efetuado possa recompensar-vos, *Ele* melhor do que ninguém!

Este trabalho foi financiado pela *'Fundação para a Ciência e a Tecnologia'* (FCT) através da Bolsa de Doutoramento com a referência SFRH/BD/45590/2008.

Conteúdo

CONTEÚDO

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

> "Progress isn't made by early risers. It's made by lazy men trying to find easier ways to do something."
>
> *Robert Heinlein*

Este capítulo tem como objetivo apresentar uma visão geral sobre a área de recuperação de informação, com ênfase nos tópicos multimédia. A ilustração de conteúdos textuais com o auxílio de grandes coleções de imagens é descrita sob as perspetivas de desempenho e satisfação de resultados. A enunciação da tese foca-se sobretudo nas abordagens adotadas para a resolução desta tarefa, sendo seguida por um sumário da estrutura do documento apresentado.

1.1 Motivação

Pesquisar grandes coleções de dados e encontrar a informação necessária é hoje uma atividade essencial do nosso quotidiano. Desde a pesquisa efetuada sobre o histórico de navegação ou caixa de correio eletrónico até aos motores de pesquisa web avançados de que dispomos atualmente, houve um desenvolvimento gradual de ferramentas especificamente desenhadas para permitir responder às mais variadas necessidades de informação. A investigação na área de recuperação de informação textual tem sido constante, produzindo resultados significativos. As palavras e frases representam um recurso rico em informação semântica intrínseca, dado que a escrita

está fortemente ligada às estruturas do pensamento humano.

No entanto, atualmente a informação disponível não é apenas textual. Os conteúdos multimédia representam hoje uma parte significativa da informação armazenada em vários domínios como agências noticiosas e bases de dados médicas. O aparecimento e crescente importância das comunidades web e redes sociais contribuiu também para a necessidade de partilhar um número cada vez mais significativo de recursos multimédia, não se restringindo ao envio de mensagens de texto. O aumento da capacidade de armazenamento permitiu a criação de enormes coleções multimédia, bases de dados públicas e privadas contendo quantidades consideráveis de informação que requer uma gestão adequada, não só a nível de preservação mas também a nível de consulta e exploração.

A criação e consumo de recursos multimédia tem sido significativo ao longo das últimas décadas. Áreas como o entretenimento, medicina e conteúdos noticiosos, entre outras, deparam-se com uma dificuldade crescente em gerir grandes coleções de dados [LSDJ06]. Os utilizadores podem beneficiar de sistemas avançados capazes de auxiliar ou mesmo executar tarefas repetitivas, sistemas estes que terão de ser capazes de lidar com informação multimédia em larga-escala [DJLW08].

Atualmente, é muito comum visitar e utilizar sítios web desenhados especificamente para grandes quantidades de fotos (Flickr[1], Photo.net[2], Photobucket[3],...) e vídeos (Youtube[4], Vimeo[5], Hulu[6]...). A pesquisa nestes sites é baseada nas anotações manuais ou *tags* que os utilizadores inserem durante o processo de submissão, e na informação adicional que utilizadores registados podem adicionar posteriormente para enriquecer as descrições. Esta informação pode ser restrita a vocabulários específicos do domínio, ou texto livre sem qualquer tipo de limitação. A mesma estratégia pode ser usada em coleções pessoais ou grandes bases de dados empresariais, por exemplo.

Os algoritmos e estratégias de recuperação de informação textual podem ser adaptados à pesquisa de recursos multimédia. Os dados relativos a fotos e vídeos, *tags* inseridas por utilizadores, e texto que rodeia imagens em documentos web são

[1] `http://www.flickr.com`
[2] `http://photo.net`
[3] `http://photobucket.com`
[4] `http://www.youtube.com`
[5] `http://www.vimeo.com/`
[6] `http://www.hulu.com`

algumas das fontes de informação usadas para indexar e pesquisar estes conteúdos multimédia, permitindo a sua pesquisa com base em interrogações textuais.

Uma estratégia alternativa designa-se por recuperação baseada no conteúdo, permitindo interrogações focadas em semelhança. Por exemplo, pesquisas com base em exemplos, como rascunhos (no caso de imagens) ou assobios (músicas), são algumas das possibilidades com análise de conteúdo. No entanto, os requisitos de processamento poderão ser mais elevados quando comparados com a pesquisa de texto tradicional, tornando-se assim um dos maiores obstáculos na adoção deste tipo de pesquisa. Outro obstáculo está relacionado com o facto de que, na ausência de anotações textuais, as características de baixo nível extraídas dos conteúdos multimédia (como cor, textura e ritmo) poderão não estar diretamente ligadas aos conceitos de alto nível presentes nesses mesmos recursos [SWS+00].

Uma das alterações mais recentes na forma como a recuperação de informação multimédia é abordada está relacionada com uma mudança de foco. A investigação foi durante largos anos focada na perspetiva de sistema, assumindo que eventualmente a análise de conteúdo seria suficiente e que os programas mais complexos seriam capazes de reconhecer conteúdos audiovisuais da mesma forma que os utilizadores comuns, permitindo substituí-los e retirá-los dos processos associados às tarefas pretendidas. No entanto, investigação mais recente tem analisado com maior ênfase o *feedback* dos utilizadores nos processos de recuperação de informação, considerando os sistemas como um meio para a melhoria das capacidades de pesquisa na exploração e visualização dos conteúdos multimédia. A recolha de *feedback* dos utilizadores representa assim uma fonte de validação e indicação de melhoramentos que não poderia ser obtida através da perspetiva inicial. Torna-se necessário conceber e desenvolver meios adequados à extração de características, indexação, exploração e visualização em tempo-real para suportar sistemas interativos com interfaces intuitivas e exploratórias, respeitando tempos de resposta aceitáveis.

Estamos habituados a exprimir as nossas interrogações utilizando palavras em vez de imagens ou excertos musicais, principalmente porque os sistemas de pesquisa atuais foram concebidos para a inserção de texto e não para a apresentação de exemplos. Investigação recente focou-se na relação existente entre descrições textuais e características visuais de forma a melhorar a anotação automática de fotos e descoberta de tópicos. Através de uma análise estatística complexa, as palavras textuais

e características visuais co-ocorrentes podem ser associadas e agrupadas para encontrar conceitos semânticos latentes em bases de dados multimédia. Avanços nesta linha de investigação têm permitido melhoramentos na propagação de *tags* em coleções de larga-escala; na depuração de dicionários, eliminando palavras frequentes e irrelevantes; na pesquisa de imagens inicialmente sem qualquer tipo de descrições ou informação pesquisável por texto; e na descoberta de novas ligações e "comunidades" de recursos multimédia que não seriam visíveis através de um agrupamento baseado apenas nos metadados existentes. A análise de conteúdo multimédia permite também considerar cenários de fusão de pesquisas, onde resultados textuais e visuais são combinados de forma a obter novas respostas passíveis de conter elementos mais interligados e diversos do que as respostas baseadas apenas num único tipo de informação.

Apesar da investigação já efetuada na área de pesquisa por conteúdo ao longo dos últimos anos, ainda não existem sistemas capazes de compreender os recursos multimédia da mesma forma que um indivíduo comum. No entanto, surgiram desenvolvimentos recentes focados na estratégia de recuperação transversal de informação [RCPC$^+$10], onde texto e imagem são inicialmente correlacionados de forma a produzir resultados diferentes dos obtidos com fusões de listas de resultados unimodais em fases posteriores do processo de pesquisa.

Os avanços nas capacidades de armazenamento permitem a existência de coleções multimédia massivas, com milhões de documentos. O acesso a esta informação permite a execução de tarefas que de outra forma poderiam ser extremamente morosas ou até mesmo impossíveis caso fossem executadas de forma manual [DJLW08]. A tarefa de geração de cenas, onde as fotos são enriquecidas com partes de outras imagens, requer a análise de secções visuais muito semelhantes existentes em outras imagens para substituição na original [HE07]. Dada uma coleção suficientemente grande de fotos e algoritmos de similaridade adequados, esta tarefa torna-se acessível já que a probabilidade de encontrar segmentos de imagem convincentes aumenta com o número de recursos disponíveis e com a diversidade de escolha.

Duas tarefas quotidianas importantes que se baseiam em conceitos multimédia são designadas por anotação de imagens e ilustração de texto (esta última por vezes designada como "ilustração de histórias"). Enquanto que a primeira tarefa está relacionada com a intenção de encontrar descrições textuais adequadas para o conteúdo

de cada imagem, a segunda tarefa foca-se no problema inverso, isto é, dado um documento de texto, pretende-se encontrar uma ou mais imagens adequadas para o enriquecer. Este enriquecimento de informação poderá ser obtido através do enquadramento visual das fotos, cativando o leitor, e pelo reforço exemplificativo das ações descritas no texto, facilitando a sua interpretação. Ambas as tarefas dependem de uma extração e descrição corretas dos conteúdos multimédia, acompanhadas por métodos adequados de indexação, determinação de similaridade e correlação entre conteúdos textuais e visuais.

A tarefa de ilustração automática de conteúdo textual é condicionada pela resolução de dois grandes desafios. O primeiro está relacionado com a pesquisa de grandes coleções de conteúdos multimédia, de forma a produzir bons resultados em domínios generalizados. Um cenário exemplificativo é a ilustração de textos jornalísticos, onde os eventos relatados nas notícias podem abranger um conjunto extremamente diversificado de tópicos como política, guerra, temas sobre a educação ou entretenimento. O segundo desafio, ainda mais significativo, é a necessidade de ultrapassar o fosso semântico [SWS+00], a separação existente entre os conceitos de alto-nível da mente humana e as características de baixo-nível extraídas das imagens, como cor e textura.

Jornalistas que preparam os seus artigos noticiosos, "bloggers" que colocam diariamente conteúdo nos seus sítios web pessoais, famílias que gerem coleções de fotos das viagens e férias, a preparação de histórias educativas para crianças ou ilustrativas para idosos são apenas algumas das instâncias da tarefa apresentada que podem beneficiar substancialmente de desenvolvimentos nas técnicas de ilustração automática de textos com recurso a sistemas de recuperação de informação multimédia. Além disso, considerando a existência do fosso semântico e o facto de que o desempenho destes sistemas deve ser também avaliado pelos utilizadores finais, uma abordagem centrada nos utilizadores torna-se importante na medida em que beneficia das capacidades destes para guiar o sistema.

Joshi et al. [JWL06] apresentaram abordagens não-supervisionadas para a ilustração automática de textos, onde palavras-chave são extraídas das descrições das fotos e usadas para a pesquisa sobre uma pequena base de dados de imagens, sendo utilizado posteriormente um esquema de ordenamento de imagens baseado em características visuais elementares extraídas paralelamente, aquando da indexação dos conteúdos multimédia. Delgado et al. [DMC10] abordaram também a tarefa de se-

leção de imagens em função de um texto apresentado, mas a sua vertente de investigação foca-se na ilustração de histórias. Nesta variante da tarefa, o objetivo é descrever da forma mais aproximada e clara possíveis as ações e fluxo do conteúdo textual para aumentar a compreensão visual do contexto, sendo este um objetivo relevante na elaboração de histórias infantis e na apresentação de notícias a utilizadores idosos com necessidades específicas de aquisição de conhecimento.

1.2 Tese

A tarefa de ilustração de textos baseada em recuperação de informação multimédia, envolvendo interrogações textuais semanticamente ricas e respostas compostas por conjuntos contextualmente relevantes de imagens como desenhos simplificados e fotos de alta qualidade, pode ser executada com tempos de resposta adequados e recursos de hardware bem dimensionados através de um processo de pesquisa transversal usando algoritmos avançados de extração, descrição, indexação e pesquisa de recursos multimédia orientados ao conteúdo.

O trabalho apresentado segue a abordagem principal de ilustração de textos, isto é, o propósito de auxiliar os criadores de conteúdos a enriquecê-los com os recursos multimédia disponíveis, não havendo uma necessidade específica de melhorar a sua compreensão ou criar redundância entre o conteúdo textual e o visual. O sistema desenvolvido tira partido de descritores avançados de conteúdo visual, métodos de indexação para pesquisa aproximada, e reordenação de resultados com base em recuperação transversal de informação multimédia em coleções de larga-escala. Este sistema foi desenvolvido ao longo de todo o processo de investigação, passando por várias iterações que resultaram em publicações e demonstrações [CR10, CR11c, DCNR12, CDR13].

O processo de ilustração de textos é exemplificado na Figura 1.1. O utilizador fornece ao sistema fragmentos de texto e obtém como resposta uma lista de imagens recomendadas para ilustrar esses mesmos excertos. Não se limitando a uma simples resposta, o processo permite ainda um refinamento interativo dos resultados, quer através da reordenação dos mesmos com base nas suas características visuais, quer na pesquisa exploratória de toda a coleção usando o conteúdo visual de uma das

fotos como exemplo.

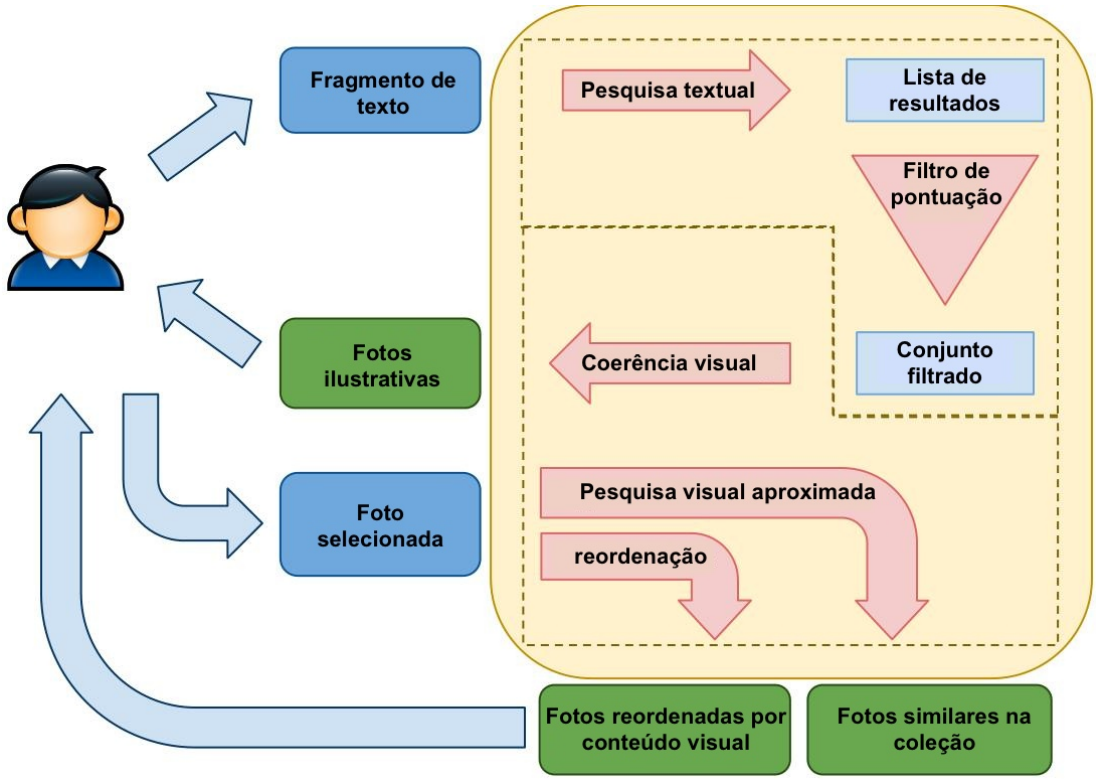


Figura 1.1: O processo de ilustração de textos

1.3 Metodologia

A coleção multimédia principal utilizada para a investigação desenvolvida, nomeadamente a coleção fotojornalística SAPO-Lusa, contendo 1.5 milhões de fotos e respetivas descrições fornecidas pela Agência Lusa ao SAPO, é regularmente pesquisada por jornalistas e pelo público em geral. Tirando partido do acesso a esta coleção representativa de um caso de uso válido para a tarefa abordada, foi concebido um sistema de ilustração de texto capaz de lidar com coleções de larga-escala e seguindo metodologias de desenvolvimento e avaliação focados nas perspetivas de desempenho e qualidade dos resultados produzidos.

Do ponto de vista técnico, a escolha de algoritmos efetuada, a combinação de métodos de recuperação transversal de informação e a integração de tecnologias de

estado da arte permitiram lidar com os pontos críticos de escalabilidade e desempenho apresentados pela tarefa e coleção de dados adotada. Do ponto de vista dos utilizadores, a satisfação com os resultados obtidos foi avaliada através de *ground-truth* e *crowdsourcing*, permitindo validar as metodologias e abordagens adotadas para a resolução da tarefa no cenário de utilização especificado.

Primeiramente foi desenvolvida uma plataforma de avaliação para a realização de experiências e testes comparativos em coleções de imagens médicas e fotos genéricas. Mais especificamente, foram testados descritores visuais globais na coleção IRMA-2007 (11.000 radiografias) e na coleção MIRFlickr-25k (25.000 fotos). A utilização de coleções pertencentes a domínios tão distintos permitiu observar a informação extraída pelos descritores, a sua adequação e desempenho face a imagens com propriedades visuais distintas, e o impacto da combinação de informação para determinação da similaridade entre imagens. Os descritores globais revelaram um desempenho superior na coleção médica, dado o ambiente controlado: fundo neutro, objetos de interesse centrados na imagem e características visuais distintas entre classes. Na coleção de fotos genérica, o desempenho dos descritores manteve-se aceitável apesar do ambiente não controlado em que foram obtidas. Foi possível obter imagens visualmente similares em contextos distintos, dado que a análise global de conteúdo visual não considera informação textual semântica ou de reconhecimento visual de objetos. Dado que se pretende a recuperação de informação multimédia em larga-escala sem a utilização de elevados recursos de hardware, a escolha de descritores globais provou ser a mais acertada, apresentando um compromisso entre desempenho e qualidade dos resultados produzidos.

Baseado nos resultados obtidos no decorrer do trabalho preliminar, assim como na literatura existente sobre o estado da arte em recuperação de informação multimédia, foram escolhidas abordagens específicas para investigação e integração nos protótipos desenvolvidos. O foco no problema específico de ilustração automática de textos jornalísticos e o acesso à coleção SAPO-Lusa permitiu a análise e exploração de uma coleção multimédia de larga-escala com a respetiva extração, descrição e indexação da informação disponível através de métodos avançados.

Os protótipos desenvolvidos permitiram uma recolha de *feedback* sobre o desempenho e qualidade dos resultados, validação das funcionalidades disponibilizadas e tempos de resposta pretendidos. Este *feedback* foi essencial para validar um aspeto

crucial da tese aqui apresentada, nomeadamente a influência positiva da informação visual no processo de ilustração de texto face à abordagem tradicional baseada apenas nos metadados e conteúdo textual disponível em legendas ou *tags*. Ao longo deste processo, os protótipos processaram coleções de tamanho crescente, permitindo assim um acompanhamento da evolução dos aspetos de larga-escala e de respostas em tempo aceitável, de forma a permitir a obtenção de um sistema de ilustração de texto capaz de estabelecer rapidamente a similaridade entre milhões de recursos multimédia.

O processo de avaliação de resultados é extremamente importante nas metodologias de recuperação de informação multimédia, dada a natureza imprecisa das tarefas. Representa assim um problema complexo perante a necessidade de garantir tempos de execução adequados e uma utilização regrada dos recursos de hardware disponíveis (avaliação de desempenho), bem como a satisfação das necessidades de informação específicas de cada utilizador, o que evidencia uma avaliação personalizada e que considere o utilizador como iniciador da interrogação e indicador de relevância dos resultados gerados (avaliação qualitativa).

Na tarefa de ilustração de conteúdos textuais, o utilizador desempenha um papel importante na validação dos sistemas de recuperação de informação multimédia. A metodologia apresentada contempla este aspeto, traduzindo-se na realização de tarefas de avaliação por *crowdsourcing* e análise dos tempos de resposta médios das várias funcionalidades disponibilizadas aos utilizadores, as quais permitiram determinar e melhorar o desempenho e usabilidade dos protótipos.

O trabalho desenvolvido foi apresentado e discutido em várias conferências e workshops, com ênfase na demonstração de funcionalidades de pesquisa multimédia e avaliação de resultados com base no *feedback* dos utilizadores, preparação e manutenção de *groundtruth*. Algumas das coleções utilizadas pertencem ao domínio público, o que permitiu comparações diretas e indiretas dos avanços alcançados ao longo da investigação efetuada.

O desenvolvimento foi efetuado de forma iterativa, com acompanhamento constante para garantir que os objetivos estabelecidos seriam alcançados. Os pontos de situação mais críticos, após o desenvolvimento de cada protótipo, validaram as afirmações que compõem a tese apresentada e permitiram a análise do feedback obtido

até ao momento, estabelecendo prioridades para as fases de investigação e desenvolvimento subsequentes.

Os protótipos iniciais foram alvo de uma avaliação essencialmente focada na comparação dos diferentes descritores e estratégias de pesquisa associadas. Após o acesso e utilização de coleções multimédia de maior escala, foi possível determinar o comportamento dos algoritmos adotados e respetivos melhoramentos a nível dos tempos de resposta, com a possibilidade de efetuar um elevado número de pesquisas paralelas. O impacto do recurso a técnicas de indexação para pesquisa aproximada foi também observado, dado que através desta abordagem são efetuadas apenas comparações de similaridade com os candidatos mais prováveis. Com base nos desempenhos obtidos, foi explorada também a vertente de exploração de coleções de larga-escala com base em imagens-exemplo. Desta forma, o processo sequencial de ilustração de fragmentos de texto seguido da reordenação dos resultados ou exploração interativa da coleção disponibiliza aos utilizadores uma forma mais rica de execução da tarefa abordada.

A avaliação de desempenho do sistema contempla a componente técnica da plataforma desenvolvida. Esta faceta da avaliação considera não apenas os tempos de resposta das interrogações lançadas, mas também as necessidades de armazenamento e processamento associado à indexação e pesquisa da informação. Existem vários pontos a considerar em todo o processo de recuperação de informação, de forma a disponibilizar a capacidade de pesquisa de coleções de larga-escala em tempo útil. Esta avaliação permite ainda validar as escolhas tecnológicas efetuadas e determinar a complexidade real dos algoritmos considerados.

A avaliação dos resultados obtidos considera a validação face à existência de *groundtruth*, e ao *feedback* dos utilizadores. Os testes efetuados permitiram recolher dados relativamente à qualidade dos resultados face a necessidades de informação específicas, bem como orientar a interface de utilizador concebida e as funcionalidades disponibilizadas. Foi considerado um número elevado de avaliadores no teste de *crowdsourcing*, teste esse projetado para ser simples, intuitivo e isento, não fornecendo pistas que revelassem os algoritmos testados.

Posteriormente, os conceitos definidos e abordagens adotadas foram aplicados ao domínio da recuperação de informação musical, nas tarefas de criação de *playlists* e

exploração de grandes coleções de conteúdo áudio. Desta forma, foi possível validar a transversalidade das técnicas propostas e demonstrar a aplicabilidade do processo de recomendação e exploração de conteúdo multimédia a domínios distintos.

1.4 Contribuições e publicações

As principais contribuições deste trabalho incidem sobretudo na componente de pesquisa em larga-escala usando recursos limitados, isto é, quando a dimensão das coleções utilizadas ultrapassa significativamente a capacidade de processamento e pesquisa em série mantendo tempos de resposta adequados às tarefas em questão:

- Apresentação de uma abordagem de recuperação de informação multimédia adequada ao processamento, gestão e pesquisa da informação em larga-escala usando recursos limitados, para a execução de tarefas de ilustração de textos e pesquisa interativa de coleções de imagens;
- Demonstração da transversalidade da abordagem a domínios visualmente distintos, como imagens médicas, fotojornalísticas ou de âmbito geral, bem como tarefas usando recursos distintos, como geração de *playlists* com base no conteúdo áudio das músicas;
- Análise do impacto da aplicação de filtros de abstração, diminuindo a informação necessária para a determinação de similaridade entre imagens, o que contribui para a redução das necessidades de armazenamento e processamento das coleções de larga-escala a pesquisar.

1.4.1 Publicações

O trabalho desenvolvido traduziu-se em várias publicações em conferências e workshops estabelecidas na área de recuperação de informação multimédia, com contribuições na pesquisa em larga-escala, ilustração de texto e avaliação com utilizadores:

- ***Impacto dos Filtros de Abstração no Reconhecimento Facial em Imagens (Pedro Pontes, Filipe Coelho e Cristina Ribeiro, Simpósio INFORUM, 2013)***: O reconhecimento facial em imagens constitui uma área de investigação em aberto, principalmente se considerarmos situações de

captura de imagens não-controladas. Os filtros de abstração atuam como ferramentas de remoção de informação redundante existente nas imagens. Foi desenvolvido um sistema de reconhecimento facial de personalidades, baseado em código aberto, onde é utilizada a abstração de imagens juntamente com tarefas de pré-processamento paralelas, de forma a analisar o seu impacto no processo de reconhecimento. A avaliação foi efetuada com recurso à coleção de imagens Labeled Faces in the Wild, sob duas perspetivas, Closed-set Identification e Image Retrieval, e utilizando nove cadeias de pré-processamento de imagens distintas. Os resultados demonstram que a aplicação de filtros de abstração no processo de reconhecimento resulta no compromisso entre a diminuição dos requisitos de armazenamento das imagens e a ligeira redução da eficácia da identificação. A deteção e segmentação das faces presentes nas imagens revelou ser a etapa de pré-processamento com maior importância para um reconhecimento eficaz. O desempenho foi avaliado através dos algoritmos Eigenfaces, Fisherfaces e Local Binary Patterns Histograms, tendo o último revelado o melhor desempenho em termos globais.

- ***Juggle: Large-scale Discovery in Music Recommendation (Filipe Coelho, José Devezas e Cristina Ribeiro, International Conference on Open research Areas in Information Retrieval, 2013)***: Today's offer of audio content exceeds the human capability of manually searching datasets with hundreds of songs, demanding automated tools capable of handling music recommendation when faced with large-scale collections. In this work, we address the playlist generation and song discovery tasks with large-scale datasets. It is possible to quickly obtain playlists and explore collections with example-based queries using audio features, lyrics and tags. We developed a music discovery prototype to demonstrate this content based approach. This demo is based on the Million Song Dataset, a large-scale collection of audio features and associated text data comprising almost 300 GB of information.

- ***Large-scale Crossmedia Retrieval for Playlist Generation and Song Discovery (Filipe Coelho, José Devezas and Cristina Ribeiro, International Conference on Open research Areas in Information Retrieval, 2013)***: Today's offer of audio content exceeds the human capability of manually searching datasets of hundreds of songs, demanding automated tools to aid in music recommendation when faced with large-scale collections. Collaborative filtering recommenders rely on user feedback, facing limitations

when there is a lack of users or a bias for certain popular genres, enclosing users in an information bubble. Content-based analysis is a reliable source of audio similarity, as seen in music classification, but proper indexing and retrieval methods are required to effectively use it in highly interactive tasks. In this work, we address the playlist generation and song discovery tasks with large-scale datasets. Using audio features, lyrics and tags, it is possible to quickly generate playlists and explore collections with example-based queries. We demonstrate the use of approximate indexing and crossmedia reranking for creating playlists, as well as mapping content-based similarity to textual representations that can be handled by information retrieval libraries. We explored the feasibility of this content-based approach in the Million Song Dataset, a large-scale collection of audio features and associated text data comprising almost 300GB of information. The proposed strategy can be used independently as a content-based music retrieval system, and as a component for hybrid recommender systems.

- ***Image Abstraction in Crossmedia Retrieval for Text Illustration (Filipe Coelho and Cristina Ribeiro, European Conference on Information Retrieval, 2012)***: Text illustration is a multimedia retrieval task that consists in finding suitable images to illustrate text fragments such as blog entries, news reports or children stories. In this paper we describe a crossmedia retrieval system which, given a textual input, selects a short list of candidate images from a large media collection. This approach makes use of a recently proposed method to map metadata and visual features into a common textual representation that can be handled by traditional information retrieval engines. Content-based analysis is enhanced by visual abstraction, namely the Anisotropic Kuwahara Filter, which impacts feature information captured by the Joint Composite and Speeded Up Robust Features visual descriptors. For evaluation purposes, we used the well-established MIRFlickr photo collection, with 25,000 photos and user tags collected from Flickr as well as manual annotations provided as image retrieval groundtruth. Results show that image abstraction can improve visual retrieval as well as significantly reduce processing and storage requirements, even more when paired with Google's WebP image format. We conclude that applying a visual rerank after an initial text retrieval step improves the quality of results, and that the adopted text mapping method for visual descriptors provides an effective crossmedia approach for text illustration.

- ***Studying a Personality Coreference Network in a News Stories Photo Collection (José Devezas, Filipe Coelho, Sérgio Nunes and Cristina Ribeiro, European Conference on Information Retrieval, 2012)***: We build and analyze a co-reference network based on entities from photo descriptions, where nodes represent personalities and edges connect people mentioned in the same photo description. We identify and characterize the communities in this network and propose taking advantage of the context provided by community detection methodologies to improve text illustration and general search.

- ***Automatic Illustration with Crossmedia Retrieval in Large scale Collections (Filipe Coelho and Cristina Ribeiro, International Conference on Content Based Multimedia Indexing, 2011)***: In this paper, we approach the task of finding suitable images to illustrate text, from specific news stories to more generic blog entries. We have developed an automatic illustration system supported by multimedia information retrieval, that analyzes text and presents a list of candidate images to illustrate it. The system was tested on the SAPO-Lusa media collection, containing almost two million images with short descriptions, and the MIRFlickr-25k collection, with photos and user tags from Flickr. Visual content is described by the Joint Composite Descriptor and indexed by a Permutation-Prefix Index. Illustration is a three-stage process using textual search, score filtering and visual clustering. A preliminary evaluation using exhaustive and approximate visual searches demonstrates the capabilities of the visual descriptor and approximate indexing scheme used.

- ***Dpikt - Automatic Illustration System for Media Content (Filipe Coelho and Cristina Ribeiro, International Workshop on Content-Based Multimedia Indexing, 2011)***: Journalists and bloggers need to find useful images to illustrate news stories and blog entries with high quality photos. The dpikt text illustration system uses multimedia information retrieval to assist this content enrichment task. Users query the system with text fragments and get collections of candidate photos. Images in the results can be visually sorted according to a selected photo, or be used as a seed for interactive searches over the entire collection. Dpikt incorporates a recent visual descriptor, the Joint Composite Descriptor, and an approximate indexing

scheme designed for large-scale image collections, the Permutation-Prefix Index. We have used the SAPO-Lusa large-scale news stories photo collection, containing almost two million high quality photos with short descriptions, as the resource for the illustration task.

- ***Characterization of the SAPO-Lusa news stories photo collection (Filipe Coelho and Cristina Ribeiro, Technical Report, 2011)***: The SAPO-Lusa news stories photo collection is a multimedia collection containing approximately 1.5 million journalistic photos accompanied by short news descriptions. It covers 85 years of portuguese photojournalism and provides a broad overview of national and international events. The database contains medium-size photos, thumbnails, and all the available metadata, including the news descriptions, titles and manually added tags. The available resources are valuable for multimedia retrieval tasks such as automatic text illustration, cross-media retrieval and content based search.

- ***Hierarchical Medical Image Annotation Using SVM-based Approaches (Igor Amaral, Filipe Coelho, Joaquim Costa and Jaime Cardoso, International Conference on Information Technology and Applications in Biomedicine, 2010)***: Automatic image annotation or image classification can be an important step when searching for images from a database. Common approaches to medical image annotation with the Image Retrieval for Medical Applications (IRMA) code make poor or no use of its hierarchical nature, where different dense sampled pixel based information methods outperform global image descriptors. In this work we address the problem of hierarchical medical image annotation by building a Content Based Image Retrieval (CBIR) system aiming to explore the combination of three different methods using Support Vector Machines (SVMs): first we concatenate global image descriptors with an interest points Bag-of-Words (BoW) to build a feature vector; second, we perform an initial annotation of the data using two known methods, disregarding the hierarchy of the IRMA code, and a third that takes the hierarchy into consideration by classifying consecutively its instances; finally, we make use of pairwise majority voting between methods by simply summing strings in order to produce a final annotation. Our results show that although almost all fusion methods result in an improvement over standalone classifications, none clearly outperforms each other. Nevertheless,

these are quite competitive when compared with related works using an identical database.

- ***Evaluation of Global Descriptors for Multimedia Retrieval in Medical Applications (Filipe Coelho and Cristina Ribeiro, International Conference on Database and Expert Systems Applications, 2010)***: In this paper, global descriptors from MPEG7, GIST and Compact Composite Descriptors are evaluated for image retrieval in the IRMA-2007 medical collection. This evaluation tests descriptors using every image from each class instead of a small group of representative images. The evaluation results obtained by Mean-Average Precision (MAP) and precision@N indicate that MPEG7 EH, GIST and Fuzzy BTDH outperform the other global descriptors analyzed by a large margin, even more when combined by late-fusion rank aggregation. A multimedia retrieval evaluation system was developed to support the experiment and offers the possibility of textual, visual and combined searches over the medical collection.

- ***Temporal Analysis of Terms in Blogs (Filipe Coelho, Doctoral Simposium on Informatics Engineering, 2009)***: Blogs are becoming extremely popular, revealing the most relevant topics for their social communities on a daily basis. The work presented here has focused on the temporal analysis of terms usage in blogs, specifically the Portuguese SAPO Blogs collection, to find the most relevant terms occurred during the first half of 2008. The gathered information was stored and processed by means of a data warehouse, which facilitated the necessary calculations for terms analysis by the relevance and interestingness ranking algorithms. Term clouds were used to show the comparison between these algorithms, allowing us to quickly determine that interestingness ranking produced the best results for this collection.

1.5 Estrutura da Tese

O documento apresentado encontra-se organizado em 7 capítulos:

- Introdução (capítulo atual);

- Revisão do estado da arte na área de recuperação de informação multimédia (capítulo 2);

- Extração de informação, incluindo as coleções de dados usadas, descritores multimédia usados, e o processo de abstração de imagens (capítulo 3);
- Recuperação em coleções de larga-escala, com ênfase na pesquisa aproximada e transversal de informação multimédia (capítulo 4);
- Implementação da solução apresentada, com detalhes sobre as estratégias adotadas, desenvolvimento dos protótipos e respetiva avaliação de resultados (capítulo 5);
- Aplicação dos conceitos de pesquisa em larga-escala à área de recuperação de informação musical, com ênfase na recomendação e descoberta de músicas (capítulo 6);
- Conclusões sobre a investigação efetuada, com uma discussão geral sobre as contribuições efetuadas e ramificações (capítulo 7).

Capítulo 2

Uma visão geral sobre recuperação multimédia

> "If you know the enemy and know yourself, you need not fear the result of a hundred battles."
>
> *Sun Tzu, The Art of War*

O objetivo da recuperação de informação multimédia (RIM) é facilitar a captura, armazenamento, pesquisa e utilização de conteúdos digitais no quotidiano [LSDJ06]. Apesar de as pesquisas tradicionais se restringirem a informação apenas textual, os métodos de pesquisa baseados no conteúdo são necessários quando não existem descrições ou metadados, ou mesmo quando estes estão incompletos. Os pressupostos iniciais da investigação nesta área sugeriam que as características de baixo nível diretamente extraídas dos recursos seriam suficientes para aumentar a precisão e relevância dos resultados produzidos e evidenciar aspetos importantes das coleções multimédia processadas. Inteligência artificial, teoria da otimização, visão computacional e reconhecimento de padrões são apenas algumas das áreas que influenciaram significativamente os fundamentos matemáticos usados em recuperação de informação multimédia.

O trabalho inicialmente desenvolvido em RIM baseou-se essencialmente em conceitos de visão computacional; por exemplo, na pesquisa de conteúdos vídeo o foco estava colocado na deteção robusta de limites entre cenas. Perante os resultados

obtidos, rapidamente os investigadores se aperceberam de que a similaridade entre documentos baseada apenas nas características de conteúdo não era suficiente. Houve a necessidade de criar sistemas capazes de compreender ambas as vertentes, nomeadamente a riqueza semântica contida nos dados e metadados textuais e as características audiovisuais extraídas do conteúdo multimédia. A existência do fosso semântico é ainda hoje um obstáculo que envolve investigação contínua na sua resolução. No entanto, os avanços obtidos pela análise de conteúdo permitiram o desenvolvimento de sistemas capazes de determinar a similaridade em vários domínios, como a pesquisa em bases de dados de logótipos e de conteúdos musicais para deteção de infrações de patentes, por exemplo.

Os sistemas de RIM carecem de foco em dois aspetos importantes das necessidades de informação apresentadas pelos utilizadores: a pesquisa transversal por conteúdos interligando os diferentes tipos de média com o texto existente, e a possibilidade de exploração de coleções de larga-escala baseada em exemplos, com respostas em tempo útil e usando recursos de hardware limitados. Infelizmente, os sistemas atuais ainda não são capazes de compreender vocabulários de grandes dimensões criados pelas comunidades de utilizadores (designados por folksonomias), e responder com elevado grau de satisfação às interrogações fornecidas pelos mesmos. Desta forma, alguns tópicos de investigação adquirem uma prioridade elevada de forma a ultrapassar eventualmente o fosso semântico e tornar os sistemas de RIM acessíveis para uso diário pelos consumidores e indústria. Estes tópicos são a computação centrada no utilizador, a exploração de características multimédia de mais alto nível, a análise de novos tipos de média, a pesquisa exploratória e a avaliação de sistemas RIM do ponto de vista não apenas técnico, mas também de satisfação com os resultados obtidos.

2.1 Compilações

Datta et al. elaboraram uma compilação detalhada de artigos [DJLW08] apresentando o estado da arte na recuperação de informação baseada em conteúdo (RIBC). O uso de semânticas de alto nível na pesquisa de imagens é também discutido por Liu et al [LZLM07].

Um painel de investigadores [HLMS08] discutiu a importância da recuperação de informação multimédia e as dificuldades que representam um obstáculo à sua adoção no quotidiano. Vários tópicos foram abordados, principalmente a necessidade de "aplicações de topo", ou "killer applications". Kankanhalli et al.[KR08] analisaram o impacto das aplicações existentes de RIM e as tendências de investigação que estão a influenciar os trabalhos atuais e futuros. Hanjalic et al.[HLMS08] apresentaram uma discussão alargada sobre o facto de a pesquisa multimédia, apesar da sua crescente importância, não ter ainda encontrado a sua verdadeira identidade e propósito na atual sociedade de informação.

Heesch[Hee08] demonstrou vários modelos para a exploração de documentos multimédia baseada no conteúdo, descrevendo métodos tradicionais de recuperação centrada nos metadados assim como métodos inovadores suportados por redes de similaridade NN^k ("nearest neighbors"). Disponibilizou também uma compilação de publicações bastante completa sobre modelos de interação e *feedback* de relevância em pesquisa de conteúdos visuais [HR07]. Kennedy et al.[KCN08] conduziram um inquérito alargado sobre estratégias de pesquisa adaptativas em domínios de aplicação variados, onde os mecanismos internos de recuperação usados para a obtenção de resultados se adaptam em resposta à previsão antecipada das necessidades de informação de cada utilizador, com base nas suas interações prévias com o sistema. Vasconcelos[Vas07] apresenta uma visão sobre a evolução dos sistemas de recuperação, principalmente através da utilização de características de baixo nível e a sua combinação de forma a extrair conhecimento dos documentos analisados, nomeadamente imagens e vídeos.

A recuperação de informação multimédia representa uma área de investigação extremamente diversa e englobando uma variedade considerável de tipos de dados, problemas de investigação e diferentes metodologias de extração, descrição, indexação e pesquisa de resultados. Wang et al.[WBDB$^+$06] discutiram a importância desta mesma diversidade para o crescimento da investigação na área. Jaimes et al.[JCG$^+$05] responderam a várias questões sobre o que define o conceito de pesquisa multimédia, a forma como a recuperação multimédia é diferente dos restantes tipos de recuperação de informação, os desafios técnicos mais significativos na área, as aplicações de topo, oportunidades de investigação e direções futuras para exploração.

2.2 Recuperação textual e visual

Wang et al. [WZZ08] abordaram o problema da existência de um fosso semântico [HLES06] entre as características visuais de baixo nível e os conceitos semânticos de alto nível, um dos obstáculos mais críticos na pesquisa de imagens. Guiada pela informação textual que normalmente acompanha imagens presentes em sítios web, a plataforma proposta tenta adquirir uma medida de distância no espaço visual permitindo posteriormente usar essa medida para encontrar resultados semanticamente similares a uma qualquer imagem fornecida como exemplo. Para atenuar o ruído introduzido pela variedade de *tags* nas imagens, e de forma a utilizar totalmente a informação textual disponível, é introduzido um modelo de texto ao nível dos tópicos baseado em "Latent Dirichlet Allocation" para definir a semelhança semântica entre pares de imagens. A medida de distância adquirida pode ser aplicada em ambos os contextos, isto é, pesquisa de imagens por conteúdo ou anotação de imagens.

Uma formulação probabilística de anotação e pesquisa semântica de imagens foi proposta por Carneiro et al. [CCMV07]. A anotação e recuperação de imagens são abordados como tarefas de classificação, onde cada classe é definido como o conjunto de imagens às quais foi atribuída uma etiqueta semântica comum. Foi demonstrado que, ao estabelecer uma correspondência direta entre etiquetas e classes semânticas, é possível obter um erro mínimo de probabilidade na anotação e recuperação de imagens usando algoritmos conceptualmente simples e computacionalmente eficientes, não exigindo uma segmentação semântica prévia das imagens usadas para treino. Os benefícios de uma formulação supervisionada quando comparada com modelos mais complexos e populares é exemplificada através de argumentação teórica e realização de várias experiências.

Os desafios colocados pelos paradigmas de pesquisa baseada exclusivamente em conteúdo textual e metadados tem inspirado uma investigação contínua na área de recuperação de imagens por conteúdo. Além da necessidade de abordar a interação com os utilizadores, nomeadamente através da especificação de palavras-chave ou imagens exemplo para iniciar pesquisas visuais, os sistemas de recuperação de informação baseados no conteúdo têm de obrigatoriamente lidar com o fosso semântico e com as limitações sensoriais existentes. Veltman et al. [VWN09] reavaliam abordagens tradicionalmente aceites para lidar com estes problemas, e demonstram as limitações inerentes à utilização de apenas uma dimensão dos dados (textual, visual,

...) para a obtenção de resultados satisfatórios.

Stottinger et al. [SBP+09] evidenciam o facto de que tirar partido de estudos com utilizadores com ênfase nos requisitos funcionais pode orientar a seleção de características visuais adequadas para os sistemas de pesquisa desejados. Os autores testaram esta hipótese através de um estudo usado para melhorar um sistema de recuperação de imagens jornalísticas apenas baseado em pesquisa textual. Um dos resultados obtidos através dos comentários dos utilizadores permitiu concluir que estes preferem características visuais compreensíveis e fáceis de especificar pelos próprios jornalistas.

Leuken et al. [vLGOvZ09] propuseram métodos para a diversificação visual dos resultados produzidos por pesquisas visuais. Os algoritmos de recuperação textual procuram garantir a relevância contextual dos resultados, mas a similaridade (ou variedade) visual é necessária para alargar o alcance das pesquisas efetuadas, sobretudo em coleções e larga-escala.

Através da combinação de técnicas de modelação estatísticas recentes com as ontologias existentes no serviço WordNet, Datta et al. [DGLW07] apresentaram uma abordagem promissora para a pesquisa de imagens utilizando um processo de anotação automática de imagens como base de suporte.

Zheng et al. [ZG08] apresentaram uma analogia entre a pesquisa visual e a recuperação de informação textual e propuseram uma abordagem baseada em "frases visuais" para obter imagens contendo objetos específicos, ou seja, para efetuar pesquisa de objetos baseada em conteúdo. As frases visuais são definidas como pares de segmentos locais adjacentes e co-ocorrentes, e construídas utilizando algoritmos de "clustering". Neste trabalho são também apresentados métodos para a construção de frases visuais e respetiva indexação.

Nister et al. [NS06] aplicaram um esquema de reconhecimento de objetos escalável, indexando milhares de objetos. A eficiência e qualidade dos resultados foi visível através de uma demonstração em tempo-real capaz de reconhecer capas de cd's de música num universo de 40.000 imagens de álbuns populares. O esquema apresentado é baseado em técnicas de indexação de descritores locais extraídos de regiões salientes das imagens, tornando-se assim robusto a ruído visual. Os descri-

tores locais são quantizados hierarquicamente numas árvore de vocabulário, a qual permite a utilização eficiente de vocabulários complexos e de grande dimensão. Foi demonstrado que este esquema permite uma melhoria substancial nos resultados produzidos por pesquisas de objetos em imagens. Uma característica importante da árvore está relacionada com o facto de que esta define o nível de quantização. Desta forma, a quantização e indexação estão totalmente integradas, combinando ambos os processos no mesmo algoritmo. A qualidade de reconhecimento foi avaliada numa coleção com um milhão de imagens para as quais já havia sido definido o *groundtruth*.

A investigação realizada na área de recuperação de informação multimédia tem dado origem a novas sub-áreas de pesquisa, sendo uma delas a análise de emoções produzidas pelas imagens, a pesquisa semântica emocional. Wang et al. [WH08] introduziram esta perspetiva emergente à comunidade, apresentando uma visão geral da investigação preliminar e plataformas já desenvolvidas. Neste campo foram discutidos três aspetos cruciais a abordar, nomeadamente a representação semântica das emoções, a extração de características visuais relevantes, e a identificação de emoções com base nas mesmas, sendo propostas algumas abordagens promissoras e respetivos desafios.

A dimensão estética, no contexto da arte e fotografia, refere-se às características de beleza presentes em imagens, sendo que a sua avaliação em todas as perspetivas representa uma tarefa altamente subjetiva. Assim sendo, não existe ainda um padrão unanimemente acordado para a determinação exata do valor estético de uma imagem. Contudo, apesar da inexistência de regras bem definidas, existem algumas características que podem ser definidas objetivamente e que se constatou estarem fortemente ligadas à noção geral de beleza.

Datta et al. [DJLW06] abordaram a tarefa de inferir automaticamente a qualidade estética de fotos como um problema de aprendizagem computacional. Usando como fonte de dados um sítio web de partilha de fotos com avaliação efetuada pelos utilizadores, extraíram características visuais específicas baseando-se na sua própria intuição e assumindo que as mesmas seriam suficientes para identificar fotos esteticamente agradáveis. Foram construídos classificadores automáticos usando algoritmos de aprendizagem computacional, nomeadamente máquinas de vetores de suporte e árvores de classificação. A técnica de regressão linear aplicada aos termos polinomiais das características auxiliou a geração de pontuações em escalas numéricas. Esta

abordagem explora a relação existente entre as emoções provocadas pelas imagens em indivíduos, e as respetivas características de conteúdo. As potenciais aplicações da capacidade de reconhecimento e avaliação de características visuais na perspetiva estética podem influenciar positivamente a recomendação e pesquisa por conteúdo de fotos.

Do ponto de vista da pesquisa cooperativa de imagens, Maree et al. [MDWG10] apresentaram uma plataforma de recuperação de imagens preparada para casos concretos onde os documentos se encontram distribuídos por múltiplos servidores. O método proposto segue a abordagem de descrição de pontos de interesse e geração de palavras visuais, mas utiliza estratégias de indexação randomizadas e independentes das coleções analisadas. Desta forma, a pesquisa é efetuada por várias máquinas sobre os dados associados, existindo partilha de resultados e cooperação na votação para identificação dos candidatos mais relevantes. A similaridade visual entre imagens é computada de forma distribuída, exigindo apenas uma quantidade mínima de dados transferidos entre nós do grupo. As experiências efetuadas em vários tipos de coleções de imagens demonstraram que a plataforma apresentada está apta a lidar com coleções distribuídas e heterogéneas mantendo resultados considerados satisfatórios pelos avaliadores.

2.3 Ilustração de texto e recuperação transversal

Joshi et al. [JWL06] apresentaram uma abordagem não- supervisionada para auxiliar a tarefa de ilustração automática de texto. O texto fornecido pelo utilizador é analisado para deteção e extração de palavras-chave, as quais são usadas para uma pesquisa textual inicial. Em seguida, é utilizado um esquema de pontuação de imagens que combina os resultados da pesquisa textual com características visuais rudimentares de baixo-nível. As anotações presentes nas imagens foram previamente processadas com o auxílio do serviço online Wordnet, enquanto que as imagens são comparadas entre si utilizando um esquema de reforço mútuo de similaridade. O sistema implementado, designado por *Story Picturing Engine* (Figura 2.1) foi avaliado em coleções de pequena escala (centenas de imagens) através de um estudo com utilizadores.

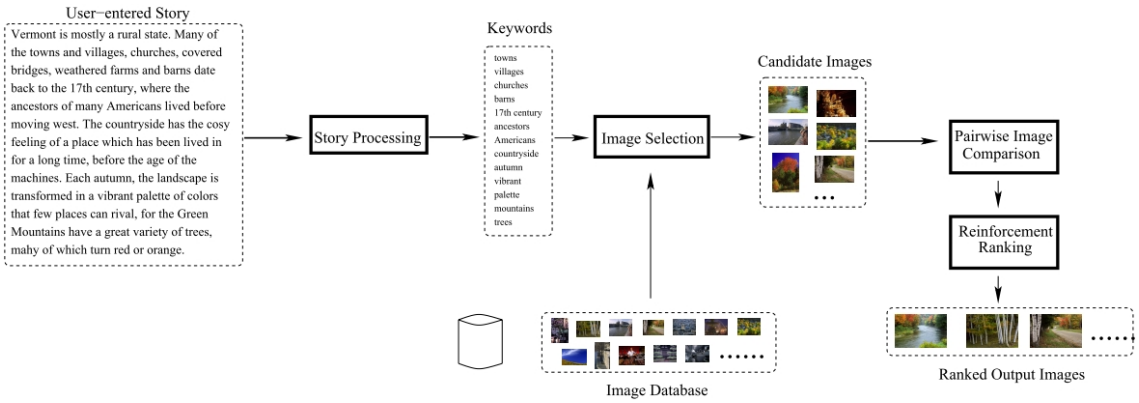


Figura 2.1: Esquema do motor de ilustração "Story Picturing Engine" (detalhado em [JWL04])

Al-Phine et al. [APBC+09] consideraram o processo de enriquecimento de conteúdos e pesquisa de informação multimédia propondo duas abordagens de processamento híbrido de informação textual e visual, as quais podem ser diretamente generalizadas para cenários multimodais. Ambas as abordagens discutidas enquadram-se na categoria de recuperação transversal de informação multimédia, auxiliadas por *feedback* de relevância por parte dos utilizadores. A primeira abordagem propõe um modelo de misturas de componentes agregados, efetivamente considerando-os como um único conceito de relevância. Na segunda abordagem, são definidas similaridades transversais multimédia como uma agregação das similaridades monomodais dos elementos agregados e o novo objeto multimodal. São também apresentadas as respetivas medidas de similaridade monomodal para texto e imagem, as quais servem de base para as medidas de similaridade transversais propostas. Os autores argumentam que uma grande variedade de tarefas de recuperação de informação multimédia podem ser enquadradas por esta perspetiva genérica, nomeadamente tarefas como anotação e legendagem de fotos, ilustração de textos, pesquisa multimédia e agrupamento de documentos (Figura 2.2). Para demonstração das potencialidades das abordagens, são ainda apresentadas duas aplicações: um sistema de auxílio à ilustração de blogs sobre viagens, e um explorador de conteúdo multimédia existente na Wikipédia.

2.4 Anotação de imagens

A anotação automática de imagens representa uma tarefa complexa, adquirindo uma importância crescente face à existência de coleções multimédia de cada vez

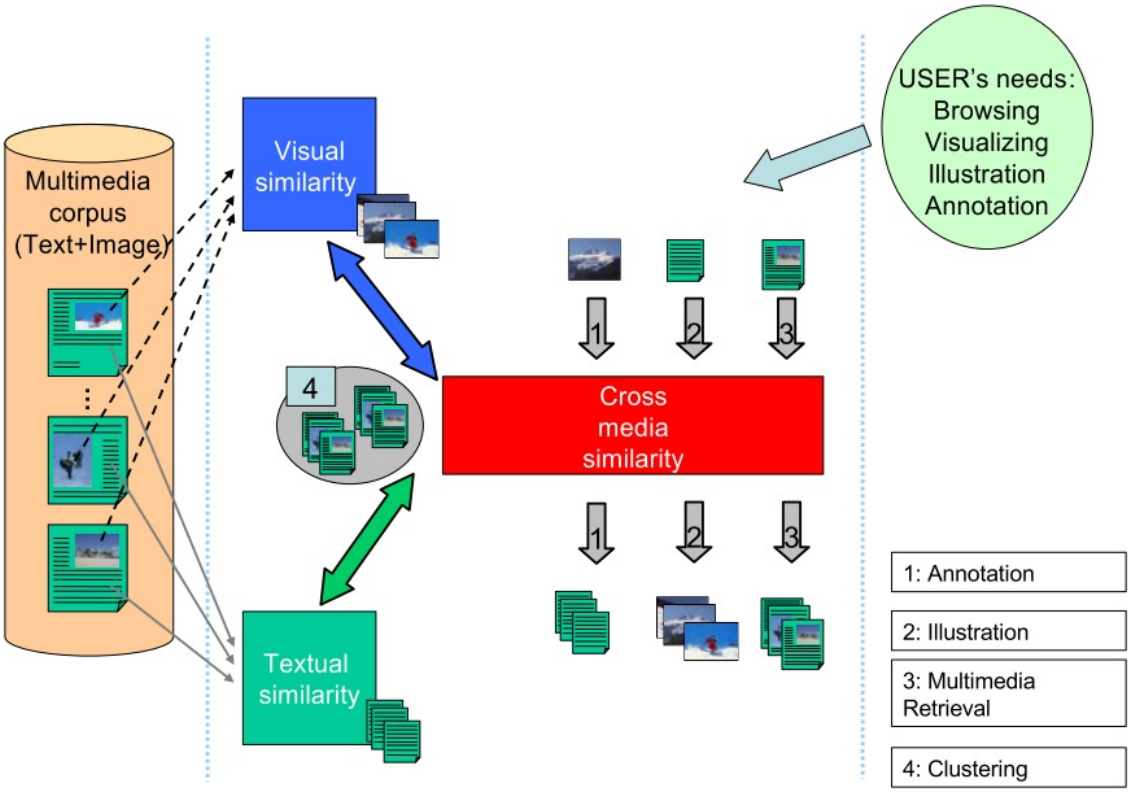


Figura 2.2: Cenários de aplicação para um sistema de recuperação transversal (detalhado em [APBC+09])

maior escala. Lindstaedt et al. [LMS+09] descreveram técnicas de anotação automática auxiliadas por bases de dados de imagens anotadas de forma colaborativa. Foram usados algoritmos de classificação para anotar imagens segundo um vocabulário controlado, seguidos de uma fase de propagação de etiquetas pelas imagens visualmente mais similares no seu conteúdo. As experiências realizadas com fotos descarregadas do serviço online Flickr demonstraram o aumento de precisão e eficiência dos métodos propostos, os quais também foram considerados por Wu et al. [WYYH09] para propagar as etiquetas mais relevantes entre fotos do Flickr.

A inserção de etiquetas por utilizadores em sítios web com forte componente social tem disponibilizado informação importante para análise e recuperação de informação multimédia de grande escala. No entanto, os atuais sistemas de recomendação baseiam-se apenas na modalidade de co-ocorrência de etiquetas, ignorando por exemplo a correlação visual entre imagens. Nesse sentido, Wu et al.[WYYH09] propuseram uma técnica de recomendação multimodal baseada em ambas as perspetivas (etiquetas e conteúdo visual), formulando a tarefa de sugestão e propagação

de etiquetas como um problema de aprendizagem computacional.

Liu et al. [LHY+09] propuseram um esquema de pontuação de etiquetas, com o objetivos de classificarem automaticamente as etiquetas associadas a uma determinada imagem de acordo com a relevância ao seu conteúdo. As etiquetas são habitualmente associadas a uma imagem de forma aleatória, sem qualquer tipo de informação sobre a sua relevância, o que limita a sua potencial eficácia na realização de pesquisas e propagação das mesmas.

Wang et al. [WMGH09] apresentaram uma abordagem à tarefa de anotação automática de imagens combinando características globais, locais e contextuais segundo um modelo de relevância transversal. Ao contrário dos métodos que até então utilizavam exclusivamente um tipo de características, ignorando a informação contextual presente em anotações e outros conteúdos textuais, a abordagem adotada considerou as três fontes de informação existentes para descrever o significado semântico existente nas fotos.

A técnica de saco de palavras visuais [NJT06] é uma técnica popular de representação de imagens que tem sido utilizada para a tarefa de anotação automática de conteúdo visual. Herve et al. [HB09] expandiram esta representação de forma a incluir a noção de informação geométrica através do uso de pares de palavras visuais. Foi demonstrado em algumas coleções de teste frequentemente utilizadas para avaliação de desempenho que o impacto da utilização de palavras visuais para deteção de objetos melhorou significativamente o desempenho do sistema de anotação automática concebido.

Com base nos pontos de interesse extraídos das zonas salientes de uma imagem, esta pode ser descrita segundo o modelo de saco de palavras visuais, adequado à tarefa de classificação de imagens. A escolha do tamanho, seleção e relevância das palavras visuais definidas por este paradigma é crucial para o nível de desempenho esperado durante a classificação, o que requer uma investigação aprofundada na determinação dos valores mais significativos para cada coleção de dados. Dada a analogia entre esta representação e a representação original do saco de palavras em documentos de texto, Yang et al. [YJHN07] aplicaram técnicas habitualmente utilizadas em categorização de texto, incluindo atribuição de pesos a termos, remoção de palavras frequentes, escolha de características, entre outras, de forma a gerar repre-

sentações de imagens que variassem segundo estes pontos. O estudo disponibilizou assim uma base empírica para a conceção de representações visuais mais robustas para o reconhecimento.

A anotação de imagens tem sido um tema de investigação bastante ativo nos últimos anos devido ao seu impacto significativo na compreensão de informação visual, reconhecimento e pesquisa de imagens. Modelos existentes, baseados na relevância probabilística da combinação de palavras e conteúdo visual co-ocorrente, tem sido explorados na vertente de fusão de conteúdo multimédia. No entanto, a existência do fosso semântico e a dependência criada na escolha das coleções de dados multimédia analisados têm restringido o desempenho e escalabilidade deste tipo de modelos. Liu et al. [LWL+07] propuseram um modelo dual de relevância multimédia transversal concebido para a anotação automática de imagens, o qual permite estimar a probabilidade conjunta baseada na expectativa das palavras em função de um léxico previamente definido. Este modelo é baseado na relação entre imagens e palavras, e na co-ocorrência das palavras visuais e textuais entre si, o que permite reforçar a informação de co-ocorrência através das ligações multimédia intrínsecas.

A necessidade do desenvolvimento de métodos mais eficazes para a anotação de imagens tem colocado sérios desafios aos investigadores desta área. A capacidade de anotação automática por computadores pode originar avanços significativos em vários domínios e tarefas, como a catalogação e pesquisa de conteúdos web em larga-escala, a descrição automática de conteúdo em bases de dados médicas, jornalísticas e pessoais, entre outros. Dada a forte ligação com as tarefas de reconhecimento e aprendizagem automática, Liu et al.[LHY+09] desenvolveram novas técnicas de estimativa e otimização para solucionar problemas comuns. Estas técnicas serviram de base para o desenvolvimento de um protótipo de anotação de fotos, nomeadamente o método de agrupamento D2, inspirado no algoritmos “k-means” e concebido para agrupar objetos representados por vetores de palavras com pesos distintos, e a técnica generalizada de mistura de modelos usando mapeamento local para dados não-vetoriais.

Estudos iniciais demonstraram que a inferência automática de alto nível da qualidade estética de imagens é uma tarefa muito complexa. A capacidade de um sistema efetuar essa avaliação provou ser de importância significativa para várias áreas. Nesse sentido, Datta et al. [DLW08] definiram o “fosso estético”, numa alusão ao “fosso

semântico" anteriormente definido pela comunidade de investigadores em pesquisa de conteúdo visual, e apresentaram também tópicos essenciais para a resolução da tarefa e inferência de emoções presentes em fotos genéricas. Foram também introduzidas as questões-base para a investigação na área, bem como abordagens para a resolução e avaliação dos problemas discutidos.

A anotação automática de imagens pode também ser abordada como a atribuição de etiquetas a um conjunto de pixels de uma imagem, etiquetas essas que indicam a presença de classes previamente estabelecidas num conjunto finito, conforme apresentado por Dumont et al. [DMWG09]. O método proposto baseia-se na extração de amostras visuais de um conjunto de imagens previamente anotado, de forma a treinar um modelo de anotação de segmentos de imagens. Este modelo traduz-se num conjunto de "árvores extremamente randomizadas", concebido para lidar com espaços vetoriais de grandes dimensões. A anotação do pixel de uma imagem é feito com base na agregação das anotações das amostras visuais contendo esse pixel. Esta abordagem foi comparada com outras mais básicas, que classificam um pixel com base numa janela centrada no mesmo, e com métodos estatísticos mais complexos. Relativamente à precisão, o método discutido apresentou vantagens claras relativamente aos mais básicos, e um desempenho competitivo face aos mais complexos, sendo mais genérico, conceptualmente mais simples e computacionalmente mais eficaz que estes últimos.

2.5 Larga-escala

Jegou et al. [JDSP10] propuseram uma abordagem para a pesquisa de imagens em larga-escala, considerando simultaneamente três pontos essenciais: a precisão dos resultados obtidos, a eficiência e desempenho do sistema, e a quantidade de informação presente nas representações visuais. É apresentado um algoritmo de agregação de descritores visuais em vetores de dimensões reduzidas, podendo ser visto como uma simplificação das representações com kernels Fisher. Foi demonstrado como otimizar simultaneamente a escolha da dimensão dos vetores e das técnicas de indexação, preservando a qualidade das representações existentes. Os testes efetuados revelaram que a pesquisa de imagens foi significativamente acelerada sem grande perda de desempenho, permitindo a recuperação de informação visual em coleções

com milhões de fotos.

Descritores visuais locais como SIFT e SURF permitem a deteção de objetos presentes em imagens com grande precisão e rapidez, mas exigem recursos computacionais que limitam a sua eficácia em máquinas tradicionais, portáteis, e ainda mais em dispositivos móveis. Nesse sentido, Calonder et al. [CLF+09] apresentaram melhoramentos a integrar numa técnica baseada em classificação estatística de geração de descrições visuais. Esta técnica, inicialmente concebida para produzir descrições de uma forma muito rápida, exigia quantidades consideráveis de memória, tornando-a desadequada para a sua utilização em dispositivos móveis. Foi demonstrado que é possível explorar a esparsividade das descrições de forma a compactá-las, acelerar a sua computação e reduzir significativamente a quantidade de memória necessária, baseando-se na teoria de sensibilidade à compressão. A eficácia do método foi também evidenciado através da sua incorporação em sistemas SLAM (*"simultaneous localization and mapping"*).

Salakhutdinov et al. [SH09] demonstraram como efetuar a aprendizagem de um modelo complexo baseado em grafos tendo por base os vetores de contagem de termos obtidos de um conjunto alargado de documentos. De acordo com os resultados obtidos, os valores das variáveis latentes na camada mais profunda do modelo são fáceis de inferir e providenciam uma representação mais rica de cada documento do que a técnica mais comum de análise semântica latente. Quando a camada mais profunda é forçada a usar um número relativamente pequeno de variáveis binárias, o modelo de grafos efetua na realidade o que é designado por "semantic hashing": os documentos analisados são mapeados para endereços de memória de forma a que documentos semanticamente semelhantes sejam colocados em endereços próximos. Desta forma, os documentos mais semelhantes a um documento-exemplo de pesquisa podem ser encontrados nos endereços que difiram apenas alguns "bits" do endereço atribuído ao documento-exemplo. Esta forma de propagar a eficiência dos algoritmos de "hashing" à pesquisa aproximada é substancialmente mais rápida do que aplicando a técnica de "locality sensitive hashing", tradicionalmente aplicada nestes contextos. Experiências subsequentes permitiram concluir que a filtragem dos resultados obtidos por TF-IDF através desta técnica foram mais relevantes do que os produzidos pela pesquisa inicial.

A técnica de "Semantic Hashing" tem como objetivo a geração de códigos binários compactos representativos de documentos originais de forma a que a distância de Hamming entre códigos esteja fortemente correlacionada com a similaridade semântica. Weiss et al. [WTF08] demonstraram que o problema em encontrar a melhor codificação para um dataset específico está diretamente relacionado com o problema do particionamento de grafos e que pode ser considerado "NP- hard". Através da simplificação das restrições inerentes ao problema inicial, foi obtido um método espetral cujas soluções estão contidas num subconjunto dos vetores próprios do Laplaciano do grafo. Baseando-se nos resultados obtidos previamente na convergência de vetores próprios dos Laplacianos de grafos, demonstraram que é possível calcular o código de um novo documento de forma eficaz. Combinados, os métodos de aprendizagem dos códigos e da sua aplicação a novos documentos da coleção resultaram em melhoramentos na simplificação do problema e consequente aumento de desempenho.

Capítulo 3

Características multimédia

> "A picture is worth a thousand words."
>
> *Chinese proverb*

Neste capítulo são apresentadas em detalhe as coleções multimédia de larga-escala adotadas para a experimentação e validação da tese apresentada. Foi efetuada uma análise das representações internas das características visuais presentes nas imagens, e a comparação de vários métodos de extração e descritores concebidos para a rápida aquisição de informação visual e subsequente pesquisa. O impacto do pré-processamento das imagens é também explorado, demonstrando-se a importância crucial desta fase inicial para a obtenção de resultados qualitativos face às restrições temporais da tarefa abordada e das necessidades de informação presentes nos utilizadores finais.

3.1 Coleções multimédia

Ao longo do trabalho desenvolvido, foram utilizadas coleções de cada vez maior escala, desde as dezenas de milhares de fotos *grayscale* no domínio médico, até à riqueza das cores e texturas presentes em quase dois milhões de fotos jornalísticas. Cada coleção apresentou desafios específicos, como a análise dos metadados existentes, a determinação da informação a recolher e a definição de *groundtruth* para

avaliação, terminando na escolha de descritores e respetivos algoritmos de comparação associados.

3.1.1 A coleção de imagens médicas IRMA-2007

Para a avaliação dos descritores visuais em cenários da área de medicina, foi obtida a coleção IRMA-2007 [LGT+04], a qual contém 11.000 radiografias pertencentes a 116 categorias. Alguns exemplos de imagens desta coleção podem ser observados na Figura 3.1(a). Estas imagens foram obtidas durante procedimentos médicos de rotina no *RWTH Aachen University Hospital*. Cada imagem possui um código associado relativo à tipologia da imagem, o qual foi traduzido para anotações textuais de forma a permitir a pesquisa textual das radiografias, conforme pode ser observado na Figura 3.1(b).

3.1.2 A coleção de imagens MIRFlickr-25k

A coleção de fotos MIRFlickr-25k [HL08] foi publicamente disponibilizada na conferência *2008 ACM International Conference on Multimedia Information Retrieval*[1]. O seu objetivo é providenciar à comunidade de investigadores na área de recuperação de informação multimédia uma vasta coleção de fotos de alta qualidade de âmbito genérico, com as *tags* e metadados respetivos associados, para uso livre e contendo um conjunto diverso de conceitos e características existentes em ambientes não-controlados.

A Figura 3.2 apresenta amostras da coleção e uma visualização em *nuvem de palavras* contendo as *tags* mais frequentes, e a Figura 3.3 contém uma listagem dos tópicos e conceitos presentes na coleção.

3.1.3 A coleção de fotos jornalísticas SAPO-Lusa

As Figuras 3.4 e 3.5 revelam uma amostra dos itens presentes na coleção de imagens fotojornalísticas. Esta coleção multimédia contém 1.490.168 fotos acompanhadas por legendas detalhadas, e cobre 85 anos do fotojornalismo português,

[1] `http://press.liacs.nl/mir2008/`

permitindo assim uma visão alargada de acontecimentos nacionais e internacionais [CR11b].

As Tabelas 3.1 e 3.2 contém um resumo dos recursos visuais e metadados presentes nesta coleção, respetivamente.

Tabela 3.1: Recursos Visuais

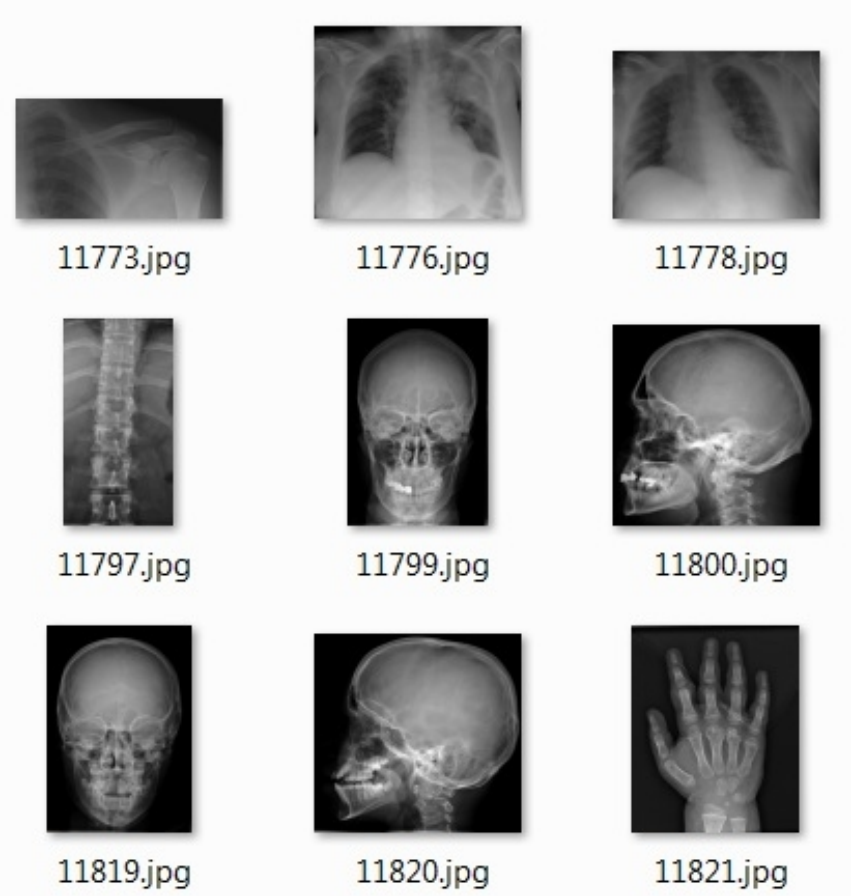
Característica	Valor
Número de documentos	1.490.168 fotos
Espaço em disco / Formato	145 GB / JPEG
tamanho mín./méd./máx. ficheiro	5 / 102 / 244 KB
largura mín./méd./máx. foto	134 / 453 / 500 px
altura mín./méd./máx. foto	62 / 399 / 500 px

Tabela 3.2: Recursos textuais

Característica	Valor
Número de documentos	1.490.168 legendas
Espaço em disco / Formato	840 MB / CSV
tamanho mín./méd./máx. título	1 / 4 / 12 palavras
tamanho mín./méd./máx. legenda	2 / 59 / 272 palavras
número mín./méd./máx. *tags*	1 / 8 / 28 *tags*

As Figuras 3.6, 3.7 e 3.8 representam as nuvens de palavras existentes nos títulos, legendas e *tags* das fotos, respetivamente. Em alguns casos, palavras específicas (evidenciadas por um "*" nas tabelas seguintes) não foram consideradas na geração das nuvens, dado que a sua frequência é exageradamente elevada e afetaria a representação de frequência das restantes palavras, e consequentemente a visualização e interpretação das nuvens.

As tabelas 3.3, 3.4 e 3.5 contêm listas estendidas das palavras existentes nos metadados previamente referidos, juntamente com as respetivas frequências.

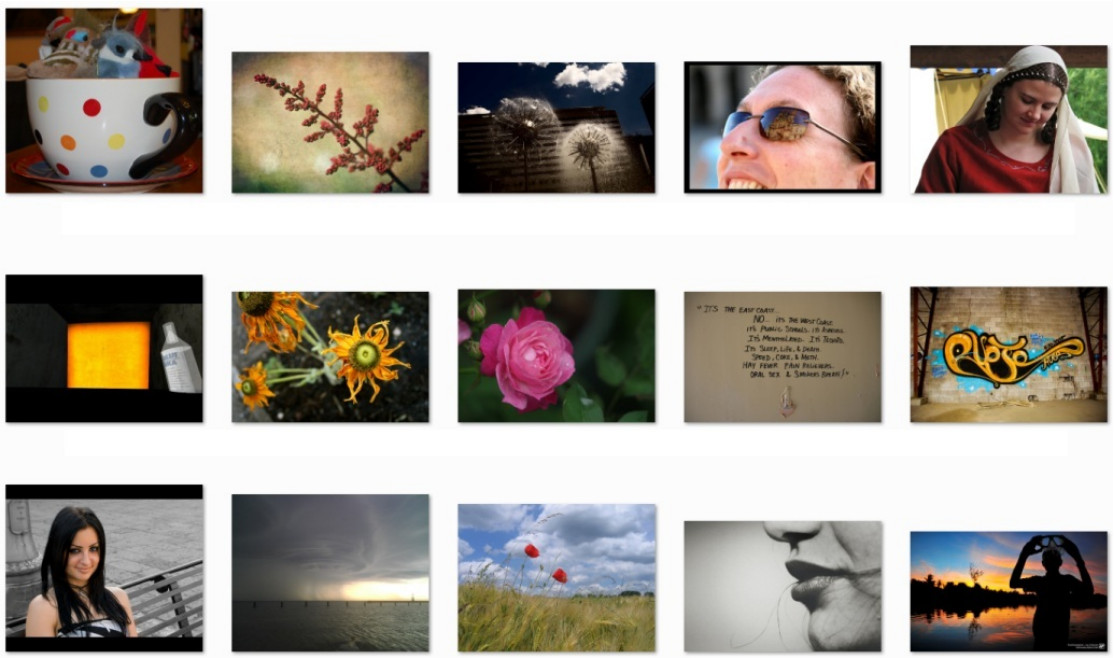


(a) Exemplos da coleção

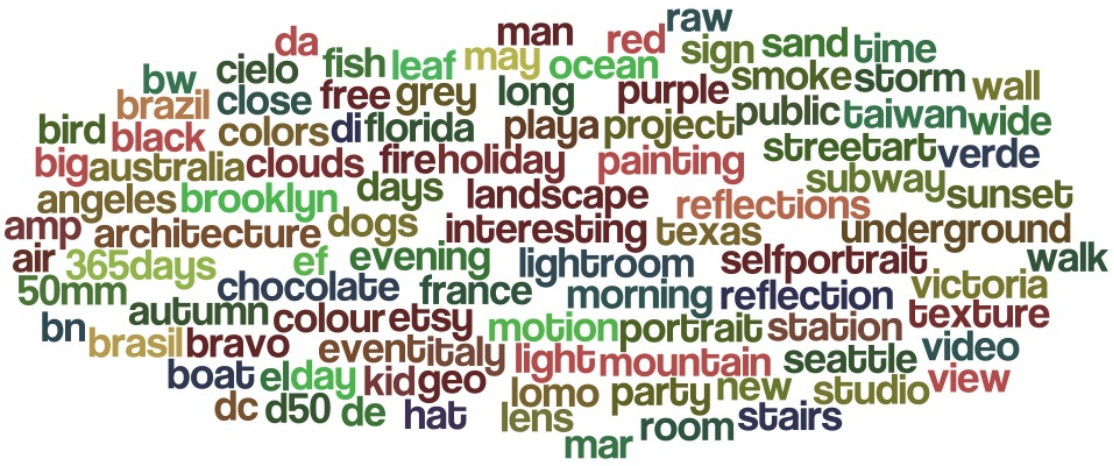
No	Rank	Field	Text	No	Rank	Field	Text
1	11000	<annotations>	radiography	21	1943	<annotations>	posteroanterior
2	11000	<annotations>	plain	22	1930	<annotations>	lateral
3	11000	<annotations>	analog	23	1874	<annotations>	arm
4	11000	<annotations>	x-ray	24	1857	<annotations>	left
5	7028	<annotations>	coronal	25	1704	<annotations>	PA
6	6778	<annotations>	system	26	1690	<annotations>	lower
7	6469	<annotations>	overview	27	1680	<annotations>	leg
8	6469	<annotations>	image	28	1594	<annotations>	right-left
9	6248	<annotations>	musculosceletal	29	1509	<annotations>	right
10	5085	<annotations>	AP	30	1280	<annotations>	spine
11	5085	<annotations>	anteroposterior	31	1051	<annotations>	cranium
12	4521	<annotations>	energy	32	1010	<annotations>	hand
13	4521	<annotations>	beam	33	952	<annotations>	inspiration
14	4520	<annotations>	unspecified	34	805	<annotations>	joint
15	4346	<annotations>	chest	35	721	<annotations>	orientation
16	4222	<annotations>	high	36	721	<annotations>	other
17	3554	<annotations>	extremity	37	604	<annotations>	mediolateral
18	2996	<annotations>	sagittal	38	528	<annotations>	lumbar
19	2166	<annotations>	coronalsupine	39	515	<annotations>	knee
20	1990	<annotations>	upper	40	497	<annotations>	cervical

(b) Anotações de tipologia

Figura 3.1: A coleção IRMA-2007

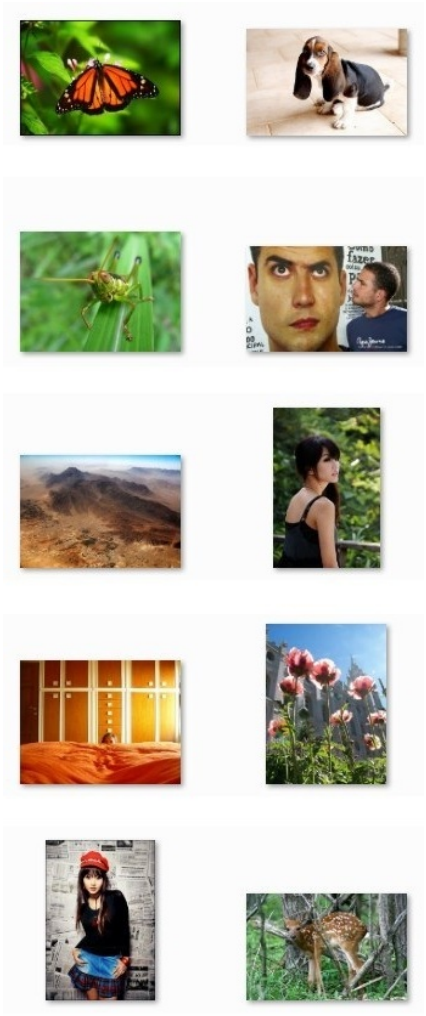


(a) Exemplos da coleção



(b) Nuvem de palavras

Figura 3.2: A coleção MIRFlickr-25k



(a) Fotos

Tag	# Images	Tag	# Images
sky	845	people	330
water	641	city/urban	308/247
portrait	623	sea	301
night	621	sun	290
nature	596	girl	262
sunset	585	snow	256
clouds	558	food	225
flower/flowers	510/351	bird	218
beach	407	sign	214
landscape	385	car	212
street	383	lake	199
dog	372	building	188
architecture	354	river	175
graffiti/streetart	335/184	baby	167
tree/trees	331/245	animal	164

General topic	Subtopics
sky	clouds
water	sea/ocean, river, lake
people	portrait, boy/man, girl/woman, baby
night	
plant life	tree, flower
animals	dog, bird
man-built structures	architecture, building, house, city/urban, bridge, road/street
sunset	
indoor	
transport	car

(b) *Tags* e conceitos

Figura 3.3: Metadados presentes na coleção

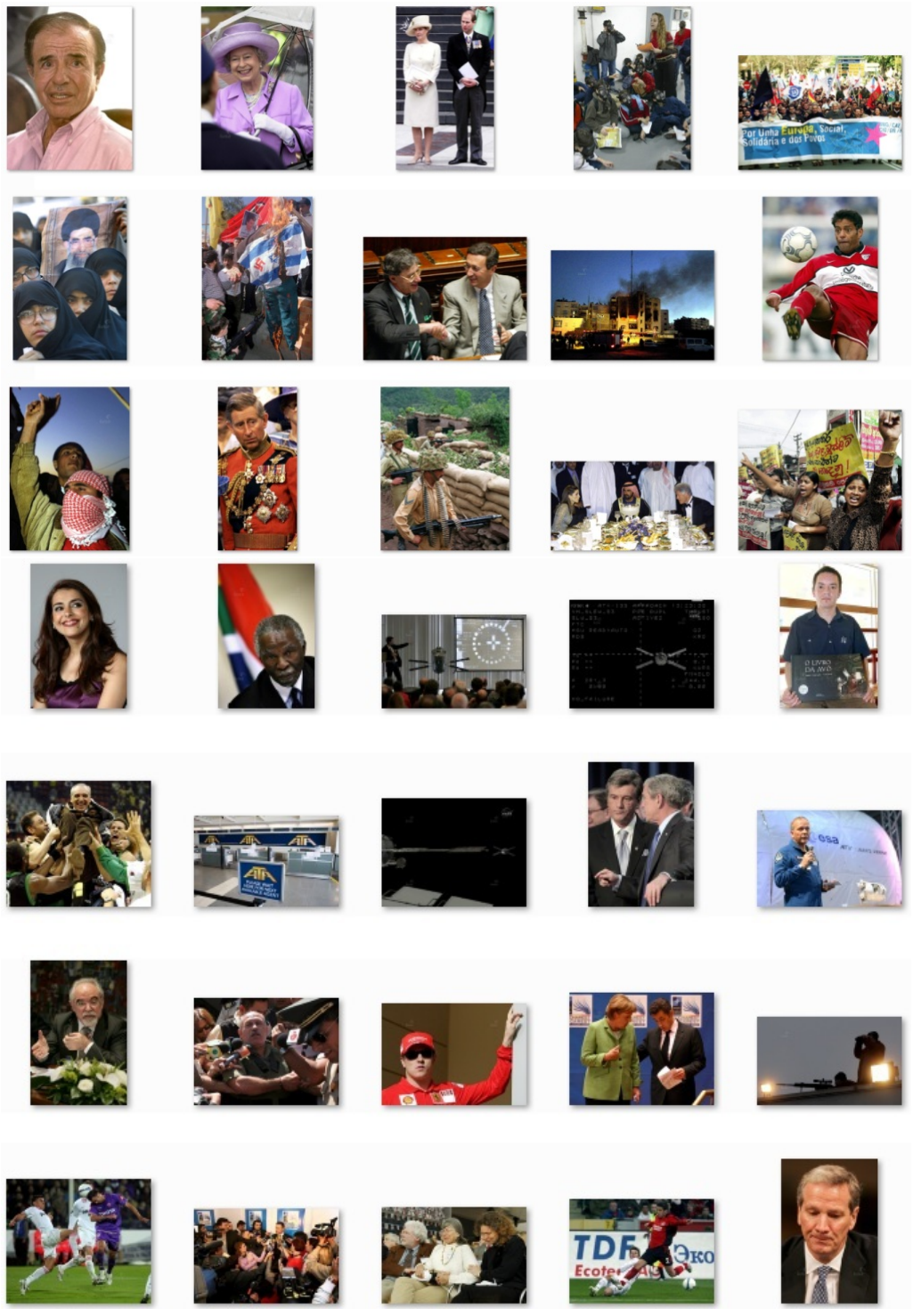


Figura 3.4: A coleção SAPO-Lusa

Portuguese Prime-Minister, José Sócrates, together with french President, Nicolas Sarkozy, portuguese Foreign Afffairs Minister, Luis Amado, and his french counterpart, Bernard Kouchner, and french Prime-Minister, François Fillon, at the starting of the cerimonies of signing of the Treaty of Lisbon at the Jeronimos Monastery, in Lisbon, Portugal, 13 december 2007. TIAGO PETINGA/LUSA PRT LISBON LUSA © 2007

Figura 3.5: Detalhes das fotos

Figura 3.6: Nuvem de palavras dos títulos

Tabela 3.3: As 20 palavras mais frequentes nos títulos das fotos

Palavra	Frequência
usa	171213
portugal	139389
soccer	123905
germany	93750
france	70591
spain	68650
tennis	66938
china	64436
cup	60435
world	60418
britain	52938
italy	51718
mideast	46842
iraq	42588
2008	42295
switzerland	41227
open	38652
russia	35846
israel	35497
us	35098

Figura 3.7: Nuvem de palavras das legendas

Tabela 3.4: As 20 palavras mais frequentes nas legendas

Palavra	Frequência
epa(*)	1148747
european(*)	661130
pressphoto(*)	618120
agency(*)	616132
during	589175
2007	436990
2008	374634
2006	252867
l	252128
after	244918
from	244840
r	239435
his	239239
lusa	203372
2005	200668
portugal	196793
match	191585
usa	185519
agência	169898
sa	169629

Figura 3.8: Nuvem de *tags*

Tabela 3.5: As 20 *tags* mais frequentes

Palavra	Frequência
agencialusa	1490137
lusa	1490133
desporto	496485
2008	366592
política	351757
2007	300525
2006	248379
2005	207183
arte	162244
entretenimento	162244
cultura	162244
2004	137390
conflitos	117417
futebol	116280
guerras	116007
2003	77563
diplomacia	72776
economia	71505
negócios	69301
finanças	69301

3.2 Descritores textuais

O conteúdo textual existente nas coleções, como as legendas, descrições, *tags*, títulos e restante informação associada, são processados por bibliotecas padrão de recuperação de informação textual.

A representação de documentos de texto é feita através de vetores de frequências de termos. O conteúdo textual é preprocessado para remoção das palavras mais frequentes (*stopwords*), as quais habitualmente não representam informação semântica relevante nas pesquisas por palavras-chave e podem ser descartadas, aumentando a eficiência do sistema a nível de armazenamento e processamento de informação. Nesta fase ocorre também a substituição de carateres acentuados, identificação e separação de termos.

Os termos são contabilizados por documento, com a geração de um índice invertido: cada termo possui uma lista de documentos que o contém e a respetiva frequência. A importância do termo na coleção é também considerada, dado que um termo que esteja presente em poucos documentos será mais específico do que um termo frequente na coleção. Esta noção de importância de um termo pela sua frequência no documento e frequência inversa na coleção é designada por TF-IDF (*term frequency - inverse document frequency*). Com base nesta noção é possível gerar os vetores de frequências de termos para cada documento e compará-los através da distância euclideana ou similaridade do cosseno.

Neste trabalho, a escolha incidiu sobre a biblioteca *Apache Lucene*, por ser flexível e integrável como módulo de um sistema de recuperação multimédia. A biblioteca Terrier [OAP+06] é também muito utilizada em ambientes de recuperação de informação textual, mas o seu ênfase na avaliação de coleções já estabelecidas e a estruturação do próprio código torna-a pouco flexível para posterior integração em sistemas.

3.3 Descritores visuais

O objetivo de analisar e compreender informação visual tendo por base apenas métodos de extração e indexação puramente automáticos está ainda fora do alcance das capacidades dos sistemas tradicionais. No entanto, têm-se verificado avanços significativos na área da recuperação de informação visual baseada no conteúdo das imagens [DJLW08, LZLM07]. Estas descobertas são aplicáveis não a casos de pesquisa genérica, mas em cenários específicos com restrições bem definidas e com metadados relevantes. Um desses cenários é representado pelo domínio médico, onde a complementaridade entre a recuperação de informação textual com a determinação da similaridade visual tem sido considerada um melhoramento crucial quando

comparada com a pesquisa restrita à informação contida em metadados [MMBG04].

Os descritores de características visuais que capturam informação representativa de toda a imagem, como cor, textura e contorno, são designados por descritores globais. Por sua vez, os descritores que capturam informação de uma zona específica da imagem são designados por descritores locais. Foi demonstrado que descritores locais como *scale-invariant feature transform* (SIFT) contribuem para uma melhor precisão no reconhecimento de objetos presentes em imagens, do que os descritores globais mais simples, mas exigem também recursos significativos a nível de memória e capacidade de processamento [DKN08].

Nesse sentido, um compromisso viável pode ser obtido através de descritores compostos, os quais agregam informação de vários descritores numa única representação. Com o contínuo aumento de tamanho e qualidade das imagens, os sistemas de recuperação de informação que incluam características visuais necessitam de permanecer capazes de gerar respostas em tempo aceitável. Assim sendo, se a ênfase da tarefa abordada não envolver reconhecimento e aprendizagem computacional, a utilização de estratégias que estabeleçam um bom compromisso entre desempenho e qualidade dos resultados obtidos torna-se assim não apenas desejável mas essencial.

3.3.1 MPEG7

Os descritores MPEG7 têm sido extensivamente utilizados nos sistemas mais básicos de pesquisa visual, e representam a norma padrão de pesquisa usando descritores globais. Os vetores de características resultantes destes descritores são comparados através de métricas específicas recomendadas [Eid03], para determinação da similaridade visual.

O descritor *Scalable Color* (SC) é baseado no histograma calculado no espaço de cores *Hue - Saturation - Value* (HSV). Este descritor tem como objetivo representar a informação de cor presente na imagem, e a qualidade dos resultados obtida aumenta com o número de coeficientes considerado. Habitualmente são usados 64 coeficientes para a transformada de Haar, resultando num vetor de características de 64 valores reais.

O descritor *Color Layout* (CL) procura ultrapassar as limitações da análise global da cor presente nas imagens através da subdivisão das mesmas em blocos, os quais são posteriormente processados e concatenados. Este descritor opera no espaço de cores YCrCb e resulta num vetor de características contendo 12 valores.

O descritor *Edge Histogram* (EH) analisa a distribuição espacial de quatro cantos direcionais e um não-direcional, capturando informação visual sob a forma de um "rascunho" da imagem, sendo esta informação semanticamente relevante para a deteção de contornos. As implementações tradicionais deste descritor devolvem um vetor de 80 valores para representar cada imagem.

3.3.2 Descritores Compostos Compactos

Os descritores compostos compactos combinam várias características num único histograma [CAB10]. Estes descritores são comparáveis através do cálculo do Coeficiente de Tanimoto, o qual procura introduzir pesos nas características visuais de forma semelhante ao cálculo TF-IDF (frequência de um termo nos documentos multiplicada pela sua frequência inversa na coleção) usado na similaridade textual. Dadas duas imagens com vetores de características A e B, a sua similaridade, de acordo com o Coeficiente de Tanimoto é dada pela seguinte equação:

$$T(A,B) = \frac{A \cdot B}{||A||^2 + ||B||^2 - A \cdot B}$$

Os vetores podem também ser comparados através da distância euclideana, a qual devolve uma similaridade aproximada usando uma fração do tempo de processamento necessário para o cálculo do Coeficiente de Tanimoto. Os descritores compostos compactos estão disponíveis para investigação e implementação de sistemas de recuperação de informação multimédia através da biblioteca *img(Rummager)* [CBL09].

O descritor *Joint Composite Descriptor* (JCD) foi concebido para a análise de imagens genéricas, sendo semelhante a nível de utilização e objetivos aos descritores *MPEG7* e *GIST*. Este descritor resulta ele próprio da concatenação de informação

produzida por outros dois descritores compostos, nomeadamente os descritores *Color and Edge Directivity Descriptor* (CEDD) e *Fuzzy Color and Texture Histogram (FCTH).*

O descritor CEDD utiliza uma versão *fuzzy* dos quatro filtros digitais propostos pelo descritor MPEG7 *Edge Histogram* para descrever a informação de textura presente nas imagens. Este descritor usa também dois sistemas *fuzzy* para mapear as cores existentes para uma palete customizada de 24 cores, resultando num vetor de características de 144 valores [CB08a].

O descritor FCTH captura informação de cor usando este mesmo processo, mas extrai informação de textura analisando as bandas de alta frequência da transformada *wavelet* de Haar através de um sistema *fuzzy*. Este descritor gera um vetor de 192 valores contendo as características extraídas das imagens [CB08b].

Estão disponíveis versões compactas destes dois descritores, nomeadamente como CCEED (60 valores) e CFCTH (80 valores), os quais sacrificam detalhe na representação das características visuais para gerar vetores mais compactos e fáceis de processar.

O descritor *Brightness and Texture Directionality Histogram* (BTDH) foi especificamente concebido para representar informação visual presente em radiografias [CB10]. Este descritor combina características de luminosidade e de textura, representando a sua distribuição espacial através de um sistema de dois módulos *fuzzy* que gera um vetor de características compacto. A informação de luminosidade é obtida através do primeiro módulo pela classificação dos valores de intensidade dos pixeis agrupados em *clusters*, cujos centros são calculados usando o classificador*fuzzy Gustafson Kessel*. O vetor resultante, apesar de compacto relativamente à informação contida, contém 2.048 valores.

3.3.3 Outros descritores

Foram considerados vários descritores globais neste trabalho, representativos do estado da arte. A medida de similaridade usada, a distância euclideana, permitiu

comparar os vetores de características e estabelecer uma noção generalizada de semelhança visual entre as imagens analisadas.

O descritor *GIST* foi desenvolvido no contexto do reconhecimento de cenas visuais [OT01]. Este descritor descreve a organização espacial da imagem através da captura de características visuais semanticamente ricas como a naturalidade, abertura, expansão, profundidade, complexidade e simetria, entre outras. Foi projetado para ser aplicado a miniaturas das imagens, com tamanhos comuns variando entre 32x32 e 128x128 pixeis, o que facilita o processamento mas reduz a informação visual capturável. Este descritor aborda a questão da análise de informação de cor através da sua aplicação aos três canais RGB em separado, com uma concatenação final que resulta num vetor de características visuais de 960 valores.

O descritor *Simple Color Histogram*, que incide sobre o espaço de cores RGB, representa a distribuição quantizada da informação de cor presente em imagens com cor, e a informação de intensidade luminosa em imagens *grayscale*. O vetor resultante contém 256 valores.

O descritor *Tamura Textures* representa a informação de textura presente nas imagens, nomeadamente a granularidade, contraste e direcionalidade das texturas existentes, concatenando-a num vetor de 18 valores. O processo de captura de informação de textura do descritor composto compacto BTDH é baseado numa abordagem *fuzzy* do histograma de direcionalidade deste descritor.

O vetor gerado pelo descritor *Auto Color Correlogram* [HKM$^+$97] é obtido com base numa matriz de co-ocorrência que representa a frequência de cada par de pixels de uma cor ou intensidade separados por uma certa distância e segundo uma certa direção na imagem. Este descritor gera um vetor de 4.096 valores.

3.4 Comparação de descritores

A avaliação de descritores é tipicamente baseada em coleções de imagens já classificadas, as quais são usadas para treino e teste de algoritmos de aprendizagem computacional, tendo como objetivo o reconhecimento de objetos em imagens. No entanto, este esquema de avaliação pode ser facilmente adaptado à perspetiva de

avaliação da recuperação de informação multimédia. Dada uma coleção de imagens agrupadas em categorias, é possível usar estas como exemplos de interrogações, e a classificação de cada imagem como *groundtruth* para determinar a sua relevância para cada uma das categorias. As métricas utilizadas têm assim ênfase na recuperação de informação existente e não na eficácia de classificação das imagens usadas para pesquisa.

Para suportar esta avaliação inicial, foi adotada a coleção IRMA-2007, dada a categorização detalhada das radiografias contidas na mesma. Foi concebida uma plataforma de avaliação com base na classificação disponível de cada imagem e a informação textual que as acompanha, e usando cada radiografia como interrogação visual para comparação dos descritores. Esta plataforma resultou posteriormente no primeiro protótipo de recuperação de informação multimédia desenvolvido neste trabalho, permitindo a pesquisa de radiografias com base na informação textual, visual e combinada, e a exploração interativa da coleção em tempo real.

A pesquisa de imagens por conteúdo requer a definição de medidas de similaridade de forma a ser possível estabelecer comparações entre os vetores de características visuais de dois itens Para a geração dos vetores relativos às radiografias, foram considerados os descritores MPEG7 Scalable Color, Color Layout e Edge Histogram, assim como os descritores compostos compactos e o descritor GIST. Todos estes descritores produzem vetores fixos de características globais, um para cada imagem, contendo valores reais comparáveis pela distância euclideana ou por medidas de similaridade específicas associadas aos próprios descritores.

A comparação efetuada envolveu um total de 15 descritores, incluindo variantes de alguns descritores, os quais extraíram características visuais de baixo nível e as representaram em vetores com elevado número de dimensões. As implementações utilizadas estão livremente disponíveis através das bibliotecas *Pyleargist* e *img(Rummager)* [CBL09].

As imagens da coleção IRMA-2007 encontram-se catalogadas em 116 classes disjuntas, com códigos específicos indicando a pertença de cada imagem a uma classe. Ao considerar que as imagens de uma classe específica representam os resultados relevantes esperados de uma pesquisa visual que use qualquer uma dessas mesmas imagens, é possível definir assim um *groundtruth* adequado através do qual seja pos-

sível testar o comportamento dos descritores visuais numa perspetiva de recuperação de informação multimédia.

Embora esta coleção especifique à partida um conjunto de 1.000 imagens de teste para avaliação do reconhecimento e classificação das radiografias, foram utilizadas as 11.000 imagens definidas como treino, para as quais dispomos de informação relativa à sua categorização. Além disso, foi necessário testar exaustivamente o comportamento dos descritores, especificamente a sua resposta em função das imagens-exemplo fornecidas, bem como o desempenho dos mesmos quer em qualidade de resultados quer em tempo de resposta.

Foram utilizadas duas medidas de avaliação de listas ordenadas de resultados, nomeadamente a *mean average precision* (MAP) e a *R-precision* (P@R), dado que estas determinam diferentes aspetos do comportamento dos descritores nesta coleção. Enquanto que o MAP calcula a média da precisão e *recall* ao longo da lista de resultados, a P@R define um ponto de *corte*. Numa interrogação para a qual se sabe existirem R imagens relevantes na coleção, a P@R traduz-se na precisão da lista aos R documentos de topo, ignorando-se a existência e posição de eventuais resultados relevantes posteriores a este ponto de corte.

Foram consideradas duas variantes das medidas apresentadas. Na primeira versão, cada imagem é considerada como uma interrogação independente e a média é calculada sobre todas as interrogações. Nesta versão, o comportamento dos descritores em classes com maior número de imagens assume um peso maior do que em classes para as quais existam poucos exemplos. A segunda versão (com prefixo "c-") envolve um passo intermédio de cálculo da média por classe, o que atenua a situação previamente referida.

Nesse sentido, os 15 descritores e variantes foram testados sobre as 11.000 imagens presentes na coleção, usando cada uma como interrogação visual. Este processo resultou em 165.000 interrogações ao longo de 8 horas de execução. As listas de resultados foram armazenadas para serem avaliadas posteriormente, num total de 20 GB de dados.

A Tabela 3.6 mostra as pontuações obtidas nas duas medidas de avaliação adotadas, sob as perspetivas de interrogações independentes ou estabelecendo primeira-

mente a média de resultados por cada classe. Nesta tabela, os descritores encontram-se ordenados por ordem decrescente de média de resultados nas quatro métricas.

Tabela 3.6: Comparação dos descritores visuais (valores em percentagem)

Descritor global	c-P@R	P@R	c-MAP	MAP	Média
MPEG7 EH	25.6	43.2	24.3	44.2	**34.3**
$GIST_{32}$	23.3	41.3	21.7	41.4	**31.9**
BTDH	23.5	39.9	21.8	40.2	**31.4**
$GIST_{64}$	22.5	40.5	20.9	40.8	31.2
$GIST_{128}$	22.0	40.1	20.2	40.6	30.5
MPEG7 CL	14.1	26.8	12.6	25.3	19.7
CCEDD	11.1	26.4	9.6	25.3	18.1
CEDD	11.1	26.4	9.6	25.3	18.1
JCD	11.2	25.6	9.7	24.5	17.8
Tamura Textures	10.9	20.0	9.3	18.0	14.6
RGB Histogram	7.8	18.3	6.9	17.3	12.6
Color Correlogram	6.7	18.7	6.3	17.5	12.3
CFCTH	5.7	17.8	4.1	16.4	11.0
FCTH	5.7	17.8	4.1	16.4	11.0
MPEG7 SC	1.4	11.8	1.6	11.3	6.5

Primeiramente, os resultados são consistentes nas métricas adotadas, MAP e P@R, em ambas as vertentes de avaliação (interrogações independentes e com média por classe). Tendo obtido pontuações acima dos 30%, os descritores MPEG7 Edge Histogram, GIST e BTDH ultrapassaram claramente todos os outros nesta coleção e segundo a metodologia de avaliação adotada, revelando que a informação de contorno e forma capturada por estes descritores se traduz no aspeto mais importante para a pesquisa de imagens de uma classe específica seguindo a perspetiva de recuperação de informação multimédia.

O descritor GIST produziu melhores resultados em miniaturas mais pequenas das imagens, o que se traduz num aspeto positivo deste descritor, dado que quanto menor a dimensão das imagens menor o espaço de armazenamento necessário e tempo de processamento.

Curiosamente, o descritor Color Layout conseguiu a quarta melhor pontuação logo a seguir aos descritores MPEG7 EH, BTDH e GIST, o que poderá estar relacionado com a informação espacial da cor (neste caso, luminosidade) capturada por

este descritor.

Os descritores compostos compactos CEDD e FCTH, agrupando informação de cor e contorno, obtiveram uma pontuação menor do que o esperado. Estes resultados podem indicar que a ausência de cor afetou negativamente a informação de contorno extraída, produzindo resultados piores.

Por último, os descritores Tamura Textures, os descritores que recolhem apenas informação de cor, e os descritores compactos combinando cor e textura, obtiveram os piores resultados, o que evidencia a importância de informação de contorno na análise e comparação de imagens nesta coleção.

É também importante verificar o comportamento das versões reduzidas CCEED e CFCTH, as quais obtiveram exatamente os mesmos resultados que as versões originais de maiores dimensões. Este facto pode ser explicado pela característica essencial desta coleção, a ausência de cor. A redução de dimensionalidade aplicada às versões compactas não causou perda de informação descritiva, dado que a quantização das características de cor considerando apenas a intensidade é suficiente para determinar a similaridade das imagens. Nesta situação, é preferível a utilização das versões mais compactas, dado que produzem os mesmos resultados de forma mais rápida e exigindo menores quantidades de memória e armazenamento, sem qualquer perda de precisão.

Após a obtenção dos resultados usando cada descritor em separado, foi efetuada uma segunda experiência de comparação através da combinação dos resultados individuais. Para combinar as listas ordenadas dos descritores *D1* and *D2*, foram adicionadas as posições de cada imagem e reordenadas segundo esta soma. Esta soma é matematicamente equivalente ao cálculo da média das posições de uma imagem, evitando uma divisão comum na comparação e ordenação das mesmas. As métricas adotadas foram recalculadas para as novas listas combinadas, gerando os resultados que podem ser observados na Tabela 3.7. Nesta tabela apenas estão representadas as combinações que produziram melhores resultados do que o melhor descritor isolado apresentado na tabela anterior.

A combinação dos três descritores de topo da Tabela 3.6 produziu os melhores resultados nas métricas estabelecidas, indicando claramente que cada um capturou

Tabela 3.7: Combinação de descritores visuais (valores em percentagem)

Combinação	c-P@R	P@R	c-MAP	MAP	Média
EH+BTDH+$GIST_{32}$	29.4	48.3	28.3	49.9	**39.0**
EH+BTDH	28.6	47.1	27.5	48.7	38.0
EH+$GIST_{32}$	27.9	46.3	26.7	47.8	37.2
BTDH+$GIST_{32}$	26.5	44.8	24.9	45.6	35.5

informação específica de características visuais, e que a sua combinação pode produzir melhores resultados em recuperação multimédia do que considerando apenas um isoladamente. Um pormenor importante nesta experiência é o facto de que cada descritor teve o mesmo peso no resultado final da combinação de listas.

É importante reforçar que a comparação de descritores efetuada considerou a perspetiva de recuperação de informação multimédia e não a perspetiva de reconhecimento de objetos e aprendizagem computacional através de exemplos positivos e negativos. Na tarefa de reconhecimento, os algoritmos de classificação obtêm melhores resultados à custa de se adaptarem aos exemplos disponibilizados para treino dos mesmos. A abordagem de recuperação de informação é mais genérica, baseando-se sobretudo na qualidade da informação extraída e representada pelos descritores, e pelas medidas de similaridade definidas, tornando-se mais genérica e adaptável a qualquer coleção. A avaliação sob a perspetiva de recuperação multimédia tende a valorizar a existência de resultados positivos no início das listas de resultados e a tolerar (até certo ponto) a existência de documentos similares mas não relevantes.

Com a combinação de descritores seguindo um processo de *late fusion* foi possível alcançar melhores resultados do que os produzidos pelo melhor descritor isolado, confirmando as expectativas de que os descritores compostos agregam informação mais rica e portanto passível de produzir melhores resultados, desde que as características analisadas não se penalizem entre si. Dado que o processo de combinação de listas é inerentemente paralelizável, é possível obter resultados no mesmo tempo necessário para obter resultados usando apenas um descritor, assumindo a existência de recursos suficientes.

3.5 Abstração de Imagens

A abstração de imagens tem sido tradicionalmente utilizada como uma técnica de *renderização não-fotorealista* que gera imagens visualmente agradáveis a partir de fotos originais, atribuindo-lhes um aspeto *cartoonesco*, assemelhando-se a um quadro pintado a aguarelas. Neste trabalho, foi analisado o impacto deste tipo de filtros de pré-processamento na pesquisa de imagens com base no conteúdo visual, para determinar se seria possível obter melhorias a nível de qualidade de resultados e armazenamento das próprias imagens.

O filtro *Anisotropic Kuwahara Filter* (AKF) adotado traduz-se numa generalização do filtro de Kuwahara adaptado ao contorno local dos objetos presentes nas cenas processadas [KKD09]. Este filtro aplica um efeito visual semelhante a um quadro pintado, incidindo sobre as características de direção das texturas mas preservando os limites de contorno dos objetos. O *ruído visual* é atenuado, tornando a imagem mais percetível, à semelhança de filmes de animação e banda desenhada, e de pinturas a óleo dos quadros clássicos, como pode ser observado na Figura 3.9.

A implementação utilizada recorre às potencialidades das placas gráficas mais modernas (GPU), sendo capaz de aplicar o filtro a vídeo em tempo-real [KKD10], o que permite, no domínio das imagens, a sua aplicação rápida a um conjunto considerável de fotos em pouco tempo, reduzindo o impacto de desempenho que a sua utilização poderia introduzir na fase de extração de características de um sistema de recuperação de informação multimédia.

Os ficheiros JPEG das imagens filtradas são em média 33% mais pequenos do que os ficheiros das imagens originais, conforme pode ser visto na Tabela 3.8. Usando um formato de imagens mais recente, como o formato WebP[2] da Google, foi possível aplicar o filtro à coleção MIRFlickr-25k e obter uma coleção de fotos abstraídas que ocupa um décimo do tamanho da coleção original. Este passo produz assim ficheiros mais pequenos e simplificados que reduzem o custo de armazenamento e processamento durante o processo de extração e representação das características pelos descritores visuais.

[2] `http://code.google.com/speed/webp/`

Figura 3.9: Imagens originais (esquerda) e abstraídas (direita)

Tabela 3.8: Requisitos de armazenamento da coleção MIRFlickr-25k (valores melhores a negrito)

Característica	JPEG	JPEG c/AFK	WebP	WebP c/AFK
espaço necessário	2877 MB	1915 MB	652 MB	**303 MB**
tamanho médio	118 KB	79 KB	27 KB	**12 KB**
redução obtida	–	-33.4%	-77.3%	**-89.5%**

Capítulo 4

Pesquisa em larga-escala para ilustração interativa

> "If we knew what it was we were doing, it would not be called research, would it?"
>
> *Albert Einstein*

Neste capítulo é apresentada em detalhe a tarefa abordada neste trabalho, nomeadamente a ilustração automática de conteúdo textual com recurso a coleções multimédia de larga-escala. Os algoritmos de indexação e pesquisa multimodal recorrem à informação disponibilizada pelos descritores visuais discutidos previamente, e permitem a execução da tarefa nas vertentes de recomendação de imagens por conteúdo e exploração interativa das coleções multimédia processadas.

4.1 Descrição da tarefa

O paradigma da pesquisa multimodal explora a possibilidade de mapeamento e interligação entre diferentes tipos de conteúdos multimédia. As abordagens adotadas podem utilizar técnicas de aprendizagem computacional, como classificação e agrupamento supervisionado, ou técnicas de recuperação de informação como a extração e indexação de termos textuais, características visuais e propriedades áudio.

A interpretação de imagens, envolvendo o reconhecimento de objetos ou propagação de etiquetas, representa um exemplo de tarefas que enriquecem o espaço

visual com conteúdo textual. Na perspetiva oposta, a tarefa de ilustração de textos enriquece as descrições, parágrafos e excertos com conteúdo multimédia como fotos, diagramas, música e sons numa tentativa de melhoramento das experiências de consumo ou mesmo para clarificar a sua mensagem.

Rasiwasia et al. [RCPC$^+$10] apresentaram uma estratégia de combinação de conteúdos para correlacionar propriedades textuais e visuais, permitindo que os sistemas possam aprender características específicas para o mapeamento entre dimensões. Esta estratégia produziu resultados bastante promissores em coleções de pequena escala, na ordem dos milhares de documentos multimédia.

De forma a lidar com coleções de grande escala, com milhões de recursos o trabalho aqui discutido segue uma abordagem em *cascata*. A análise textual é usada primeiramente para delimitar o contexto da interrogação, devolvendo um conjunto de potenciais imagens candidatas para a ilustração dos textos introduzidos. Posteriormente, este conjunto de fotos é reordenado através da determinação da similaridade entre as imagens e consequente reordenação de resultados.

É possível observar que a tarefa de ilustração de textos, além de interativa, é sobretudo uma tarefa subjetiva. Assim sendo, o foco deste trabalho é na disponibilização de funcionalidades que auxiliem a tarefa de ilustração disponibilizando o acesso, pesquisa e navegação em grandes quantidades de dados multimédia, com tempos de resposta adequados à interatividade subjacente.

A tarefa de compreensão de documentos textuais incide na necessidade essencial de descrever o curso da ação. Torna-se especialmente relevante quando aplicada a histórias infantis, incentivando o estímulo da aprendizagem através de componentes visuais e auditivas. Na faixa etária oposta, existe também a necessidade de transmissão e compreensão de conteúdos noticiosos a utilizadores mais idosos. Delgado et al. [DMC10] apresentou resultados promissores na descrição visual de textos, focando-se essencialmente na compreensão dos eventos aí relatados.

Numa perspetiva relacionada mas com objetivos diferentes, o trabalho apresentado procura auxiliar os criadores de conteúdo no enriquecimento dos documentos, sendo a sua melhor compreensão um bónus colateral. Os algoritmos de pesquisa apresentados nas secções seguintes procuram organizar as sugestões de fotos para

ilustração de forma a que os utilizadores possam efetuar diferentes escolhas também em função da análise visual e não apenas dos metadados que acompanham as imagens.

Com a escolha de algoritmos de análise textual que não envolvam a complexidade dos métodos de processamento de linguagem natural, as abordagens não-supervisionadas de ilustração de texto podem tornar-se aptas a lidar com coleções de recursos multimédia de larga-escala, tirando partido da interatividade com os utilizadores para garantir a execução da tarefa.

Joshi et al. [JWL06] apresentaram um motor de ilustração de textos que efetua a extração de palavras-chave e as usa para pesquisar pequenas coleções de imagens anotadas. Estas coleções são específicas para o contexto visual associado a cada documento a ser ilustrado (paisagens, objetos artísticos e eventos históricos, ente outros). É usado um esquema de pontuação para determinar a importância de cada imagem, tendo em consideração as anotações e o resultado da aplicação de alguns descritores básicos de conteúdo visual. O trabalho aqui discutido expande esta abordagem no sentido de aplicar a coleções de grande escala, na ordem dos milhões de imagens, coleções essas com conteúdo genérico abrangendo um conjunto diverso de tópicos como política, conflitos, entretenimento e desporto, entre outros. São necessários algoritmos de pesquisa capazes de lidar com dados desta dimensão, assim como esquemas e avaliação que contemplem o desempenho e a qualidade dos resultados produzidos pela execução da tarefa de ilustração de textos usando coleções multimédia com milhões de fotos de alta qualidade.

A estratégia seguida neste trabalho tira partido da aplicação de filtros às imagens (Secção 3.5), descritores visuais compostos compactos (Secção 3.3.2), algoritmos de indexação e pesquisa aproximada (Secção 4.2) e reordenação transversal (Secção 4.6) para efetuar a ilustração automática de texto com recurso a coleções multimédia de grande escala. As funcionalidades associadas à tarefa foram demonstradas através da implementação de dois protótipos de exploração de coleções de grande escala por conteúdo e recomendação de fotos para ilustração [CR10, CR11c].

As duas fases do processo podem ser observadas em pormenor na Figura 4.1 e Figura 4.2. Os utilizadores fornecem fragmentos de texto para obter listas de imagens recomendadas para ilustração. Ao percorrer essas listas, podem refinar os resultados

do processo de pesquisa através da reordenação por similaridade com uma imagem escolhida, ou explorar a coleção com base no conteúdo visual dessa mesma imagem.

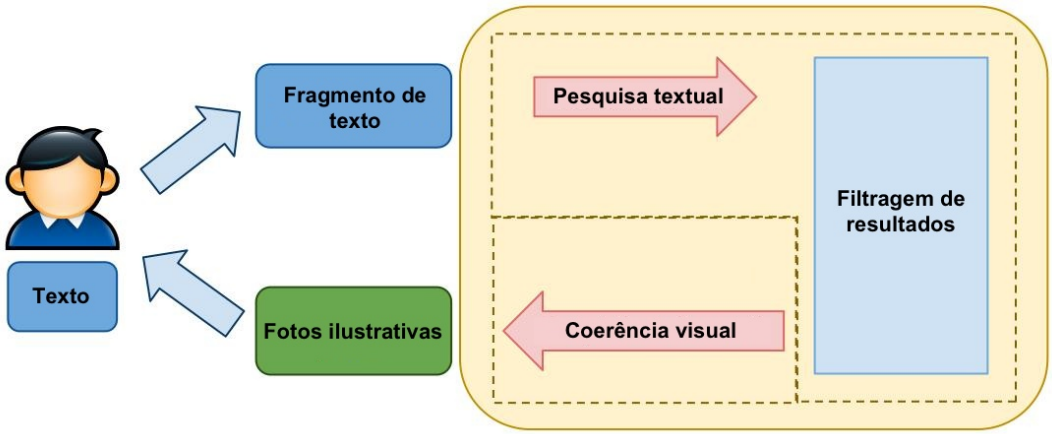


Figura 4.1: Recomendação de fotos

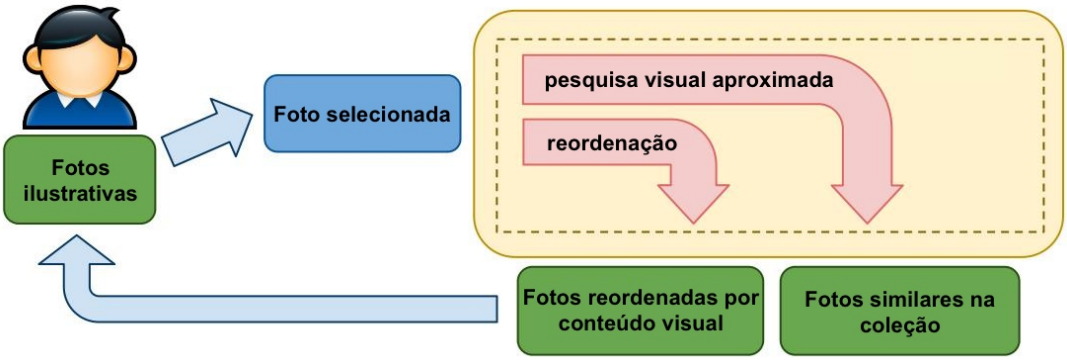


Figura 4.2: Exploração visual

4.2 Indexação em larga-escala

Na indexação de conteúdos visuais, envolvendo descritores globais e características de baixo nível, cada imagem é representada por um vetor de características de elevada dimensão, sendo a similaridade determinada com base em cada uma das características extraídas. Uma abordagem simples e direta de pesquisa consiste na comparação exaustiva do vetor de características da imagem usada como interrogação com cada vetor armazenado no índice, de forma a encontrar as imagens mais

similares.

Apesar dos contínuos avanços nas capacidades de processamento e acesso a dados, a comparação exaustiva torna-se proibitiva para coleções de imagens de grande escala, resultando em tempos de resposta desadequados para tarefas interativas. Os recursos visuais requerem um armazenamento e processamento eficientes de forma a permitir um acesso rápido à informação relevante, dado que, para responder às tarefas consideradas, os sistemas de pesquisa multimédia devem retornar um conjunto sucinto de resultados provenientes de coleções com milhões de possíveis candidatos.

A investigação na área de pesquisa em larga-escala focou-se na criação de métodos de pesquisa aproximados, de forma a simplificar a recuperação de dados e obter tempos de resposta adequados. A estratégia mais adotada, designada por *Locality Sensitive Hashing* [DIIM04], consiste no agrupamento de imagens recorrendo a tabelas de dispersão. Imagens similares são colocadas no mesmo grupo, e a localização destes grupos é tanto mais próxima quanto maior a similaridade das imagens que pertencem a esses mesmos grupos. Isto é, imagens similares encontrar-se-ão próximas, em posições de memória contíguas.

A dificuldade na aplicação bem sucedida desta estratégia está na determinação de uma função de dispersão que seja capaz de posicionar corretamente os grupos e distribuir de forma eficiente as imagens consoante a sua similaridade, o que introduz um custo a nível de utilização de memória, o que pode comprometer a sua eficiência se não existirem recursos computacionais suficientes para albergar coleções multimédia de grande escala.

Uma alternativa igualmente eficaz na obtenção de grupos de imagens similares, mas que requer menos recursos de memória e processamento, designa-se por "Metric Index" (M-Index), e utiliza um número reduzido de documentos (textos, imagens, músicas, ...) como pontos de referência, denominados "pivôs" [AS08]. Após a sua escolha, estes pontos de referência são usados para definir a localização de cada documento no espaço vetorial de grandes dimensões criado pelos vetores de características associados. Nesse sentido, cada imagem é então representada pela similaridade ordenada aos pivôs, sendo o seu vetor de características necessário apenas para uma fase posterior, de determinação de similaridades de forma mais exata.

Este conceito de determinação de similaridades baseado na distância ordenada a pontos de referência tem sido progressivamente melhorado. Uma variante, o "Permutation Prefix Index" (PP-Index) [Esu09], considera apenas, para cada imagem, os primeiros N pivôs da lista ordenada, com base na dimensão da coleção. Recorre também a vários conjuntos de pontos de referência, ao contrário o M-Index que apenas utiliza um único conjunto de pivôs para indexar os documentos. A Figura 4.3 exemplifica a indexação de um conjunto de documentos com base na sua distância aos pivôs.

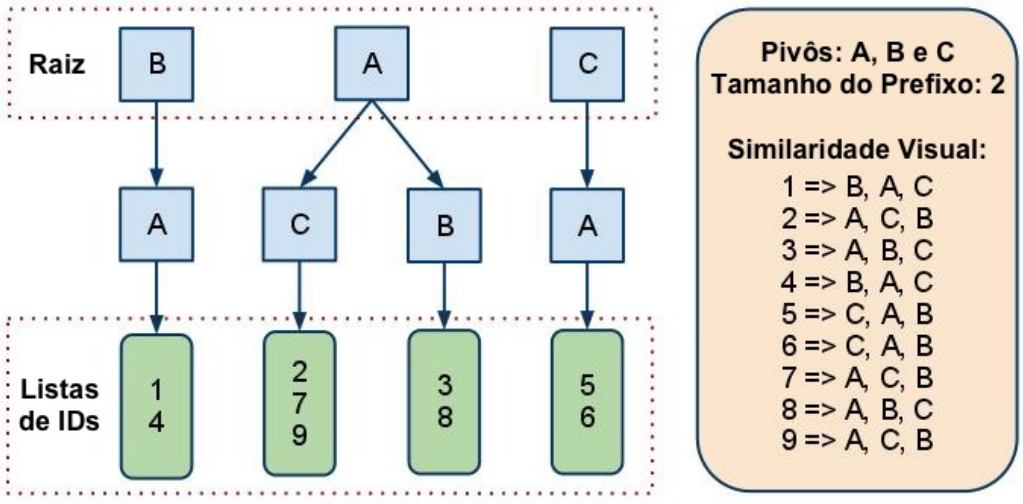


Figura 4.3: Indexação baseada em pontos de referência

4.3 Escolha dos pontos de referência

A estratégia de indexação e pesquisa aproximada adotada engloba um número considerável de variáveis, desde a determinação do número de pontos de referência, a utilização de uma ou mais árvores de prefixos, e envolve também a escolha dos pivôs, a qual está intrinsecamente ligada à capacidade de estabelecer rapidamente a similaridade entre imagens.

Nesse sentido, foi efetuada uma comparação de estratégias de escolha de pontos de referência, bem como uma análise do impacto do número de conjuntos de pivôs

no desempenho da pesquisa aproximada. Relativamente aos parâmetros testados, foram determinados valores com base na documentação disponível e na realização de experiência preliminares. Para uma coleção de D documentos, o PP-Index recorre a $\log D$ árvores (conjuntos de pivôs), $\sqrt[3]{D}$ pivôs por árvore, e o prefixo foi limitado a 3 pivôs para indexar cada imagem. Estas definições resultaram em 6 árvores e 114 pivôs por árvore para a coleção SAPO-Lusa, e 4 árvores com 29 pivôs cada para a coleção MIRFlickr.

Para a análise da escolha de pivôs, foram definidas duas estratégias, nomeadamente a escolha aleatória de pontos de referência, e a utilização de um algoritmo de diversificação. Este algoritmo determina inicialmente o "medóide", isto é, a imagem mais central da coleção, considerando o espaço vetorial em questão. Este ponto inicial é adicionado a uma lista vazia, e determinado em seguida o ponto mais distante do medóide, o qual é também adicionado à lista de pivôs. O algoritmo vai encontrando e adicionando sucessivamente os pontos mais distantes ao conjunto até aí obtido, até atingir o número de elementos pretendido. O objetivo deste algoritmo é tirar partido do espaço vetorial e "espalhar" os pivôs de forma equitativa.

As experiências seguintes, de avaliação de desempenho, foram realizadas localmente num computador portátil equipado com um processador Intel Quad-Core i7 a 1.6GHz, 8GB de RAM DDR3 e um disco rígido SATA a 5400rpm.

4.3.1 Experiências na coleção MIRFlickr-25k

O primeiro passo consistiu na geração dos índices para esta coleção. O PP-Index recorre a 4 árvores com 29 pivôs, e o M-Index possui apenas uma árvore, com 116 pivôs. A Tabela 4.1 apresenta os atributos de cada índice usando as duas estratégias de seleção de pivôs: a escolha aleatória ou a seleção de pontos de referência com ênfase na diversificação.

O índice M-Index foi construído mais rapidamente e exigiu um menor espaço de armazenamento que o PP-Index. O esquema de armazenamento é efetuado usando técnicas de serialização dos dados e compressão dos ficheiros gerados, o que permite carregar as estruturas de dados de forma eficaz.

Tabela 4.1: Construção dos índices - Coleção MIRFlickr-25k

Índice Pivôs	PP-I aleatórios	PP-I diversif.	M-I aleatórios	M-I diversif.
tempo de construção (s)	145	158	**43**	72
armazenamento (KB)	271	222	124	**74**
RAM utilizada (MB)	6	**3**	6	4

Relativamente à escolha de pivôs, o algoritmo de diversificação resultou em requisitos de armazenamento e memória mais reduzidos do que selecionando aleatoriamente os pontos de referência, mas necessita de mais tempo na fase de construção de ambos os índices.

De forma a determinar os potenciais melhoramentos introduzidos pela pesquisa aproximada com base em pontos de referência, foram utilizadas as anotações existentes nas coleções adotadas para avaliação deste trabalho. O *groundtruth* das classes visuais fornecido com a coleção MIRFlickr-25k foi criado manualmente por um anotador experiente, dado que esta coleção foi concebida para apoio à investigação na área de pesquisa visual baseada no conteúdo.Esta coleção inclui 24 conceitos visuais (Figure 3.3) como *animal*, *people* e *sunset*, e listas de imagens em que esses conceitos estão presentes. Alguns exemplos desses conceitos podem ser vistos na Figura 4.4.

O próximo passo na avaliação do processo de pesquisa aproximada consistiu em iterar todas as listas das fotos pertencentes a cada conceito e usar cada imagem como interrogação visual. Dado que as fotos podem conter um ou mais conceitos, estas podem estar presentes em várias listas. As 73.342 interrogações visuais efetuadas excederam largamente o número de fotos contidas nesta coleção.

Com base nos dados de avaliação disponibilizados com a MIRFlickr-25k, foram analisadas as pontuações médias a 10 e 100 resultados, assim como o tempo necessário para efetuar as pesquisas. Nesta avaliação, uma imagem é considerada relevante para uma interrogação visual específica se estiver contida na lista de *groundtruth* do conceito pesquisado associado a essa interrogação. Os resultados obtidos podem ser observados na Tabela 4.2.

Nesta experiência de avaliação da pesquisa com base em pontos de referência, o índice M-Index produziu resultados mais rapidamente e obtendo resultados simila-

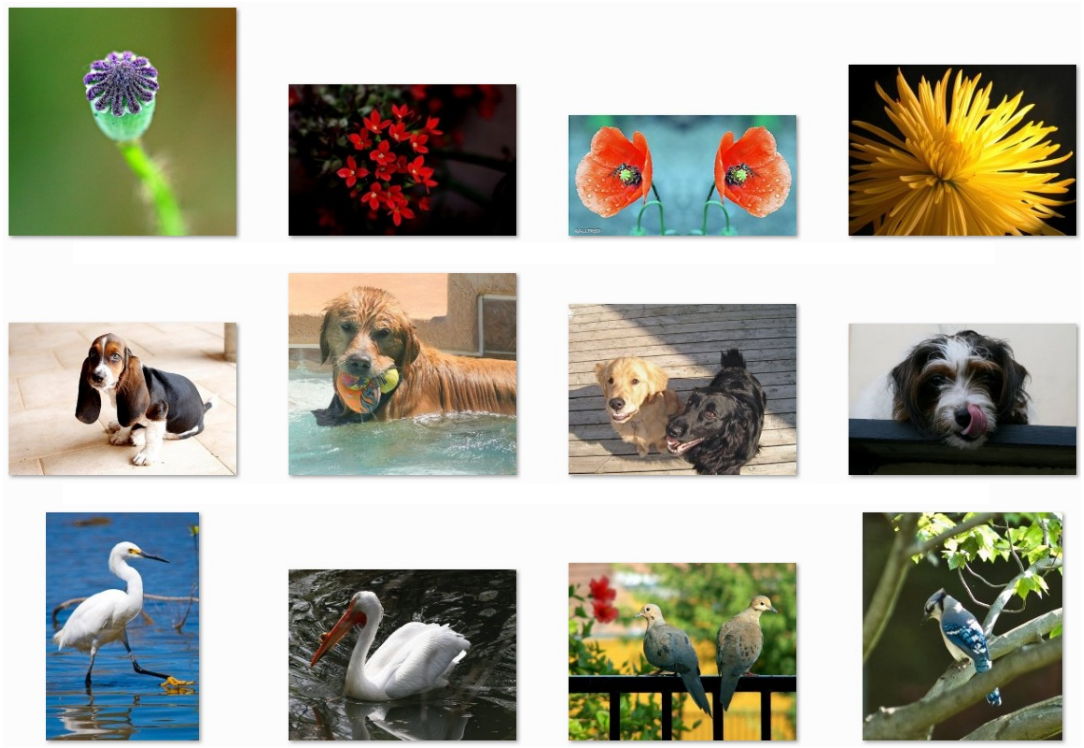


Figura 4.4: Exemplos de conceitos visuais. Primeira fila: plant. Fila do meio: dog. Fila de baixo: bird

Tabela 4.2: Pesquisa - Coleção MIRFlickr-25k

Índice Pivôs	PP-I aleatórios	PP-I diversif.	M-I aleatórios	M-I diversif.
tempo de pesquisa (ms)	10	19	**4**	**4**
Precisão@10 (%)	43	43	**45**	43
Precisão@100 (%)	34	34	**40**	35

res ou melhores do que o PP-Index. A combinação do M-Index com uma escolha aleatória de pivôs produziu os melhores resultados de precisão, mas a estratégia de diversificação de pontos de referência usou menos recursos de memória e armazenamento.

4.3.2 Experiências na coleção SAPO-Lusa

Foram também efetuadas experiências de avaliação na coleção SAPO-Lusa, de forma a verificar o comportamento destes índices quando confrontados com uma coleção duas ordens de grandeza maior do que a coleção MIRFlickr-25k. Nesta coleção,

o índice PP-Index usa 6 árvores com 114 pivôs cada, e o M-Index contém 684 pivôs numa única árvore. A Tabela 4.3 apresenta os atributos de cada índice usando uma escolha aleatória ou diversificada de pontos de referência.

Tabela 4.3: Construção dos índices - Coleção SAPO-Lusa

Índice Pivôs	PP-I aleatórios	PP-I diversif.	M-I aleatórios	M-I diversif.
tempo de construção (min)	49	104	**21**	58
armazenamento (MB)	27	19	6	**3**
RAM usada (MB)	149	95	138	**20**

Quando comparados com os resultados obtidos na coleção MIRFlickr-25k, é possível constatar que o índice M-Index continua a exigir menos espaço de armazenamento, e que a estratégia de diversificação de pontos de referência resulta em índices mais pequenos a nível de memória e de espaço em disco. Relativamente aos tempos de construção dos índices, a fase de escolha de pivôs assume uma ênfase acentuada, representando praticamente metade do tempo total de construção do índice.

Para a tarefa de pesquisa, a definição de *groundtruth* foi feita com base nos metadados disponíveis, dado que esta coleção não possui julgamentos de relevância e não foi preparada com o objetivo de avaliar algoritmos de pesquisa visual baseados no conteúdo. Assim sendo, foi definido um conjunto de 100 personalidades distintas com o propósito de serem usadas como base para interrogações visuais num processo semelhante ao proposto para a coleção MIRFlickr-25k. As fotos são acompanhadas de legendas e títulos previamente inseridos para indexação manual da coleção. O título das fotos foi utilizado para estabelecer o *groundtruth* de cada personalidade, isto é, uma imagem é considerada relevante para uma interrogação visual se a personalidade mencionada no título da foto for a mesma da imagem usada para interrogação. Alguns exemplos das personalidades existentes na coleção SAPO-Lusa podem ser observados na Figura 4.5.

Com a criação das listas de *groundtruth* das personalidades, cada foto foi usada como interrogação visual. Neste caso, o número total de interrogações efetuadas foi de 19.815, com os resultados evidenciados na Tabela 4.4.

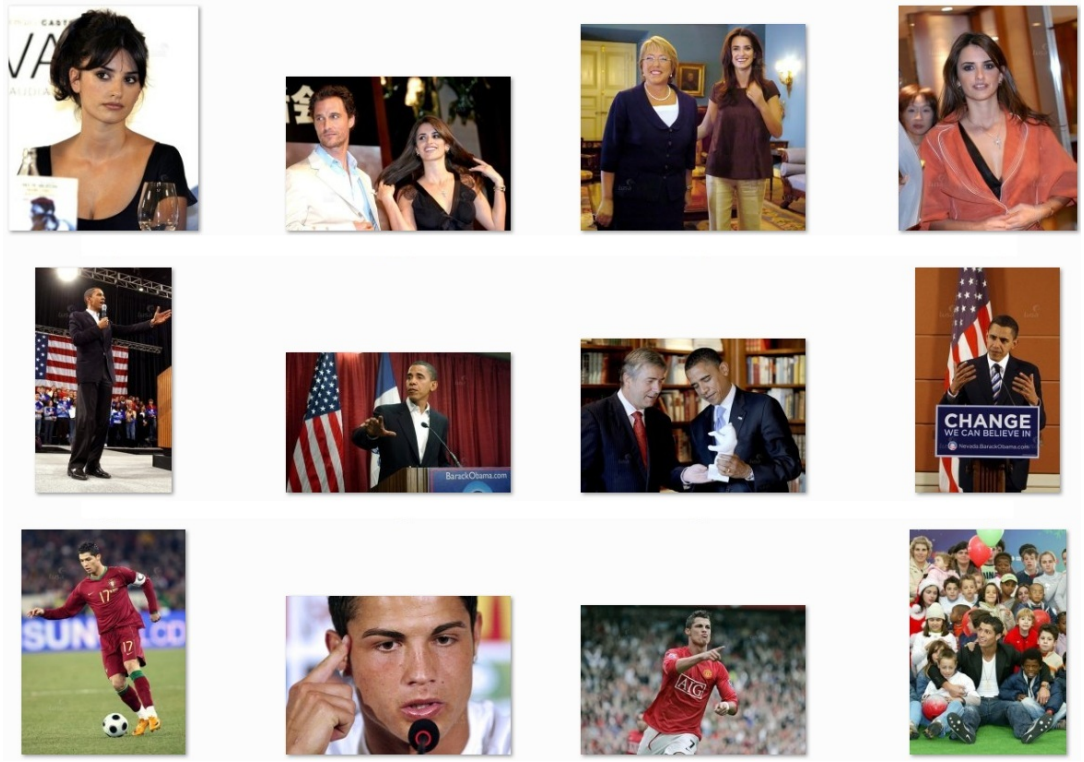


Figura 4.5: Exemplos de fotos de personalidades. Fila de cima: atriz Penélope Cruz. Fila do meio: presidente Barack Obama. Fila de baixo: futebolista Cristiano Ronaldo

Tabela 4.4: Pesquisa - Coleção SAPO-Lusa

Índice Pivôs	PP-I aleatórios	PP-I diversif.	M-I aleatórios	M-I diversif.
tempo de pesquisa (ms)	25	643	**2**	85
Precisão@10 (%)	12	**13**	**13**	12
Precisão@100 (%)	2	2	**5**	2

Com base nos resultados obtidos, é possível enunciar as mesmas conclusões previamente obtidas para a coleção MIRFlickr-25k. Os melhores resultados de precisão são obtidos com o índice M-Index, e efetuando uma escolha aleatória de pivôs. É possível obter índices necessitando de poucos recursos de armazenamento e memória através do algoritmo de diversificação de pontos de referência, com uma pequena perda de precisão e aumento do tempo de pesquisa.

Relativamente às baixas pontuações de precisão, uma das desvantagens da coleção SAPO-Lusa é a ausência de julgamentos de relevância manuais. A estratégia adotada, usando o título das fotos como fonte separada de informação, representou

a forma mais direta de geração de listas de *groundtruth* aproximadas.

Os exemplos contidos na Figura 4.6 demonstram que imagens visualmente similares podem corresponder à mesma personalidade. No entanto, numa coleção onde estão presentes grupos de fotos relativas a um evento específico, o descritor JCD adotado, e consequentemente a indexação baseada em pontos de referência, promovem imagens de diferentes personalidades em contextos semelhantes, recorrendo a características visuais de baixo nível.

4.3.3 Análise de estratégias

Através da comparação efetuada entre índices baseados em pontos de referência, foi possível determinar a qualidade e a rapidez na obtenção de resultados em pesquisas aproximadas, bem como constatar os requisitos de armazenamento e memória necessários.

Foi possível concluir que os índices apresentados aceleram de forma substancial as tarefas de pesquisa de imagens por conteúdo, exigindo um custo de armazenamento mínimo quando comparado com o tamanho das coleções indexadas. Nos testes efetuados, usando os parâmetros previamente determinados, o índice M-Index produzir melhores resultados do que o índice PP-Index, sugerindo que uma árvore com um número elevado de pivôs e prefixos curtos é suficiente para obter bons resultados nestas coleções.

A escolha de pontos de referência produz os melhores resultados quando é aleatória, exigindo tempos de resposta mais curtos à custa de memória e espaço de armazenamento. O algoritmo de diversificação proposto reduz de forma significativa a memória e espaço em disco usados pelos índices, introduzindo apenas uma pequena redução na qualidade e aumento dos tempos de pesquisa.

4.4 Binarização dos vetores de características

Uma forma de acelerar o processo de determinação da similaridade entre imagens, baseado na distância entre os vetores de características respetivos, está relacionada

Figura 4.6: Exemplos de exploração visual. As imagens no topo foram usadas como interrogações visuais.

com a simplificação dos próprios vetores.

Considerando que as coleções são estáticas, ou que a sua dimensão e diversidade de fotos é de tal forma elevada que incluirá um conjunto suficientemente representativo das possíveis imagens a pesquisar, é possível estabelecer a existência de um vetor "médio", um vetor de características cujos valores representem a média dos valores em cada característica ou dimensão.

Com a obtenção deste vetor médio, é possível binarizar cada um dos vetores representativos das fotos existentes na coleção. Para cada característica, o seu valor é substituído por 0 se for igual ou inferior ao valor médio, ou por 1 se for superior. Quando o número de características é igual ou inferior a 64, é possível armazenar os vetores binarizados em inteiros de 32 ou 64 bits. Este ponto torna-se especialmente relevante quando se verifica que uma comparação de vetores binarizados se traduz numa operação de "ou exclusivo" (XOR), seguida de uma contagem do número de bits diferentes. Esta sequência é muito mais eficiente a nível de processamento do que um cálculo complexo de distância euclideana entre dois vetores.

4.5 Mapeamento de características

Foram apresentadas recentemente soluções *opensource* que exploram a natureza paralelizável da pesquisa visual [HSD11]. Baseadas na deteção de pontos de interesse, descritores SIFT/SURF e geração de sacos de palavras visuais, estas soluções encontram-se adaptadas para o reconhecimento de objetos em imagens, tirando partido de plataformas distribuídas como *Apache Hadoop* [1] para a gestão de *clusters* de máquinas capazes de processar coleções de dados de grande escala.

No entanto, a solução apresentada neste trabalho pode evitar o recurso a índices visuais específicos ao tirar partido da eficiência dos motores de indexação textuais já existentes, através de um mapeamento de características visuais para um formato textual [GABS10, ABF+11], designado por *Surrogate Text Representation* (STR). Estas representações são armazenadas como um campo adicional dos documentos no conjunto indexado pelos motores de pesquisa, permitindo a pesquisa de imagens por similaridade textual, visual e combinada, juntamente com os metadados já existentes.

[1] `http://hadoop.apache.org/`

4.5.1 Mapeamento de características visuais

No esquema de pesquisa aproximada baseada em pontos de referência apresentado anteriormente, cada imagem é comparada com um conjunto predefinido de pivôs, resultando num vetor de similaridade ordenada que efetivamente substitui o vetor de características visuais durante a fase de pesquisa. De forma a traduzir esta similaridade numa representação indexável por motores de pesquisa textuais, é necessário atribuir um identificador (id) a cada pivô. Para gerar a STR a partir de um vetor com P pivôs ordenados, é inicialmente criada uma string vazia à qual cada identificador irá ser adicionado $P-R+1$ vezes, onde R representa a ordem de similaridade do pivô.

Este algoritmo pode ser explicado através de um exemplo adequado: para uma imagem com vetor de similaridade ordenada [B, C, A], isto é, sendo o vetor B o mais próximo da imagem e o vetor A o mais afastado, a STR correspondente será "B B B C C A". Quando analisada por um motor de indexação textual tradicional, baseado na representação vetorial de termos em documentos, a STR representará efetivamente o peso de cada ponto de referência no cálculo de similaridade entre imagens, traduzido na comparação dos campos STR de cada uma [ABF+11].

A vantagem desta estratégia reside na robustez e popularidade das plataformas de recuperação de informação textual atualmente disponíveis, algumas já preparadas para processamento distribuído e paralelo dos dados. Assim sendo, o mapeamento de características visuais em conteúdo textual acelera a introdução das capacidades de pesquisa visual por conteúdo em sistemas de recuperação textual. Esta possibilidade torna-se mais importante em ambientes nos quais é difícil ou até mesmo impossível de instalar e integrar soluções baseadas em software adicional específico, que carece de verificação e validação a nível de segurança e compatibilidade. As STRs das imagens podem ser geradas de forma independente e indexadas como um campo extra juntamente com os metadados já existentes, dado que as pesquisas visuais se traduzem em pesquisas textuais focadas nos campos que armazenam as representações visuais.

Neste trabalho, a determinação do número de pivôs e do tamanho dos prefixos foram efetuadas da seguinte forma: para uma coleção com D documentos, foram escolhidos 10 grupos de $P = \sqrt{D}$ pivôs. Determinou-se a distância interna de cada grupo, isto é, a soma das distâncias entre cada pivô, e foi escolhido o grupo com

o maior valor associado, evidenciando uma maior cobertura do espaço vetorial associado. Após a geração dos vetores de similaridade ordenada para cada imagem, estes foram truncados aos primeiros $\sqrt{P}$ pivôs, estabelecendo-se assim o tamanho dos prefixos. Finalmente, foram geradas as STRs das imagens, sendo armazenadas numa base de dados para posterior indexação com os restantes metadados.

4.6 Reordenação transversal multimédia

Com o conteúdo multimédia totalmente indexável por plataformas de recuperação de informação textual tradicionais, é possível tirar partido das estratégias atualmente disponíveis para melhorar e acelerar as interrogações, tais como a utilização de "caching".

Neste trabalho, é apresentado um modelo em "cascata" para a resolução da tarefa de ilustração de texto. Neste modelo, o conteúdo textual é utilizado para delimitar elementos de contexto tais como entidades, locais e datas contidas nas legendas das fotos e títulos. As propriedades visuais são usadas para ordenar as imagens por *coerência visual*. Esta coerência é definida como a soma das similaridades de uma imagem a todas as outras obtidas no conjunto resultante da pesquisa textual. As fotos mais similares entre si serão agrupadas no topo da lista de resultados, enquanto que as fotos mais distintas no conjunto serão colocadas no final da lista. Considerando uma perspetiva de análise de teoria de grafos, assumindo que cada imagem é efetivamente um nó, conforme a Figura 4.7, e que as similaridades representam ligações pesadas entre esses nós, a coerência visual traduz-se na determinação da *weighted degree centrality* [OAS10] de cada foto.

O processo tri-faseado de recuperação de informação multimédia (ilustrado na Figura 4.1) efetuado sobre as coleções indexadas resulta em listas de imagens candidatas à ilustração do texto introduzido. O primeiro passo traduz-se numa pesquisa textual com um máximo de 100 imagens.

O segundo passo inclui uma filtragem dessas imagens que, acima de um mínimo de 10 itens, exclui todos os que obtiveram uma pontuação inferior a 50% da pontuação do primeiro item. Este filtro procura reforçar o objetivo de que o sistema selecione apenas itens altamente relacionados com o texto introduzido, de acordo

com a informação textual e metadados associados a cada foto.

O terceiro passo consiste na ordenação visual dos resultados obtidos após os dois passos anteriores. Cada imagem é comparada com todas as outras através da sua STR, vetores binários ou vetores de características originais.

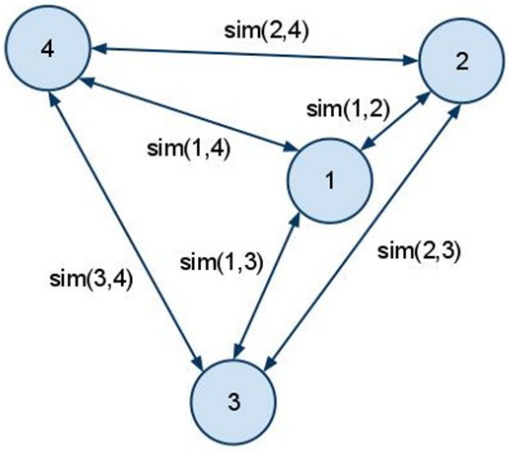


Figura 4.7: Perspetiva de teoria de grafos

Ao adicionar as pontuações de similaridade para cada imagem, determinando o seu grau de centralidade, e ordenando-as por valor decrescente do mesmo, as imagens mais representativas do conjunto obtido serão atraídas para o topo da lista de resultados, enquanto que os "outliers" serão arrastados para o final da mesma. As fotos de eventos específicos terão características visuais similares, enquanto que fotos visualmente distintas do conjunto poderão pertencer a eventos não-relacionados que não foram removidos com sucesso pela filtragem de pontuações efetuada no segundo passo.

Capítulo 5

Implementação e avaliação

> "Done is better than perfect."
>
> *Scott Allen*

Neste capítulo são apresentados os dois protótipos de recuperação de informação multimédia desenvolvidos ao longo deste trabalho com o objetivo de auxiliarem a tarefa de ilustração automática de conteúdos textuais.

5.1 Recuperação de informação visual

O protótipo inicial, com ênfase na exploração de coleções de imagens de larga-escala, incidiu sobre as coleções de fotos MIRFlickr-25k e de imagens médicas IRMA-2007. O segundo protótipo, já designado como sistema interativo de ilustração automática de texto, teve por base a coleção SAPO-Lusa e serviu como prova de conceito dos algoritmos de pré-processamento, extração, descrição, indexação e pesquisa multimodal baseada no conteúdo e contexto dos documentos multimédia existentes.

5.1.1 Protótipo de pesquisa de imagens médicas

Para a indexação da coleção IRMA-2007, nomeadamente dos metadados textuais resultantes da conversão dos códigos IRMA existentes em anotações textuais descritivas, foi utilizada a biblioteca de recuperação de informação textual *Lucene*[1]. Este módulo do sistema indexa as anotações, efetuando o *parsing* e normalização das

[1] `http://lucene.apache.org/`

palavras.

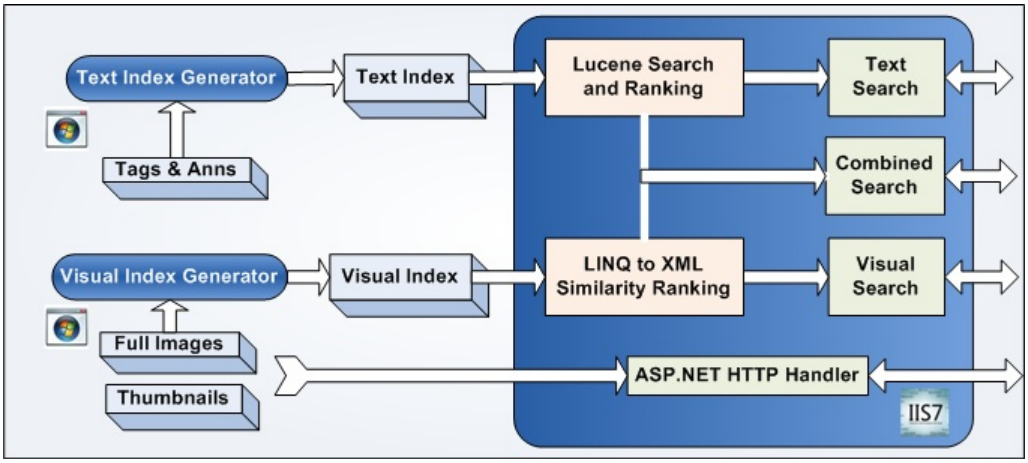
Os vetores de características, representativos das propriedades visuais extraídas de cada imagem, foram gerados pela biblioteca *img(Rummager)* [CBL09], de forma a tornar possível a determinação da similaridade e ordenação de resultados durante a fase de pesquisa visual com base em conteúdo. O módulo de descrição visual gera os vetores e armazena-os num único ficheiro XML.

Para efeitos de desempenho do sistema, o ficheiro XML contendo os descritores foi posteriormente convertido num ficheiro serializado, o que permitiu acelerar de forma considerável o carregamento do índice visual. Testes efetuados demonstraram que, em média, as pesquisas visuais baseadas na leitura de conteúdo XML demoravam 3 a 15 segundos, dependendo do tamanho dos descritores escolhidos. As pesquisas usando o ficheiro serializado atingiam os 0.1 a 1 segundos, evitando o *parsing* do ficheiro XML e a conversão dos valores armazenados em objetos da plataforma de desenvolvimento.

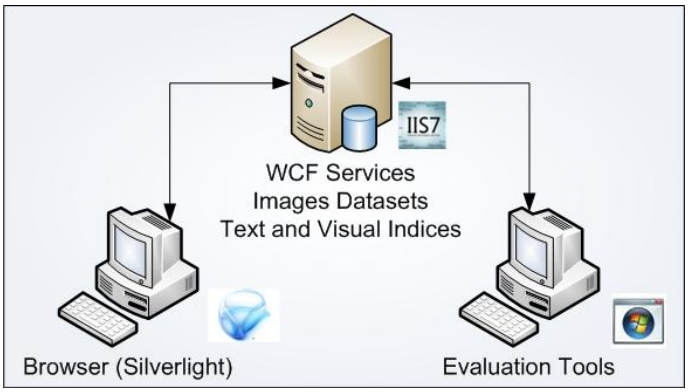
De forma a exemplificar os detalhes da arquitetura adotada, a Figura 5.1(a) demonstra as aplicações e dados necessários ao funcionamento do protótipo de recuperação de informação multimédia. Duas aplicações de linha de comandos geram os índices visuais e textuais que alimentam os três serviços de pesquisa. São também disponibilizadas miniaturas das imagens, de forma a acelerar a visualização dos resultados de pesquisa e reduzir o tempo de espera.

A Figura 5.1(b) exibe alguns aspetos da arquitetura do protótipo, com um servidor responsável pelos dados e lógica de negócio e por disponibilizar os serviços de interação para pesquisas (*searches*), avaliação e validação das funcionalidades do sistema.

A interface de utilização do sistema é executada nas máquinas dos utilizadores e tira partido dos serviços web disponíveis. A Figura 5.2 demonstra a interface do protótipo desenvolvido. A pesquisa textual está disponível, representando esta um passo obrigatório para efetuar posteriormente pesquisas visuais. É possível selecionar os descritores pretendidos para comparação de conteúdos visuais, assim como combinar os resultados das pesquisas textuais e visuais. A lista de resultados, do lado esquerdo, contém miniaturas, códigos e anotações das imagens, e o lado direito



(a) Arquitetura de *backend*



(b) Visão geral do sistema

Figura 5.1: O processo de recuperação de informação multimédia

exibe a imagem selecionada, escalada de forma a manter o seu formato inicial.

5.1.2 Protótipo de ilustração de texto

Os desenvolvimentos subsequentes efetuados no protótipo incidiram sobre a tarefa de ilustração de conteúdos textuais jornalísticos. a Após o processamento da informação de conteúdo e metadados existentes na coleção SAPO-Lusa, o protótipo é capaz de permitir a realização de pesquisas textuais e visuais de forma a efetuar ou auxiliar a ilustração automática de texto, disponibilizando ainda a capacidade de exploração visual desta coleção de larga-escala.

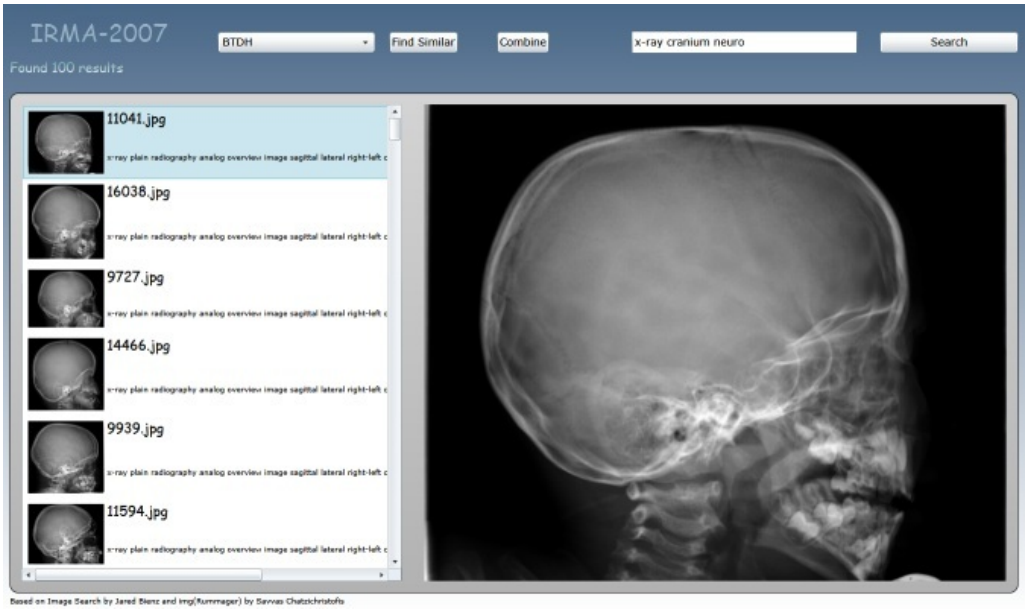


Figura 5.2: Interface de utilizador

Os utilizadores podem interagir com o protótipo de forma a refinar os resultados da pesquisa para ilustração, usando as imagens obtidas de uma execução automática. Se uma foto for considerada relevante, é possível reordenar as restantes por similaridade visual, na tentativa de agrupar fotos do mesmo evento, contexto, tema ou apenas devido às suas características visuais específicas.

Além disso, os utilizadores podem escolher uma foto específica e explorar a coleção com base nas suas propriedades visuais, ultrapassando a barreira contextual imposta pelo processo de pesquisa de fotos para ilustração baseada apenas no conteúdo textual associado. Por exemplo, fotos genéricas de um estádio ou campo de futebol poderão ser consideradas relevantes para textos associados a um jogador de futebol. Estas imagens, que podem eventualmente ter sido excluídas no processo de filtragem de resultados por contexto, podem ser recuperadas através da exploração visual da coleção.

A utilização dos índices de similaridade aproximada descritos no capítulo anterior permite obter resultados de uma forma muito mais rápida do que através de comparações sequenciais exaustivas percorrendo toda a coleção. O resultado final é ordenado visualmente por coerência visual, permitindo ao utilizador encontrar as imagens mais representativas no topo dos resultados, e as imagens mais distintas no final dessa lista.

O protótipo possui uma interface simples e intuitiva, exibida na Figura 5.3. As duas opções de recuperação de informação disponíveis são "Ilustrar", a qual efetua apenas uma pesquisa textual com filtragem de resultados por contexto, e "Ilustrar e Agrupar", a qual efetua também o passo de ordenação por coerência visual (Secção 4.6), combinando assim a pesquisa textual e as propriedades visuais das fotos.

A exploração das características visuais é disponibilizada através das opções "Reordenação visual", a qual reordena os resultados por similaridade com uma foto específica, e "Encontrar Semelhantes", a qual se baseia exclusivamente nas propriedades visuais das fotos para pesquisar toda a coleção por imagens semelhantes.

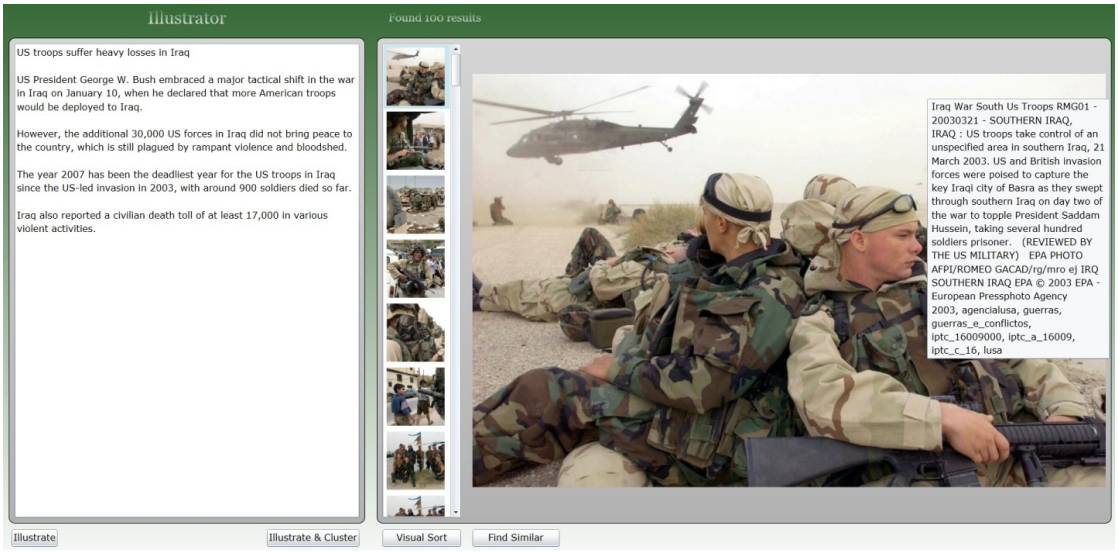
A nível técnico, a implementação da interface for realizada em Silverlight[2], o que permitiu a criação de uma aplicação a executar dentro dos browsers, comunicando com os serviços web WCF[3]. Foi posteriormente disponibilizada uma demonstração pública deste protótipo usando a coleção SAPO-Lusa, através do projeto "Dpikt"[4].

5.2 Avaliação

Foram realizadas experiências de avaliação para comprovar a eficácia da arquitetura proposta, nas perspetivas de qualidade dos resultados e desempenho do sistema. A realização em paralelo de um estudo com utilizadores, recorrendo a uma plataforma de *crowdsourcing*, permitiu averiguar se os resultados provenientes das estratégias de validação seriam coincidentes com os obtidos por utilizadores reais. O impacto dos filtros de abstração foi analisado através da realização de pesquisas visuais, usando o *groundtruth* disponível para determinar o seu impacto e potenciais benefícios no armazenamento e processamento de imagens.

As experiências de avaliação incidiram sobre as coleções MIRFlickr-25k e SAPO-Lusa, sendo estas bastante diferentes na sua natureza e objetivos, conforme discutido na Secção 3.1. A primeira experiência procura avaliar a qualidade dos resultados, validando-os em cada uma das três fases do processo de recomendação de fotos. A

[2] http://www.silverlight.net/
[3] http://msdn.microsoft.com/en-us/library/ms731082.aspx
[4] http://labs.sapo.pt/2012/12/dpikt-feup/



(a) Interface de utilizador

(b) Opções de recuperação e exploração visuais

Figura 5.3: Protótipo de ilustração automática e interativa

segunda experiência está focada na avaliação do desempenho do esquema de pesquisa aproximada proposto para a rápida exploração de coleções de grande escala através de similaridade visual.

5.2.1 Experiências na coleção MIRFlickr-25k

Para efeitos de avaliação, a coleção MIRFlickr-25k representa uma coleção de fotos frequentemente utilizada na área de avaliação de sistemas de recuperação de informação multimédia, mas que não possui conteúdo textual significativo para cada foto. Os 24 conceitos pré-estabelecidos (Figure 3.2(b)) para esta coleção foram usados como interrogações.

As anotações, efetuadas manualmente, denotam a existência de conceitos visuais específicos presentes nas imagens. Estas incluem também *tags* inseridas pelos utilizadores, o que permitiu definir a estratégia da primeira experiência de avaliação: efetuar pesquisas por conceitos nas *tags* associadas às fotos (Figura 3.3(b)), considerando as anotações relativas a conceitos como *groundtruth*. Uma foto é considerada

relevante se o conceito pesquisado fizer parte da sua lista de conceitos.

5.2.1.1 Qualidade da recomendação de fotos

A experiência de avaliação da recomendação de fotos para ilustração utilizou os 24 conceitos como interrogações. Os resultados foram avaliados após cada uma das etapas do processo de recuperação de informação: pesquisa textual sobre os metadados indexados; pesquisa refinada pelo filtro de pontuação, e a aplicação do algoritmo de coerência visual.

As medidas de avaliação usadas, nomeadamente a precisão nos 10 primeiros resultados (Prec@10), precisão nos R primeiros resultados (R-Prec) e a precisão média (MAP) permitiram obter os valores representados na Tabela 5.1, para a coleção MIRFlickr-25k. Verifica-se que os conceitos visuais definidos nas anotações estão presentes nas *tags* dos utilizadores, dado que o valor de precisão é superior a 70%. A precisão aumenta com a aplicação do filtro de pontuação, e ainda mais com a aplicação do algoritmo de coerência visual.

Tabela 5.1: Recomendação de fotos na coleção MIRFlickr-25k

Avaliação	textual	c/ filtro$_{pontuacao}$	c/ filtro & coerência$_{visual}$
Prec@10	70%	70%	**73%**
R-Prec	71%	71%	71%
MAP	72%	73%	**78%**

5.2.1.2 Desempenho da exploração visual

A experiência de avaliação da exploração visual, através de pesquisas baseadas somente no conteúdo das fotos, compara a execução de interrogações visuais usando os esquemas de pesquisa sequencial e aproximada apresentados neste trabalho. O conjunto de interrogações visuais é formado pelas fotos pertencentes ao grupo de imagens de cada conceito. Uma imagem é considerada relevante para uma pesquisa visual se pertencer ao *groundtruth* do conceito associado.

Dado que cada imagem pode incluir vários conceitos visuais, os quais são analisados individualmente, a redundância de interrogações é evitada através da uma

verificação inicial. Desse modo foram realizadas apenas 25.000 interrogações e não 73.342, reduzindo o tempo necessário para a obtenção de resultados.

As posições das imagens nos resultados das pesquisas sequenciais e aproximadas foram comparadas através do coeficiente de correlação de Spearman. Para obter o seu valor, são usadas as posições r_{seq} sequencial e r_{approx} aproximada das imagens nas listas de resultados, e o tamanho n dessas listas:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$
$$\text{onde } d_i = r_{seq_i} - r_{approx_i}$$

Após a realização das interrogações, os tempos médios de resposta foram analisados para a coleção MIRFlickr-25k, com os resultados visíveis na Tabela 5.2. As pesquisas usando os algoritmos de pesquisa aproximada obtiveram um valor de correlação de 99% com os resultados obtidos através das pesquisas exaustivas sequenciais.

Tabela 5.2: Exploração visual na coleção MIRFlickr-25k

Avaliação	pesquisa$_{exaustiva}$	pesquisa$_{aproximada}$
Correlação	**99%**	
Tempo médio de resposta	8 seg	**1 seg**

5.2.2 Experiências na coleção SAPO-Lusa

A coleção SAPO-Lusa, uma coleção de fotos jornalísticas de larga-escala, utilizada para projetos de investigação realizados no âmbito dos Laboratórios SAPO[5], resulta de um acordo estabelecido entre a agência de notícias *Agência Lusa* e a empresa SAPO (Secção 3.1.3).

Esta coleção torna-se interessante do ponto de vista da tarefa de ilustração de conteúdos textuais dado que as fotos foram enriquecidas com descrições detalhadas, mas sem o objetivo de a preparar para efeitos de avaliação de algoritmos de pesquisa. Os metadados associados e anotações existentes foram utilizados para estabelecer

[5] `http://labs.sapo.pt/`

o *groundtruth* das interrogações a realizar, dado que seria extremamente custoso defini-lo manualmente numa coleção de tão grande escala como a SAPO-Lusa, com cerca de dois milhões de fotos.

Dado que a coleção não se encontrava preparada para o processo de avaliação, foram definidas 100 personalidades do domínio público como referência para as pesquisas. As fotos contêm legendas e títulos associados, o que permite usar estes últimos, mais específicos, assumindo que uma imagem é considerada relevante quando o seu título contém o nome da personalidade usada para pesquisa. Exemplos das personalidades escolhidas como interrogações podem ser observados na Figura 4.5.

5.2.2.1 Qualidade da recomendação de fotos

Esta experiência é similar à efetuada sobre a coleção MIRFlickr-25k. O sistema respondeu a interrogações relativas às 100 personalidades, produzindo resultados para os três tipos de pesquisa, nomeadamente pesquisa textual simples, pesquisa textual filtrada pelo filtro de pontuação, e pesquisa textual filtrada seguida de ordenação por coerência visual.

Como pode ser observado na Tabela 5.3, os resultados após filtragem de pontuações obtiveram uma maior precisão do que os provenientes de pesquisas textuais simples. A aplicação de coerência visual também demonstrou um aumento de precisão relativamente às pesquisas textuais originais, mas não tão elevado como aplicando apenas o filtro de pontuação.

Tabela 5.3: Recomendação de fotos na coleção SAPO-Lusa

Avaliação	textual	c/ filtro$_{pontuacao}$	c/ filtro & coerência$_{visual}$
Prec@10	38%	**43%**	39%
R-Prec	40%	**42%**	41%
MAP	39%	**47%**	43%

5.2.2.2 Desempenho da exploração visual

A segunda parte dos testes comparou as interrogações visuais baseadas em pesquisas exaustiva sequenciais e recorrendo aos índices de pesquisa aproximada. Foi

efetuada uma interrogação visual para cada uma das imagens pertencentes às personalidades. O módulo de avaliação realizou um total de 19.815 interrogações (100 personalidades × fotos respetivas), que originaram os resultados exibidos na Tabela 5.4.

Tabela 5.4: Exploração visual na coleção SAPO-Lusa

Avaliação	pesquisa$_{exaustiva}$	pesquisa$_{aproximada}$
Correlação	**99%**	
Tempo médio de resposta	60 seg	**3 seg**

5.2.3 Experiências com *crowdsourcing*

O objetivo do algoritmo de coerência visual é agrupar imagens similares no topo da lista de resultados, colocando as fotos visualmente mais distintas no final da mesma. Foi necessário avaliar este conceito através de utilizadores reais, tirando partido de uma das plataformas de *crowdsourcing* disponíveis atualmente, designada por CrowdFlower[6]. Foi concebida uma experiência de avaliação baseada na comparação de listas de sugestões de fotos para ilustração para fragmentos textuais específicos.

Foram definidos 12 fragmentos, evidenciados na Tabela 5.5, os quais procuram abranger a diversidade de tópicos presentes na coleção SAPO-Lusa.

Tabela 5.5: Crowdsourcing - Interrogações para ilustração

ID	interrogação
1	president barack obama speaks to the audience in a speech
2	the france presidential elections in 2007
3	EU family photo after conferences
4	Cristiano Ronaldo playing in the portuguese national team
5	soccer matches during the european 2008 championship
6	swimming during the Olympic Games
7	movie actress Angelina Jolie
8	Grammy Awards in 2006
9	stage performance during Rock In Rio
10	US soldiers during the Iraq invasion
11	soldiers and vehicles in a military parade
12	the North and South Korea conflict

[6] `http://crowdflower.com/`

Para cada tarefa de ilustração, foram geradas 5 listas de resultados com 10 fotos cada: Aleatória, Pesquisa Textual, Reordenada, Coerência e Coerência Inversa:

- A lista Aleatória contém 3 fotos resultantes da pesquisa textual, misturadas com 7 fotos escolhidas aleatoriamente na coleção. A lista Pesquisa Textual resultou dos primeiros 10 resultados da pesquisa textual efetuada.
- A lista Reordenada foi criada reordenando as fotos por similaridade ao primeiro resultado da lista completa de pesquisa textual.
- A lista Coerência foi gerada com base no algoritmo de coerência visual, que agrupa as fotos mais similares entre si no início da lista de resultados, limitando-os posteriormente aos primeiros 10 itens.
- A lista Coerência Inversa contém os 10 últimos resultados da lista Coerência completa, por ordem inversa, ou seja, as fotos visualmente mais distintas entre os resultados obtidos.

A lista Aleatória for criada essencialmente para validar os avaliadores na plataforma de *crowdsourcing* adotada. Foram criadas tarefas "gold", idênticas às tarefas de ilustração, onde é exibido o conteúdo textual a ilustrar e duas listas de resultados. Nestas tarefas de validação, uma das listas é a Aleatória, sendo que se espera que o avaliador atento opte sempre pela lista alternativa. Desta forma, é possível excluir do processo de avaliação da tarefa avaliadores incapazes de efetuar julgamentos de relevância, respondendo aleatoria ou incoerentemente às tarefas de validação que lhes são propostas.

A experiência de *crowdsourcing* é composta por 48 tarefas de ilustração, isto é, 12 conteúdos textuais com quatro comparações cada:

- lista Aleatória e lista Pesquisa Textual (tarefa "gold");
- lista Pesquisa Textual e lista Reordenada;
- lista Pesquisa Textual e lista Coerência;
- lista Coerência e lista Coerência Inversa.

Foi estabelecido um mínimo de 7 julgamentos por tarefa, ou seja 252 (7 * 36, dado que as tarefas "gold"não são consideradas) para obter *feedback* considerado satisfatório. Foram recolhidos 1.255 julgamentos, dos quais 276 foram considerados válidos pela plataforma, e 888 julgamentos foram invalidados face às respostas erradas desses avaliadores nas tarefas "gold". Este facto revela que uma grande percentagem não foi considerada, sendo excluída do processo e do respetivo pagamento pelas avaliações desempenhadas.

O processo ocorreu durante 3 horas. As tabelas 5.6, 5.7 e 5.8 apresentam resumidamente os resultados obtidos:

Tabela 5.6: Crowdsourcing - Pesquisa textual (A) e Ordenada (B)

ID	Lista A	%	Lista B	%	Indiferente	%
1	16	55%	7	24%	6	21%
2	15	44%	10	29%	9	26%
3	18	55%	9	27%	6	18%
4	9	36%	8	32%	8	32%
5	12	40%	10	33%	8	27%
6	7	58%	5	42%	0	0%
7	8	40%	6	30%	6	30%
8	15	48%	9	29%	7	23%
9	11	52%	6	29%	4	19%
10	10	40%	10	40%	5	20%
11	21	49%	12	28%	10	23%
12	16	41%	13	33%	10	26%
média	–	47%	–	31%	–	22%
desvio	–	7%	–	5%	–	8%

5.2.4 Análise de resultados

A fase de avaliação produziu um número elevado de dados, entre os quais valores relativos ao desempenho do sistema e *feedback* dos utilizadores. Nesta secção são discutidas as conclusões mais significativas, os tópicos mais relevantes e as perspetivas de aplicação destas técnicas para o suporte da tarefa de ilustração automática de conteúdos textuais.

Tabela 5.7: Crowdsourcing - Pesquisa Textual (A) e Coerência Visual (B)

ID	Lista A	%	Lista B	%	Indiferente	%
1	9	60%	5	33%	1	7%
2	6	35%	6	35%	5	29%
3	11	58%	6	32%	2	11%
4	10	45%	8	36%	4	18%
5	12	60%	6	30%	2	10%
6	9	64%	3	21%	2	14%
7	15	45%	9	27%	9	27%
8	13	39%	13	39%	7	21%
9	10	42%	8	33%	6	25%
10	46	40%	29	40%	25	20%
11	12	39%	10	32%	9	29%
12	19	49%	12	31%	8	21%
média	–	49%	–	32%	–	20%
desvio	–	10%	–	5%	–	8%

Tabela 5.8: Crowdsourcing - Coerência Visual (A) e Coerência Visual Inversa (B)

ID	Lista A	%	Lista B	%	Indiferente	%
1	12	52%	6	26%	5	22%
2	18	67%	6	22%	3	11%
3	16	43%	14	38%	7	19%
4	7	50%	4	29%	3	21%
5	15	58%	9	35%	2	8%
6	16	57%	7	25%	5	18%
7	6	43%	6	43%	2	14%
8	18	51%	12	34%	5	14%
9	16	64%	6	24%	3	12%
10	12	46%	9	35%	5	19%
11	8	40%	7	35%	5	25%
12	16	42%	11	29%	11	29%
média	–	51%	–	31%	–	18%
desvio	–	9%	–	6%	–	6%

5.2.4.1 Recomendação de fotos

A experiência de avaliação da recomendação de fotos é adequada para a coleção MIRFlickr-25k, concebida para investigação na área de recuperação de informação visual. No entanto, esta coleção apresenta limitações relevantes relativamente à tarefa de ilustração de conteúdos textuais: as interrogações estão limitadas a conceitos exprimíveis apenas por palavras-chave, o que não é representativo dos textos espe-

rados.

Por seu lado, a pesquisa de personalidades efetuada sobre a coleção SAPO-Lusa revelou ser um bom modelo de testes para a tarefa considerada. Uma coleção fotojornalística desta natureza possui informação diversificada sobre as personalidades pesquisadas, e as legendas das fotos permitem uma pesquisa textual eficaz durante o primeiro passo do processo de ilustração.

A limitação no uso desta coleção está relacionado com a inexistência de conceitos, classes ou interrogações, com o respetivo *groundtruth*. Esta dificuldade foi ultrapassada através da definição de interrogações baseadas nas personalidades existentes, e do *groundtruth* com base no uso do título das fotos como fonte de informação separada, para validação da existência de personalidades nas mesmas. Esta abordagem resultou em valores de precisão baixos, os quais, após observação atenta dos resultados das pesquisas, se devem ao facto de que várias legendas de fotos, apesar de mencionarem as personalidades pesquisadas, não justificam a sua presença nas fotos, mas sim eventos e personalidades relacionadas.

Em ambas as coleções, os resultados de utilização dos passos relativos aos filtros de pontuação e coerência visual ultrapassam os obtidos efetuando apenas uma pesquisa textual. Na coleção SAPO-Lusa, a aplicação da coerência visual após o filtro de pontuação não melhorou os resultados obtidos, o que pode ser explicado pela influência do *groundtruth* definido e a natureza da reordenação visual. Os exemplos apresentados na Figura 4.6 mostram que imagens visualmente similares podem corresponder à mesma personalidade, mas numa coleção onde fotos do mesmo evento possam estar presentes, a coerência visual pode promover, baseada nos critérios de conteúdo visual, imagens de diferentes personalidades em contextos visuais semelhantes.

5.2.4.2 Exploração visual

A comparação efetuada entre os resultados obtidos com as pesquisas exaustivas e as baseadas nos índices aproximados demonstrou que não houve perda significativa na qualidade das respostas. Os testes de desempenho utilizaram um número elevado de interrogações visuais, e o *groundtruth* já apresentado. Os valores de precisão a 10 resultados foram aproximadamente 43% na coleção MIRFlickr-25k e 13% para a

coleção SAPO-Lusa. Esta diferença detetada no comportamento do descritor JCD usando o *groundtruth* definido pode ser explicada pelo facto de que é mais fácil associar os conceitos genéricos presentes na coleção MIRFlickr-25k com as características visuais de baixo nível como cor e textura, do que detetar personalidades através da mesma abordagem.

O descritor visual foi concebido para ser mais eficaz quando as classes a identificar são visualmente distintas, com elevada similaridade entre as imagens da mesma classe, e quando os objetos de interesse ocupam uma porção considerável das fotos. A Figura 4.6 demonstra exemplos de exploração visual efetuados sobre a coleção SAPO-Lusa, evidenciando que as imagens podem ser visualmente muito similares no seu contexto visual e não nas personalidades que nelas podem ser observadas.

5.2.5 Pesquisa visual em imagens abstraídas

Para testar o impacto do filtro de abstração, foi utilizada a coleção MIRFlickr-25k e o descritor JCD, seguindo a abordagem previamente usada na comparação de descritores com a coleção IRMA-2007. Os tópicos pré-estabelecidos, exemplificados na Figura 3.2(b) resultam de anotações disponibilizadas juntamente com as fotos da coleção e foram manualmente atribuídos por um anotador responsável por determinar a presença ou ausência dos tópicos nas imagens.

O conjunto de interrogações visuais para um tópico é formado por todas as imagens pertencentes ao mesmo, definidas no seu *groundtruth*. Uma foto é considerada relevante na lista de resultados de uma pesquisa específica se a mesma pertencer ao *groundtruth* do tópico correspondente. Dado que nesta coleção uma imagem pode estar anotada com vários tópicos, esta pode pertencer a vários conjuntos de interrogações visuais resultando em pesquisas duplicadas mas com *groundtruth* distinto. No total, foram efetuadas 73.342 interrogações visuais, cujos resultados estão representados na Tabela 5.9, nomeadamente as pontuações MAP.

A Figura 5.4 mostra exemplos das pesquisas visuais, onde as fotos no topo de cada coluna representam a imagem usada para interrogação visual. Os resultados demonstram que o processo de abstração de imagens não diminui significativamente a qualidade dos resultados das pesquisas visuais segundo o paradigma de avaliação

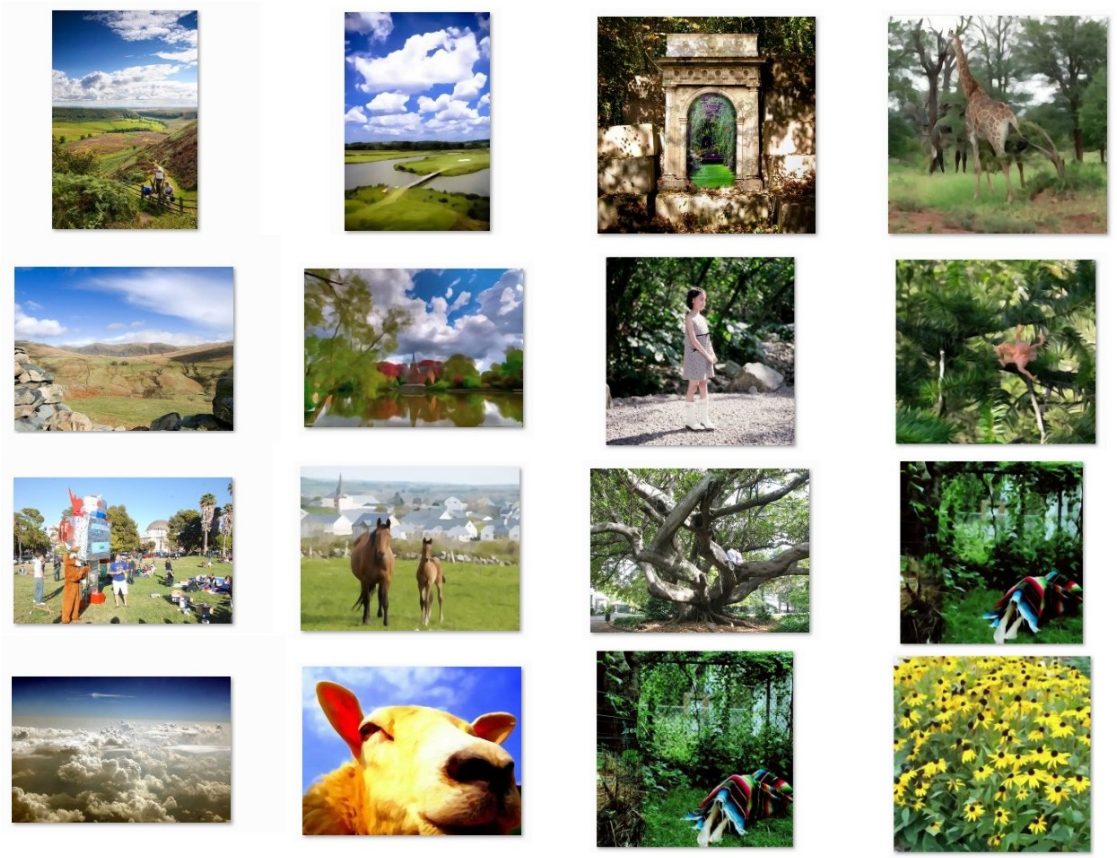
Tabela 5.9: Resultados das pesquisas por conteúdo (melhores resultados a negrito)

Tópico	JCD	JCD c/ AKF
animals	**28.4**%	28.0%
baby	**65.2**%	62.8%
bird	**41.5**%	40.1%
car	47.4%	**47.8**%
clouds	31.7%	**34.1**%
dog	36.5%	**38.0**%
female	**31.2**%	30.6%
flower	**34.8**%	34.5%
food	34.5%	**34.6**%
indoor	**45.1**%	44.9%
lake	33.7%	**34.4**%
male	28.8%	**29.2**%
night	**35.7**%	34.0%
people	**44.1**%	43.8%
plantlife	**49.6**%	48.9%
portrait	**32.4**%	31.8%
river	**64.2**%	63.4%
sea	47.5%	**49.8**%
sky	46.2%	**46.9**%
structures	**50.7**%	49.2%
sunset	32.7%	**34.0**%
transport	27.5%	*27.6*%
tree	**34.2**%	33.7%
water	31.0%	**31.3**%
Average	**39.8**%	39.7%

adotado, e em alguns casos pode até obter resultados marginalmente melhores.

Com base nos resultados obtidos, o pré-processamento de imagens não degradou o processo de recuperação de informação. Desta forma, através do filtro de abstração, é possível reduzir substancialmente os requisitos de armazenamento de uma coleção multimédia sem comprometer a qualidade dos resultados obtidos.

(a) Interrogações



(b) Resultados

Figura 5.4: Exemplos de pesquisa visual: imagens originais (1ª e 3ª colunas) e imagens abstraídas (2ª e 4ª colunas)

Capítulo 6

Adaptação para recomendação musical

> "There is nothing new under the sun."
>
> *Bible, Ecclesiastes 1:9*

Os algoritmos, técnicas e metodologias propostas e analisadas nos capítulos anteriores foram aplicados de forma transversal na área de recuperação de informação musical em coleções de larga-escala, com ênfase na recomendação e descoberta musicais. O terceiro protótipo desenvolvido serviu de base para a realização de um projeto de investigação na área de recomendação híbrida musical, denominado *Juggle*[1], combinando análise de conteúdo, contexto e filtragem colaborativa para a geração de playlists, em colaboração com o colega de doutoramento José Luís Devezas[2] desde Dezembro de 2012.

6.1 Recomendação

As redes sociais introduziram uma mudança na análise de informação de gostos musicais, através do *feedback* dos utilizadores e não diretamente do processamento de conteúdo áudio, num esforço de amenizar os requisitos de hardware e tirar partido da esparsidade dos dados.

[1] `http://labs.sapo.pt/2012/12/juggle-feup/`
[2] `http://josedevezas.com/`

No entanto, a consequência imediata do foco na informação dos utilizadores, na forma de dados implícitos (utilização do software) e explícitos (pontuações), é a de que os sistemas de recomendação que dependem exclusivamente de algoritmos de filtragem colaborativa sofrem do designado problema de *cold start*, isto é, a dificuldade em obter resultados iniciais na ausência de informação sobre os utilizadores. O sistema fica incapacitado de modelar os gostos dos seus utilizadores sem pontuações e preferências, e consequentemente de recomendar itens relevantes.

Além disso, ao ignorar informação adicional de contexto e conteúdo, este tipo de sistemas de recomendação tem a tendência para restringir os utilizadores aos seus gostos, não sendo capazes de sugerir itens potencialmente relevantes que estes gostariam mas sobre os quais não têm ainda conhecimento.

Mesmo que os problemas de processamento possam ser atenuados com base na esparsidade dos dados relativos aos utilizadores, a diversificação de resultados e exploração de potenciais itens relevantes são algumas das estratégias que podem melhorar significativamente a experiência de utilização e satisfação dos utilizadores com sistemas de recomendação. Estas áreas podem usufruir da pesquisa baseada em conteúdos, de forma a expandir os horizontes dos gostos dos utilizadores.

A área de recuperação de informação musical [LME+11], com ênfase nas tarefas de geração de playlists [Fie11] e descoberta de músicas, permitiu demonstrar a transversalidade dos algoritmos de pesquisa aproximada e reordenação multimédia propostos para a ilustração automática de textos com coleções de imagens de larga-escala. A abordagem seguida é a de análise de conteúdo multimédia, a qual foi testada com sucesso na exploração visual e recomendação de fotos [CR11a].

Um problema inerente aos sistemas de recomendação está relacionado com a complexidade dos processos de extração e indexação de características multimédia, exigindo um poder de processamento e armazenamento bastante superiores às abordagens centradas no *feedback* dos utilizadores, de forma a ser possível analisar e pesquisar os dados disponíveis. A estratégia adotada pode ser vista como um protótipo independente de pesquisa de informação musical, e como base para um sistema híbrido de recomendação, capaz de considerar a informação de conteúdo e de filtragem colaborativa.

A visão proposta está representada na Figura 6.1. Esta abordagem é iniciada com uma interrogação textual onde os utilizadores inserem palavras-chave, títulos de canções, nomes de artistas ou mesmo excertos das músicas. As *playlists* resultantes agrupam músicas semelhantes, adaptando o conceito de coerência visual à dimensão musical. As músicas mais semelhantes entre si aparecem no início da *playlist*, movendo as músicas mais díspares para o final da lista. Seguindo um procedimento idêntico à reordenação visual apresentada nos protótipos anteriores, os utilizadores podem reordenar as *playlists* por proximidade. Desta forma, cada música é seguida pela mais similar que ainda não tenha sido ouvida, encadeando a similaridade áudio. Por último, de forma a permitir a descoberta de músicas novas, os utilizadores podem selecionar uma canção específica e pesquisar toda a coleção com base nas características áudio, ou usando a similaridade textual das letras e *tags*.

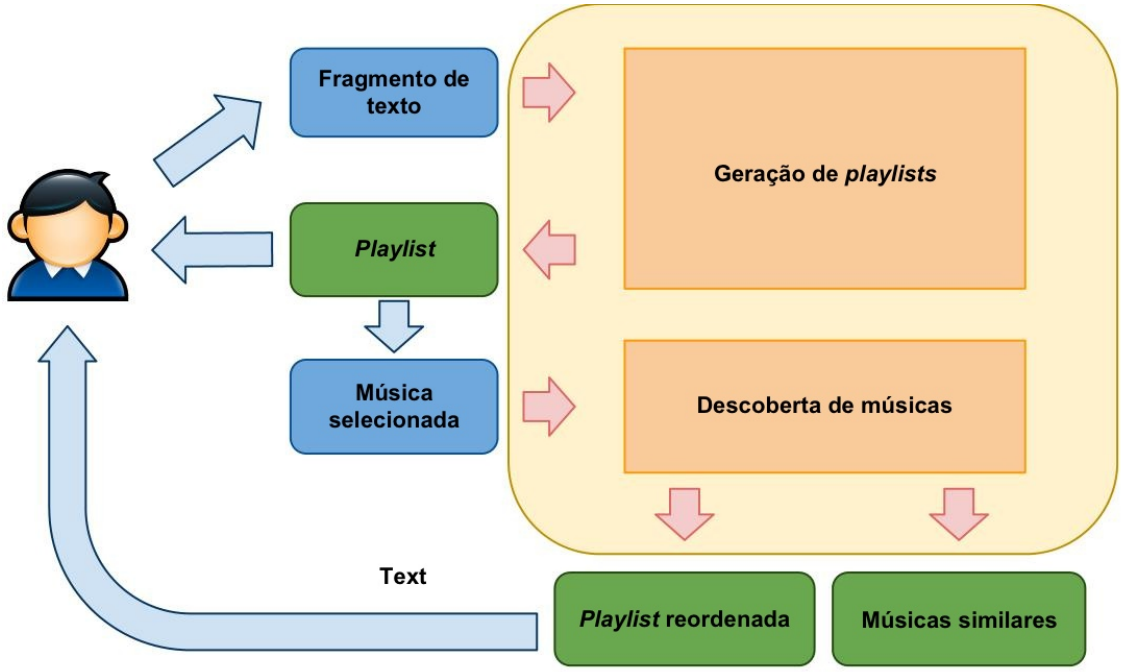


Figura 6.1: Recomendação musical

6.2 A coleção *Million Song Dataset*

A coleção *Million Song Dataset*[3] (MSD), disponibilizada em 2011, é uma coleção livre de características áudio e metadados relativos a um milhão de músicas popula-

[3] http://labrosa.ee.columbia.edu/millionsong/

res contemporâneas [BMEWL11]. Esta coleção multimédia representa uma avanço significativo na área de recuperação de informação musical, com o objetivo de encorajar a investigação de algoritmos de pesquisa em larga-escala, e tornar-se uma coleção de referência para avaliação de resultados na recomendação e descoberta de conteúdos musicais. Apesar de não incluir os ficheiros áudio originais, por questões de licenciamento, fornece as características áudio e metadados extraídos e analisados pelo software *The Echo Nest*[4].

6.2.1 Metadados adicionais e características áudio

Esta coleção é acompanhada de dois conjuntos de dados adicionais, os quais incluem dados textuais relativos às músicas.

O conjunto de dados *Last.fm*[5] representa a informação oficial de *tags* e similaridade entre as músicas existentes na coleção MSD. Este conjunto abrange 94% das músicas, sendo que mais de metade possuem pelo menos uma tag associada. No trabalho desenvolvido, a informação de similaridade foi considerada para efeitos de avaliação, por se basear no *feedback* dos utilizadores, e a informação das *tags* foi utilizada como característica textual para indexar as músicas.

O conjunto de dados *musiXmatch*[6] inclui as letras das canções existentes na MSD. Estas letras foram armazenadas em forma de "saco de palavras", devido a restrições de licenciamento. Dada a existência de géneros musicais com pouca ou nenhuma informação relativa a letras, este conjunto abrange na realidade apenas 24% da coleção MSD, o que representa ainda assim um conjunto substancial de informação relevante para pesquisa dos dados.

Relativamente à informação de conteúdo áudio, foram escolhidas características específicas do conjunto disponibilizado na coleção MSD, nomeadamente *time signature*, *tempo*, *mode*, *loudness*, *key*, *duration*, *pitch coefficient average* $\times$ 12, e *timbre coefficient average* $\times$ 12.

A Tabela 6.1 mostra algumas propriedades dos conjuntos de dados utilizados. Não foi possível gerar os descritores áudio de todas as músicas existentes devido

[4] `http://the.echonest.com/`
[5] `http://labrosa.ee.columbia.edu/millionsong/lastfm`
[6] `http://labrosa.ee.columbia.edu/millionsong/musixmatch`

à ausência de características. As canções afetadas foram excluídas do conjunto de dados indexado, que manteve no entanto a sua propriedade de larga-escala.

Tabela 6.1: A coleção *Million Song Dataset*.

Músicas	**Tamanho**	**Artistas**	***Tags***	**Termos das letras**
961.493	262 GB	44.263	214.809	4.920

6.3 Aplicações

Existe já um número elevado de trabalhos na área de avaliação de *playlists* [SW06, BOL09, FRd10], sendo que os estudos envolvendo utilizadores são normalmente considerados em fases posteriores, para uma validação mais próxima da realidade.

Dada a natureza subjetiva das tarefas de geração de *playlists* e descoberta musical, foram obtidos resultados para interrogações predefinidas de forma a observar o comportamento da abordagem baseada na análise do conteúdo áudio das canções.

6.3.1 Geração de *playlists*

Após a inserção de texto por parte do utilizador, como o nome de uma música ou artista, um excerto da letra ou palavras-chave referentes a estados de espírito como "felicidade" e "traição", o sistema efetua uma pesquisa textual sobre os campos indexados. O resultado traduz-se numa *playlist* inicial limitada a um máximo de 20 itens

Uma possível lista de resultados para a interrogação "coldplay live" é exibido na Tabela 6.2. "Pontuação" refere-se ao valor obtido pelo resultado na pesquisa efetuada com recurso à biblioteca Lucene. Neste exemplo, os utilizadores pretendem músicas de concertos ao vivo da banda *Coldplay*. O protótipo permite a pesquisa por géneros musicais específicos, caracterizados pelas palavras-chave e *tags* associadas às canções respetivas.

Após a obtenção da *playlist* inicial, é aplicada uma reordenação por coerência áudio, baseada na similaridade entre músicas. Através da distância euclideana, usando

Tabela 6.2: Playlist inicial

Interrogação: "coldplay live"	**Pontuação**
See You Soon (Live In Sydney) - Coldplay	5.1
Shiver (Live In Sydney) - Coldplay	5.1
One I Love (Live In Sydney) - Coldplay	5.1
Amsterdam (Live In Sydney) - Coldplay	4.2
You Only Live Twice (Live Norway) - Coldplay	4.0
Daylight - Coldplay Tribute	3.9
Moses (Live In Sydney) - Coldplay	3.7
Yellow (Live In Sydney) - Coldplay	3.4
Speed Of Sound (Live) - Coldplay	3.2
Fix You (Live) - Coldplay	3.2
Sleeping Sun - Coldplay	3.2
The World Turned Upside Down - Coldplay	3.2
One I Love	3.2
Pour Me (Live At The Hollywood Bowl)	3.2
High Speed - Coldplay	3.2
Low	3.1
Clocks - Coldplay Tribute	3.1
Warning Sign	3.1
Fix You	3.1
We Never Change	3.1

os vetores de características, as distâncias entre canções são adicionadas para determinar um total. Músicas com um valor inferior, isto é, com uma similaridade maior a todas as outras, tornam-se assim "centrais" na *playlist*. As canções mais similares entre si são então agrupadas no início da *playlist*, enquanto que as mais distintas são deslocadas para o final. O resultado da aplicação deste método pode ser observado na Tabela 6.3.

Outra opção disponível está relacionada com a possibilidade de reordenar a *playlist* com base na "música mais próxima", conforme exemplificado na Tabela 6.4. Após o utilizador escolher uma música específica na *playlist*, o sistema procura a música mais similar nessa lista, e vai sucessivamente procurando a música mais semelhante à anterior. Desta forma, pretende-se que o utilizador escolha uma música preferida e consiga ouvir a *playlist* com o mínimo de perturbação, atenuando as diferenças nas passagens entre canções.

Tabela 6.3: Playlist reordenada

Após reordenação áudio	**Pontuação**
Shiver (Live In Sydney) - Coldplay	5.1
Moses (Live In Sydney) - Coldplay	3.7
One I Love (Live In Sydney) - Coldplay	5.1
One I Love	3.2
Pour Me (Live At The Hollywood Bowl)	3.2
See You Soon (Live In Sydney) - Coldplay	5.1
Warning Sign	3.1
Fix You	3.1
Daylight - Coldplay Tribute	3.9
The World Turned Upside Down - Coldplay	3.2
Amsterdam (Live In Sydney) - Coldplay	4.2
You Only Live Twice (Live Norway) - Coldplay	4.0
We Never Change	3.1
Fix You (Live) - Coldplay	3.2
Clocks - Coldplay Tribute	3.1
Low	3.1
Yellow (Live In Sydney) - Coldplay	3.4
High Speed - Coldplay	3.2
Sleeping Sun - Coldplay	3.2
Speed Of Sound (Live) - Coldplay	3.2

6.3.2 Descoberta de músicas

A maior vantagem da abordagem baseada no conteúdo é a sua independência relativamente à popularidade das músicas, permitindo a descoberta de músicas que possam ser do agrado dos utilizadores. Conforme pode ser observado na Tabela 6.5, ao efetuar uma pesquisa com base na atuação ao vivo da música "One I Love", foi possível obter a versão original usando as características áudio, o que indica que o algoritmo de pesquisa aproximada, o esquema de binarização de vetores e as características áudio escolhidas capturam informação suficiente para permitir a descoberta de músicas semelhantes em coleções de grande escala.

A Tabela 6.6 contém os resultados da mesma interrogação, mas utilizando a informação contida nas letras e *tags* das músicas para efetuar uma pesquisa por contexto. A *playlist* resultante é assim composta por canções de géneros musicais semelhantes e com um conteúdo lírico aproximado.

As tarefas de geração de *playlists* e descoberta de música são efetuadas em segundos usando a biblioteca *Lucene*. O índice final, incluindo metadados e características áudio, ocupa cerca de 12 GB em disco, aproximadamente 4% do tamanho total da

Tabela 6.4: Playlist começada por uma música específica

Música escolhida: "Sleeping Sun"	**Pontuação**
Sleeping Sun - Coldplay	3.2
You Only Live Twice (Live Norway) - Coldplay	4.0
Shiver (Live In Sydney) - Coldplay	5.1
Moses (Live In Sydney) - Coldplay	3.7
One I Love (Live In Sydney) One I Love	3.2
Pour Me (Live At The Hollywood Bowl)	3.2
The World Turned Upside Down - Coldplay	3.2
Low	3.1
Speed Of Sound (Live) - Coldplay	3.2
Fix You (Live) - Coldplay	3.2
Fix You	3.1
Daylight - Coldplay Tribute	3.9
See You Soon (Live In Sydney) - Coldplay	5.1
Warning Sign	3.1
Amsterdam (Live In Sydney) - Coldplay	4.2
Clocks - Coldplay Tribute	3.1
High Speed - Coldplay	3.2
We Never Change	3.1
Yellow (Live In Sydney) - Coldplay	3.4

coleção, e apenas pode ser carregado para memória em servidores com essa capacidade. No entanto, a biblioteca *Lucene* dispõe de mecanismos que permitem manter em memória as estruturas e dados relevantes para as pesquisas mais recentes.

Para coleções de ainda maior escala, é possível recorrer às facilidades de distribuição do índice e conteúdos por várias máquinas, e à possibilidade de separação dos tipos de dados por plataformas diferentes, adequadas às suas características. Por exemplo, armazenar as músicas, fotos e respetivas descrições em servidores de conteúdo estático, guardar os metadados em bases de dados, e manter no índice apenas a informação identificativa dos documentos, de forma a ser possível reagrupar toda a informação durante a fase de interação com os utilizadores.

6.4 Avaliação

A avaliação das *playlists* tirou partido da coleção de dados *The Echo Nest Taste Profile Subset*[7] para determinar a qualidade dos resultados produzidos para algu-

[7] `http://labrosa.ee.columbia.edu/millionsong/tasteprofile`

Tabela 6.5: Músicas semelhantes – características áudio

Interrogação: "One I Love (Live In Sydney)"	**Pontuação**
One I Love (Live In Sydney) - Coldplay	17.0
Banda De Rock & Roll - Ratones Paranoicos	17.0
Silver Strand (Album Version) - The Corrs	17.0
Time (24-Bit Digitally Remastered 05) - Blind ...	14.8
Amplified Ohm - Melting Euphoria	14.7
Tomorrow Is Coming - Ocha la Rocha	12.5
Jalousi (Igen!) - Peter Sommer	12.2
Are You Anywhere? (edit) - Padded Cell	11.5
One I Love - Coldplay	11.5
Let The Sky Fall - Ten Years After	11.5
Jane Says (Live)(LP Version) - Jane's ...	11.2
Black - Pearl Jam	11.0
To Bække Små - Peter Sommer	10.9
Our Addictions - Art In Manila	10.7
Genius - The Lovetones	10.6
Whip a Rose - Thomas Jefferson Slave ...	10.5
Il Compositore Di Nuvole - Le Vibrazioni	10.5
A Decade Without a Death - Ghost of the ...	10.5
Meet Us Here - The Glorious Unseen	10.3
All For You (Ruff & Jam Midnight Mix) - Kate ...	10.2

mas interrogações predefinidas. A estratégia adotada partiu da hipótese de que uma *playlist* automaticamente gerada terá maior qualidade quanto maior o número de músicas que co-ocorram em *playlists* geradas por utilizadores. Desta forma, a pontuação de uma *playlist* resultará da soma do número de pares de músicas que tenham sido ouvidas por pelo menos um utilizador. Dada uma *playlist* de n canções, a pontuação máxima que esta poderá obter será o número total de pares na lista, ou seja, $n(n-1)/2$.

Foram geradas 20 *playlists* com $n = 20$ músicas cada, baseadas nas interrogações exibidas na Tabela 6.7. Os resultados obtidos demonstraram que as interrogações respondidas com base nos metadados textuais continham mais pares de músicas ouvidas em conjunto pelos utilizadores. As pesquisas com base nas letras das músicas e *tags* obtiveram resultados intermédios, sendo que os resultados das pesquisas baseadas no conteúdo áudio atingiram os valores mais baixos, mas ainda assim positivos.

Este esquema de avaliação permitiu demonstrar que a informação de contexto continua a ser a melhor solução para obter uma *playlist* inicial aceitável, mas as

Tabela 6.6: Músicas semelhantes – letras e *tags*

Interrogação: "One I Love (Live In Sydney)"	**Pontuação**
One I Love - Coldplay	4.7
One I Love (Live In Sydney) - Coldplay	3.9
You - Mr. Sancho	3.1
Desperado - Journey South	3.1
You Are The One Lalala - Morten Abel	2.8
You Are Everything - Dru Hill	2.8
Blame It On Me - Aaron Watson	2.7
Just The Way (Explicit) - Alfonzo Hunter	2.7
Sprung - B2K	2.6
If Work Permits - The Format	2.6
Home - Edward Sharpe & The Magnetic ...	2.6
Wish List - Jets To Brazil	2.6
Willie - Cat Power	2.6
Southern State - Bright Eyes	2.6
You Are The Light - Jens Lekman	2.6
Christmas TV - Slow Club	2.5
Stand Up - Forty Deuce	2.5
Come On Home - Everything But The Girl	2.5
My Way Home Is Through You [B-Side] - ...	2.5
You Came Back - Pete Townshend	2.5

características áudio conseguem encontrar músicas similares. Quanto maior for a pontuação obtida, maior será a capacidade de o sistema gerar *playlists* com músicas co-ocorrentes em *playlists* criadas por utilizadores, validando os métodos de similaridade por conteúdo.

Esta informação pode ser usada para produzir recomendações equilibradas, conjugando a "segurança" da pesquisa baseada no contexto com a potencial "serendipidade" da pesquisa baseada no conteúdo, isto é, a capacidade de descoberta de músicas que não seriam encontradas usando apenas métodos baseados na pesquisa dos metadados. Dado que o protótipo apresentado contempla a informação baseada no conteúdo, não corre o risco de ficar enviesado às preferências dos utilizadores, introduzindo implicitamente uma potencial diversificação dos resultados.

Tabela 6.7: Avaliação das *playlists*

Playlist	**Interrogação**	**Pont.**$_{texto}$	**Pont.**$_{audio}$	**Pont.**$_{tags}$
1	coldplay live	0.7316	0.0000	0.1368
2	metallica slayer heavy metal	0.3526	0.0737	0.4105
3	nirvana days of the new grunge alice in chains	0.3474	0.0053	0.2842
4	jason mraz i'm yours	0.7263	0.0263	0.4684
5	happy good vibe	0.0105	0.0368	0.0842
6	sad depressing doom dark	0.0368	0.0421	0.3947
7	britney spears rihanna madonna	0.4105	0.0579	0.1158
8	norah jones diana krall jamie cullum	0.7632	0.0053	0.3158
9	miles davis john coltrane classic jazz	0.0316	0.0263	0.0526
10	frank sinatra new york	0.0368	0.0000	0.0947
11	bob marley reggae summer happy positive	0.2421	0.0789	0.2263
12	pop rock avril lavigne	0.4000	0.0474	0.0263
13	indiana jones soundtrack	0.0842	0.0105	0.3211
14	led zeppelin the who classic rock	0.1000	0.0053	0.2053
15	rockabilly 50s elvis presley	0.0000	0.0158	0.0421
16	country bluegrass bill monroe banjo	0.0158	0.0105	0.0368
17	dubstep skrillex new beat	0.1737	0.0263	0.0947
18	electronic aphex twin creative	0.0632	0.0000	0.0947
19	house techno trance bestof	0.0474	0.0105	0.1421
20	blues muddy waters robert johnon jimi hendrix	0.0684	0.0263	0.0316
Média		0.2 ± 0.3	0.03 ± 0.02	0.2 ± 0.1

Capítulo 7

Conclusões

> "The pessimist complains about the wind; the optimist expects it to change; the realist adjusts the sails."
>
> *William Arthur Ward*

Neste capítulo são apresentadas as principais conclusões obtidas ao longo deste trabalho, validando a tese enunciada e investigação paralela que permitiu contemplar ramificações futuras dos resultados disponibilizados. Ao longo desta investigação, foram desenvolvidos trabalhos de aplicação das metodologias analisadas em áreas distintas como a análise de conteúdo visual e a recomendação musical.

No sentido de explorar novas possibilidades de enriquecimento da informação existente nas coleções multimédia, foram consideradas metodologias na área das redes sociais e análise de gráfos para auxiliar a deteção de entidades consideradas relevantes. Foram também contempladas áreas como o pré-processamento das fotos recorrendo a filtros de abstração visual no reconhecimento facial, para identificação de personalidades.

7.1 Ilustração suportada por recuperação de informação multimédia em larga-escala

Foi apresentada uma arquitetura de sistema para responder à tarefa de ilustração automática de conteúdos textuais, a qual recorre a algoritmos de recuperação de informação para extração e descrição de conteúdos multimédia, bem como pesquisa aproximada e reordenação por coerência.

A indexação é efetuada sobre os metadados, legendas e *tags* associadas, e sobre as descrições visuais das fotos existentes nas coleções de grande escala adotadas. O objetivo principal, de recomendação de imagens para ilustração de conteúdos textuais, é atingido com recurso aos mecanismos de determinação de similaridade textual no contexto, e similaridade visual no conteúdo das fotos. A capacidade de resposta em tempo útil, face à dimensão dos dados e capacidades de processamento existentes, permite ainda a exploração das coleções com base num exemplo fornecido.

A utilização de componentes *opensource*, bem como o mapeamento de características visuais para conteúdo textual, indexável por plataformas de recuperação de informação largamente disseminadas e estabelecidas, tornam a estratégia proposta facilmente integrável em sistemas já existentes.

A utilização do conceito de coerência entre documentos, e a aplicação das metodologias a áreas distintas como a ilustração de texto, geração de playlists musicais, exploração de grandes coleções de fotos e descoberta de novas músicas, foram extensivamente avaliados comprovando-se a sua eficácia no compromisso entre precisão e diversidade dos resultados obtidos, face a tarefas tão subjetivas e dependentes da interação com o utilizador e respetiva satisfação das necessidades de informação.

7.2 Ramificações

As secções seguintes descrevem sucintamente aspetos analisados em paralelo, os quais têm o potencial de enriquecer ainda mais os resultados produzidos e contribuir para a investigação nesta área.

7.2.1 Deteção de comunidades

As comunidades constituem grupos de pessoas que partilham um contexto específico, como interesses comuns, padrões de comportamento ou simplesmente o idioma ou etnicidade. Através de técnicas de reconhecimento de entidades mencionadas, foi possível identificar e extrair personalidades presentes nas legendas das fotos jornalísticas que constituem a coleção SAPO-Lusa.

Nesse sentido, assume-se que quando duas personalidades são co-referidas na descrição de uma foto, contribuem para o mesmo contexto. Uma representação baseada em grafos, em que as personalidades são representadas por nós e a sua presença simultânea nas legendas das fotos representadas por ligações entre esses nós, permite obter uma rede de co-referência de personalidades da coleção SAPO-Lusa.

A rede gerada foi analisada de forma a verificar a sua validade como estrutura de comunidade, fator importante para a sua utilização no melhoramento das capacidades de pesquisa e ilustração de texto. A estrutura de comunidade contribui para evidenciar nos resultados a presença de personalidades que partilhem o mesmo contexto, permitindo a desambiguação de entidades e associação de fotos a eventos relacionados.

7.2.1.1 Criação da rede de entidades

A criação da rede de personalidades sobre a coleção SAPO-Lusa, com 1.5 milhões de fotos e respetivas legendas, envolveu o auxílio de ferramentas específicas. As descrições das fotos foram analisadas por um serviço de recolha de entidades em notícias portuguesas, disponibilizado pelo SAPO como parte do projeto SAPO Verbetes. Desta forma, foi possível comparar os dados da coleção com uma lista de personalidades regularmente atualizada.

Para cada entidade encontrada, através do "match" do nome nos conteúdos textuais de uma foto, esta informação é armazenada numa base de dados através dos identificadores de cada uma. Posteriormente, com o auxílio do software R Project[1], é gerada uma matriz de adjacências em que as ligações entre personalidades assumem

[1] `www.r-project.org`

um peso maior consoante o número de fotos em que sejam mencionadas simultaneamente. Esta matriz de adjacência é finalmente convertida para um ficheiro GraphML, permitindo a sua importação na ferramenta de visualização de grafos Gephi[2].

7.2.1.2 Análise das ligações entre personalidades

A rede de personalidades pode ser observada na Figura 7.1, onde se verifica a existência de 7 comunidades bem delimitadas, com ênfase nas áreas de Política e Desporto. Esta observação permite concluir que as fotos contidas na coleção SAPO-Lusa incidem maioritariamente sobre estas duas vertentes das notícias nacionais e internacionais.

A Tabela 7.1 exibe algumas propriedades das comunidades indentificadas no grafo gerado, nomeadamente a sua densidade, o grau de conetividade, e o valor de PageRank [BP98, For10].

Tag da Comunidade	**Densidade**	**Grau**		**PageRank**	
		Média	Mediana	Média	Mediana
Políticos port.	0.60%	7.025	3.00	0.0009	0.0004
Políticos intern.	1.22%	10.160	3.00	0.0012	0.0006
Finanças	7.37%	4.866	3.00	0.0145	0.0113
Futebol	1.70%	**11.880**	**6.00**	0.0014	0.0009
Ténis / Fórmula 1	1.84%	6.195	3.00	0.0030	0.0020
Ciclismo	6.45%	3.614	2.00	0.0175	0.0129
Basquetebol	**12.12%**	4.000	3.00	**0.0294**	**0.0214**

Tabela 7.1: Análise de Comunidades (valores máximos para cada coluna destacados a negrito

A estrutura de comunidade desta rede foi identificada recorrendo ao algoritmo de otimização de modularidade [BGLL08] presente no software *Gephi*. Cores diferentes evidenciam comunidades diferentes, com o tamanho do nó a representar o seu valor de *PageRank*.

Após efetuar um "sampling" de algumas personalidades de cada comunidade, foi efetuada uma pesquisa no *Google* e *Wikipedia* para determinar a sua profissão, o que permitiu chegar rapidamente à melhor definição para as comunidades detetadas.

[2] `http://gephi.org/`

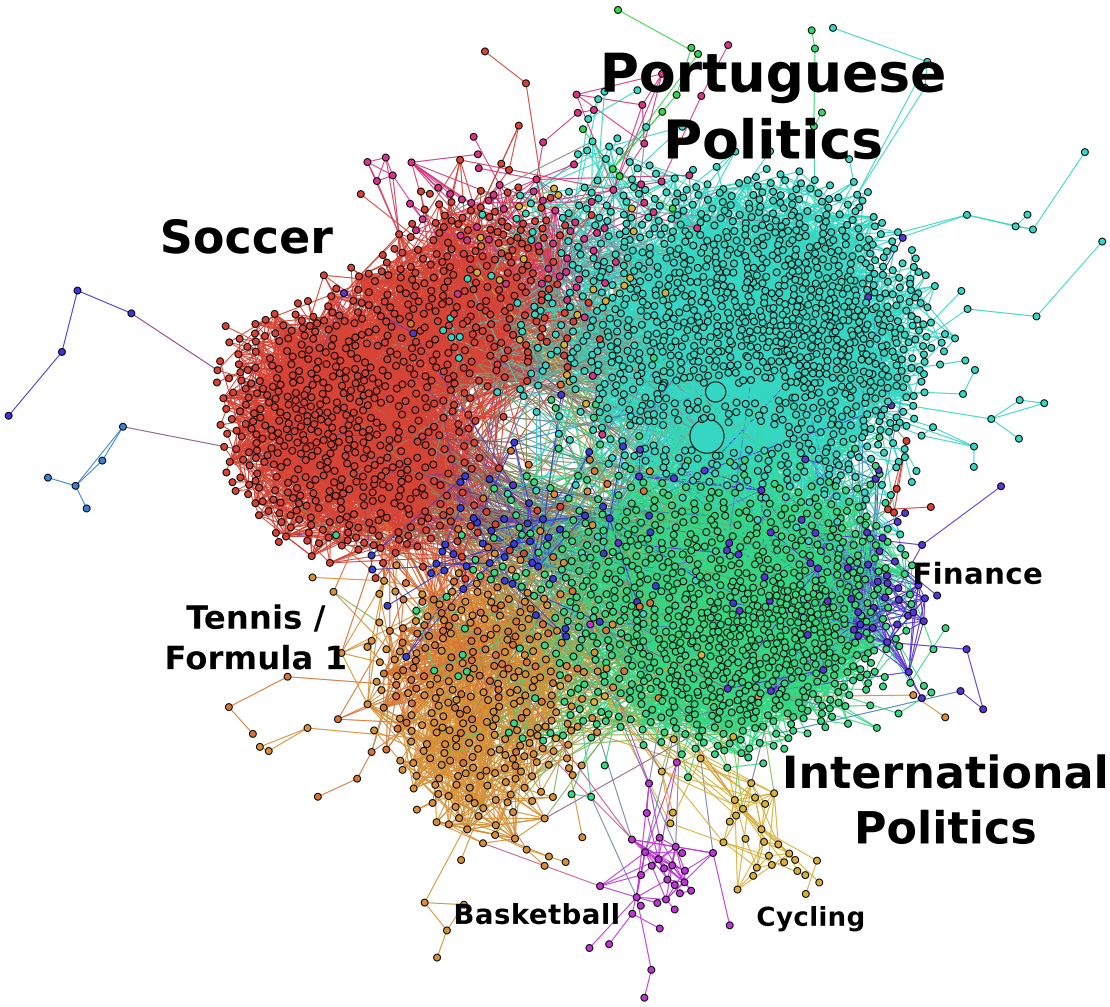


Figura 7.1: Rede de personalidades

As *tags* foram validadas através de uma agregação das notícias de cada comunidade, após remoção das palavras portuguesas e inglesas mais frequentes, e identificação das frequências dos termos.

A Tabela 7.2 contém as personalidades mais representativas de cada comunidade, bem como o seu grau e valor de *PageRank*. É possível concluir que, por exemplo, as personalidades relacionadas com Futebol são normalmente mencionadas com várias personalidades dessa área (grau de conetividade máximo), enquanto que as personalidades ligadas ao Basquetebol aparecem normalmente associadas a entidades muito importantes nesse contexto (valor máximo de *PageRank*).

Rank	Personalidade	Grau	Personalidade	PageRank
Políticos portugueses				
1	Cavaco Silva	268	Jorge Sampaio	0.029835838
2	Jorge Sampaio	260	Cavaco Silva	0.029552041
3	José Sócrates	173	José Sócrates	0.017865690
4	Jaime Gama	87	Manuel Pinho	0.008565179
Políticos internacionais				
1	George W. Bush	186	George W. Bush	0.02080142
2	Angela Merkel	135	Nicolas Sarkozy	0.01398886
3	Nicolas Sarkozy	130	Angela Merkel	0.01352073
4	Javier Solana	129	Vladimir Putin	0.01244500
Finanças				
1	Christine Lagarde	20	Christine Lagarde	0.05494546
2	Christian Noyer	17	Didier Reynders	0.04656112
3	Jean Claude Trichet	17	Peer Steinbrueck	0.04198738
4	Axel Weber	15	Pedro Solbes	0.04117235
Futebol				
1	Stamford Bridge	111	Stamford Bridge	0.010989572
2	Thierry Henry	93	Thierry Henry	0.009253354
3	Cristiano Ronaldo	91	Cristiano Ronaldo	0.009160757
4	Michael Ballack	82	Borussia Dortmund	0.008084869
Ténis/ Fórmula 1				
1	Roger Federer	44	Roger Federer	0.014216810
2	Michael Schumacher	35	Michael Schumacher	0.012195036
3	Rafael Nadal	31	**Martin Scorsese**	0.010473509
4	Fernando Alonso	30	**Thomas Gottschalk**	0.010239316
Ciclismo				
1	Tom Boonen	14	Alberto Contador	0.05277021
2	Alberto Contador	13	Tom Boonen	0.05171201
3	Alejandro Valverde	12	Alejandro Valverde	0.04549064
4	Ivan Basso	9	**Carlos Pereira**	0.03587219
Basquetebol				
1	Kobe Bryant	13	Tony Parker	0.08251185
2	Tony Parker	12	Kobe Bryant	0.08140955
3	Carmelo Anthony	10	David Stern	0.07534336
4	Paul Pierce	10	Carmelo Anthony	0.05919088

Tabela 7.2: As 4 entidades mais representativas de cada comunidade

7.2.2 Deteção e reconhecimento facial

O trabalho desenvolvido no âmbito do estudo do impacto dos filtros de abstração no processo de reconhecimento facial foi executado pelo finalista e investigador Pedro

Tiago Pontes[3] no projeto *Visage*[4]. Este projeto teve por base os resultados obtidos na aplicação destes filtros na recuperação de informação visual, discutidos previamente.

O reconhecimento facial representa uma tarefa efetuada diariamente de forma transparente por seres humanos. O reconhecimento facial em imagens é um tema atual e onde se tem verificado um interesse crescente devido às suas múltiplas áreas de aplicação, assim como ao elevado valor comercial tradicionalmente associado a este tipo de soluções. No entanto, a construção de sistemas automáticos de reconhecimento facial engloba um conjunto de sub-problemas característicos, como a deteção e segmentação das faces presentes nas imagens ou vídeos, a sua normalização e a extração de características distintivas das faces, com o objetivo de efetuar o reconhecimento das pessoas representadas. A execução bem sucedida destas etapas requer previamente a resolução de um conjunto de desafios a nível da variação das poses, iluminação e expressão das identidades contidas nas imagens.

Estes sub-problemas, aliados as múltiplas áreas de aplicação do reconhecimento facial, fazem com que exista uma a grande variação do desempenho dos diversos sistemas existentes, a qual se encontra diretamente relacionada com as condições de utilização dos mesmos, nomeadamente ao no que diz respeito às galerias de imagens utilizadas. A este nível, em situações onde as condições de captura das imagens são controladas e existe uma cooperação ativa por parte dos utilizadores os resultados obtidos são muito satisfatórios.

Os desafios colocados pelo processo de reconhecimento, e consequentemente a vasta gama de aplicações onde a identificação de indivíduos é necessária, como o controlo de acesso a informação, segurança, entretenimento e a gestão de conteúdos multimédia, entre outros, despoletou a atenção da comunidade científica ao longo das últimas décadas. Verificou-se assim uma evolução notável ao nível da eficácia dos sistemas desenvolvidos, considerando-se mesmo que o problema de reconhecimento facial em cenários cooperativos e com condições de captura de imagens controladas se encontra praticamente resolvido (Figura 7.2).

Por outro lado, em cenários não cooperativos e onde se regista uma variação não controlada da captura das imagens, esta é ainda uma área de investigação em aberto.

[3] http://www.pedrotiagopontes.com
[4] http://labs.sapo.pt/2012/12/visage-feup/

(a) Apenas uma face

(b) Múltiplas faces

Figura 7.2: Deteção de faces

Os filtros de abstração, descritos na secção anterior, constituem uma forma moderna de simplificação do conteúdo visual, permitindo remover informação redundante e

dar destaque à mensagem visual a transmitir. Um exemplo pode ser observado na Figura 7.3.

Pretendeu-se estudar o impacto do uso de filtros de abstração no processo de reconhecimento facial automático através do desenvolvimento de um protótipo de reconhecimento facial de personalidades, baseado em código aberto, envolvendo a utilização da abstração de imagens juntamente com um conjunto de outras tarefas de pré-processamento sobre as imagens a reconhecer. Para uma dada imagem fornecida ao protótipo implementado, é aplicada sobre ela uma cadeia de pré-processamento, na qual a abstração de imagens se encontra incluída, e é efetuado posteriormente o seu reconhecimento, sendo devolvida uma lista ordenada de possíveis entidades contidas na imagem original. Foi utilizada a coleção de imagens *Labelled Faces in the Wild*[5], para análise do comportamento dos algoritmos de reconhecimento disponíveis na plataforma *OpenCV*[6], nomeadamente *Eigenfaces* [TP91], *Fisherfaces* [BHK97] e *Local Binary Patterns Histograms* [AHP04].

Os resultados demonstram que a aplicação de filtros de abstração no processo de reconhecimento resulta num compromisso entre a diminuição dos requisitos de armazenamento das imagens e uma ligeira diminuição da eficácia da identificação. Estes filtros constituem uma forma moderna e computacionalmente eficaz de abstração de informação, sendo tradicionalmente utilizados para comunicar mais eficazmente uma mensagem visual. Para além disso, o uso destes filtros para a pesquisa baseada em conteúdos com vista a ilustração automática de texto demonstrou resultados positivos ao nível da informação retornada. A este nível destaca-se o filtro *Kuwahara* anisotrópico, o qual representa o maior grau de abstração dos filtros utilizados, permitindo uma diminuição considerável do tamanho da galeria processada, mas que regista também um maior impacto no desempenho do reconhecimento.

Ao nível das avaliações efetuadas, é possível concluir que a deteção e segmentação correta das faces constituem as etapas de pré-processamento mais relevantes para a obtenção de resultados positivos no reconhecimento dos indivíduos, independentemente do algoritmo de reconhecimento utilizado. Por outro lado, a normalização do contraste das imagens através da equalização do seu histograma, revela uma melhoria significativa nos resultados obtidos com particular ênfase no algoritmo *Eigenfaces*.

[5] `http://vis-www.cs.umass.edu/lfw/`
[6] `http://opencv.org/`

(a) Original

(b) Após filtro AKF

Figura 7.3: Aplicação do filtro anisotrópico de Kuwahara

Por último, dos três algoritmos analisados, o algoritmo *Local Binary Patterns Histograms* revelou ter o melhor desempenho na maioria dos conjuntos analisados.

Referências

[ABF+11] Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro e Fausto Rabitti. Combining local and global visual feature similarity using a text search engine. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 49–54. IEEE, 2011.

[AHP04] Timo Ahonen, Abdenour Hadid e Matti Pietikäinen. Face recognition with local binary patterns. In *Computer Vision-ECCV 2004*, pages 469–481. Springer, 2004.

[APBC+09] Julien Ah-Pine, Marco Bressan, Stephane Clinchant, Gabriela Csurka, Yves Hoppenot e Jean-Michel Renders. Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications*, 42(1):31–56, 2009.

[AS08] Giuseppe Amato e Pasquale Savino. Approximate similarity search in metric spaces using inverted files. In *Proceedings of the 3rd international conference on Scalable information systems*, page 28. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte e Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[BHK97] Peter N. Belhumeur, João P Hespanha e David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[BMEWL11] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman e Paul Lamere. The million song dataset. In *ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida*, pages 591–596. University of Miami, 2011.

[BOL09] Luke Barrington, Reid Oda e Gert RG Lanckriet. Smarter than genius? human evaluation of music recommender systems. In *ISMIR*, volume 9, pages 357–362, 2009.

[BP98] Sergey Brin e Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[CAB10] Savvas A Chatzichristofis, Avi Arampatzis e Yiannis S Boutalis. Investigating the behavior of compact composite descriptors in early fusion, late fusion, and distributed image retrieval. *Radioengineering*, 19(4):725–733, 2010.

[CB08a] Savvas A Chatzichristofis e Yiannis S Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Computer Vision Systems*, pages 312–322. Springer, 2008.

[CB08b] Savvas A Chatzichristofis e Yiannis S Boutalis. Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 191–196. IEEE, 2008.

[CB10] Savvas A Chatzichristofis e Yiannis S Boutalis. Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. *Multimedia Tools and Applications*, 46(2-3):493–519, 2010.

[CBL09] Savvas A Chatzichristofis, Yiannis S Boutalis e Mathias Lux. Img (rummager): An interactive content based image retrieval system. In *Similarity Search and Applications, 2009. SISAP'09. Second International Workshop on*, pages 151–153. IEEE, 2009.

[CCMV07] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno e Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):394–410, 2007.

[CDR13] Filipe Coelho, José Devezas e Cristina Ribeiro. Juggle: large-scale discovery in music recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 219–220, 2013.

[CLF+09] Michael Calonder, Vincent Lepetit, Pascal Fua, Kurt Konolige, James Bowman e Patrick Mihelich. Compact signatures for high-speed interest point description and matching. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 357–364. IEEE, 2009.

[CR10] Filipe Coelho e Cristina Ribeiro. Evaluation of global descriptors for multimedia retrieval in medical applications. In *Database and Expert Systems Applications (DEXA), 2010 Workshop on*, pages 127–131. IEEE, 2010.

[CR11a] Filipe Coelho e Cristina Ribeiro. Automatic illustration with cross-media retrieval in large-scale collections. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 25–30. IEEE, 2011.

[CR11b] Filipe Coelho e Cristina Ribeiro. Characterization of the SAPO-Lusa news stories photo collection. Technical report, Faculdade de Engenharia da Universidade do Porto, 2011. `http://www.inescporto.pt/~fcoelho/web/_media/files/2011sapolabs.pdf`.

[CR11c] Filipe Coelho e Cristina Ribeiro. dpikt: Automatic illustration system for media content. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 241–244. IEEE, 2011.

[DCNR12] José Devezas, Filipe Coelho, Sérgio Nunes e Cristina Ribeiro. Studying a personality coreference network in a news stories photo collection. In *Advances in Information Retrieval*, pages 485–488. Springer, 2012.

[DGLW07] Ritendra Datta, Weina Ge, Jia Li e James Ze Wang. Toward bridging the annotation-retrieval gap in image search. *IEEE MultiMedia*, 14(3):24–35, 2007.

[DIIM04] Mayur Datar, Nicole Immorlica, Piotr Indyk e Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM, 2004.

[DJLW06] Ritendra Datta, Dhiraj Joshi, Jia Li e James Z Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.

[DJLW08] Ritendra Datta, Dhiraj Joshi, Jia Li e James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[DKN08] Thomas Deselaers, Daniel Keysers e Hermann Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, 2008.

[DLW08] Ritendra Datta, Jia Li e James Ze Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 105–108. IEEE, 2008.

[DMC10] Diogo Delgado, Joao Magalhaes e Nuno Correia. Assisted news reading with automated illustration. In *Proceedings of the international conference on Multimedia*, pages 1647–1650. ACM, 2010.

[DMWG09] Marie Dumont, Raphaël Marée, Louis Wehenkel e Pierre Geurts. Fast multi-class image annotation with random windows and multiple output randomized trees. In *Proc. International Conference on Computer Vision Theory and Applications (VISAPP) Volume*, volume 2, pages 196–203, 2009.

[Eid03] Horst Eidenberger. Distance measures for mpeg-7-based retrieval. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 130–137. ACM, 2003.

[Esu09] Andrea Esuli. Pp-index: Using permutation prefixes for efficient and scalable approximate similarity search. *Proceedings of LSDS-IR*, 2009, 2009.

[Fie11] Benjamin Fields. *Contextualize your listening: the playlist as recommendation engine*. PhD thesis, Department of Computing Goldsmiths, University of London, 2011.

[For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[FRd10] Ben Fields, Christophe Rhodes e Mark d'Inverno. Using song social tags and topic models to describe and compare playlists. In *1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain*, 2010.

[GABS10] Claudio Gennaro, Giuseppe Amato, Paolo Bolettieri e Pasquale Savino. An approach to content-based image retrieval based on the lucene search engine library. In *Research and Advanced Technology for Digital Libraries*, pages 55–66. Springer, 2010.

[HB09] Nicolas Hervé e Nozha Boujemaa. Visual word pairs for automatic image annotation. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 430–433. IEEE, 2009.

[HE07] James Hays e Alexei A Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007.

[Hee08] Daniel Heesch. A survey of browsing models for content based image retrieval. *Multimedia Tools and Applications*, 40(2):261–284, 2008.

[HKM+97] Jing Huang, S Ravi Kumar, Mandar Mitra, Wei-Jing Zhu e Ramin Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997.

[HL08] Mark J Huiskes e Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.

[HLES06] Jonathon S Hare, Paul H Lewis, Peter GB Enser e Christine J Sandom. Mind the gap: another look at the problem of the semantic gap in image retrieval. In *Electronic Imaging 2006*, pages 607309–607309. International Society for Optics and Photonics, 2006.

[HLMS08] Alan Hanjalic, Rainer Lienhart, W-Y Ma e JOHN R Smith. The holy grail of multimedia information retrieval: So close or yet so far away? *Proceedings of the IEEE*, 96(4):541–547, 2008.

[HR07] Daniel Heesch e Stefan Rüger. Interaction models and relevance feedback in image retrieval. *Semantic-Based Visual Information Retrieval*, pages 160–186, 2007.

[HSD11] Jonathon S Hare, Sina Samangooei e David P Dupplaw. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 691–694. ACM, 2011.

[JCG+05] Alejandro Jaimes, Mike Christel, Sébastien Gilles, Ramesh Sarukkai e Wei-Ying Ma. Multimedia information retrieval: what is it, and why isn't anyone using it? In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 3–8. ACM, 2005.

[JDSP10] Hervé Jégou, Matthijs Douze, Cordelia Schmid e Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[JWL04] Dhiraj Joshi, James Z Wang e Jia Li. The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 119–126. ACM, 2004.

[JWL06] Dhiraj Joshi, James Z Wang e Jia Li. The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):68–89, 2006.

[KCN08] Lyndon Kennedy, Shih-Fu Chang e Apostol Natsev. Query-adaptive fusion for multimodal search. *Proceedings of the IEEE*, 96(4):567–588, 2008.

[KKD09] J.E. Kyprianidis, H. Kang e J. Döllner. Image and Video Abstraction by Anisotropic Kuwahara Filtering . *Computer Graphics Forum*, 28(7):1955–1963, 2009. Special issue on Pacific Graphics 2009.

[KKD10] JE Kyprianidis, H Kang e J Döllner. Anisotropic kuwahara filtering on the gpu. *GPU Pro-Advanced Rendering Techniques*, pages 247–264, 2010.

[KR08] Mohan S Kankanhalli e Yong Rui. Application potential of multimedia information retrieval. *Proceedings of the IEEE*, 96(4):712–720, 2008.

[LGT+04] Thomas M Lehmann, MO Gold, Christian Thies, Benedikt Fischer, Klaus Spitzer, Daniel Keysers, Hermann Ney, Michael Kohnen, Henning Schubert e Berthold B Wein. Content-based image retrieval in medical applications. *Methods of Information in Medicine*, 43(4):354–361, 2004.

[LHY+09] D. Liu, X.S. Hua, L. Yang, M. Wang e H.J. Zhang. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pages 351–360. ACM New York, NY, USA, 2009.

[LME+11] Cynthia Liem, Meinard Müller, Douglas Eck, George Tzanetakis e Alan Hanjalic. The need for music information retrieval with user-centered and multimodal strategies. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 1–6. ACM, 2011.

[LMS+09] Stefanie Lindstaedt, Roland Mörzinger, Robert Sorschag, Viktoria Pammer e Georg Thallinger. Automatic image annotation using visual content and folksonomies. *Multimedia Tools and Applications*, 42(1):97–113, 2009.

[LSDJ06] Michael S Lew, Nicu Sebe, Chabane Djeraba e Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.

[LWL+07] Jing Liu, Bin Wang, Mingjing Li, Zhiwei Li, Weiying Ma, Hanqing Lu e Songde Ma. Dual cross-media relevance model for image annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 605–614. ACM, 2007.

[LZLM07] Ying Liu, Dengsheng Zhang, Guojun Lu e Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[MDWG10] Raphaël Marée, Philippe Denis, Louis Wehenkel e Pierre Geurts. Incremental indexing and distributed image search using shared randomized vocabularies. In *Proceedings of the international conference on Multimedia information retrieval*, pages 91–100. ACM, 2010.

[MMBG04] Henning Müller, Nicolas Michoux, David Bandon e Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International journal of medical informatics*, 73(1):1–23, 2004.

[NJT06] Eric Nowak, Frédéric Jurie e Bill Triggs. Sampling strategies for bag-of-features image classification. In *Computer Vision–ECCV 2006*, pages 490–503. Springer, 2006.

[NS06] David Nister e Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE, 2006.

[OAP+06] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald e Christina Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*, pages 18–25. Citeseer, 2006.

[OAS10] Tore Opsahl, Filip Agneessens e John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[OT01] Aude Oliva e Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[RCPC+10] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy e Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, pages 251–260. ACM, 2010.

[SBP+09] Julian Stottinger, Jana Banova, Thomas Ponitz, Nicu Sebe e Allan Hanbury. Translating journalists' requirements into features for image search. In *Virtual Systems and Multimedia, 2009. VSMM'09. 15th International Conference on*, pages 149–153. IEEE, 2009.

[SH09] Ruslan Salakhutdinov e Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

[SW06] Malcolm Slaney e William White. Measuring playlist diversity for recommendation systems. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 77–82. ACM, 2006.

[SWS+00] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta e Ramesh Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[TP91] Matthew Turk e Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[Vas07] Nuno Vasconcelos. From pixels to semantic spaces: Advances in content-based image retrieval. *Computer*, 40(7):20–26, 2007.

[vLGOvZ09] Reinier H van Leuken, Lluis Garcia, Ximena Olivares e Roelof van Zwol. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*, pages 341–350. ACM, 2009.

[VWN09] Melanie A Veltman, Michael Wirth e JingBo Ni. Impediments to general purpose content based image search. In *Proceedings of the 2nd Canadian Conference on Computer Science and Software Engineering*, pages 257–265. ACM, 2009.

[WBDB+06] James Z Wang, Nozha Boujemaa, Alberto Del Bimbo, Donald Geman, Alexander G Hauptmann e Jelena Tesić. Diversity in multimedia information retrieval research. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 5–12. ACM, 2006.

[WH08] Weining Wang e Qianhua He. A survey on emotional semantic image retrieval. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 117–120. IEEE, 2008.

[WMGH09] Yong Wang, Tao Mei, Shaogang Gong e Xian-Sheng Hua. Combining global, regional and contextual features for automatic image annotation. *Pattern Recognition*, 42(2):259–266, 2009.

[WTF08] Yair Weiss, Antonio Torralba e Rob Fergus. Spectral hashing. In *Advances in neural information processing systems*, pages 1753–1760, 2008.

[WYYH09] Lei Wu, Linjun Yang, Nenghai Yu e Xian-Sheng Hua. Learning to tag. In *Proceedings of the 18th international conference on World wide web*, pages 361–370. ACM, 2009.

[WZZ08] Changhu Wang, Lei Zhang e Hong-Jiang Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 355–362. ACM, 2008.

[YJHN07] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann e Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.

[ZG08] Qing-Fang Zheng e Wen Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(1):7, 2008.