Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Recuperação de Informação Multimédia em Memórias Pessoais

Rui Manuel Feliciano de Jesus
(Mestre)

Dissertação apresentada para a obtenção do Grau de Doutor em Informática pela Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.

Lisboa
Setembro 2009

Esta dissertação foi desenvolvida sob orientação do Professor Doutor Nuno Manuel Robalo Correia do Departamento de Informática da Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa e sob co-orientação do Professor Doutor Arnaldo Joaquim de Castro Abrantes do Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa.

À minha filha Beatriz

Agradecimentos

Em primeiro lugar, queria agradecer aos meus orientadores, Prof. Nuno Correia e Prof. Arnaldo Abrantes, pela forma como acompanharam e se envolveram na orientação desta tese. Algumas vezes foram mais do que orientadores científicos. A sua competência, rigor científico e espírito crítico foram fundamentais para a realização deste trabalho.

Os meus agradecimentos a todas as pessoas que me apoiaram ao longo do desenvolvimento desta tese, incluindo familiares e amigos, colegas do ISEL e da Faculdade de Ciências e Tecnologia.

Queria também agradecer às três instituições que de forma diferente contribuíram para a concretização deste trabalho. Ao ISEL, nomeadamente ao DEETC, por me ter libertado de outras tarefas permitindo assim que eu estivesse mais disponível para o desenvolvimento desta tese. À Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT/UNL) que através do Centro de Investigação em Informática e Tecnologias da Informação (CITI) disponibilizou todas as condições para a execução deste trabalho e, finalmente, um agradecimento à Fundação para a Ciência e Tecnologia pelo apoio financeiro.

Em particular, queria agradecer às pessoas do Grupo de Multimédia Interactiva (IMG) do Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa que colaboraram comigo no desenvolvimento deste trabalho, Rute Frias, Tiago Martins, Edgar Santos, Ricardo Dias, Duarte Gonçalves e Filipe Grangeiro.

Agradeço a todos que contribuíram com fotografias pessoais para as colecções de fotos utilizadas nesta dissertação, Manuela Vieira, Paula Louro, João Martins, Pedro Jorge, Gonçalo Marques, Rute Frias, Rossana Santos e Nuno Correia.

Um agradecimento também para todos os restantes membros do IMG pela troca de conhecimentos, ideias, pelas conversas e lanches que também foram contributos válidos para o desenvolvimento deste trabalho. Em especial, quero agradecer ao Diogo Cabral pelas grandes discussões sobre diversos temas, ao Hugo Vieira, ao Duarte Gonçalves e ao Filipe Grangeiro pelas discussões sobre futebol e às colegas do IMG pelos bolos que faziam e partilhavam com os restantes membros, nomeadamente, a Carmen, a Rossana e a Guida.

À minha mãe pelo apoio dado ao longo de toda a minha vida. Finalmente, um agradecimento muito especial à minha filha por me "aborrecer" de uma forma tão reconfortante ao longo deste trabalho.

Resumo

Esta dissertação descreve soluções para a recuperação e anotação de informação multimédia em memórias pessoais. Estas propostas incluem métodos de recuperação e anotação de fotografias, baseados na extracção de informação semântica em imagens, e aplicações de recuperação e anotação em três cenários diferentes: ambientes domésticos, cenários móveis e em actividades de entretenimento.

Os métodos propostos para recuperar e anotar fotos utilizam informação multimodal, nomeadamente, características visuais, informação de áudio e metadados contextuais obtidos no instante de captura. Esta metodologia baseia-se na análise semântica de imagens, obtida utilizando estes dados, para recuperar e anotar imagens automaticamente. Os metadados contextuais utilizados são o instante de captura da foto e a sua localização geográfica.

Para ambientes domésticos, é proposta uma aplicação para partilha de momentos relevantes do passado baseada na pesquisa e visualização de fotos pessoais. A interface inclui uma linguagem visual baseada em ícones para definir a pesquisa e permite interrogar a colecção pessoal com objectos físicos. Para cenários móveis, nomeadamente em actividades turísticas, é descrita uma aplicação para partilha de fotos no momento da visita a locais de interesse, por exemplo, museus ou lugares históricos. A partilha de imagens é baseada no método de recuperação proposto. A aplicação permite a captura de fotografias e a sua anotação com informação de áudio obtida segundos após a captura e com coordenadas de localização geográfica obtidas pelo receptor GPS (Global Positioning System) incluído. Para corrigir erros produzidos pelo método automático de recuperação e anotação é proposta uma aplicação para anotação semi-automática de imagens. Esta aplicação inclui um jogo de computador para anotar imagens baseado numa interface gestual de modo a motivar os utilizadores para a tarefa da anotação.

A tese apresenta as soluções referidas descrevendo a metodologia de concepção utilizada no desenvolvimento das aplicações, incluindo os resultados obtidos nos testes de usabilidade efectuados. São também apresentados e discutidos resultados da avaliação efectuada para validar os métodos de recuperação e anotação em cada uma das aplicações.

Abstract

This dissertation describes solutions to retrieve and annotate multimedia information in personal memories. It presents methods to retrieve and annotate personal pictures based on the extraction of semantic information in images. Applications to retrieve and annotate photos are also proposed in three different scenarios: domestic environments, mobile contexts and in entertainment activities.

The methods proposed use multimodal information namely visual features, audio information and contextual metadata obtained at capture time. This data is used to semantically analyze pictures. The proposed approaches are based on this semantic information to retrieve and annotate images automatically. The contextual metadata used are the capture time and the geographic location of the picture.

To share past relevant moments in domestic environments it is proposed an application based on the search and the visualization of personal pictures. The interface includes a visual language based on icons to define queries and it allows the user to query the database with physical objects. For mobile environments, namely in tourism activities, it is proposed an application to share pictures when visiting points of interest, for instance, historical sites or museums. Images are shared based on the retrieval method proposed. The application allows to capture pictures and their annotation with audio information obtained at capture time and with geographic location information captured by the GPS receiver included. To correct errors generated by the automatic retrieval and annotation methods, it is proposed an application to annotate images in a semi-automatic way. This application includes a computer game for image annotation using a gesture interface to motivate users for the annotation task.

This dissertation presents the mentioned solutions describing the design methodology used in the development of the applications, including the usability tests performed. The results of the evaluation provided to validate the retrieval and annotation methods in each application are also presented and discussed.

Lista de Símbolos

A	Parâmetro da sigmóide
$A_{j,i}$	Anotação na imagem I_j do conceito w_i
a	Comprimento do eixo maior de uma elipse
a_j	Peso da $j^{ésima}$ característica
B	Parâmetro da sigmóide
b	Comprimento do eixo menor de uma elipse
C	Componente de cor
Cob	Cobertura
$C_{group}(m)$	Função que calcula a confiança do grupo numa anotação
C_{img}	Colecção de imagens
$C_{player}(n)$	Função que calcula a confiança no jogador
C_i	Cobertura por imagem
C_w	Cobertura por palavra
c	Vector de coeficientes da função discriminante do RLSC
c_n	Componente de cor normalizada
d	Vector de ocorrências de termos
d_x^j	Vector de ocorrências de termos visuais da $j^{ésima}$ característica
d_{gps}	Direcção
d_{max}	Distância temporal máxima entre imagens
$d(t_i)$	Distância temporal entre a imagem capturada no instante t_i e a imagem seguinte
e_i	Erro de classificação
$f_i(d_x^j)$	Função discriminante do classificador binário para o conceito i e característica d_x^j
$g_{m,n}$	Resposta impulsiva de um filtro de Gabor
$h(d)$	Função discriminante utilizada no método específico para "Faces"
$h_i(d)$	Classificador fraco
I	Imagem
$K(d_i, d)$	*Kernel* Gaussiano

K	Número de conceitos da interrogação
K_{moves}	Número de jogadas correctas necessárias para chegar ao valor máximo de confiança
k_{conf}	Confiança máxima atribuída a um jogador
k_g	Parâmetro da exponêncial utilizada para calcular a confiança no grupo
k_p	Constante utilizada para incrementar a confiança do jogador
L	Subconjunto de imagens da colecção utilizado em cada nível do jogo Tag Around
L_{gps}	Subconjunto de imagens da colecção que respeitam um determinado critério geográfico
lat	Latitude
lon	Longitude
M	Número de imagens do conjunto de treino
m	Número de vezes que um conceito foi anotado numa imagem
m_{con}	Vector de metadados contextuais
N	Número de imagens da colecção
N_A	Número de anotações de um conceito
N_{col}	Número de colunas de pixels de uma imagem
NI_{auto}	Número de imagens anotadas automaticamente
NI_{corr}	Número de imagens anotadas correctamente
NI_{manu}	Número de imagens anotadas manualmente
$N_{l,k}$	Número de ocorrências do termo t_{wl} na imagem I_k
N_{lin}	Número de linhas de pixels de uma imagem
N_{orient}	Número de orientações
N_{pixels}	Percentagem de pixels de uma região da imagem
N_{scale}	Número de escalas
N_t	Número de termos do vocabulário
N_{upd}	Número de anotações necessários para actualizar um conceito
NW_{corr}	Número de palavras anotadas correctamente numa imagem
NW_{manu}	Número de palavras anotadas manualmente
n	Número de anotações correctas efectuadas por um utilizador
n_{rec}	Número de imagens recuperadas
o	Orientação
P	Número de conceitos do vocabulário
P_{gps}	Ponto de coordenadas geográficas
$Prec$	Precisão
P_w	Precisão por palavra
$p(w/I)$	Probabilidade de w dada a imagem I
$p_t(w_i/I)$	Probabilidade do conceito w_i na imagem I utilizando informaçao temporal
Q	Interrogação com conceitos
Q_g	Interrogaçao geográfica
q	Interrogação representada por um vector de ocorrências
R	Número de características visuais
Rel	Conjunto das imagens relevantes para uma interrogação
r	Lista com as posições das imagens relevantes na lista das imagens recuperadas
r_{earth}	Raio da terra
r_{query}	Raio da região definida na interrogação geográfica

$Sim(Q,I)$	Função que calcula a semelhança entre a Q e I
S_m	Conjunto de treino
$S_{new}(I,w,n)$	Função que calcula a pontuação para a primeira anotaçao de um conceito numa imagem
$S_{total}(I,w,n,m)$	Função que calcula a pontuação atribuída a uma jogada
s	Escala
T	Vector com os instante de captura das imagens da colecção
t	Instante de tempo
th	limiar utilizado para anotar um conceito numa imagem
t_W	Termos reconhecidos a partir de áudio
V_{con}	Vocabulário de conceitos semânticos
V_{sc}	Subconjunto de conceitos
V_t	Vocabulário de termos
w	Variável aleatória de Bernoulli
$W_{j,i}$	Peso do termo j no documento i
X_{img}	Matriz termo-documento
x_k	Vector de características visuais de uma imagem
xr_k	Vector de características visuais de uma região da imagem
Y	Vector composto pelo y_i de um conjunto de imagens
y_i	Etiqueta que define a classe de uma imagem
α_i	Peso atribuído a uma característica seleccionada
α_{t_i}	Peso de uma anotação numa imagem capturada no instante t_i
μ	Média
σ	Desvio padrão

Acrónimos e Abreviaturas

AP	Average Precision
CBIR	Content Based Image Retrieval
DoG	Difference-of-Gaussian
DTW	Dynamic Time Warping
EM	Expectation-Maximization
EMD	Earth Movers Distance
EXIF	Exchangeable Image File Format
GPRS	General Packet Radio Service
GPS	Global Positioning System
HMM	Hidden Markov Model
HSV	Hue, Saturation e Value
JPEG	Joint Photographic Experts Group
LSA	Latent Semantic Analysis
LSCOM	Large Scale Concept Ontology for Multimedia
MAP	Mean Average Precision
MDL	Minimum Description Length
MDS	Multi-Dimensional Scaling
Memex	Memory Extender
MIL	Multiple Instance Learning
MPEG	Moving Picture Experts Group
NLS	oN Line System
OGRE	Object-oriented Graphics Rendering Engine
PC	Personal Computer
PCA	Principal Component Analysis
PDA	Personal Digital Assistant
PDH	Personal Digital Historian
PIM	Personal Information Management
PLSA	Probabilistic Latent Semantic Analysis

RGB Red, Green e Blue
RLS Regularized Least Squares
RLSC Regularized Least Squares Classifier
SIFT Scale-Invariant Feature Transform
SVD Singular Value Decomposition
SVM Support Vector Machine
SVT Semantic Visual Templates
TF-IDF Term Frequency-Inverse Document Frequency
WLAN Wireless Local Area Network
WWMX World Wide Media eXchange

Conteúdo

Lista de Figuras

Lista de Tabelas

1

Introdução

Conteúdo

Este capítulo introduz o trabalho desenvolvido no âmbito desta dissertação. Apresenta as motivações para abordar as memórias pessoais e as principais dificuldades identificadas pela comunidade científica na recuperação destas memórias. Termina com a descrição das contribuições e com a organização da tese.

1.1 Enquadramento e Motivação

A partilha de memórias e experiências é uma actividade fundamental dos seres humanos, praticada ao longo de milhares de anos, que permite a troca de conhecimento entre gerações e culturas. A fotografia, porque inclui informação visual, é uma das formas mais ricas não só para registar momentos relevantes mas também para transportar esta informação entre familiares, amigos ou até desconhecidos.

Avanços recentes na tecnologia têm vindo a alterar os conceitos de captura, de armazenamento e de partilha da informação visual. Cada vez existem mais dispositivos móveis com câmara fotográfica integrada, com considerável capacidade de armazenamento e que permitem a comunicação de informação através de diversos métodos. Este progresso, nomeadamente a integração de dispositivos de captura em telemóveis, aliado ao baixo custo das máquinas fotográficas digitais, dos cartões de memória e de toda a tecnologia relacionada têm vindo a contribuir para um aumento da popularidade da fotografia digital. Actualmente, a tecnologia existente permite que as pessoas possam tirar fotografias em qualquer momento e em qualquer lugar e através da World Wide Web partilhar esta informação com outras pessoas. Como consequência, uma vasta quantidade de informação pessoal tem vindo a ser produzida e armazenada. Por exemplo, durante o ano de 2006 foram colocadas no Flickr cerca de 250 milhões de fotos. Com este crescimento das memórias pessoais, mais experiências são preservadas mas também maiores dificuldades surgem no acesso à informação que se pretende encontrar.

Assim, para partilhar uma experiência será necessário desenvolver sistemas para recuperar de forma eficiente imagens de uma colecção de memórias pessoais e será preciso conceber aplicações baseadas nestes sistemas, que permitam ao cidadão comum recuperar as suas memórias e partilhá-las em qualquer momento e em qualquer lugar. No caso do Flickr, o problema da recuperação é parcialmente resolvido através de anotações produzidas manualmente pelos utilizadores mas nem todos estão dispostos a realizar esta tarefa, que pode ser penosa para grandes colecções de imagens.

A tecnologia existente permite pensar no desenvolvimento de aplicações para diversos tipos de utilizadores, nomeadamente os idosos que à partida serão as pessoas com maior dificuldade em lidar com a tecnologia, mas que também são os utilizadores com mais experiências para partilhar. Por outro lado, as actuais características dos dispositivos móveis também permitem pensar na concepção de aplicações de partilha de imagens fora do ambiente doméstico, por exemplo, em situações ocasionais ao encontrar um conhecido na rua ou num restaurante ou em momentos de lazer ao visitar um local de interesse. Em qualquer actividade os dispositivos móveis podem ser excelentes auxiliares porque permitem a pesquisa no local e no momento do acontecimento.

Estes desafios científicos com memórias pessoais seguem a visão de Vannevar Bush publicada no artigo "As We May Think" [Bush45] em 1945 e proposta com as limitações da tecnologia existente na altura. Mais recentemente, vários têm sido os trabalhos apresentados [Gray03,Fitzgibbon03,Rowe05] apontando os principais problemas da captura e reutilização de informação pessoal e referindo o tópico das memórias pessoais como um grande desafio científico para os anos seguintes. Este tópico foi também considerado como grande desafio científico para as ciências da computação em 2002 no Reino Unido, sendo criada um rede financiada de investigadores de várias áreas designada por Memories for Life, que desde essa data desenvolve

trabalho científico em memórias pessoais.

Esta tese apresenta contribuições nos aspectos relacionadas com memórias pessoais constituídas por fotografias digitais, nomeadamente no desenvolvimento de soluções para o problema da recuperação de fotos em colecções pessoais e na concepção de aplicações de recuperação de informação, que permitam reviver e partilhar momentos importantes do passado em ambientes domésticos e em locais de interesse. Um dos problemas apontados na recuperação de fotos é a anotação de imagens, por isso este trabalho também foca o desenvolvimento de métodos para melhorar esta tarefa. As soluções desta tese, para a recuperação e anotação de imagens, baseiam-se num modelo de análise semântica de imagens que visa a realização destas tarefas de forma automática ou semi-automática.

1.2 Aproximação e Objectivos

Depois do enquadramento e da apresentação das motivações do tópico fundamental desta tese, nesta secção são descritos os objectivos principais e a estratégia seguida para atingir esses objectivos, numa perspectiva inicial do trabalho a desenvolver. Considerando o domínio de aplicação e tendo em conta a caracterização do cenários prováveis em ambientes domésticos e ambientes móveis, nomeadamente em actividades de turismo, o primeiro passo consiste na especificação de um modelo conceptual de análise semântica de imagens, utilizando informação multimodal. São definidas as características visuais a utilizar no desenvolvimento de um modelo probabilístico genérico para qualquer conceito semântico. Este modelo também inclui características visuais específicas para conceitos mais complexos. Ainda nesta fase de definição do modelo conceptual, integram-se os metadados contextuais obtidos no instante de captura e a informação de áudio. Depois de definido, implementa-se o modelo de análise semântica multimodal de imagens.

No passo seguinte, definem-se conceitos aplicáveis às memórias pessoais e treinam-se modelos para os conceitos semânticos definidos. Esta fase inclui a extracção das características visuais e de áudio que são utilizadas no treino dos modelos. Também são desenvolvidas características específicas para detectar faces em imagens. A seguir, desenvolvem-se métodos automáticos de recuperação e anotação de imagens utilizando estes modelos semânticos. Com base nestes métodos, implementa-se uma biblioteca que disponibiliza um conjunto de ferramentas de programação para serem utilizadas em aplicações de recuperação ou pesquisa e anotação de imagens.

Utilizando a biblioteca, realiza-se o estudo e a concepção de uma aplicação de recuperação para partilhar memórias pessoais em ambientes domésticos, através de um computador pessoal, e outra aplicação para dispositivos móveis para partilhar fotos no momento da visita a locais de interesse. São requisitos principais no desenvolvimento destas interfaces a forma como o sistema disponibiliza a definição da pesquisa e a visualização dos resultados. Em ambos os requisitos é considerado o comportamento da memória humana. Para atenuar os erros dos métodos de recuperação e anotação automáticos, desenvolve-se uma aplicação semi-automática para anotação de imagens baseada num jogo de computador. Os pressupostos na concepção desta interface baseiam-se na integração de técnicas automáticas de anotação com métodos manuais através de uma tarefa que motiva o utilizador.

Para finalizar este estudo, experimentam-se e avaliam-se os métodos de recuperação e ano-

tação de imagens e as aplicações referidas. Os métodos de recuperação e anotação são avaliados nas aplicações, utilizando várias colecções pessoais com métricas conhecidas e regularmente utilizadas na área de recuperação de informação. Os modelos semânticos são testados individualmente para cada característica utilizada e são comparados para várias combinações de características. No caso das aplicações, utilizam-se metodologias típicas de desenvolvimento de interfaces, em processos iterativos, desde a definição de funcionalidades aos testes de usabilidade com utilizadores incluindo protótipos em papel. Estes testes visam a concepção da aplicação e a melhoria da interface em todos os seus aspectos e a validação, por parte dos utilizadores, dos resultados obtidos utilizando os métodos propostos para recuperação e anotação de imagens.

1.3 Panorâmica da Área Científica

Esta tese propõe soluções para recuperação de informação multimédia em memórias pessoais e apresenta aplicações para partilha de experiências pessoais através de fotografias baseadas na metodologia proposta. A captura, o armazenamento e a reutilização de memórias pessoais foi identificada em [Gray03, Rowe05, Fitzgibbon03] como um dos grandes desafios para as áreas das ciências computacionais. Das aproximações que se dedicaram à captura, armazenamento e recuperação, as que mais influenciaram o trabalho desenvolvido foram propostas pela Microsoft, o SenseCam [Gemmell04] para a captura e o MyLifeBits [Gemmell02] para gerir a informação pessoal, e pela universidade de Tóquio [Hori03, Tancharoen05]. Estas propostas integram uma câmara de vídeo com vários sensores para capturar informação e desenvolveram sistemas para recuperar a informação.

Os trabalhos iniciais de recuperação de imagens baseada em conteúdo centravam-se na comparação entre o vector de características da interrogação (imagem ou esboço) e representações idênticas das imagens da colecção [Smeulders00], contudo, devido ao problema da falha semântica [Lew06], foram propostas soluções semi-automáticas baseadas na técnica de retroacção de relevância [Zhou03] e soluções automáticas baseadas em conceitos semânticos [Datta08]. Um dos trabalhos pioneiros a associar automaticamente conceitos semânticos a imagens foi proposto por Mori *et al.* [Mori99]. Durante os últimos anos, várias aproximações têm sido propostas para estimar conceitos semânticos [Duygulu02, Carneiro05, Fan07]. A maioria utiliza apenas a informação visual para treinar os modelos mas em [Luo06] foi proposta uma aproximação que combina a informação visual com a informação contextual tal como a solução proposta nesta tese.

Em relação às aplicações para computador pessoal, existem vários exemplos comerciais, iPhoto, Picasa, ACDSee ou Adobe Photoshop Album, para gerir memórias pessoais que se baseiam na utilização de anotação manual com o objectivo de melhor organizar e recuperar documentos em colecções de fotos. O Fotofile [Kuchinsky99] é um dos primeiros trabalhos a integrar técnicas de anotação automática baseada em conteúdo com anotação manual para organizar fotos pessoais. Mais recentemente, foi proposto o MediAssist [Hare05] que inclui informação contextual combinada com o conteúdo da imagem para recuperação. Para a Web, o Flickr [Flickr04] é uma aplicação muito usada mas não inclui anotação automática de imagens, ao contrário do ALIPR [Li06] que utiliza modelos semânticos para anotação automática em tempo real na Internet.

Para dispositivos móveis têm sido propostas interfaces de pesquisa para vários tipos de aplicação, por exemplo, aplicações de realidade aumentada [Kim05], aplicações de partilha de experiências [Anguera08, Gurrin05] ou guias de turismo [Fockler05]. A maioria utiliza tecnologia CBIR (Content Based Image Retrieval) para recuperar imagens. Os sistemas MAMI [Anguera08] e PhoneGuide [Fockler05] processam as imagens localmente ao contrário da maioria dos restantes que enviam as fotos para um servidor. O sistema MediAssist utiliza informação contextual para pesquisar fotos e é o único que utiliza recuperação de fotos com informação semântica como é proposto neste documento.

Para tornar a anotação manual de imagens uma tarefa que motiva o utilizador, Luis von Ahn e Laura Dabbish propuseram o ESP GAME [VonAhn04]. O jogo ESP é jogado por dois jogadores escolhidos aleatoriamente utilizando a Web. Sempre que ambos os jogadores escrevam a mesma palavra para a mesma imagem ganham pontos. Seguindo a mesma ideia em [Nicholas07] foi proposto o PhotoPlay, um jogo para ser jogado num ecrã horizontal (*tabletop*) com todos os jogadores no mesmo local a utilizar um controlador de jogo para anotar imagens e em [Tuulos07] foi proposto o Manhattan Story Mashup, com o objectivo de contar histórias com imagens sobre Manhattan. Estes trabalhos não incluem métodos automáticos ou semi-automáticos para anotar imagens de forma divertida como é proposto nesta tese.

1.4 Contribuições

Considerando os objectivos referidos, as principais contribuições desta tese são as seguintes:

- Concepção de um modelo de análise multimodal de imagens para aplicações de recuperação e anotação automática com conceitos semânticos. As características mais relevantes são:
  - O modelo de análise semântica serve de base para os métodos de recuperação e anotação de imagens propostos nesta tese;
  - O modelo combina informação baseada no conteúdo com informação contextual e extraída de áudio. São integradas características visuais com informação de áudio, temporal e espacial (geo-referenciada) obtida no instante de captura;
  - O modelo é composto por um método para treinar conceitos genéricos e permite a inclusão de métodos adicionais específicos para conceitos mais complexos;
  - O método genérico é baseado no classificador Regularized Least Squares e numa função sigmóide;
  - O modelo inclui um método específico para detectar faces baseado em características visuais adequadas ao conceito "Faces".
  - O método específico para o conceito "Faces" permite treinar conceitos adicionais, por exemplo o conceito "Género" (masculino/feminino).
- Desenvolvimento de um algoritmo semi-automático para anotação de imagens baseado no método proposto para anotação automática e nas intervenções dos utilizadores num jogo de computador.
- Desenvolvimento de uma técnica de definição de pesquisas baseada no *drag & drop* de ícones para um espaço reservado para a construção da interrogação, onde o utilizador

pode combinar através de uma linguagem visual diversos tipos de elementos para definir a pesquisa.

- Desenvolvimento de uma aplicação para gerir memórias pessoais em ambientes domésticos baseada no sistema de recuperação de imagens com conceitos semânticos. A aplicação tem as seguintes características:
  - A recuperação de imagens utiliza a técnica de definição de pesquisas baseada na linguagem visual que é proposta nesta tese;
  - A aplicação permite interrogar a base de dados com objectos físicos;
  - A aplicação inclui dois modos de anotação de imagens com áudio. No modo "Slideshow" todos os comentários são gravados sem a intervenção directa do utilizador. No outro modo de anotação, o utilizador indica explicitamente que pretende capturar áudio para anotar uma imagem.
- Concepção de uma aplicação para dispositivos móveis de partilha de fotos no momento da visita a locais de interesse, através de um motor de pesquisa baseado no sistema de recuperação proposto. As características principais desta aplicação são:
  - Aplicação que guia os visitantes de um local de interesse com fotos partilhadas;
  - Recuperação de imagens que utiliza a técnica de definição de pesquisas baseada na linguagem visual que é proposta nesta tese;
  - Aplicação de recuperação de fotos para dispositivos móveis que inclui pesquisa com conceitos semânticos;
  - Anotação da foto no instante de captura através da gravação em ficheiro de áudio de comentários do utilizador.
- Concepção de uma aplicação semi-automática para anotação de imagens baseada num jogo de computador. As componentes mais relevantes da aplicação são:
  - O jogo é baseado numa interface gestual e utiliza o reconhecimento facial para identificação do jogador;
  - A aplicação inclui o método de anotação semi-automática de imagens baseado nos conceitos semânticos;
  - O jogo pode ser jogado em lugares públicos, dado que o computador não precisa de estar visível e não requer dispositivos adicionais para ser jogado.
- Implementação de uma biblioteca com os algoritmos propostos para recuperação e anotação de imagens que é utilizada nas aplicações referidas e também numa aplicação para Web, com funções idênticas às características da aplicação proposta para dispositivos móveis.
- Contribuição no desenvolvimento de projectos em curso ou de projectos propostos, nomeadamente:
  - Projecto InStory 2 - desenvolvido no CITI/Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT/UNL) em

colaboração com a fundação Cultursintra/Quinta da Regaleira, Sintra, Portugal. O objectivo do projecto é a definição de uma arquitectura e a implementação de uma plataforma tecnológica adequada à navegação em narrativas espaciais, acesso a informação geo-referenciada e participação em actividades lúdicas através de dispositivos móveis. O método de recuperação de imagens proposto nesta tese foi utilizado numa aplicação para Web e numa aplicação desenvolvida para dispositivos móveis.

- Projecto Comunicação Pública da Arte - projecto a decorrer numa colaboração entre a Faculdade das Ciências Sociais e Humanas da Universidade Nova de Lisboa (FCSH/UNL) e o CITI/Departamento de Informática da FCT/UNL. O projecto inclui o desenvolvimento de uma versão do jogo proposto para anotação de imagens, adaptado para ser jogado por vários visitantes de um museu utilizando diversas formas de interacção.
- Projecto VideoFlow - projecto a realizar em parceria entre uma empresa produtora audiovisual, a Duvideo, e o CITI/Departamento de Informática da FCT/UNL com o objectivo de desenvolver um sistema suportando interfaces avançadas para o acesso a arquivos de vídeo, com base em metadados extraídos com uma combinação de processos automáticos e conhecimento humano. Neste projecto os modelos semânticos são aplicados a vídeo.
- Projecto CRUSH - projecto financiado pela Fundação para a Ciência e a Tecnologia, que visa desenvolver uma nova abordagem para recuperar clip-arts, independentemente do seu formato (raster e vectorial) e que combina as potencialidades das duas técnicas. A solução permitirá a recuperação de clip-arts utilizando esboços como interrogações. São utilizadas as técnicas de processamento de imagem e recuperação de informação, desenvolvidas no âmbito desta dissertação, para descrever o conteúdo de clip-arts. O projecto envolve a colaboração de duas instituições de investigação, o Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa do Instituto Superior Técnico da Universidade Técnica de Lisboa (INESC ID/IST/UTL) e o CITI/Departamento de Informática da FCT/UNL.

1.5 Publicações

Nesta secção são apresentadas as publicações que difundiram os resultados do trabalho de investigação realizado e que abordam os aspectos mais relevantes do trabalho desenvolvido nesta tese. As publicações são agrupadas em quatro tópicos principais:

- **Análise semântica de imagens** - inclui o método de recuperação proposto, o método automático e semi-automático para anotação de imagem e as características específicas para treinar faces e detectar o género da pessoa. As publicações neste tópico são as seguintes:

  - Filipe Grangeiro, Rui Jesus, Nuno Correia, "Face Recognition and Gender Classification in Personal Memories". IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pages 1945-1948, Taipei, Taiwan, April 2009.

- Filipe Grangeiro, Rui Jesus, Nuno Correia, "Detecção e Reconhecimento de Faces para Aplicações Multimédia", 4ª Jornadas de Engenharia de Electrónica e Telecomunicações e de Computadores, Lisboa, Portugal 2008, (*Prémio de melhor artigo de estudante*).
- Rui Jesus, Duarte Goncalves, Arnaldo Abrantes, Nuno Correia, "Playing Games as a Way to Improve Automatic Image Annotation". CVPRW 08: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Workshop on Semantic Learning Applications in Multimedia (SLAM08), pages 1-8, Anchorage, Alasca, EUA, June 2008.
- Rui Jesus, Ricardo Dias, Rute Frias, Arnaldo Abrantes, Nuno Correia, "Sharing Personal Experiences while Navigating in Physical Spaces". ACM SIGIR Conference on Research and Development in Information Retrieval, Multimedia Information Retrieval Workshop, Amsterdam, The Netherlands, July 2007.
- Rui Jesus, Arnaldo Abrantes, Nuno Correia, "Photo Retrieval from Personal Memories using Generic Concepts". Advances in Multimedia Information Processing - PCM 2006, Springer LNCS, 2006, volume 4261: p. 633-640. Hangzhou, China.

- **Aplicações para dispositivos móveis** - as publicações da aplicação proposta para dispositivos móveis de recuperação e anotação de imagens no instante de captura são as seguintes:

  - Rui Jesus, Ricardo Dias, Rute Frias, Arnaldo Abrantes, Nuno Correia, "Memoria Mobile: Sharing Pictures of a Point of Interest". AVI 08: Proceedings of the Working Conference on Advanced Visual Interfaces, pages 412-415, ACM, Naples, Italy, 2008.
  - Rui Jesus, Ricardo Dias, Rute Frias, Nuno Correia, "Geographic Image Retrieval in Mobile Guides". GIR 07: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, pages 37-38, Lisbon, Portugal, 2007.
  - Ricardo Dias, Rui Jesus, Rute Frias, Nuno Correia, "Mobile Interface of the Memoria Project". SIGIR 07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 904-904, Amsterdam, The Netherlands, 2007.
  - Rui Jesus, Tiago Martins, Rute Frias, Arnaldo Abrantes Nuno Correia, "PhotoNav: a System to Capture, Share and Access Personal Memories". Memories for Life Colloquium, London, England, 2006.

- **Aplicações de entretenimento** - as publicações referentes ao jogo de computador para anotação semi-automática de imagens são apresentadas a seguir:

  - Duarte Gonçalves, Rui Jesus, Filipe Grangeiro Teresa Romão, Nuno Correia, "Tag around: a 3D Gesture Game for Image Annotation". ACE 08: Proceedings of the 2008 ACM International Conference on Advances in Computer Entertainment Technology, pages 259-262, Yokohama, Japan, 2008.
  - Duarte Gonçalves, Rui Jesus, Nuno Correia, "A Gesture Based Game for Image Tagging". CHI 08: ACM CHI 08 Extended Abstracts on Human Factors in Computing Systems, pages 2685-2690, Florence, Italy, 2008.

– Duarte Gonçalves, Rui Jesus, Filipe Grangeiro, Nuno Correia, "Tag Around - Interface Gestual para Anotação de Imagens". 3rd Portuguese Conference on Human Computer Interaction, Évora, Portugal, 2008.

- **Aplicações para ambientes domésticos** - inclui a aplicação para recuperação e anotação para computadores pessoais.

  Rui Jesus, Edgar Santos, Rute Frias, Nuno Correia, "An Interface to Explore Personal Memories". 15th Portuguese Computer Graphics Group Conference (EPCG) , Porto Salvo, Portugal, 2007.

1.6 Organização da Tese

Esta secção apresenta a estrutura deste documento. A tese para além deste capítulo de introdução e dos seguintes que abordam o domínio científico e o trabalho relacionado, é constituída por quatro capítulos com o trabalho desenvolvido (um para cada tópico principal), por um capítulo de avaliação e por um capítulo com as conclusões e trabalho futuro. Estes capítulos estão organizados da seguinte forma:

- Capítulo 1. Introdução: Enquadra e motiva o problema, apresenta um resumo do trabalho relacionado e as contribuições mais relevantes do trabalho desenvolvido.

- Capítulo 2. Domínio, Objectivos e Conceitos: apresenta o domínio de aplicação, os objectivos da tese e descreve os conceitos mais relevantes do domínio científico.

- Capítulo 3. Trabalho Relacionado: neste capítulo são apresentados os trabalhos relacionados com as várias áreas de interesse para a tese, nomeadamente, a captura e recuperação de memórias pessoais, a recuperação baseada em conteúdo de imagens, a anotação de imagens e aplicações de recuperação e anotação de imagens em vários dispositivos.

- Capítulo 4. Recuperação e Anotação de Imagens: descreve os vários blocos dos métodos de recuperação e anotação de imagens propostos nesta tese. Apresenta as técnicas utilizadas para extrair vários tipos de informação de uma colecção de memórias pessoais, desde a informação visual à informação contida nos metadados contextuais, e descreve o modelo de análise semântica multimodal.

- Capítulo 5. Recuperação de Imagens em Ambientes Domésticos: apresenta a aplicação de recuperação e anotação de memórias pessoais desenvolvida para reviver momentos do passado em ambientes familiares. Inclui a descrição da linguagem visual para definir pesquisas e os métodos de anotação de imagens com áudio. Também descreve a metodologia utilizada na concepção da aplicação.

- Capítulo 6. Recuperação de Imagens em Locais de Interesse: descreve a aplicação para dispositivos móveis de partilha de memórias pessoais em qualquer lugar, nomeadamente em locais de interesse turístico. Apresenta também a técnica de definição de pesquisa utilizando a linguagem visual baseada em ícones e a metodologia utilizada no desenvolvimento da aplicação.

- Capítulo 7. Aplicação Semi-Automática de Anotação: descreve a aplicação para anotação semi-automática de imagens, principalmente o jogo baseado em gestos e reconhecimento facial. Apresenta a metodologia utilizada para integrar o método de anotação automática com o jogo, de forma a construir um método semi-automático para anotação.
- Capítulo 8. Avaliação: Este capítulo apresenta os resultados obtidos em todas as experiências realizadas nesta tese, nomeadamente, os testes efectuados para avaliar os métodos de anotação e recuperação propostos e os resultados dos testes realizados com utilizadores para avaliar as aplicações propostas.
- Capítulo 9. Conclusões e Perspectivas Futuras: A tese termina com a apresentação das conclusões resultantes dos testes realizados e com a descrição das perspectivas para trabalho futuro.

2

Domínio, Objectivos e Conceitos

Conteúdo

Este capítulo tem como finalidade enquadrar o trabalho em memórias pessoais proposto nesta tese, nas áreas de investigação de recuperação de informação multimédia e interacção pessoa máquina. São também apresentados os objectivos da tese e os conceitos mais relevantes.

2.1 Introdução

O trabalho proposto nesta tese está inserido na área de investigação de recuperação de informação multimédia aplicada ao domínio das memórias pessoais. O domínio de aplicação estabelece um conjunto de requisitos que condicionam o desenvolvimento das aplicações. Assim, este capítulo começa por introduzir o tópico, memórias pessoais, abordando a importância da recuperação de informação para as aplicações neste domínio. A seguir, são descritos os requisitos necessários para a construção de um sistema de recuperação de memórias pessoais e são apresentados os objectivos da tese. O capítulo termina com a descrição dos conceitos principais da área científica.

2.2 Memórias Pessoais

Ao longo da história, as pessoas sempre utilizaram objectos físicos (por exemplo, livros, jornais ou fotografias em papel) como referências de momentos relevantes, que mais tarde servem como auxiliares de memória para relembrar e partilhar experiências do passado. Com o avanço tecnológico, estes auxiliares externos transformaram-se em memórias digitais guardadas em formato electrónico em computador, por exemplo, na forma de mensagens de correio electrónico, fotos digitais ou vídeos. O domínio digital permite registar maior quantidade de informação e com maior exactidão. Por exemplo, a SenseCam [Gemmell04] permite a captura passiva (o utilizador apenas tem de a usar) de informação durante um dia inteiro e inclui vários tipos de sensores para ajudar a caracterizar melhor o momento. Este tipo de aplicações, dada a quantidade de informação capturada, levanta vários desafios científicos ao nível da gestão desta informação. Contudo, a procura de soluções para estes problemas é motivada pela utilidade desta informação não só na partilha de experiências pessoais mas também em outras aplicações em diversas áreas [Czerwinski06]:

- Auxiliar de memória humana nas actividades diárias, por exemplo, para encontrar objectos perdidos, rever reuniões de trabalho ou relembrar nomes de pessoas ou lugares;
- Partilha de experiências pessoais para reviver experiências com pessoas ou melhorar a comunicação entre gerações;
- Análise e reflexão pessoal, para compreender o desenvolvimento pessoal ou melhorar a saúde através de monitorização médica [Pratt06];
- Segurança, para provar álibis ou para detectar possíveis actos de terrorismo em locais públicos.

Actualmente, a fotografia digital é uma das formas mais populares para registar os momentos mais relevantes, não só devido ao baixo custo e à elevada capacidade de armazenamento dos vários dispositivos de captura que estão acessíveis à maioria das pessoas, mas também porque a maioria dos dispositivos móveis inclui uma câmara digital. Alguns destes dispositivos, começam a incluir sensores adicionais (por exemplo, receptores de GPS) para melhor capturar o momento e podem ter a capacidade de capturar informação de forma passiva através da utilização de aplicações para o efeito. No caso das aplicações para recordar momentos de lazer, a

ausência de captura passiva pode não representar uma desvantagem por permitir que o indivíduo seleccione os momentos relevantes, mas a inclusão de mais sensores é uma mais valia na gestão da informação capturada.

A gestão das memórias pessoais inclui o seu armazenamento e organização de forma a que o acesso a esta informação seja feito em tempo útil. Como é discutido em [Dix02], actualmente é possível gravar em vídeo com boa qualidade a vida inteira de um indivíduo. São necessários 100kbit/s e admitindo que em média uma pessoa vive 70 anos ($2.2x10^9$ segundos), serão necessários cerca de 27.5 Terabytes de memória. Se a Lei de Moore continuar a prevalecer nos próximos 70 anos poderá ser possível colocar a vida de uma pessoa num dispositivo electrónico da dimensão de um grão de areia. Assim, no contexto da captura não passiva, como é o caso das memórias relativas a momentos de lazer, a questão do armazenamento não constitui um problema.

A organização da informação apresenta uma maior relevância na resolução do problema. Qualquer solução terá de equilibrar os aspectos relacionados com o utilizador, nomeadamente a sua memória, com os recursos computacionais existentes. O utilizador, para recuperar uma experiência do passado, precisa ter algumas pistas sobre o evento, por isso é necessário ter em conta os aspectos funcionais da memória humana no desenvolvimento de sistemas que a possam complementar.

As categorias de memória humana podem ser subdivididas em [Endel02]:

- Memória não declarativa, composta por uma colecção de capacidades expressas através de actividades que não exigem acesso a nenhuma memória consciente (por exemplo, andar de bicicleta ou desenhar com precisão).

- Memória declarativa, permite a recordação consciente de factos e eventos ocorridos na vida do indivíduo. Divide-se em dois tipos: memória episódica e memória semântica. A memória episódica permite realizar uma viagem mental através do tempo e recordar as suas experiências prévias [Endel02]. A memória semântica refere-se á capacidade de relembrar factos e conhecimentos gerais sobre o mundo.

As memórias pessoais são recuperadas utilizando a memória episódica que está relacionada com a memória dos eventos, da data, do local e de outras pistas acerca da experiência do passado. Assim, as aplicações para recuperar memórias pessoais devem disponibilizar ao utilizador formas para aceder à informação utilizando as diversas pistas que a memória episódica usa para relembrar o passado. Por outro lado, o sistema tem de ter capacidade para indexar a informação através das pistas referidas, neste caso a tarefa da anotação da informação assume um papel fundamental no sistema de recuperação de memórias pessoais.

A interface é outra questão importante na concepção do sistema de recuperação de memórias pessoais. Neste tópico é preciso ter em conta as características dos utilizadores tipo e também o contexto (por exemplo, ambientes móveis ou ambientes domésticos) em que é efectuada a recuperação do evento do passado.

2.3 Requisitos

No desenvolvimento de aplicações para reviver experiências do passado utilizando fotografias, as tarefas de recuperação e anotação têm um papel preponderante na organização da informa-

ção, de modo a que a aplicação responda às pistas geradas pela memória episódica. Por outro lado, a interface da aplicação tem relevância ao nível dos requisitos do utilizador e do contexto de aplicação. Assim, a concepção de uma plataforma no domínio das memórias pessoais conduz aos seguintes requisitos iniciais:

- Recuperação de informação multimédia em complemento à memória humana;
- Anotação baseada nas funcionalidades da memória humana;
- Registo de metadados no instante de captura;
- Dependência do cenário de aplicação;
- Sensibilidade à faixa etária e ao nível tecnológico do utilizador;
- Características dos dispositivos de visualização;

Estes requisitos são a base da metodologia utilizada para a concepção do sistema de recuperação de memória pessoais. A seguir é apresentado com mais detalhe cada um dos requisitos.

2.3.1 Recuperação e Memória Humana

Como referido anteriormente, as pessoas utilizam a memória episódica para se recordarem dos eventos do passado. Um sistema de recuperação de memórias que procure ser um complemento à memória humana terá de dispor, por um lado, de mecanismos que permitem que o utilizador use as pistas, obtidas utilizando a memória episódica, para pesquisar por memórias digitais do evento e por outro terá de, em tempo útil, mostrar ao utilizador a informação requerida, organizada de forma perceptível.

2.3.2 Anotação e Memória Humana

De modo a que o sistema de recuperação encontre a informação do passado utilizando as pistas solicitadas pela memória humana, as fotografias digitais têm de estar anotadas com a informação referente a estas pistas. A metodologia a seguir para realizar esta tarefa deverá ser automática ou semi-automática uma vez que é um dado adquirido por diversos estudos [Frohlich02, Wenyin01] que as pessoas em geral não executam esta tarefa manualmente. Para anotação automática, o sistema a desenvolver terá de utilizar o conteúdo das imagens ou a meta-informação anotada no instante de captura.

2.3.3 Metadados no Instante de Captura

Um dos problemas mais citado pela comunidade científica dedicada à recuperação de imagens baseada em conteúdo é a falha semântica, isto é, a diferença entre a correlação entre imagens identificada pelos humanos e a semelhança medida entre os respectivos vectores de características visuais. Uma das causas deste problema é atribuída à falha sensorial [Davis04], a diferença entre a realidade e a sua representação através de um conjunto de pixels. Uma solução é usar mais sensores, por exemplo, um receptor de GPS ou um microfone que em algumas máquinas fotográficas já aparecem incorporados, para capturar informação adicional no instante de captura e assim ajudar a compensar as falhas do sensor de imagem.

2.3.4 Cenários de Aplicação

Com a tecnologia actual, a partilha de experiências pessoais através de fotografia pode ser realizada em diversos locais (por exemplo, em locais públicos ou em ambientes domésticos) por isso, no processo de concepção de novas aplicações, é necessário incluir o cenário de utilização da interface. Em ambientes domésticos é importante respeitar as características familiares e de convívio social usuais nestes casos. Em locais públicos é necessário ter em conta as diversas distracções que o utilizador pode ter com o ambiente.

2.3.5 Tipos de Utilizadores

Actualmente a fotografia digital atingiu um grau de popularidade elevado que atravessa utilizadores de várias faixas etárias e com diversas capacidades para lidar com tecnologia. Para manter, com a fotografia digital, os hábitos de partilha de experiências utilizados com as fotos em papel os sistemas actuais requerem algum conhecimento tecnológico da parte do utilizador. O desenvolvimento de novas aplicações tem de ter em conta que as pessoas mais idosas, com mais experiências do passado, têm maior dificuldade em lidar com tecnologia. Por isso, novas aplicações deverão proporcionar interacções mais apropriadas. Por outro lado, há que contar também com as características dos utilizadores mais jovens.

2.3.6 Dispositivos de Visualização

Outro aspecto a ter em conta são os dispositivos de visualização de fotos pessoais. Desde o ecrã do computador pessoal, ao ecrã de dimensões reduzidas dos dispositivos móveis passando pelo ecrãs públicos, todos eles introduzem condicionantes na visualização de imagens. Os dispositivos móveis, por exemplo, por terem o ecrã pequeno, são os que mais condicionam a apresentação de fotos. Por outro lado, o ecrã do computador pessoal pode apresentar limitações se o número de participantes no reviver da experiência for elevado. Assim, as novas aplicações terão de contar com as características físicas do dispositivos de visualização e com a forma como são usados.

2.4 Objectivos

O trabalho proposto neste documento tem como objectivo estudar e propor soluções para a recuperação de informação em memórias pessoais constituídas por fotografias. Exploram-se as memórias pessoais utilizando interfaces que facilitem a interacção do utilizador em computador pessoal e em dispositivos móveis. Depois de analisado o domínio de aplicação e os requisitos necessários, tendo em conta as diversas áreas de investigação envolvidos, são definidos os objectivos principais do trabalho proposto nesta tese:

- Definição de um modelo conceptual de análise semântica multimodal em imagens;
- Desenvolvimento de um sistema de recuperação de informação multimédia baseado no modelo semântico proposto;
- Desenvolvimento de um sistema automático de anotação de imagens baseado no modelo de análise semântico

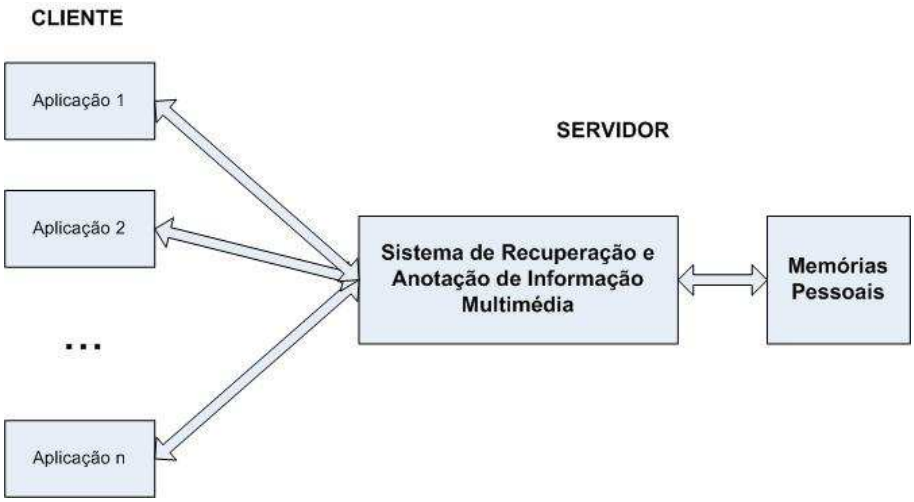


Figura 2.1: Arquitectura Cliente/Servidor.

- Desenvolvimento de algoritmos semi-automáticos para anotação semântica de imagens;
- Implementação de uma infra-estrutura para disponibilizar os sistemas de anotação e recuperação de imagens para diversas aplicações;
- Estudo e concepção de aplicações para recuperação de memórias pessoais baseadas em fotos para ambientes domésticos;
- Análise e desenvolvimento de aplicações para motivar o utilizador para a tarefa de anotação de imagens;
- Estudo e concepção de aplicações para explorar memórias pessoais de forma colaborativa em locais de interesse.

2.4.1 Infra-estrutura

É também objectivo desta tese criar uma infra-estrutura que ligue as diversas partes, nomeadamente, as aplicações e os métodos de recuperação e anotação que são comuns às diversas aplicações. Nesta infra-estrutura, temos um conjunto de aplicações independentes para dispositivos com diferentes recursos computacionais e que partilham os métodos de recuperação e anotação de informação e a colecção de fotos. Dada a natureza dos problemas a ter em conta no desenvolvimento de cada um destes blocos, a sua concepção é realizada de forma separada. A concepção das interfaces é feita pensando no cenário de aplicação, nos dispositivos e nos utilizadores. O sistema de recuperação e anotação depende da informação extraída das imagens e da forma como é utilizada. As arquitecturas cliente/servidor e *peer to peer* são duas soluções. A figura 2.1, mostra um exemplo de uma arquitectura típica para suportar estas funcionalidades.

2.4.2 Recuperação e Anotação de Imagens

O sistema de recuperação e anotação de informação multimédia é baseado em análise semântica multimodal. A análise semântica inclui como fontes de informação o conteúdo da imagem, a informação de áudio e os metadados anotados no instante de captura. O desenvolvimento e

avaliação deste sistema inclui o estudo e a implementação de métodos para treinar modelos semânticos automaticamente e também a investigação e a proposta de métodos semi-automáticos para anotação de informação multimédia. É também objectivo testar e avaliar os algoritmos desenvolvidos para anotação e recuperação de imagens nas diversas aplicações.

2.4.3 Aplicações

A recuperação de memórias pessoais é realizada em aplicações de três áreas diferentes: em ambientes domésticos, em aplicações para turismo ou actividades de lazer e em aplicações de entretenimento. Neste trabalho, estudam-se e desenvolvem-se novas interfaces para recuperação e anotação de imagens em memórias pessoais para diversos tipos de utilizadores e diversos contextos de aplicação. Inclui-se neste estudo o desenvolvimento de interfaces para computadores pessoais e dispositivos móveis, considerando a utilização de diversas técnicas de interacção.

2.5 Conceitos

O trabalho proposto cruza tópicos de várias áreas científicas que são utilizados no contexto do domínio das memórias pessoais. Para melhor enquadrar os objectivos desta tese, nesta secção são definidos alguns conceitos relevantes para o desenvolvimento do trabalho proposto. A tese envolve um sistema de recuperação e anotação de informação multimédia, utilizado em várias áreas de aplicação incluindo o turismo, entretenimento ou ambientes domésticos. O sistema de recuperação e anotação inclui o processo de extracção de informação visual e meta-informação dos dados multimédia e o desenvolvimento das aplicações inclui a metodologia de concepção da interface e a sua avaliação com utilizadores. Estas áreas definem um conjunto vasto de conceitos que são usados ao longo da tese.

As **memórias pessoais** são representadas por objectos físicos ou representações em formato electrónico que nos ligam por alguma razão a uma experiência do passado. Nesta tese, as memórias pessoais são representadas por fotografias digitais como elementos que registam um momento e que permitem mais tarde recordar esse momento. A captura do momento pode ser realizada de forma **passiva**, por exemplo com uma câmara ao pescoço, sendo possível registar um dia inteiro sem intervenção do utilizador, ou de forma **activa** em que o utilizador selecciona os momentos relevantes para capturar. As memórias obtidas pela captura passiva são mais ricas e podem ser utilizadas em mais aplicações, por exemplo para monitorizar problemas de saúde, mas representam maiores dificuldades para o sistema de recuperação. No caso da captura activa, as aplicações relacionadas com a partilha de momentos de lazer são exemplos de utilização das memórias obtidas.

As fotos capturadas nos momentos de lazer são utilizadas na maior parte das vezes para **partilhar** a experiência com amigos ou familiares. Tradicionalmente, as fotos em papel são partilhadas na presença das pessoas em **ambientes domésticos**. As fotos digitais, por um lado, são mais fáceis de partilhar à distância através da World Wide Web e mais fáceis de partilhar em qualquer lugar e em qualquer momento, por exemplo em **locais de interesse turístico**, através dos diversos dispositivos móveis que estão à disposição de todos. Porém, a fotografia digital introduziu dificuldades a alguns tipos de utilizadores, nomeadamente os utilizadores com maior dificuldade em lidar com a tecnologia. Para além disso, a actual tecnologia (computa-

dor pessoal) utilizada também alterou os métodos de partilha tradicionalmente utilizados na fotografia em papel. Por exemplo, as pessoas estavam habituadas a estarem viradas umas para as outras num ambiente social mais propício à partilha da experiência e partilha de ideias. Em redor do computador, é mais difícil criar este ambiente.

Os problemas introduzidos pela fotografia digital podem ser resolvidos através da utilização de novas tecnologias com novos paradigmas de **interacção pessoa máquina**. Este conceito diz respeito ao estudo da interacção entre as pessoas e o computador, geralmente efectuada através de uma **interface** que permite a comunicação entre ambos. Por exemplo, o utilizador pode utilizar objectos físicos através de uma **interface tangível**, pode utilizar movimentos dos braços ou da cabeça para uma câmara (**interface gestual**) ou falar para o computador de modo a interagir com a aplicação (**interface de voz**). Podem também ser utilizadas várias técnicas numa **interface multimodal**.

A aplicação de partilha de fotos envolve, em geral, a selecção de um subconjunto de fotos da colecção através de uma pesquisa na base de dados. Estas operações estão incluídas na área de investigação de **recuperação de informação multimédia** que diz respeito ao estudo do armazenamento, da organização, da representação e do acesso à informação multimédia. A representação e a organização devem providenciar uma recuperação dos elementos pretendidos pelo utilizador da base de dados.

Um documento de texto é representado por palavras mas para uma imagem a descrição do seu conteúdo por palavras não é uma tarefa fácil. Uma imagem é constituída por pixels com cor que representam texturas e formas de objectos. Um dos trabalhos pioneiros a utilizar a cor e a forma dos objectos para recuperar imagens foi publicado em [Hirata92] por Hirato e Kato que utilizaram o termo CBIR (Content Based Image Retrieval) para descrever as suas experiências. Este termo tem sido utilizado desde essa altura para designar o trabalho proposto em **recuperação de imagens baseada em conteúdo**.

A figura 2.2 apresenta a estrutura típica de um sistema CBIR. Um sistema de recuperação de imagens baseado em conteúdo envolve três entidades: um subsistema de extracção de **características visuais** (por exemplo, cor, textura ou forma) para obter uma representação de cada imagem, o bloco da **interrogação** onde o utilizador define a pesquisa que pretende fazer na base de dados e um **motor de busca** que apresenta ao utilizador as imagens relevantes para a sua pesquisa. A pesquisa pode ser definida utilizando uma imagem exemplo ou através de um esboço das imagens pretendidas no bloco da interrogação. Em ambos os casos é necessário extrair a mesma representação das imagens na base de dados. O motor de busca inclui um subsistema de **indexação** para criar a lista ordenada com os resultados.

Para alguns casos, o sistema CBIR descrito em cima não apresenta um bom desempenho, nomeadamente quando a **interrogação** é exigente do ponto de vista semântico. Nestas condições os resultados apresentados ao utilizador nem sempre são os esperados, por isso, alguns sistemas incluem a técnica de **retroacção de relevância** para tornar a pesquisa interactiva. Depois de apresentados os primeiros resultados, o utilizador é chamado a intervir para atribuir relevância a cada imagem. O sistema aprende com esta informação fornecida e procura apresentar melhores resultados. Este processo pode repetir-se várias vezes.

A necessidade de incluir o utilizador na pesquisa deriva das dificuldades que o sistema tem em representar, através de uma medida de semelhança entre vectores de características visuais, a correlação entre imagens identificada pelos humanos. Esta dificuldade do sistema

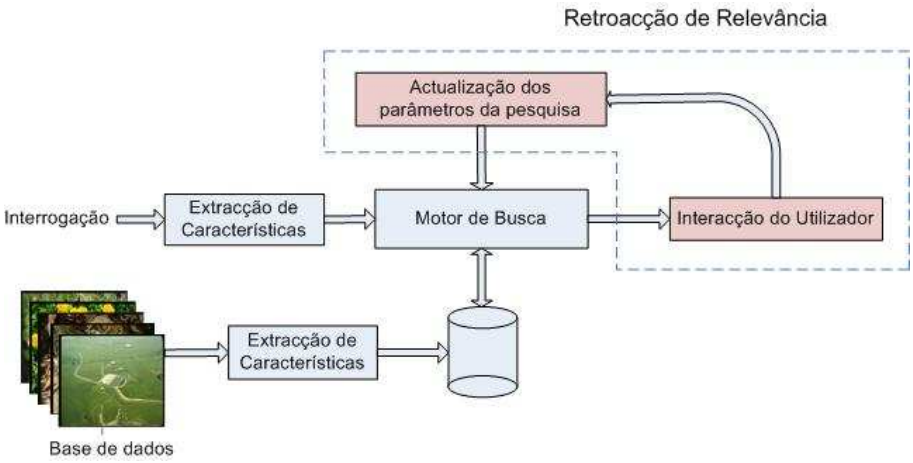


Figura 2.2: Estrutura geral de um sistema de recuperação de imagem baseada em conteúdo (CBIR).

CBIR é designada na literatura da área por **falha semântica**. Na verdade, quando se compara a análise que é feita de forma automática da realidade a partir de uma imagem com a percepção humana do mundo real são encontradas três falhas [Datta08]: a **sensorial**, a **semântica** e a **subjectiva**.

A **falha sensorial** está relacionada com a diferença entre o objecto real e a informação descrita numericamente a partir da imagem capturada da realidade. Uma das soluções consiste em introduzir mais sensores (por exemplo, um receptor de GPS ou um microfone) para conseguir capturar mais informação para ajudar a caracterizar melhor a cena no mundo real. A **falha semântica** como foi referido, diz respeito à diferença entre a informação numérica (vector de características) que representa o conteúdo da imagem e a percepção humana da imagem. A **falha subjectiva** é a diferença entre os conceitos detectados numa imagem e a percepção que cada indivíduo tem da imagem fruto da combinação do conjunto de conceitos detectados individualmente.

Como referido anteriormente, a técnica de retroacção de relevância é uma das formas para diminuir a falha semântica. Outra hipótese consiste em efectuar uma **análise semântica** da imagem, de forma a treinar automaticamente modelos para vários **conceitos semânticos** que representam um domínio semântico do qual se pode inferir palavras utilizando, por exemplo, uma **ontologia**.

A análise semântica é realizada com base nas características visuais extraídas na imagem e com base na informação obtida no instante de captura por sensores adicionais. Desta forma, a classificação da imagem é efectuada usando informação proveniente de diversos sensores e, por isso, é designada por análise semântica multimodal por se basear em **informação multimodal**.

Na bibliografia da área alguns autores chamam à informação adicional obtida no local de captura, **informação contextual**, por ser informação que ajuda a definir o ambiente em que a fotografia foi capturada. Contudo, também é possível definir o **contexto** utilizando o conteúdo da imagem.

Estes conceitos semânticos podem ser utilizados para recuperar imagens ou para anotar automaticamente imagens. Aliás, a **anotação** manual de palavras descrevendo o conteúdo da imagem é uma técnica alternativa para recuperar fotos, utilizando técnicas idênticas aos méto-

dos usados na pesquisa de documentos de texto. No entanto, esta técnica representa uma tarefa difícil de concretizar, para colecções com um número elevado de fotos, por ser dispendiosa em termos de tempo e aborrecida. Os **jogos de computador**, por serem divertidos, podem ser uma hipótese para tornar a anotação manual uma tarefa divertida.

2.6 Síntese

Este capítulo começa por abordar o domínio de aplicação do trabalho desenvolvido nesta tese. Foram identificados os problemas e motivadas as necessidades do domínio relativamente às aplicações de recuperação de informação multimédia. Desta forma, foram definidos os requisitos necessários para o desenvolvimento de aplicações de recuperação de fotografias. A terminar, foram também apresentados os objectivos da tese e definidos os conceitos mais importantes das áreas científicas abordadas nesta tese.

3

Trabalho Relacionado

Conteúdo

O capítulo descreve o trabalho relacionado em memórias pessoais, anotação e recuperação de imagens. São descritos os métodos de anotação e recuperação baseados em conteúdo e em metadados contextuais e são apresentadas as interfaces utilizadas para recuperar imagens em computador pessoal e em dispositivos móveis.

3.1 Introdução

O trabalho apresentado nesta dissertação tem como objectivo o desenvolvimento de um sistema para recordar e partilhar experiências do passado utilizando fotos. Assim, são necessárias aplicações que permitam encontrar numa colecção de fotos as imagens de uma dada experiência, com base em pistas que as pessoas normalmente utilizam para se lembrarem do passado. Estas aplicações devem disponibilizar interfaces para que o utilizador possa definir o que pretende através de diversas pistas e devem apresentar os resultados numa forma em que seja simples a sua interpretação.

Uma data, um local, um evento, um cenário, um objecto ou uma pessoa são algumas das pistas normalmente utilizadas para recordar o passado. Para recuperar informação com base nestas pistas as imagens têm de estar anotadas com esta informação. Assim, uma aplicação para recordar e partilhar imagens de experiências do passado é essencialmente constituída por três componentes: (1) um sistema de anotação, (2) um sistema de recuperação e (3) interfaces de acesso aos resultados em diversos locais e por diversos tipos de pessoas.

O capítulo é organizado com uma secção sobre memórias pessoais e uma secção de trabalho relacionado para cada uma das componentes referidas. A secção de memórias pessoais apresenta um resumo do trabalho desenvolvido para capturar e armazenar memórias pessoais, a secção de anotação apresenta as técnicas propostas para anotação desde as manuais até aos sistemas automáticos (inclui os sistemas baseados em conteúdo), a secção de recuperação completa a anterior mas apenas para os sistemas automáticos e semi-automáticos (porque são estes os sistemas que são propostos nesta tese) e a secção destinada às interfaces descreve o trabalho relacionado com as interfaces em computador pessoal e em dispositivos móveis.

3.2 Memórias Pessoais

Desde sempre que o ser humano gosta de guardar informação acerca de acontecimentos importantes da sua vida, para mais tarde recordar, trocar experiências, fazer histórias pessoais ou simplesmente para registar informação pessoal. As memórias pessoais tradicionais são representadas por artefactos físicos, incluindo jornais, diários, livros, álbuns de fotografia ou discos de vinil. Estes são guardados como algo que fica associado a uma experiência e que permite recordá-la. O avanço tecnológico dos últimos anos permitiu que as memórias pessoais possam ser constituídas por informação em formato digital, por exemplo, através de correio electrónico, ficheiros, páginas da Web, mensagens, músicas, imagens ou vídeos [Beagrie05]. Com a evolução da capacidade de armazenamento em formato digital, actualmente é possível guardar todos os aspectos da vida de um indivíduo em formato digital como foi demonstrado pelo projecto Microsoft MyLifeBits [Gemmell02, Gemmell04, Gemmell06].

MyLifeBits é uma realização da visão de Vannevar Bush proposta no artigo "As We May Think" [Bush45] em 1945. O artigo propõe o Memex (Memory Extender), um sistema constituído por um repositório de informação pessoal, incluindo notas pessoais, fotografias e esquemas, mecanizado de forma a que a consulta a qualquer documento seja bastante rápida e flexível e que possa ser um complemento à memória humana. No Memex propunham-se também câmaras montadas na cabeça dos indivíduos para gravar as experiências e microfilmes para armazenar estas experiências. Memex foi uma visão de Bush que influenciou os sistemas

que surgiram mais tarde.

Uma primeira implementação inspirada nestas ideias foi realizada no fim da década de 1960. Douglas Engelbart e a sua equipa do Augmentation Research Center, Stanford Research Institute em Menlo Park, USA, apresentaram uma demonstração dos conceitos com o sistema NLS (oN Line System). Mais tarde foi apresentada uma versão melhorada com o nome de Augment [Engelbart68]. Também inspirado nas ideias de Bush, Ted Nelson que inventou o termo "hypertext" [Nelson65], apresentou uma nova infra-estrutura computacional [Nelson99]. Nelson concebeu o projecto Xanadu com a intenção de armazenar um conjunto de documentos como um conjunto inter-relacionado, com ligações, e para fornecer acesso instantâneo a qualquer documento.

Na década de 80, foi utilizada pela primeira vez a designação PIM (Personal Information Management) [Lansdale88], que ainda hoje é usada para referir o trabalho relacionado com a visão de Bush. Este conceito refere-se à pratica e ao estudo das actividades realizadas pelas pessoas para adquirir, organizar e recuperar informação para usar no dia a dia [Jones05]. Desde a década de 80, várias aproximações em PIM têm sido propostas. Em [Jones05] são descritas algumas das propostas mais relevantes [Freeman96, Lansdale89, Dourish00, Huynh02, Dumais03] e alguns dos estudos [Malone83, Boardman04, Whittaker06] realizados que mais contribuíram para o desenvolvimento de aplicações PIM.

Malone contribuiu fazendo uma análise da forma como os utilizadores organizam a informação nas secretárias e Boardman e Sasse estudaram o comportamento dos utilizadores na utilização de aplicações específicas para um tipo de informação. Este estudo foi realizado para vários tipos, com o objectivo de propor estratégias para desenvolver aplicações com múltiplos tipos de informação. Também no estudo apresentado em [Whittaker06] sobre interfaces para gerir mensagens de correio electrónico, as conclusões vão no sentido do desenvolvimento de aplicações para a integração de vários tipos de informação.

Em relação às aplicações mais relevantes, Haystack [Huynh02] e Stuff I've Seen [Dumais03] são duas propostas que permitem gerir vários tipos de informação enquanto que Memoirs [Lansdale89] e LifeStreams [Freeman96] são aplicações focadas na organização da informação numa sequência de eventos, utilizando os atributos temporais. Dourish *et al.* [Dourish00] propõem o Placeless Documents, um sistema baseado em propriedades que facilitam a gestão de documentos em vez da hierarquia de directorias. Estes trabalhos foram desenvolvidos na perspectiva de que a informação pessoal já está armazenada em formato digital e os desafios estão no acesso e na organização.

Na visão de Bush surge também a ideia da captura audiovisual permanente para registar informação que possa ajudar a memória humana em diversas actividades [Gemmell06, Czerwinski06]. Um dos trabalhos pioneiros em captura passiva de imagens e vídeos, para registar experiências pessoais, surge na década de 1980 no MIT proposto por Steve Mann. Em 1996 foi publicada [Mann96] uma versão melhorada deste sistema (*wearable*), o Smart Clothing. Este sistema, para além da câmara de vídeo montada na cabeça, inclui também um microfone para gravar informação áudio, sensores para detectar a força e a velocidade nos sapatos, o ritmo cardíaco, a respiração e a resistência da pele.

Na década de 90, o Rank Xerox EuroPARC desenvolveu um projecto com memórias digitais compostas por pequenos vídeos [Lamming92]. Neste sistema, as câmaras de vídeo estão distribuídas pelas várias zonas de um edifício. Cada utilizador usa um Active Badge que indica

a sua presença num zona de um edifício e activa um sensor para capturar um pequeno vídeo nessa zona. Esta estratégia tem a vantagem de diminuir a carga que o utilizador transporta mas não permite que o utilizador veja o vídeo que é gravado no instante de captura.

Ambas as propostas capturam de forma passiva todos os movimentos do utilizador mas nem todos os movimentos têm o mesmo grau de relevância. Ainda na década de 90, foi proposta a StartleCam [Healey98], uma câmara para ser usada pelo utilizador com sensores de condutividade da pele com o objectivo de evitar a captura de todos os vídeos. Estes sensores permitem detectar a atenção do utilizador para algo e consequentemente activar a captura de imagens.

Para explorar a informação capturada de forma passiva, em [Clarkson02] foi apresentada uma proposta para encontrar padrões de vida em memórias digitais para ajudar a prever situações futuras. Mais recentemente, foram propostos vários trabalhos [Gemmell06, Hori03] para captura contínua de imagens ou vídeos através de dispositivos compostos por vários sensores e que também apresentam soluções para gerir a informação adquirida.

A Microsoft propôs o MyLifeBits [Gemmell02] para gerir a informação e a SenseCam [Gemmell04] para a captura de informação. Esta é constituída por uma câmara para adquirir imagens, um dispositivo para captura de GPS e sensores de luz, temperatura e infra-vermelhos para detectar indivíduos. A proposta da universidade de Tóquio [Hori03, Tancharoen05] usa um receptor GPS, um giroscópio, um acelerómetro e um sensor de onda cerebral que produziu resultados promissores [Aizawa01] na detecção de cenários em que a atenção do utilizador a algum elemento do cenário foi mais intensa. Este trabalho também inclui um sistema de recuperação de vídeo baseado em informação contextual, obtida através de diversos sensores do sistema (*wearable*) utilizado (por exemplo, informação temporal, de localização, de movimento e detecção de faces).

Outra perspectiva para capturar e analisar experiências humanas é apresentada em [Sumi04, Hagita03]. Este trabalho captura as interacções entre vários indivíduos que usam um sistema *wearable* (câmara de vídeo, microfone e sensores no corpo) num ambiente fechado com vários objectos, incluindo um robot, com sensores devidamente identificados para facilitar a construção posterior de histórias, sumários e qualquer outra tarefa para recuperar a experiência.

A captura de imagens e vídeos de experiências é importante para auxiliar a memória humana mas para tirar partido da informação capturada, também é importante desenvolver aplicações que possam recuperar esta informação de forma útil. As próximas secções apresentam trabalhos que propõem soluções para a recuperação e anotação de informação multimédia.

3.3 Anotação

Actualmente, a utilização de fotografias digitais para guardar experiências é uma actividade muito popular. No entanto, para recordar essas memórias, é necessário que as fotos estejam devidamente anotadas com informação que permita a um utilizador comum recuperá-las. Para mostrar o papel da anotação nas colecções pessoais compostas por imagens, em [Kustanowitz05] são apresentadas várias formas para visualizar a informação recuperada, só possíveis quando as imagens estão anotadas, por exemplo, recuperar e representar as faces da família, analisar a evolução dos filhos ao longo de vários anos ou construir histórias de viagens. Várias aproximações têm sido propostas com o objectivo de anotar imagens com palavras

chave que descrevem o seu conteúdo. Propomos as seguintes categorias para as classificar:

- **Manual** - utilizador atribui manualmente palavras chave a imagens;
- **Colaborativa** - vários utilizadores contribuem com anotações para as mesmas imagens;
- **Anotação com áudio** - anotação com palavras reconhecidas utilizando aplicações de reconhecimento automático de fala;
- **Anotação com aplicações de entretenimento** - anotação envolvida numa tarefa divertida;
- **Semi-automática** - parte do processo da anotação é automática e outra parte requer intervenção do utilizador;
- **Automática** - anotação através de análise automática da imagem.

Estas categorias não são exclusivas, por exemplo a anotação colaborativa pode ser também manual. Na tabela 3.1, são apresentadas algumas características das técnicas de anotação referidas anteriormente. A forma mais eficiente de anotar consiste na associação manual de palavras chave [Shneiderman00] a imagens. A principal desvantagem deste método está relacionada com o esforço humano necessário para anotar colecções com elevado número de imagens. Em geral, as pessoas não gostam de realizar esta tarefa [Frohlich02, Wenyin01]. Mais eficiente poderá ser a anotação manual obtida de forma colaborativa [Flickr04]. Isto pode acontecer porque vários utilizadores anotando as mesmas imagens adicionam um conjunto mais rico de anotações e porque o esforço humano necessário é menor (ver tabela 3.1). Mais fácil para o utilizador é a anotação obtida através de palavras reconhecidas automaticamente a partir de ficheiros de áudio [Rodden03]. O problema deste método são os erros de reconhecimento que podem frustrar o utilizador. Estes métodos requerem esforço humano mas são os mais eficientes.

Para realizar anotação automática de imagens, é necessário extrair características do conteúdo visual ou usar os metadados referentes aos parâmetros da câmara no instante de captura (por exemplo, instante de captura, informação de GPS ou distância ao sujeito) e que são anotados no cabeçalho do EXIF (Exchangeable Image File Format) [Exif98] do ficheiro JPEG (Joint Photographic Experts Group) da imagem. Estes metadados representam informação útil para definir o contexto em que a fotografia foi tirada. Para recuperar imagens com informação mais complexa (por exemplo, pessoas e edifícios) é necessário incluir características extraídas do conteúdo visual. Em sistemas que utilizam a informação visual, a interrogação é geralmente constituída por imagens, o que pode ser uma vantagem para o utilizador porque as imagens são mais descritivas do que as palavras chave. Contudo, a maior complexidade que as imagens representam para o sistema de recuperação é a principal desvantagem. A informação usada pelo sistema é constituída pelas características visuais automaticamente extraídas ou por modelos semânticos estimados a partir destas características (ver [Lew06, Datta08], dois artigos recentes que apresentam o estado da arte neste tópico). A anotação automática apresenta um desempenho mais fraco do que o processo manual (ver tabela 3.1) dado que algumas dificuldades permanecem sem solução, como é expresso no relatório do TRECVID 2006 [Over06]. Os métodos semi-automáticos procuram resolver algumas destas dificuldades incluindo o utilizador no processo [Wenyin01]. Estes métodos aumentam a eficiência da anotação mas também aumentam o esforço humano quando comparados com a anotação automática.

Outra opção, consiste em transformar a anotação de imagens numa tarefa divertida. Esta ideia foi proposta em [VonAhn04], tendo Luis von Ahn e Laura Dabbish convertido a anotação manual num jogo de computador para imagens da Web. O esforço humano é idêntico ao necessário na anotação manual, mas é utilizado de forma divertida mantendo-se o elevado desempenho. Nas próximas secções, são apresentadas as propostas mais relevantes de cada tipo de anotação.

	Características			
Técnicas de Anotação	Esforço Humano	Desempenho	Input	Informação
Manual	alto	alto	texto	palavras chave
Colaborativa	médio	alto	texto	palavras chave
Áudio	médio	médio	áudio	palavras chave
Semi-Automática	médio	médio	imagens	características visuais e contextuais
Entretenimento	baixo	alto	texto	palavras chave
Automática	baixo	baixo	imagens	características visuais e contextuais

Tabela 3.1: Comparação entre várias técnicas de anotação relativamente ao esforço humano necessário, desempenho, informação dada pelo utilizador e informação utilizada pelo sistema.

3.3.1 Manual

A anotação manual é actualmente a forma mais eficiente de associar imagens a palavras chave descrevendo o seu conteúdo. É também a mais utilizada pelas pessoas para organizarem as suas colecções pessoais. Várias aplicações comerciais incluindo iPhoto, Picasa, ACDSee e Adobe Photoshop Album e várias aplicações desenvolvidas no meio académico incluindo Photofinder, Fotofile ou PhotoMesa utilizam a anotação manual com o objectivo de melhor organizar e recuperar fotos em colecções pessoais.

Em geral, as aplicações referidas permitem categorizar uma ou mais imagens com palavras que foram inseridas pelo utilizador e algumas com categorias já definidas por omissão. A maior parte das interfaces guarda as anotações para futura utilização de forma a diminuir o esforço humano. O iPhoto permite associar teclas especiais a algumas categorias para facilitar a anotação, o Photoshop Album e o Fotofile [Kuchinsky99] permitem definir relações hierárquicas entre as categorias. O Picasa possibilita anotar a localização geográfica da foto através do Google Earth e o AcdSee e o Photofinder [Shneiderman00] utilizam a técnica do *drag & drop* para associar palavras a imagens. Em [Shneiderman00] foi proposta uma técnica para anotar imagens com o nome de Direct Annotation que permite que nomes de pessoas possam ser colocados directamente nas fotos. O utilizador escolhe um nome de uma lista e arrasta-o directamente para a foto perto da região onde se encontra a pessoa na foto a anotar. A lista de nomes é criada manualmente uma única vez. O WWMX (World Wide Media eXchange) [Toyama03] é outra aplicação que também utiliza a técnica do *drag & drop* mas para localizar imagens em mapas, de forma a associar a informação de localização às imagens.

Para medir o esforço humano é importante quantificar o tempo necessário para anotar uma imagem. Trabalho publicado recentemente [Yan07] propõe modelos para quantificar este tempo. Nesta proposta a anotação manual é dividida em dois tipos:

- **Navegação**, é escolhida uma palavra e depois o utilizador navega na base de dados anotando todas as imagens que podem ser descritas pela palavra (mais eficiente para pala-

vras muito frequentes);

- **Etiquetagem**, é seleccionada uma imagem e depois são atribuídas palavras a essa imagem que pertencem a um determinado vocabulário (apropriado para palavras menos frequentes).

Em [Yan07] são propostos dois modelos para quantificar o tempo dispendido nos dois tipos de anotação anteriores e um modelo híbrido que utiliza os dois tipos de anotação baseado na frequência de cada palavra.

Para além do esforço humano, existe também o problema da anotação não ser realizada por especialistas. Embora os utilizadores tenham melhor conhecimento das suas colecções pessoais, no caso de estas serem pesquisadas por outros é necessário uma uniformização na anotação. A aplicação BabelVision [Haase04] procura resolver este problema. O utilizador escreve uma palavra ou uma frase para anotar uma imagem. Depois o sistema retorna vários conceitos relacionados com a anotação e o utilizador escolhe os conceitos adequados. O sistema inclui um vocabulário estruturado que contém as relações entre termos (ontologias) que ajuda a melhorar as descrições feitas pelos utilizadores não especialistas. PhotoStuff [Halaschek05] é outra aplicação que usa ontologias para anotação de regiões de imagens para a Web semântica.

Apesar do esforço das várias aproximações referidas para diminuir a interacção humana na anotação de imagens, a anotação manual de imagens continua a ser uma tarefa fastidiosa para o utilizador porque tem a conotação de trabalho a realizar. Com o objectivo de atenuar este problema, várias estratégias têm sido propostas no sentido de obter as anotações manuais com menor esforço da parte do utilizador. As estratégias mais frequentes consistem em obter as anotações através de tarefas cujo o objectivo principal não é anotar imagens e através de tarefas colaborativas:

- Descrições em páginas na Web (Google Image Search e o Yahoo Image Search são dois exemplos de aplicações que utilizam esta técnica) - para imagens na Web é utilizado o texto que é inserido junto das imagens nos sites;

- Texto em correio electrónico - para fotos que são enviadas através de correio electrónico [Lieberman01];

- Partilha de fotos na Web - em aplicações de partilha é exigida alguma anotação para que os outros utilizadores possam ter acesso, por exemplo, o Flickr [Flickr04] ou o Riya [Riya05];

- Anotação colaborativa - aplicações comerciais (Fototagger [Fototagger06]) ou académicas [Walter07, Russell08] que permitem que vários utilizadores façam anotações sobre as mesmas imagens;

- Aplicações para contar ou construir histórias [Balabanovic00] - algumas aplicações para construir histórias utilizam fotografias em conjunto com descrições textuais.

- Aplicações de entretenimento - o utilizador associa texto a imagens quando está a realizar uma tarefa divertida [VonAhn04, VonAhn06, Tuulos07, Nicholas07], por exemplo, um jogo de computador.

Google Image Search é uma aplicação muito utilizada para pesquisar fotos na Internet. As palavras chave utilizadas para a procura são baseadas no nome do ficheiro, no texto da hiperligação que aponta para a imagem e no texto adjacente à imagem. Esta informação é escrita manualmente mas com o objectivo de construir uma página na Web.

O Flickr [Flickr04] é uma aplicação para a Web que foi desenvolvida com a intenção de proporcionar a partilha de fotos pessoais entre amigos que estão distantes ou entre desconhecidos. Esta aplicação encoraja os utilizadores a anotar as suas imagens, porque estas anotações são vistas como uma forma de facilitar o acesso. É uma aplicação colaborativa porque permite que qualquer pessoa faça anotações em imagens públicas. Outra aplicação colaborativa é o LabelMe [Russell08] que inclui uma ferramenta para segmentar objectos em imagens e realizar a respectiva anotação. Foi concebido com o objectivo de obter uma base de dados suficientemente genérica para testar e avaliar algoritmos de visão por computador, dado que as que existem não exploram todas as situações.

Noutro contexto, surgem as aplicações para contar histórias [Balabanovic00]. Algumas pessoas gostam de associar histórias às suas fotos e esta informação pode servir como metadados, isto é, as narrativas dos eventos capturados pelas fotos podem ser utilizadas como uma fonte para melhor organizar e anotar as fotos. A anotação torna-se no processo de contar histórias que é uma actividade mais atractiva.

Mais divertida pode ser a utilização de jogos de computador para gerar anotações. Esta ideia foi proposta por Luis von Ahn e Laura Dabbish através do jogo ESP GAME [VonAhn04]. Neste artigo o problema da anotação foi convertido num jogo de computador com base em conteúdos na Web. O jogo ESP (ver figura 3.1) é jogado por dois jogadores, escolhidos aleatoriamente utilizando a Web. Sempre que ambos os jogadores escrevam a mesma palavra para a mesma imagem ganham pontos, dado o facto que as palavras vindas de pessoas diferentes são mais robustas e descritivas do que anotadas apenas individualmente. Pares de palavras/imagens são consideradas válidas quando pelo menos um par de jogadores as tenha associado. O objectivo é tentar entrar na mente do outro e anotar as mesmas palavras na mesma imagem o mais rapidamente possível. Mais tarde, Luis von Ahn *et al.* propuseram outro jogo, o Peekaboom [VonAhn06], uma aplicação semelhante ao LabelMe mas envolvendo o utilizador num jogo. Também é jogado por dois jogadores na Web e o objectivo é anotar objectos da imagem. Um jogador escreve uma palavra relacionada com a imagem seleccionada que não é visível para o outro utilizador. Este tenta adivinhar a palavra escrita, visualizando apenas as partes da imagem que são seleccionadas pelo primeiro jogador. Quando acertam na mesma palavra ganham pontos. O Manhattan Story Mashup [Tuulos07] foi outro jogo proposto mas com o objectivo de contar histórias com imagens sobre Manhattan. Vários jogadores, usando a Web, telemóveis e ecrãs públicos, tiram fotografias e anotam-nas ao mesmo tempo com o objectivo de produzir histórias.

Os jogos referidos foram desenvolvidos principalmente para serem jogados entre jogadores que estão à distância. O PhotoPlay [Nicholas07] é um jogo com o mesmo objectivo mas para ser jogado num ecrã horizontal (*tabletop*) com todos os jogadores no mesmo local a utilizar um controlador de jogo. É um jogo semelhante aos jogos de palavras (Scrabble), que inclui uma grelha de letras e quatro fotos rotativas que são escolhidas aleatoriamente. Os jogadores seleccionam letras para formar palavras e anotar fotos ao mesmo tempo utilizando o controlador de jogo. Ao fim de cada jogada, os jogadores validam as anotação feitas pelos outros jogadores.

Figura 3.1: Jogo para anotação de imagens.

Esta avaliação promove a interacção social e resulta num processo de anotação mais eficaz.

3.3.2 Semi-Automática

Na anotação semi-automática uma parte da anotação é feita de forma automática e a restante com intervenção do utilizador. Neste contexto, porque estão directamente relacionados com o trabalho proposto nesta tese, são abordados dois tipos de sistemas: (1) os que fazem anotação através do mecanismo de retroacção de relevância [Zhou03] e (2) os que permitem anotação através de áudio. A retroacção de relevância é uma técnica utilizada para melhorar os resultados de uma pesquisa. Em geral, estes sistemas permitem que o utilizador possa validar os resultados obtidos através de uma interrogação. O sistema adiciona esta informação providenciada pelo utilizador e apresenta novos resultados para validação. O processo repete-se várias vezes. Quando o sistema inclui palavras chave na interrogação estas são associadas às imagens consideradas relevantes. Desta forma, é realizada a anotação semi-automática com palavras [Wenyin01]. No caso dos sistemas que utilizam áudio para anotação, o utilizador faz comentários relativos a uma foto usando um microfone e depois esta informação é transcrita para texto através de aplicações de reconhecimento automático de fala [Rodden03].

3.3.2.1 Retroacção de Relevância

O conceito de retroacção de relevância foi introduzido e estudado detalhadamente na área de recuperação de documentos de texto [Salton86, Yates99]. As primeiras técnicas de retroacção de relevância apresentadas com imagens foram inspiradas no trabalho de Rocchio [Rocchio71] proposto no contexto da pesquisa de documentos. Esta técnica consiste na reformulação da interrogação utilizando a informação providenciada pelo utilizador, isto é, desloca a interrogação na direcção das imagens relevantes no espaço de características através da equação,

$$Q_{t+1} = \alpha Q_t + \beta \left(\frac{1}{|R|} \sum_{x \in R} x \right) - \gamma \left(\frac{1}{|N|} \sum_{x \in N} x \right), \tag{3.1}$$

onde α, β e γ são constantes, N e R representam conjuntos de vectores de características de imagens relevantes e não relevantes.

Os trabalhos iniciais em recuperação de imagem envolveram a optimização de uma das

componentes do sistema CBIR, a reformulação da interrogação [Rui97,Ishikawa98,Porkaew99] ou a modificação da medida de semelhança [Rui98]. Os métodos de reformulação da interrogação dividem-se em métodos que deslocam a interrogação [Rui97, Ishikawa98] no espaço de características e em métodos de expansão da interrogação [Porkaew99].

As técnicas que deslocam a interrogação baseiam-se na técnica proposta por Rocchio, isto é, movimentam o vector que representa a interrogação na direcção das imagens relevantes no espaço de características. Em [Ishikawa98] e [Rui00] foram propostas aproximações para estimar a interrogação óptima, isto é, o ponto que minimiza a soma das distâncias a todas as imagens relevantes. A solução encontrada foi a média pesada das representações das imagens relevantes,

$$q_{opt} = \frac{\sum_{i=1}^{M} s_i x_i}{\sum_{i=1}^{M} s_i}, \tag{3.2}$$

onde M é o número de imagens relevantes, x_i denota o vector da i-ésima imagem e s_i representa um peso relacionado com a relevância atribuída pelo utilizador.

Em relação aos métodos que modificam a medida de semelhança, uma das aproximações pioneiras foi proposta por Rui et al. [Rui98]. No modelo hierárquico proposto para representar cada imagem, o peso de cada componente é inversamente proporcional ao desvio padrão da componente no conjunto das imagens relevantes. Assim, uma componente que apresente uma grande variação entre as imagens relevantes terá um peso baixo dado que não é um grande ajuda na discriminação entre imagens relevantes e não relevantes.

Mais recentemente, foram propostas várias técnicas de retroacção de relevância baseadas em técnicas de aprendizagem [Wood98,Vasconcelos00,Cox00,Tong01,Chen01,Zhang01,Jing03, Tieu04,Ferecatu05] que apresentam métodos sistemáticos para inferir o objectivo da pesquisa. Neste tipo de aproximação têm sido extensivamente utilizadas as SVMs (Support Vector Machines) [Tong01,Chen01,Zhang01,Ferecatu05]. Estas apresentam uma série de vantagens sobre os outros classificadores que as tornam adequadas para o problema da retroacção de relevância. Em geral, estes métodos estimam a função de densidade de probabilidade que melhor representa as imagens relevantes ou classificam as imagens da base de dados como relevantes ou não relevantes. Um dos aspectos mais importantes nestes sistemas de retroacção de relevância é o tempo necessário para o treino em cada iteração. Em [Yu07] é proposto uma aproximação que optimiza o tempo necessário para o processamento efectuando-o em memória. Um trabalho recente [Giacinto07], baseado no paradigma do vizinho mais próximo mas com maior capacidade de generalização, apresentou um desempenho comparável com as SVMs.

Outros trabalhos [Boldareva03, Dong03, Hoi04] têm sido propostos que acrescentam informação obtida entre várias pesquisas, sessões e utilizadores para melhorar a recuperação. Não utilizam palavras no processo e por isso não há uma anotação semântica explícita, embora esteja associada uma semântica ao conjunto de imagens indicadas como relevantes. Têm a vantagem de incluírem mais informação (outras sessões e outros utilizadores) para construírem os modelos semânticos.

Os trabalhos anteriores utilizam a técnica de retroacção para melhorar os resultados de uma pesquisa. Para utilizar esta técnica, para anotar imagens com palavras, é necessário proceder à associação de imagens com palavras durante o processo. Vários trabalhos [Chang98,Wenyin01, Zhou02, Lu03, Yang05, Jing05, Davis04] seguiram esta estratégia.

Uma proposta inicial propõe o treino semi-automático de modelos semânticos [Chang98]. Chang *et al.* apelidaram os modelos de SVT (Semantic Visual Templates). Estes associam um conjunto de interrogações (ícones) a um conceito semântico. As interrogações são obtidas através de um algoritmo semi-automático. O utilizador faz um esboço e define algumas características e o sistema gera automaticamente um conjunto de ícones. Depois o utilizador atribui relevância aos resultados obtidos por cada um dos ícones e são escolhidos os doze ícones que maximizam a cobertura e a precisão do sistema.

As propostas [Wenyin01, Zhou02, Lu03] utilizam a validação do utilizador sobre os resultados obtidos automaticamente para actualizar os pesos de uma rede semântica que associa imagens a conceitos. As interrogações são formadas por palavras, sendo por isso necessário que algumas imagens sejam anotadas previamente. Em [Yang05] todas as imagens são representadas por uma expressão semântica que pode ser obtida a partir de palavras previamente anotadas na imagem ou utilizando as características visuais.

Jing *et al.* [Jing05] propuseram um sistema com retroacção de relevância para conceitos semânticos (ver figura 3.2) que se aproxima de uma das propostas desta tese. O sistema permite a definição de interrogações com palavras chave e imagens exemplo. Previamente são treinados modelos semânticos usando uma SVM para cada palavra chave existente no sistema. É necessário ter algumas imagens anotadas manualmente com palavras para treinar os modelos, usando as características visuais. Depois, com estes modelos é feita a propagação das palavras chave pelas outras imagens. Quando a interrogação é definida por palavras, os primeiros resultados são obtidos através dos modelos das palavras. Quando a interrogação é formada por imagens exemplo, os primeiros resultados são obtidos pelas características visuais. Em ambas as situações é treinado um modelo durante a pesquisa com as imagens relevantes indicadas pelo utilizador durante o processo de validação de resultados. O modelo final será uma combinação linear do modelo previamente treinado com o modelo treinado durante a pesquisa.

Outra perspectiva, proposta por [Davis04], consiste em anotar imagens no instante de captura. Segundo Davis *et al.*, a anotação no instante de captura deverá ser mais precisa porque o conteúdo da imagem está presente no ambiente. Este sistema guarda no servidor automaticamente, a localização, o tempo, dados do utilizador e a foto tirada. A seguir, esta informação é comparada com outras fotos tiradas nas mesmas condições (por exemplo, no mesmo local ou pelo mesmo utilizador) e são propostas várias anotações que serão consideradas após validação do utilizador.

3.3.2.2 Reconhecimento de Palavras em Áudio

Como referido, outra alternativa para anotar fotos, consiste em utilizar as palavras reconhecidas em ficheiros de áudio, gravados com comentários descrevendo o conteúdo das imagens. Numa primeira fase, o utilizador é chamado a intervir para gravar os ficheiros de áudio e depois é utilizada uma aplicação de reconhecimento automático de fala, por isso, classificamos esta técnica como semi-automática.

O SHOEBOX [Rodden03] é uma das aplicações proposta para gerir memórias pessoais e que inclui o método de anotação baseado em palavras reconhecidas com áudio para gerar informação semântica relativa a imagens. Esta aplicação procura motivar as anotações na fase em que é feita a cópia das fotos para o computador pessoal. Assim, as anotações podem ser feitas com boas condições acústicas. O SHOEBOX também integra técnicas de processamento

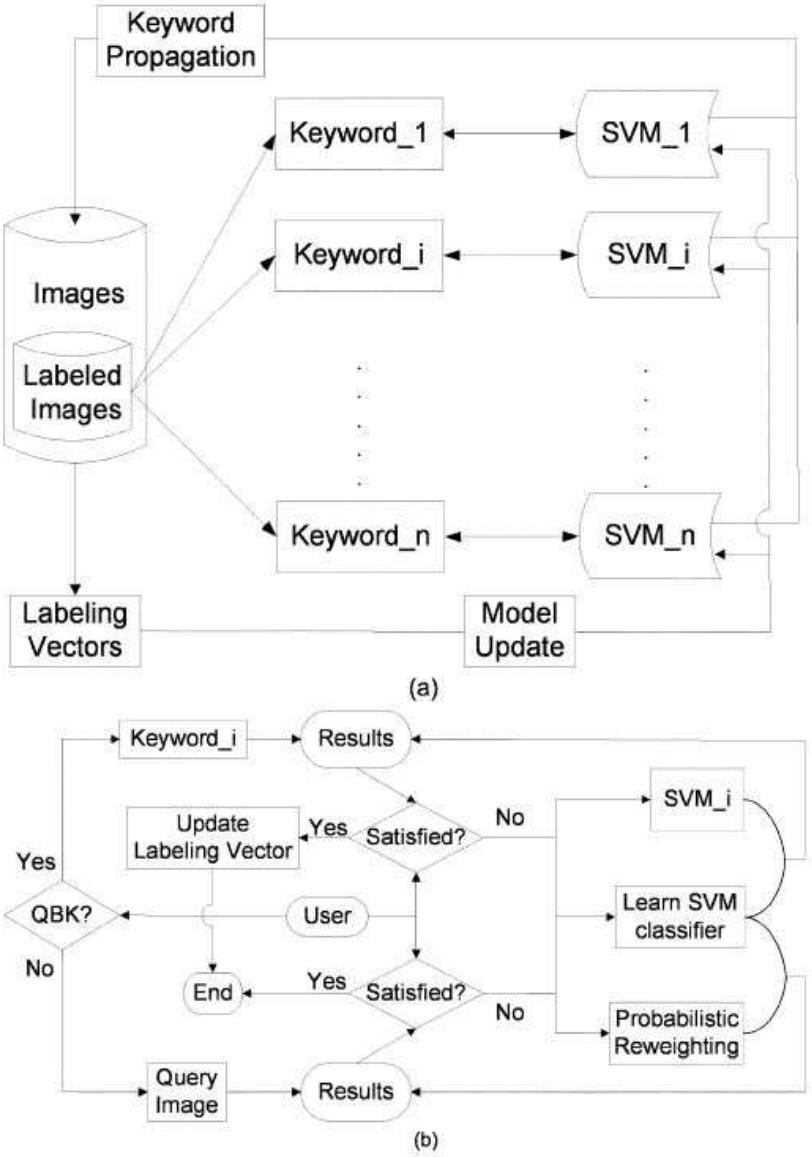


Figura 3.2: Arquitectura do sistema proposto em [Jing05]: a) Construção prévia do modelo das palavras chave; b) Interrogação utilizando palavras chave e imagens exemplo.

de imagem para indexar imagens. O ShowTell [Srihari00] é outro sistema que utiliza técnicas de anotação áudio mas para imagens médicas e de satélite. Esta anotação é usada em conjunto com métodos que usam informação visual para indexar imagens de forma idêntica ao SHOEBOX.

Pelo contrário, o sistema proposto em [Jiayi03] para indexar imagens, recupera as fotos através de anotações de áudio realizadas directamente na máquina fotográfica digital. Como o áudio não é obtido nas melhores condições, são propostas várias técnicas para atenuar os erros do reconhecimento. É utilizado um algoritmo de cancelamento de ruído e o método da expansão da interrogação para seleccionar as melhores palavras reconhecidas, de modo a atenuar os erros de reconhecimento.

Trabalhos recentes [Tokela08,Frohlich08] propõem aplicações para dispositivos móveis para contar histórias utilizando o microfone do dispositivo e as fotografias capturadas. Também nestas aproximações, o áudio é transformado em texto utilizando aplicações de reconhecimento automático de fala. Apesar do objectivo ser contar histórias, esta informação também providencia anotação de imagens.

3.3.3 Anotação Automática

Tal como referido, para efectuar anotação automática em imagens é necessário extrair informação do seu conteúdo (por exemplo, a cor, a textura ou a forma) ou utilizar informação contextual (por exemplo, a data, o local ou a distância ao sujeito) obtida no instante de captura e gravada no EXIF [Exif98] do ficheiro JPEG. Esta secção é dedicada aos sistemas CBIR, nomeadamente, à extracção de características de baixo nível e à estimação de modelos semânticos treinados utilizando a informação visual e contextual para classificação automática.

3.3.3.1 Características Visuais

O desempenho de um sistema CBIR (Content-Based Image Retrieval) depende essencialmente de duas componentes [Smeulders00,Lew06,Datta08]: (1) das características visuais e (2) da medida de semelhança entre imagens. A primeira componente tem como objectivo representar o conteúdo da imagem matematicamente e a segunda é utilizada para definir a lista ordenada (de acordo com a interrogação) de imagens que representam os resultados da pesquisa. Na maior parte das vezes as duas componentes estão relacionadas. Nesta secção, são apresentadas as características visuais mais relevantes e na secção 3.4.1 são descritas as medidas de semelhança utilizadas.

O termo CBIR foi utilizado pela primeira vez em [Hirata92], por Hirata e Kato, para descrever as suas experiências em recuperação de informação visual utilizando a cor e a forma dos objectos em imagens. Nos últimos anos, com o objectivo de permitir a descrição do conteúdo de imagens de uma forma normalizada, foram testadas e avaliadas um conjunto de características de cor, textura e forma para serem incluídas na norma MPEG-7 (Moving Picture Experts Group) [Pereira01]. Fora do contexto da norma MPEG-7 têm sido propostas características visuais que reflectem a cor, a textura, a forma dos objectos e pontos de interesse numa imagem.

No caso da cor, o espaço utilizado é relevante para o desempenho das características no que diz respeito à representação e comparação de imagens. O espaço de cor RGB é o mais utilizado para visualização em dispositivos, mas não é perceptualmente uniforme e, por isso, não é o mais indicado para comparar imagens de forma a satisfazer o utilizador. Outros espaços

utilizados são os de Munsell e os espaços Lab porque são uniformes do ponto de vista perceptual. Os espaços Lab foram definidos de forma a que a distância euclidiana entre duas imagens modele as diferenças de cor capturadas pela visão humana. O espaço HSV é utilizado devido às suas características de invariância à iluminação e à direcção da câmara e, por isso, é adequado para recuperar objectos. Em [Bimbo99] são descritos e comparados com mais detalhe os espaços de cor mais utilizados.

Histogramas de cor [Swain91], cores dominantes [Gong94], momentos de cor [Stricker95] e o correlograma de cor [Huang97] são as propostas mais representativas para descrever a cor em imagens. O histograma de cor é robusto à translação e rotação mas não inclui informação espacial. O correlograma de cor acrescenta esta informação ao histograma. Pelo contrário, os métodos das cores dominantes e dos momentos de cor reduzem a representação da imagem em relação ao histograma sem degradarem significativamente o seu desempenho e, por essa razão, são soluções mais equilibradas relativamente ao compromisso entre a quantidade de informação necessária para representar uma imagem e a eficiência.

Em relação à textura, a extracção de características estatísticas de níveis de cinzento foi um dos primeiros métodos utilizados para as classificar. Haralick [Haralick73] propôs a utilização de matrizes de co-ocorrências de níveis de cinzento para extrair estatísticas de segunda ordem de uma imagem. Outra aproximação inicial foi proposta por Tamura [Tamura78]. Nesta proposta, em cada pixel são calculadas seis características que, de acordo com estudos psicológicos, correspondem à percepção humana da textura. Mais recentemente foram propostas as características obtidas utilizando o filtro de Gabor [Manjunath96] e a transformada de Wavelet [Do02]. No caso do filtro de Gabor, a imagem é filtrada por um banco de filtros com seis orientações e quatro escalas e o objectivo é detectar texturas em várias direcções e a diferentes escalas. Utilizando a transformada de Wavelet também é aplicado um banco de filtros a várias escalas. Em ambos os casos, a imagem é representada pela média e desvio padrão de cada imagem filtrada. O banco de filtros de Gabor e a transformada de Wavelet são os mais utilizados [Datta08].

As características anteriores podem ser extraídas globalmente em toda a imagem ou localmente num conjunto de pixels. As características globais têm a vantagem de serem computacionalmente mais eficientes no que diz respeito à extracção e ao cálculo da semelhança entre imagens, no entanto, a extracção global não apresenta muita sensibilidade para detectar aspectos visuais localizados. Para capturar estes aspectos, é necessário fazer uma extracção local dividindo, por exemplo, a imagem em blocos iguais e procedendo à extracção individual em cada bloco. Contudo, esta divisão cria descontinuidades que são mais relevantes na presença de objectos. A alternativa é proceder à segmentação dos objectos na imagem. Dadas as dificuldades da segmentação [Sebe02, Cheng01] e uma vez que as aplicações de recuperação de imagens toleram alguns erros, vários métodos [Shi00, Carson02, Comaniciu02] têm sido propostos para fazer uma segmentação grosseira de modo a dividir a imagem em regiões mais coerentes. Os algoritmos k-médias, Normalized Cuts [Shi00] e Mean-Shift [Comaniciu02] são algumas das soluções utilizadas. Em [Carson02] foi proposta uma aproximação baseada no algoritmo EM (Expectation-Maximization) [Dempster77], que utiliza um modelo de mistura de Gaussianas para formar *blobs*, com o objectivo de definir interrogações locais e recuperar imagens com base nessas regiões. O algoritmo Mean-Shift têm sido utilizado com sucesso em várias aplicações (por exemplo, segmentação, filtragem ou seguimento de objectos em vídeo).

Ao contrário das características de cor e textura, as características de forma necessitam que as imagens sejam previamente segmentadas em regiões ou objectos. Contudo, uma vez que a tarefa da segmentação ainda apresenta algumas dificuldades [Cheng01], as características de forma têm sido pouco utilizadas. Na segmentação de objectos, as técnicas propostas dividem-se em algoritmos globais ou baseados em toda a região do objecto, por exemplo, momentos estatísticos [Flickner95, Sebe02] ou algoritmos locais baseados nos contornos do objecto [Mehrotra95, Bartolini05]. Sebe e Lew [Sebe02] usaram contornos activos para segmentação e momentos invariantes para medir a forma dos objectos. Recentemente, Bartolini *et al.* sugeriu a utilização da fase de Fourier e a distância DTW (Dynamic Time Warping). Esta aproximação apresentou resultados positivos na correspondência entre objectos.

Na última década tem havido um crescente interesse na extracção de características visuais baseadas em regiões [Datta08] mas, como já foi referido, existem ainda alguns problemas para resolver em relação à segmentação de imagens. Em alternativa, surgiram várias aproximações baseadas na detecção de pontos de interesse [Weber00, Sebe03, Mikolajczyk04, Lowe04] em objectos ou regiões. A ideia principal consiste na detecção de pontos que caracterizam uma região de interesse numa primeira fase e depois na extracção de descritores da região (por exemplo, de cor, de textura ou de forma). A maior parte destas aproximações detecta pontos e regiões que são invariantes à escala, rotação, translação e iluminação.

O detector de cantos e círculos de Forstner [Weber00], o detector de Harris-Laplace [Mikolajczyk04], a transformada de Wavelet [Sebe03] e filtros de diferenças de Gaussianas (DoG - Difference-of-Gaussian) [Lowe04] são alguns dos métodos utilizados para detectar pontos de interesse em imagens. Em [Mikolajczyk04] e [Sebe03] são apresentados resultados que mostram que no problema do reconhecimento de objectos, o detector de Harris-Laplace é a melhor solução porque detecta menos pontos do cenário que os restantes métodos. Para a aplicação de recuperação de imagens o método descrito em [Sebe03] apresentou melhores resultados, nomeadamente em imagens de cenários onde não existem objectos.

Na fase de extracção de características localizadas no pontos de interesse, várias técnicas têm sido utilizadas, incluindo as características de cor, textura e forma descritas anteriormente. O descritor SIFT (Scale-Invariant Feature Transform) [Lowe04] proposto por Lowe tem apresentado bons resultados em várias aplicações, incluindo no reconhecimento de objectos, na realidade aumentada ou em sistemas CBIR. O descritor é invariante à iluminação, rotação, escala e apresenta uma elevada capacidade de distinção. Numa região em torno do ponto, são calculados vários histogramas locais do gradiente obtido em todos os pontos da região. Estes histogramas definem o descritor.

Em [Bosch06] são comparados vários descritores na aplicação de classificação de imagens, tendo o descritor SIFT obtido o melhor desempenho. No problema da classificação, cada imagem é representada por um vector de ocorrências de palavras visuais (*bag of features*) pertencentes a um vocabulário que é obtido utilizando o algoritmo k-médias e descritores SIFT detectadas em toda a colecção de imagens. Esta estratégia tem sido utilizada com sucesso nos últimos anos [Nowak06] para representar imagens. Em [Nowak06] são apresentados vários trabalhos que seguem esta ideia e são discutidos e comparados vários aspectos, por exemplo os vários métodos de detectar pontos, os descritores e as diferentes estratégias para construir o vocabulário.

Mais informação sobre as técnicas utilizadas para extracção de características pode ser con-

sultada em [Smeulders00, Lew06, Datta08].

3.3.3.2 Anotação Semântica

Para algumas interrogações, a correlação entre imagens identificada pelos humanos é difícil de representar através de uma medida de semelhança entre vectores de características visuais. Este é o problema, designado por falha semântica [Smeulders00, Lew06, Datta08], que é apontado por toda a comunidade científica como a maior dificuldade dos sistemas CBIR. Uma das soluções para enfrentar esta dificuldade consiste no treino de modelos semânticos, representativos de palavras-chave, utilizando características visuais de baixo nível. Estes modelos permitem reduzir a falha semântica e também são uma solução que se aproxima das preferências dos utilizadores. Em geral, as pessoas gostam de definir interrogações semânticas utilizando descrições textuais para encontrar imagens relevantes [Rodden03], mas evitam a anotação manual [Frohlich02].

Várias aproximações têm sido propostas no sentido de automatizar a anotação de imagens com descrições textuais. Nesta secção, são apresentados e discutidos os trabalhos representativos das soluções propostas. Estas podem classificar-se em quatro categorias:

1. **Modelação conjunta de palavras com características visuais** - representam a informação visual de uma forma semelhante ao texto e modelam em conjunto esta informação através de métodos de computação linguística;

2. **Modelo probabilístico** - estimam um modelo probabilístico para cada conceito semântico;

3. **Classificação binária** - transformam o problema da anotação num problema de classificação;

4. **Modelação conjunta de informação visual e contextual** - incluem informação visual e contextual para anotar imagens automaticamente.

O objectivo dos métodos pertencentes à primeira categoria (**modelação conjunta de palavras com características visuais**) é modelar a informação visual em conjunto com o texto, utilizando métodos de computação linguística usados com sucesso no domínio do texto. A maioria destas aproximações constrói um vocabulário de características visuais, obtidas aplicando métodos de agrupamento às características detectadas em todas imagens (regiões) da base de dados. De seguida, são aplicadas técnicas utilizadas na recuperação de texto. O modelo de tradução automática [Duygulu02], o modelo generativo hierárquico [Barnard01], o modelo de relevância entre tipos de dados [Jeon03] e o método PLSA (Probabilistic Latent Semantic Analysis) [Monay03] são alguns dos métodos utilizados para modelar a informação textual e visual em conjunto.

Um dos trabalhos mais relevantes foi proposto por Duygulu *et al.* [Duygulu02]. Nesta proposta, a anotação de imagens é transformada num problema de alinhamento entre um vocabulário de *blobs* e um vocabulário de palavras. Para traduzir o conjunto de *blobs* em palavras são estimadas as probabilidades $p(a_{nj} = i)$ que maximizam a função de verosimilhança no conjunto de treino,

$$p(w|b) = \prod_{n=1}^{N} \prod_{j=1}^{M_n} \sum_{i=1}^{L_n} p(a_{nj} = i) t(w = w_{nj} | b = b_{ni}), \tag{3.3}$$

onde $p(a_{nj} = i)$ representa a probabilidade de na imagem n, um determinado *blob* b_i ser associado com uma palavra w_j, N específica o número de imagens, M_n e L_n denotam o número de palavras e de *blobs* associados com a imagem n e $t(w = w_{nj}|b = b_{ni})$ é a probabilidade de obter uma instância da palavra w dada uma instância do *blob* b. Estas probabilidades são estimadas utilizando o algoritmo EM. Este método consegue predizer um elevado número de palavras com precisão.

Outra forma de anotar automaticamente imagens com conceitos consiste em estimar um **modelo probabilístico** para cada palavra chave w pertencente ao vocabulário D. Uma palavra w é anotada numa imagem x de acordo com a função de densidade de probabilidade $p(w|x)$ calculada utilizando o teorema de Bayes,

$$p(w|x) = \frac{p(x|w)p(w)}{p(x)}, \tag{3.4}$$

onde $p(w)$, representa a função de densidade de probabilidade à *priori* da classe semântica w, $p(x)$ denota a função de densidade de probabilidade de uma imagem e $p(x|w)$ é a função de densidade de probabilidade de x condicionada à classe w.

O modelo de co-ocorrências [Mori99], técnicas não paramétricas [Yavlinsky05], modelos baseados na máxima entropia [Magalhaes07], métodos semi-supervisionados [Carneiro05] e modelos 2D de Markov não observáveis multi-resolução [Li03] são algumas técnicas propostas para modelar esta função de densidade condicionada.

A proposta de Mori *et al.* [Mori99] é uma das propostas pioneiras na anotação automática de palavras em imagens. Propõe a divisão da imagem em blocos rectangulares iguais, representados por características de cor e textura. Usa quantificação vectorial para encontrar os grupos representativos destes blocos nas imagens de treino e depois utiliza o modelo de co-ocorrências para anotar as imagens de teste.

Yavlinsky *et al.* [Yavlinsky05] mostraram que técnicas não paramétricas conseguem resultados idênticos aos outros métodos mais elaborados mas requerem muitos recursos computacionais. Magalhães e Rueger [Magalhaes07] utilizam o critério MDL (Minimum Description Length) para encontrar a representação óptima no espaço de características. Carneiro e Vasconcelos [Carneiro05] apresentaram os melhores resultados publicados com o conjunto de dados fornecidos por Duygulu *et al.* [Duygulu02].

Os mesmos autores de [Li03] propõem em [Li06] o ALIPR (ver figura 3.3), um sistema para a Web baseado em modelos 2D de Markov não observáveis multi-resolução e técnicas de optimização para anotação de imagens em tempo real. Esta técnica explora dependências entre as características visuais em blocos adjacentes e para várias resoluções da imagem.

Os métodos que se baseiam na **classificação binária** representam outra categoria de técnicas propostas para anotação automática de imagens. A ideia principal consiste em detectar a presença ou ausência de objectos e cenários em imagens através de uma colecção de classificadores binários. A maioria dos trabalhos propostos baseia-se em SVMs [Chapelle99, Li03a, Chang03, Yang06, Fan07, Papadopoulos07] porque as SVMs apresentam um conjunto de características apropriadas para este problema e têm apresentado os melhores resultados como classificadores binários [Muller01].

Dado um conjunto de treino $S_m = \{(x_i, y_i)_{i=1}^m\}$ onde $y_i \in \{-1, 1\}$ e x_i representa um vector de características de uma imagem, a fronteira de decisão entre as duas classes (por exemplo, interior/exterior) é obtida através de,

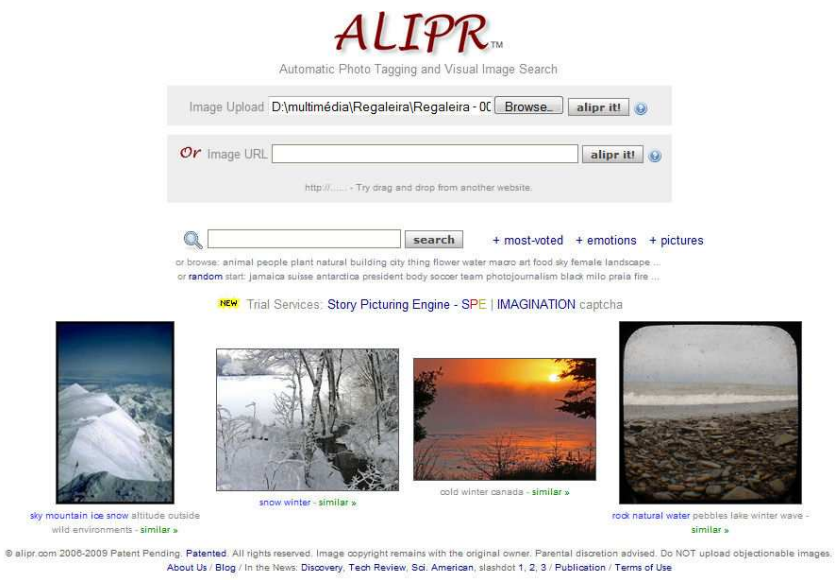


Figura 3.3: ALIPR - Aplicação na Web para anotação de novas imagens.

$$f(x) = \sum_{i=1}^{m} y_i c_i K(x_i, x) + \beta, \tag{3.5}$$

onde m é o número de imagens de treino, c_i e β são coeficientes obtidos por optimização matemática e $K(x_i, x)$ representa um *kernel*, por exemplo o *kernel* Gaussiano, $K(x_i, x) = e^{-\gamma \|x_i - x\|^2}$. Uma imagem x é classificada numa classe se $f(x) \geq th$, onde th é um limiar pré-definido. Se a condição não é satisfeita, a imagem pertence à outra classe.

Um dos primeiros trabalhos a utilizar as SVMs foi proposto por Chapelle *et al.* [Chapelle99]. Neste trabalho, as SVMs são utilizadas pela sua capacidade de generalização mesmo para espaços de características de elevada dimensão como o espaço definido por histogramas de cor, a representação escolhida para cada imagem. Trabalhos mais recentes [Yang06, Fan07, Papadopoulos07], utilizam as SVMs em aproximações que as integram com outras técnicas de aprendizagem. Em [Fan07], as SVMs são utilizadas num algoritmo hierárquico de *boosting* para anotação automática em vários níveis (por exemplo, areia, praia, sol) de imagens em conjunto com uma ontologia de conceitos. As SVMs também têm sido utilizadas para anotação de regiões em imagens [Yang06, Papadopoulos07]. Em [Yang06] é proposto um algoritmo para treinar as SVMs sobre a plataforma MIL (Multiple Instance Learning) [Maron98], uma variação da aprendizagem supervisionada, em que o objectivo é aprender um conceito a partir de dados (imagens) que são representados por uma sequência de instâncias (regiões), umas positivas outras negativas e cada uma descrita através de um conjunto de vectores.

Também foram propostos trabalhos que não utilizam SVMs. Um trabalho representativo e pioneiro foi proposto em [Vailaya01] com o objectivo de classificar imagens de férias, através de um algoritmo hierárquico composto por vários classificadores Bayesianos binários. Também utilizaram o critério MDL para escolher a dimensão óptima do dicionário de palavras de código obtido para cada classificador binário através de quantificação vectorial. Esta técnica apresentou bons resultados na detecção de cenários (por exemplo, cidade, campo, interior e exterior).

As técnicas descritas anteriormente para anotação automática utilizam apenas informação

visual. Recentemente, a combinação de **informação contextual** com a informação visual para estimar modelos semânticos de modo a anotar imagens tem recebido mais atenção da comunidade científica [Chang05, Luo06, Davis06, Choi08, Joshi08], nomeadamente no caso particular das colecções pessoais de fotos digitais. A informação contextual é a informação que ajuda a definir o contexto em que foi tirada a fotografia, por exemplo, a data, o local ou os parâmetros da câmara no instante em que foi capturada. Estes metadados são anotados automaticamente no EXIF [Exif98] do ficheiro JPEG da imagem. A informação de localização é obtida através de um receptor de GPS que vem incorporado em algumas máquinas fotográficas ou é colocado à *posteriori* utilizando software de anotação [Toyama03].

Alguns trabalhos [Naaman05, Hare05] utilizam apenas a informação contextual, nomeadamente, a data e a localização de captura. Em [Hare05] esta informação é utilizada para detectar os conceitos "Dia" ou "Noite" em imagens e para anotar palavras referentes às condições atmosféricas, recorrendo a páginas na Web que guardam a informação meteorológica ao longo do tempo e para diversos locais. Em [Naaman05] esta informação é usada para identificar pessoas em fotos pessoais. Dadas algumas imagens anotadas manualmente (conjunto de treino), este trabalho agrupa as imagens por evento e sugere anotações de indivíduos presentes em imagens com base em padrões de co-ocorrência de pessoas diferentes em locais diferentes.

Para combinar informação visual e contextual a maioria das propostas optam por utilizar aproximações probabilísticas [Chang05, Luo06, Joshi08, Choi08]. Luo *et al.* [Luo06] utilizam SVMs para anotação automática de alguns conceitos e depois utilizam os modelos de Markov não observáveis para integrar os resultados obtidos com as SVMs e as pistas contextuais obtidas no instante de captura. A data, o tempo de exposição, a distância ao sujeito e o nível do flash são os metadados utilizados nesta aproximação. Em [Chang05] é proposta uma aproximação probabilística próxima das redes bayesianas para combinar informação contextual com visual. Também integram uma ontologia semântica no modelo. O algoritmo foi avaliado no reconhecimento de pessoas e marcas. Em [Joshi08] são utilizados vectores de ocorrência (*bag of words*) mas para informação geográfica e são usadas imagens geo-referenciadas retiradas do Flickr [Flickr04] para detectar eventos e actividades. No caso das colecções de fotografias pessoais, a detecção de pessoas e a anotação de faces são tarefas relevantes dado que a maior parte das fotos contém pessoas. Em [Choi08] é proposto um algoritmo para anotação de faces que junta aos usuais algoritmos de reconhecimentos de faces a informação temporal.

Noutro contexto, Davis *et al.* [Davis06] propuseram uma aproximação para detecção automática de faces e locais em dispositivos móveis. Neste trabalho, a informação contextual é dividida em três partes: temporal, espacial e social. A informação social inclui o nome da pessoa que tirou a foto, as pessoas que estavam presentes e no caso da foto ser partilhada, as pessoas que a receberam. No artigo [Davis06] mostra-se que juntando esta informação à visual o desempenho do algoritmo proposto aumenta.

Em geral, estas propostas apresentam resultados que mostram que combinar informação contextual à informação visual é um tipo de aproximação que melhora desempenho da anotação semântica.

3.4 Recuperação de Informação Multimédia

As secções anteriores discutiram e apresentaram vários métodos para anotação de imagem, uma tarefa fundamental para os sistemas de recuperação de informação. O tipo de anotação disponível condiciona a forma como o utilizador pode formular a interrogação e por consequência como são recuperados os documentos. Em geral, esta recuperação consiste na construção de uma lista ordenada de documentos que satisfazem a interrogação. Para tal, é necessário medir a relevância que cada documento tem para a interrogação e utilizar um sistema de indexação que permita aceder aos documentos de forma eficiente. Em [Bimbo99] são apresentadas e discutidas as técnicas mais usadas em bases de dados multimédia. No trabalho proposto nesta tese são comparados todos os elementos de forma sequencial com a interrogação. Para o número de imagens utilizadas na colecção (aproximadamente 5000), não houve necessidade de desenvolver um sistema de indexação mais eficiente. Assim, nesta secção um sistema de recuperação é analisado como sendo essencialmente composto por um bloco de anotação onde são analisados os dados multimédia e por uma medida de semelhança entre os documentos. Nesta secção, são apresentadas as métricas mais utilizadas para comparar imagens e os sistemas CBIR mais representativos.

3.4.1 Medidas de Semelhança

Em geral, os sistemas que utilizam uma imagem exemplo como interrogação recorrem a dois tipos de métricas:

1. Distâncias entre dois vectores - imagem representada por um vector de características;

2. Distâncias entre vários vectores - imagem representada por um conjunto de vectores (um vector de características por cada região).

As métricas do tipo 1 foram as métricas utilizadas nos sistemas CBIR iniciais [Smeulders00] enquanto que as métricas do tipo 2 são mais recentes e apareceram em conjunto com as características visuais baseadas em regiões (ver secção 3.3.3.1). No primeiro caso, uma medida pioneira foi a intersecção entre histogramas de cor proposta por [Swain91],

$$S(q,d) = \frac{\sum_{i=1}^{M} min[H_i(q), H_i(d)]}{\sum_{i=1}^{M} H_i(d)}, \tag{3.6}$$

onde $H(q)$ e $H(d)$ representam dois histogramas (interrogação e um documento) com M intervalos de quantificação. Swain e Ballard mostraram que se o número de pixels fôr igual para todas as imagens esta medida de semelhança apresenta as mesmas propriedades da distância de Manhattan [Bimbo99].

A distância de Manhattan e a distância Euclideana foram as mais utilizadas nos sistemas propostos inicialmente [Veltkamp00]. Outra medida relevante e muito utilizada nos sistemas de pesquisa com retroacção de relevância [Zhou03], é a distância Euclideana pesada [Wu01] que permite atribuir pesos diferentes às características.

Nos casos em que a imagem é representada pela função de densidade de probabilidade do vector de características, a medida mais utilizada é a divergência de Kullback-Leibler. Esta estratégia é mais utilizada em métodos que utilizam a textura [Do02] para representar imagens.

Para os casos onde é utilizado o modelo do espaço vectorial [Yates99] para representar uma imagem tal como um documento [Salton86] na pesquisa de texto, a métrica mais utilizada para medir a correlação entre vectores é a medida de coseno. Nestes trabalhos, as imagens são representadas por vectores de ocorrências de termos visuais de um dicionário previamente construído (ver secção 3.3.3.1). Nesta representação está presente o conceito de uma imagem ser representada por um conjunto de regiões, no entanto, porque duas imagens continuam a ser comparadas com base em dois vectores, incluímos estes trabalhos nesta categoria.

As métricas do tipo 2 têm sido utilizadas mais recentemente para comparar imagens representadas por um conjunto de regiões [Datta08]. Cada região é tipicamente representada por um vector de características e por um peso. Considerando que uma imagem é representada por M regiões cada uma representada por um vector de características $z_i^{(m)}$ e pelo respectivo peso $p_i^{(m)}$, $I_m = \{(z_1^{(m)}, p_1^{(m)}), (z_2^{(m)}, p_2^{(m)}), ..., (z_{M_m}^{(m)}, p_{M_m}^{(m)})\}$, uma forma natural para comparar uma imagem I_q, que é utilizada como interrogação, com um documento da base de dados, I_d, consiste em combinar as distâncias entre os dois conjuntos de vectores,

$$D(q,d) = \sum_{i=1}^{M_q} \sum_{j=1}^{M_d} s_{i,j} d(z_i^{(q)}, z_j^{(d)}), \tag{3.7}$$

onde $s_{i,j}$ representam os pesos associados às distâncias entre $z_i^{(q)}$ e $z_j^{(d)}$. Vários trabalhos têm sido propostos utilizando esta medida [Datta08]. Estas aproximações diferem entre si na forma como estimam os pesos $s_{i,j}$. Minimizando a equação 3.7 e considerando as restrições $\sum_j s_{i,j} = p_i^{(q)}$, $\sum_i s_{i,j} = p_j^{(d)}$ e $s_{i,j} \geq 0$, temos a distância de Mallows [Mallows72].

A EMD (Earth Movers Distance) [Rubner00] é um das distâncias que apresentou os melhores resultados na comparação entre imagens representadas por características de cor das regiões detectadas [Datta08]. Esta técnica consiste em estimar a energia mínima necessária para "mover" uma componente do histograma de cor de uma imagem para uma componente da outra imagem. Para o caso em que $p_i^{(q)}$ e $p_j^{(d)}$ representam probabilidades, esta distância é equivalente à distância de Mallows.

Os sistemas que se baseiam em modelos semânticos, em geral, criam a lista ordenada das imagens relevantes para a interrogação através dos resultados obtidos pelo modelo que caracteriza o conceito. As diversas variantes dos modelos, apresentadas na secção 3.3.3.2, dividem-se em dois tipos de modelos: probabilístico ou baseado em classificadores. Os modelos probabilísticos [Mori99, Li03, Yavlinsky05] ordenam a lista de imagens de acordo com as probabilidades dos vectores de características das novas imagens serem geradas pelo modelo. Nas aproximações que modelam o problema da recuperação de informação como um problema de classificação [Chapelle99, Li03a, Chang03, Yang06, Fan07, Papadopoulos07], a lista ordenada com os resultados da pesquisa é obtida através do valor da saída do classificador. No caso dos trabalhos baseados em SVMs, imagens mais afastadas da fronteira de decisão são posicionadas no início da lista porque a confiança do classificador é maior para essas imagens. As imagens próximas da fronteira são colocadas no fim da lista.

3.4.2 Sistemas

Na última década e meia têm sido propostos muitos sistemas que utilizam a tecnologia CBIR [Veltkamp00,Lew06,Datta08]. Os sistemas propostos nos anos iniciais (até ao ano 2000) baseiam-

se nas características visuais extraídas automaticamente das imagens e no paradigma de interrogação por exemplo. Em [Veltkamp00] são apresentados alguns destes sistemas. Os sistemas mais recentes [Lew06, Datta08] caracterizam-se pela inclusão de modelos semânticos treinados utilizando o conteúdo da imagem. No caso dos sistemas propostos no domínio das memórias pessoais, é também incluída a informação contextual [Hare05].

Na tabela 3.2, são apresentados por ordem cronológica os sistemas mais representativos e algumas das suas características principais. Na tabela, a base de dados que é utilizada em cada sistema serve para indicar o domínio da proposta. O campo "Informação" descreve a informação utilizada para representar as imagens (características visuais, metadata contextual ou palavras). RR indica se o sistema inclui o mecanismo de retroacção de relevância e a última coluna indica se o sistema utiliza modelos semânticos treinados com características visuais para recuperar imagens.

Os sistemas apresentados podem organizar-se em três grupos. O primeiro é composto pelos trabalhos iniciais, Photobook [Sclaroff94], VisualSEEK [Smith96a], WBIIS [Wang97], NeTra [Ma97], NEC AMORE [Mukherjea99] e Blobworld [Carson02] que são baseados nas características visuais (interrogação através de imagens exemplo) e pelos sistemas QBIC [Flickner95], PicHunter [Cox96], MARS [Rui97], VIRAGE [Gupta97], Pictoseek [Gever00] que incluem retroacção de relevância. Os sistemas Photobook e QBIC são actualmente usados em aplicações reais. A tecnologia de reconhecimento de faces do Photobook tem sido utilizada pela Viisage Technology (software FaceID) que é utilizado em vários departamentos da polícia americana. O motor de busca do QBIC é utilizado na página Web do museu Hermitage em Saint Petersburg, Rússia, para pesquisa de arquivos de arte internacional. Os sistemas PicHunter e MARS são duas referências nas aproximações que utilizam retroacção de relevância. O Blobworld foi um dos primeiros sistemas a disponibilizar interrogações baseadas em regiões de imagens.

O segundo grupo inclui os trabalhos que foram desenvolvidos no domínio das fotos pessoais, Fotofile [Kuchinsky99], MiAlbum [Wenyin00], SHOEBOX [Rodden03], MediAssist [Hare05], Riya (Web) 2005 [Riya05] e ALIPR [Li06]. Estes trabalhos têm como objectivo disponibilizar ao utilizador um conjunto de funcionalidades para gerir fotos pessoais e por essa razão, na fase de concepção e desenvolvimento das aplicações foi dada maior importância à usabilidade das interfaces. O Fotofile foi um dos primeiros trabalhos a apresentar soluções para detecção e reconhecimento de faces em memórias pessoais. O SHOEBOX permite anotar imagens através de áudio e o MiAlbum utiliza retroacção de relevância para anotação semi-automática de imagens com palavras.

O terceiro grupo integra os trabalhos que treinam modelos semânticos para anotação automática de imagens utilizando informação visual. IFind (2000) [Zhang00], SIMPLIcity [Wang01], Cortina (2004) [Quack04], MediAssist, PARAgrab [Joshi06] e o ALIPR são as propostas apresentadas na tabela 3.2. O iFind e SIMPLIcity são duas referências relevantes no que diz respeito à utilização de modelos semânticos. O segundo é utilizado em duas aplicações comerciais na Web, o Airliners.Net [Airliners05] e o GlobalMemoryNet [MemoryNet06]. O MediAssist e o ALIPR são duas propostas que apresentam soluções para melhorar as dificuldades apresentadas pelo desempenho dos modelos semânticos. O MediAssist combina a informação visual com informação contextual (data e local) e o ALIPR sugere anotações automáticas e permite que estas sejam corrigidas pelo utilizador. O sistema aprende com a informação fornecida pelo utilizador e assim são treinados modelos que apresentam melhor desempenho.

Ano	Nome	Plataforma	Base de dados	Informação	Interrogação	RR	Semântica
1994	Photobook	Desktop	VisTex e Feret	Visual	Imagens	N	N
1995	QBIC	Desktop	-	Visual	Imagens e esboços	S	N
1996	VisualSEEK	Web	-	Visual	Esboços	N	N
1996	PicHunter	Desktop	Corel	Visual	Imagens	S	N
1997	WBIIS	Desktop	-	Visual	Imagens e esboços	N	N
1997	MARS	Desktop	Museu Fowler	Visual	Imagens e Esboços	S	N
1997	VIRAGE	Desktop	-	Visual	Imagens e Esboços	S	N
1997	NeTra	Web	Corel	Visual	Regiões de Imagens	N	N
1999	AMORE	Web	Web	Visual e textual	Imagens e texto	N	N
1999	Fotofile	Desktop	Pessoais	Visual e textual	Imagens e texto	N	N
2000	SHOEBOX	Desktop	Pessoais	Visual e Áudio	Imagens e texto	N	N
2000	MiAlbum	Desktop	Pessoais	Visual e textual	Imagens e texto	S	N
2000	Pictoseek	Web	-	Visual	Imagens e esboços	S	N
2000	iFind	Web	Corel	Visual e textual	Imagens e texto	S	S
2001	SIMPLIcity	Desktop	Corel	Visual	Imagens	N	S
2002	Blobworld	Web	Corel	Visual	Regiões de Imagens	N	N
2004	Cortina	Web	Web	Visual e textual	Imagens e texto	S	S
2005	MediAssist	Desktop e móvel	Pessoais	Visual e contextual	Imagens e localizações	N	S
2005	Riya	Web	Pessoais	Visual e textual	Texto	N	N
2006	PARAgrab	Web	Web	Visual e textual	Imagens e texto	N	S
2006	ALIPR	Web	Pessoais	Visual	Imagens e texto	S	S

Tabela 3.2: Sistemas CBIR

3.5 Interfaces

Em geral, um sistema de recuperação de imagens interpreta o pedido do utilizador através da interrogação e apresenta os resultados numa forma em que o utilizador tenha a percepção da relevância desses resultados em relação à interrogação. Na perspectiva do utilizador [Eakins04], a forma como se define a interrogação, a qualidade dos resultados da pesquisa e como são visualizados são os factores a ter em conta no desenvolvimento dos sistemas de recuperação de imagens. Nesta secção, são descritas e discutidas várias interfaces para definir interrogações e para visualização dos resultados. A forma como estes resultados são calculados foi abordada nas secções anteriores. A secção é dividida em duas partes, uma parte dedicada às aplicações em computador pessoal (inclui Web e mesas tangíveis) e outra para descrever sistemas em dispositivos móveis.

3.5.1 Computador Pessoal

A definição da interrogação e a visualização dos resultados obtidos são duas funções importantes na concepção de aplicações para recuperar imagens que podem estar relacionadas, por exemplo, uma interrogação de localização pode envolver a visualização de imagens em mapas ou uma interrogação temporal num calendário. Para analisar a informação que o utilizador pretende e como essa informação está relacionada com a interrogação, vários estudos têm sido realizados sobre interrogações feitas na Web [Jansen00, Cunningham04] ou em sistemas de recuperação de imagens da história americana [Choi02, Choi03]. Nestes estudos as interrogações analisadas foram realizadas através de palavras.

Em [Rodden03] foi publicado outro estudo no contexto das memórias pessoais e que per-

mite analisar interrogações baseadas em conteúdo. Fotos de eventos, fotos com pessoas, ou uma foto de algo específico são os objectivos das interrogações realizadas neste estudo. Em [Cunningham04] são analisadas interrogações feitas na Web para pesquisar imagens com objectos de arte, por exemplo, quadros e esculturas. As interrogações mais utilizadas são as que contêm metadados bibliográficos (por exemplo, artista, data do trabalho, título do trabalho, ou nacionalidade do autor) e as que podem ser realizadas pela tecnologia CBIR, por exemplo, esboços com objectos ou imagens exemplo.

De uma forma geral, os estudos referidos apontam como mais populares as interrogações com elevada incidência em termos que representam especificamente objectos ou pessoas, termos geográficos ou termos cronológicos. Estes estudos mostram também que para pesquisar informação multimédia nem sempre é fácil descrever com palavras o que se pretende. A maior parte das aplicações para recuperação de imagens que têm sido propostas coincidem com as conclusões dos estudos anteriores, no que diz respeito à definição da interrogação e visualização de resultados.

A maioria dos sistemas propostos fora do contexto das memórias pessoais e que utilizam a tecnologia CBIR concentram-se em três categorias para definir interrogações:

- Imagem exemplo [Veltkamp00];
- Esboço [Smith96, Rui97, Buijs99, Retrievr06, Matkovic04];
- Conceitos semânticos [Li06].

A interrogação através de imagem exemplo é a que requer menor interacção, no entanto, exige que o utilizador tenha conhecimento relativamente à funcionalidade do sistema, nomeadamente, em relação à forma como o sistema interpreta a imagem e representa a interrogação (vector de características).

A maioria dos sistemas propostos até ao ano 2000 [Veltkamp00] baseavam-se nesta técnica. A terceira categoria é a mais recente e permite ao utilizador procurar por imagens como se procura por documentos de texto, através de palavras chave. Em algumas situações, o utilizador pode ter dificuldades em exprimir por palavras as imagens que pretende [Jansen00]. Neste aspecto, a interrogação por esboço é a que permite ao utilizador maior liberdade para definir a interrogação mas é também a que exige mais interacção. O sistema MARS [Rui97] utiliza esta técnica, o utilizador define as cores a partir de uma tabela de cores e texturas a partir de um conjunto de padrões pré-definidos, para construir um esboço das imagens que pretende encontrar. No sistema VisualSeek [Smith96a], o utilizador faz um esboço com várias regiões e define as suas cores, dimensões e posições. Outra aplicação idêntica mas para a Web é o Retrievr [Retrievr06], que permite pesquisar imagens da Web através de esboço utilizando regiões de cor. Uma estratégia diferente é utilizada no ImageScape [Buijs99], o esboço das imagens pretendidas é efectuado através da técnica de *drag & drop* de ícones que representam conceitos para uma zona destinada à interrogação (ver figura 3.4a). Esta estratégia facilita o trabalho do utilizador porque não tem de desenhar, mas limita a interrogação ao conceitos disponíveis.

No sentido de tornar a interacção mais divertida e para abranger mais grupos de utilizadores incluindo crianças e idosos, em [Matkovic04] é proposta uma interface tangível para definir o esboço da interrogação (ver figura 3.4b). O utilizador coloca e organiza cubos coloridos numa mesa para definir um esboço de cores. Múltiplos utilizadores podem beneficiar desta estratégia para construir interrogações colaborativamente. A dificuldade em construir objectos mais

(a) (b)

Figura 3.4: Interrogação através de esboço: a) ImageScape: Interrogação através de esboço utilizando ícones; b) Pesquisa utilizando uma interface tangível [Matkovic04]

complexos com cubos é a principal desvantagem deste método, reconhecida pelos testes de usabilidade efectuados a esta técnica.

No contexto das colecções pessoais de fotografias, para além dos tipos de interrogação referidas anteriormente, os sistemas propostos apresentam também as interrogações geográficas [Toyama03] e temporais [Kang00, Flickr04, Toyama03]. Em [Toyama03] as interrogações geográficas podem ser definidas através de um clique numa região do mapa ou indicando a região por palavras. No PhotoFinder [Kang00] são utilizados operadores booleanos para combinar vários atributos (por exemplo, a data, a localização ou pessoas). O Flickr [Flickr04] permite definir num calendário interrogações temporais e interrogações através de palavras, tirando partido da anotação colaborativa dos utilizadores na Web. O utilizador pode escrever um conjunto de palavras ou seleccioná-las de um conjunto pré-definido com as palavras mais populares. Nesta interface, a popularidade dos conceitos é proporcional ao tamanho com que são apresentadas ao utilizador.

A maioria das propostas organizam espacialmente as imagens resultantes de uma pesquisa utilizando um dos três critérios: (1) relevância em relação à interrogação, (2) relação entre as imagens ou (3) as duas hipóteses em simultâneo. Estes cenários ocorrem independentemente do tipo de interrogação incluindo interrogação visual, com palavras chave, data ou localização. As propostas dedicadas à navegação variam entre a organização usando a informação temporal, a informação espacial ou a semelhança visual.

No caso das visualizações utilizando a informação temporal, para navegação ou pesquisa, as imagens têm sidos organizadas numa linha temporal (por exemplo, no Picasa e no Personal Digital Historian) ou na forma de um calendário [Rodden03, Flickr04, Graham02]. A maior dificuldade destas aproximações está relacionada com a apresentação simultânea de um elevado número de imagens, mantendo visível a estrutura temporal ou vice-versa. Uma das soluções consiste em agrupar as imagens em eventos e depois apresentar na estrutura temporal apenas as imagens representativas de cada grupo [Graham02]. Graham *et al.* propõem um algoritmo hierárquico para agrupar as imagens temporalmente e criar um sumário da colecção. A figura 3.5 mostra a interface desenvolvida para visualização de fotos organizadas na forma de um calendário. Do lado esquerdo da interface estão os grupos, organizados hierarquicamente de

Figura 3.5: Navegação numa colecção de imagens organizadas hierarquicamente pela data.

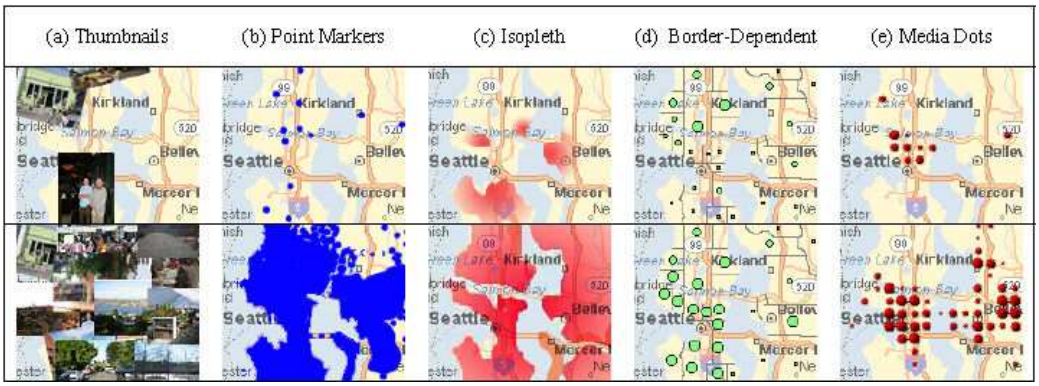


Figura 3.6: WWMX - Vários métodos para apresentar fotos em mapas.

forma semelhante a uma estrutura de directorias.

O WWMX [Toyama03] é uma aplicação para visualização de imagens em mapas. Esta interface apresenta vários formas de representar imagens em mapas (ver figura 3.6). A visualização espacial também enfrenta a dificuldade da visualização temporal. Neste caso, ou temos uma boa resolução para a localização ou para muitas imagens. Na figura 3.6 são apresentados os vários modos de visualização da aplicação WWMX. A visualização em tamanho reduzido é a única que permite ver as imagens mas limita a resolução da localização. Outras técnicas incluem a utilização de pontos ou círculos com diâmetro proporcional ao número de imagens na região.

A apresentação das imagens por ordem de relevância em relação à interrogação é o modo de visualização de resultados mais comum [Veltkamp00]. Todos os sistemas descritos na tabela 3.2 recorrem a esta solução. Em [Nakazato01] é proposto um método de visualização 3D que permite apresentar as imagens ordenadas de acordo com a semelhança de cor, textura e estrutura dos contornos (ver figura 3.7). O utilizador tem mais informação para visualizar mas em contrapartida a interacção é mais complexa.

Na figura 3.8, é apresentada a interface IGroup [Wang07], como exemplo 2D ilustrativo de visualização de imagens ordenadas pela relevância em relação à interrogação. Esta aplicação também agrupa por contexto as imagens da lista de resultados (coluna no lado esquerdo da figura 3.8). Estes grupos são representados por algumas imagens que caracterizam cada grupo. Desta forma, o utilizador pode decidir em que contexto formulou a interrogação. A aplica-

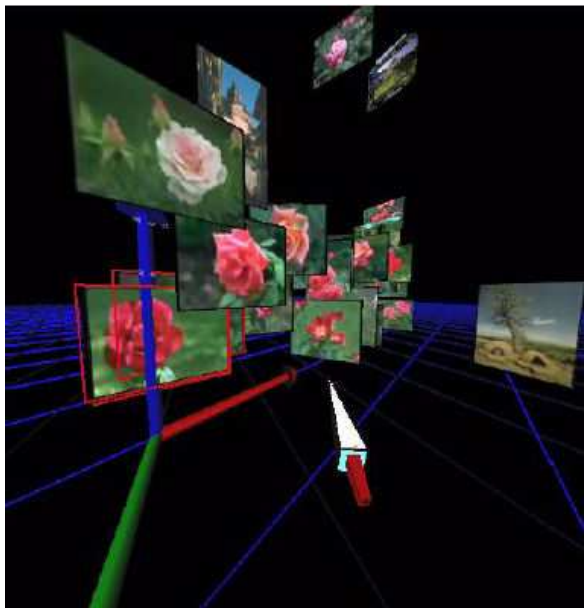


Figura 3.7: MARS 3D - Interface para visualização de imagens num espaço tridimensional.

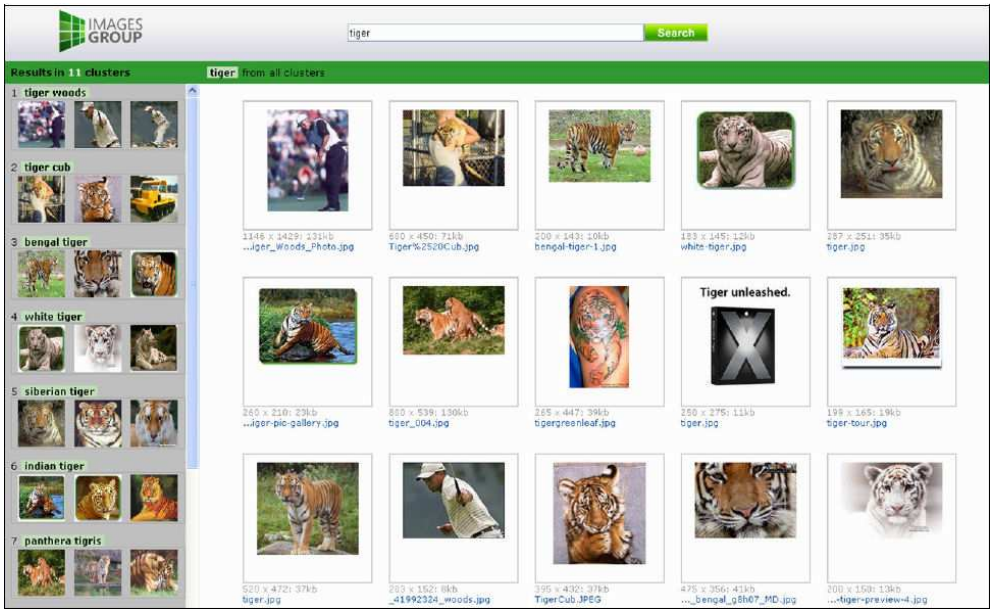


Figura 3.8: Interface da aplicação IGroup.

Figura 3.9: Visualização das imagens com base na cor utilizando o EMD e o MDS.

ção IGroup também organiza os resultados relacionando as imagens por algum critério (por exemplo, contextual, directorias ou semelhança visual), por isso, o IGroup é um exemplo da organização dos resultados de uma interrogação, utilizando as duas estratégias, relevância em relação à interrogação e relação entre as imagens resultantes.

Um estudo realizado em [Rodden01] mostra a utilidade na pesquisa da organização por semelhança dos resultados. O objectivo da organização dos resultados, relacionando as imagens mais relevantes por determinado critério, é fornecer mais informação visual ao utilizador. Por exemplo, no caso do IGroup esta estratégia permite eliminar as ambiguidades. Na figura 3.8, a interrogação apresentada pelo utilizador é "tiger" e os resultados apresentados revelam as dúvidas do sistema que o utilizador pode eliminar, com um clique no respectivo grupo. Os critérios utilizados para agrupar os resultados variam desde o contexto [Wang07], semelhança visual [Rubner97, Urban03, Heesch04, Moghaddam04] e directorias do sistema de ficheiros [Bederson01].

Uma das aproximações relevantes na utilização de informação visual foi proposta em [Rubner97]. Este trabalho utiliza o algoritmo EMD para medir a semelhança de cor entre duas imagens e o MDS (Multi-Dimensional Scaling) para mapear as imagens num espaço bidimensional. Desta forma, é possível visualizar as imagens no ecrã de acordo com a sua semelhança de cor (ver figura 3.9). Este método é mais adequado para navegar na colecção porque permite uma visualização global, como é representado na figura 3.9. Em relação à visualização das imagens resultantes de uma pesquisa, apenas as mais relevantes são apresentadas mas mantendo a mesma posição espacial, perdendo-se desta forma a noção de relevância para a interrogação. Para resolver este problema, Heesch e Rueger [Heesch04] propõem a organização dos resultados em "olho de peixe" (ver figura 3.10). As imagens mais relevantes são apresentadas no centro do ecrã e ocupando mais espaço, enquanto as menos relevantes são apresentadas na periferia do ecrã e com dimensões reduzidas. Este trabalho utiliza o algoritmo do vizinho mais próximo para k características visuais (NN^k).

Os trabalhos [Urban03, Moghaddam04] também utilizam várias características visuais mas incluem a interrogação na forma como apresentam as imagens no ecrã. Em [Urban03] é usado

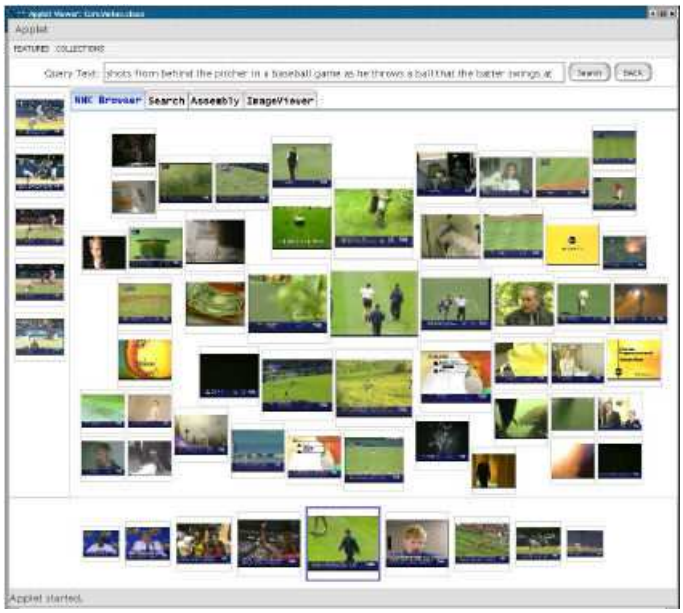


Figura 3.10: Visualização dos resultados de uma interrogação organizados de acordo com o método NN^k e apresentados no modo "olho de peixe".

o modelo ostensivo [Campbell00] para fazer pesquisas adaptativas à base de dados. Dada a interrogação inicial, são apresentadas as imagens mais relevantes ao utilizador que escolhe uma delas de acordo com as suas preferências. A interrogação passa a ser representada por duas imagens e novas imagens são apresentadas (ver figura 3.11). O processo repete-se e as imagens escolhidas permanecem no ecrã ligadas entre si. No fim, é possível ver o conjunto de imagens que definiram a interrogação e as imagens mais próximas da interrogação.

Outra forma de visualizar as imagens no ecrã é proposta por [Bederson01] e aplicada no sistema PhotoMesa. O objectivo não é apresentar os resultados de uma interrogação mas mostrar o maior número possível de imagens no ecrã, numa resolução baixa, e depois utilizar operações de *zoom* de acordo com o interesse do utilizador (ver figura 3.12). PhotoMesa é uma aplicação para navegar em directorias de imagens que utiliza o algoritmo Quantum Treemap para visualizar conjuntos de imagens organizadas no ecrã. Este algoritmo preenche o espaço 2D do ecrã com grupos de imagens, dividindo recursivamente o espaço em rectângulos com áreas proporcionais à dimensão das imagens. O número de rectângulos e a relação entre as dimensões dos rectângulos são parâmetros do algoritmo.

Outra questão que se coloca na visualização de memórias pessoais está relacionada com a quebra dos hábitos que as pessoas tinham, quando mostravam à família vários albums de fotografia em papel na sala em ambiente de convívio familiar [Frohlich02]. Vários trabalhos [Moghaddam04,Apted06] apresentam soluções para partilha de fotos em ecrãs horizontais, utilizando uma tecnologia semelhante ao Surface da Microsoft [Surface07], no sentido de tornar a partilha de foto digital mais próxima da fotografia em papel. Moghaddam *et al.* [Moghaddam04] propõem o projecto PDH (Personal Digital Historian) para partilha de histórias pessoais com fotos utilizando uma ecrã horizontal circular multitoque. Este protótipo foi desenhado para facilitar a conversação entre pessoas que partilham uma experiência. Permite fazer pesquisas baseadas no modelo de interrogação dos quatro "Ws" ("who", "what", "when", "where"). Na pesquisa "who", são apresentadas imagens de todas as pessoas presentes na base de dados (ver figura 3.13a), e o utilizador selecciona uma para encontrar imagens dessa pessoa. A

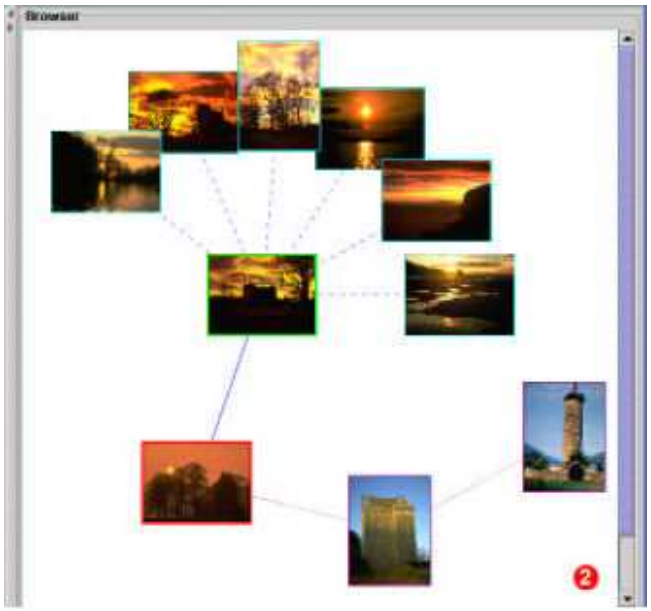


Figura 3.11: Interface com a interrogação e os resultados utilizando o modelo ostensivo adaptativo.

Figura 3.12: PhotoMesa - visualização de imagens utilizando o algoritmo Treemap para organizar o espaço do ecrã.

(a) (b)

Figura 3.13: Aplicações em ecrãs horizontais: a) PHD, interrogação "Who"; b) SharePic.

pesquisa "what" é igual mas para eventos. Na pesquisa de localização, as imagens são representadas num mapa e o utilizador clica no mapa para escolher um região a pesquisar. Na pesquisa "when", as imagens são apresentadas numa linha temporal. Um trabalho idêntico é proposto em [Apted06] mas para ser utilizado por idosos (ver figura 3.13b).

3.5.2 Dispositivo Móvel

O desenvolvimento da tecnologia no últimos anos tem contribuído para o enriquecimento das capacidades dos dispositivos móveis. Ao mesmo tempo, novas aplicações têm surgido utilizando estas potencialidades. No caso da fotografia, várias aplicações têm sido propostas para recuperação de fotos, nomeadamente, interfaces para partilha local ou partilha entre utilizadores e interfaces de pesquisa para vários tipos de aplicação (por exemplo, realidade aumentada, partilha de experiências ou guias de turismo).

O Pocket ACDSee [Acdsee01] é uma interface de navegação que permite a visualização de imagens de uma directoria. Também inclui a possibilidade de anotar imagens manualmente e com áudio. Mais elaborado é o Pocket PhotoMesa [Khella04] que, à semelhança da versão para computador pessoal, inclui uma interface de visualização de imagens baseada em operações de *zoom* e no algoritmo Quantum Treemaps. Este método permite navegar com base no nome de várias directorias de imagens. Em [Harada04] foi proposta outra interface para visualizar imagens que, para além de permitir a navegação em directorias, também inclui uma interface para visualização de fotos organizadas hierarquicamente de acordo com a informação temporal.

Outra interface, na linha das anteriores, foi proposta em [Cho07], diferenciando-se na forma como o utilizador interage com a aplicação. Esta interface é baseada em movimentos do dispositivo móvel. Na figura 3.14, são apresentadas duas formas de navegação numa colecção de fotos, uma local (figura 3.14a) e outra global (figura 3.14b). O modo de navegação local permite visualizar duas fotos com mais detalhe. Se houver um movimento do dispositivo para a direita como indicado na figura, o ângulo de inclinação do dispositivo indica o número de fotos que são deslocadas no ecrã. No modo de visualização global, este ângulo indica quantas fotos é preciso deslocar o cursor para a direita.

As interfaces anteriores para navegação em colecções de imagens foram propostas para permitir a visualização e partilha, em qualquer local e circunstância, e não tiram partido das capacidades de comunicação disponíveis no dispositivo. O Zurfer [Hwang07], por exemplo,

(a) (b)

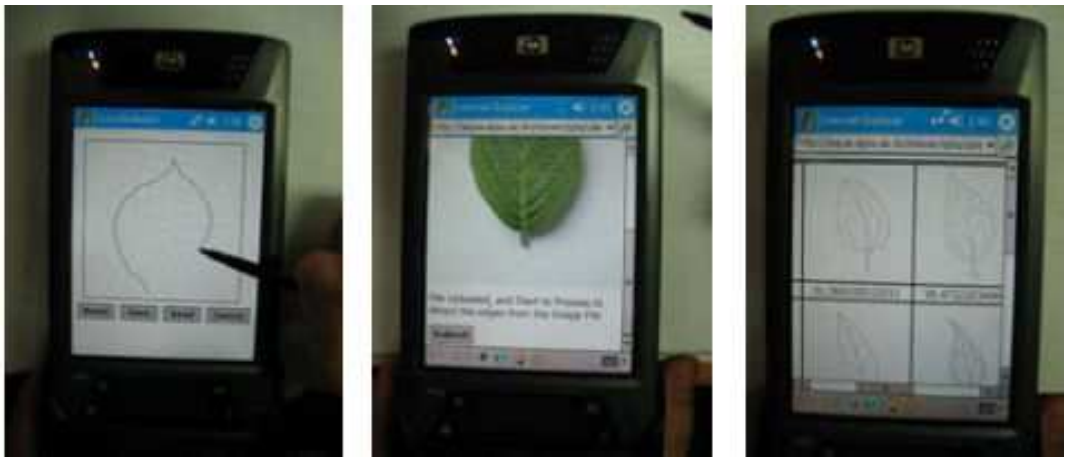
Figura 3.14: Visualização em interface baseada em movimentos do dispositivo móvel: a) local; b) global.

já inclui esta característica, permitindo visualizar fotos do Flickr. Com esta funcionalidade, o Zurfer permite a partilha de fotos entre utilizadores e a visualização em diversos contextos, por exemplo, social para ver fotos dos utilizadores amigos, espacial para ver fotos partilhadas por outros do local onde o utilizador se encontra ou por tópicos interessantes (utilizando as anotações do Flickr). Para partilha de fotos foram propostas outras interfaces [Sarvas04, Clawson08]. O Mobiphos [Clawson08] é uma aplicação recente que permite que um grupo de utilizadores possa capturar, partilhar e explorar em conjunto um local.

Em relação ao sistemas de pesquisa, aplicações de realidade aumentada [Noda02, Yu04, Sonobe04, Yeh05, Kim05], para guiar pessoas [Fockler05, Beeharee06, Chevallet07] ou simplesmente para visualizar e partilhar memórias com amigos [Fan05, Gurrin05, Anguera08] são as categorias mais representativas das aplicações que têm sido propostas. Exceptuando o PhoneGuide [Fockler05] e o MAMI [Anguera08], todas as restantes aplicações são baseadas na arquitectura cliente/servidor para permitir que as funções mais exigentes do ponto de vista computacional possam ser executadas no servidor.

Em geral, as aplicações de realidade aumentada com modelos de informação pré-definidos são baseadas num sistema de recuperação de imagens. A interrogação é uma imagem capturada pelo dispositivo móvel quando o utilizador está a realizar uma determinada actividade e necessita de informação adicional. Esta imagem é enviada para o servidor para ser processada e para indexar informação para enviar para o cliente. Esta estratégia tem sido utilizada para aumentar a informação disponível no instante de captura em várias aplicações, por exemplo, saber mais sobre flores [Noda02], peixes [Sonobe04], folhas de plantas [Kim05] ou pirilampos em aulas de ecologia [Yu04]. O IDeixis [Yeh05] foi proposto para procurar informação adicional sobre o local onde a foto foi capturada. Através de uma foto de um objecto característico do local são pesquisadas páginas na Web com imagens semelhantes. A informação adjacente e as fotos semelhantes são enviadas para o utilizador. As aplicações anteriores utilizam interrogações através de imagem exemplo (capturada), enquanto o mClover [Kim05] também permite fazer interrogações por esboço (ver figura 3.15), neste caso, esboços de folhas de plantas.

As aplicações para guiar pessoas em museus, cidades ou em outros pontos de interesse (para mais detalhes sobre guias móveis consultar [Baus05]) seguem a mesma estratégia das aplicações de realidade aumentada, contudo, a informação enviada tem como objectivo guiar o utilizador. O PhoneGuide é uma aplicação para guiar visitas em museus com a particularidade de não necessitar de servidor. Identificado o objecto é conhecido o local e é enviada informação para guiar o visitante. O SnapToTell [Chevallet07] segue a mesma estratégia mas utiliza também a informação de localização para melhorar os resultados do reconhecimento. A localização restringe esta operação a um conjunto mais pequeno de imagens. Em [Beeharee06]



(a) (b) (c)

Figura 3.15: mCLOVER: a) interrogação por esboço; b) interrogação através de imagem exemplo; c) resultados.

Figura 3.16: MediAssist: Interface para pesquisa de fotos pessoais.

é proposta outra aplicação para guiar pessoas, neste caso em cidades, com duas diferenças em relação às anteriores: (1) a interrogação é realizada usando a localização e (2) as pessoas são guiadas usando apenas fotos visualizadas no mapa do local.

O Photo-to-Search [Fan05], o MediAssist [Gurrin05] e o MAMI [Anguera08] são projectos desenvolvidos com o objectivo de melhorar a pesquisa de fotos pessoais em dispositivos móveis. Em cada trabalho é proposta uma técnica diferente para construir a interrogação. O Photo-to-Search e o MAMI utilizam interrogações multimodais. O primeiro utiliza uma imagem e texto inserido manualmente e o segundo utiliza uma imagem e áudio inserido pelo utilizador para indexar imagens. Ambos propõem estratégias para combinar listas de imagens ordenadas por vários tipos de informação incluindo, textual, visual ou informação de áudio. O MAMI ao contrário dos restantes, não precisa de servidor o que pode ser uma vantagem. Contudo e dado que realiza todo o processamento localmente, existe a necessidade de restringir a complexidade dos algoritmos. O MediaAssist é uma aplicação que utiliza a informação temporal e de localização para definir interrogações contextuais ao nível semântico (ver figura 3.16), como por exemplo, condições meteorológicas ou estações do ano.

3.6 Síntese

Neste capítulo foi apresentado o trabalho relacionado com memórias pessoais, anotação e recuperação de imagens e com as respectivas interfaces. No caso da anotação, os trabalhos apresentados mostram que uma das soluções, para convencer os utilizadores a anotarem manualmente as suas fotos, é torná-la numa tarefa motivante. Na anotação automática com conceitos semânticos, um caminho promissor é a combinação de vários tipos de informação. No caso das interfaces para memórias pessoais, são necessárias aplicações que permitam aceder a fotos através de vários tipos de informação. Para terminar, as diversas propostas apresentadas com aplicações para dispositivos móveis que utilizam sistemas de recuperação de imagens, são uma indicação da relevância que estes sistemas poderão ter no desenvolvimento de novas aplicações.

4

Recuperação e Anotação de Imagens

Conteúdo

Neste capítulo é descrito o método proposto para anotação e recuperação semântica de imagens utilizando vários tipos de informação. É também proposto um algoritmo semi-automático de anotação com o objectivo de melhorar o desempenho do método automático.

4.1 Introdução

A eficiência na recuperação de fotos em colecções com um elevado número de imagens depende da forma como estas são indexadas. Os métodos de indexação dependem dos metadados associados a cada imagem (anotação). A forma natural do utilizador exprimir o que pretende é através de palavras, exceptuando os casos em que a componente semântica é baixa (por exemplo, a cor ou a textura são suficientes) ou o utilizador tem dificuldades em definir por palavras o que pretende. Nestas condições, uma imagem exemplo ou um esboço são os meios mais adequados para definir a interrogação. A anotação manual de imagens com palavras descrevendo o seu conteúdo é a forma mais eficiente (ver capítulo 3) mas é uma tarefa evitada pelos utilizadores porque o esforço humano necessário para anotar uma base de dados com muitas imagens é elevado. A alternativa consiste em desenvolver técnicas automáticas para realizar a anotação. Neste caso, o problema reside no tipo de informação a extrair da imagem e como utilizar estes dados para efectuar anotações automáticas ao nível semântico. O conteúdo da imagem é a informação mais utilizada (ver capítulo 3) mas, actualmente, a maioria das máquinas fotográficas permitem fazer anotações áudio, registar informação referente ao instante de captura no EXIF do ficheiro JPEG e algumas têm mesmo receptores de GPS para registar a localização. É também expectável que no futuro mais sensores sejam incluídos nos dispositivos de captura, à semelhança da SenseCam [Gemmell04], para gravar mais informação quando a fotografia é capturada.

Neste capítulo, é descrito um método sistemático para anotação e recuperação semântica de imagens utilizando vários tipos de informação e com capacidade para incluir novos tipos de dados provenientes de novos sensores. A próxima secção apresenta a arquitectura global do sistema proposto. A seguir, são descritos os métodos propostos para recuperação e anotação de imagens. Na secção referente à anotação também é apresentado um método semi-automático. Segue-se uma secção onde é feita a descrição da análise semântica efectuada em imagens, isto é, onde são treinados os modelos semânticos que são utilizados para recuperação e anotação. Finalmente, são descritos os métodos utilizados para extrair a informação das imagens.

4.2 Arquitectura

Nesta dissertação é proposto um sistema multimodal para recuperar e anotar fotos em colecções pessoais. Este sistema é utilizado em várias aplicações em diversas áreas. Assim, de modo a disponibilizar as funcionalidades do sistema de recuperação e anotação para várias aplicações foi desenvolvida uma infra-estrutura cliente/servidor (ver figura 4.1). O servidor inclui o sistema proposto e a colecção de fotos enquanto que os clientes são interfaces de pesquisa e anotação desta informação. Esta estrutura também facilita o desenvolvimento de aplicações em dispositivos com menores recursos computacionais uma vez que os algoritmos mais pesados computacionalmente são implementados no servidor. Na figura 4.1, são apresentadas as aplicações desenvolvidas que utilizam o sistema de recuperação e anotação de imagens. Para ser utilizada em ambientes familiares, em casa, foi implementada a aplicação Memoria Desktop. Em actividades de turismo, quando é efectuada uma visita a um local de interesse, geralmente são capturadas muitas fotografias. Nesta tese são propostas as aplicações Memoria Mobile e Memoria Web para complementar as experiências vividas durante e após a visita. Finalmente,

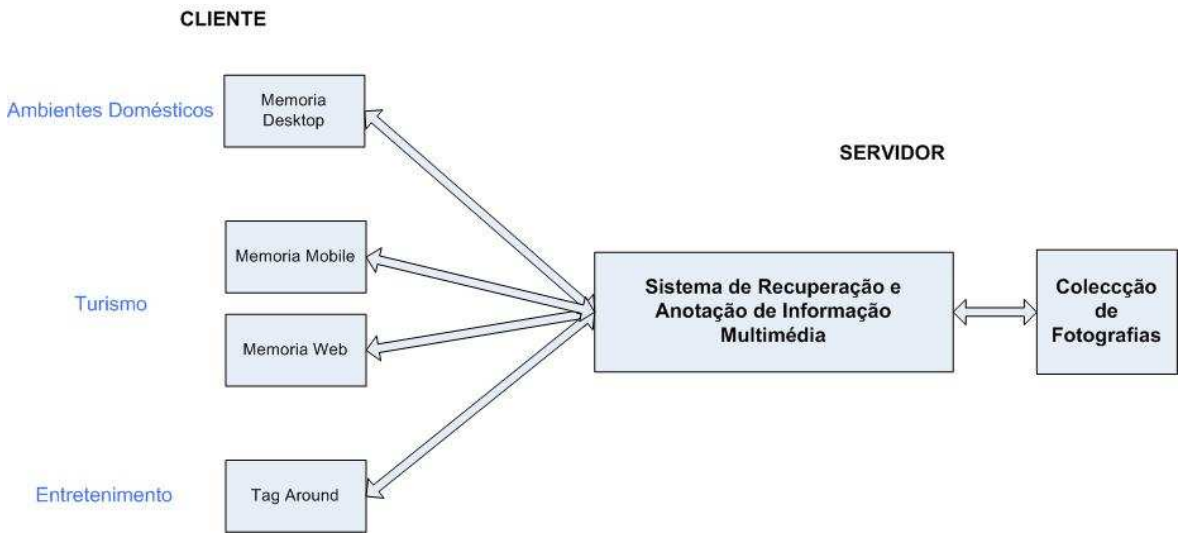


Figura 4.1: Arquitectura da infra-estrutura proposta para recuperação e anotação de informação multimédia.

foi concebido um jogo de computador, designado por Tag Around para anotar imagens de forma semi-automática. Estas aplicações são descritas nos capítulos seguintes. Neste capítulo é abordada a componente do servidor, isto é, o sistema de recuperação e anotação de imagens.

O sistema proposto nesta tese para recuperar e anotar imagens é baseado em modelos semânticos treinados utilizando informação multimodal. O sistema é compostos por três blocos (ver figura 4.2):

- Recuperação e anotação de imagens - aplicações da análise semântica;
- Análise semântica - estimação dos modelos semânticos;
- Extracção de informação - extracção de características visuais, informação de áudio e metadados contextuais (data e localização).

O primeiro bloco, recuperação e anotação de imagens, destina-se a aplicações que utilizem os conceitos semânticos estimados no bloco de análise semântica. A recuperação de imagens utilizando interrogações semânticas e a anotação de imagens com palavras são as duas tarefas utilizadas nesta tese para validar a análise semântica proposta [Jesus06, Jesus07, Jesus07a]. Ambas as tarefas utilizam modelos probabilísticos, representados por $p(w|I)$ para recuperar e anotar fotos, onde w denota um conceito semântico do vocabulário definido com P conceitos, $V_{con} = \{w_1, w_2, ..., w_P\}$ e I representa uma imagem da colecção composta por N imagens, $C_{img} = \{I_1, I_2, ..., I_N\}$.

No bloco da análise semântica, os modelos propostos são treinados com informação multimodal (ver figura 4.2), isto é, informação visual e de áudio e informação contextual obtida no instante de captura. O método proposto baseia-se em classificação binária para detectar a presença ou a ausência de um conceito numa imagem. Neste processo é utilizado o classificador RLS (Regularized Least Squares) [Poggio03] e uma função sigmóide para obter os modelos definidos pela $p(w|I)$.

No bloco da extracção de informação são apresentadas as técnicas utilizadas para representar cada imagem da base de dados. Admite-se que cada documento é constituído por uma

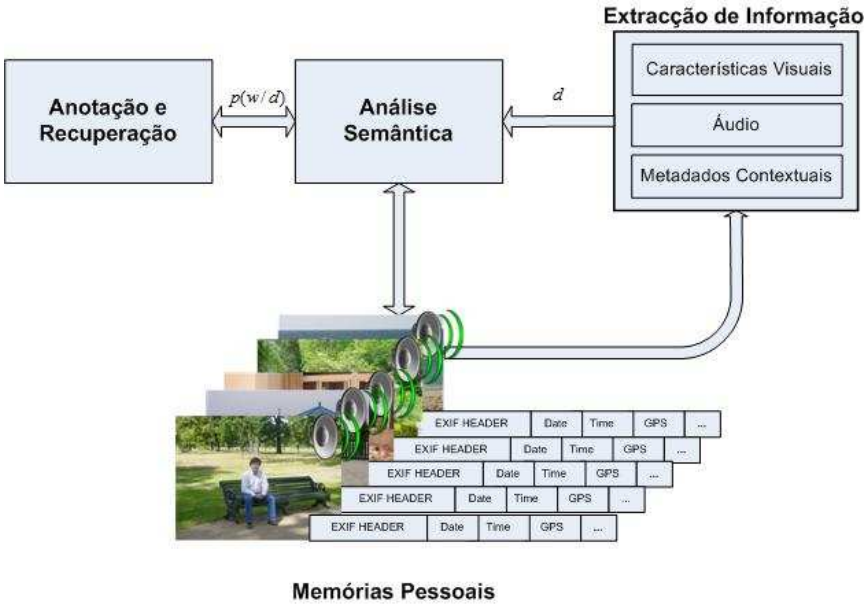


Figura 4.2: Recuperação e anotação baseada em análise semântica de imagens.

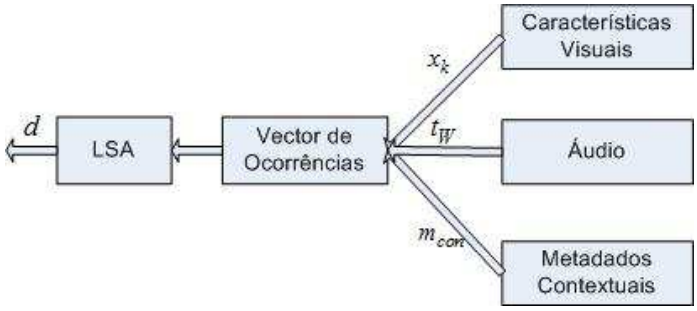


Figura 4.3: Vector de ocorrências - Representação de uma imagem obtida no bloco de extracção de informação.

imagem, um ficheiro de áudio com comentários dos utilizadores descrevendo a imagem e metadados obtidos no instante de captura e gravados na estrutura EXIF.

Os modelos propostos baseados em análise semântica para anotação e recuperação de imagens são treinados utilizando cor, textura, faces, o instante e localização de captura e informação extraída de áudio gravado no instante de captura. A figura 4.3 apresenta os blocos principais da metodologia utilizada para representar cada imagem da colecção. As características de cor e textura são detectadas globalmente, em regiões ou em pontos de interesse. O instante de captura t é obtido através da data e hora e a informação espacial é obtida através das coordenadas geográficas obtidas por GPS (latitude e longitude), registadas no instante de captura. Na figura 4.3, m_{con} representa esta informação. As características visuais x_k e os termos t_W reconhecidas a partir de áudio são representadas através de um vector de ocorrências d (*bag of words*) e é utilizado o LSA (Latent Semantic Analysis) para explorar as relações entre termos.

Em algumas situações, os modelos semânticos apresentam dificuldades e por isso é também proposto um algoritmo semi-automático de anotação que visa melhorar a anotação automática. A seguir, são descritos em detalhe cada um destes blocos em secções individuais.

4.3 Recuperação de Imagens com Informação Multimodal

A recuperação de informação é uma das aplicações da análise semântica descrita na secção 4.5. O objectivo desta aplicação consiste em criar uma lista ordenada de imagens de acordo com a relevância para a interrogação. Considerando que a interrogação é definida por K conceitos, $Q = \{w_1, w_2, ..., w_K\}$, que descrevem o cenário e alguns dos objectos presentes nas imagens desejadas (por exemplo, "Indoor", "People" e "Computers"), a posição na lista ordenada de uma imagem I, é obtida utilizando informação de áudio, visual, espacial e temporal através da medida de semelhança [Jesus07],

$$Sim(Q,I) = f_{gps}(Sim_{visual+time}(Q,I) + Sim_{audio}(Q,I)), \tag{4.1}$$

onde f_{gps} é uma função para seleccionar ou eliminar a imagem de acordo com o critério espacial (por exemplo, uma região ou uma direcção). O Sim_{audio} representa a semelhança entre a interrogação e a imagem utilizando a informação de áudio e $Sim_{visual+time}$ representa a mesma semelhança mas utilizando as características visuais e temporais. Estes termos são explicados nas secções seguintes. A lista de imagens resultante da interrogação Q é obtida ordenando as imagens de acordo com a função $Sim(Q,I)$.

4.3.1 Visual

A semelhança de uma imagem em relação a uma interrogação Q, constituída por K conceitos e considerando múltiplas características visuais, é a soma das semelhanças obtidas individualmente para cada característica [Jesus06],

$$Sim_{visual+time}(Q,I) = \sum_{j=1}^{R} a_j Sim_{visual+time}(Q, d_x^j), \tag{4.2}$$

onde R denota o número de características, d_x^j representa o vector de ocorrências da $j^{ésima}$ característica extraída da imagem I e a_j é o peso associado a cada característica assumindo $\sum_{j=1}^{R} a_j = 1$. Para interrogações Q com K conceitos a semelhança é obtida pela probabilidade conjunta,

$$Sim_{visual+time}(Q, d_x^j) = p(w_1, w_2, ..., w_K | d_x^j). \tag{4.3}$$

Admitindo independência entre os conceitos, a probabilidade conjunta dos conceitos dada uma imagem é,

$$p(w_1, w_2, ..., w_K | d_x^j) = \prod_{i=1}^{k} p(w_i | d_x^j). \tag{4.4}$$

As probabilidades $p(w_i|d_x^j)$ são obtidas com as equações 4.9 e 4.10, utilizando um classificador diferente $f_i(d_x^j)$ para cada conceito e característica visual.

4.3.2 Áudio

As palavras obtidas, utilizando aplicações de reconhecimento de fala, são utilizadas para anotar imagens (ver secção 4.4). A recuperação de imagens é realizada utilizando técnicas habitualmente usadas na recuperação de documentos de texto [Yates99].

Uma imagem I é representada por um vector de ocorrências de termos (radicais das palavras reconhecidas) $d_i = [W_{1,i}, W_{2,i}, ..., W_{N_t,i}]^T$ que são pesados utilizando a técnica designada por TF-IDF (Term Frequency- Inverse Document Frequency) [Jones03]. É também aplicado o método LSA (ver secção 4.6.1). Dada uma interrogação Q, também representada por um vector de ocorrências de termos $q = [W_{1,q}, W_{2,q}, ..., W_{N_t,q}]^T$, associados aos conceitos w através de uma ontologia, a semelhança de uma imagem com a interrogação é medida através do coseno do ângulo formado entre os dois vectores [Yates99],

$$Sim_{audio}(q, d_i) = \frac{\sum_{j=1}^{N_t} W_{j,i} W_{j,q}}{\sqrt{\sum_{j=1}^{N_t} W_{j,i}^2} \sqrt{\sum_{j=1}^{N_t} W_{j,q}^2}}, \tag{4.5}$$

em que $W_{j,i}$ e $W_{j,q}$ são os pesos (TD-IDF) dos termos dos vectores. Com esta medida é definido o termo $Sim_{audio}(Q, I)$ da equação 4.1.

Cada conceito do vocabulário V_{con} define um domínio semântico. Quando o utilizador anota uma imagem com áudio, pode usar várias palavras pertencentes ao domínio semântico, por isso, a interrogação definida por um conceito semântico é representada por um vector de ocorrências de termos pertencentes a um vocabulário específico de cada conceito e que é definido previamente. Esta estratégia segue o método publicado em [Haase04] que utiliza ontologias simples.

4.3.3 Metadados Contextuais

A informação de localização (coordenadas geográficas obtidas por GPS) das fotos também é utilizada para recuperar imagens [Jesus07a]. Nesta tese, esta informação é utilizada quando a interrogação inclui informação geográfica para definir um subconjunto de imagens da colecção. Dado que o conjunto a ordenar é mais pequeno, espera-se que o desempenho do sistema de recuperação aumente. São utilizados dois tipos de interrogação geográfica: (1) interrogação definindo uma região de interesse e (2) interrogação definindo uma direcção de interesse em relação a um ponto específico. Em ambas as situações são utilizadas as coordenadas GPS. No primeiro caso, são seleccionadas as imagens que estão dentro da região definida, isto é, as imagens que estão dentro da circunferência definida pela região. Para tal, é necessário calcular a distância entre as coordenadas obtidas por GPS de todas as imagens com o centro da região. A distância entre a localização de uma imagem Ig e o centro da região seleccionada Qg, é obtida utilizando a distância do Grande Círculo [Sinnott84],

$$dist(Qg, Ig) = r_{earth} acos[sin(lat_{Qg}) sin(lat_{Ig}) + cos(lat_{Qg}) cos(lat_{Ig}) cos(lon_{Ig} - lon_{Qg})]. \tag{4.6}$$

O $r_{earth} = 6378,7$Km, é o raio da terra, lat representa a latitude e lon representa a longitude. Todas as imagens que satisfaçam a condição,

$$dist(Qg, Ig) < r_{query}, \tag{4.7}$$

são ordenadas de acordo com a equação 4.6. O r_{query} é o raio do círculo obtido a partir do centro da região definida.

A interrogação através da definição de uma direcção selecciona as imagens que estão na

direcção em relação a um ponto indicado. São consideradas quatro direcções, "Norte", "Sul", "Este" e "Oeste". Assim, dado um ponto P_{gps} com coordenadas GPS (lat, lon) o subconjunto de imagens L_{gps} de C_{img} na direcção d_{gps} é:

- Se $d_{gps} =$ "$Norte$" então $L_{gps} = \{\forall I_g \in C_{img} : lat_{I_g} > lat\}$;
- Se $d_{gps} =$ "Sul" então $L_{gps} = \{\forall I_g \in C_{img} : lat_{I_g} < lat\}$;
- Se $d_{gps} =$ "$Este$" então $L_{gps} = \{\forall I_g \in C_{img} : lon_{I_g} > lon\}$.
- Se $d_{gps} =$ "$Oeste$" então $L_{gps} = \{\forall I_g \in C_{img} : lon_{I_g} < lon\}$;

As imagens na direcção desejada são ordenadas utilizando a equação (4.6).

Para os dois casos de interrogação geográfica, só é utilizada a ordenação por distância geográfica quando a pesquisa inclui apenas informação geográfica. Nesta condição, a informação geográfica deixa de ser apenas um filtro e passa a definir a ordem das imagens. Nas restantes situações, a informação espacial é utilizada como critério para seleccionar um subconjunto de imagens. Esta selecção define a função f_{gps} da equação 4.1.

4.4 Anotação

A anotação de imagens é outra tarefa que utiliza os modelos semânticos descritos na secção 4.5. Nesta tese, a anotação de imagens é encarada como uma tarefa necessária para a recuperação de imagens, porém, o método descrito nesta secção é independente do método apresentado para recuperação. A técnica de anotação também pode ser utilizada em outras aplicações, por exemplo para construir histórias de viagens [Kustanowitz05]. A associação de conceitos em imagens é realizada utilizando a informação visual e temporal através dos modelos semânticos propostos ou através de palavras reconhecidas em áudio. Os modelos podem ser utilizados de forma automática (descrito em baixo) ou num algoritmo semi-automático (descrito na secção seguinte).

4.4.1 Automática

O modelo de análise semântica de imagens, descrito na secção 4.5, é utilizado para recuperação de imagens (ver secção 4.3) mas também pode ser usado para associar conceitos a imagens (anotação). A probabilidade definida na equação 4.16, no caso da recuperação, é utilizada para posicionar uma imagem na lista ordenada criada como resultado da interrogação e no caso da anotação define uma medida para anotar um conceito numa imagem.

Dado um conceito w_i pertencente ao vocabulário V_{con} e utilizando a probabilidade $p_t(w_i|I_j)$ (ver equação 4.16), a anotação do conceito w_i na imagem I_j, $A_{j,i} = (I_j, w_i)$ é efectuada se,

$$p_t(w_i/I) > th, \tag{4.8}$$

onde th é obtido empiricamente de forma a maximizar o desempenho da anotação. Neste caso, a probabilidade $p(w_i|I)$ é obtida utilizando informação visual e temporal.

A aplicação Memoria Mobile proposta nesta tese (ver capítulo 6), permite que o utilizador grave, durante alguns segundos, comentários acerca da fotografia capturada. Após a captura, a aplicação liga o microfone para gravar comentários do utilizador e sons característicos do local,

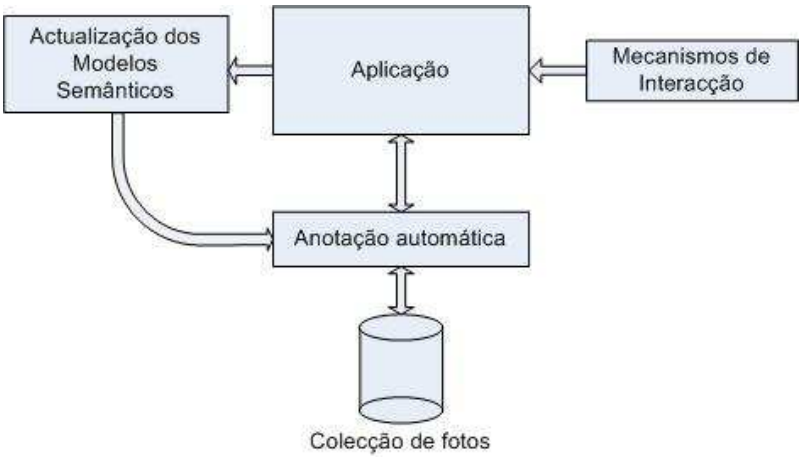


Figura 4.4: Anotação semi-automática.

por exemplo, o ruído produzido por automóveis ou produzido por animais. Nesta tese são utilizados apenas os comentários do utilizador para anotar imagens. Este áudio é convertido para texto usando reconhecimento automático de fala. Este texto é utilizado para anotação automática de imagens.

4.4.2 Semi-Automática

O método automático proposto para análise e anotação semântica em imagens apresenta dificuldades na detecção de alguns conceitos mais complexos por exemplo, "Snow" ou "Party". Isto acontece porque em algumas situações as características de baixo nível não são as mais adequadas para o conceito, porque não existe uma imagem idêntica no conjunto de treino ou porque a imagem é ambígua.

Para atenuar estas dificuldades, é proposto um algoritmo semi-automático [Jesus08] para anotar imagens que envolve uma aplicação (por exemplo um jogo de computador), o método automático para sugerir anotações e o utilizador para as corrigir (ver figura 4.4). Este algoritmo é baseado nos princípios da técnica de retroacção de relevância mas associando palavras às imagens relevantes. Esta técnica é geralmente utilizada quando o utilizador executa uma pesquisa ou navega na base de dados. No entanto, estes princípios, isto é, a inclusão do utilizador para validar os resultados obtidos pelo método automático e a utilização desta validação para melhorar as técnicas automáticas, podem ser usados noutras aplicações, para além da pesquisa ou navegação na colecção de imagens.

Nesta tese, é proposto um jogo de computador baseado em gestos (ver capítulo 7) em que as jogadas do utilizador validam o método automático e este é usado para calcular a pontuação. A figura 4.4 apresenta o diagrama de blocos do sistema proposto para anotação semi-automática. No bloco da aplicação é utilizado o jogo que disponibiliza ao utilizador um conjunto de imagens e conceitos. O jogador procura associar correctamente estas imagens com os conceitos (mecanismos de interacção). Estas anotações, quando correctas, são usadas para melhorar o desempenho da anotação automática (actualização dos modelos semânticos). Finalmente, o bloco da anotação automática contribui para o cálculo da pontuação que mede a correcção da anotação.

Dado um conjunto $L \subset C_{img}$ com N_l imagens, e um conjunto $V_{sc} \subset V_{con}$ com N_{con} conceitos, o algoritmo proposto é definido pelos seguintes passos:

1. Os subconjuntos L e V_{sc} são apresentados na interface da aplicação;
2. O utilizador escolhe uma imagem $I_l \in L$ e um conceito $w_k \in V_{sc}$;
3. O utilizador faz uma anotação $A_{l,k} = (I_l, w_k)$;
4. É calculada a pontuação utilizando os modelos automáticos, a confiança no jogador e as jogadas de jogadores anteriores (ver capítulo 7);
5. Para todos os conceitos $w_k \in V_{sc}$, se o $|\{A_{1,k}, A_{2,k}, ..., A_{N_A,k}\}| > N_{upd}$ para um conceito w_k então é actualizado o conjunto de treino de w_k e estimado um novo modelo semântico;
6. Ir para 2.

Um modelo semântico é estimado novamente quando o número de anotações diferentes do respectivo conceito é superior a N_{upd}. Uma anotação é considerada válida quando é efectuada pelo menos por dois utilizadores diferentes. Depois de validado, o par imagem/conceito (anotação) é incluído no conjunto de treino.

Como resultado deste algoritmo temos o conjunto total de anotações e os modelos dos conceitos do conjunto V_{con} estimados com um conjunto de treino maior. Assim, quantas mais iterações, mais anotações temos e melhores modelos são obtidos.

O subconjunto de imagens de C_{img} e o subconjunto de conceitos de V_{con} usados em cada nível do jogo são escolhidos no bloco de anotação automática (ver figura 4.4) e por isso, o método automático pode conduzir o processo de aprendizagem. Desta forma, é possível utilizar as técnicas habituais nos métodos de retroacção de relevância (por exemplo a aprendizagem activa).

O jogo proposto para o bloco da aplicação é descrito no capítulo 7.

4.5 Análise Semântica

Na recuperação de documentos de texto, a maioria dos métodos propostos com melhores resultados [Yates99] são baseados no modelo do espaço vectorial [Salton75]. O modelo do espaço vectorial é um modelo algébrico para representar documentos através de vectores de termos. Por exemplo, dado o vocabulário $V = \{$"*papel*","*natureza*","*pessoa*","*exterior*",...$\}$, o documento de texto apresentado na figura 4.5a é representado numa forma simplificada (requer normalização) pelo vector $d_T = [1,0,3,0,...]^T$, com o número de ocorrências de cada palavra. Este vector indica quais os termos presentes e quantas vezes ocorrem. O método proposto para recuperação e anotação de imagens segue a mesma estratégia utilizada na recuperação de texto. O objectivo é estimar modelos semânticos para permitir uma representação idêntica para as imagens. Por exemplo, representar a imagem da figura 4.5b pelo vector, $d_I = [0,1,3,1,...]^T$.

Um documento de texto é representado por palavras e, por isso, a detecção de palavras e o cálculo da importância de cada uma no documento é menos complexa do que numa imagem que descreve uma cena que é caracterizada por um conjunto de objectos (por exemplo, pessoas ou edifícios) e o cenário (por exemplo, urbano, cenário de montanha ou de praia) onde a fotografia foi tirada.

O método proposto para descrever uma imagem baseia-se numa combinação de detectores individuais que, à semelhança do texto, permite obter a informação dos conceitos que estão

Memórias Pessoais

As memórias pessoais tradicionais são constituídas por papel ou formatos analógicos, por exemplo, jornais, diários pessoais, livros, álbuns de fotografia ou discos de vinil. No entanto, o avanço tecnológico permitiu que as memórias pessoais possam ser constituídas por informação em formato digital por exemplo, através de *emails*, ficheiros, páginas da web, mensagens, músicas, imagens ou vídeos [5].

Actualmente estão à disposição vários dispositivos móveis que permitem armazenar uma elevada quantidade de informação e em qualquer lugar. Mas esta facilidade na captura e armazenamento da informação só será útil, se for possível usar esta informação para apoiar as actividades realizadas no dia a dia. Assim, é preciso de alguma forma gerir esta informação. Este problema foi abordado por Vannevar Bush em 1945 em [6]. Neste artigo Bush considera um dispositivo futuro com o nome "Memex" para armazenar todos os livros, registos e comunicações de um indivíduo. Um dispositivo mecanizado de forma a que a consulta seja bastante rápida e flexível e que possa ser um complemento à memória humana. As ideias de Bush foram realizadas na década de 1960 usando tecnologia desenvolvida por Engelbart e publicadas em [7] e, mais tarde, com uma nova infra-estrutura computacional por [8].

(a) (b)

Figura 4.5: Informação: a) documento de texto; b) imagem.

presentes numa imagem. Nesta tese é utilizado o RLSC como classificador binário para detectar a presença/ausência de um conceito (objecto ou cenário) numa imagem e uma função sigmóide para normalizar os resultados obtidos pelo classificador. Após esta classificação, é analisada a correlação temporal entre imagens. Esta informação é incluída no modelo para melhorar o processo de classificação.

Na fase de análise do documento, o áudio é utilizado separadamente em relação à informação visual e temporal. Para a informação de áudio são utilizadas técnicas habitualmente usadas na área de reconhecimento de fala. A integração dos diversos tipos de dados é efectuado nos métodos de recuperação e anotação.

Dada a colecção C_{img} e o vocabulário de conceitos, V_{con}, o objectivo da análise semântica é estimar a probabilidade $p_t(w_i|I_j)$ do conceito w_i dada a imagem I_j, para todos os conceitos pertencentes a V_{con} e todas as imagens da colecção C_{img}.

Nas secções seguintes são descritas as contribuições de cada tipo de informação utilizada.

4.5.1 Visual

O método de análise semântica de imagens, no caso da informação visual, utiliza duas estratégias (ver figura 4.6): uma genérica para qualquer conceito [Jesus06] e outra específica para conceitos mais complexos que requerem a utilização de uma metodologia específica para a sua detecção numa imagem. Em ambas as estratégias, as imagens são representadas por vectores de características visuais e são utilizados classificadores. O resultado da classificação pode ser utilizado para recuperar ou anotar imagens automaticamente, contudo, esta medida não está normalizada e por isso não é adequada para combinar diferentes tipos de informação (por exemplo, informação visual, áudio, informação temporal ou espacial) ou vários conceitos. Assim, para normalizar a classificação é utilizada uma função sigmóide.

No caso da estratégia para conceitos específicos, é proposta uma técnica para detectar faces baseada no algoritmo AdaBoost proposto por Viola e Jones [Viola04] e num filtro de pele [Grangeiro08, Grangeiro09]. Este método é também utilizado para classificação do género (masculino/feminino) da pessoa. A detecção de faces e a classificação de género são descritas na secção seguinte. Em relação ao método genérico é utilizado o classificador RLS.

Para converter a saída do classificador numa probabilidade, várias estratégias têm sido propostas. Em [Wahba99] é proposto um método baseado na regressão logística para produzir

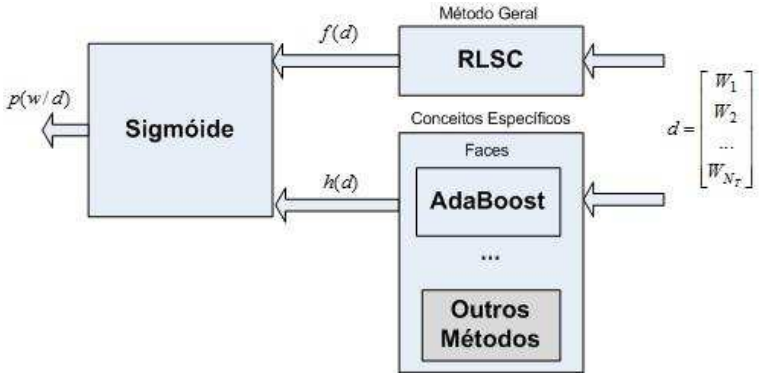


Figura 4.6: Diagrama de blocos do sistema de análise semântica utilizando informação visual.

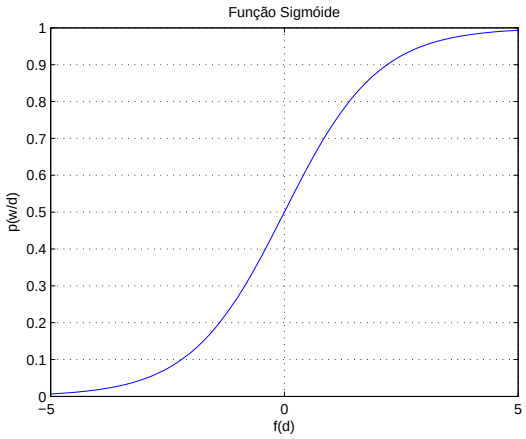


Figura 4.7: Função sigmóide, $f(d)$ representa a função discriminante do classificador.

valores probabilísticos a partir de valores de classificação obtidos por métodos baseados em *Kernel*. Platt [Platt99], em vez de estimar a função de densidade condicionada à classe como em [Wahba99], propõe um modelo paramétrico para estimar directamente a probabilidade *a posteriori*. O modelo proposto nesta tese para detectar conceitos em imagens utiliza a estratégia de Platt adaptada ao RLSC.

Considerando w uma variável aleatória de Bernoulli, em que uma realização pode ter como resultado um de dois conceitos disjuntos (por exemplo,"Indoor"/"Outdoor"), a probabilidade de w dada uma imagem representada por d_x, $p(w|d_x)$ pode ser obtida utilizando o resultado da classificação obtido por $f(d_x)$ e a função sigmóide (ver figura 4.7) de forma semelhante a [Platt99]. Definindo o espaço amostral de w, $\omega = \{-1,1\}$ temos para a saída positiva (por exemplo "Indoor"),

$$p(w=1|d_x) = \frac{1}{1+e^{-Af(d_x)+B}}, \tag{4.9}$$

e para a saída negativa (por exemplo "Outdoor"),

$$p(w=-1|d_x) = 1 - p(w=1|d_x) = \frac{e^{-Af(d_x)+B}}{1+e^{-Af(d_x)+B}}, \tag{4.10}$$

onde A e B são dois parâmetros para ajustar a curva da sigmóide (ver figura 4.7). Estes parâmetros são obtidos empiricamente mas podem ser estimados utilizando o método de validação cruzada. Em [Platt99] são apresentados e discutidos vários métodos para estimar os parâme-

tros da sigmóide.

A função $f(d_x)$ representa a função discriminante do classificador. Em geral, os modelos semânticos propostos utilizam o classificador RLS, contudo, para os conceitos mais complexos (por exemplo, "Face" ou "Party"), é utilizada outra metodologia, que pode incluir outro classificador, mais adequado para lidar com as dificuldades relacionadas com esses conceitos (ver secção 4.5.1.1).

Dado um conjunto de treino $S_m = \{(d_i, y_i)_{i=1}^M\}$ com $y_i \in \{-1, 1\}$ e sendo d_i um vector de características da imagem, a fronteira de decisão entre duas classes (por exemplo, "Indoor" e "Outdoor"), no caso geral (classificador RLS), é obtida pela função discriminante,

$$f(d) = \sum_{i=1}^{M} c_i K(d_i, d), \tag{4.11}$$

onde $K(d_i, d)$ representa o *Kernel* Gaussiano $K(d_i, d) = e^{-\frac{\|d_i - d\|^2}{2\sigma^2}}$, M é o número de pontos de treino e $c = [c_1, ..., c_M]^T$, é o vector de coeficientes estimado pelo método dos mínimos quadrados [Poggio03],

$$(m\gamma I + K)c = Y, \tag{4.12}$$

onde I é a matriz identidade, K é uma matriz quadrada definida positiva com os elementos $K_{i,j} = K(d_i, d_j)$, Y é um vector com as coordenadas y_i e γ é um parâmetro de regularização. Os valores óptimos de σ e γ foram obtidos pelo método de validação cruzada.

Um ponto d_i com $f(d_i) \leq 0$, é classificado na classe negativa ($y_i = -1$) e um ponto com $f(d_i) > 0$ é classificado na classe positiva ($y_i = 1$).

4.5.1.1 Faces

Em geral, o método proposto para extrair a informação semântica em imagens pode ser utilizado para estimar um modelo para qualquer conceito semântico. A única exigência é a existência de um conjunto de treino devidamente anotado. No entanto, para alguns conceitos mais complexos, por exemplo faces, são necessárias metodologias específicas para lidar com as dificuldades impostas pela detecção destes conceitos. Assim, é utilizado o método proposto por Viola e Jones [Viola04] para detectar faces em memórias pessoais compostas por fotos. Tal como referido, para lidar com as dificuldades relacionadas com a elevada variabilidade de faces neste domínio, é proposto um filtro de pele [Grangeiro08a] para corrigir alguns erros do método de Viola e Jones. O filtro de pele é aplicado sobre as faces detectadas pelo método de Viola e Jones e utiliza uma SVM e a sigmóide usada no método proposto para estimar os modelos semânticos (ver figura 4.6).

No método proposto em [Viola04] a classificação é realizada por uma cascata de classificadores cada um treinado pelo algoritmo Adaboost que calcula o classificador forte através da combinação de classificadores fracos,

$$h(d) = \begin{cases} 1 & \sum_{i=1}^{R} \alpha_i h_i(d) \geq \frac{1}{2} \sum_{i=1}^{R} \alpha_i \\ 0 & \text{para outros casos.} \end{cases}, \tag{4.13}$$

onde d denota o vector de características que, para este caso, não são representadas por vectores de ocorrências, R representa as hipóteses utilizando uma única característica, $\alpha_i = \log(\frac{1-e_i}{e_i})$ é

o peso da característica que é obtido utilizando o erro de classificação e_i e $h_i(d)$ representa o classificador fraco que apresenta o erro mais baixo. Na fase inicial, a cascata de classificadores é composta apenas por um único classificador. À medida que se progride pelos vários blocos da cascata a complexidade vai aumentando progressivamente até chegar à fase final onde é confirmada a presença de uma face numa sub-janela da imagem.

Nas faces detectadas pelo método de Viola e Jones é aplicado o filtro de pele [Grangeiro08a] para confirmar ou rejeitar a presença da face. Assumindo que a face está orientada com os eixos e conhecendo as coordenadas do ponto central (P_x, P_y) da face detectada, cada face é representada por medidas estatísticas dos pixels contidos na elipse,

$$\frac{(x-P_x)^2}{a^2} + \frac{(y-P_y)^2}{b^2} = 1, \tag{4.14}$$

onde a e b são os comprimentos dos semi-eixos maior e menor respectivamente. Uma face é representada pelo vector $x_k = [\mu_r, \mu_g, \sigma_r, \sigma_g]^T$, onde μ representa a média da cor e σ é o desvio padrão da cor dos pixels contidos na elipse. Neste trabalho, esta representação da face é obtida utilizando o espaço de cor RGB normalizado (ver [Vezhnevets03] para mais informações acerca do desempenho dos espaços de cor em filtros de pele) definido por,

$$c_n = \frac{C}{R+G+B}, \tag{4.15}$$

onde c_n denota a componente de cor normalizada e C a componente de cor. O vector x_k é utilizado como vector de características no treino de uma SVM [Muller01] para detectar a presença ou ausência da cor de pele na faces detectadas pelo método de Viola e Jones. Ao resultado obtido pelo classificador SVM é aplicada a função sigmóide descrita nas equações 4.9 e 4.10.

Classificação de Género

O género (masculino ou feminino) das pessoas detectadas em fotografias é outro conceito importante porque pode ajudar a encontrar uma pessoa ou um conjunto de pessoas de um género específico. Uma das formas para fazer esta classificação baseia-se na detecção e análise da face [Moghaddam00, Baluja07, Lapedriza06]. A técnica proposta [Grangeiro09, Grangeiro08a] também segue esta estratégia. Em primeiro lugar, é realizada a detecção da face utilizando o método apresentado na secção anterior. Depois é aplicado o filtro de pele para eliminar alguns falsos positivos e é utilizado o método PCA (Principal component analysis) [Turk91] para representar cada face num espaço de menor dimensão. Esta representação é utilizada no treino de uma SVM para detectar o género, masculino ou feminino da face detectada. Finalmente, é aplicada a sigmóide (ver equações, 4.9 e 4.10) para normalizar os resultados do classificador.

4.5.2 Áudio

A informação de áudio é obtida utilizando uma das técnicas com melhores resultados obtidos no reconhecimento automático de fala, os modelos de Markov não observáveis (HMMs) [Rabiner90]. Dado um ficheiro de áudio, foi utilizada uma aplicação fornecida pela Microsoft Language Development Center em Portugal [MLDCPort05] para reconhecer palavras a partir de áudio. Estas palavras são associadas a imagens e são utilizadas em conjunto com o modelo obtido com a informação visual e temporal nos métodos de recuperação e anotação de imagens.

(a) (b) (c)

Figura 4.8: Imagens consecutivas capturadas com um intervalo de 10s.

4.5.3 Metadados Contextuais

A maioria das aplicações para gerir memórias pessoais utiliza a informação temporal para organizar fotografias de modo a permitir que o utilizador possa navegar na colecção pessoal, através da data de captura de cada foto (ver capítulo 3). No entanto, a informação temporal também pode ser utilizada para inferir anotações numa imagem, utilizando as anotações de imagens capturadas em instantes temporalmente próximos pelo mesmo utilizador. Por exemplo, a figura 4.8 mostra três imagens capturadas com um intervalo de 10s. Nas duas imagens apresentadas nas figuras 4.8a e 4.8c não deverá ser difícil identificar o conceito "Beach" mas o mesmo poderá não acontecer na figura 4.8b. Nestas condições, a proximidade temporal pode ajudar a inferir o conceito "Beach" na figura 4.8b utilizando as outras duas imagens.

O modelo semântico apresentado na secção 4.5 é ajustado utilizando a informação temporal obtida no instante de captura. Seja $T = [t_1, t_2, ..., t_N]$ o vector ordenado com os instantes de captura das imagens da colecção C_{img}, a probabilidade de um conceito w dada uma imagem I_{t_i} capturada no instante t_i é,

$$p_t(w|I_{t_i}) = \frac{\alpha_{t_{i-1}} p(w|I_{t_{i-1}}) + p(w|I_{t_i}) + \alpha_{t_i} p(w|I_{t_{i+1}})}{1 + \alpha_{t_{i-1}} + \alpha_{t_i}}, \tag{4.16}$$

onde α_{t_i} e $\alpha_{t_{i-1}}$ pesam a importância da anotação de w nas imagens capturadas nos instantes anterior e posterior à imagem I_{t_i}. Estes pesos são inversamente proporcionais à distância temporal entre as imagens,

$$\alpha_{t_i} = 1 - \frac{d(t_i)}{d_{max}}, \tag{4.17}$$

onde d_{max} é uma constante obtida empiricamente que representa a distância temporal máxima permitida para que as anotações de uma imagem possam influenciar a outra imagem e $d(t_i)$ é a distância temporal entre a imagem capturada no instante t_i e a imagem capturada no instante a seguir,

$$d(t_i) = \begin{cases} |t_{i+1} - t_i|, & |t_{i+1} - t_i| < d_{max} \\ d_{max}, & \text{para outros casos.} \end{cases} . \tag{4.18}$$

Este método, que se baseia na ideia de utilizar a correlação temporal entre imagens para melhorar as anotações, também pode ser aplicado utilizando a correlação espacial através das coordenadas geográficas obtidas por GPS.

4.6 Extracção de Informação

Para representar as imagens da colecção de fotos é extraída informação de quatro tipos:

1. Características visuais - cor e textura;
2. Informação temporal - obtida a partir da data e hora de captura da foto;
3. Informação espacial - localização obtida utilizando informação extraída do GPS na forma de coordenadas geográficas;
4. Informação de áudio - palavras reconhecidas a partir de áudio.

As características visuais são representadas num vector de ocorrências de padrões visuais pertencentes a um vocabulário, que é obtido a partir do agrupamento de características detectadas em toda a colecção. A seguir, é apresentado o vector de ocorrência de termos que é aplicado às palavras reconhecidas a partir de áudio e às características visuais. Nas secções seguintes são descritas as características visuais, a informação de áudio e a informação contextual em secções individuais.

4.6.1 Vector de Ocorrências

Nos últimos anos, a representação de imagens por um vector de ocorrências de padrões visuais pertencentes a um vocabulário tem sido utilizada com sucesso na classificação de imagens [Nowak06], seguindo a estratégia popularizada na recuperação de documentos de texto [Salton86, Yates99]. Nesta tese, também é utilizada esta representação para as palavras reconhecidas a partir de áudio e para as características visuais. Um factor importante neste método é o vocabulário. No caso das palavras, é construído um vocabulário com os radicais da língua portuguesa. Para as características visuais a descrição de uma imagem com um vector de ocorrências encontra três dificuldades principais:

1. Construção do vocabulário - qual a melhor estratégia para encontrar os termos visuais representativos de toda a colecção?
2. Representação dos termos visuais - qual o melhor descritor?
3. Informação a extrair - quais as zonas da imagem a extrair?

Em [Nowak06] são discutidas e comparadas várias estratégias para solucionar estas dificuldades. Para a primeira questão, a solução mais utilizada é o algoritmo k-médias que poderá ser incremental por motivos computacionais. Relativamente à segunda questão, os autores de [Nowak06] apontam o descritor SIFT [Lowe04] como uma das soluções com melhor desempenho. Estas duas soluções foram também seguidas nesta tese. O mesmo não acontece na terceira questão. Nowak *et al.* apresentam gráficos de desempenho que sugerem que uma das melhores soluções é a extracção de descritores visuais (por exemplo, de cor ou textura) em pontos amostrados regularmente. No trabalho proposto nesta tese não é seguida esta estratégia porque em zonas onde não há objectos ou texturas relevantes não é necessário extrair vários pontos. Em alternativa, utilizou-se o detector de pontos de interesse proposto em [Lowe04] para representar texturas e a imagem foi segmentada para representar as cores mais importantes.

Em geral, uma imagem é representada por um vector de ocorrências de termos (podem ser radicais de palavras ou descritores visuais representativos). Estes vectores são normalizados e depois é aplicado o método de LSA [Deerwester90] à matriz de ocorrências para explorar a semântica latente entre os termos nos documentos e para baixar a característica da matriz.

Dada a colecção de imagens C_{img} e um vocabulário com N_t termos $V_t = \{t_{w_1}, t_{w_2}, ..., t_{w_{N_t}}\}$, o vector de ocorrências para uma imagem I_k é representado pelo histograma que conta o número de ocorrências, $N_{l,k} = n(t_{w_l}, I_k)$, de cada termo t_{w_l} na imagem I_k,

$$d_k(l) = \sum_{i=0}^{N_t} N_{l_k} \delta(l-i), \tag{4.19}$$

onde $\delta(l-i)$ representa a função de Knonecker.

Em [Salton86] são discutidas e apresentadas várias estratégias para pesar estes termos, sendo o mais utilizado o TF-IDF que é também usado nesta tese. Assim, a matriz termo-documento da colecção C_{img} é representada por,

$$X_{img} = \begin{pmatrix} W_{1,1} & W_{1,2} & \dots & W_{1,N-1} & W_{1,N} \\ W_{2,1} & W_{2,2} & \dots & W_{2,N-1} & W_{2,N} \\ \dots & \dots & \dots & \dots & \dots \\ W_{N_t-1,1} & W_{N_t-1,2} & \dots & W_{N_t-1,N-1} & W_{N_t-1,N} \\ W_{N_t,1} & W_{N_t,2} & \dots & W_{N_t,N-1} & W_{N_t,N} \end{pmatrix}, \tag{4.20}$$

onde $W_{l,k} = \frac{N_{l,k}}{\sum_l N_{l,k}} \log(\frac{N}{n_l})$, são os pesos obtidos aplicando o TF-IDF. O n_l representa o número de imagens onde ocorre o termo l.

4.6.1.1 Latent Semantic Analysis

Em geral, a matriz X_{img} é uma matriz esparsa de dimensões elevadas e por isso a sua manipulação pode ser limitada do ponto de vista computacional. Por outro lado, a matriz termo-documento, descrita na secção anterior, apresenta limitações ao lidar com a sinonímia e a polissemia, problemas fundamentais no processamento de linguagem natural [Yates99]. Para resolver estas dificuldades é aplicado o método LSA [Deerwester90] na matriz X_{img}. O objectivo deste método é encontrar uma aproximação da matriz termo-documento com uma característica mais baixa, utilizando a decomposição em valores singulares (SVD) para representar os documentos no espaço de menor dimensão designado pelo espaço de conceito. Aplicando o método SVD, algumas dimensões são combinadas e passam a depender de mais de um termo como é ilustrado na figura 4.9. Na figura 4.9a, são exemplificadas as relações termo-documento (matriz X_{img}) e na figura 4.9b são apresentadas as mesmas relações mas depois de aplicar o LSA. Os termos e os documentos passam a relacionar-se através de um espaço intermédio, o espaço de conceito. Na figura 4.9, o documento d_2 pode ser recuperado com uma interrogação com o termo t_2 por partilhar a ocorrência de t_1 com o documento d_1 que contém ocorrências de t_2. Assim, é possível atenuar o problema da sinonímia e da polissemia ao relacionar os termos e documentos num espaço de semântica latente.

No caso das características visuais, o objectivo é relacionar aspectos visuais que estão contidos na interrogação e que não ocorrem em algumas imagens relevantes. Se uma imagem relevante, com aspectos visuais contidos na interrogação, partilhar aspectos visuais que não estão na interrogação com imagens relevantes então, através das relações dos termos, é possível

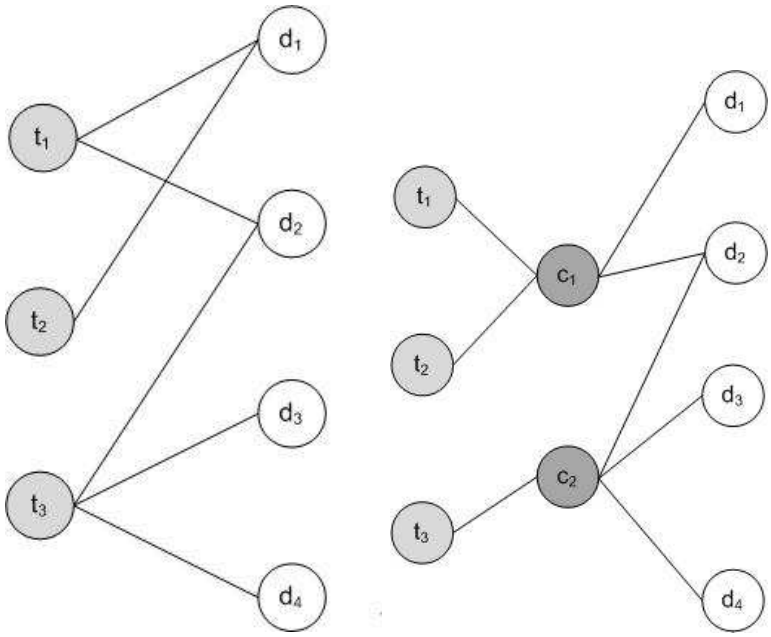


Figura 4.9: LSA: a) Espaço termo-documento; b) Espaço termo-tópico-documento.

recuperar estas imagens. A formalização matemática do método pode ser consultada em [Deerwester90], onde também são apresentados mais detalhes sobre o LSA.

4.6.2 Características Visuais

Nesta secção são apresentados os descritores visuais utilizados nesta tese:

- Momentos de cor no espaço HSV;
- Regiões de cor no espaço LUV;
- Características de textura obtidas com o banco de filtros de Gabor;
- Descritor SIFT.

No caso particular da detecção de faces, são utilizadas características específicas obtidas aplicando o método proposto em [Viola04]. Viola e Jones detectam um número elevado de características baseadas nas funções de Haar e utilizam o algoritmo Adaboost para escolher um número reduzido de características e estimar o classificador forte com base nas características seleccionadas. Para analisar a imagem toda, de forma eficiente, é utilizada uma cascata de classificadores. A complexidade dos classificadores aumenta progressivamente até chegar à fase final onde é confirmada a presença de uma face numa sub-janela da imagem. O objectivo passa pela eliminação do maior número de sub-janelas possível nas etapas iniciais, fazendo que a passagem pelas últimas e mais complexas etapas seja um acontecimento pouco comum. Desta forma, a maioria das sub-janelas é rapidamente eliminada nas etapas iniciais aumentando a eficiência computacional.

As características utilizadas variam entre descritores de cor (momentos de cor e regiões de cor) e de textura (filtro de Gabor e descritor SIFT). Também são aplicadas várias técnicas para extrair os descritores das imagens: globalmente (filtro de Gabor), em zonas rectangulares

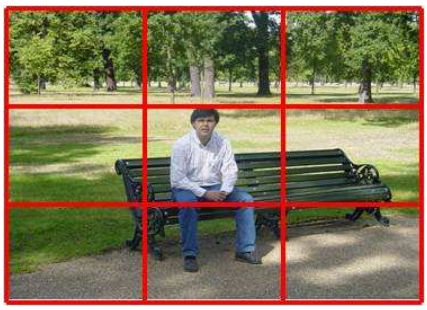


Figura 4.10: Características de cor em 9 regiões.

(momentos de cor) dividindo a imagem em 9 blocos iguais, em pontos de interesse (filtro de Gabor e descritor SIFT) e em objectos após segmentação (regiões de cor).

4.6.2.1 Momentos de Cor

O objectivo desta característica é a identificação de padrões de cor no espaço HSV [Bimbo99] em blocos da imagem (ver figura 4.10). As fotos pessoais são capturadas em diversas situações, por isso, o espaço HSV, pelas suas características de invariância, é o indicado para analisar a cor. A divisão da imagem em blocos tem como objectivo a captura de cores localizadas. A imagem é dividida em 9 blocos (ver figura 4.10). Em cada bloco é calculada a média e variância de cada componente de cor. A média no bloco k para a componente de cor c é obtida por,

$$\mu_{k,c} = \frac{1}{N_{col}N_{lin}} \sum_{i=1}^{N_{col}} \sum_{j=1}^{N_{lin}} I_{k,c}(i,j), \tag{4.21}$$

onde N_{col} e N_{lin} representam o número de colunas e o número de linhas do bloco k da imagem. A variância é obtida por,

$$\sigma^2_{k,c} = \frac{1}{N_{col}N_{lin}} \sum_{i=1}^{N_{col}} \sum_{j=1}^{N_{lin}} [I_{k,c}(i,j) - \mu_{k,c}]^2. \tag{4.22}$$

Cada $I_k \in C_{img}$ é representada pelo vector $x_k = [\mu_{1,1}, \sigma^2_{1,1}, ..., \mu_{1,N_c}, \sigma^2_{1,N_c}, ..., \mu_{N_b,N_c}, \sigma^2_{N_b,N_c}]^T$.

4.6.2.2 Regiões de Cor

A divisão da imagem efectuada na característica anterior produz descontinuidades, por exemplo em objectos, que se reflectem nas características extraídas. Para contornar este problema é efectuada uma segmentação de cor utilizando um algoritmo baseado no Mean-Shift [Comaniciu02]. Na figura 4.11b, é apresentado um exemplo com a imagem da figura 4.11a segmentada utilizando este algoritmo (para mais detalhes consultar [Comaniciu02]).

Cada região detectada na imagem é representada pela média μ_c e variância σ^2_c de cada componente de cor, pela média μ_x, μ_y e variância σ^2_x, σ^2_y das coordenadas dos pixels da região e pela percentagem de pixels N_{pixels}. O vector $xr_k = [\mu_{c_1}, \sigma^2_{c_1}, ..., \mu_{c_n}, \sigma^2_{c_n}, \mu_x, \mu_y, N_{pixels}, \sigma^2_x, \sigma^2_y]^T$ representa uma região.

4.6.2.3 Filtro de Gabor

Para detectar texturas em imagens são extraídas medidas estatísticas em imagens filtradas pelo banco de filtros de Gabor [Manjunath96]. O banco de filtros de Gabor permite analisar imagens

(a) (b)

Figura 4.11: Regiões de cor utilizando o algoritmo Mean Shift: a) Imagem Original; b) Imagem Segmentada.

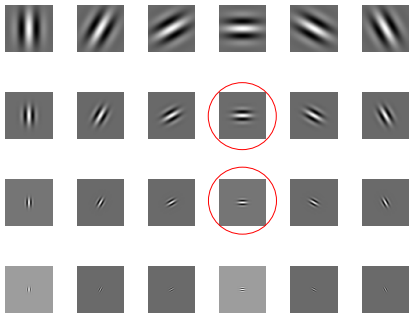


Figura 4.12: Filtro de Gabor - Banco de filtros.

em várias escalas e rotações. Cada imagem I da colecção C_{img} é filtrada utilizando a equação,

$$W_{m,n}(x,y) = \int I(x,y) g^*_{m,n}(x-x_1, y-y_1) d_{x_1} d_{y_1}, \tag{4.23}$$

onde $g_{m,n}$ denota um filtro de Gabor. Na figura 4.12 é representado o banco de filtros obtidos para 6 rotações e 4 escalas. Cada imagem é representada pela média $\mu_{o,s}$ e variância $\sigma^2_{o,s}$ do módulo da imagem filtrada $W_{m,n}$ para cada orientação o e escala s, no vector, $x_k = [\mu_{1,1}, \sigma^2_{1,1}, ..., \mu_{N_{orient},N_{scale}}, \sigma^2_{N_{orient},N_{scale}}]^T$.

Na figura 4.14, são apresentadas as imagens filtradas pelo banco de filtros de Gabor da imagem presente na figura 4.13, para exemplificar o tipo de informação que é extraída pelo banco de filtros de Gabor. A imagem (figura 4.13) inclui uma pessoa com uma camisola às riscas orientadas horizontalmente. Nas figuras 4.12 e 4.14, verifica-se que as imagens filtradas (inseridas num círculo) que mais reflectem as riscas da camisola correspondem aos filtros orientados horizontalmente.

Figura 4.13: Filtro de Gabor - Imagem original.

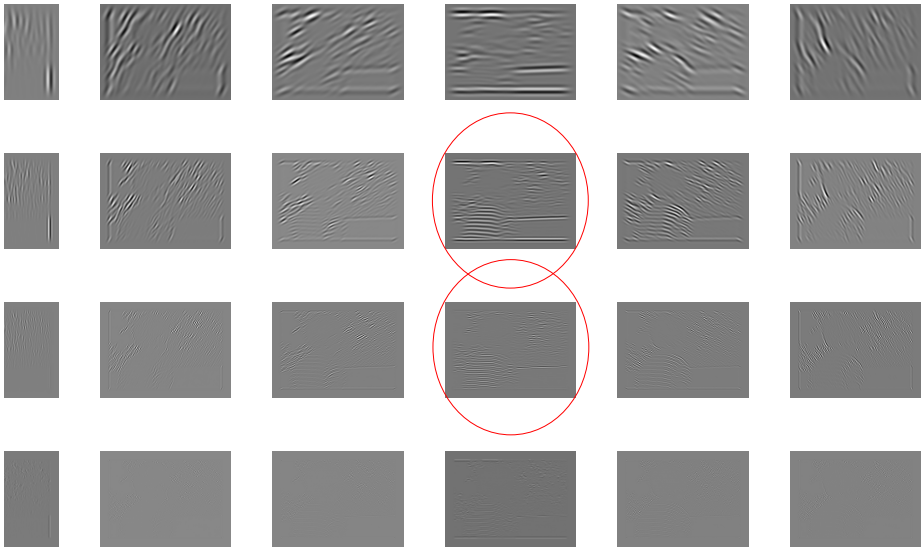


Figura 4.14: Filtro de Gabor - Imagens filtradas.

4.6.2.4 SIFT

As características obtidas utilizando o método SIFT [Lowe04] têm sido utilizadas com sucesso em várias áreas [Nowak06, Lew06, Datta08]. O método inclui duas fases que podem ser utilizadas separadamente: (1) detecção de pontos de interesse e (2) extracção do descritor visual. Na primeira fase, o objectivo consiste em encontrar zonas da imagem que possam ser relevantes e que sejam estáveis em relação às mudanças de escala e rotação. É utilizada uma estratégia baseada em filtros de diferenças de Gaussianas para detectar estes pontos. Na figura 4.15, é apresentada uma imagem com os pontos de interesse calculados por este método. São também apresentados os vectores correspondentes à orientação da região que também são calculados pelo algoritmo. O descritor SIFT é calculado utilizando esta orientação. Em primeiro lugar, é obtido o gradiente em cada ponto numa região de 16x16 pixels em torno de um ponto de interesse (ver figura 4.16a). Depois, em cada bloco de 4x4 pixels, é calculado o histograma com 8 direcções do gradiente (ver figura 4.16b). Cada região é representada por 128 valores (16 regiões de 4x4 pixels vezes 8 pontos do histograma). O descritor apresenta como características mais importantes a invariância à iluminação, rotação, escala e apresenta uma elevada capacidade de distinção [Nowak06, Bosch06]. Esta última propriedade permite pensar no descritor como "palavras visuais" que descrevem imagens tal como palavras descrevem documentos de textos.

4.6.3 Áudio

A informação de áudio é obtida utilizando uma API de reconhecimento de fala para a língua portuguesa, disponibilizada pelo Microsoft Language Development Center em Portugal que utiliza técnicas de análise de áudio para reconhecimento de palavras.

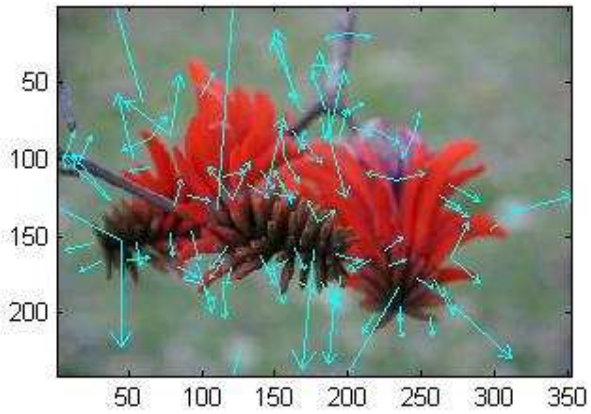
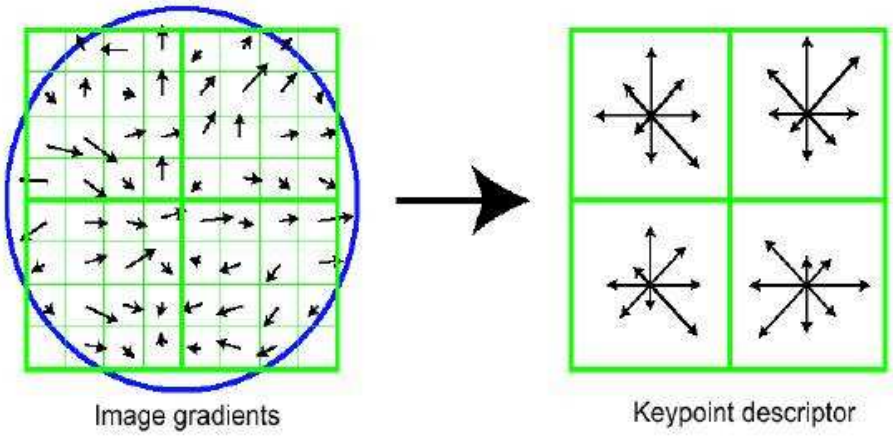


Figura 4.15: SIFT - *Keypoints* detectados.



(a) (b)

Figura 4.16: Exemplo do descritor SIFT numa região de 8x8 pixels: a) Gradiente b) Descriptor.

4.6.4 Metadados Contextuais

Em relação à informação temporal, cada foto é representada pelo instante de captura em segundos. Esta informação é obtida a partir da data e da hora de captura disponível no EXIF do ficheiro da imagem.

A informação de localização é capturada pelo dispositivo de GPS e é também registada no EXIF do ficheiro da imagem. A localização é constituída por três coordenadas, longitude, latitude e altitude mas no método proposto são utilizadas apenas a longitude e a latitude, dado serem suficientes para localizar geograficamente as imagens na metodologia proposta.

4.7 Síntese

O capítulo descreve um método para análise semântica de imagens baseado em informação multimodal. A técnica proposta utiliza a localização geográfica da fotografia, informação temporal, áudio gravado no instante de captura e informação visual para anotar e recuperar imagens com conceitos semânticos. São utilizadas técnicas habitualmente usadas para reconhecimento de áudio, a localização e a data de captura são obtidas no EXIF do ficheiro e são extraídas características visuais de cor e textura. Estas são extraídas de três formas: globalmente utilizando toda a imagem, em regiões e em pontos de interesse. Os próximos capítulos apresentam aplicações de recuperação e anotação que utilizam o método proposto para análise semântica de imagens.

5

Recuperação de Imagens em Ambientes Domésticos

Conteúdo

O capítulo descreve uma aplicação para partilha de experiências pessoais com fotografia em ambientes domésticos. Esta aplicação utiliza o método de recuperação baseado em conceitos semânticos e inclui a técnica proposta para definir interrogações através de uma linguagem visual baseadas em ícones.

5.1 Introdução

A partilha de experiências com fotografias em papel é usualmente realizada em casa com amigos e familiares [Frohlich02]. Nos últimos anos, a fotografia digital tem vindo a ganhar popularidade e, dadas as suas características, a promover diferentes formas de partilhar experiências. A partilha continua a ser realizada maioritariamente em ambientes domésticos, utilizando computadores pessoais de forma semelhante à partilha com fotos em papel [Kim06]. Porém, como é sugerido no estudo publicado em [Lindley06], novas aplicações para partilha de imagens devem ser desenvolvidas que se adaptem aos diversos tipos de utilizadores (por exemplo, idosos ou pessoas com poucos conhecimentos tecnológicos), às divisões da casa com maior actividade social e mantenham os mesmos hábitos utilizados na partilha de fotos em papel (por exemplo, permitir um diálogo frontal entre as pessoas ou permitir um fácil controlo da apresentação quer pelo apresentador quer pela audiência).

No contexto doméstico, a recuperação das experiências depende da forma como são realizadas as tarefas de visualização e de pesquisa ou navegação. No caso da pesquisa, o utilizador utiliza pistas que lhe fazem lembrar a experiência para definir a interrogação e pedir ao sistema imagens de um momento do passado. A visualização permite ao utilizador relembrar a experiência e partilhar com outros. Estas duas tarefas dependem da anotação, isto é, a forma como as imagens são recuperadas e visualizadas depende da informação que é possível anotar em cada imagem, por exemplo imagens anotadas com informação geográfica podem ser recuperadas definindo interrogações geográficas e podem ser visualizadas em mapas. Da mesma forma, uma interrogação através de uma imagem exemplo ou um esboço requer uma anotação baseada no conteúdo da imagem. Assim, para o desenvolvimento de aplicações de uso em ambientes domésticos, para recuperar e visualizar imagens é necessário ter em conta os diversos tipos de interrogação e a respectiva anotação.

Neste capítulo, é descrita a aplicação Memoria Desktop [Jesus07b] para explorar memórias pessoais em ambientes domésticos. Esta aplicação suporta vários tipos de interrogação e por essa razão é proposta uma linguagem visual que permite definir interrogações combinadas. A técnica de interacção escolhida baseia-se no *drag & drop* dos diversos elementos para uma "Query Box". Adicionalmente, a aplicação providencia várias formas de anotar imagens automaticamente. Para além dos metadados de localização e temporais, extraídos automaticamente do EXIF da imagem, é utilizado o conteúdo visual para anotar as imagens com conceitos semânticos e a informação de áudio para anotar com palavras reconhecidas.

Na próxima secção, é apresentada a arquitectura da aplicação proposta e nas secções seguintes são descritos os módulos da aplicação, a interface e o sistema de recuperação e anotação. O capítulo termina com a apresentação dos passos realizados na concepção da aplicação. De realçar que esta última secção apresenta a metodologia de desenvolvimento mas não apresenta os resultados dos testes efectuados com utilizadores. Os resultados da avaliação da interface com utilizadores são descritos no capítulo 8.

5.2 Memórias Pessoais em Ambientes Domésticos

Nesta secção é apresentado o diagrama de blocos da estrutura que serve de suporte ao Memoria Desktop, isto é, o sistema para partilha de experiências pessoais com fotografias digitais em

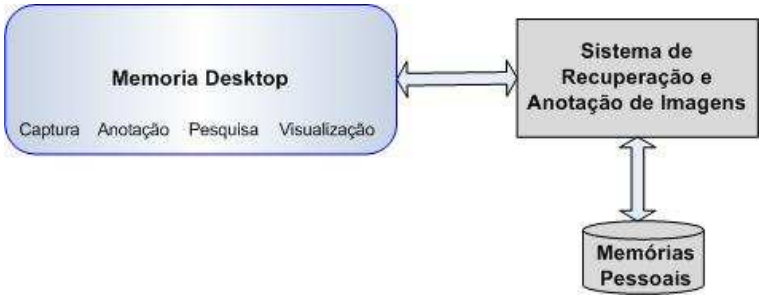


Figura 5.1: Memoria Desktop - aplicação para partilha em ambientes domésticos.

ambientes familiares. Esta aplicação é uma instância do módulo "Anotação e Recuperação" da figura 4.2 e por isso utiliza os modelos semânticos propostos. A figura 5.1 mostra as componentes principais desta aplicação, essencialmente o sistema de recuperação e anotação de imagem e a interface Memoria Desktop. Na interface o utilizador define vários tipos de interrogação e no sistema de recuperação é criada a respectiva lista ordenada com os resultados da pesquisa.

A interface é caracterizada por quatro funcionalidades:

- Captura - para capturar imagens utilizando uma câmara Web. Estas imagens podem ser utilizadas para pesquisar imagens relacionadas;
- Visualização - permite ao utilizador navegar e visualizar imagens da colecção pessoal. As imagens podem ser organizadas por diversos critérios (por exemplo, o nome do ficheiro ou a data de gravação do ficheiro);
- Anotação - para anotar imagens automaticamente com informação obtida a partir de áudio;
- Pesquisa - permite recuperar imagens utilizando uma linguagem para definir interrogações com vários tipos de informação (por exemplo, conceitos, imagens ou partes de mapas).

A funcionalidade mais relevante é a pesquisa de imagens, sendo as restantes tarefas auxiliares da pesquisa. As imagens capturadas podem ser utilizadas na pesquisa, a visualização serve para apresentar os resultados das pesquisas e a anotação é tarefa essencial na recuperação de imagens.

As próximas secções apresentam os dois módulos da estrutura com maior ênfase na interface dado que o sistema de recuperação e anotação é uma aplicação do modelo proposto no capítulo 4.

5.3 Interface

A interface Memoria Desktop permite ao utilizador realizar a gestão das suas experiências pessoais através de fotografias em ambiente familiar. A interface gráfica é dividida em duas secções principais no ecrã (ver figura 5.2): a secção destinada à apresentação de imagens e a secção reservada para a definição da interrogação. Na figura 5.2, a secção de apresentação de imagens localiza-se na parte superior do ecrã e é essencialmente usada para a visualização de imagens resultantes de uma interrogação ou imagens de uma directoria. A área de visualização

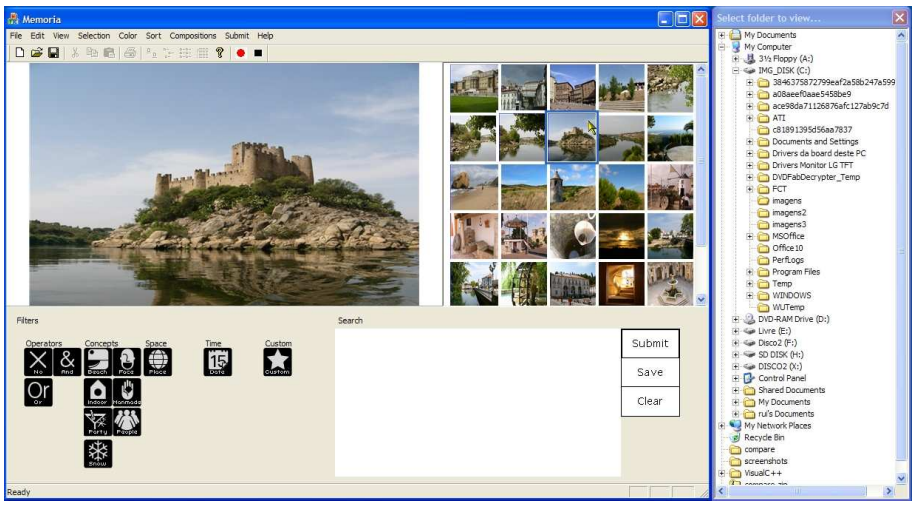


Figura 5.2: Interface do Memoria Desktop.

é composta por uma zona onde é apresentada uma lista de dimensão variável de imagens em tamanho reduzido e por uma outra zona onde é exibida uma imagem de maior dimensão (modo "Preview") que pode ser uma foto capturada pela câmara de vídeo, um "Slideshow" da lista de imagens ou uma versão da imagem seleccionada da lista. A região destinada à definição da interrogação contém uma área reservada a ícones de filtros e operadores e contém a "Query Box" que é a designação atribuída à zona utilizada para arrastar os filtros e os operadores de modo a definir interrogações.

A interface permite recuperar imagens através de pesquisas e navegação na colecção pessoal, permite a visualização de um conjunto de imagens, a anotação com informação de áudio e a captura de memórias utilizando uma câmara. As próximas secções descrevem as acções principais necessárias para realizar estas tarefas.

5.3.1 Captura

A interface permite capturar imagens através de uma câmara. A aplicação serve para partilha de fotos em ambientes domésticos e a inclusão desta função possibilita que o utilizador possa capturar fotos desses momentos. Para além disso, as fotos capturadas podem ser utilizadas para pesquisa de imagens semelhantes. Desta forma, é possível que um grupo de amigos possa relembrar uma experiência do passado onde estiveram juntos, tirando uma ou mais fotos que a seguir podem ser utilizadas como interrogações para recuperar imagens dessa experiência. Esta funcionalidade também possibilita que o utilizador faça pesquisas com objectos físicos de modo a facilitar a recuperação de experiências de utilizadores com poucos conhecimentos tecnológicos. Na figura 5.3 é apresentado um exemplo desta forma de recuperar imagens. O utilizador pretende encontrar fotos de festas e através da colocação de uma garrafa em frente da câmara, é possível recuperar imagens de festas onde aparecem garrafas. A imagem capturada é arrastada para a "Query Box" e enviada para o sistema de recuperação de imagens onde é processada (extracção de características e análise semântica) e é calculada a lista ordenada de imagens semelhantes.

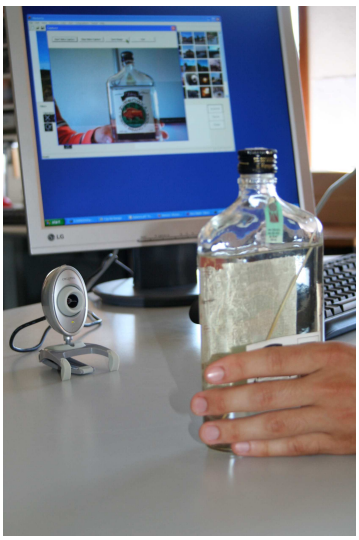


Figura 5.3: Captura com uma câmara Web.

5.3.2 Visualização

Na componente da interface reservada à visualização de imagens é possível apresentar os resultados de uma pesquisa ou as imagens de uma directoria em três modos diferentes:

- "Preview" - permite visualizar com mais detalhe e anotar com áudio (ver secção 5.3.3) uma imagem seleccionada da lista apresentada em modo "List" ou uma foto capturada. Na figura 5.2, é apresentado um exemplo no qual a imagem que aparece em modo "Preview" é a que está seleccionada na lista no lado direito da interface (contida num rectângulo azul);
- "List" - para visualizar um conjunto de imagens da colecção pessoal representadas em tamanho reduzido (ver figura 5.2). Permite mostrar várias imagens em simultâneo ordenadas por diferentes critérios: relevância para a interrogação, nome do ficheiro, data e dimensão do ficheiro;
- "Slideshow" - para apresentar uma sequência de imagens, uma de cada vez em intervalos constantes de tempo. Este modo de visualização pode ser efectuado ocupando todo o ecrã (ver figura 5.4) ou na janela destinada ao modo"Preview". Em ambos os casos o microfone é aberto para gravar possíveis comentários do utilizador que são utilizados para anotação de imagens;

As imagens podem ser visualizadas simultaneamente em dois modos, no modo "List" e no modo "Preview" ou "Slideshow" na janela do "Preview". Os modos "Preview" e "Slideshow" permitem fazer anotação de imagens com áudio e no modo "List" e "Slideshow" é possível visualizar as imagens ordenadas por diferentes critérios.

5.3.3 Anotação com Áudio

A anotação de imagens com palavras é um dos métodos mais utilizados pela maioria das aplicações comerciais (ver capítulo 3). Geralmente, o utilizador é o responsável pela inserção e associação a imagens das palavras que são utilizadas para recuperar imagens através de pesquisas. A nossa proposta nesta aplicação simplifica esta tarefa, permitindo que a anotação

Figura 5.4: Anotação com áudio no modo de visualização "Slideshow".

possa ser realizada através de palavras reconhecidas de áudio gravado pelo utilizador e da extracção automática de características visuais. Nesta secção são abordadas as técnicas que usam áudio. Quando as pessoas partilham as suas fotos com amigos ou familiares em modo "Slideshow" ou da forma tradicional com fotos em papel, geralmente fazem alguns comentários sobre os momentos vividos no passado e relembrados através das imagens [Frohlich02]. Estes comentários podem ser úteis para explicar o momento capturado e o contexto relacionado e assim úteis para a tarefa de anotação de imagens.

Esta aplicação disponibiliza duas técnicas para anotação de imagens:

- Anotação de forma explícita - o utilizador selecciona uma imagem, carrega no botão para gravar áudio e fala para o microfone descrevendo com palavras a imagem. Esta informação é gravada em ficheiro para ser analisada e utilizada para anotação.
- Anotação de forma implícita - quando em modo "Slideshow", o microfone é aberto e todos os comentários feitos são gravados em ficheiro e associados à imagem presente no momento. Na figura 5.4 é apresentada uma ilustração desta técnica.

Os ficheiros de áudio gravados são analisados *a posteriori* através de aplicações de reconhecimento automático de fala que, dado um dicionário, reconhecem palavras que são associadas à imagem. Estas palavras são integradas nos modelos semânticos propostos no capítulo 4 e utilizadas no sistema de recuperação de imagem.

5.3.4 Pesquisa de Imagem

Para recuperar imagens a aplicação oferece duas aproximações de pesquisa: uma baseada em conceitos semânticos e outra baseada na composição de uma imagem através de partes de imagens. Ambas as aproximações utilizam a linguagem visual para definir interrogações proposta nesta tese. Também é possível combinar as duas aproximações para construir interrogações. As interrogações são definidas arrastando (*drag & drop*) para a "Query Box" ícones que representam vários tipos de elementos e operadores da linguagem visual. Conceitos, elementos temporais e espaciais, operadores lógicos e partes de imagens são os elementos disponíveis para construir interrogações utilizando a linguagem visual, tal como descrito de seguida.

5.3.4.1 Linguagem Visual para Definir Interrogações

A linguagem visual para realizar pesquisas baseia-se em elementos das seguintes categorias:

- Imagens - para pesquisar por fotos semelhantes na colecção pessoal (podem ser utilizadas imagens ou partes de imagens que são analisadas com técnicas de processamento de imagem);
- Elementos temporais - para pesquisar por eventos através da data de acontecimento do evento;
- Elementos geográficos - para pesquisar por imagens dada uma localização, por exemplo uma região do mapa;
- Conceitos - para pesquisar por imagens utilizando conceitos baseados em modelos que são treinados automaticamente.

Para combinar estes elementos são utilizados operadores lógicos. Os operadores disponíveis são definidos pelo conjunto,

$$L_{operadores} = \{AND, OR, NOT\}. \tag{5.1}$$

O operador AND expressa a conjunção ou a intersecção e o operador OR designa a disjunção ou união. O operador NOT representa a negação e permite definir o conceito oposto.

O utilizador pode escolher os conceitos relevantes para a sua pesquisa entre o universo pré-estabelecido pelo conjunto de conceitos,

$$Conceitos = \{Beach, Face, Indoor, ManMade, Party, Snow, People\}. \tag{5.2}$$

Utilizando o operador NOT é definido o conjunto de conceitos,

$$NOTConceitos = \{NoBeach, NoFace, Outdoor, Nature, NoParty, NoSnow, NoPeople\}. \tag{5.3}$$

A localização da experiência e o instante temporal em que ocorre representam informação relevante na recuperação de memórias pessoais, por isso são também definidos elementos temporais pertencentes ao conjunto de todos os intervalos de tempo possíveis no contexto da colecção pessoal,

$$Intervalos_{Temporais} = \{T_{int_1}, T_{int_2}, , ...T_{int_{Nint}}\}, \tag{5.4}$$

e o conjunto de todas as regiões possíveis dentro das localizações disponíveis na colecção pessoal,

$$Regioes = \{R_1, R_2, , ...R_{Mreg}\}. \tag{5.5}$$

Imagens ou partes de imagens são obtidas a partir das imagens do repositório pessoal, C_{img}.

Recorrendo a estes elementos e operadores é possível produzir interrogações como, por exemplo, NOT Indoor AND NOT ManMade. Ao definir esta interrogação o utilizador pretende encontrar imagens do exterior com elementos da natureza (Nota: NOT ManMade = Nature and NOT Indoor = Outdoor).

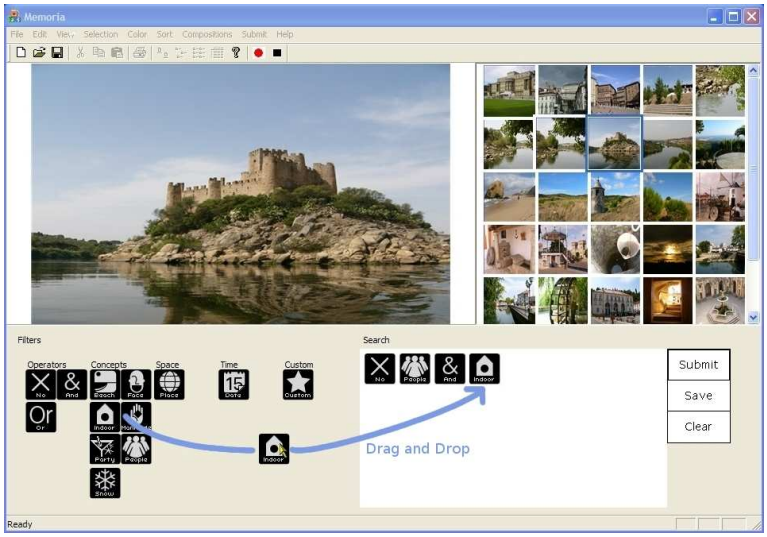


Figura 5.5: *Drag & drop* para a "Query Box".

A combinação destes operadores de diferentes tipos permite definir interrogações mais adequadas à recuperação de memórias pessoais, uma vez que existem elementos que representam as pistas que geralmente a memória humana utiliza [Endel02]. As interrogações são definidas arrastando ícones que representam os elementos apresentados para a "Query Box" (ver figura 5.5).

5.3.4.2 Pesquisa por Conceitos Semânticos

Para pesquisar imagens utilizando a interrogação através de conceitos, o utilizador formula a interrogação escolhendo de entre um conjunto pré-definido os conceitos que são relevantes para a pesquisa. Por exemplo, se a selecção consistir em Outdoor AND No People AND Nature, o sistema retorna o conjunto de imagens com maior probabilidade de acordo com a combinação dos conceitos indicados na interrogação. As imagens resultantes são apresentadas ao utilizador e podem ser utilizadas para definir outro tipo de interrogação.

Ao contrário das aplicações semelhantes não é necessário digitar nenhuma palavra para construir uma interrogação. O utilizador arrasta os elementos para a "Query Box" e clica no botão "submit" para enviar a interrogação para o sistema de recuperação de imagem. Em qualquer altura, os items presentes na "Query Box" podem ser reordenados ou apagados. Esta solução é mais simples e rápida do que a introdução de palavras mas o utilizador está limitado ao conjunto de conceitos existente. A figura 5.5 mostra uma exemplo de uma interrogação através de conceitos. Os conceitos "People" e "Indoor" e os operadores NOT e AND estão na "Query Box" para pesquisar por imagens interiores (dentro de casa) com pessoas. Não existem precedências e a interrogação é analisada da esquerda para a direita. Sempre que o utilizador não coloca um operador entre conceitos é utilizado o operador AND por omissão.

5.3.4.3 Pesquisa por Composição

A interrogação através da composição de uma imagem com partes de várias imagens é outro método disponível na aplicação para recuperar memórias (ver figura 5.6). Com este tipo de interrogação, o utilizador pesquisa por imagens com características visuais semelhantes aos

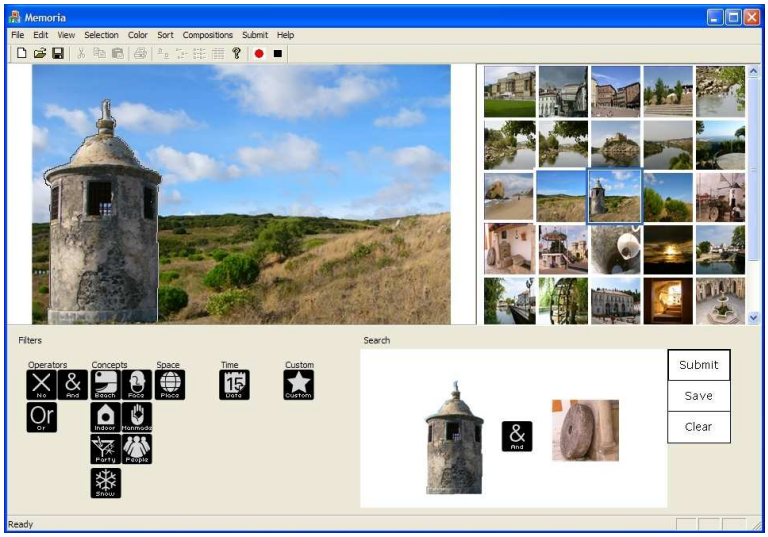


Figura 5.6: Pesquisa por composição de uma imagem.

objectos que foram seleccionadas para compor a interrogação. Este tipo de interrogação funciona de forma semelhante à interrogação por esboço [Buijs99] mas a imagem exemplo é definida usando partes de imagens. Esta técnica tem como objectivo ajudar o utilizador que não consegue definir com conceitos a pesquisa que pretende e também não tem disponível uma imagem que descreva o que pretende encontrar. Desta forma, o utilizador pode compor o tipo de imagens que pretende procurar.

As partes são seleccionadas na imagem presente na janela de "Preview" e podem ter forma rectangular ou podem ser desenhadas à mão livre. As várias partes são organizadas na "Query Box" onde é efectuada a composição da imagem. A aplicação também permite gravar uma composição para reutilização. Existe um menu "Compositions" com uma lista das composições existentes. O utilizador pode seleccionar uma composição da lista ou construir uma nova para recuperar imagens da base de dados.

Depois de formulada e submetida, a imagem composta é enviada para o sistema de recuperação onde são extraídas as palavras visuais e é construída a lista ordenada com imagens semelhantes. Na figura 5.6, é apresentado um exemplo de pesquisa por composição. Duas partes de imagens e o operador AND foram arrastados para a "Query Box" para pesquisar imagens semelhantes.

5.4 Sistema de Recuperação de Imagens

A interface descrita neste capítulo permite pesquisar na base de dados através de dois métodos: pesquisa por conceitos e pesquisa por composição. O primeiro método utiliza mais informação para fazer a pesquisa porque, em geral, os conceitos são treinados com centenas de imagens. O segundo, apesar de dar mais liberdade ao utilizador para personalizar a interrogação, fornece pouca informação ao sistema, que apenas tem uma imagem para analisar. Assim, é difícil ao sistema interpretar a imagem e o que o utilizador pretende através de um vector de características (problema conhecido como falha semântica).

A pesquisa por conceitos é baseada nos conceitos semânticos estimados utilizando o modelo semântico proposto na secção 4.3. A pesquisa por composição é baseada na extracção

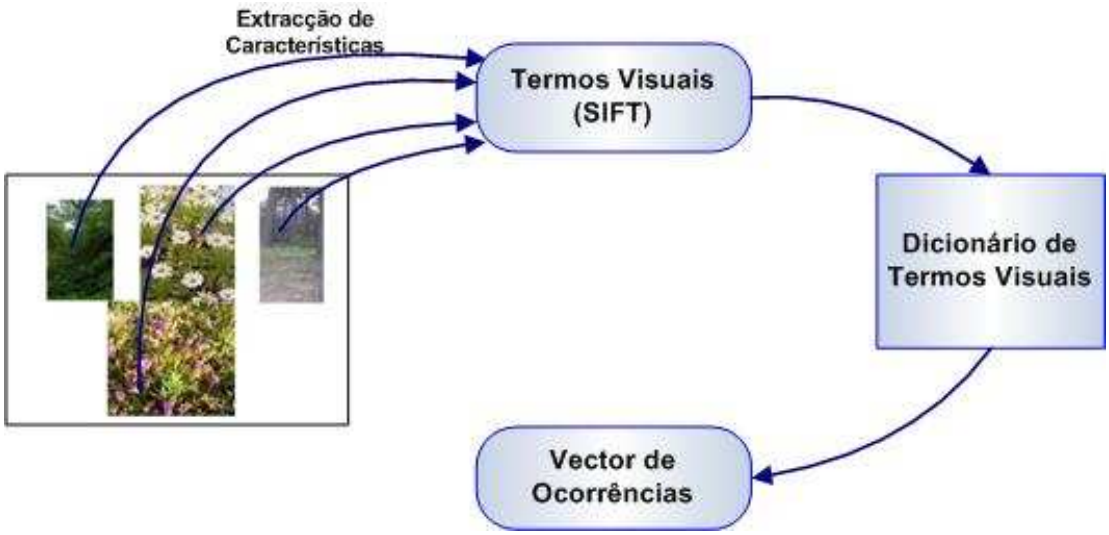


Figura 5.7: Vector de ocorrências de termos visuais.

automática de características visuais (ver secção 4.6) nas partes de imagens que compõem a interrogação. A interrogação é representada por um vector de ocorrências de termos visuais pertencentes a um vocabulário (ver figura 5.7). É aplicado o método LSA e é utilizada a medida de coseno para construir a lista ordenada com as imagens relevantes, como é explicado na secção 4.6.

5.5 Concepção

A aplicação Memoria Desktop tem as mesmas funcionalidade que a aplicação Memoria Mobile (ver capítulo 6) mas adaptadas para um computador pessoal e para serem utilizadas em ambientes domésticos. Uma vez que a aplicação Memoria Mobile foi desenvolvida em primeiro lugar, a experiência obtida no desenvolvimento desta aplicação permitiu avançar mais rapidamente nos passos iniciais na concepção da aplicação Memoria Desktop, apesar da interacção em computador pessoal e o contexto de utilização serem diferentes. Assim, não houve necessidade de repetir os passos iniciais da definição de funcionalidades e dos protótipos em papel. Foram desenvolvidos os ecrãs utilizando o Adobe Photoshop que depois de refinados em várias iterações, deram origem a um protótipo inicial de alta fidelidade. Este protótipo foi submetido a testes de usabilidade e refinado até ao protótipo final. Uma descrição mais completa da metodologia de concepção e dos testes efectuados e dos seus resultados é apresentada no capítulo 8.

5.5.1 Análise e Protótipos de Alta Fidelidade

A fase de análise e definição das funcionalidades foi encurtada pelo trabalho que já tinha sido feito na concepção do Memoria Mobile. Em geral, as funcionalidades da aplicação móvel, incluindo a recuperação com diversos tipos de elementos, a anotação com áudio, a visualização e a captura mantêm-se na aplicação em ambientes domésticos (como foi referido anteriormente). A técnica de arrasto de elementos para uma "Query Box" também se mantém, sendo a interacção em computador mais facilitada. Os modos de visualização de imagens são idênticos nas duas interfaces, com a vantagem de haver mais espaço na interface Memoria Desktop.

Desta forma, é possível visualizar mais informação em simultâneo. Há pequenas diferenças nos conceitos de captura e de anotação. No caso da captura, o contexto é limitado ao ambiente familiar (por exemplo, quarto ou sala) e é utilizada para pesquisa com objectos físicos. A anotação, apesar de ser feita no Memoria Desktop no modo "Slideshow", mantém a mesma filosofia da anotação no instante de captura realizada no Memoria Mobile. Feita esta análise, foram criados vários ecrãs da aplicação em Adobe Photoshop que foram refinados utilizando o resultado de avaliações heurísticas [Nielsen90]. A partir destes protótipos de alta fidelidade foi desenvolvido o protótipo em computador pessoal.

5.5.2 Testes de Usabilidade

O protótipo em computador foi submetido a testes de usabilidade com o objectivo de avaliar as decisões tomadas nas fases anteriores da concepção da interface. Decidimos realizar estes testes com um número elevado de utilizadores (58) para permitir realizar uma análise mais completa da aplicação. Estes testes foram guiados por um questionário onde é pedido ao utilizador para realizar determinada tarefa e dar a sua opinião sobre a experiência. O questionário inclui uma parte de caracterização do perfil do utilizador, nomeadamente, dados pessoais e questões sobre a forma como gere as suas memórias pessoais. Depois, é pedido ao utilizador para realizar quatro tarefas e para responder a várias questões relacionadas com aplicação, nomeadamente, o aspecto visual da interface, a utilidade da aplicação, se é simples de usar e se os métodos propostos para recuperar memórias são úteis e fáceis de usar, principalmente a linguagem visual para definir interrogações (mais detalhes no capítulo 8).

5.6 Síntese

Neste capítulo, foi apresentada uma proposta para partilha de memórias pessoais compostas por imagens em ambientes familiares. Os aspectos mais relevantes são a linguagem visual para realizar pesquisas com combinações de diferentes tipos de informação e os dois modos de anotação de imagens com informação extraída de áudio. De salientar também a funcionalidade de captura, que permite a interacção com a colecção pessoal através de objectos físicos e que é parte de uma interface tangível a desenvolver no futuro para relembrar memórias pessoais em casa (ver perspectivas futuras no capítulo 9). O objectivo desta aplicação será tornar o computador invisível, utilizando o televisor ou outro ecrã para visualização e um objecto de decoração para integrar a câmara e servir para aceder à colecção pessoal com objectos que façam recordar eventos do passado.

6

Recuperação de Imagens em Locais de Interesse

Conteúdo

O capítulo apresenta soluções para partilha de fotografias em locais de interesse baseada no sistema de recuperação de imagens proposto. São descritas duas aplicações de pesquisa de imagens baseadas em conceitos semânticos, uma aplicação para partilha no momento da experiência e outra para a Web para complementar a visita.

6.1 Introdução

Durante os últimos anos, os avanços na tecnologia móvel têm contribuído para melhorar o processo de captura, partilha e armazenamento de memórias pessoais constituídas por fotografias digitais. Actualmente, está à disposição de todos um conjunto de dispositivos móveis com uma elevada capacidade computacional, com câmaras fotográficas de qualidade incorporadas, com receptores de GPS integrados, com a possibilidade de utilizarem cartões de memória com elevada capacidade de armazenamento e com várias tecnologias de comunicação sem fios (por exemplo, GPRS, Bluetooth e Wi-Fi). Em geral, é possível qualquer indivíduo capturar fotos ou vídeos em qualquer lugar e depois partilhar esta informação através da World Wide Web utilizando, por exemplo, o Flickr [Flickr04] ou o YouTube [Youtube05]. A partilha da experiência pode ser realizada na presença da pessoa, quando encontra um amigo ou familiar em qualquer local utilizando o dispositivo móvel ou, à distância, através da Web. As aplicações na Web para partilha à distância permitem ainda a partilha de fotos ou vídeos com desconhecidos, o Flickr e o YouTube são dois exemplos de sucesso dessas aplicações. A quebra de privacidade pode ser uma desvantagem desta estratégia mas, em compensação, é possível melhorar ou complementar experiências idênticas no futuro, por exemplo, a experiência de uma pessoa que pretende visitar um museu ou outro ponto de interesse qualquer. Aliás, visitas a museus ou a lugares históricos ou outras actividades de lazer estão entre as situações onde a maioria das fotos pessoais são obtidas. Neste contexto e dadas as potencialidades dos dispositivos móveis, há duas questões que se colocam:

- Haverá alguma vantagem em partilhar as fotos no momento da experiência ou seja, durante a visita ao local de interesse?
- Será que as pessoas estão dispostas a partilhar fotografias nestas situações?

Em relação à primeira questão, uma colecção melhorada de fotografias da experiência e mais informação sobre local que estão a visitar são motivos vantajosos, contudo, existe também a possibilidade desta informação distrair o visitante. Na segunda questão, para além das questões de privacidade, é expectável que os visitantes estejam na disposição de partilhar fotos com desconhecidos uma vez que isso já é prática comum na Web.

Este capítulo descreve a aplicação Memoria Mobile [Jesus06a, Dias07, Jesus08a], a nossa proposta para partilhar imagens nas condições referidas. É uma aplicação para PDAs (Personal Digital Assistant) para capturar, partilhar e aceder a memórias pessoais compostas por fotos no momento da visita a um ponto de interesse. É também apresentada a interface Memoria Web que permite conhecer o local antes da visita e complementar a experiência após a visita.

Na próxima secção, é descrita a arquitectura geral do sistema e nas secções seguintes são descritas as duas interfaces.

6.2 Sistema de Partilha para Locais de Interesse

O sistema proposto para partilha de memórias pessoais de um local de interesse [Jesus07] é baseado na arquitectura cliente/servidor (ver figura 6.1). A arquitectura é composta pelo sistema de recuperação de imagens (ver capítulo 4) no servidor e por dois clientes (duas aplicações de recuperação de imagem): Memoria Mobile e Memoria Web. O servidor também inclui o

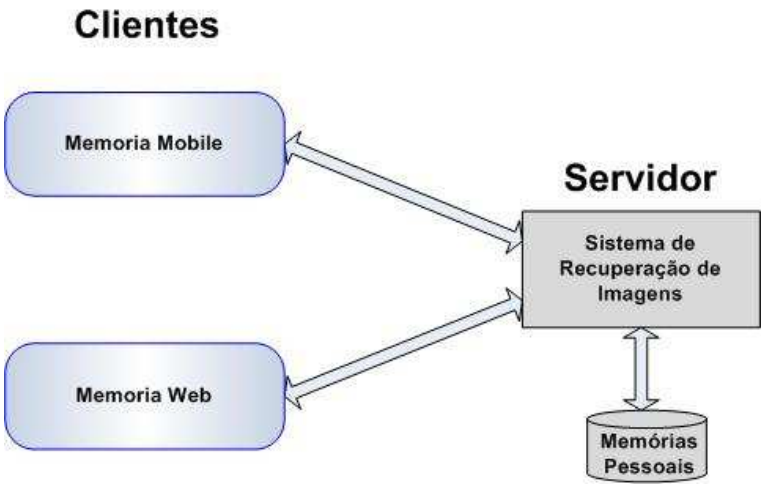


Figura 6.1: Arquitectura do sistema de partilha de fotos de um local de interesse.

conjunto de memórias constituído pelas imagens disponíveis para partilhar com os visitantes do local. A opção por esta arquitectura permite centralizar o conjunto de memórias a partilhar pelos clientes e possibilita a execução dos algoritmos de recuperação de imagem numa plataforma com mais recursos computacionais.

O cliente móvel permite capturar, anotar, recuperar imagens do conjunto de memórias e visualizar os resultados. Em qualquer altura da visita, quando o utilizador tira uma foto as coordenadas geográficas (obtidas por GPS) do local onde é capturada e a informação áudio relacionada com possíveis comentários, providenciados pelo utilizador, são automaticamente anotados na imagem. A aplicação Memoria Mobile acede às memórias do local através de uma ligação sem fios (WLAN ou GPRS) ao servidor. Os visitantes podem aceder as estas memórias (imagens) definindo interrogações na aplicação móvel e submetendo-as ao servidor que inclui o sistema de recuperação de imagens. O visitante pode pesquisar imagens por semelhança, proximidade ou de acordo com um contexto definido pelo utilizador (por exemplo, imagens com pessoas ou edifícios).

As pesquisas por semelhança incluem o envio de uma imagem para o servidor. As imagens enviadas que não pertencem ao conjunto de memórias, por exemplo uma imagem tirada pelo utilizador, são incluídas na base de dados. Desta forma, o visitante submete ao servidor imagens para partilhar com outros visitantes e em troca recebe mais imagens do local que está a visitar. Esta estratégia permite que o conjunto de memórias seja construído por todos os visitantes do local que estejam dispostos a partilhar as suas experiências. Os resultados das pesquisas enviados pelo servidor também podem ser usados pelos visitantes para melhorar a visita.

A aplicação Memoria Web permite realizar mais funções que a aplicação móvel mas inclui o mesmo tipo de pesquisas. O objectivo é criar uma rede social com visitantes de um ponto de interesse. A aplicação permite que um utilizador possa explorar o local antes da visita e que possa complementar a experiência depois da visita. O utilizador pode copiar imagens para o servidor, pode anotar as suas imagens com texto ou coordenadas GPS ou construir a história da sua visita com imagens, vídeos e texto. No âmbito do trabalho proposto nesta tese apenas estão incluídas as funções de anotação e de pesquisa de imagens por isso, na secção dedicada a esta aplicação, apenas são apresentadas estas funcionalidades.

6.3 Memoria Mobile

A aplicação Memoria Mobile [Dias07] é constituida por quatro funcionalidades principais:

1. Máquina fotográfica - para registar a experiência pessoal;
2. Anotação automática - anotação de imagens no instante de captura com informação de localização e de áudio;
3. Recuperação de imagens - para aceder ao conjunto de memórias partilhadas;
4. Visualização de imagens - para visualizar as fotos capturadas e os resultados das pesquisas.

A interface desenvolvida para realizar estas funcionalidades é constituída essencialmente por (ver figura 6.2):

- Uma barra com o nome e um menu para executar funcionalidades (canto superior esquerdo da interface);
- Uma barra com indicação do estado, por exemplo, bateria ou rede;
- Uma barra de navegação para alternar entre modos de visualização e entre imagens pessoais ou todas as imagens (canto superior direito);
- Filtros (por exemplo, conceitos e direcções);
- "Query Box", espaço para definir interrogações (canto inferior esquerdo).

A interface inclui também um espaço reservado à visualização de imagens (zona central da interface) e vários botões. O botão "More Filters" mostra todos os filtros disponíveis para fazer pesquisas na base de dados. A "Query Box" é utilizada para definir interrogações através do arrasto para este espaço de filtros. Foi escolhida esta forma para definir interrogações para pesquisar por imagens porque permite combinar e seleccionar vários tipos de elementos (por exemplo, imagens, texto ou botões). Quatro tipos de elementos podem ser arrastados para a "Query Box" [Dias07]:

- Imagens - para recuperar imagens semelhantes;
- Regiões do mapa - para pesquisar por imagens localizadas;
- Direcções - para procurar imagens numa direcção em relação à posição do utilizador ("Norte", "Sul", "Este", e "Oeste");
- Conceitos semânticos - para recuperar imagens de acordo com interesses específicos, por exemplo, procurar por fotos com objectos feitos pelo homem ou fotos que incluem pessoas.

Outros tipos de interrogação podem ser efectuados através da combinação destes elementos, por exemplo pesquisar por imagens semelhantes (utilizando o conteúdo da imagem) na base de dados mas limitar a pesquisa a uma região incluindo uma zona do mapa na interrogação. A seguir, são descritas as principais funções da interface.

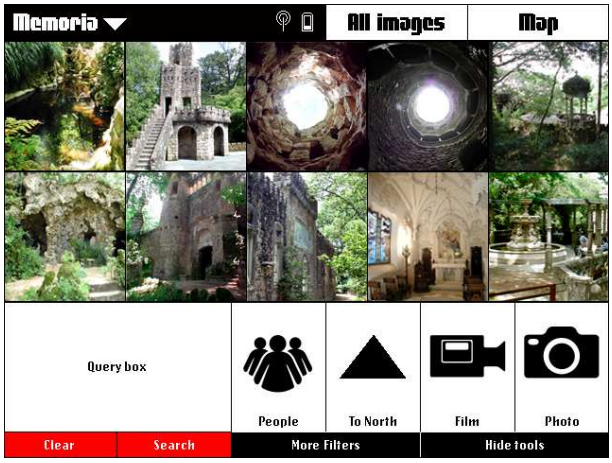


Figura 6.2: Aplicação Memoria Mobile: visualização de uma lista de imagens (modo "Grid").

6.3.1 Captura

A aplicação cliente, que é executada no dispositivo móvel, pode ser utilizada como uma máquina fotográfica normal com a vantagem de capturar as coordenadas geográficas obtidas por GPS e informação de áudio no instante de captura da foto. Para obter a fotografia, o utilizador pode usar o botão que existe fisicamente no dispositivo móvel, semelhante ao de uma máquina fotográfica e que pode ser utilizado por qualquer outra aplicação, ou o botão da aplicação Memoria Mobile situado no canto inferior direito. Na figura 6.2, o botão com uma máquina fotográfica desenhada serve para seleccionar o modo de captura. Também existe um botão para captura de vídeo. Durante o desenvolvimento da aplicação, o critério mais relevante na integração desta funcionalidade foi o de permitir que o utilizador capture o momento da mesma forma que faria se estivesse a utilizar uma máquina fotográfica. A informação de localização e de áudio que são obtidos no instante de captura são gravados em ficheiro e enviados para o servidor para análise e anotação semântica da imagem.

6.3.2 Visualização

A visualização de fotos é a funcionalidade que permite visualizar e partilhar as fotos utilizando três modos de apresentação de imagens, dois que existem normalmente na maioria das aplicações para gerir fotos e um terceiro que permite a visualização espacial de fotos. Os três modos de visualização são:

- Modo "Slideshow", permite apresentar um conjunto de imagens sequencialmente;
- Modo "Grid", permite visualizar uma lista de imagens organizadas numa grelha (ver figura 6.2);
- Modo de visualização no mapa, permite visualizar um conjunto de imagens no mapa do local (ver figura 6.3).

As imagens visualizadas podem ser resultados de pesquisas ou imagens de uma directoria. Também é possível seleccionar entre todas as imagens de uma directoria apenas as imagens tiradas pelo utilizador. O modo de visualização no mapa dá ao visitante um contexto espacial porque é apresentada a posição do utilizador, o caminho percorrido e as fotos nas localizações

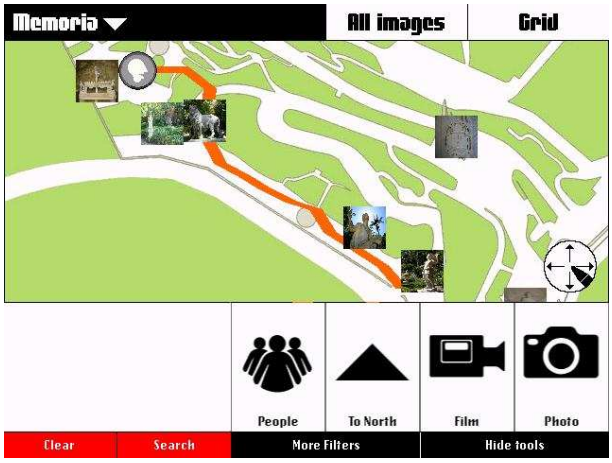


Figura 6.3: Percurso realizado pelo utilizador durante a visita.

onde foram obtidas durante o percurso da visita. Este modo de visualização permite exibir simultaneamente informação espacial, através da apresentação no mapa das imagens nos locais de captura, e informação visual contida nas próprias imagens. As figuras 6.3 e 6.4 exemplificam a utilização do modo de apresentação de imagens no mapa. Na figura 6.3, são mostradas as imagens tiradas pelo utilizador durante o percurso efectuado bem como o percurso e a posição do utilizador. Na figura 6.4, são apresentados os resultados de uma pesquisa. De notar que o modo de visualização no mapa também facilita a definição de interrogações geográficas (ver secção 6.3.3).

6.3.3 Recuperação de Imagens

A aplicação Memoria Mobile permite fazer pesquisas de imagens partilhadas no conjunto de fotos presentes no servidor. No dispositivo móvel são definidos os pedidos do utilizador e visualizados os resultados das pesquisas. O sistema de recuperação de imagens é executado no servidor para onde são encaminhados os pedidos dos clientes móveis. Do lado do cliente, podem ser definidas interrogações utilizando quatro tipos de elementos referidos anteriormente.

Para definir a interrogação, é utilizada uma versão simplificada da linguagem visual definida na aplicação Memoria Desktop (ver capítulo 5). As interrogações são definidas arrastando para a "Query Box" ícones que representam os quatro tipos de elementos referidos anteriormente. Na interface móvel não se incluem os operadores lógicos para combinar os diversos elementos, simplificando assim a linguagem visual por causa das dimensões do ecrã. Uma vez que também por esta razão se limita o número de imagens a visualizar resultantes das pesquisas e o número de elementos visíveis para definir interrogações, não era coerente ocupar mais espaço com os ícones dos operadores lógicos. Desta forma, a linguagem visual para definir interrogações com ícones é composta apenas por um operador, o "AND" que é o operador utilizado por omissão e por isso não é necessário incluí-lo na interrogação. O operador "AND" é utilizado para combinar os resultados das interrogações geográficas (direcções e regiões) com as interrogações visuais (imagens e conceitos). A lista ordenada de imagens é obtida utilizando a equação 4.1. São somadas as listas obtidas a partir da informação visual e de áudio e depois é aplicada a função f_{gps} para seleccionar um conjunto de imagens. Esta função representa a aplicação do operador "AND" entre o subconjunto de imagens seleccionadas pela informação geográfica e o conjunto de todas as imagens.

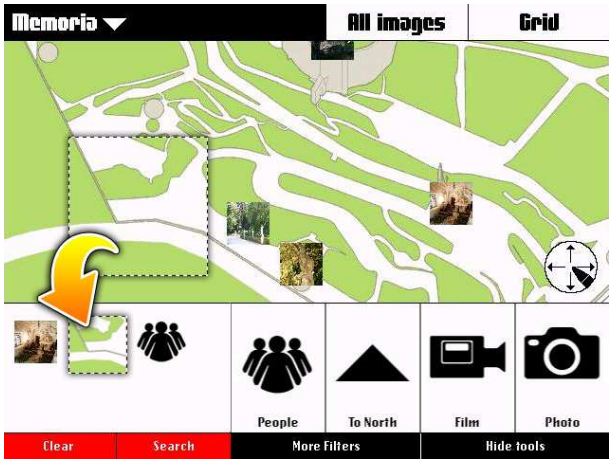


Figura 6.4: Definição de interrogação através do arrasto para a "Query Box" de imagens, conceitos, direcção ou regiões de mapas.

Na figura 6.4, é apresentado um exemplo de uma interrogação, definida com um conceito, uma imagem e uma região do mapa. O objectivo é pesquisar por imagens semelhantes visualmente e que estejam dentro da região definida. Neste exemplo seriam imagens da capela com pessoas.

Depois de definida e submetida a interrogação para o servidor, a lista ordenada com os resultados é obtida utilizando o sistema de recuperação de imagens proposto no capítulo 4. No caso da interrogação ser composta apenas por elementos geográficos, a lista de imagens é ordenada utilizando a equação 4.6. Para a interrogação com uma região do mapa são consideradas as imagens que estão dentro do círculo definido pela região (ver equação 4.7). Na interrogação definida por uma direcção são seleccionadas as imagens que estão na direcção em relação a um ponto indicado. Dadas as coordenadas geográficas do ponto onde se encontra o utilizador e uma direcção ("Norte", "Oeste", "Sul" e "Este"), o conjunto de imagens na direcção indicada é obtido comparando as coordenadas geográficas das imagens e do ponto indicado, tal como é descrito na secção 4.3.3.

Quando a interrogação não inclui elementos geográficos, a lista ordenada a devolver ao cliente é obtida aplicando a equação 4.1 sem aplicar a função f_{gps}. No caso particular da interrogação incluir imagens exemplo, é calculada uma lista ordenada de acordo com a distância entre o vector de características (ver secção 4.6) das imagens exemplo e os vectores de características das imagens da base de dados.

6.3.4 Anotação de Imagens

No instante em que a foto é capturada também são obtidas e gravadas em ficheiro as coordenadas do local onde é tirada. Durante alguns segundos após a captura da foto, a aplicação liga o microfone para poder gravar comentários efectuados pelo utilizador. Assim, quando uma fotografia é obtida, a informação visual (a imagem), as coordenadas geográficas e a informação de áudio são gravados em ficheiros associados. Quando a imagem é enviada para o servidor, o conteúdo visual da imagem e o ficheiro de áudio são analisados para providenciar a anotação da imagem como é descrito na secção 4.4. No caso do conteúdo visual, são extraídas as características e é efectuada a análise semântica. Em relação ao áudio, são utilizadas aplicações de reconhecimento de fala para converter o áudio em texto e assim providenciar anotações

semânticas baseadas em palavras reconhecidas.

O áudio obtido nas condições referidas pode não ter muita qualidade porque pode ser gravado em ambientes com ruído. Contudo, como é informação adicional, qualquer informação reconhecida sem erros permite melhorar o desempenho do sistema.

6.3.5 Concepção

A metodologia utilizada para conceber a aplicação baseou-se inicialmente nos conhecimentos e estudos prévios de interacção pessoa máquina em dispositivos móveis [Dix00], em estudos etnográficos [Frohlich02], em informação sociológica relacionada com estudos em museus [Levasseur83], em estudos referentes à utilização da tecnologia móvel em turismo [Brown03] e, mais em particular, em estudos relacionados com aplicações de pesquisa de fotos em dispositivos com ecrã de dimensões reduzidas [Patel06]. Para além disso, também foi útil a experiência adquirida no desenvolvimento da plataforma do projecto InStory [Correia05] num local cultural como a Quinta da Regaleira em Sintra, Portugal.

Esta informação e os cenários criados foram representados através de informação textual e de imagens. Depois de definidas algumas funcionalidades foram criados esboços da interface em computador que refinados heuristicamente deram origem a protótipos em papel, para avaliar e optimizar a aplicação proposta em testes com utilizadores. Ao mesmo tempo, também foi avaliado um protótipo em PDA para testar as características difíceis de avaliar com protótipos em papel. Com a informação resultante, o protótipo em PDA foi refinado até chegar à versão final apresentada nas secções anteriores. A seguir são apresentados mais detalhes do processo de concepção da aplicação Memória Mobile [Jesus08a].

6.3.5.1 Estudos de Campo

No âmbito de um projecto anterior, InStory [Correia05], foram realizadas várias visitas à Quinta da Regaleira um lugar histórico e cultural em Sintra, Portugal, local onde também foi testada a aplicação Memoria Mobile. Estas visitas permitiram obter uma interpretação significativa dos problemas que caracterizam as visitas a lugares de lazer desconhecidos. Por exemplo, um dos problemas dos visitantes da Quinta da Regaleira, uma propriedade do século XIX com jardins, caves, torres, igrejas e um palácio, é a dificuldade em encontrar algumas direcções e alguns lugares importantes que não estão representados em forma de fotografia no mapa do local. As pessoas têm dificuldade em se localizar nos imensos jardins e acabam por perder alguns espaços valiosos do local. Esta informação e os cenários criados foram recolhidos na forma de descrições textuais e esboços com imagens (*storyboards*) dos percursos realizados pelos visitantes.

6.3.5.2 Protótipos de Alta Fidelidade

Depois do estudo de campo, foram definidas um conjunto de funcionalidades e interações relacionadas que levaram à construção de vários esboços da interface de alta fidelidade utilizando o Adobe Photoshop. Embora estes esboços sejam geralmente reservados para as fases finais, estes protótipos iniciais sem interacção mas com alguns detalhes visuais são úteis por várias razões: reduzem a complexidade, permitem fazer avaliações heurísticas [Nielsen90], melhoram a definição dos protótipos em papel e permitem melhorar os aspectos visuais da aplicação de

(a) (b)

Figura 6.5: Protótipos em papel

forma interactiva. Estes esboços permitiram recolher dados de várias avaliações heurísticas e permitiram ficar com uma melhor percepção da aplicação a desenvolver. A partir destes esboços da interface não funcionais, foram desenvolvidos os protótipos de baixa fidelidade em papel e um protótipo de alta fidelidade em PDA destinado a testar a técnica de usabilidade proposta (*drag & drop*) para definir interrogações.

6.3.5.3 Protótipos em Papel e em PDA

No sentido de validar e refinar as opções tomadas nas fases anteriores de desenvolvimento da interface, foram realizados testes com protótipos em papel (ver figura 6.5) e protótipos iniciais em PDA com utilizadores de perfis diferentes. Em geral, os protótipos de baixa fidelidade permitiram testar as sequências de acções necessárias para cumprir as tarefas pedidas enquanto que o protótipo em PDA tinha como objectivo avaliar a técnica de *drag & drop*. Foi necessário produzir um protótipo inicial em PDA dado que constatamos que era díficil avaliar o *drag & drop* com protótipos em papel. Este protótipos foram testados com vários utilizadores de diferentes perfis sequencialmente, isto é, em primeiro lugar o utilizador usava os protótipos em papel e depois testava o protótipo em PDA. Os utilizadores foram entrevistados antes e depois dos testes e em qualquer momento podia fazer comentários em relação à aplicação. Os protótipos finais em PDA foram obtidos analisando esta informação e os vídeos dos testes que foram gravados. Os resultados dos testes de usabilidade e mais detalhes sobre o procedimento adoptado podem ser consultados no capítulo 8.

6.4 Memoria Web

Para complementar a aplicação móvel foi desenvolvida uma aplicação para a Web, designada por Memoria Web [Mweb08]. O objectivo desta aplicação é criar uma rede social para que os visitantes de um ponto de interesse possam partilhar as suas experiências e enriquecer as visitas de futuros visitantes. Esta aplicação inclui várias funcionalidades, a maior parte fora do contexto desta tese, por exemplo, a construção de histórias. Por esta razão, esta interface não é apresentada em detalhe como a aplicação Memoria Mobile. São apresentadas as funcionalidades referentes à anotação e recuperação de imagens.

No que diz respeito à anotação e recuperação de imagens, a aplicação Web providencia funções semelhantes às funções descritas para a aplicação móvel. O utilizador pode copiar para o

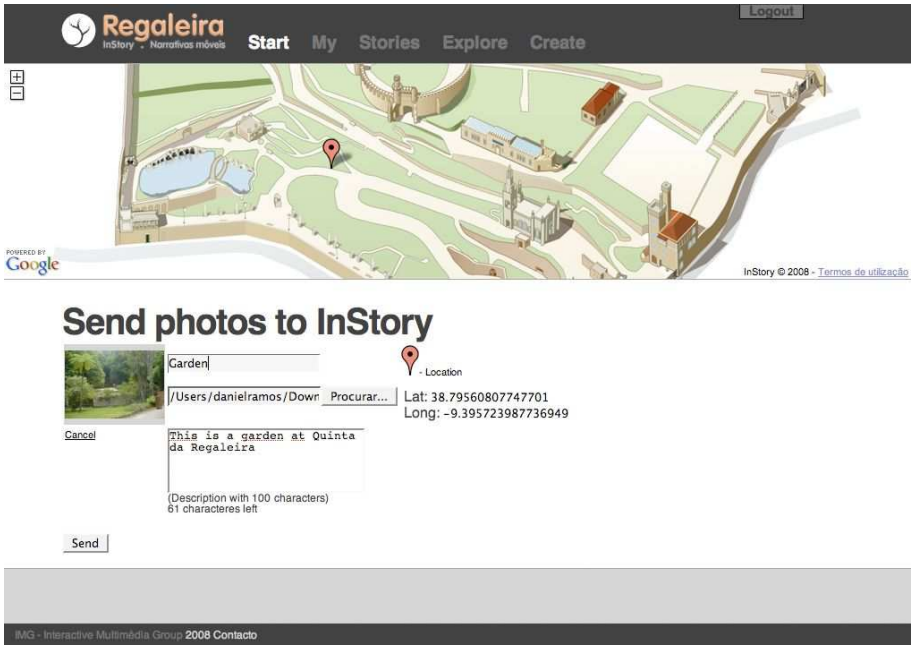


Figura 6.6: Memoria Web - anotação de imagens com coordenadas GPS.

servidor fotos da Quinta da Regaleira e nesta fase pode proceder à sua anotação mas manualmente (ver figura 6.6), isto é, o utilizador pode fazer descrições textuais que são usadas para anotação. Também pode anotar a imagem com as coordenadas geográficas do local representado na foto utilizando o mapa disponível (ver figura 6.6). De notar aqui uma diferença em relação à aplicação móvel que anota com as coordenadas de onde é capturada a foto.

Para a recuperação de imagens é utilizado o sistema proposto no capítulo 4 tal como no dispositivo móvel e que é executado no servidor. No cliente Web é definida a pesquisa (ver figura 6.7) e no servidor é calculada a lista ordenada com as imagens resultantes da pesquisa. Nesta aplicação apenas estão disponíveis as interrogações por imagem exemplo e por conceito. A definição da interrogação segue a mesma estratégia utilizada na aplicação móvel, isto é, arrasto para a "Query Box" de vários tipos de elementos para realizar interrogações combinadas (ver figura 6.7).

6.5 Síntese

Neste capítulo são descritas duas aplicações para partilha de experiências obtidas durante a visita a um local de interesse. Memoria Mobile é a aplicação proposta para partilhar imagens durante a visita. Esta imagens servem também para guiar a visita. Memoria Web é uma aplicação para visitar o local virtualmente e construir e partilhar histórias sobre as experiências vividas durante a visita. Ambas as aplicações utilizam o sistema multimodal de recuperação de imagens proposto no capítulo 4. As aplicações descritas têm a particularidade de permitirem que o utilizador combine vários tipos de elementos para pesquisar por imagens relacionadas com um local de interesse.

Figura 6.7: Memoria Web - Pesquisa de Imagens.

7

Aplicação Semi-Automática de Anotação

Conteúdo

O capítulo descreve a plataforma proposta para anotação semi-automática baseada no método de anotação automática e na intervenção do utilizador, por via de um jogo de computador com interacção baseada em gestos e reconhecimento facial.

7.1 Introdução

Os sistemas iniciais de recuperação de imagem utilizando o conteúdo eram baseados em características de baixo nível (por exemplo, cor, textura ou forma) [Smeulders00]. Contudo, para algumas interrogações a correlação entre as imagens, identificada pela visão humana, é difícil de representar através de medidas de semelhança entre características de baixo nível. Isto acontece porque estas características não capturam o significado semântico da cena descrita na imagem. Uma solução para este problema é a inclusão da intervenção humana no processo de recuperação [Zhou03]. A informação adicional providenciada pelo utilizador durante a pesquisa permite melhorar os resultados. Porém, quando os resultados apresentados ao utilizador não incluem exemplos relevantes é difícil melhorar os resultados.

A anotação com palavras chave descrevendo o seu conteúdo é uma solução para o acesso eficaz a imagens [Kustanowitz05]. Esta tarefa pode ser efectuada através de conceitos semânticos [Lew06], treinados com informação de baixo nível, extraída automaticamente de imagens, para anotação automática de imagens com palavras. Uma vez que este treino pode incluir centenas de imagens, a recuperação baseada em semântica pode obter melhores resultados que os obtidos utilizando uma imagem exemplo. Contudo, como é apresentado no relatório do TRECVID 2006, algumas dificuldades persistem [Over06].

A anotação automática não é tão precisa como o processo manual mas os humanos têm a tendência de evitar a anotação manual [Frohlich02]. Em geral, a captura de fotos é relativamente agradável mas a tarefa de estar sentado ao computador em casa a associar palavras a imagens é uma actividade pouco interessante [Wenyin01]. Existe falta de motivação da parte das pessoas para anotar imagens e, por isso, esta tarefa acaba por ser encarada como um trabalho a desempenhar sendo esquecida a componente de entretenimento.

O jogo ESP proposto em [VonAhn04] introduziu uma nova aproximação no processo de anotação de imagens. Luis von Ahn e Laura Dabbish propõem utilizar a capacidade computacional humana para anotar imagens num jogo de computador. Desta forma, o utilizador é envolvido numa aplicação de entretenimento para efectuar uma actividade penosa.

Este capítulo descreve uma plataforma para anotação semi-automática de imagens que explora os benefícios de cada um dos paradigmas de anotação [Jesus08]. O método proposto tira partido dos resultados da anotação automática, utiliza a capacidade computacional humana para corrigir os erros do método automático e envolve o utilizador numa actividade de entretenimento com o objectivo de o motivar. É também descrita e apresentada a metodologia utilizada no desenvolvimento do jogo Tag Around [Goncalves08c], a nossa proposta para o módulo de aplicação da plataforma.

Na secção seguinte é apresentada a plataforma para anotação semi-automática. As secções subsequentes descrevem o jogo Tag Around, a forma de interacção escolhida e como são actualizados os modelos automáticos com a informação fornecida pelo utilizador. O capítulo termina com a apresentação da metodologia utilizada para conceber e avaliar a aplicação proposta.

7.2 Anotação Semântica Semi-Automática

De uma forma geral, os algoritmos de retroacção de relevância [Zhou03] incluem o utilizador no processo de pesquisa com o objectivo de corrigir os erros gerados pela pesquisa automá-

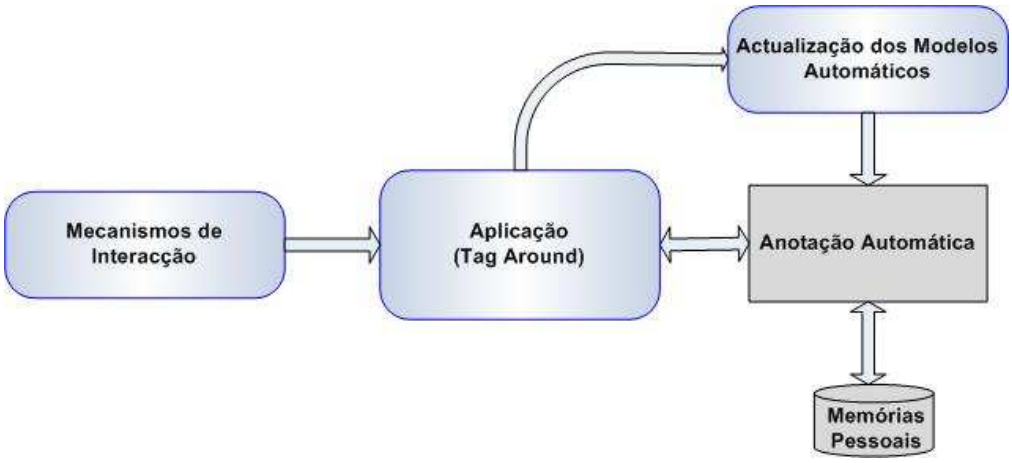


Figura 7.1: Plataforma para anotação semântica semi-automática

tica. O sistema aprende com esta informação e apresenta novos resultados ao utilizador. A plataforma proposta segue esta estratégia com duas diferenças (ver figura 7.1). Em primeiro lugar, a aplicação utilizada não é um aplicação de pesquisa de imagens. Na nossa proposta, esta tarefa é substituída por uma aplicação que anota imagens mas envolvendo o utilizador numa tarefa divertida, tal como em [VonAhn04] onde é utilizado um jogo, ou em [Tuulos07] onde é utilizada uma aplicação para construir histórias com imagens. A outra diferença está relacionada com a validação da informação obtida do utilizador, apenas depois de cumpridas algumas condições (por exemplo a validação da anotação) e não em cada iteração do algoritmo.

Na figura 7.1, é apresentado o diagrama de blocos da plataforma para anotação semântica semi-automática de imagens [Jesus08]. A proposta é composta por 4 blocos principais:

- Aplicação - bloco destinado a uma tarefa para anotar imagens, de preferência uma tarefa que seja atractiva para o utilizador;
- Mecanismos de interacção - bloco para lidar com a interacção do utilizador na aplicação;
- Actualização dos modelos automáticos - bloco para voltar a estimar os modelos para anotar imagens incluindo a informação fornecida pelo utilizador nas suas intervenções;
- Anotação automática - bloco onde são estimados os modelos para anotar imagens (descritos no capítulo 4).

O módulo da aplicação é a componente principal da metodologia proposta. Nesta tese, é proposto o jogo Tag Around (descrito na secção 7.3) como aplicação para anotar imagens. Este jogo é baseado numa interface 3D e num motor de jogo responsável pela análise das jogadas e cálculo da pontuação. No bloco de interacção são geridas as intervenções do utilizador. O Tag Around é jogado através de gestos com as mãos e reconhecimento facial. O módulo da actualização dos modelos automáticos tem como funcionalidade a actualização dos parâmetros dos modelos automáticos, utilizando a informação obtida nas jogadas efectuadas pelo utilizador. O bloco da anotação automática refere-se aos modelos semânticos descritos no capítulo 4.

Com os módulos da figura 7.1, foi definido um algoritmo semi-automático para anotação de imagens descrito na secção 4.4.2. Inicialmente, um conjunto de imagens, previamente anotadas pelos modelos automáticos, é apresentado ao jogador para anotação. Em cada jogada o utilizador associa uma palavra a uma imagem e é calculada uma pontuação. Se um conceito

for anotado correctamente em mais de N_{upd} imagens (ver algoritmo na secção 4.4.2) então o respectivo modelo semântico é actualizado. Depois de várias jogadas, é esperado que a precisão dos modelos semânticos aumente.

O jogo Tag Around é diferente do jogo ESP [VonAhn04] porque utiliza os modelos automáticos e é baseado numa interface gestual que permite que o jogo possa ser jogado por diversos tipos de utilizadores e em diversos locais, por exemplo locais onde as pessoas estão à espera e por isso têm tempo disponível (por exemplo, aeroportos ou hospitais). As próximas secções descrevem os diferentes módulos da plataforma para anotação semântica semi-automática.

7.3 Tag Around

Nesta secção é descrito o jogo Tag Around [Goncalves08, Goncalves08b], como exemplo para o bloco de aplicação da figura 7.1.

O jogo é jogado através de gestos em frente a uma câmara de vídeo. Estes gestos servem para movimentar e associar conceitos a imagens e são detectados no vídeo capturado quando o utilizador está a jogar. Também é usada uma interface baseada em reconhecimento facial para efectuar o *login* no jogo. A aplicação é dividida em vários módulos (ver figura 7.2):

- Interface gráfica - interface 3D desenvolvida utilizando o OGRE (Object-oriented Graphics Rendering Engine);
- Motor de jogo - gere a dinâmica do jogo, analisa as jogadas e calcula a pontuação;
- Detecção de movimento - detecta os gestos do jogador analisando o vídeo do utilizador a jogar;
- Reconhecimento de faces - detecção e reconhecimento de faces para registar novos utilizadores e para fazer *login* na aplicação.

O jogo é organizado nestes módulos para permitir fácil adaptação a diferentes cenários. Os módulos da interface gráfica, detecção de movimento e reconhecimento de faces estão directamente relacionados com as técnicas de interacção escolhidas e por isso podem sofrer adaptações de acordo com o cenário escolhido e o tipo de utilizadores. O motor de jogo mantém-se inalterado. As secções seguintes explicam cada um destes módulos.

7.3.1 Interface do Jogo

O objectivo é apresentar aos utilizadores um cenário 3D onde os jogadores interagem com imagens e palavras para fazer anotações e ao mesmo tempo percebem se estão a fazer boas ou más anotações. A interface é constituída por quatro ecrãs principais:

- "Inicial" - primeiro ecrã apresentado aos jogadores;
- "Highscores" - para apresentar os jogadores com melhor pontuação;
- "Jogo" - utilizado na fase em que o utilizador está a jogar;
- "Login" - para entrar no jogo.

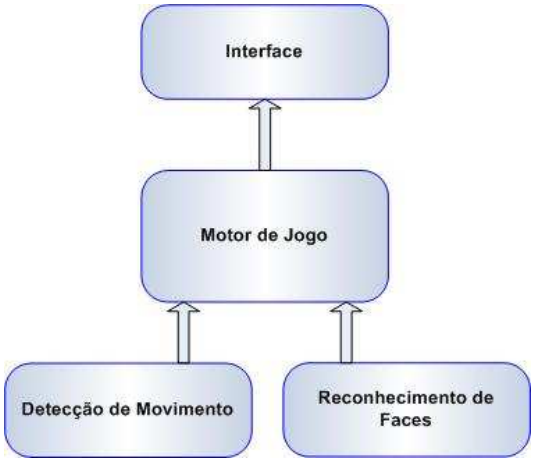


Figura 7.2: Diagrama de blocos do jogo Tag Around.

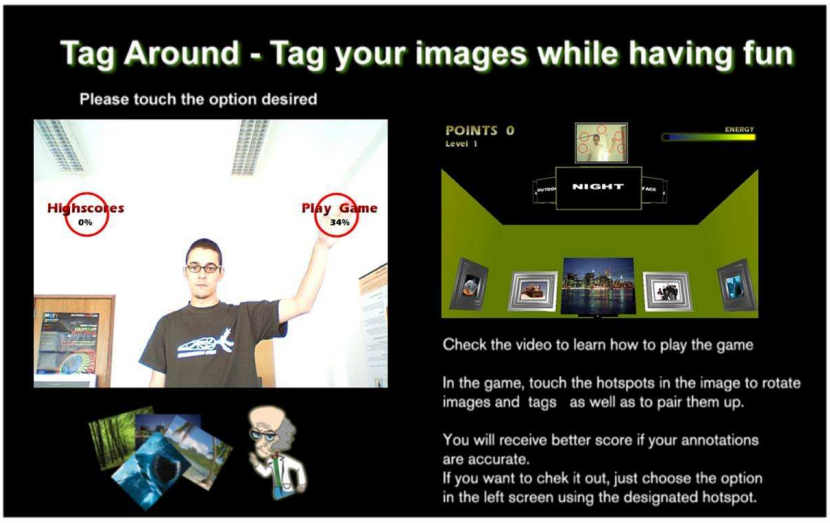


Figura 7.3: Menu inicial da aplicação.

O ecrã "Inicial" é composto por duas opções diferentes: "Play Game" e "Highscores"(ver figura 7.3). O utilizador pode escolher uma das opções movimentando as mãos em frente das zonas definidas para cada opção (dentro dos círculos a vermelho na figura 7.3). O utilizador tem que fazer movimentos até que a percentagem indicada seja 100.

Caso a opção seja "Highscores", o utilizador pode ver os cinco jogadores com maior pontuação identificados pela sua fotografia (ver figura 7.4). Dado que o jogo não inclui um teclado ou um rato não há forma de digitar o nome do utilizador e por isso utiliza-se a sua foto.

Quando o utilizador entra no modo "Play Game", o ecrã de "Login" com a interface de reconhecimento facial é apresentado ao utilizador para que este se registe ou entre no jogo (ver figura 7.5). A seguir, o utilizador entra no modo "Play Game" e pode começar a jogar (ver figura 7.6).

O ecrã "Jogo", representado na figura 7.6 é composto por vários elementos representados no ecrã:

- A imagem do jogador com as diferentes marcas para interagir;
- Um conjunto pré-definido de palavras colocadas numa plataforma rotacional;

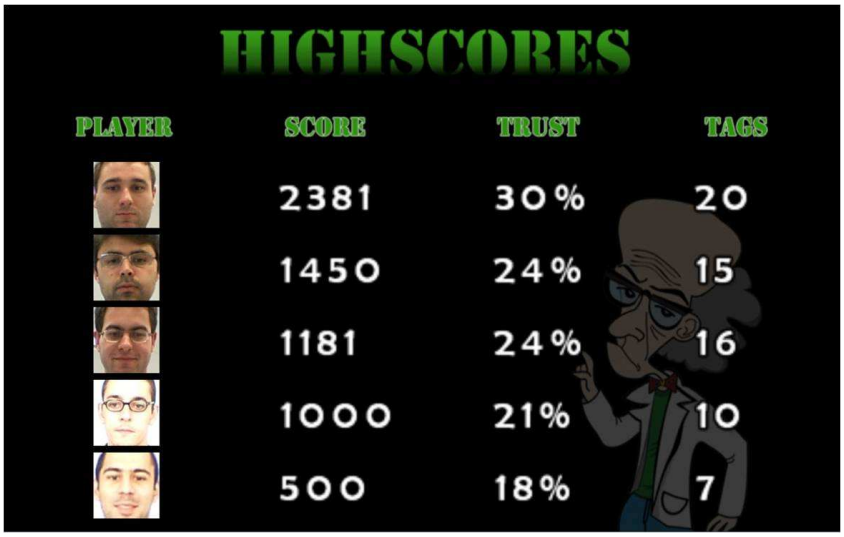


Figura 7.4: "Highscores" - face para identificar o utilizador.

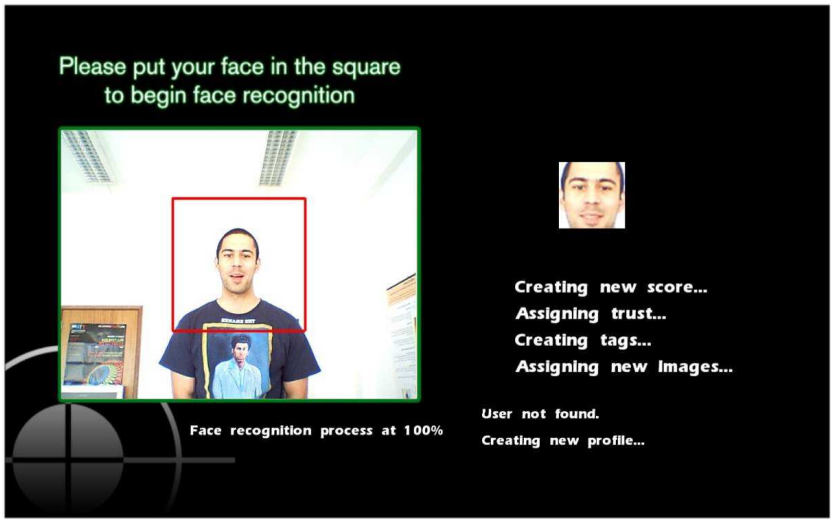


Figura 7.5: Interface para *login* utilizando técnicas de reconhecimento de faces.

- Um conjunto de imagens para anotar com palavras que são apresentadas em baixo no ecrã;
- Uma barra de energia para indicar quando termina o jogo;
- A pontuação que depende da qualidade das anotações (correcta ou incorrectas) de conceitos a imagens efectuadas em cada jogada pelo utilizador;
- Uma lista de palavras que já foram anotadas na imagem seleccionada (imagem colocada no centro do ecrã).

Quando o jogo termina, a barra de energia desaparece do ecrã, a pontuação, o número de anotações realizadas pelo utilizador e a confiança que o jogo tem no utilizador são apresentados no ecrã. Esta informação também é associada ao perfil do utilizador e gravada para jogos futuros.

Figura 7.6: Interface do Jogo.

7.3.2 Motor de Jogo

O motor de jogo controla a dinâmica do jogo, recebe a informação referente às acções do utilizador dos módulos de detecção de movimento e reconhecimento facial e envia a informação para o bloco da interface, para que graficamente seja apresentada a resposta do jogo às acções do jogador. É também responsável pelo cálculo da pontuação em cada jogada. A dinâmica do jogo é resumida nos seguintes passos:

- Quando o jogo começa, é iniciada uma contagem de tempo e é apresentado ao utilizador um conjunto de imagens (seleccionadas aleatoriamente) e um conjunto de conceitos pré-definidos;

- A seguir, o jogador utilizando marcas designadas (ver secção 7.4) tem de associar o maior número de imagens a palavras. Quanto maior for o número de anotações correctas mais pontos ganha o jogador e mais tempo joga porque a barra de energia diminui mais lentamente;

- O jogo termina quando acabar a energia, isto é, a barra de energia desaparecer completamente.

Durante o jogo, o desempenho do jogador condiciona a passagem para os níveis seguintes e anotações incorrectas atrasam a mudança de nível. Sempre que ocorre uma mudança de nível, o conjunto de imagens a anotar é alterado mas o conjunto de palavras mantém-se. O tempo disponível para que o jogador faça anotações diminui à medida que o jogador progride nos níveis do jogo. É importante realçar que as anotações correctas melhoram a pontuação porque fazem aumentar a energia e por isso o utilizador dispõe de mais tempo para anotar imagens. Em geral, o tempo disponível num nível vai diminuindo mas quando são efectuadas anotações correctas o tempo disponível aumenta. Pelo contrário, anotações incorrectas penalizam o utilizador diminuindo a energia e por consequência o tempo disponível.

7.3.2.1 Cálculo da Pontuação

Um dos aspectos mais relevantes num jogo é a pontuação obtida pelos jogadores porque é uma das formas de premiar o desempenho. O objectivo principal do jogo Tag Around é anotar correctamente imagens com palavras, por isso os jogadores que se esforçam para fazer anotações correctas devem obter a pontuação mais alta. De notar que o jogo foi concebido para resolver o problema da falta de motivação do utilizador para realizar a tarefa da anotação manual, portanto admite-se que as imagens não têm nenhuma anotação inicial. Esta opção torna difícil a análise das primeiras anotações numa imagem porque não há forma de saber exactamente se a anotação é correcta ou não. Neste caso, a pontuação é baseada na confiança no jogador que nas suas primeiras jogadas não é relevante e na classificação obtida pelo modelos automáticos.

Depois de analisados vários tipos de jogos (cooperativos e não cooperativos) e de entrevistar utilizadores foi definida uma fórmula para calcular a pontuação de cada jogada (anotação) [Jesus08], que foi testada utilizando simulações feitas na ferramenta Matlab. Assim, uma anotação feita por um jogador é avaliada utilizando três factores distintos:

- Probabilidade obtida pelo algoritmo automático proposto no capítulo 4. Para novas anotações com novos utilizadores é a única informação disponível para avaliar a jogada;
- A confiança do sistema no jogador obtida através do desempenho do utilizador no jogo;
- A confiança no grupo de utilizadores que anteriormente fez a mesma anotação.

No caso de ser a primeira vez que o jogador utiliza a aplicação, a confiança no jogador é nula e nestas circunstâncias o algoritmo automático tem mais relevância.

Dado um conjunto de imagens $L = \{I_1, ..., I_{N_l}\}$ ($L \subset C_{img}$) e um conjunto de conceitos $V_{sc} = \{w_1, ..., w_{N_{con}}\}$ ($V_{sc} \subset V_{con}$), a pontuação associada à anotação do conceito w na imagem I é obtida por,

$$S_{total}(I,w,n,m) = C_{group}(m) + [1 - C_{group}(m)]S_{new}(I,w,n), \tag{7.1}$$

onde n representa o número de anotações correctas efectuadas pelo utilizador, m é o número de vezes que o conceito w foi anotado na imagem I, $S_{new}(I,w,n)$ é um valor que avalia a anotação a partir do algoritmo automático e da confiança no utilizador (equação 7.3) e $C_{group}(m)$ representa a confiança no grupo obtida por,

$$C_{group}(m) = 1 - e^{-(\frac{m}{k_g})}, \tag{7.2}$$

onde k_g é um parâmetro da exponencial que é calculado para que a partir de m anotações a confiança no grupo seja aproximadamente 1. Consideramos que três utilizadores ($m = 3$) a fazer a mesma anotação representa uma confiança na anotação elevada e por isso k_g é obtido admitindo esta hipótese. O ESP GAME [VonAhn04] valida uma anotação com dois jogadores. Com a equação 7.2, quando o $m = 2$, não é atribuído o valor máximo da pontuação mas o valor obtido é suficiente para que o sistema a classifique de correcta.

Quando um conceito w é anotado pela primeira vez numa imagem I a pontuação é,

$$S_{new}(I,w,n) = C_{player}(n) + [1 - C_{player}(n)]p(w|I), \tag{7.3}$$

onde $p(w|I)$ é a probabilidade obtida pelo método automático (ver capítulo 4) e C_{player} é a confiança no jogador que expressa a qualidade das anotações anteriores do mesmo jogador,

$$C_{player}(n) = \begin{cases} k_p n, & n < K_{moves} \\ k_{conf}, & n \geq K_{moves} \end{cases} \tag{7.4}$$

K_{moves} é uma constante com o número de jogadas correctas necessárias para chegar ao valor máximo de confiança k_{conf} e k_p é uma constante que é utilizada para incrementar a confiança no jogador.

Uma anotação é considerada correcta (n é incrementado) caso a confiança no grupo seja diferente de zero e a pontuação obtida para esta anotação seja superior a um limiar. Quando a confiança no grupo é zero significa que a pontuação é obtida utilizando apenas o algoritmo automático e a confiança no jogador. Assim, é difícil classificar a anotação de correcta. Quando a pontuação é inferior a um outro limiar a anotação é considerada incorrecta (n é decrementado).

7.3.3 Detecção de Movimento

Neste módulo são detectados e interpretados os gestos para que o motor de jogo possa dar sequência ao jogo de acordo com a interacção do utilizador. Para jogar, o jogador faz movimentos com as mãos em cinco zonas específicas da imagem capturada. Foram experimentados dois algoritmos para detectar movimentos no vídeo do jogador, um algoritmo baseado em fluxo óptico [Lucas81] e outro em detecção de movimento (baseado na subtracção de imagens consecutivas). Pelos testes efectuados o algoritmo com melhor relação entre desempenho computacional e eficácia na captura de movimento foi o de detecção de movimento. Por isso, é utilizado o algoritmo de detecção de movimento.

7.3.4 Reconhecimento de Faces

A identificação do jogador, no sentido de associar a informação referente ao seu desempenho para apresentação no ecrã "Highscores" ou para calcular a confiança de modo a obter a pontuação, é realizada utilizando reconhecimento facial.

Este método é dividido em três tarefas [Grangeiro08a]: detecção, normalização e reconhecimento de faces. Em primeiro lugar, é preciso detectar a presença de uma face na imagem capturada pela câmara. O método utilizado é baseado no sistema descrito em [Viola04] complementado por um método de detecção de pele para confirmar a presença de faces. Para normalizar as imagens das faces utilizou-se a equalização de histogramas, de modo a resolver os problemas de iluminação e a detecção de olhos para regularizar a posição dos mesmos na imagem da face detectada. Finalmente, para o reconhecimento facial utilizou-se uma técnica para representação das faces [Turk91] e outra para classificação da identidade das faces através do método descrito em [Muller01]. Para melhorar este processo é também implementada uma técnica de estimação da pose facial baseada no método proposto em [Viola04] para que o reconhecimento seja feito comparando faces com a mesma pose.

7.4 Mecanismos de Interacção

Para suportar um jogo com uma interacção acessível a vários tipos de utilizadores (sem a necessidade de usarem dispositivos como o teclado ou o rato) e para tornar a aplicação mais divertida, a aplicação proposta utiliza um interface baseada em gestos e uma interface baseada em reconhecimento facial. Para jogar o utilizador efectua movimentos com as mãos em frente a uma câmara de vídeo e para se registar o jogador utiliza a face. Desta forma o jogo não necessita de um computador visível, apenas um ecrã e um câmara de vídeo e pode ser utilizado em locais públicos.

7.4.1 Interface Baseada em Gestos

Para jogar o Tag Around o utilizador tem de realizar movimentos com as mãos em zonas específicas. Na figura 7.6, estas zonas são representadas por círculos a vermelho na imagem do utilizador (em cima). No ecrã "Inicial" existem duas zonas para escolher entre jogar o jogo ou ver os "Highscores". No ecrã "Jogo" estão disponíveis cinco zonas específicas na imagem e o utilizador tem de movimentar as mãos nestas zonas para movimentar os objectos no ecrã. Em baixo, existem duas zonas no ecrã (ver figura 7.6) para rodar as imagens de modo a seleccionar uma delas. Em cima estão mais duas zonas para rodar os conceitos. Em cima e em baixo, as zonas do lado esquerdo servem para rodar as imagens ou os conceitos para o lado esquerdo e as do lado direito para rodar para o lado direito. A quinta zona está situada por cima do utilizador e serve para associar a imagem seleccionada ao conceito escolhido.

7.4.2 Interface Baseada em Reconhecimento Facial

Para guardar informação referente a cada utilizador é necessário registar a sua identificação. Geralmente, esta identificação é realizada através do registo do nome do utilizador. Contudo, nesta aplicação é difícil manter e actualizar informação deste tipo uma vez que a interacção não se faz através das formas mais habituais como o rato ou o teclado. Assim, utiliza-se uma interface baseada no reconhecimento facial para registar os dados do utilizador. A utilização desta interface passa pela colocação da face numa área limitada por um quadrado durante 10 segundos para que o sistema proceda ao seu reconhecimento. Durante esse tempo, é mostrado o estado de evolução do processo sob a forma de percentagem (ver figura 7.5). A aplicação de um método de reconhecimento de faces neste sistema depara com os seguintes problemas: dificuldade no reconhecimento da identidade da pessoa devido à quantidade e variabilidade reduzida de fotos de cada pessoa nos primeiros *logins* e reconhecimento indevido das pessoas que, por exemplo, estão a assistir ao jogo. As soluções encontradas para resolver os problemas referidos são:

- Reconhecimento do utilizador durante 10 segundos, isto é, utilizando cerca de 300 imagens do vídeo capturado. Esta opção permite capturar mais faces do utilizador para o algoritmo de reconhecimento de faces. Espera-se que o utilizador durante os 10 segundos não esteja sempre na mesma posição de modo a aumentar a variabilidade das imagens capturadas.
- Armazenamento de novas fotos do utilizador a cada *login*. Desta forma, é garantida uma maior variabilidade das faces do indivíduo contidas na base de dados;

- Limitação do reconhecimento de faces a uma área quadrada indicada (ver figura 7.5). Desta forma, melhoram-se os resultados porque é reduzida a área da imagem a processar.

7.5 Actualização dos Modelos Automáticos

Com os blocos da figura 7.1, é implementado um algoritmo semi-automático para anotar imagens com conceitos semânticos (descrito na secção 4.4.2). Neste bloco é verificado se estão reunidas as condições para estimar novamente algum dos conceitos pré-definidos, de acordo com o algoritmo descrito na secção 4.4.2. Caso um conceito tenha sido anotado correctamente com mais de N_{upd} imagens então o respectivo modelo semântico é actualizado, isto é, é estimado novamente o modelo semântico mas com as N_{upd} imagens acrescentadas ao conjunto de treino. Espera-se que com um conjunto de treino maior o algoritmo consiga melhorar os seus resultados e classificar imagens com mais precisão. Desta forma, a parte da pontuação referente ao algoritmo aumentará progressivamente a sua eficácia e, por consequência, são obtidas pontuações mais adequadas.

7.6 Anotação Automática

Os modelos semânticos estimados, como é descrito no capítulo 4, são utilizados neste módulo para calcular a probabilidade do conjunto de conceitos pré-definido estar presente ou ausente nas imagens da base de dados. Estas probabilidades são utilizadas para calcular a pontuação (ver secção 7.3.2.1) e têm um papel importante no cálculo da pontuação das primeiras anotações de um conceito numa imagem. Ao longo do tempo, esta importância diminui porque os restantes factores utilizados no cálculo da pontuação, que dependem dos utilizadores, passam a ter valores mais fiáveis.

7.7 Concepção

O projecto e a implementação da interface do jogo para anotação de imagens foi feito de forma iterativa. No início, foram definidas as ideias principais e vários cenários de aplicação. De seguida, foram realizados vários testes com protótipos em papel e finalmente foram realizados testes de usabilidade. O protótipo computacional foi refinado em cada uma das fases de forma iterativa. Nesta secção, apenas são apresentados os passos e as questões mais relevantes do processo de concepção da aplicação. Mais detalhes e os resultados dos testes realizados são apresentados no capítulo 8.

7.7.1 Análise e Definição das Funcionalidades

A ideia de implementar um jogo como o Tag Around foi inspirada no paradigma dos jogos com um propósito proposto por Luis Von Ahn [VonAhn06a], no modo de interacção utilizado no EyeToy [EyeToy05], no algoritmo semi-automático para anotação proposto em [Jing05] e na experiência obtida durante o desenvolvimento e experimentação de uma versão multi-utilizador e colaborativa do jogo Pong criado em 1972 por Nolan Bushnell e Ted Dabney. Luis Von Ahn introduziu a ideia de converter a anotação de imagens numa tarefa divertida através de um jogo, o EyeToy mostrou que técnicas de interacção diferentes do teclado e do rato tornam a

aplicação mais divertida e em [Jing05] é apresentada uma proposta para anotação com modelos automáticos corrigidos pelo utilizador. A combinação destas três ideias mais a experiência obtida na versão Pong foram peças chave para a definição jogo Tag Around e das suas funcionalidades. A versão do Pong, desenvolvida por um grupo de alunos do curso de Engenharia Informática do DI/FCT-UNL, é jogada com dois conjuntos de pessoas em duas salas diferentes. Um sala joga contra a outra e a raqueta é deslocada através de gestos (com um cartão de uma cor na mão) da parte da sala para onde se deve mover a raqueta. Durante as experiências notou-se que com uma interacção simples (sem a utilização de tecnologia), num espaço público sem um computador visível e colocando várias pessoas a colaborarem, esta versão do Pong se tornou numa actividade social muito divertida. Com base nestas premissas foram definidas várias funcionalidades e equacionados vários cenários de aplicação:

- Jogo para locais públicos onde as pessoas permanecem algum tempo sem nada para fazer (por exemplo, aeroportos ou hospitais);
- Aplicação para ser utilizada por crianças em escolas de modo enriquecer o seu vocabulário;
- Interface para pacientes com problemas, por exemplo de afasia para melhorarem a linguagem com a ajuda de fotografias;
- Jogo para ambientes domésticos para o utilizador anotar as suas fotos em casa;

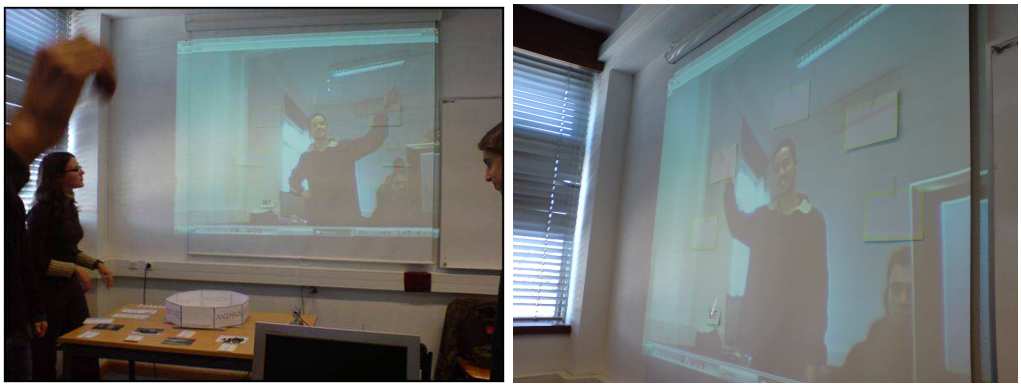
O jogo Tag Around não foi desenvolvido especificamente para nenhum destes cenários. Foi concebido de forma genérica para que possa ser posteriormente adaptado a qualquer um dos cenários. Nesta fase foi definido que a aplicação seria desenvolvida para utilização em locais públicos, com uma interface gestual e com anotação semi-automática.

7.7.2 Protótipos em Papel

Com o objectivo de desenvolver uma versão inicial do jogo e com base na análise efectuada na secção anterior foram realizados vários protótipos em papel para serem testados com vários utilizadores [Goncalves08]. Uma das questões fulcrais da aplicação é o uso de uma câmara e a utilização de interacção baseada em gestos do utilizador com o objectivo de anotar imagens. Esta foi a maior preocupação nesta fase, de modo a construir os ecrãs principais da aplicação e definir a sequência de acções adequada para lidar com este tipo de interacção.

Para testar os protótipos em papel, houve necessidade de utilizar alguma tecnologia (ver figura 7.7) porque não era fácil testar a interacção baseada em gestos só com estes protótipos. Na figura 7.7, são apresentadas duas imagens que descrevem o cenário utilizado. Foi colocada uma câmara de vídeo a capturar os movimentos do utilizador e o vídeo capturado era projectado. Para definir as zonas de interacção foram coladas marcas com folhas de papel no plano de projecção do vídeo de modo a indicar as zonas onde o utilizador tinha de realizar movimentos com as mãos (ver figura 7.7b). Esta estratégia permitiu ajustar a posição das zonas de interacção. Os ecrãs eram representados em cima de uma mesa (ver figura 7.7b).

Depois de realizados testes com vários utilizadores e de refinados os protótipos em papel foi implementado um primeiro protótipo em computador.



(a) (b)

Figura 7.7: Protótipos em papel: a) Cenário construído para realizar os testes; b) Marcas de papel sobre o vídeo do utilizador para definir zonas de interacção.

7.7.3 Testes de Usabilidade

O jogo Tag Around também foi sujeito a testes de usabilidade [Goncalves08a] com o objectivo de avaliar a complexidade da interface, a utilidade, os aspectos visuais, para perceber se é fácil de usar e para analisar a componente de diversão do jogo. Alguns destes aspectos foram testados com protótipos em papel e voltaram a ser avaliados nesta fase. A componente de diversão do jogo, um ponto fulcral para os objectivos da aplicação, foi nesta etapa o aspecto mais relevante. Os testes foram realizados com vários utilizadores que eram entrevistados antes de experimentarem o jogo e respondiam a um questionário depois de jogar o jogo. Depois de analisada toda esta informação o protótipo foi refinado até chegar à estrutura descrita na secção 7.3.

7.8 Síntese

Neste capítulo é apresentada uma proposta para anotação semântica semi-automática, que utiliza o modelo de anotação semântica proposto no capítulo 4. A plataforma é composta por um bloco de aplicação, um módulo de interacção e um bloco de actualização dos métodos de anotação automática. Primeiro, é descrita uma plataforma para anotação que pode ser utilizada com vários tipos de aplicações e depois é descrito o jogo Tag Around, uma proposta específica para o bloco de aplicação da plataforma. O capítulo termina com a descrição dos passos utilizado na concepção da aplicação.

8

Avaliação

Conteúdo

Este capítulo descreve as experiências realizadas para avaliar as aplicações dos modelos semânticos propostos. São apresentados e discutidos os resultados dos testes realizados para avaliar a recuperação e anotação de imagens, bem como os testes de usabilidade para avaliar as aplicações.

8.1 Introdução

Nesta tese é proposto um método para análise semântica de imagens que pode ser utilizado para anotação ou recuperação com base em conceitos semânticos. Este modelo é testado em três aplicações de memórias pessoais em contextos diferentes: ambientes domésticos, turismo e entretenimento. Cada aplicação utiliza a anotação ou a recuperação de imagens aplicada a estes contextos e utiliza um tipo de interacção diferente. Na aplicação para ambientes domésticos é utilizada uma interface para PC e o utilizador interage com o teclado e o rato. A aplicação turística é baseada numa interface para dispositivos móveis com interação através de uma caneta específica para tocar no ecrã do dispositivo. A aplicação de entretenimento utiliza uma interface baseada em gestos e reconhecimento facial.

Dado que cada aplicação tem características específicas, o método para análise semântica é avaliado separadamente para cada aplicação. As aplicações desenvolvidas são também avaliadas através de testes de usabilidade com utilizadores. Assim, para cada aplicação é realizada uma avaliação dos resultados de recuperação ou anotação baseada em medidas de desempenho (ver seccão 8.2) e uma avaliação baseada em testes de usabilidade para testar e melhorar as funcionalidade da aplicação. Os resultados desta avaliação são apresentados neste capítulo.

Para avaliar o método de análise semântica são necessárias três componentes: uma colecção de fotos anotada manualmente, uma estratégia para treinar e testar o sistema e um conjunto de medidas de avaliação de desempenho para métodos de recuperação e anotação de imagens. A próxima secção apresenta as medidas de avaliação utilizadas. A colecção de imagens e a estratégia de treino e teste são descritas nas secções seguintes, destinadas à apresentação da avaliação efectuada a cada uma das aplicações. Estas secções descrevem a colecção pessoal, apresentam os testes e as avaliações realizadas com o método de análise semântica e descrevem também os testes com utilizadores.

8.2 Medidas de Avaliação

No trabalho desenvolvido a recuperação de imagens é realizada através de pesquisas. Dada uma interrogação, o sistema constrói uma lista ordenada de imagens relevantes para essa interrogação. No caso do sistema ideal, esta lista deve incluir no topo as imagens relevantes seguidas pelas imagens não relevantes. Na prática isto nem sempre acontece, as imagens relevantes aparecem misturadas com as imagens não relevantes. Várias medidas têm sido utilizadas para medir a capacidade de um sistema recuperar imagens. As mais utilizadas são a precisão e a cobertura. A precisão mede a capacidade do sistema recuperar apenas documentos relevantes, a cobertura é uma medida da capacidade do sistema recuperar todos os documentos relevantes. Assim, a precisão é dada por,

$$Prec = \frac{número\ de\ imagens\ relevantes\ nas\ primeiras\ n_{rec}\ recuperadas}{n_{rec}}, \tag{8.1}$$

e a cobertura é obtida por,

$$Cob = \frac{número\ de\ imagens\ relevantes\ nas\ primeiras\ n_{rec}\ recuperadas}{|Rel|}, \tag{8.2}$$

onde *Rel* representa o conjunto das imagens relevantes. A precisão e a cobertura medem o

desempenho do sistema para n_{rec} imagens apresentadas ao utilizador. Em geral, um sistema com elevada precisão apresenta uma cobertura baixa e vice-versa. Uma medida global do sistema que mede a capacidade de colocar os documentos relevantes no topo da lista ordenada com os resultados é a Average Precision (AP),

$$AP = \frac{\sum_{i=1}^{|Rel|} Prec[r(i)]}{|Rel|}, \tag{8.3}$$

onde r representa uma lista com as posições das imagens relevantes nos resultados da pesquisa. Esta medida calcula a média da precisão obtida sempre que é recuperado um novo documento relevante, isto é, existe uma alteração da cobertura. A Average Precision é uma medida para avaliar o resultado de uma interrogação muito utilizada para avaliar conceitos semânticos [Over06]. Mean Average Precision (MAP) é a média da Average Precision obtida para vários conceitos,

$$MAP = \frac{\sum_{w \in V_{con}} AP(w)}{|V_{con}|}, \tag{8.4}$$

onde V_{con} representa o conjunto de conceitos semânticos.

O desempenho de um sistema de anotação automática é avaliado através da comparação das anotações automáticas com as anotações produzidas manualmente. A medida MAP é também utilizada para medir o desempenho da anotação de imagens quando os conceitos são utilizados para recuperar imagens. Porém, também existe um conjunto de métricas [Feng04, Li03] específicas para avaliar a anotação de imagens. Nesta tese são utilizadas três medidas: a precisão por palavra, a cobertura por palavra e a cobertura por imagem. Estas medidas aplicam os conceitos de precisão e cobertura referidos anteriormente para avaliar o sistema de recuperação de imagens mas considerando conjuntos diferentes. Assim, a precisão para cada palavra é obtida por,

$$P_w = \frac{NI_{corr}}{NI_{auto}}, \tag{8.5}$$

onde NI_{corr} representa o número de imagens anotadas correctamente e NI_{auto} o número de imagens anotadas automaticamente. Esta medida representa a exactidão com que o sistema anota imagens de um conceito. A cobertura por palavra é obtida,

$$C_w = \frac{NI_{corr}}{NI_{manu}}, \tag{8.6}$$

em que NI_{manu} é o número de imagens anotadas manualmente. A cobertura por palavra mede a capacidade do sistema anotar correctamente imagens. Outra medida utilizada é a cobertura por imagem,

$$C_i = \frac{NW_{corr}}{NW_{manu}}, \tag{8.7}$$

onde NW_{corr} representa o número de palavras anotadas correctamente numa imagem e NW_{manu} é o número de palavras anotadas manualmente. A cobertura média é uma medida do número de palavras anotadas correctamente por imagem.

Figura 8.1: Fotos da colecção pessoal do autor utilizadas na aplicação Memoria Desktop e na Aplicação para Anotação Semi-Automática.

No processo de avaliação das aplicações são utilizados questionários e para algumas questões as respostas são dadas numa escala tipo Likert [Likert32] com 5 hipóteses de resposta. Em geral, nesta escala o 1 representa uma resposta negativa e o 5 uma resposta positiva. Para avaliar as respostas dos utilizadores são calculadas três medidas estatísticas: a média, o desvio padrão e a moda.

8.3 Recuperação de Imagens em Ambientes Domésticos

Nesta secção é apresentada a avaliação da aplicação proposta para partilha de experiências através de fotografia em ambientes domésticos. Primeiro, é caracterizada a colecção de fotografias utilizada nos testes, de seguida são apresentados os resultados obtidos com as experiências realizadas para avaliar o sistema de recuperação e o método de anotação. A secção termina com a descrição dos resultados da avaliação efectuada com utilizadores, com o objectivo de aperfeiçoar a interface e validar as propostas.

8.3.1 Caracterização da Colecção de Imagens

A aplicação Memoria Desktop permite pesquisar imagens em dois modos principais: pesquisa utilizando conceitos ou pesquisa por composição de uma imagem através de partes de outras. Para avaliar estes métodos de pesquisa e a técnica de anotação proposta é utilizada a colecção pessoal do autor da tese com cerca de 5000 fotos. Estas memórias pessoais são compostas essencialmente por fotos de pessoas, paisagem naturais e urbanas, fotos de viagens, férias e de outros momentos importantes do passado. A figura 8.1 mostra algumas fotos para ilustrar a colecção utilizada.

Utilizando os modelos semânticos propostos foram treinados sete classificadores binários para serem avaliados nesta colecção de fotos:

- "People" *versus* "No People";
- "Indoor" *versus* "Outdoor";

- "Manmade" *versus* "Nature";
- "Face" *versus* "No Face";
- "Snow" *versus* "No Snow";
- "Beach" *versus* "No Beach";
- "Party" *versus* "No Party".

O conjunto de treino para estimar estes classificadores binários foi seleccionado da base de dados do TRECVID2005 [Trecvid05], dos CDs da Corel Stock Photo e do Flickr [Flickr04] de forma a construir um conjunto de treino genérico.

Para avaliar os resultados obtidos pelos métodos propostos, as imagens da colecção pessoal foram anotadas manualmente com os seguintes conceitos:

- "Beach", 207 imagens na colecção;
- "Face", 1597 fotos com faces visíveis na colecção;
- "Indoor", 1271 imagens da colecção obtidas em interiores;
- "Manmade", 2756 fotos na base de dados com objectos artificiais (feitos pelo homem);
- "Nature", 1890 fotografias da colecção capturadas com paisagens naturais;
- "Outdoor", 3439 imagens de exteriores no repositório pessoal;
- "Party", 264 fotos de eventos festivos;
- "People", 2286 imagens com pessoas;
- "Snow", 125 imagens na base de dados.

Uma colecção pessoal é caracterizada por fotos que reflectem as preferências de cada indivíduo e por isso, para cada colecção pessoal, é possível definir um conjunto específico de conceitos. Por uma questão de generalidade são utilizados os conceitos listados em cima. Estes conceitos representam um subconjunto dos 449 conceitos da LSCOM (*Large Scale Concept Ontology for Multimedia*) [Naphade06] que entendemos adequados para memórias pessoais. O projecto LSCOM tinha como objectivo definir uma ontologia de conceitos para serem utilizados na anotação e recuperação de informação multimédia.

8.3.2 Sistema de Recuperação de Imagens

Para testar os modelos semânticos na aplicação de recuperação de imagens foram realizadas várias pesquisas utilizando a linguagem visual descrita na secção 5.3.4.1 e os conceitos referidos na secção 8.3.1, treinados com as características visuais e temporais (ver secção 4.6). Na figura 8.2, mostram-se os resultados obtidos para a pesquisa "No Indoor AND No Manmade", como exemplo ilustrativo das experiências efectuadas para testar e avaliar os conceitos semânticos. A maior parte das imagens apresentadas na figura 8.2 satisfazem os critérios definidos na pesquisa e, por isso, consideramos que os resultados para esta pesquisa são bons.

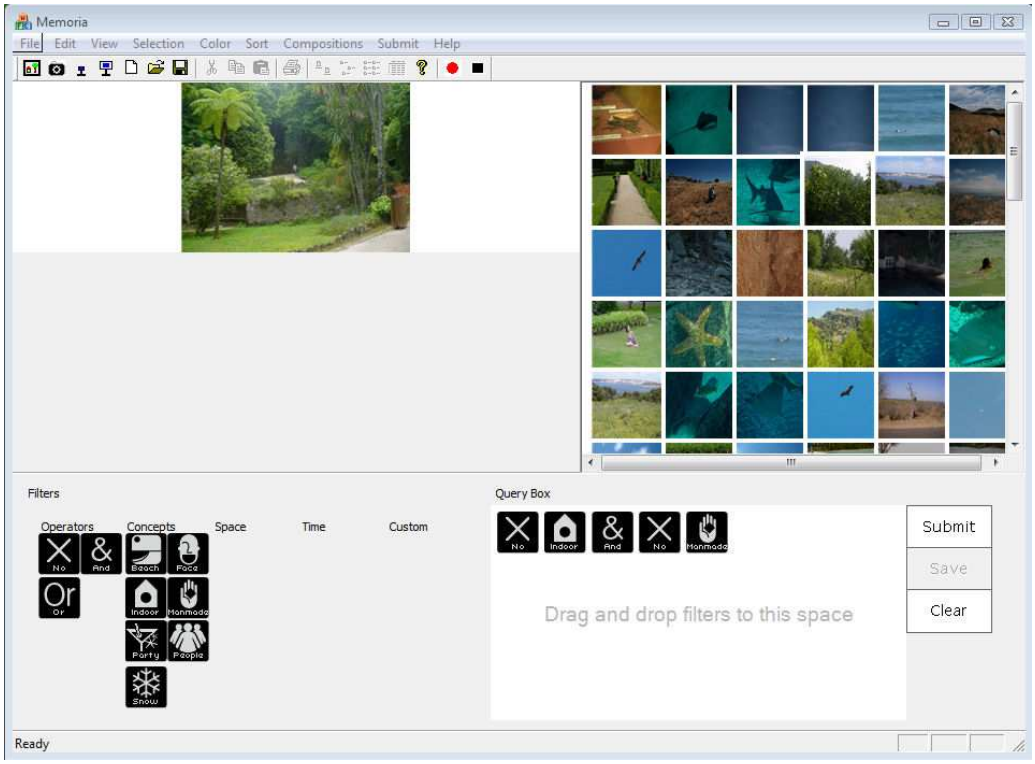


Figura 8.2: Pesquisa por conceito - resultados obtidos para a interrogação "No Indoor AND No Manmade".

Para avaliar globalmente os modelos semânticos, repetimos a experiência anterior para cada conceito do conjunto definido anteriormente e sobre a lista ordenada foi calculado o Average Precision para cada conceito e depois foi obtido o MAP. A tabela 8.1 apresenta os resultados obtidos [Jesus06] sendo que nesta primeira experiência o objectivo é comparar o desempenho individual dos vários descritores visuais utilizados. O descritor baseado em regiões de cor e os descritores SIFT são representadas em vectores de ocorrências de termos visuais pertencentes a vocabulário de cor e textura respectivamente (ver secção 4.6).

Conceitos	Momentos de cor	Regiões de cor	Filtro de Gabor	SIFT
People	0,67	0,68	0,68	0,76
Face	0,38	0,46	0,51	0,38
Outdoor	0,89	0,80	0,89	0,89
Indoor	0,46	0,41	0,57	0,59
Nature	0,34	0,44	0,48	0,66
Manmade	0,57	0,59	0,62	0,80
Snow	0,32	0,11	0,05	0,03
Beach	0,35	0,24	0,10	0,23
Party	0,09	0,26	0,12	0,11
MAP	**0,45**	**0,44**	**0,45**	**0,49**

Tabela 8.1: MAP para vários conceitos utilizando diversas características visuais. As regiões de cor e as características SIFT são representadas num vector de ocorrências e é utilizado o LSA.

Os valores de MAP apresentados na tabela 8.1 permitem avaliar na globalidade o desempenho de cada descritor. Como se pode observar, o descritor SIFT (MAP=0,49) apresenta os melhores resultados quando comparado com os restantes descritores (MAP=0,44 ou 0,45). Em geral, os descritores apresentaram maiores dificuldades nos conceitos, "Snow", "Beach", "Party"

e "Face". As características visuais utilizadas são a principal razão para justificar estas dificuldades, nomeadamente para conceitos mais complexos como "Party" e "Face" que requerem características específicas. Outro motivo está relacionado com a frequência de cada conceito na colecção sendo que os modelos de conceitos mais raros têm maior dificuldade na obtenção de um valor alto de Average Precision.

De notar o bom desempenho do descritor SIFT em relação aos restantes descritores nos conceitos "Indoor", "Manmade", "Nature" e "People". Estes conceitos caracterizam-se por pequenos detalhes de textura o que justifica estes resultados. Do lado oposto, no que diz respeito à eficiência computacional, aparece o descritor de momentos de cor que apresenta melhores resultados em dois conceitos difíceis de recuperar, "Beach" e "Snow". Isto acontece porque as imagens com estes conceitos exibem cores características mais fáceis de capturar no espaço HSV. Os descritores regiões de cor e banco de filtros de Gabor superam os restantes nos conceitos "Party" e "Face", respectivamente.

Depois da avaliação individual de cada descritor visual, foram realizados os mesmos testes mas combinando cor, textura e informação temporal. A tabela 8.2 apresenta os resultados obtidos [Jesus06], combinando os momentos de cor com o banco de filtros de Gabor, as regiões de cor com o descriptor SIFT e estes últimos com a informação temporal. A inclusão da informação temporal tem como o objectivo explorar a correlação temporal entre imagens. Utilizou-se o modelo descrito em 4.5.3.

As combinações de cor com textura obtiveram resultados melhores que os resultados obtidos por cada característica da combinação individualmente. Introduzindo o tempo no modelo, o desempenho global é melhor que utilizando qualquer descritor individual ou combinação. A inclusão do tempo supera a combinação regiões de cor com SIFT em todos os conceitos e apresenta melhores resultados que as outras alternativas (incluindo os resultados obtidos individualmente por cada descritor) nos conceitos "Indoor" e "Beach". Os maiores incrementos acontecem nos conceitos "Snow", "Party" e "Beach" que se referem a eventos onde as pessoas permanecem algum tempo num determinado local e tiram fotos com características idênticas. De salientar também que os melhores resultados para os conceitos "Outdoor" e "Face", foram obtidos pela combinação mais simples do ponto de vista computacional, momentos de cor e banco de filtros de Gabor.

Conceitos	Momentos cor + Filtro Gabor	Regiões cor + SIFT	Regiões cor + SIFT + Tempo
People	0,69	0,75	0,75
Face	0,58	0,44	0,45
Outdoor	0,91	0,87	0,89
Indoor	0,59	0,57	0,60
Nature	0,45	0,57	0,58
Manmade	0,61	0,71	0,73
Snow	0,17	0,09	0,13
Beach	0,26	0,34	0,42
Party	0,14	0,22	0,26
MAP	**0,49**	**0,51**	**0,53**

Tabela 8.2: MAP para diversos conceitos combinando várias características visuais e informação temporal. Foi considerada uma distância temporal máxima entre imagens de $d_{max} = 240$ segundos.

Na tabela 8.2, considerou-se a existência de correlação temporal para uma distância tempo-

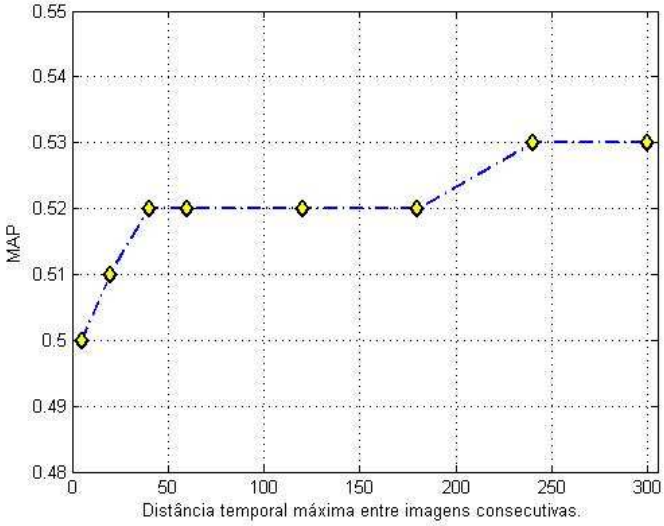


Figura 8.3: MAP para vários valores de d_{max}.

ral máxima de 4 minutos entre imagens consecutivas, valor obtido empiricamente. Na figura 8.3, mostra-se um gráfico com o MAP obtido considerando vários valores para a distâncias temporais. O MAP estabiliza para o valor $d_{max} = 240$ segundos.

Os resultados anteriores foram obtidos utilizando o método geral para treinar conceitos com várias características visuais. A técnica proposta também inclui métodos específicos, nomeadamente para o conceito "Face" (ver secção 4.5.1.1). Este modelo foi avaliado nas mesmas condições que os anteriores. O desempenho do método específico para detectar faces superou os resultados obtidos pelo método geral com as diversas características gerais. Este método obteve um AP=0,66 acima de 0,58 que foi o melhor resultado obtido com o método geral. A seguir, combinamos os momentos de cor e o banco de filtros de Gabor com o método específico e o resultado melhorou para AP=0,71. Estes resultados justificam a opção de incluir no modelo de análise semântica métodos específicos para conceitos mais complexos.

Com base no método específico para detectar faces, foi desenvolvido também um modelo para calcular o género da pessoa (masculino/feminino). Este método foi avaliado na base de dados "Labeled Faces in the Wild" [Huang07] que contém 13233 imagens com faces de 5749 indivíduos diferentes (1490 do género feminino e 4259 do género masculino). Para treinar o modelo foram utilizadas 490 de cada classe e as restantes 1000 foram utilizadas para teste. A percentagem de faces classificadas correctamente com o género foi de 81,40 (79,5 para o género feminino e 83,3 para o género masculino).

Para avaliar a pesquisa através da composição de uma interrogação com várias partes de imagens, foram testadas quatro interrogações diferentes com os seguintes objectivos:

- Procurar por fotos com edifícios com uma arquitectura específica - a interrogação é composta por várias partes de imagens que descrevem um conjunto diferente de características dos edifícios;

- Procurar por imagens com plantas de uma determinada espécie - a interrogação é composta por várias plantas de uma espécie;

- Procurar por imagens de neve - a interrogação é composta por partes de imagens de neve

e objectos relacionados;

- Procurar por imagens da casa do autor da tese - a interrogação é composta por vários objectos da casa.

Estes casos foram avaliados utilizando a medida precisão nas primeiras 100 imagens e os resultados foram comparados com o desempenho individual de cada imagem que contribui com uma parte para a imagem composta (ver tabela no apêndice C). Com a interrogação composta obtiveram-se resultados idênticos aos obtidos por uma das imagens usada para construir a composição. Esperava-se melhor desempenho com a composição de partes de imagens, dada a maior diversidade de informação fornecida, mas porque são criadas descontinuidades ao cortar parte de uma imagem os resultados ficaram abaixo do esperado. Contudo, esta técnica permite maior liberdade ao utilizador para construir a interrogação.

8.3.3 Anotação

O método de análise semântica proposto é também utilizado para anotação de imagens. Se a probabilidade de um conceito dada uma imagem, $p(w_i/I)$, for superior a um limiar, th, então a imagem é anotada automaticamente com o conceito (ver secção 4.4). A primeira experiência realizada visa encontrar o melhor th. Para avaliar os resultados foram utilizadas as medidas cobertura por palavra, precisão por palavra e cobertura por imagem. Na tabela 8.3, é apresentada a média da cobertura e da precisão por palavra para vários valores de th. São também apresentados os resultados do método utilizado em [Feng04, Li03] para anotação. Este método consiste em anotar uma imagem com os 5 conceitos com maior probabilidade, $p(w_i/I)$. Dado que os nossos testes são com 9 conceitos, apresentamos também os resultados do método utilizado em [Feng04, Li03] mas anotando os melhores 3 e os melhores 4 conceitos em cada imagem.

Métodos	Cobertura/palavra	Precisão
3 Melhores	0,22	0,59
4 melhores	0,30	0,55
5 melhores	0,40	0,53
Superior a th=0,5	**0,50**	**0,50**
Superior a th=0,6	0,41	0,52
Superior a th=0,7	0,32	0,56

Tabela 8.3: Média da cobertura e precisão obtida para várias palavras utilizando diversos métodos.

Os resultados da tabela mostram que a técnica proposta para anotação com $th = 0,5$ apresenta o melhor desempenho tendo em conta o compromisso entre a cobertura e a precisão por palavra. Em relação ao método utilizado em [Feng04, Li03], a técnica de escolher os "5 melhores" é a que se aproxima do método proposto para $th = 0,5$ mas com um registo pior, admitindo o critério anterior (igual importância entre cobertura e precisão). As restantes técnicas melhoram a precisão mas diminuem a cobertura.

Na tabela 8.4 apresenta-se a avaliação do desempenho para os métodos referidos anteriormente mas utilizando a cobertura média por imagem como medida. Mais uma vez, o método proposto com $th = 0,5$ supera os restantes. Este resultado (cobertura/imagem = 0,55) significa que em média cerca de metade das anotações realizadas em cada imagem são correctas. Com $th = 0,6$ os resultados são idênticos aos "5 melhores".

Métodos	Cobertura/Imagem
3 melhores	0,26
4 melhores	0,35
5 melhores	0,45
Superior a th=0,5	**0,55**
Superior a th=0,6	0,45
Superior a th=0,7	0,36

Tabela 8.4: Média da cobertura obtida em cada imagem da colecção pessoal utilizando vários métodos.

Na figura 8.4 são apresentadas algumas fotos da colecção pessoal com as anotações produzidas pelo método proposto para ilustrar os resultados obtidos. A figura 8.4 mostra pelo menos um exemplo correcto para cada conceito e várias imagens com várias anotações correctas (mínimo de 2 anotações correctas e máximo de 5). São também ilustrados erros nos conceitos "People", "Indoor", "Face", "Nature" e "Manmade".

Para finalizar a avaliação dos modelos semânticos na aplicação de anotação de imagens com esta colecção de fotos, são apresentados os resultados com a técnica proposta mas para cada conceito na tabela 8.5. Os conceitos "Manmade" e "People" obtiveram bons resultados admitindo o critério de compromisso entre as duas medidas. O conceito "People" consegue a melhor cobertura por palavra (93%) com um valor razoável de precisão por palavra (53%), o que não acontece com os conceitos "Face" e "Indoor" que conseguem uma cobertura por conceito alta mas produzem muitos falsos positivos. O conceito "Outdoor" apresenta o melhor valor de precisão por palavra mas um baixo valor de cobertura por palavra, principalmente por ser um acontecimento com uma frequência alta na colecção. O conceito "Snow" mantém um desempenho fraco tal como na recuperação de imagens.

Conceitos	Cobertura/palavra	Precisão
Beach	0,20	0,66
Face	0,66	0,42
Indoor	0,85	0,31
Manmade	0,64	0,59
Nature	0,36	0,40
Outdoor	0,32	0,85
Party	0,37	0,41
People	0,93	0,53
Snow	0,19	0,35
Média	**0,50**	**0,50**

Tabela 8.5: Cobertura e precisão por conceito utilizando o método proposto com $th = 0,5$.

8.3.4 Avaliação da Aplicação

Na secção 5.5 do capítulo 5, foi descrita a metodologia de concepção da aplicação para recuperar memórias pessoais em ambientes familiares. A metodologia é constituída pelas fases iniciais de definição de funcionalidade e testes com protótipos de baixa fidelidade, pelos protótipos de alta fidelidade e pelos testes de usabilidade utilizando um protótipo em PC. Como já foi referido, os passos iniciais não foram realizados para esta interface dado que as funcionalidades são idênticas às funções da aplicação Memoria Mobile que foi desenvolvida em

Anotação automática: Nature Outdoor People

Anotação automática: Nature Outdoor

Anotação automática: Face Nature Outdoor

Anotação automática: Beach Indoor ManMade People

Anotação automática: Indoor ManMade People Snow

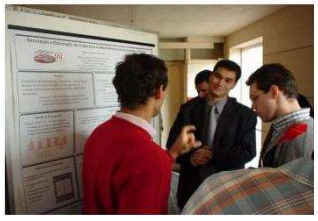
Anotação automática: Indoor ManMade Snow

Anotação automática: Beach Face ManMade Outdoor People

Anotação automática: ManMade Outdoor People

Anotação automática: Face ManMade Outdoor People

Anotação automática: Face Indoor Nature Party People



Anotação automática: Face Indoor ManMade Party People

Anotação automática: Indoor ManMade Snow

Figura 8.4: Anotação Automática - resultados obtidos utilizando o método "superior a th=0,5" para a colecção pessoal utilizada na aplicação Memoria Desktop.

primeiro lugar. Assim, a definição de funcionalidades e os protótipos em papel foram efectuados para a aplicação móvel e essa experiência foi útil na concepção da aplicação Memoria Desktop. A seguir, foram construídos vários *ecrãs* de alta fidelidade (não funcionais) que foram testados e consequentemente refinados por vários membros do nosso grupo. Nestes testes os utilizadores simulavam um conjunto de acções para cumprir as funcionalidades e com base em heurísticas [Nielsen90] sugeriam correcções na aplicação. Estes protótipos, depois de refinados, serviram de base para construir um protótipo inicial em computador pessoal que foi submetido a testes de usabilidade. A secção é dedicada a estes testes de usabilidade.

8.3.4.1 Testes de Usabilidade

Para analisar o protótipo em computador pessoal e testar a usabilidade da aplicação foram efectuados vários testes com utilizadores. Estes testes foram guiados por um questionário onde é pedido ao utilizador para realizar determinada tarefa e dar opinião sobre a sua experiência. De uma forma geral, os objectivos principais deste estudo são:

- Adquirir mais conhecimento sobre as necessidades dos utilizadores em relação à gestão de memórias pessoais;
- Avaliar as propostas para pesquisa e visualização de imagens;
- Produzir um conjunto detalhado de opiniões de utilizadores sobre a aplicação Memoria Desktop para guiar futuros melhoramentos da aplicação.

A seguir, é apresentada a metodologia utilizada para efectuar os testes, são caracterizados os utilizadores e são descritos os questionários usados neste processo. A secção termina com a apresentação dos resultados obtidos.

Método

Os testes foram realizados pelos estudantes da disciplina de Interacção Pessoa-Máquina do Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (DI/FCT/UNL). Os testes foram feitos individualmente mas a aplicação foi testada simultaneamente por vários alunos numa sala de aula utilizando a rede de computadores disponível.

No início dos testes foi feita uma pequena descrição dos objectivos aos utilizadores e foi dado a cada participante um questionário para o guiar durante o teste. O questionário, como é explicado na secção seguinte, descreve as tarefas que os utilizadores devem cumprir e apresenta questões relacionadas a que os participantes devem responder depois de completarem a tarefa correspondente.

Duas pessoas (facilitadores/observadores) supervisionaram os testes, encorajando os alunos a pensarem alto, ajudando-os no essencial, tirando notas sobre o desempenho e registando todos os problemas explicitamente mencionados. Os testes tiveram uma duração mínima de 20 minutos e uma duração máxima de 30 minutos, dependente da estratégia de cada utilizador em explorar a aplicação. Os objectivos principais destes testes dividem-se em dois tipos:

- Avaliação geral da aplicação
    1. Verificar se é fácil aprender a utilizar a interface;

2. Analisar os aspectos visuais da interface;
3. Medir a utilidade da aplicação.

- Avaliação específica para a recuperação de imagens

1. Verificar a utilidade das técnicas de pesquisa propostas;
2. Analisar a técnica proposta para definir interrogações (*drag & drop* para "Query Box");
3. Avaliar os resultados das pesquisas;
4. Verificar se os ícones utilizados são perceptíveis;

Esta informação foi recolhida com o objectivo de refinar a interface Memoria Desktop de acordo com as respostas dos utilizadores.

Participantes
Os testes foram realizados por 58 participantes voluntários, alunos da disciplina de Interacção Pessoa-Máquina do DI/FCT/UNL. Dez destes alunos eram do sexo feminino. A idade dos participantes desta experiência variou entre os 21 e os 31 anos com uma média de 23,5. A maioria dos participantes desta experiência tem máquina fotográfica digital e, em média, utilizam-na 2,7 vezes por mês e 24 vezes por ano. Quando questionados sobre o número de fotos da sua colecção pessoal as respostas diferem entre um mínimo de 10 e um máximo de 15000, com uma média de 2592 e uma moda de 1000. Apenas 14 participantes (24,1%) declaram que normalmente anotam as suas fotos.

No que diz respeito à forma como pesquisam fotos na sua colecção pessoal, a maioria dos participantes (67,2%) disse que explora as directorias do sistema utilizando o nome da directoria que é o nome do evento que procuram (este nome é atribuído manualmente na altura da cópia das fotos para o sistema). Alguns utilizadores (13,7%) também referem que utilizam a data do evento como critério de pesquisa. Apenas uma minoria dos utilizadores (8,6%) afirmam que usam software específico para gerir as suas colecções pessoais.

Quando questionados sobre qual o objectivo das pesquisas, as respostas enquadram-se no conjunto de objectivos resultantes do estudo realizado em [Rodden03]. O objectivo é encontrar um foto específica ou um conjunto de fotos, seja de um evento ou de vários eventos partilhando uma ou várias características. As resposta dividem-se pelos seguintes objectivos:

- Datas importantes, por exemplo, festas de Natal, aniversários ou casamentos (53%);
- Actividades de lazer, por exemplo, férias, viagens ou visitas a museus (47%);
- Pessoas, por exemplo, uma pessoa ou grupos de pessoas (26%);
- Locais, por exemplo, países, cidades ou locais de interesse (10%).

As pistas utilizadas na interrogação para satisfazer estes objectivos são a data (60,3%), o local (48,3%), o tipo de evento (20,7%) e as pessoas (12,1%).

Questionário
O questionário utilizado é constituído por três partes diferentes. A primeira parte destina-se a

recolher dados pessoais. Para além da idade e do género, a secção dos dados pessoais também inclui perguntas relacionadas com a utilização da máquina fotográfica digital, a dimensão da colecção pessoal e questões sobre a forma como o participante gere a sua colecção de fotos.

A segunda parte guia o utilizador na exploração da interface, explicando as tarefas que deve cumprir e capturando a informação referente à experiência do utilizador com a aplicação. O questionário apresenta quatro tarefas diferentes que os participantes devem realizar:

- Navegação na árvore de directorias;
- Visualização de fotos;
- Pesquisa de imagens utilizando a técnica de *drag & drop* de conceitos e operadores lógicos para a "query box";
- Pesquisa de imagens por composição de esboços com partes de imagens.

Depois de completar cada tarefa, os utilizadores têm de responder a várias questões relacionadas com a sua experiência a utilizar a interface. Esta parte do questionário é composta por dois tipos de questões: questões abertas e questões de resposta numa escala tipo Likert.

A última parte do questionário é reservada para o utilizador fazer uma avaliação global da aplicação. Esta parte é constituída por quatro afirmações relacionadas com a experiência adquirida a utilizar a aplicação Memoria Desktop. Cada participante indica o seu nível de concordância com a afirmação através da colocação de um círculo na resposta. As questões são de resposta numa escala tipo Likert. Nesta escala, a resposta 1 significa desacordo e a resposta 5 significa que o utilizador concorda com a afirmação.

Resultados

Como referido anteriormente, a segunda parte do questionário guia o utilizador através das quatro tarefas. Nesta secção são analisados os resultados destas tarefas e das questões finais de avaliação global. A seguir, são descritos os resultados das duas primeiras tarefas e depois os resultados das tarefas de pesquisa. A apresentação dos resultados termina com a análise dos resultados das questões de avaliação global.

Navegação na árvore de directorias e visualização de fotos

Na primeira tarefa, depois de utilizar a árvore de directorias para navegar na colecção pessoal, era pedido aos utilizadores para indicarem as dificuldades que encontraram enquanto executavam a tarefa. A maioria dos participantes referiu que não tinha tido dificuldades (74,1%). A dificuldade mais vezes referida pelos restantes participantes está relacionada com o facto da árvore de directorias não estar integrada na janela principal, sendo uma janela à parte que se pode sobrepor à janela principal.

Em relação à segunda tarefa, que consistiu na visualização de grupos de imagens em modo "Slideshow", 36,2% dos participantes afirmou não ter dificuldades e 48,3% considerou que os ícones não eram intuitivos e por isso tiveram dificuldades em encontrar o botão correcto para iniciar o "Slideshow". Alguns utilizadores sugeriram o uso de etiquetas (*tooltips*) para indicar as funções de cada botão. Outra crítica foi a ausência de métodos para controlar o "Slideshow", nomeadamente para avançar e recuar.

Pesquisa de fotos

A terceira tarefa era composta por várias pesquisas de imagens através da técnica de *drag & drop* de conceitos e operadores lógicos para "Query Box". Para completar esta tarefa, os utilizadores têm de realizar as seguintes pesquisas de imagens:

- Fotos com pessoas;
- Fotos ilustrando paisagens naturais;
- Imagens exteriores excluindo fotos tiradas em praias;
- Fotografias com pessoas ou paisagens naturais.

Para realizar as pesquisas, os utilizadores têm de construir interrogações, arrastando vários ícones representando conceitos e operadores lógicos para uma região designada por "Query Box". Depois de cada pesquisa, os utilizadores tinham de classificar os resultados utilizando uma escala com cinco valores, onde 1 representa mau e 5 excelente. Esta classificação é sumarizada na tabela 8.6.

Queries	Média	desvio Padrão	Moda
People	3,9	1	3
Nature	3,8	1	4
Outdoor AND No Beach	4	1	4
People Or Nature	3,5	1	4

Tabela 8.6: Avaliação feita pelos utilizadores aos resultados das pesquisas.

Em geral, os resultados obtidos nas pesquisas foram bons. Isto significa que os valores obtidos anteriormente, utilizando as medidas de avaliação (ver secção 8.3.2), para "People" e "Nature" são valores aceitáveis para os utilizadores. Na aplicação utilizada neste estudo o modelo do conceito "People" foi estimado usando os momentos de cor (AP=67%) e para o conceito "Nature" utilizou-se o banco de filtros de Gabor (AP=48%).

Depois de avaliar os resultados das pesquisas foram efectuadas uma série de questões relacionadas com a técnica de definição de interrogações. Primeiro, foi perguntado aos participantes se tinham tido dificuldades ao efectuar as pesquisas, no sentido de avaliar a técnica de definição de interrogações. Esta foi uma questão aberta e por isso foram registadas muitas opiniões diferentes. Enquanto 24% dos participantes não tiveram dificuldades os restantes apontaram alguns problemas. As dificuldades mais citadas foram:

- A obrigatoriedade de construir a interrogação sequencialmente. Se houvesse necessidade de inserir um conceito no meio de uma interrogação tinham de a reconstruir novamente (ausência de um comando para cancelar a última alteração);
- Os participantes indicaram algumas dificuldades na utilização dos operadores porque não sabiam as prioridades (ausência de definição de prioridades nos operadores);
- Alguns participantes gostariam de ter mais conceitos disponíveis (número de conceitos disponíveis baixo).

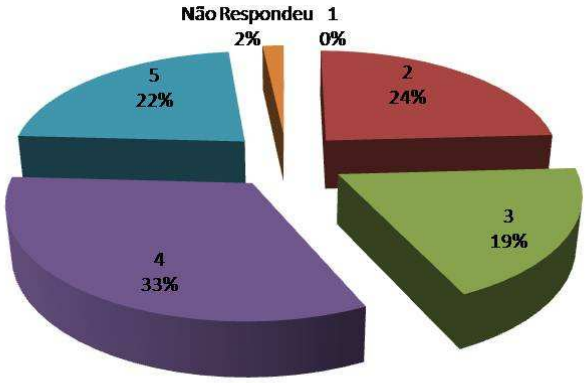


Figura 8.5: Resultados obtidos com a questão "Na sua opinião, o *drag & drop* é adequado para a tarefa de definição de interrogações?"

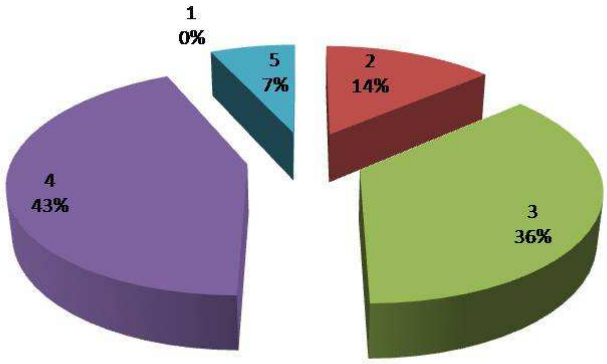


Figura 8.6: Resultados obtidos com a questão 'Na sua opinião, é perceptível a forma como deve combinar os ícones para obter um determinado resultado?'

A seguir, foi pedido aos participantes para responderem à questão "Na sua opinião, o *drag & drop* é adequado para a tarefa de definição de interrogações?" Numa escala com 5 valores, onde 1 significa inadequado e 5 muito adequado, a maioria dos utilizadores considerou adequado ou muito adequado (Moda = 4; Média = 3,54; Desvio Padrão = 1,1). A figura 8.5 mostra mais informação acerca das respostas dadas pelos participantes.

Na figura 8.6, são apresentados os resultados obtidos com outra questão realizada para validar a nossa opção na definição de interrogações, "Na sua opinião, é perceptível a forma como deve combinar os ícones para obter um determinado resultado?". Numa escala idêntica à questão anterior, a resposta mais vezes escolhida pelos utilizadores foi afirmativa (43%, adequado).

Um elemento importante na proposta para definir interrogações são os ícones que se arrastam para a "Query Box", por isso também foram efectuadas questões relacionadas com estes elementos. Em primeiro lugar, perguntamos aos utilizadores se os ícones representavam os conceitos adequadamente. A maioria dos participantes (81%) considerou os ícones perceptíveis enquanto que 17,2% deram a resposta oposta.

Os icones mais perceptíveis foram os que representam os conceitos "People" (escolhido por 23 participantes) e "Snow", que foi escolhido por 20 participantes. Pelo contrário, os menos perceptíveis foram os conceitos "Manmade" e "Beach", seleccionados por 40 e 38 participantes

respectivamente.

Os utilizadores elegeram o conceito "People" como o mais útil para fazer pesquisas de imagens em colecções pessoais (seleccionado por 34 participantes). Os conceitos "Party" com 12 e "Indoor" com 10 são os mais referidos em termos de utilidade dos conceitos. Alguns utilizadores fizeram sugestões para novos conceitos, por exemplo, "Sports", "Sea", "Day", "Night", "Winter", "Summer", "Spring" e "Autumn".

Na quarta tarefa era pedido aos participantes para executarem duas pesquisas utilizando como interrogação uma imagem composta por partes de várias imagens. Na primeira, os utilizadores tinham de procurar por imagens de edifícios com diferentes arquitecturas e o esboço tinha de ser constituído com partes de imagens rectangulares. Na segunda pesquisa, os participantes podiam seleccionar partes sem restrições geométricas na forma das partes (modo "Freehand") e o objectivo era procurar por faces de uma pessoa.

Depois de completarem a tarefa, os participantes tinham de responder a duas questões com resposta numa escala de 5 valores (1 significa mau e 5 excelente) para avaliar os resultados das pesquisas e a uma questão aberta com o objectivo de classificar a utilidade deste tipo de pesquisas. Na primeira questão, referente à avaliação da primeira pesquisa, a maioria dos participantes escolheu a resposta 3 (Moda = 3; Média = 2,59; Desvio Padrão= 0,97) e na segunda a maioria seleccionou a resposta 2 (Moda = 2; Média = 2,2; Desvio Padrão= 0,93). Tal como na avaliação anterior (ver secção 8.3.2) também os utilizadores não acharam os resultados muito bons. O desempenho deste método de pesquisa é inferior à pesquisa com conceitos, como era de esperar dado que a informação fornecida ao sistema é inferior. Mesmo assim, na questão aberta sobre a utilidade da funcionalidade para pesquisar imagens, a maioria dos participantes (67,2%) considerou-a útil porque permite interrogar visualmente a colecção de fotos e apenas 15,5% tiveram opinião contrária. Cinco participantes dos que deram resposta positiva acharam a funcionalidade adequada para procurar faces ou pessoas.

Avaliação geral da interface

A última parte do questionário tem como objectivo obter opiniões dos participantes sobre a aplicação na sua globalidade. Foi pedido aos utilizadores para avaliarem a interface no que diz respeito, à sua utilidade, à sua componente visual e à sua perceptibilidade.

A maioria dos participantes considerou a informação fornecida pelo sistema útil (Moda = 4, Média = 3,53; Desvio Padrão = 0,79). Os resultados foram semelhantes para a afirmação, "É fácil aprender a usar a aplicação", (Moda = 3, Média = 3,34; Desvio Padrão = 0,83) embora a resposta mais vezes atribuída se aproxime da posição neutra (Moda = 3).

Em relação às afirmações "O aspecto estético da interface agrada-me" e "Eu utilizaria esta aplicação para gerir as minhas fotos pessoais", os resultados são idênticos para o primeiro caso (Moda = 3; Média = 3,19; Desvio Padrão = 0,91) e para o segundo (Moda = 3 ; Média = 3,12; Desvio Padrão = 1,08). Os resultados são mais baixos que os valores obtidos pelas afirmações anteriores mas continuam próximos da posição neutral. Na figura 8.7, mostra-se mais informação sobre os resultados obtidos com a segunda afirmação, "Eu utilizaria esta aplicação para gerir as minhas fotos pessoais". Cerca de 26% das respostas (1 e 2) são negativas e 38% (4 e 5) são positivas. Apesar dos resultados não serem negativos esperava-se maior consenso nas respostas.

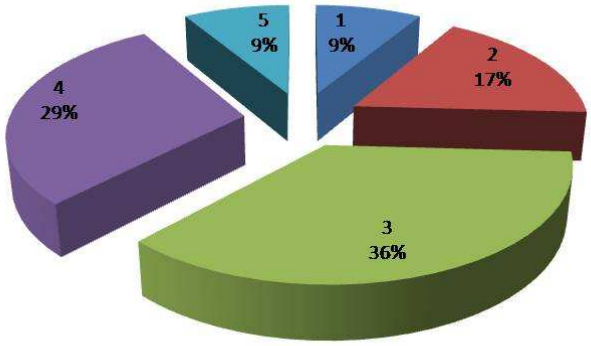


Figura 8.7: Resultados obtidos para a afirmação, "Eu utilizaria esta aplicação para gerir as minhas fotos pessoais"

Em geral, os resultados dos testes de usabilidade foram positivos e forneceram informação útil para melhoramentos futuros da interface. Em relação aos conceitos apresentados, alguns foram reconhecidos pelos utilizadores como úteis mas, da análise dos resultados, surge a ideia que os participantes preferiam escolher os seus próprios conceitos o que parece sensato dado que cada colecção pessoal tem características próprias. Outra aspecto a melhorar é a técnica proposta para definir interrogações. É necessário tornar mais versátil a correcção de erros na interrogação e definir prioridades nos operadores.

8.4 Recuperação de Imagens em Locais de Interesse

Esta secção descreve a avaliação da aplicação proposta para partilhar imagens durante a visita a locais de interesse. O estudo realizado para avaliar esta aplicação foi idêntico ao efectuado para o Memoria Desktop, por isso a secção tem uma organização semelhante à secção anterior: caracterização da base de dados, avaliação do sistema de recuperação, avaliação do sistema de anotação e avaliação da aplicação com testes de usabilidade. Foi utilizada a Quinta da Regaleira, um local histórico e cultural em Sintra, Portugal, como local de interesse para testar o Memoria Mobile e o PDA Fujitsu Siemens Pocket Loox 720.

8.4.1 Caracterização da Colecção de Imagens

Esta aplicação permite recuperar imagens disponibilizando ao utilizador vários tipos de pesquisa que podem ser combinadas numa "Query Box". Para avaliar o sistema de recuperação e o método de anotação de imagens foi utilizada uma base de dados com cerca de 1500 fotos da Quinta da Regaleira. Estas fotografias foram capturadas por vários membros do nosso grupo de investigação durante várias visitas à Quinta da Regaleira no âmbito do projecto InStory [Correia05]. A Quinta da Regaleira, na forma actual, foi construída no início do século XX com jardins, caves, torres, igrejas e um palácio. A figura 8.8 mostra o mapa da Quinta da Regaleira com alguns destes elementos. Na figura 8.9 são apresentadas algumas imagens das base de dados para ilustrar as características do local e da colecção pessoal.

Esta colecção de fotos tem características diferentes da base de dados utilizada no Memoria Desktop devido à natureza da aplicação. A colecção reflecte maior diversidade nas suas ca-

Figura 8.8: Mapa da Quinta da Regaleira.

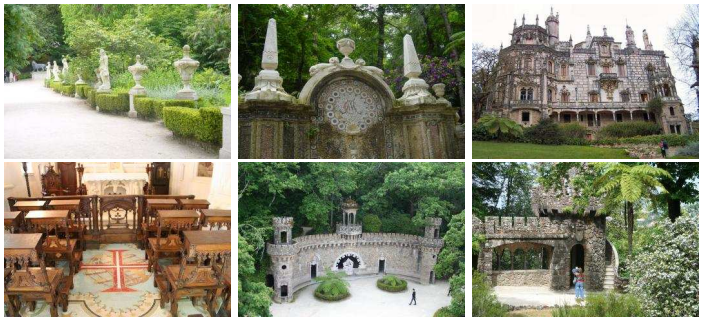


Figura 8.9: Fotos da colecção da Quinta da Regaleira.

racterísticas, dado que a captura foi obtida com diversos dispositivos. Cerca de 275 imagens têm informação de áudio, providenciada pelo utilizadores no instante de captura utilizando o Memoria Mobile. Deste conjunto, 205 foram anotadas pelo menos com uma palavra reconhecida. Em relação à informação de localização, a colecção é constituída por 250 imagens com dados obtidos pelo receptor de GPS. Todas as imagens da colecção são representadas por características visuais.

Os testes realizados para avaliar o sistema de recuperação incidem essencialmente em dois tipos de interrogação: pesquisa por imagem exemplo e pesquisa por conceitos (pode incluir também informação de localização). Por isso, houve necessidade de anotar manualmente as imagens da colecção para as interrogações utilizadas na avaliação. No caso das pesquisas por imagem exemplo, definiu-se um conjunto de tópicos relevantes para o local (alguns são idênticos aos utilizados nos conceitos com a diferença de que são representados por uma imagem como interrogação):

- "People", 226 fotos com pessoas;
- "Palace", 102 fotos do palácio da Quinta da Regaleira;
- "Nature", 869 imagens da colecção com paisagens naturais;
- "Tower", 57 fotos da torre;
- "Sculpture", 265 fotografias com esculturas;
- "Tiles", 31 imagens de azulejos;

Em relação às pesquisas através de conceitos, foi usado um subconjunto dos conceitos utilizados no Memoria Desktop, nomeadamente, "Outdoor", com 1110 imagens na colecção, "Indoor", com 314 imagens, "People", com 226 imagens, "Manmade", com 846 imagens e "Nature" com 869 imagens no colecção da Quinta da Regaleira. Os conceitos escolhidos poderiam ser adaptados para o local mas optou-se por usar um subconjunto dos conceitos da aplicação Memoria Desktop, continuando assim a usar conceitos incluídos no vocabulário definido pela LSCOM.

8.4.2 Sistema de Recuperação de Imagens

Esta secção divide-se numa primeira parte com os resultados das experiências realizadas para avaliar a pesquisa por imagem, isto é, as características visuais utilizadas e numa segunda parte onde são apresentados os testes efectuados com os conceitos utilizando informação visual, de áudio e de localização geográfica [Jesus07]. As medidas usadas para avaliar são a precisão nas primeiras 10 imagens (no Memoria Mobile são apresentadas 10 de cada vez) e o MAP. Escolhemos estas medidas dado que o ecrã do dispositivo móvel é pequeno, por isso não é possível apresentar muitas imagens e as que são apresentadas devem ser relevantes, até porque neste cenário (visita a um local histórico) os utilizadores também devem ter pouca tolerância para ver imagens que não correspondem às suas expectativas. Assim, quanto maior a precisão melhor será o desempenho desta aplicação.

Para avaliar as interrogações por imagem exemplo, foram seleccionadas quatro imagens descrevendo cada um dos tópicos ("People", "Palace", "Nature", "Tower", "Sculpture" e "Tiles") que foram utilizadas como imagens exemplo para pesquisar a base de dados. Na figura

Figura 8.10: Resultados obtidos com o vector de ocorrências de termos SIFT para uma imagem exemplo utilizada para recuperar esculturas.

8.10, são apresentados os primeiros resultados obtidos para uma imagem exemplo com uma escultura utilizando o vector de ocorrências de descritores SIFT para representar cada imagem da colecção. Para este exemplo, as quatro imagens correspondem ao tópico pretendido. Na tabela 8.7 é apresentada a média da precisão para os tópicos referidos e para várias características de textura. De notar que o valor de precisão apresentado para cada tópico é a média da precisão obtida pelas quatro imagens exemplo seleccionadas. Na primeira coluna de resultados da tabela 8.7 temos a precisão obtida com o banco de filtros de Gabor aplicados globalmente na imagem, na segunda temos o mesmos filtros de Gabor mas obtidos localmente nos pontos de interesse detectados pela SIFT e representados num vector de ocorrências. Nas restantes colunas mostram-se os resultados obtidos com o vector de ocorrências com descritores SIFT sem aplicar o LSA e aplicando o LSA.

Imagens (Interrogações)	Filtro de Gabor	Bag Gabor	Bag SIFT	LSA SIFT
Pessoas	0,40	0,50	0,55	0,55
Palácio	0,60	0,70	0,78	0,80
Natureza	0,98	0,95	0,93	0,85
Torre	0,13	0,25	0,35	0,40
Azulejo	0,17	0,10	0,17	0,30
Escultura	0,38	0,37	0,45	0,60
Média	**0,44**	**0,48**	**0,54**	**0,58**

Tabela 8.7: Resultados com texturas - precisão utilizando imagens exemplo como interrogação e considerando 10 imagens recuperadas.

Fizemos a mesma experiência para as características de cor (tabela 8.8) e combinando as características de cor e textura (tabela 8.9). Na tabela 8.8, são apresentados os resultados utilizando os momentos de cor no espaço HSV dividindo a imagem em 9 blocos iguais, utilizando um vector de ocorrências da característica anterior e o vector de ocorrência de regiões de cor no espaço LUV com e sem LSA. Na tabela 8.9, apresentamos os resultados obtidos combinando os momentos de cor com o banco de filtros de Gabor e os descritores SIFT com as regiões de cor com e sem o método LSA.

Analisando as três tabelas concluímos que:

- Quando combinadas as características de cor e textura (ver tabela 8.9), a diferença no desempenho entre usar vector de ocorrências e não utilizar aumenta (8%). Da mesma forma, também ao aplicar o LSA, o incremento no desempenho aumenta (11%). Em resumo, o desempenho é maior para vectores de ocorrência de descritores SIFT combinados com regiões de cor e aplicando o LSA;

Imagens (Interrogações)	Momentos	Bag Momentos	Bag Regiões	LSA Regiões
Pessoas	0,60	0,45	0,45	0,50
Palácio	0,48	0,25	0,53	0,50
Natureza	1,0	0,95	1,0	1,0
Torre	0,40	0,13	0,18	0,15
Azulejo	0,18	0,60	0,74	0,74
Escultura	0,65	0,73	0,60	0,70
Média	**0,55**	**0,52**	**0,58**	**0,60**

Tabela 8.8: Resultados com cor - precisão utilizando imagens exemplo como interrogação e considerando 10 imagens recuperadas.

Imagens (Interrogações)	Momentos Cor+Gabor	Regiões Cor+SIFT	LSA Regiões+ SIFT
Pessoas	0,45	0,50	0,60
Palácio	0,60	0,60	0,60
Natureza	1,0	1,0	1,0
Torre	0,10	0,27	0,58
Azulejo	0,27	0,60	0,87
Escultura	0,88	0,73	0,78
Média	**0,55**	**0,63**	**0,74**

Tabela 8.9: Resultados usando cor e textura - precisão utilizando imagens exemplo como interrogação e considerando 10 imagens recuperadas.

- Aplicando as características individualmente, a representação com vectores de ocorrência é melhor exceptuando para os momentos de cor (ver tabela 8.8);
- As características SIFT mostraram melhor desempenho do que as características extraídas utilizando o banco de filtros de Gabor.

Em relação aos conceitos, foi utilizado o modelo de cada conceito para pesquisar a colecção e foi realizada a avaliação dos resultados de dois modos: (1) medindo a precisão nas primeiras 10 imagens e (2) calculando o MAP. Também foi comparado o desempenho utilizando informação visual, de áudio e de localização individualmente e combinando dois e três tipos de informação. Na figura 8.11, mostra-se um exemplo da experiência realizada neste caso com o conceito "Manmade". As medidas de avaliação são calculadas sobre a lista de imagens resultantes da pesquisa.

Em primeiro lugar, medimos a precisão nas primeiras 10 imagens utilizando apenas características visuais, apenas palavras reconhecidas de áudio e combinando as duas. Os resultados são apresentados na tabela 8.10 e mostram que a informação de áudio melhora o desempenho e que em qualquer dos casos apresentados os resultados são bons (valor mínimo de MAP=0.74). Em média, temos no máximo 3 imagens erradas em 10. As características visuais utilizadas para realizar estes testes foram os momentos de cor e o banco de filtros de Gabor que são as que apresentaram menor desempenho nas pesquisas por imagem exemplo.

A seguir, foi realizada a mesma experiência mas em vez de ser calculado o MAP para toda a base de dados, foram processadas apenas as imagens localizadas numa zona, introduzindo a informação de localização (ver tabela 8.11) e neste caso a média da precisão baixou. Repetimos a experiência para outras localizações (ver tabelas em apêndice C) e, em geral, a precisão baixa apesar de aumentar para alguns conceitos dependendo das características do local. Por exemplo, a experiência apresentada na tabela 8.11 limita a zona de imagens à direcção norte a partir

Figura 8.11: Imagens mais relevantes para o conceito “Manmade” utilizando informação visual.

da Capela (ver mapa na figura 8.8). Pelas características da região seleccionada existem poucas fotos "Outdoor" e "Nature". Portanto, para estes conceitos a precisão baixa.

Conceitos (Interrogações)	Visual	Áudio	Visual e Áudio
Outdoor	1,0	0,8	0,9
Indoor	0,8	0,7	0,9
Nature	1,0	1,0	1,0
Manmade	0,8	1,0	1,0
People	0,3	0,8	0,8
Indoor + Manmade	0,3	0,7	0,5
Outdoor + Nature	1,0	0,8	0,8
Média	**0,74**	**0,83**	**0,84**

Tabela 8.10: Precisão para vários conceitos considerando 10 imagens recuperadas.

Conceitos (Capela, direcção Norte)	GPS e Visual	GPS, Visual e Áudio
Outdoor	0,5	0,6
Indoor	0,8	0,9
Nature	0,1	0,4
Manmade	0,8	0,9
People	0,3	0,4
Indoor + Manmade	0,3	0,4
Outdoor + Nature	0,4	0,4
Média	**0,46**	**0,57**

Tabela 8.11: Recuperação de imagens considerando uma localização (entrada da Capela) e uma direcção (Norte) - precisão para vários conceitos utilizando informação geográfica, visual e de áudio considerando 10 imagens recuperadas.

Para os mesmos conceitos repetimos a experiência mas calculando o MAP, isto é, em vez de medir a capacidade do sistema em colocar imagens relevantes nas 10 primeiras imagens (caso anterior), medidos a capacidade do sistema em colocar todas as imagens relevantes no topo da lista com os resultados. Na tabela 8.12, comparam-se os resultados obtidos por cada tipo de informação individualmente e combinados dois a dois. Na tabela 8.13, apresentam-se os resultados obtidos combinando os três tipos de informação. Analisando as tabelas conclui-se que:

- Os resultados baixaram em relação à experiência anterior para a informação visual e de áudio onde foi calculada a precisão nas primeiras 10 imagens (ver tabela 8.10);
- Adicionando as palavras reconhecidas do áudio ou restringindo a região de pesquisa com informação de GPS os resultados são melhores do que usando apenas as características visuais;
- Os ganhos na combinação dos três tipos de informação em relação à utilização de apenas dois não são evidentes;
- Quando a informação geográfica obtida por GPS é utilizada, o desempenho continua a depender das características da região. Contudo, a medida MAP compensa a ausência de imagens relevantes na região porque o AP é dividido pelo número de imagens relevantes na região.

Conceitos	Visual	Áudio	Visual e Áudio	Visual e GPS Dir	Visual e GPS 60m
Outdoor	0,86	0,74	0,86	0,60	0,97
Indoor	0,33	0,33	0,37	0,64	0,09
Nature	0,68	0,68	0,69	0,33	0,84
Manmade	0,32	0,74	0,70	0,83	0,68
People	0,23	0,18	0,27	0,44	0,20
Indoor + Manmade	0,21	0,34	0,26	0,45	0,17
Outdoor + Nature	0,75	0,57	0,75	0,44	0,84
MAP	**0,48**	**0,51**	**0,56**	**0,53**	**0,54**

Tabela 8.12: MAP para vários conceitos utilizando informação de áudio, visual e informação de localização (GPS) para definir uma região de 60 metros no Patamar dos Deuses (GPS 60m) e para seleccionar um conjunto de imagens na direcção Norte a partir da Capela (GPS Dir).

Conceitos	GPS 60m, Visual e Áudio	GPS Dir, Visual e Áudio
Outdoor	0,97	0,64
Indoor	0,09	0,63
Nature	0,86	0,46
Manmade	0,71	0,86
People	0,20	0,50
Indoor + Manmade	0,16	0,48
Outdoor + Nature	0,86	0,51
MAP	**0,55**	**0,58**

Tabela 8.13: MAP para vários conceitos combinando informação de áudio, visual e informação de localização para seleccionar uma região ou uma direcção em relação a um ponto. GPS 60 significa região de 60 metros no Patamar dos Deuses e GPS Dir representa direcção Norte a partir da Capela.

8.4.3 Anotação

Nesta secção avalia-se o método proposto para anotação de imagens na colecção da Quinta da Regaleira. As medidas utilizadas foram as mesmas usadas para avaliar a colecção pessoal do Memoria Desktop, precisão por palavra, cobertura por palavra e cobertura por imagem. Foi experimentado novamente o método proposto para vários limiares e comparado com os resultados obtidos com método utilizado em [Feng04, Li03] mas anotando uma imagem com os três conceitos com maior probabilidade, em vez dos cinco, porque o número de conceitos a anotar é baixo. Também foi testada esta técnica seleccionando para anotação os dois conceitos com maior probabilidade e o conceito com maior probabilidade. Na tabela 8.14 mostra-se a média da cobertura por palavra e precisão por palavra para estas técnicas e na tabela 8.15 apresenta-se a cobertura média por imagem.

Métodos	Cobertura/Palavra	Precisão
O melhor	0,19	0,55
2 melhores	0,37	0,57
3 melhores	0,57	0,58
Superior a th=0,5	0,57	0,58
Superior a th=0,6	0,39	0,59
Superior a th=0,7	0,20	0,59

Tabela 8.14: Média da cobertura e precisão obtida para diversos conceitos utilizando vários critérios para anotar imagens com os modelos semânticos.

Métodos	Cobertura/Imagem
O Melhor	0,15
2 melhores	0,31
3 melhores	0,46
Superior a th=0,5	0,46
Superior a th=0,6	0,30
Superior a th=0,7	0,17

Tabela 8.15: Média da cobertura obtida por imagem utilizando vários critérios para anotação.

Também nesta colecção de fotos o melhor limiar é o $th = 0,5$ (comparando a cobertura por imagem e a relação entre precisão e cobertura por palavra). Porém, desta vez anotando uma imagem com os 3 conceitos mais prováveis os resultados são iguais. A cobertura por imagem baixou em relação à base de dados utilizada no Memoria Desktop, portanto o método proposto obteve um desempenho inferior mas a precisão por palavra e a cobertura por palavra melhoraram em média porque não se aplicaram nesta colecção alguns conceitos com desempenho baixo, isto é "Party", "Snow" e "Beach". Na figura 8.12, ilustra-se o desempenho do método proposto para anotação com algumas imagens e as respectivas anotações. São apresentados exemplos em que o anotador automático erra em todas as anotações e exemplos com incorrecções nos conceitos "Indoor" e "People".

A tabela 8.16 apresenta a precisão e a cobertura obtidos para cada conceito. Os conceitos "Outdoor", "Indoor" e "People" têm um desempenho aproximado ao obtido com a colecção pessoal usada na aplicação para PC. Estes resultados mostram a capacidade de generalização dos modelos semânticos.

Em resumo, nesta base de dados o método proposto para anotação semântica mantém um

Anotação automática: ManMade No People Outdoor

Anotação automática: Indoor ManMade People

Anotação automática: Indoor ManMade People

Anotação automática: Indoor ManMade People

Anotação automática: Indoor ManMade People

Anotação automática: Indoor ManMade People

Anotação automática: Indoor Nature People

Anotação automática: Nature Outdoor People

Anotação automática: ManMade Outdoor People

Anotação automática: Nature No People Outdoor

Anotação automática: Nature No People Outdoor

Anotação automática: Nature Outdoor People

Figura 8.12: Anotação automática - resultados para o critério “superior a th=0.5”.

Conceitos	Cobertura/Palavra	Precisão
Outdoor	0,37	0,87
Indoor	0,81	0,27
Nature	0,18	0,59
Manmade	0,85	0,62
People	0,89	0,20
No People	0,31	0,94
Média	**0,57**	**0,58**

Tabela 8.16: Precisão e cobertura por palavra utilizando o método proposto para anotação com $th = 0,5$

desempenho aproximado ao obtido com a colecção usada no Memoria Desktop. Desta forma, analisou-se o comportamento do método de recuperação e anotação com outra base de dados com características diferentes.

8.4.4 Avaliação da Aplicação

O desenvolvimento da aplicação Memoria Mobile passou por vários fases: análise do local e definição de cenários, definição de funcionalidades, esboços da interface em computador, protótipos em papel e protótipo em dispositivo móvel. Nesta secção apresentam-se os resultados dos testes efectuados com os protótipos em papel. Ao mesmo tempo, na parte final de cada teste com os protótipos de baixa fidelidade os utilizadores testavam também uma versão simplificada em PDA com o objectivo de avaliar a técnica de *drag & drop*. A descrição destes testes e os resultados obtidos são apresentados nas secções seguintes.

8.4.4.1 Testes de Usabilidade

Esta secção tem como objectivo descrever os testes de usabilidade com protótipos em papel em PDA realizados para avaliar a aplicação. Esta informação é apresentada a seguir, começando pela metodologia usada nos testes, seguida da caracterização dos participantes e por fim dos resultados dos testes.

Método

Os testes foram realizados no laboratório do nosso grupo de investigação (IMG/CITI/FCT) por colegas do nosso grupo não envolvidos no projecto e por alunos. Os testes foram realizados individualmente por cada participante com a supervisão de dois observadores/facilitadores e foram gravados com uma câmara de vídeo (na figura 8.13 são apresentadas algumas imagens da sequência de vídeo capturada). Os testes tiveram uma duração máxima de 70 minutos. Cada teste é constituído por três fases:

- Para começar, os utilizadores eram encorajados a explorar a interface durante 45 minutos no máximo. O objectivo desta fase era analisar se os utilizadores percebiam os objectivos principais da aplicação;
- A seguir, era pedido ao utilizador para realizar três tarefas (recuperação de imagens utilizando três tipos de interrogação diferentes) utilizando os protótipos em papel e o protótipo em PDA. Tal como referido, foi incluído o protótipo em PDA nesta altura para analisar algumas características da aplicação.

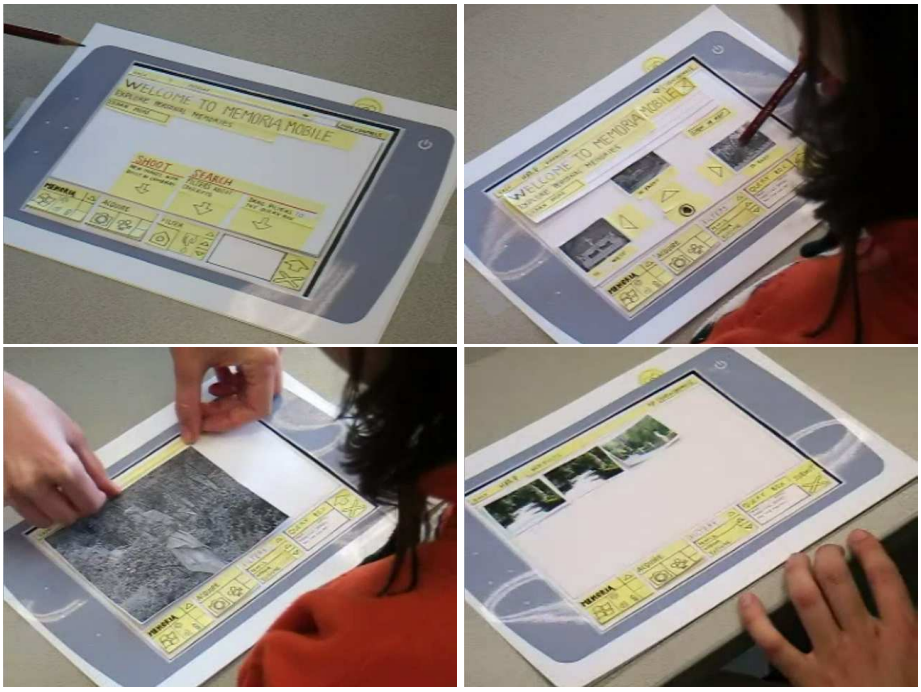


Figura 8.13: Protótipos em papel

- Finalmente, são realizadas entrevistas e os utilizadores respondem a um conjunto de questões acerca da sua experiência a usar a interface.

Depois de cada sessão é analisado o vídeo e a informação obtida pela entrevista na última fase da sessão e com estes dados são refinados os protótipos em papel. Como consequência, o último teste apresentou os melhores resultados. No final das quatro sessões toda a informação foi analisada outra vez para encontrar dificuldades comuns aos utilizadores.

Os objectivos principais destes testes foram:

- Analisar a utilidade das funcionalidades da aplicação;
- Verificar se é fácil aprender a utilizar a interface;
- Analisar o processo de recuperação de imagens, nomeadamente, a definição de interrogação e a visualização de resultados.

A seguir são caracterizados os utilizadores que participaram nos testes de usabilidade.

Participantes

Os testes com protótipos em papel foram realizados por quatro utilizadores. Três destes utilizadores, com as idades de 26, 27 e 40 anos, são membros do nosso grupo de investigação mas não participam no projecto. O quarto utilizador era estudante não universitário e tinha 16 anos. Nenhum destes utilizadores tinha na altura dos testes um PDA mas todos eram utilizadores de telemóveis e três deles trabalham em informática na área das aplicações multimédia.

Resultados

As quatro sessões de testes mostraram que:

- Os utilizadores perceberam que o objectivo principal da interface é permitir fazer pesquisas para recuperar mais informação (fotos) sobre o local;

- Nas primeiras intervenções, os utilizadores não tiveram dificuldades em interagir com a interface para realizar as funções mais comuns (capturar imagens, navegar em conjuntos de imagens, interagir com os menus e com o mapa);
- Todos os participantes consideraram interessantes as funcionalidades na perspectiva de integrarem as memórias pessoais com a visita a locais de lazer.
- Os utilizadores falharam as primeiras tentativas para fazerem pesquisa de imagens porque a técnica proposta (*drag & drop* para uma "Query Box") requer algum tempo de aprendizagem.

Após analisar os resultados, a técnica de interacção *drag & drop* foi a característica que criou problemas significativos de usabilidade na interface. Não há muita informação para apontar uma razão suficientemente forte mas parece-nos que esta técnica é lenta e difícil de utilizar em dispositivos móveis quando comparada com um simples clique como forma de interacção. No entanto, após algumas tentativas, todos os utilizadores descobriram esta técnica de interacção sem ajuda, com apenas um utilizador a falhar a sua reutilização (ocorrências exclusiva do protótipo em papel), isto é, depois de descobrir tudo se tornava mais simples. De notar que durante e no final de cada sessão a interface era refinada e que, na última sessão, o utilizador (também o mais novo) descobriu a característica *drag & drop* mais facilmente e, por isso, teve o melhor desempenho em todas as tarefas.

Com o protótipo em PDA, os utilizadores não tiveram dificuldades em descobrir a técnica *drag & drop* e nunca se esqueceram da forma como funcionava. Com base neste motivo e no efeito que a técnica produziu na percepção do sistema por parte dos utilizadores (combinação de vários elementos para definir interrogações), decidimos manter a característica *drag & drop*.

8.5 Aplicação Semi-Automática de Anotação

Esta secção visa apresentar os resultados dos testes efectuados para avaliar a aplicação Semi-Automática para Anotação. Em primeiro lugar, é avaliado o método proposto para anotação semi-automática baseado nos modelos semânticos [Jesus08], utilizando as medidas descritas na secção 8.2. A seguir, são apresentadas as experiências realizadas para validar o modelo proposto para calcular a pontuação do jogo Tag Around [Jesus08]. Finalmente, é avaliado o jogo Tag Around através de testes de usabilidade [Goncalves08a].

8.5.1 Anotação Semi-Automática

Nesta aplicação foi utilizada a colecção pessoal usada no Memoria Desktop e descrita na secção 8.3.1 e foi avaliado o desempenho dos nove modelos semânticos também utilizados na aplicação Memoria Desktop: "People", "Face", "Outdoor", "Indoor", "Nature", "Manmade", "Snow", "Beach" e "Party".

De modo a medir o desempenho do método semi-automático de anotação, aplicaram-se as medidas de avaliação, referidas anteiormente, aos resultados obtidos com o conjunto de treino inicial (CT Inicial) e aos resultados obtidos treinando os modelos com mais 20 (CT Inicial + 20) e 40 (CT Inicial + 40) imagens de cada conceito no conjunto de treino. As imagens adicionadas foram seleccionadas aleatoriamente da colecção pessoal utilizando a informação providenciada

pela anotação manual, isto é, para cada conceito eram seleccionadas imagens anotadas manualmente com o conceito. Desta forma, foi simulado o algoritmo semi-automático para anotação descrito na secção 4.4.2 que acrescenta ao conjunto de treino as imagens anotadas pelo jogador.

Em primeiro lugar avaliou-se o sistema semi-automático utilizando as medidas usadas nas aplicações de recuperação. Foi utilizado o AP para avaliar cada conceito e o MAP para uma avaliação global. Os resultados são apresentados na tabela 8.17. Como era esperado, utilizando mais imagens no conjunto de treino os modelos semânticos apresentam melhor desempenho. Os maiores incrementos são conseguidos pelos conceitos com valores mais baixos de AP. Isto significa que a inclusão de mais imagens no conjunto de treino destes conceitos permitiu aumentar a capacidade de generalização dos modelos. Globalmente, há um salto elevado no MAP quando se acrescentam as primeiras 20 imagens no conjunto de treino. O mesmo não acontece quando se acrescentam mais 20 ao conjunto anterior. Isto deve-se ao facto do conjunto de treino passar a incluir imagens da colecção de teste que estão mais correlacionadas com as restantes do que as imagens do conjunto de treino inicial.

Conceitos	CT Inicial	CT Inicial + 20	CT Inicial + 40
People	0,75	0,81	0,82
Face	0,45	0,63	0,69
Outdoor	0,89	0,96	0,96
Indoor	0,60	0,78	0,80
Nature	0,58	0,80	0,83
Manmade	0,73	0,90	0,92
Snow	0,13	0,71	0,82
Beach	0,42	0,60	0,66
Party	0,26	0,44	0,54
MAP	**0,53**	**0,74**	**0,78**

Tabela 8.17: MAP obtido utilizando vários conjuntos de treino (CT) na aprendizagem dos modelos. Conjunto de treino inicial, com mais 20 e 40 imagens de cada conceito.

A seguir, com os modelos estimados com os três conjuntos de treino referidos anteriormente, foi aplicado o método proposto para anotação automática de imagens e medido o desempenho dos modelos com a precisão por palavra, cobertura por palavra e cobertura por imagem. Na tabela 8.18, são apresentados os valores médios destas medidas. As três medidas aumentam à medida que o número de imagens no conjunto de treino aumenta. A cobertura por palavra e a cobertura por imagem têm um comportamento idêntico ao MAP da tabela 8.17, contudo, a precisão por palavra sobe apenas 2% por mais 20 imagens no conjunto de treino. O sistema semi-automático aumenta o número de anotações correctas mas produz falsos positivos e por isso os incrementos na precisão são mínimos.

Medidas	CT Inicial	CT Inicial + 20	CT Inicial + 40
Precisão/Palavra	0,50	0,52	0,54
Cobertura/Palavra	0,50	0,75	0,78
Cobertura/Imagem	0,55	0,69	0,72

Tabela 8.18: Valores médios de precisão por palavra, cobertura por palavra e cobertura por imagem obtidos utilizando vários conjuntos de treino (CT) na aprendizagem dos modelos.

Nas tabelas 8.19 e 8.20 são apresentados os resultados obtidos por cada conceito. Em relação à precisão por palavra, aumenta à medida que o número de imagens no conjunto de treino

aumenta para a maioria dos conceitos exceptuando para os conceitos "Beach" e "Party". O conceito "Snow" baixa a precisão quando se somam 20 imagens mas aumenta quando se somam 40 imagens ao conjunto de treino. No que diz respeito à cobertura por palavra, a maioria dos conceitos aumenta a cobertura por palavra com excepção dos conceitos "People" e "Face" que diminuem com o aumento de imagens no conjunto de treino. Para o conceito "Manmade" o valor da cobertura sobe para mais 20 imagens mas depois desce para mais 40.

Conceitos	CT Inicial	CT Inicial + 20	CT Inicial + 40
Beach	0,66	0,36	0,32
Face	0,42	0,46	0,51
Indoor	0,31	0,43	0,47
Manmade	0,59	0,71	0,75
Nature	0,40	0,61	0,64
Outdoor	0,85	0,93	0,93
Party	0,41	0,27	0,26
People	0,53	0,58	0,60
Snow	0,35	0,32	0,37
Média	**0,50**	**0,52**	**0,54**

Tabela 8.19: Precisão por conceito obtida para vários conjuntos de treino (CT) na aprendizagem dos modelos e utilizando o método superior a th=0,5.

Conceitos	CT Inicial	CT Inicial + 20	CT Inicial + 40
Beach	0,20	0,81	0,89
Face	0,74	0,69	0,65
Indoor	0,85	0,87	0,87
Manmade	0,64	0,76	0,75
Nature	0,36	0,56	0,66
Outdoor	0,32	0,58	0,64
Party	0,37	0,78	0,86
People	0,93	0,89	0,88
Snow	0,19	0,77	0,82
Média	**0,51**	**0,75**	**0,78**

Tabela 8.20: Cobertura por conceito obtida para vários conjuntos de treino (CT) na aprendizagem dos modelos e utilizando o método superior a th=0,5.

8.5.2 Pontuação

De forma a avaliar o modelo usado para calcular a pontuação de cada jogada foram realizadas várias simulações em Matlab. Efectuou-se uma simulação com 200 jogadas seleccionando aleatoriamente 200 imagens da colecção pessoal. Para anotar cada imagem, foram simulados dois tipos de jogadores com comportamentos diferentes:

- Tipo 1 - jogador que joga com o objectivo de conseguir a maior pontuação possível fazendo anotações correctas. Para este tipo de jogador associamos anotações correctas utilizando as anotações manuais. Simulamos aleatoriamente alguns erros (5% de anotações erradas);
- Tipo 2 - jogador com desempenho fraco que comete muitos erros nas anotações. Para este tipo de jogador simulamos aleatoriamente 50% de erros nas 200 anotações.

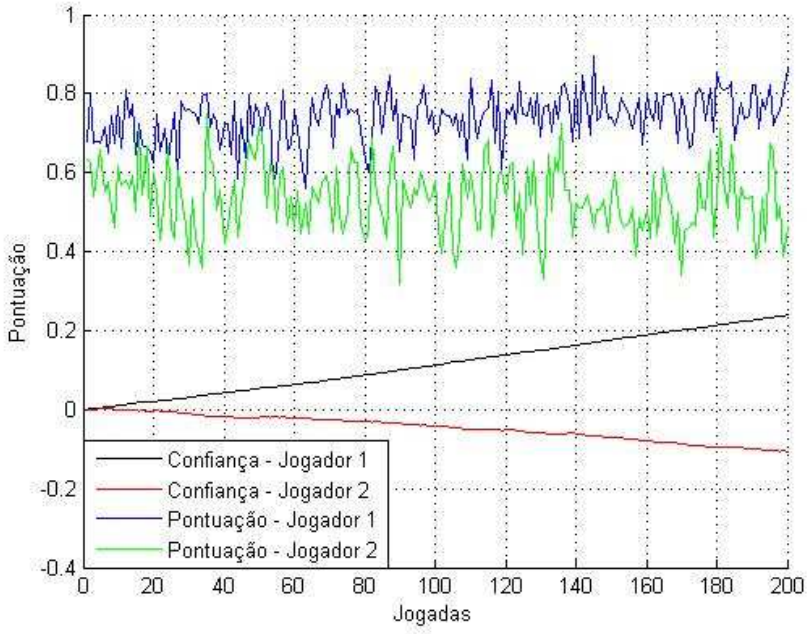


Figura 8.14: Média da pontuação e da confiança obtida por 10 jogadores do tipo 1 (5% de erros) e 10 jogadores do tipo 2 (50% de erros).

Foram feitas dez simulações com jogadores do tipo 1 e dez com jogadores do tipo 2. Na figura 8.14 são apresentadas curvas resultantes destas simulações. A curva azul representa a média da pontuação obtida pelos 10 jogadores do tipo 1 em cada jogada e a curva a verde representa o mesmo valor mas obtido pelos 10 jogadores do tipo 2. A figura mostra também a média da confiança obtida pelos 10 jogadores de cada tipo após cada jogada. Em geral, a curva dos jogadores do tipo 1 apresenta valores mais altos do que a curva dos jogadores do tipo 2. A confiança no sistema cresce para os jogadores bem comportados e decresce para os jogadores que cometem muitos erros.

Os valores da pontuação obtida em cada jogada pelo método descrito na secção 7.3.2.1 estão compreendidos entre 0 e 1 mas para tornar o jogo mais apelativo multiplicamos este valor por 100 e é este resultado que é apresentado ao utilizador. Na tabela 8.21 são apresentadas as médias das pontuações totais obtidas pelos jogadores do tipo 1 e do tipo 2, para os modelos estimados com o conjunto de treino inicial e com o conjunto de treino final (CT Final). Como esperado, a pontuação total para os jogadores do tipo 1 é mais alta do que para os jogadores do tipo 2. Quando o conjunto de treino aumenta esta diferença aumenta. Anotações correctas providenciam mais imagens para serem incluídas no conjunto de treino, por consequência são estimados modelos semânticos com melhor desempenho que permitem calcular com maior exactidão a pontuação de cada jogada.

Estas simulações mostram que o modelo proposto para calcular a pontuação se ajusta ao comportamento destes dois tipos de classes de jogadores.

8.5.3 Avaliação da Aplicação

A aplicação Tag Around em todas as suas vertentes incluindo a interface visual, os mecanismos de interacção, o motor de jogo e a fórmula de pontuação foi desenvolvida através de um modelo iterativo baseado num processo cíclico de prototipagem, testes de usabilidade, discussão

Jogadores	CT Inicial	CT Final
Tipo 1 (5% Erros)	14156	14792
Tipo 2 (50%Erros)	10984	10651

Tabela 8.21: Média da pontuação total obtida por 10 jogadores do tipo 1 e 10 jogadores do tipo 2 utilizando o conjunto de treino (CT) inicial e o conjunto de treino com mais 40 imagens anotadas pelo jogo.

de ideias e depois refinamentos do trabalho desenvolvido. As fases do projecto, desde os passos iniciais de análise e definição de funcionalidades, testes com protótipos em papel e testes com protótipo funcional em computador pessoal foram apresentadas na secção 7.7 do capítulo 7. Nesta secção são descritos e discutidos os testes que envolveram utilizadores.

8.5.3.1 Testes de Usabilidade com Protótipos em Papel

Um dos aspectos mais relevantes na aplicação é a utilização de uma câmara de vídeo como suporte da interacção baseada em gestos para anotar imagens. Para avaliar esta característica houve necessidade de utilizar tecnologia (uma câmara e um projector) em conjunto com os protótipos em papel, à semelhança do que já tinha sido feito com o Memoria Mobile. Os testes com protótipos em papel (ver figura 8.15) incluíram uma série de tarefas que os utilizadores tinham de realizar. Os utilizadores tinham conhecimento das tarefas mas desconheciam o tipo de interacção utilizada. A seguir, é apresentado o método utilizado e são descritos os participantes e os resultados [Goncalves08].

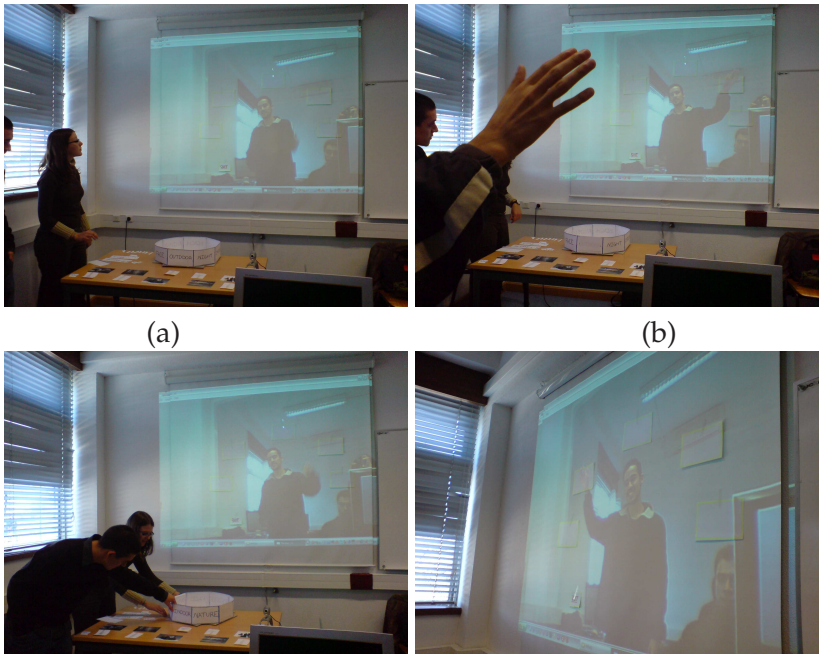
Método
Foi utilizada uma câmara de vídeo para capturar os movimentos do utilizador, sendo projectados num plano em que fiquem todos visíveis para o jogador (ver figura 8.15). Nesse plano, foram coladas marcas com folhas de papel para indicar os pontos de interacção. Desta forma, alterando a localização das folhas era possível encontrar a melhor posição para as zonas de interacção. Os esboços em papel da aplicação eram representados em cima de uma mesa como é ilustrado na figura 8.15.

Os testes são constituídos por três fases:

- Para começar, é pedido aos utilizadores para explorarem a aplicação sem nenhum conhecimento acerca dos seus objectivos nem da forma de interacção;
- A seguir, é dito aos utilizadores quais os objectivos e como é jogado o jogo. Os participantes utilizavam a aplicação sem restrições temporais nem pontuação;
- Finalmente, é pedido aos participantes para realizarem anotações correctas num tempo limite e avaliadas por uma pontuação.

A secção seguinte caracteriza os utilizadores que participaram nestes testes.

Participantes
Os testes com protótipos em papel foram realizados por cinco participantes, todos alunos da disciplina de Interacção Pessoa Máquina do curso de Engenharia de Informática do DI/FCT/UNL. Todos os participantes têm experiência em usar tecnologia e muita experiência a usar computadores. Os testes foram realizados numa sala de aula.



(a) (b)

(c) (d)

Figura 8.15: Protótipos em papel: a) Cenário construído para realizar os testes; b) Marcas de papel sobre o vídeo do utilizador para definir zonas de interacção.

Resultados

As principais ideias resultantes dos testes com protótipos em papel foram:

- Inicialmente os utilizadores tiveram dificuldades com o tipo de interacção utilizado, nomeadamente, em rodar imagens e palavras e simultaneamente controlar a pontuação e o tempo de jogo.
- Depois de algum tempo de aprendizagem, os participantes começaram a gerir melhor as diversas funções do jogo e começaram a divertir-se. Isto foi um resultado importante, os utilizadores esqueceram-se que estavam a anotar imagens;
- Alguns utilizadores indicaram dificuldades em perceber a exactidão da anotação que não era evidente consultando a pontuação. Como consequência, introduzimos um conselheiro, uma representação de um cientista que surge no ecrã e com alguma ironia faz um comentário sobre a correcção da anotação.

Depois de realizados testes com vários utilizadores e de refinados os protótipos em papel foi implementado um primeiro protótipo em computador.

8.5.3.2 Testes de Usabilidade com a Aplicação

O jogo Tag Around foi avaliado com o objectivo de medir a complexidade da interface, avaliar a utilidade da aplicação, corrigir e avaliar os aspectos visuais da interface, verificar se a interface é familiar (de fácil aprendizagem) e analisar a componente de diversão do jogo. Os resultados dos testes [Goncalves08a] são descritos de seguida. Primeiro, é explicada a metodologia usada, depois são caracterizados os participantes e feita uma descrição do questionário utilizado. No fim são apresentados os resultados.

Método

Os testes foram conduzidos por dois investigadores no papel de observadores, num laboratório da universidade e foram realizados individualmente por cada utilizador. No início era dada aos participantes uma pequena explicação sobre os objectivos do teste. Depois de uma breve descrição da aplicação e uma explicação dos objectivos a atingir durante os testes, os jogadores eram encorajados a explorar o jogo Tag Around. Os participantes tinham que iniciar a aplicação e fazer *login*.

A seguir a esta fase introdutória, os participantes começavam um jogo novo e tinham de catalogar o máximo número possível de imagens de acordo com conceitos disponíveis no sistema. Durante os testes os utilizadores eram persuadidos a pensar alto e podiam pedir ajuda aos observadores se não soubessem como avançar no jogo. Todos os comentários foram registados e analisados no final dos testes em conjunto com os questionários. Quando o jogo termina, esta fase fica concluída e é pedido aos utilizadores para preencherem um questionário onde expressam as suas opiniões sobre a aplicação que acabaram de testar.

Este teste visa perceber se a aplicação é divertida, útil, envolvente, se é fácil de usar e se é fácil aprender a utilizá-la. Para além disso, é também objectivo encontrar novas aproximações para melhorar a dinâmica da interacção e os aspectos visuais da interface.

Cada teste durou no máximo 30 minutos, dependendo do desempenho dos utilizadores, uma vez que uma anotação errada causa perda de energia e o jogo termina quando acaba a energia. No fim, toda a informação coleccionada é analisada com o objectivo final de refinar a aplicação.

Participantes

Os testes foram realizados por 15 participantes voluntários, 8 do sexo feminino. Os participantes desta experiência tinham idades entre os 18 e os 31 anos com uma média de 24 anos. Dez dos utilizadores trabalham na área das tecnologias de informação. O primeiro contacto com a aplicação por parte dos participantes foi durante os testes e nas mesmas condições.

Todos os participantes utilizam frequentemente a Internet para procurar imagens e todos afirmam usar computadores para gerir as suas imagens pessoais mas apenas 50% o faz com frequência. Os participantes também declararam que em média apenas catalogam cerca de metade das suas colecções de imagens. Quando procuram por uma foto particular nos seus computadores, todos (excepto 3 que não deram resposta) afirmam usar a árvore de directorias para fazer a pesquisa. Um dos participantes declarou usar o IPhoto. Os participantes usam as suas fotos principalmente para memória futura das experiências vividas.

Questionário

O questionário visa recolher a informação pessoal dos participantes e a informação relativa à experiência dos utilizadores com a aplicação. É composto por cinco secções: dados pessoais, motivação, dinâmica de jogo, interacção e aspectos visuais. Também inclui várias questões abertas.

Os dados pessoais incluem a idade, o género e a forma como gerem as suas fotos. A experiência de cada participante é medida por vinte e quatro questões distribuídas por quatro secções (motivação, dinâmica de jogo, interacção e aspectos estéticos). Três das perguntas são questões

abertas que visam recolher sugestões relacionadas com as alterações que possam contribuir para melhorar a interface. As restantes questões são de resposta numa escala tipo Likert, onde 1 representa desacordo total com a questão e 5 acordo total.

Resultados

Nesta secção são descritas as conclusões preliminares mais importantes e as observações obtidas durante os testes. As opções seleccionadas pelos diferentes participantes para cada questão são analisadas e é calculada a média da pontuação atribuída a cada resposta para observar se existe uma tendência geral para concordar ou discordar com a afirmação correspondente. É também calculado o desvio padrão da média para avaliar a extensão do consenso em relação à afirmação.

Facilidade em utilizar

Esta secção serve para tirar algumas dúvidas que surgiram nos testes com protótipos em papel no que diz respeito à facilidade de utilização da interface. Vários utilizadores sentiram dificuldades para encontrar a forma como começar a jogar. Alguns necessitaram da ajuda dos observadores que estavam a supervisionar os testes para avançar. Isto aconteceu porque o detector de movimento da zona destinada ao início do jogo estava calibrado para detectar movimentos muito pequenos o que à vezes confundiu o jogador acerca da forma como interagir com a aplicação (a jogar os movimentos são maiores).

Quando os utilizadores usaram a aplicação pela primeira vez, precisaram de algum tempo para perceber o paradigma de interacção utilizado. No entanto, rapidamente perceberam o que fazer. A maioria dos participantes concordou com a frase "É simples aprender a utilizar esta aplicação" (Média = 4,27; Desvio Padrão = 0,57) e também com "É simples usar esta aplicação" (Média = 4,33; Desvio Padrão = 0,60).

Interacção

Os participantes tendem a concordar com a questão "É fácil interagir com as zonas específicas usadas para rodar imagens/conceitos?", o valor da média indica uma opinião generalizada próxima da posição neutra (Média = 3,33; Desvio Padrão = 1,10), no entanto, a resposta mais vezes atribuída foi a 4 (ver figura 8.16).

A questão "A utilização deste tipo de interacção é fisicamente desgastante?" foi mais controversa (ver figura 8.17) porque a maioria dos participantes discorda (Média = 2,10; Desvio Padrão = 1,20) mas um discorda totalmente, outro discorda parcialmente e três têm uma posição neutra. O mesmo não acontece com a afirmação "Este tipo de interacção é mentalmente exigente" que obteve um conjunto de respostas mais consensuais (ver gráfico 8.18, a maioria da respostas são, discordo totalmente) .

No sentido de questionar a opção em relação à técnica de interacção usada, foi feita a questão "A aplicação seria mais intuitiva se fosse utilizado um teclado e um rato em vez dos gestos?". A maioria dos participantes discordou desta afirmação (Média = 2,53; Desvio Padrão = 1,31), contudo, esta questão foi muito controversa (ver figura 8.19): 5 participantes discordaram totalmente, 1 concordou totalmente e 3 concordaram parcialmente.

Embora este resultados não sejam tão convincentes como esperado, existe uma tendência para concordar com as técnicas de interacção propostas.

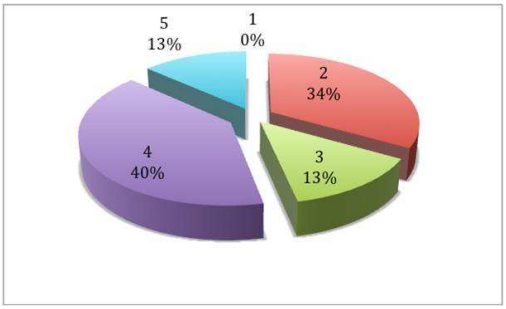


Figura 8.16: Resultados obtidos com a questão "É fácil interagir com as zonas específicas usadas para rodar imagens/conceitos"?

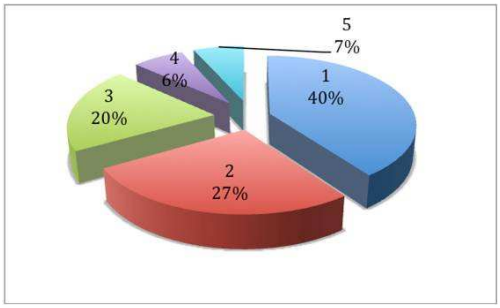


Figura 8.17: Resultados obtidos com a questão "A utilização deste tipo de interacção é fisicamente desgastante?"

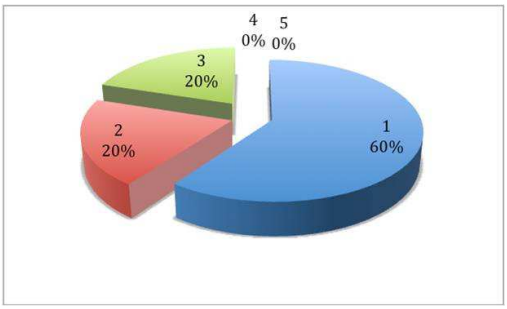


Figura 8.18: Resultados obtidos com a questão "Este tipo de interacção é mentalmente exigente?"

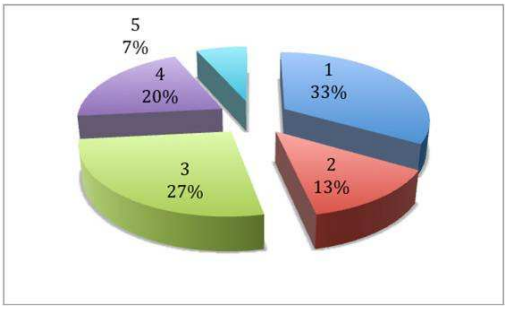


Figura 8.19: Resultados obtidos com a questão "A aplicação seria mais intuitiva se fosse utilizado um teclado e um rato em vez dos gestos?"

Características da Aplicação

Nesta secção foram avaliados vários elementos visuais e estéticos da interface. A interface inclui as imagens a serem catalogadas, os conceitos disponíveis para catalogar, as palavras já associadas com a imagem corrente, a pontuação e o tempo já decorrido. O objectivo é compreender qual a percepção do utilizador em relação a estes elementos e a toda a dinâmica da aplicação e também identificar quais as preferências do utilizador.

A tabela 8.22 apresenta o conjunto de questões realizadas e resume os resultados obtidos. A afirmação "Consigo perceber como a pontuação vai mudando ao longo do tempo" teve respostas muito diferentes da parte dos utilizadores (ver tabela 8.22). Dois participantes discordaram totalmente enquanto outros dois concordaram totalmente. O cálculo da pontuação é baseado em vários factores e dado que os utilizadores estavam mais concentrados nas suas acções, não tiveram tempo para examinar em detalhe como a pontuação é processada. Contudo, todos os participantes detectaram correlação entre a pontuação e a exactidão da anotação realizada.

Cada participante teve exactamente a mesma opinião em relação às afirmações,"As imagens deveriam estar paradas, apenas as anotações deveriam rodar" e "As anotações deveriam estar paradas, apenas as imagens deveriam rodar". A maioria dos participantes discorda totalmente de ambas as frases. Isto indica que o facto da aplicação permitir que os utilizadores rodem as imagens e as anotações parece ser uma opção adequada.

Em relação à frase "A aplicação funcionaria melhor com mais imagens", a maioria dos utilizadores (9 em 15) deu uma opinião neutral (Média = 3,07; Desvio Padrão = 0,77). Opiniões muito semelhantes foram atribuídas à afirmação "A aplicação funcionaria melhor com mais anotações". Estas respostas não foram muito conclusivas. Para adquirir mais informação devem ser realizados testes comparativos.

Questões	μ	σ
"Consigo perceber como a pontuação vai mudando ao longo do tempo"	3,2	1,26
"Percebi que estava a fazer boas ou más anotações"	4,06	0,88
"As imagens deveriam estar paradas, apenas as anotações deveriam rodar"	1,47	0,74
"As anotações deveriam estar paradas, apenas as imagens deveriam rodar"	1,47	0,74
"A aplicação funcionaria melhor com mais imagens"	3,07	0,80
"A aplicação funcionaria melhor com mais anotações"	3	0,87
"O aspecto estético da interface agrada-me"	3,8	0,68
"Considero, em termos gerais, uma interface agradável"	4,2	0,56

Tabela 8.22: Média e desvio padrão do conjunto das respostas relacionadas com as caracterização da interface.

Para concluir a avaliação das características da interface, a maior parte dos participantes afirmou gostar da interface (ver tabela 8.22).

Utilidade

Um dos objectivos principais destes testes de usabilidade é mostrar que a aplicação proposta representa uma aproximação válida para anotar imagens de forma divertida e que pode ser utilizada em diversos locais. Para avaliar a utilidade da aplicação nas condições referidas foi realizado um conjunto de questões relacionadas. Os resultados obtidos são apresentados na tabela 8.23. Para cada questão são calculados a média, o desvio padrão e também a percentagem

de respostas em que os participantes concordam ou concordam totalmente.

A maioria dos participantes declarou "É divertido utilizar esta aplicação". Quando questionados se "Usaria esta aplicação num local público para passar o tempo?", a maioria concordou totalmente ou parcialmente, apenas três mantiveram uma posição neutra. Um atitude semelhante foi detectada quando os utilizadores foram questionados se "Utilizaria esta aplicação para me divertir com amigos/família". Estes são resultados importantes porque o objectivo era construir uma aplicação para anotar imagens de forma divertida em lugares públicos.

Em relação à utilidade da aplicação para anotar imagens, a maioria dos participantes respondeu que utilizaria o jogo Tag Around para uso pessoal. A maioria concordou com a frase "Seria mais divertido usar imagens minhas com as minhas próprias anotações". Porém as respostas a uma questão relacionada "Utilizaria esta aplicação para anotar as minhas imagens?", não foram muito consensuais: duas respostas que concordam totalmente, sete concordam parcialmente, três têm opinião neutra e os restantes três discordam parcialmente.

Questões Abertas

Da análise das questões abertas e dos comentários feitos pelos participantes durante o teste, foi possível recolher algumas ideias que ajudarão a melhorar a aplicação testada. Os dados recolhidos e as soluções resultantes para refinar a aplicação são indicados de seguida.

Questões	μ	σ	**4 ou 5**
"É divertido utilizar esta aplicação"	4,47	0,64	93%
"Usaria esta aplicação num local público para passar o tempo"	4,4	0,83	80%
"Utilizaria esta aplicação para me divertir com amigos/família"	4,5	0,74	87%
"Utilizaria esta aplicação para anotar as minhas imagens"	3,53	0,99	60%
"Seria mais divertido usar imagens minhas com as minhas próprias anotações"	3,93	1,03	74%

Tabela 8.23: Média e desvio padrão para um conjunto de questões relacionadas com a utilidade do jogo. Também é apresentada a percentagem de respostas mais favoráveis (4 ou 5).

Durante os testes, notou-se que todos os jogadores se empenharam em conseguir uma boa pontuação catalogando o maior número de imagens correctamente. Alguns utilizadores mostraram interesse em descobrir como a pontuação era calculada, como a barra de energia funcionava, como se chegava ao nível mais alto e tentaram examinar todos os mecanismos da aplicação. Alguns participantes memorizaram os conceitos disponíveis para anotação e até memorizaram a ordem pela qual eram mostrados. A maioria dos utilizadores sabia quantos conceitos estavam disponíveis. Esta informação permite-nos concluir que o número de conceitos apresentados não era exagerado.

No início da aplicação alguns utilizadores não sabiam que tinham de movimentar as mãos em zonas específicas para movimentar imagens ou conceitos ou para fazer uma anotação. Contudo, os jogadores rapidamente reconheciam esse requisito. Os participantes que já tinham experimentado o EyeToy [EyeToy05] tiveram maior facilidade em utilizar a aplicação.

Os participantes perceberam que os conceitos disponíveis se moviam à volta num círculo mas alguns não reconheceram o mesmo comportamento nas imagens. Isto pode ser explicado pelo forma como as imagens são representadas e pelo facto das imagens mudarem com a mudança de nível. Este último comportamento da aplicação, que confundiu alguns utilizadores, parece ser adequado para um jogo mas não para a tarefa de anotar imagens.

Sugestões dos Utilizadores
Alguns utilizadores fizeram sugestões que indicam possíveis melhoramentos a fazer:

- Deveria existir um botão para cancelar uma anotação errada;
- Durante o *login* deveria existir um temporizador com contagem decrescente para indicar quando se pode começar a jogar;
- Os conceitos associados com a imagem corrente devem ser apresentados numa posição mais próxima da imagem correspondente;
- Deveriam ser usados sons para dar ênfase a uma boa ou má anotação;
- A pontuação e a informação do nível deviam ser mais realçados.

Em relação ao aspecto estético da interface, a principal observação está relacionada com as cores usadas. Os participantes sugeriram a utilização de cores mais apelativas que realcem a informação mais importante, por exemplo, o nível e os conceitos seleccionados para a imagem corrente. Em geral, os participantes descreveram a aplicação como sendo, útil, divertida, fácil de usar e fácil de aprender.

8.6 Síntese

Este capítulo descreve os testes realizados para avaliar o sistema de recuperação e anotação proposto nesta tese e utilizado nas três aplicações propostas: Memoria Desktop, Memoria Mobile e Aplicação Semi-Automática de Anotação. São também apresentados os resultados dos testes de usabilidade efectuados para validar as referidas aplicações. Para a aplicação Memoria Desktop foi feita uma avaliação do sistema de recuperação e do sistema de anotação utilizando uma colecção pessoal de fotos. A interface foi avaliada com testes de usabilidade que incluíram a validação dos resultados das pesquisas. Os utilizadores mostraram-se satisfeitos com resultados de pesquisa medidos com cerca de 50% de precisão. No caso da aplicação Memoria Mobile, também foi avaliado o sistema de recuperação e o sistema de anotação mas para um conjunto de memórias composto por fotos da Quinta da Regaleira. A aplicação foi avaliada utilizando testes com protótipos em papel e em PDA. O capítulo termina apresentando a Aplicação Semi-Automática para Anotação de imagens. A avaliação do sistema de anotação mostra que o desempenho do sistema melhora com o aumento do número de imagens no conjunto de treino utilizado para estimar os conceitos semânticos. Os testes de usabilidade mostraram que os utilizadores se divertiram a anotar imagens com o jogo Tag Around.

9

Conclusões e Perspectivas Futuras

Conteúdo

Neste capítulo são apresentadas as conclusões e as perspectivas de trabalho futuro. São apresentadas conclusões de âmbito geral e conclusões mais específicas produzidas pelo estudo efectuado nesta tese e as principais direcções para dar continuidade a este trabalho.

9.1 Conclusões

Esta tese enquadra-se na área de investigação em memórias pessoais que foi identificada como um "Grand Challenge for Computing Research" em [Fitzgibbon03, Rowe05] e que segue as ideias pioneiras de Vannevar Bush [Bush45] e Jim Gray [Gray03]. O principal objectivo é a reutilização de informação relativa a experiências do passado em diversos tipos de aplicações. Neste trabalho são propostas soluções no domínio das memórias pessoais, nomeadamente na recuperação e anotação de fotos em diferentes contextos de aplicação. A recuperação e anotação são uma parte do problema da reutilização de memórias pessoais.

Nesta secção, descrevem-se as conclusões relativas ao trabalho realizado. São apresentadas conclusões gerais, identificadas a partir de uma avaliação global do trabalho desenvolvido, e conclusões mais pormenorizadas em relação ao modelo semântico proposto para recuperação e anotação de imagens e às aplicações de memórias pessoais.

Para finalizar o capítulo e esta dissertação, na secção seguinte são apresentadas algumas direcções para dar continuidade ao trabalho desenvolvido. Primeiro, são descritas direcções possíveis para a recuperação e anotação de imagens e para novas aplicações e depois, são também apresentados cenários de evolução deste trabalho numa perspectiva global.

Os métodos desenvolvidos para anotação e recuperação de imagens foram experimentados em várias áreas de aplicação. Foram desenvolvidas interfaces para computadores pessoais e dispositivos móveis com diversas técnicas de interacção. O sistema de recuperação e anotação apresentado baseia-se em modelos semânticos estimados utilizando informação multimodal, isto é, características visuais, informação obtida a partir de áudio, informação temporal e dados de localização obtidos através de coordenadas espaciais. Estes métodos foram utilizados para explorar memórias pessoais em três aplicações em áreas distintas:

- Memoria Desktop - aplicação para recuperar experiências do passado em ambientes domésticos utilizando computadores pessoais;
- Memoria Mobile - aplicação para partilha de fotos (experiências) no momento da visita a locais de interesse utilizando dispositivos móveis;
- Aplicação Semi-Automática de Anotação - aplicação para anotação semântica de imagens de forma divertida através do jogo Tag Around com interface baseada em gestos.

Para avaliar o método de análise semântica em imagens foram utilizadas duas colecções de fotos: (1) colecção pessoal do autor e (2) colecção de fotos da Quinta da Regaleira. A análise semântica de imagens foi utilizada nas aplicações Memoria Desktop e Memoria Mobile para recuperar imagens através de pesquisas e foi utilizada numa aplicação semi-automática para anotar imagens através de um jogo. Os métodos de recuperação e anotação foram avaliados nas três aplicações e estas aplicações foram avaliadas na globalidade com testes de usabilidade.

A seguir, são apresentadas conclusões gerais do estudo efectuado e depois, nas secções seguintes, são apresentadas conclusões específicas dos testes realizados aos métodos de recuperação e anotação propostos e conclusões obtidas com os testes de usabilidade efectuados às aplicações.

As principais conclusões obtidas a partir dos testes efectuados para avaliar as propostas desta tese são as seguintes:

- A combinação de informação visual com informação de áudio e informação contextual (localização, data e hora) melhora o desempenho da classificação semântica de imagens;

- Os testes com utilizadores mostram que a técnica proposta, *drag & drop* de elementos para uma "Query Box", para recuperar experiências do passado satisfazem as necessidades dos utilizadores. Isto é, permite utilizar e combinar as várias pistas que a memória episódica [Endel02] utiliza para relembrar eventos do passado;

- Uma das medidas mais utilizada para avaliar o desempenho dos sistemas de recuperação é o MAP [Over06]. Os resultados apresentados nesta tese mostram que os utilizadores ficaram satisfeitos com resultados de pesquisas com MAP a rondar o valor de 0,50;

- No domínio das memórias pessoais, os resultados mostram que os utilizadores preferem ser eles a escolher o conjunto de conceitos a utilizar para recuperação e anotação de imagens nas suas colecções pessoais.

- Os resultados apresentados mostram que a anotação semi-automática de imagens através de um jogo de computador é uma opção para combater a falta de motivação dos utilizadores para a anotação manual;

9.1.1 Recuperação e Anotação de Informação Multimédia

O método proposto para recuperação de imagens foi avaliado em duas aplicações, com duas colecções de fotos diferentes. A aplicação Memoria Mobile permite definir pesquisas com imagens exemplo. A avaliação destas pesquisas permitiu comparar as várias características visuais e as suas combinações. Quando analisados individualmente o vector de ocorrências com descritores SIFT (precisão = 0,54) superou o vector de ocorrências de características obtidas com o banco de filtros de Gabor (precisão = 0,48). Nos testes efectuados com as características de cor combinadas com textura, a combinação do vector de ocorrências de regiões de cor com os vector de ocorrências de descritores SIFT obteve o melhor desempenho (precisão = 0,63). Este desempenho melhorou ainda mais quando se aplicou o método LSA (precisão = 0,74).

Na aplicação Memoria Desktop também se comparam as características visuais mas através dos modelos semânticos. Mais uma vez, a combinação vector de ocorrências de regiões de cor com vector de ocorrências de descritores SIFT obteve o melhor desempenho (MAP aplicando o método LSA = 0,51). Esta combinação foi superada quando se incluiu a informação temporal (MAP = 0,53). Conclui-se assim que explorando a correlação temporal conseguimos melhorar o desempenho do sistema.

Na aplicação Memoria Mobile, comparou-se o desempenho dos conceitos semânticos utilizando só informação visual, incluindo o áudio e utilizando também informação espacial. A combinação de informação visual com informação de áudio (MAP = 0,56) apresenta melhor desempenho do que utilizando só informação visual (MAP = 0,48). Contudo, a combinação de informação visual com informação geo-referenciada nem sempre melhorou os resultados porque o desempenho depende das características da região seleccionada.

Em relação ao método proposto para anotação de imagens, obteve-se uma cobertura por imagem de 0,55 com a colecção pessoal usada no Memoria Desktop e uma cobertura de 0,46 para a base de dados da Quinta da Regaleira.

O método semi-automático de anotação permite melhorar o desempenho do método automático (cobertura = 0,55). Incluindo 20 imagens no conjunto de treino temos cobertura = 0,70 e incluindo 40 imagens a cobertura obtida foi 0,73.

9.1.2 Aplicações

Nesta tese foram propostas três aplicações: Memoria Desktop, Memoria Mobile e Aplicação Semi-Automática de Anotação. A aplicação Memoria Desktop foi avaliada com 58 utilizadores, a aplicação Memoria Mobile com 4 utilizadores e o jogo Tag Around com 15 utilizadores.

A maioria dos utilizadores achou útil a informação que é possível observar numa colecção pessoal utilizando a aplicação Memoria Desktop (Moda = 4, sendo a resposta 5 a mais favorável) mas não afirmaram categoricamente que utilizariam a aplicação para gerir a sua colecção pessoal (Moda = 3, sendo a resposta 5 a mais favorável).

Os utilizadores fizeram 4 pesquisas e a maioria ficou satisfeita com os resultados (Moda = 4 para 3 das pesquisas, a resposta 1 significa "mau" e 5 "excelente"). Estes resultados mostram que o desempenho dos modelos semânticos na recuperação de informação é positivo (MAP a rondar o valor de 0,50).

As respostas dos utilizadores (Moda = 4, sendo 5 a mais favorável) mostram que a opção de usar a técnica de *drag & drop* de elementos para uma "Query Box" é adequada para fazer pesquisas.

Em relação à Aplicação Semi-Automática para Anotação, em particular ao jogo Tag Around, a maioria dos participantes achou o jogo divertido (93% das respostas foram 4 ou 5, sendo 5 a resposta mais positiva) e consideraram útil a aplicação para ser utilizada em casa com amigos ou família (87% das respostas foram 4 ou 5) ou para ser utilizada num local público onde tenham de esperar (80% das respostas foram 4 ou 5).

A técnica de interacção proposta para jogar o jogo Tag Around também teve aprovação da parte dos utilizadores. A maioria afirmou não ter dificuldade em interagir com as zonas específicas (Moda = 4, sendo 5 a mais favorável) e a maioria também afirmou preferir a técnica de interacção proposta em relação ao teclado ou rato.

9.1.3 Resultados

O trabalho desenvolvido nesta tese foi publicado em treze conferências, dez internacionais e três nacionais (ver secção 1.5). Estas publicações denotam o interesse da comunidade científica no trabalho desenvolvido. Estão também disponíveis três protótipos para demonstração de cada uma das aplicações propostas.

Foram realizadas demonstrações do protótipo da aplicação Memoria Mobile no Memories for life Colloquium que decorreu em Londres em 2006 e na sessão de demonstrações do ACM SIGIR Conference on Research and Development in Information Retrieval que decorreu em Amesterdão em 2007.

Em relação ao jogo Tag Around, foi apresentada uma demonstração na Conferência Nacional em Interacção Pessoa-Máquina, que decorreu em Évora em 2008, e o Memoria Desktop foi demonstrado no Media Research Workshop, que decorreu em Lisboa em 2007, no âmbito do programa UT Austin-Portugal.

Está também disponível uma biblioteca com a implementação dos algoritmos de recuperação de informação. Esta biblioteca é utilizada por todas as aplicações incluindo a aplicação Memoria Web. Esta aplicação foi desenvolvida no âmbito do Projecto InStory 2 financiado pelo POSC, em colaboração com a fundação Cultursintra/Quinta da Regaleira, Sintra, Portugal.

9.2 Perspectivas Futuras

No âmbito desta tese foi desenvolvido trabalho científico no domínio das Memórias Pessoais, área apontada em [Fitzgibbon03, Rowe05] como um "Grand Challenge for Computing Research". Este trabalho será continuado neste domínio científico mas utilizando vídeo, isto é, no desenvolvimento de novas aplicações onde é necessário o acesso a colecções pessoais de vídeo, principalmente em aplicações de telepresença que estão incluídas noutro grande desafio referido em [Rowe05, Gray03].

A seguir, é apresentado trabalho futuro focado no trabalho científico descrito nesta tese para ser realizado a curto prazo e depois são apresentadas direcções para continuar este trabalho a longo prazo.

O método de recuperação e anotação de imagens proposto nesta tese será continuado nas seguintes direcções:

- Testar e avaliar o método de anotação com um número maior de conceitos e introduzir ontologias para obter mais palavras para anotação;
- Experimentar os métodos propostos com colecções de fotos maiores para avaliar a escalabilidade do algoritmo proposto;
- Incluir mais informação referente aos metadados obtidos no instante de captura no treino dos modelos semânticos, por exemplo, o nível do flash, a distância ao objecto ou informação capturada por sensores que sejam incluídos nos dispositivos de captura à semelhança da SenseCam [Gemmell04];
- Construir um filtro 2D com a informação temporal e espacial de modo a explorar conjuntamente a correlação temporal e espacial para corrigir os modelos semânticos treinados com informação visual;
- Desenvolver características visuais específicas para os conceitos em que o método proposto obteve um desempenho baixo, por exemplo para "Party" e para outros conceitos não experimentados nesta tese mas relevantes para as memórias pessoais;
- Estimar os conceitos semânticos, explorando as dependências entre os conceitos. Por exemplo, o modelo probabilístico do conceito "Beach" deve ter em conta o modelo de "Outdoor".

O trabalho descrito nesta tese relacionado com as aplicações será continuado nas seguintes direcções:

- Melhorar as interfaces propostas introduzindo as alterações sugeridas pelos utilizadores nos testes de usabilidade, nomeadamente as sugestões relacionadas com a técnica proposta para fazer a pesquisa com diferentes tipos de informação;

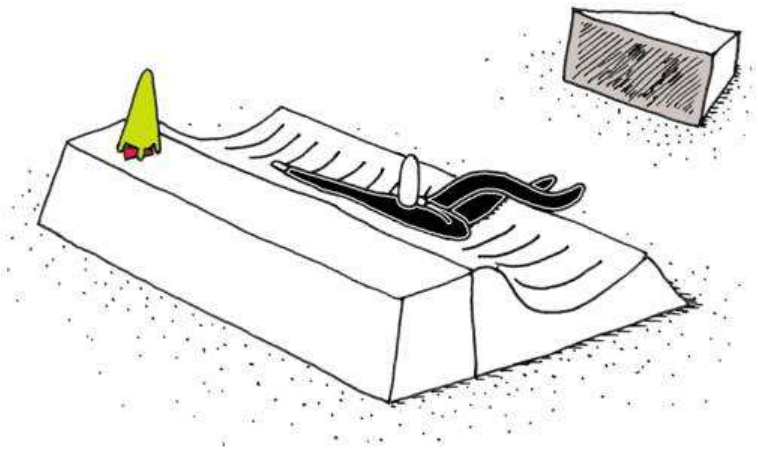


Figura 9.1: Interface tangível baseada num objecto de decoração (verde) para recuperação de memórias através de objectos (vermelho)

- Avaliar a aplicação Memoria Mobile de forma mais sistemática, com mais utilizadores quando estiverem reunidas as condições logísticas;
- Adaptar o jogo Tag Around a outros contextos, por exemplo, museus (já a decorrer) ou para ser utilizado por crianças em escolas;
- Desenvolver aplicações que permitam fazer pesquisa colaborativa, isto é, permitir que duas ou mais pessoas possam pesquisar informação diferente referente ao mesmo tópico, com diferentes formas de interacção (por exemplo, teclado, rato ou telemóvel), e partilhar os resultados no mesmo ecrã;
- Desenvolver aplicações para serem utilizadas em casa, facilitando a interacção e o convívio entre as pessoas no momento do reviver da experiência e considerando o aspecto social. A figura 9.1 apresenta uma ideia para uma interface baseada num objecto de decoração (a verde) para relembrar experiências com base em objectos colocados por baixo. O objecto de decoração inclui uma câmara de vídeo que captura uma imagem do objecto e a envia para um computador. Depois de processada a imagem capturada, que é utilizada para interrogar a base de dados, são apresentados na televisão fotos ou vídeos referentes à experiência relacionada com o objecto.

Em relação ao trabalho a realizar a longo prazo, o objectivo é desenvolver novas versões do Memoria Mobile, Memoria Web e da Aplicação Semi-Automática para Anotação no contexto da telepresença. Por exemplo, uma aplicação em que o utilizador possa fazer a visita a um local de interesse virtualmente, de forma idêntica ao Memoria Web, mas que consiga interagir com objectos locais remotamente e realizar algumas funcionalidades do Memoria Mobile que exijam a presença do utilizador no local. Em ambas as situações serão utilizadas as câmaras de vigilância do local (por exemplo, a captura de fotos). Outra alternativa, é uma aplicação para aumentar a informação disponível ao utilizador durante a visita através de uma aplicação de entretenimento para anotação semi-automática de imagens. Por exemplo, no momento da visita ao local o utilizador poderá interagir com objectos locais, virtuais e com outros visitantes através de gestos para as câmaras do local e desta forma associar imagens a palavras melhorando o seu conhecimento sobre o local.

Um versão do segundo exemplo referido atrás, está ser desenvolvida no âmbito do projecto "Comunicação Pública da Arte" financiado pela Fundação para a Ciência e a Tecnologia numa colaboração entre a FCSH/UNL e o CITI/Departamento de Informática da FCT/UNL. Esta aplicação é uma versão do jogo Tag Around mas para ser jogada por vários visitantes de um museu utilizando diversas formas de interacção.

No contexto da recuperação de informação multimédia, mas fora do domínio das Memórias Pessoais e da telepresença, está a decorrer o VideoFlow, um projecto em co-promoção de I&DT financiado pelo Quadro de Referência Estratégico Nacional (QREN). Neste projecto serão adaptados para vídeo os modelos semânticos propostos nesta tese. O VideoFlow é um projecto a realizar em parceria entre a Duvideo, uma empresa privada no domínio audiovisual, e o CITI/Departamento de Informática da FCT/UNL, com o objectivo de desenvolver um sistema suportando interfaces avançadas para acesso a arquivos de vídeo com base em metadados, extraídos com uma combinação de processos automáticos e conhecimento humano.

Está a decorrer o projecto CRUSH (Clip-art Retrieval using Sketches) financiado pela Fundação para a Ciência e a Tecnologia que visa desenvolver uma nova abordagem para recuperar clip-arts, independentemente do seu formato (*raster* e vectorial). A solução será baseada na recuperação de clip-arts utilizando esboços como interrogações. Serão adaptadas as técnicas de recuperação de imagem, desenvolvidas no âmbito desta tese, e de análise de desenhos vectoriais, para descrever o conteúdo de clip-arts. O projecto envolve a colaboração entre o INESC ID/IST/UTL e o CITI/Departamento de Informática da FCT/UNL.

Em relação à anotação de informação multimédia, está a decorrer o projecto TKB - Base de conhecimento Transmédia para Dança Contemporânea, financiado pela Fundação para a Ciência e a Tecnologia, em colaboração com várias instituições de investigação (FCSH/UNL e o Departamento de Informática da FCT/UNL são duas das instituições participantes). O projecto visa construir uma aplicação de anotação semi-automática de vídeos de dança contemporânea baseada na análise semântica de imagens proposta nesta dissertação. O método de anotação automática de imagens será adaptado para vídeo e para conceitos específicos do tema.

Finalmente, está a decorrer o projecto ARIA - Ambientes de Leitura Assistida para Idosos, financiado pela Fundação para a Ciência e a Tecnologia. O projecto envolve a colaboração entre a Fundação da Faculdade de Ciências da Universidade de Lisboa, o CITI/Departamento de Informática da FCT/UNL e o INESC ID/IST/UTL. Os métodos utilizados nesta tese para detecção de conceitos em imagens serão adaptados para detectar emoções em vídeos de faces.

Memoria - Testes de Usabilidade

A.1 Questionário

Aplicação Memoria Desktop

A aplicação Memoria Desktop destina-se à captura, anotação automática, pesquisa e visualização de memórias pessoais compostas por fotografias e vídeos digitais. Nesta fase serão avaliadas apenas duas destas funcionalidades: pesquisa e visualização de imagens. A aplicação permite navegar no sistema de directorias, visualizar fotos em modo lista ou em modo slideshow e fazer pesquisas de dois modos distintos: pesquisa por conceitos e pesquisa usando uma imagem composta por partes de outras imagens.

Através deste estudo pretendemos avaliar a usabilidade do sistema. Para o efeito, pedimos-lhe que execute as tarefas abaixo indicadas e, preenchendo este questionário, nos informe acerca das dificuldades que encontrou ao interagir com o sistema e nos dê a sua opinião acerca da interface e das funcionalidades disponíveis.

Dados pessoais

Idade:_____

Sexo:_____ (M/F)

1.1 Possui máquina digital para tirar fotografias? _____

1.2 Quantas vezes usa a máquina digital por mês?_____ Por ano? _____

1.3 Quantas fotos tem a sua colecção pessoal?_____

1.4 Costuma anotar com palavras as suas fotos?

1.5 Quando pretende pesquisar as suas fotos pessoais em formato digital o que costuma fazer?

1.6 Qual o objectivo das pesquisas que costuma fazer na sua colecção pessoal (por exemplo, fotos de uma pessoa ou de umas férias)?

1.7 Qual a pista (por exemplo, data, local ou as pessoas que estavam presentes) que mais usa para procurar fotos de um acontecimento do passado?

Tarefas

Por favor, efectue as seguintes tarefas e descreva-nos os resultados obtidos.

Exploração da interface

Tarefa 1: Navegue na colecção de fotos usando a árvore de directorias para analisar com mais detalhe algumas imagens ("Preview").

Que dificuldades encontrou para cumprir a tarefa?

Tarefa 2: Visualize um conjunto de imagens em modo "Slideshow" (normal e *fullscreen*).

Que dificuldades encontrou para cumprir a tarefa?

Pesquisa de imagens e visualização de resultados

Tarefa 3: Pesquisar imagens através do *drag & drop* de conceitos e operadores lógicos para a

"Query Box". Por favor, execute as seguintes pesquisas (utilizando os conceitos e operadores lógicos disponíveis e construindo as interrogações adequadas) e responda às questões expostas fazendo um círculo em volta do número que melhor representa a sua opinião acerca da aplicação que acaba de experimentar.

3.1 Procurar por fotos com pessoas.

Como classifica os resultados obtidos pela interrogação?

1	2	3	4	5
Maus				Excelentes

3.2 Procurar por fotos com paisagens naturais (natureza).

Como classifica os resultados obtidos pela interrogação?

1	2	3	4	5
Maus				Excelentes

3.3 Pesquisar por fotos tiradas no exterior ("Outdoor"), mas que não tenham sido tiradas em praias.

Como classifica os resultados obtidos pela interrogação?

1	2	3	4	5
Maus				Excelentes

3.4 Pesquisar por fotos com pessoas ou com paisagens naturais.

Como classifica os resultados obtidos pela interrogação?

1	2	3	4	5
Maus				Excelentes

Questões gerais relacionadas com as tarefas anteriores:

1. Quais foram as dificuldades sentidas na elaboração das interrogações?

_____

_____

_____

_____

2. Na sua opinião, o *drag & drop* é adequado para a tarefa de definição de interrogações?

1	2	3	4	5
Inadequado				Muito adequado

3. É perceptível o significado de cada um dos ícones disponíveis para definir as interrogações?

_____

Qual o ícone mais perceptível? _____

Quais os ícones que não considera perceptíveis? _____

4. Na sua opinião, é perceptível a forma como deve combinar os ícones para obter um determinado resultado?

1 2 3 4 5
Inadequado Muito adequado

5. Quais os conceitos que acha mais úteis para pesquisar imagens em colecções pessoais?

_____

_____

6. Que outros conceitos deveriam estar disponíveis para a construção das interrogações?

_____

_____

_____

_____

Tarefa 4: Pesquisar imagens através da construção de um esboço composto por partes de outras imagens da base de dados. Por favor, execute as seguintes pesquisas (escolha uma imagem, seleccione uma parte e clique na "Query Box") e responda às questões expostas fazendo um círculo em volta do número que melhor representa a sua opinião acerca da aplicação que acaba de experimentar.

4.1 Procurar por fotos de edifícios com diferentes arquitecturas; Usar partes rectangulares de imagens para construir o esboço.

Como classifica os resultados obtidos pela interrogação?

1 2 3 4 5
Maus Excelentes

4.2 Procurar por fotos com a face de uma determinada pessoa (cortar a face da pessoa em várias imagens e compor um imagem com essas faces). Usar o modo "Freehand" para cortar as faces.

Como classifica os resultados obtidos pela interrogação?

1 2 3 4 5
Maus Excelentes

Questões gerais relacionadas com as tarefas anteriores:
1 Considera útil a possibilidade de executar este tipo de pesquisas?

_____

_____

Avaliação da interface

Responda às seguintes perguntas fazendo um círculo em volta do número que melhor representa a sua opinião acerca da aplicação que acaba de experimentar.

1. Considero a informação fornecida pelo sistema útil.

1	2	3	4	5
Discordo totalmente				Concordo totalmente

2. É fácil aprender a usar a aplicação.

1	2	3	4	5
Discordo totalmente				Concordo totalmente

3. O aspecto estético da interface agrada-me.

1	2	3	4	5
Discordo totalmente				Concordo totalmente

4. Eu utilizaria esta aplicação para gerir as minhas fotos pessoais.

1	2	3	4	5
Discordo totalmente				Concordo totalmente

A.2 Resultados

Utilizadores	Idade	Género	1.1	1.2		1.3
1	23	M	S	3	36	15000
3	24	M	S	1	10	3000
5	25	M	S	2	50	200
7	25	M	S	15	-	-
9	24	M	S	2	-	1000
11	22	M	N	0	2	100
13	23	M	S	4	-	500
15	23	M	S	-	5	200
17	23	M	S	2	24	7500
19	23	M	S	-	5	-
21	27	M	S	5	60	2000
23	23	M	S	5	70	4000
25	30	M	S	3	-	10000
27	24	M	S	1	12	-
29	23	M	S	1	10	2000
31	22	M	S	2	24	2060
33	23	M	N	1	5	500
35	24	F	N	-	-	200
37	23	M	S	10	120	2000
39	21	F	S	1	12	-
Média Total	23,53					2591,76

Tabela A.1: Exemplos de respostas dos utilizadores às primeiras perguntas dos dados pessoais.

Utilizador	1.5
1	Procurar na pasta
3	Guardo as fotos em directorias com nomes explicativos depois procuro pela respectiva directoria
5	Faço browsing por data
7	Uso software adequado, ACDSee, HP software
9	Navegar nas pastas
11	Procuro na pasta respectiva
13	Estão organizadas por data, procuro a pasta por data
15	Pesquiso por data
17	Procuro na pasta com a indicação da data e local da fotos que pretendo encontrar
19	Normalmente estão organizadas por data e possuem a descrição do evento
21	Pesquiso por pasta
23	Uso um programa que organiza as fotos por ordem cronológica facilitando as pesquisas
25	Faço separação ao nível do file system
27	Pesquisa na pasta pretendida
29	Ir à pasta respectiva
31	Vou à pasta. Tenho-as organizadas por ocasião
33	Pesquiso por ano e depois por evento
35	Ir à pasta respectiva
37	Procuro o directório
39	Estão organizadas em pastas

Tabela A.2: Exemplos de respostas dos utilizadores à pergunta 1.5 dos dados pessoais.

Utilizador	1.6	1.7
1	Partilha ou impressão	Data e assunto
3	Pessoas, férias, lugares e ocasiões e data importantes	Local e data
5	Férias ou acontecimento importante	Data
7	Encontrar uma determinada foto, com algo particular	Local, data e acontecimento
9	Férias	Local e data
11	Pessoas	Data
13	Apenas visualizar	Data e acontecimento
15	Para ver se as fotos ficaram bem	Data
17	Férias	Data e local
19	Evento	Nome do evento
21	Férias	Pessoas e local
23	datas e locais	Data e locais
25	Pessoas, eventos e datas	Tipos de evento: festa, competição, passeio
27	Uma imagem em concreto de pessoas ou paisagens	data
29	Relembrar momentos passados	Local e data
31	Ocasião especial	Data e acontecimento
33	Relembrar momentos passados	Data e acontecimento
35	Aniversário e data	Local e data
37	Para enviar a alguém	Local
39	Férias ou acontecimento especifico	Data ou local

Tabela A.3: Exemplos de respostas dos utilizadores às restantes perguntas dos dados pessoais.

Utilizador	Tarefa 1
1	Nenhuma. O icone + nas pastas que não têm subdirectorias confunde um pouco.
3	Nenhuma
5	Nenhuma
7	Nenhuma
9	Poucas dificuldades
11	Nenhuma
13	Único problema foi a resolução do ecrã
15	Nenhuma
17	Nenhuma
19	Lento
21	Nenhuma
23	Não tive dificuldade
25	Nenhuma. O icone + nas pastas que não têm subdirectorias confunde um pouco.
27	Nenhuma
29	Nenhuma
31	Nenhuma
33	Nenhuma
35	Sem dificuldade
37	Nenhuma
39	Nenhuma

Tabela A.4: Exemplos de respostas dos utilizadores à pergunta referente à tarefa 1.

Utilizador	Tarefa 2
1	Botões fornecidos não são muito claros. A forma de sair do modo slideshow não é intuitiva.
3	Dificuldade em encontrar o botão slideshow no menu. Mais fácil se os botões tivessem labels.
5	Dificuldade em iniciar o slideshow.
7	Dificuldade em associar os icons ao evento slideshow
9	Nenhuma dificuldade. Bastante Intuitivo
11	Dificuldade em encontrar os botões. Dificuldade em terminar a apresentação
13	Nenhuma dificuldade
15	Nenhuma
17	Não
19	Nenhuma
21	Nenhuma
23	tive problemas quando usei o botão de slideshow
25	Dificuldades em identificar os botões com as funcionalidades respectivas
27	Dificuldade em encontrar os botões. A existência de nomes nos botões facilitaria a tarefa
29	Dificuldade em encontrar os botões. Não consegui para o slideshow
31	Algumas dificuldades iniciais no botões mas não nos menus
33	Nenhuma
35	Sem dificuldade
37	Dificuldade em identificar os botões para experimentação
39	Deveria aparecer uma legenda sobre os botões

Tabela A.5: Exemplos de respostas dos utilizadores à pergunta referente à tarefa 2.

Utilizador	Tarefa 3 - Questão 1
1	Poucos Filtros. Em caso de engano ter de refazer a query. Não poder inserir um filtro no meio de dois
3	No drag & drop
5	Poucos conceitos. Não dá para trocar conceitos dentro da query
7	Algumas queries necessitam de demasiados operadores
9	Falta Icone explicito para Natureza
11	Impossivel ordenar elementos ou inserir um elemento novo
13	Dificuldade em perceber qual o icon de natureza
15	Nenhuma
17	Não é claro como se utilizam os operadores lógicos
19	Dificuldade em entender as prioridades dos operadores
21	Nenhuma
23	Nenhuma
25	A utilização do operador Not é confusa
27	A ordem na query box e o operador No
29	Nenhuma
31	Pequenas dificuldades com o operador NO
33	Nenhuma
35	Não é possível colocar um icone no meio da query
37	Poucos conceitos
39	Conceito de Natureza não é directo

Tabela A.6: Exemplos de respostas dos utilizadores à pergunta "Quais foram as dificuldades sentidas na elaboração das interrogações?" da tarefa 3.

Utilizador	Questão 3 da tarefa 3		
1	Sim	Os icons dos operadores	Manmade
3	Não	People	Beach
5	Sim	People	Beach, Manmade
7	Sim	Snow	Beach
9	Sim	People	Beach, Manmade, Indoor
11	Sim	Snow	Beach, Manmade
13	Sim	Snow	Beach
15	Sim	Snow	Nenhum
17	Sim	Manmade, beach	party, snow, people, face, indoor
19	-	Os icons dos operadores	Beach, Manmade
21	Sim	People	Manmade
23	Sim	Snow	Beach
25	Sim	todos	Nenhum
27	Sim	Snow	Manmade
29	Sim	Face	Beach, Manmade
31	Sim	Face	Beach e os operadores
33	Sim	São todos perceptiveis	Nenhum
35	Sim	People	Beach, Manmade
37	Sim	Operadores lógicos	Beach e Manmade
39	Sim	Pessoas	Beach

Tabela A.7: Exemplos de respostas dos utilizadores à pergunta 3 da tarefa 3.

Utilizador	Questão 5 da tarefa 3
1	Indoor, People, Party
3	Beach, Snow, Face
5	Eventos Especificos e Datas
7	Paisagens e Desporto
9	People, Party, Nature
11	Data de captura
13	Beach, Snow, People
15	People, Nature, Party
17	-
19	-
21	Nome
23	People, party
25	People, Familia, férias, party
27	Data, local, Acontecimento, Parte do dia, pessoa
29	Party, People
31	tema e a data das fotos pretendidas
33	Eventos, Viagens, Épocas, Estações do ano
35	face, indoor, People, party
37	Pessoas e locais
39	Praia, pessoas, festa, interior, exterior

Tabela A.8: Exemplos de respostas dos utilizadores à pergunta 5 da tarefa 3.

Utilizador	Questão 6 da tarefa 3
1	Não concordo com a negação de conceitos; Preferia outdoor em vez de No + Indoor
3	Dia, Noite, Paisagens
5	-
7	Paisagens e Desporto
9	Natureza e Outdoor
11	Noite e dia, Data de captura, Data de cópia para disco
13	Campo, Paisagens, Edifícios
15	Amigos e Colegas
17	Reconhecimento de pessoas
19	Férias e eventos (Carnaval, Pascoa)
21	Outoor, natureza
23	Noite e dia, familiares e amigos
25	Outdoor e Nature
27	Familia, amigos
29	-
31	Data
33	Não me ocorre mais nenhum
35	Data
37	Locais especificos, estações do ano, idade das pessoas
39	dia e noite

Tabela A.9: Exemplos de respostas dos utilizadores à pergunta 6 da tarefa 3.

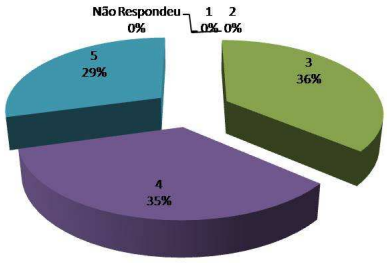


Figura A.1: Tarefa 3 - 3.1 Pesquisar por fotos com pessoas.

Utilizador	Questão 1 e da Tarefa 4
1	Não acho utilidade para uso pessoal. Pode ser util para resolver questões de segurança (faces)
3	Sim
5	Sim. No caso de não ter uma colecção bem organizada ou de querer algo especifico
7	Sim
9	Sim, embora ache que os operadores lógicos podem causar dificuldades às pessoas menos instruídas
11	Sim. Mas não obtive qualquer resultado.
13	Util para pesquisar objectos ou pessoas
15	Sim
17	Sim
19	-
21	Sim
23	-
25	Parece dar bastante trabalho para dar os resultados pretendidos
27	Nao. Para ser viável esta query teria que ter mais opções ao nível da adição de imagens
29	Sim. Dificuldade em encontrar o botão freehand
31	Em caso especificos sim
33	Sim, util para procurar pessoas
35	Sim, considero util este tipo de pesquisas
37	Sim
39	Sim, principalmente das pessoas

Tabela A.10: Exemplos de respostas dos utilizadores à pergunta 1 da tarefa 4.

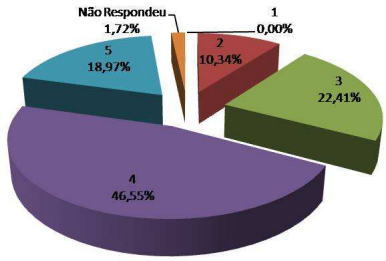


Figura A.2: Tarefa 3 - 3.2 Pesquisar por fotos com paisagens naturais (natureza).

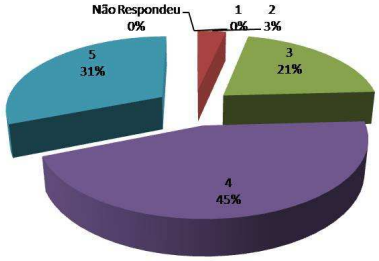


Figura A.3: Tarefa 3 - 3.3 Pesquisar por fotos tiradas no exterior (outdoor), mas que não tenham sido tiradas em praias.

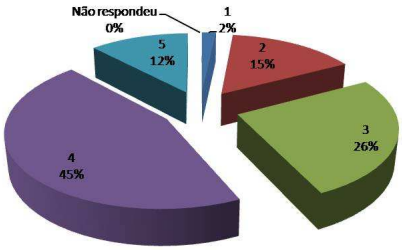


Figura A.4: Tarefa 3 - 3.4 Pesquisar por fotos com pessoas ou com paisagens naturais.

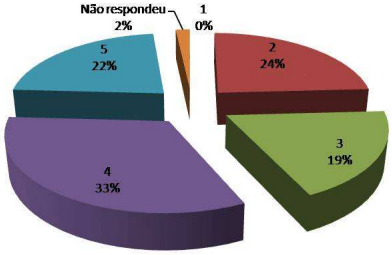


Figura A.5: Tarefa 3 - Questão 2. Na sua opinião, o drag & drop é adequado para a tarefa de definição de interrogações?

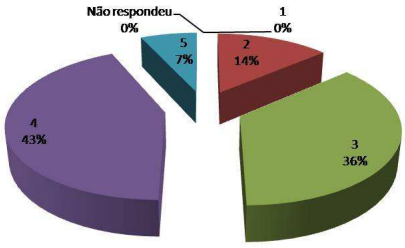


Figura A.6: Tarefa 3 - Questão 4. Na sua opinião, é perceptível a forma como deve combinar os ícones para obter um determinado resultado?

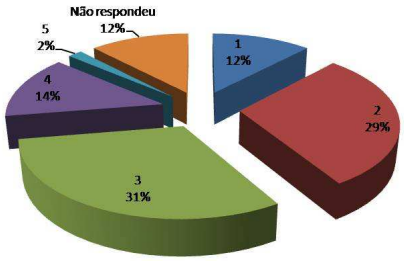


Figura A.7: Tarefa 4 - Questão 4.1 Procurar por fotos de edificios com diferentes arquitecturas; Usar partes rectangulares de imagens para construir o esboço.

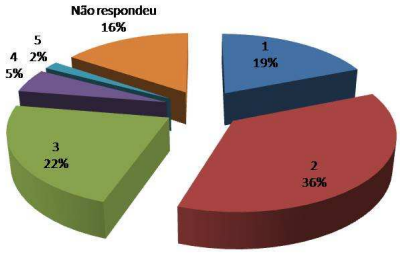


Figura A.8: Tarefa 4 - 4.2 Procurar por fotos com a face de uma determinada pessoa (cortar a face da pessoa em várias imagens e compor um imagem com essas faces). Usar o Freehand mode para cortar as faces.

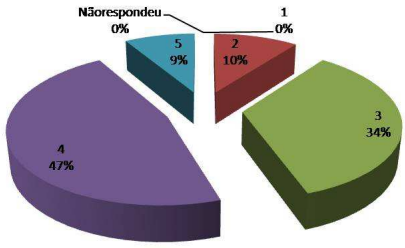


Figura A.9: Avaliação da interface - 1. Considero a informação fornecida pelo sistema útil.

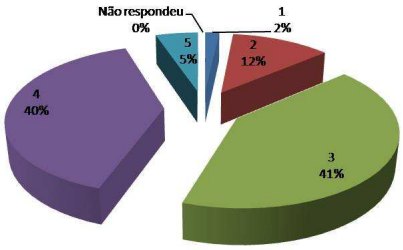


Figura A.10: Avaliação da interface - 2. É fácil aprender a usar a aplicação.

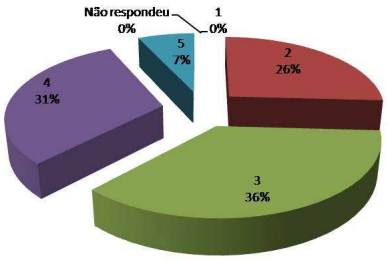


Figura A.11: Avaliação da interface - 3. O aspecto estético da interface agrada-me.

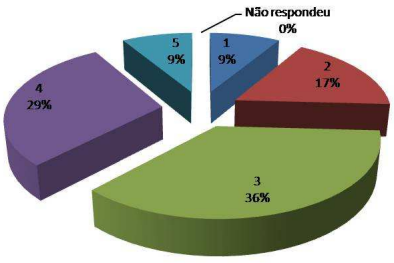


Figura A.12: Avaliação da interface - 4. Eu utilizaria esta aplicação para gerir as minhas fotos pessoais.

B

Tag Around - Testes de Usabilidade

B.1 Questionário

TAG AROUND

Aplicação 3D para a anotação de imagens

Conteúdos multimédia são trocados a todo o momento na Internet a um ritmo nunca visto. Vídeos e imagens enchem os nossos computadores, *blogs* e comunidades online espalhadas pela rede. É necessário organizar todo este conteúdo para uma melhor pesquisa e utilização do mesmo.

Tag Around é um projecto que propõe analisar a questão lúdica e os aspectos que motivam a anotação manual de imagens, propondo uma solução em que os utilizadores se divertem enquanto anotam as suas imagens. Durante esta sessão, pretendemos compreender a iteração dos utilizadores com a interface, em termos da sua complexidade, facilidade de aprendizagem, divertimento, compreensão dos objectivos propostos, e aspecto audiovisual da interface. Para isso propomos que experimente a aplicação, complete os objectivos propostos, e que acima de tudo, se divirta enquanto explora as suas potencialidades.

Aplicação

Esta aplicação consiste num jogo 3D cujo objectivo é anotar correctamente o máximo número de imagens no menor espaço de tempo possível. Como acreditamos que teclados e ratos são aborrecidos, vamos tentar interagir tanto com as imagens como com as anotações usando apenas gestos. As imagens seguintes descrevem as várias etapas da aplicação, para se familiarizarem com a mesma.

Antes de começar a jogar este jogo, o sistema precisa de identificar o jogador no sistema. Para isso, terá de colocar a sua cara dentro do quadrado encarnado, enquanto o sistema o tenta

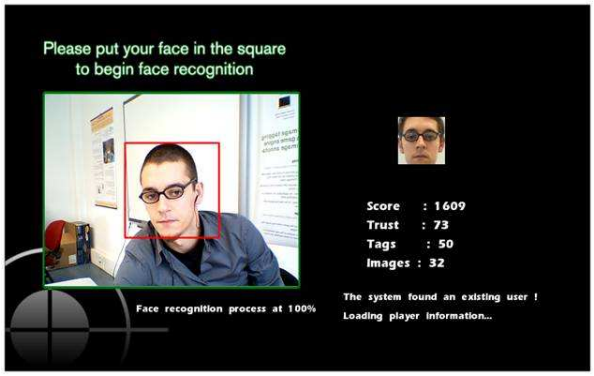


Figura B.1: Interface para fazer *login*.

Figura B.2: Interface utilizada durante a fase de jogo.

identificar. No caso de ser um jogador novo, o sistema irá criar um novo perfil (ver figura B.1). Logo após o *login*, o jogador irá começar a jogar. A figura B.2 mostra a interface utilizada na fase de jogo. A seguir são apresentados alguns elementos da interface.

Pontuação: a pontuação reflecte-se na tua perícia de anotar as imagens, associando os conceitos (em cima na imagem, às imagens em baixo).

Energia: a energia vai aumentando com boas anotações, e diminuindo com o tempo e com as más anotações.

Imagem do utilizador: O jogador irá ver-se na imagem, e terá 5 círculos vermelhos onde pode tocar. Cada um deles tem um objectivo distinto. Os círculos inferiores servem para rodar as imagens para a direita e para a esquerda, enquanto que os círculos de cima servem para rodar as anotações para a direita e para a esquerda. O círculo em cima do utilizador serve para anotar a palavra que está ao centro na imagem que também se encontra no centro.

Dados pessoais

Idade:_____

Sexo:_____ (M/F)

Ligado(a) às tecnologias de informação?_____ (S/N)

Geral - Assinale o número que melhor corresponde a sua resposta

1. Costuma utilizar a internet para fazer pesquisas de imagens?

1	2	3	4	5
Raramente				Muitas vezes

2. Costuma organizar imagens pessoais no seu computador ?

1	2	3	4	5
Raramente				Muitas vezes

3. As suas imagens pessoais/pesquisadas estão catalogadas ?

1	2	3	4	5
Nenhumas				Todas

4. De que modo utiliza as imagens guardadas no seu computador ?

Para pesquisa/trabalho _____ (S/N)

Para recordar com amigos _____ (S/N)

Para colocar em *blogs* _____ (S/N)

Outro(s):_____

5. Quando pretende pesquisar as suas fotos pessoais em formato digital o que costuma fazer?

_____

_____

_____

Motivação - Assinale o número que corresponde melhor à sua resposta, sendo o mais objectivo possível.

1. É simples aprender a utilizar esta aplicação

1	2	3	4	5
Discordo totalmente				Concordo totalmente

2. É simples usar esta aplicação

1 2 3 4 5

Discordo totalmente Concordo totalmente

3. É divertido utilizar esta aplicação

1 2 3 4 5

Discordo totalmente Concordo totalmente

4. Utilizaria esta aplicação para anotar as minhas imagens

1 2 3 4 5

Discordo totalmente Concordo totalmente

5. Usaria esta aplicação num sitio público para passar o tempo (aeroporto, cinema, hospital, etc.)

1 2 3 4 5

Discordo totalmente Concordo totalmente

6. Utilizaria esta aplicação para me divertir com amigos/família

1 2 3 4 5

Discordo totalmente Concordo totalmente

Dinâmica do jogo - Assinale o número que corresponde melhor à sua resposta, sendo o mais objectivo possível.

1. Consigo perceber como a pontuação vai mudando ao longo do tempo

1 2 3 4 5

Discordo totalmente Concordo totalmente

2. Percebi que estava a fazer boas ou más anotações

1 2 3 4 5

Discordo totalmente Concordo totalmente

3. As imagens deveriam estar paradas, apenas as anotações deveriam rodar

1 2 3 4 5

Discordo totalmente Concordo totalmente

4. As anotações deveriam estar paradas, apenas as imagens deveriam rodar

1 2 3 4 5

Discordo totalmente Concordo totalmente

5. Seria mais divertido usar imagens minhas com as minhas próprias anotações

1 2 3 4 5

Discordo totalmente Concordo totalmente

6. A aplicação seria mais fácil/intuitiva se usasse teclado / rato

1 2 3 4 5

Discordo totalmente Concordo totalmente

7. A aplicação funcionaria melhor com mais imagens

1 2 3 4 5

Discordo totalmente Concordo totalmente

8. A aplicação funcionaria melhor com mais anotações

1 2 3 4 5

Discordo totalmente Concordo totalmente

9. Quais as principais alterações que faria à interface em termos de dinâmica de jogo (objectos no jogo, pontuações, etc.) ?

_____

_____

_____

Interacção - Assinale o número que corresponde melhor à sua resposta

1. É fácil manejar os "hotspots"que rodam imagens/conceitos

1 2 3 4 5

Discordo totalmente Concordo totalmente

2. Usar este tipo de interacção é fisicamente desgastante

1 2 3 4 5

Discordo totalmente Concordo totalmente

3. Usar este tipo de interacção é mentalmente desgastante

1 2 3 4 5

Discordo totalmente Concordo totalmente

4. A imagem que mostra o utilizador/hotspots é pequena demais

1 2 3 4 5

Discordo totalmente Concordo totalmente

Estética - Assinale o número que corresponde melhor à sua resposta

1. O aspecto estético da interface agrada-me

1 2 3 4 5

Discordo totalmente Concordo totalmente

2. Considero, em termos gerais, uma interface agradável

1 2 3 4 5

Discordo totalmente Concordo totalmente

3. Utilizaria esta interface para uso pessoal

1 2 3 4 5

Discordo totalmente Concordo totalmente

4. Em termos estéticos, quais as principais alterações que faria à interface ?

_____

_____

_____

5. Em termos gerais, qual a sua opinião desta interface ?

_____

_____

_____

B.2 Resultados

	Info			**1 - General**				**2 - Motivational**					
Utilizador	Age	Sex	IT	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6
1	18	F	N	4	4	3	1,2	4	5	4	3	4	4
2	25	M	S	5	4	4	2	4	4	5	4	5	5
3	19	F	N	5	3	2	1,2	3	4	4	3	3	4
4	25	F	S	4	5	4	1,2,3	5	4	5	4	5	5
5	24	F	N	5	5	5	1,2	5	5	5	5	5	5
6	24	F	S	5	5	4	1,2	5	5	5	5	5	5
7	24	F	N	5	5	5	1,2,3	4	5	5	4	5	5
8	26	F	S	5	3	3	1,2	4	4	5	4	4	5
9	27	M	S	4	2	2	1,2	5	4	3	2	5	5
10	31	F	F	4	5	1	1,2	4	4	4	2	5	3
11	25	M	S	5	2	1	1,2	4	3	5	4	5	5
12	24	M	S	4	4	4	1,2	4	4	4	2	4	4
13	26	M	S	5	5	3	2	5	5	4	3	3	3
14	23	M	S	4	2	2	1	4	5	5	4	3	4
15	19	M	S	5	3	2	2	4	4	4	4	5	5
Average	24			4.60	3.80	3.00	1.75	4.27	4.33	4.47	3.53	4.40	4.47
SD	3.33			0.51	1.21	1.31	0.50	0.59	0.62	0.64	0.99	0.83	0.74

Tabela B.1: Respostas dos utilizadores aos grupos de perguntas do questionário dos blocos dados pessoais, geral e motivação.

	3 - Dinâmica do Jogo							
Utilizador	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8
1	5	4	2	2	4	3	3	3
2	3	5	1	1	4	3	3	3
3	4	4	2	2	3	2	3	4
4	1	5	1	1	5	4	3	3
5	3	4	1	1	5	1	3	3
6	4	4	3	3	4	1	3	3
7	5	4	1	1	4	1	4	2
8	3	4	1	1	3	1	2	3
9	4	2	1	1	2	5	2	2
10	1	5	1	1	5	3	2	2
11	2	5	1	1	5	3	3	3
12	2	5	1	1	5	4	3	4
13	4	3	1	1	4	2	3	5
14	4	3	2	2	2	4	4	3
15	3	4	3	3	4	1	5	2
Average	3.20	4.07	1.47	1.47	3.93	2.53	3.07	3.00
SD	1.26	0.88	0.74	0.74	1.03	1.36	0.80	0.85

Tabela B.2: Respostas dos utilizadores ao grupo de perguntas do bloco dinâmica do jogo.

	4 - Interacção				5 - Estética		
Utilizador	4.1	4.2	4.3	4.4	5.1	5.2	5.3
1	4	3	3	2	4	4	3
2	4	1	1	2	4	4	4
3	2	4	3	2	4	4	4
4	2	2	2	4	3	4	5
5	3	1	1	3	4	5	5
6	5	1	1	1	5	5	5
7	4	2	2	2	4	4	5
8	3	1	2	2	3	4	5
9	2	1	1	3	3	3	2
10	4	5	3	1	3	4	2
11	2	2	1	2	4	5	5
12	4	1	1	3	3	4	4
13	4	3	1	2	4	4	4
14	5	2	1	2	5	5	5
15	2	3	1	5	4	4	5
Average	3.33	2.13	1.60	2.40	3.80	4.20	4.20
SD	1.11	1.25	0.83	1.06	0.68	0.56	1.08

Tabela B.3: Respostas dos utilizadores aos grupos de perguntas dos blocos, interacção e estética.

Utilizador	Idade	Pontuação	Tempo de Jogo	Nível
1	18	360	3:20	4
2	25	1347	5:13	6
3	19	1045	4:25	6
4	25	263	2:18	3
5	24	960	4:13	6
6	24	1195	4:12	6
7	24	1385	4:39	6
8	26	1563	4:30	6
9	27	680	3:55	6
10	31	2159	5:12	6
11	25	1649	5:20	6
12	24	2511	6:00	6
13	26	2474	5:50	6
14	23	1647	5:00	6
15	19	898	3:40	4

Tabela B.4: Desempenho dos utilizadores no jogo durante os testes de usabilidade.

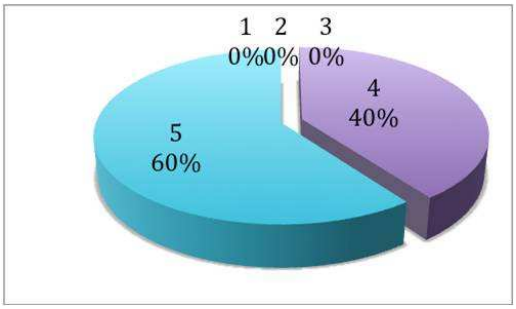


Figura B.3: Geral - 1. Costuma utilizar a internet para fazer pesquisas de imagens?

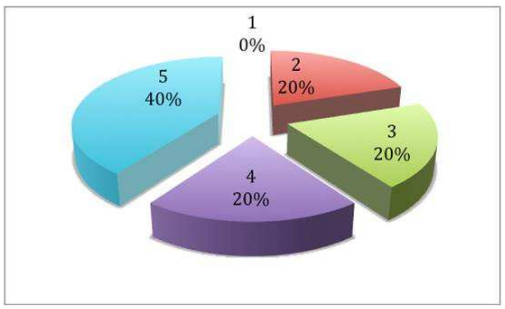


Figura B.4: Geral - 2. Costuma organizar imagens pessoais no seu computador?

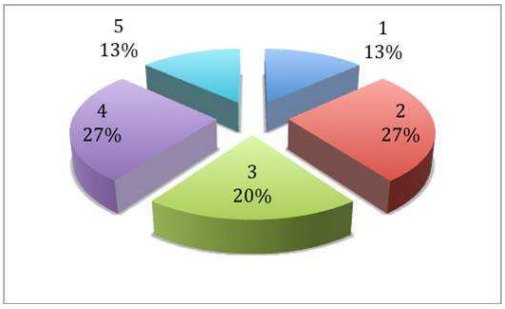


Figura B.5: Geral - 3. As suas imagens pessoais/pesquisadas estão catalogadas?

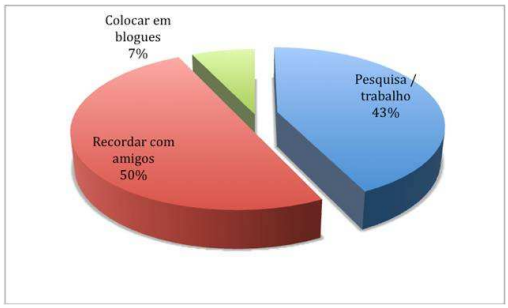


Figura B.6: Geral - 4. De que modo utiliza as imagens guardadas no seu computador?

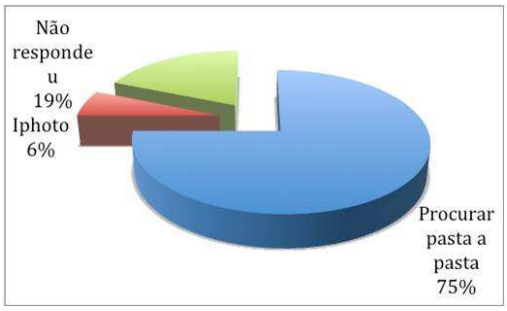


Figura B.7: Geral - 5. Quando pretende pesquisar as suas fotos pessoais em formato digital o que costuma fazer?

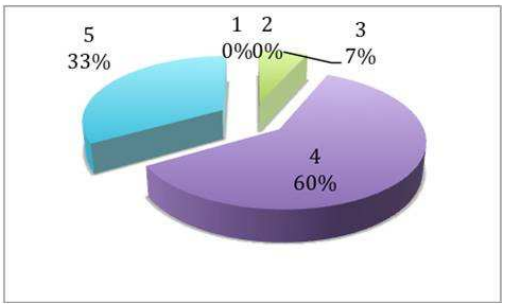


Figura B.8: Motivação - 1. É simples aprender a utilizar esta aplicação

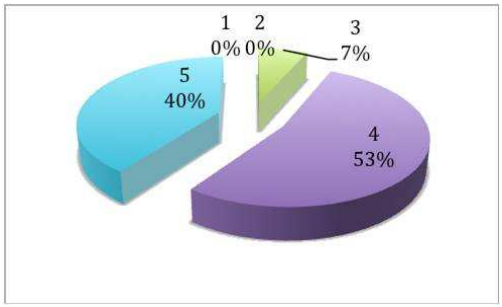


Figura B.9: Motivação - 2. É simples usar esta aplicação

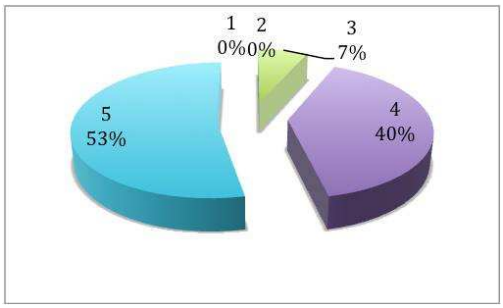


Figura B.10: Motivação - 3. É divertido utilizar esta aplicação

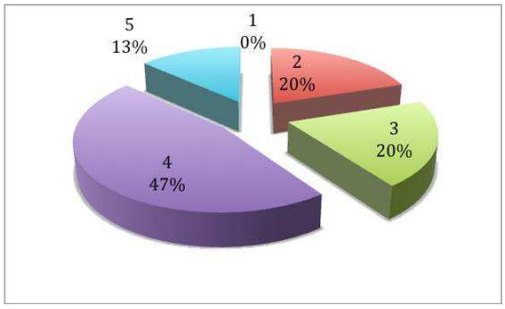


Figura B.11: Motivação - 4. Utilizaria esta aplicação para anotar as minhas imagens

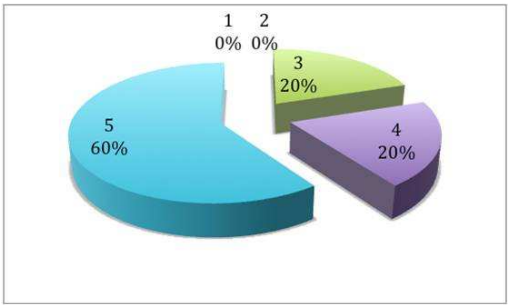


Figura B.12: Motivação - 5. Usaria esta aplicação num sitio público para passar o tempo (aeroporto, cinema, hospital, etc.)

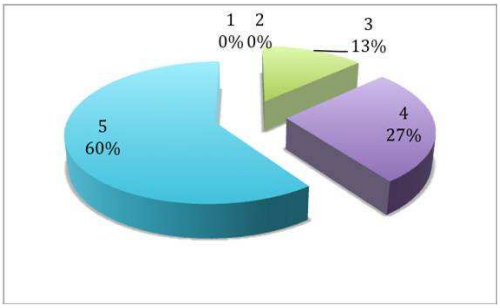


Figura B.13: Motivação - 6. Utilizaria esta aplicação para me divertir com amigos/família

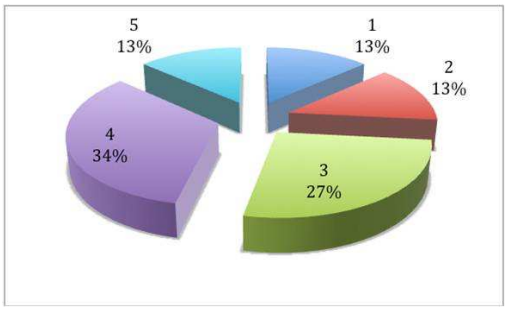


Figura B.14: Dinâmica de Jogo - 1. Consigo perceber como a pontuação vai mudando ao longo do tempo

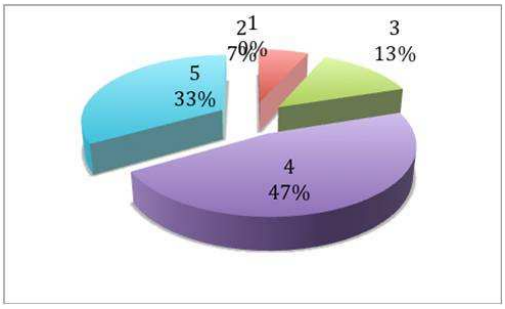


Figura B.15: Dinâmica de Jogo - 2. Percebi que estava a fazer boas ou más anotações

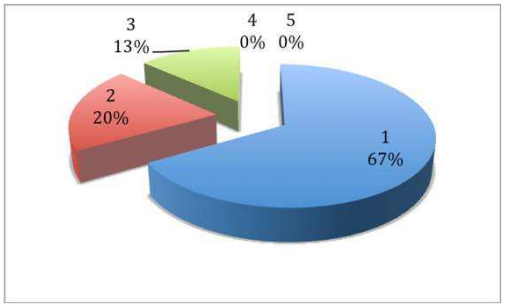


Figura B.16: Dinâmica de Jogo - 3. As imagens deveriam estar paradas, apenas as anotações deveriam rodar

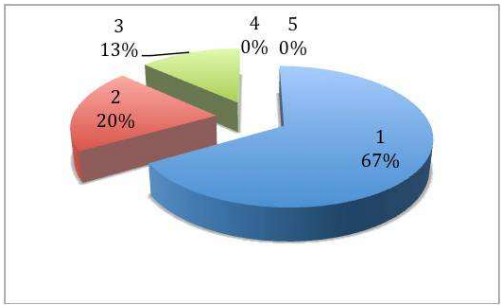


Figura B.17: Dinâmica de Jogo - 4. As anotações deveriam estar paradas, apenas as imagens deveriam rodar

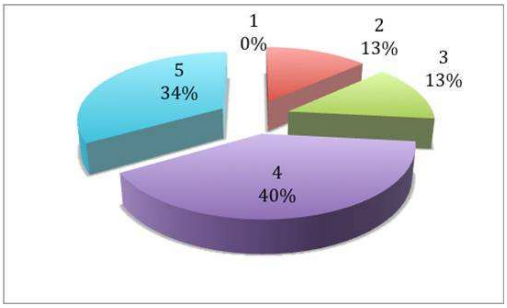


Figura B.18: Dinâmica de Jogo - 5. Seria mais divertido usar imagens minhas com as minhas próprias anotações

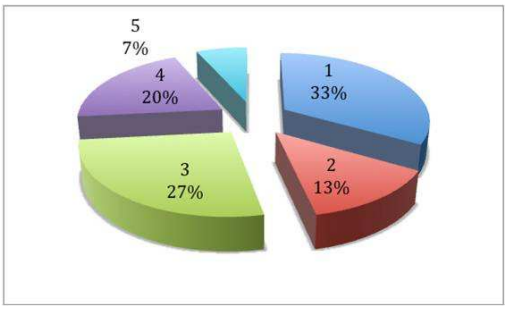


Figura B.19: Dinâmica de Jogo - 6. A aplicação seria mais fácil/intuitiva se usasse teclado / rato

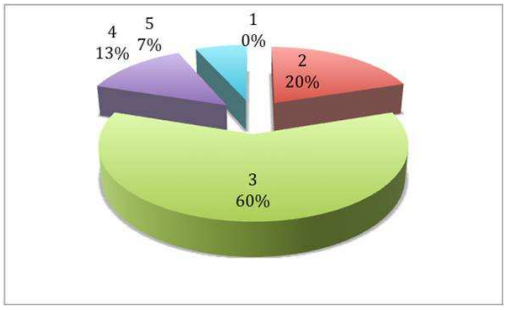


Figura B.20: Dinâmica de Jogo - 7. A aplicação funcionaria melhor com mais imagens

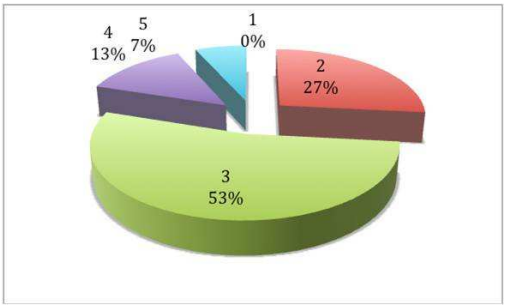


Figura B.21: Dinâmica de Jogo - 8. A aplicação funcionaria melhor com mais anotações

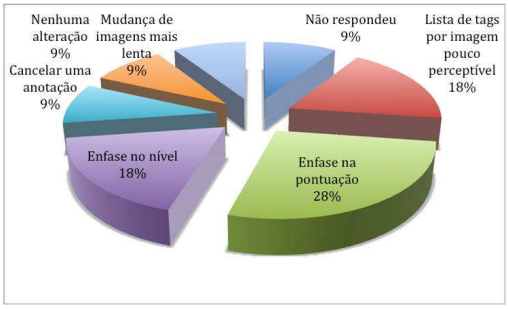


Figura B.22: Dinâmica de Jogo - 9. Quais as principais alterações que faria à interface em termos de dinâmica de jogo (objectos no jogo, pontuações, etc.) ?

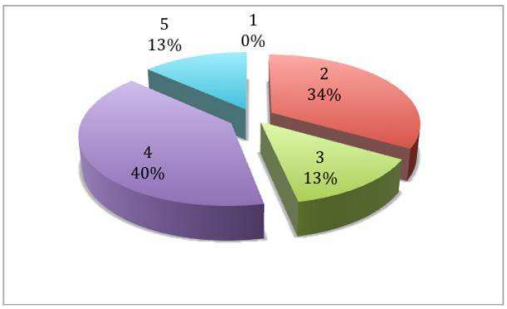


Figura B.23: Interacção - 1. É fácil manejar os "hotspots"que rodam imagens/conceitos

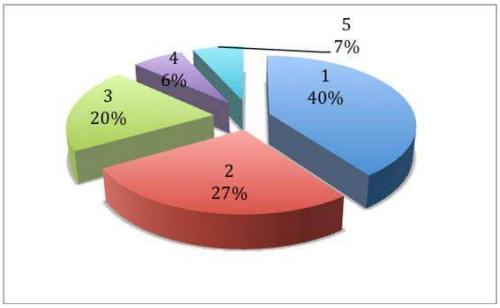


Figura B.24: Interacção - 2. Usar este tipo de interacção é fisicamente desgastante

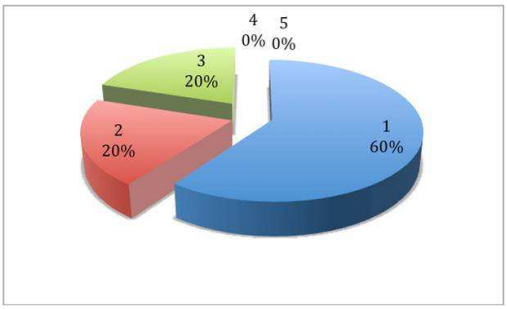


Figura B.25: Interacção - 3. Usar este tipo de interacção é mentalmente desgastante

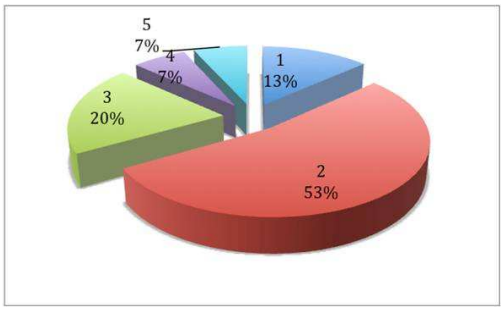


Figura B.26: Interacção - 4. A imagem que mostra o utilizador/hotspots é pequena demais

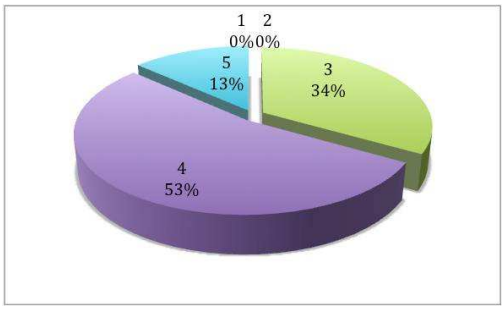


Figura B.27: Estética - 1. O aspecto estético da interface agrada-me

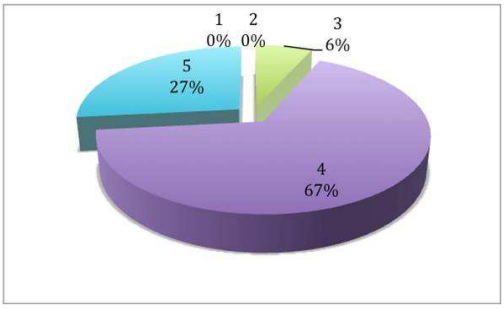


Figura B.28: Estética - 2. Considero, em termos gerais, uma interface agradável

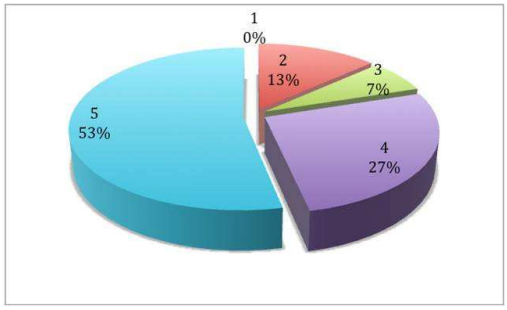


Figura B.29: Estética - 3. Utilizaria esta interface para uso pessoal

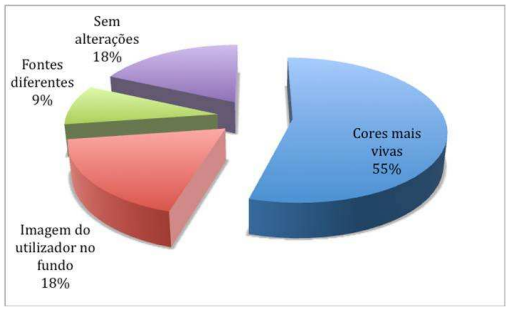


Figura B.30: Estética - 4. Em termos estéticos, quais as principais alterações que faria à interface?

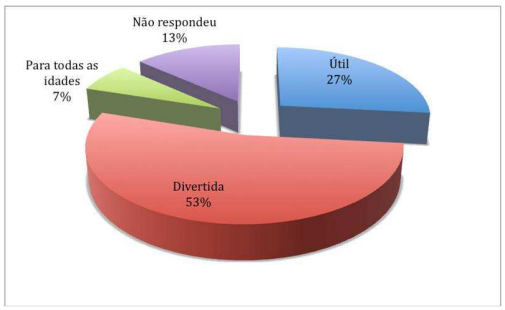


Figura B.31: Estética - 5. Em termos gerais, qual a sua opinião desta interface?

C

Recuperação de Imagens - Resultados

C.1 Pesquisa por Composição

	Imagem composta	Imagem 1	Imagem 2	Imagem 3
Edifícios com uma arquitectura específica	0,56	0,45	0,44	0,33
Plantas de uma determinada espécie	0,81	0,85	0,69	0,37
Imagens de neve	0,13	0,13	0,05	0,04
Fotos da casa do autor da tese	0,19	0,15	0,11	-
Média	0,42	0,40	0,32	0,25

Tabela C.1: Precisão obtida nas primeiras 100 imagens por várias pesquisas com imagens compostas.

C.2 Características Visuais e Metadados Contextuais

Conceitos	Momentos de cor x Gabor	Momentos de cor + Gabor + Momentos de cor x Gabor
People	0,66	0,71
Face	0,58	0,53
Outdoor	0,90	0,90
Indoor	0,55	0,59
Nature	0,44	0,39
Manmade	0,61	0,59
Snow	0,31	0,25
Beach	0,33	0,31
Party	0,13	0,15
MAP	**0,50**	**0,49**

Tabela C.2: MAP para vários conceitos utilizando as características, momentos de cor e banco de filtros de Gabor concatenados num vector (Momentos de cor x Gabor).

Conceitos	Melhores 5	Superior a th=0,5
Beach	0,76	0,66
Face	0,47	0,42
Indoor	0,33	0,31
Manmade	0,58	0,59
Nature	0,44	0,40
Outdoor	0,87	0,85
Party	0,42	0,41
People	0,54	0,53
Snow	0,38	0,35
Média	**0,53**	**0,50**

Tabela C.3: Precisão por conceito para duas técnicas de anotação: "Os melhores 5" e "Superior a th=0,5".

Conceitos	Melhores 5	Superior a th=0,5
Beach	0,09	0,20
Face	0,50	0,66
Indoor	0,77	0,85
Manmade	0,48	0,64
Nature	0,25	0,36
Outdoor	0,25	0,32
Party	0,23	0,37
People	0,88	0,93
Snow	0,11	0,19
Média	**0,40**	**0,50**

Tabela C.4: Cobertura por conceito para duas técnicas de anotação: "Os melhores 5" e "Superior a th=0,5".

Conceitos	GPS + Visual	GPS + Visual + Áudio
Outdoor	0,8	0,9
Indoor	0,1	0,2
Nature	0,6	0,5
Manmade	0,6	0,7
People	0	0,3
Indoor + Manmade	0,3	0,3
Outdoor + Nature	0,6	0,7
Média	**0,43**	**0,51**

Tabela C.5: Pesquisa de imagens numa direcção - precisão para vários conceitos utilizando informação de GPS, áudio e características visuais e considerando 10 imagens recuperadas. A pesquisa foi realizada utilizando como localização a entrada da Capela e a direcção SUL.

Conceitos	GPS 60m + Visual	GPS 60m + Visual + Audio
Outdoor	1,0	1,0
Indoor	0,2	0,2
Nature	0,9	0,9
Manmade	0, 7	0,7
People	0,1	0,1
Indoor + Manmade	0,2	0,2
Outdoor + Nature	0,8	0,8
Média	**0,53**	**0,53**

Tabela C.6: Pesquisa de imagens no Patamar dos Deuses considerando um raio de 60 metros - precisão para vários conceitos utilizando informação de GPS, áudio e características visuais e considerando 10 imagens recuperadas.

Conceitos	GPS 60m + Visual	GPS 60m + Visual + Audio
Outdoor	0,6	0,6
Indoor	0,8	0,7
Nature	0,4	0,5
Manmade	0, 8	1,0
People	0,1	0,4
Indoor + Manmade	0,3	0,4
Outdoor + Nature	0,4	0,4
Média	**0,43**	**0,57**

Tabela C.7: Pesquisa de imagens à entrada da Capela considerando um raio de 60 metros - precisão para vários conceitos utilizando informação de GPS, áudio e características visuais e considerando 10 imagens recuperadas.

Conceitos	GPS + Visual	GPS + Visual + Audio
Outdoor	0,91	0,91
Indoor	0,11	0,19
Nature	0,69	0,72
Manmade	0,48	0,49
People	0,09	0,11
Indoor + Manmade	0,12	0,11
Outdoor + Nature	0,71	0,73
MAP	**0,44**	**0,47**

Tabela C.8: Pesquisa de imagens numa direcção - MAP obtido por vários conceitos utilizando informação de GPS, áudio e características visuais e considerando 10 imagens recuperadas. A pesquisa foi realizada utilizando como localização a entrada da Capela e a direcção SUL.

Conceitos	GPS 60m + Visual	GPS 60m + Visual + Audio
Outdoor	0,67	0,70
Indoor	0,52	0,55
Nature	0,41	0,50
Manmade	0,80	0,84
People	0,27	0,40
Indoor + Manmade	0,33	0,41
Outdoor + Nature	0,49	0,53
Média	**0,50**	**0,56**

Tabela C.9: Pesquisa de imagens à entrada da Capela considerando um raio de 60 metros - MAP obtido por vários conceitos utilizando informação de GPS, áudio e características visuais.

Bibliografia

[Acdsee01] Pocket ACDSee. www.acdsee.com, 2001, *último acesso 4/01/2009*.

[Airliners05] Airliners.net homepage. http://www.airliners.net, 2005, *último acesso 26/01/2009*.

[Aizawa01] Kiyoharu Aizawa, Kenichiro Ishijima, and Makoto Shiina. Summarizing wearable video. In *Proceedings of the International Conference on Image Processing*, Thessaloniki, Greece, volume 3, pages 398–401, 2001.

[Anguera08] Xavier Anguera, JieJun Xu, and Nuria Oliver. Multimodal photo annotation and retrieval on a mobile phone. In *MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, Vancouver, British Columbia, Canada, pages 188–194, 2008.

[Apted06] Trent Apted, Judy Kay, and Aaron Quigley. Tabletop sharing of digital photographs for the elderly. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, pages 781–790, 2006.

[Balabanovic00] Marko Balabanović, Lonny Chu, and Gregory Wolff. Storytelling with digital photographs. In *CHI '00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, The Hague, The Netherlands, pages 564–571, 2000.

[Baluja07] Shumeet Baluja and Henry Rowley. Boosting sex identification performance. *International Journal of Computer Vision*, 71(1):111–119, Kluwer Academic Publishers, 2007.

[Barnard01] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, Vancouver, British Columbia, Canada, volume 2, pages 408–415, 2001.

[Bartolini05] Ilaria Bartolini, Marco Patella, and Paolo Ciaccia. WARP: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance. *IEEE Transactions on Pattern Analysis and Machine Intellenge*, 27(1):142–147, 2005.

[Baus05] Jörg Baus, Christian Kray, and Keith Cheverst. A Survey of Map-based Mobile Guides. In Liquiu Meng, Alexander Zipf, and Tumasch Reichenbacher, editors, *Map-based Mobile Services*, pages 197–216. Springer, Berlin, Heidelberg, New York, 2005.

[Beagrie05] Neil Beagrie. Plenty of Room at the Bottom? Personal Digital Libraries and Collections. *Dlib Magazine*, 11(6), June 2005.

[Bederson01] Benjamin B. Bederson. PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps. In *UIST '01: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, Orlando, Florida, USA, pages 71–80, 2001.

[Beeharee06] Ashweeni Beeharee and Anthony Steed. A natural wayfinding exploiting photos in pedestrian navigation systems. In *MobileHCI '06: Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*, Helsinki, Finland, pages 81–88, 2006.

[Bimbo99] Alberto Del Bimbo. *Visual information retrieval*. Morgan Kaufmann Publishers Inc., 1999.

[Boardman04] Richard Boardman and Martina Sasse. "Stuff goes into the computer and doesn't come out": a cross-tool study of personal information management. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria, pages 583–590, 2004.

[Boldareva03] Lioudmila Boldareva, Djoerd Hiemstra, and Willem Jonker. Relevance feedback in probabilistic multimedia retrieval. In *DELOS Workshop on Multimedia Contents in Digital Libraries*, 2003.

[Bosch06] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, Graz, Austria, volume 3954 of *Lecture Notes in Computer Science*, pages 517–530. Springer, 2006.

[Brown03] Barry Brown and Matthew Chalmers. Tourism and mobile technology. In *ECSCW'03: Proceedings of the eighth Conference on European Conference on Computer Supported Cooperative Work*, Helsinki, Finland, pages 335–354, 2003.

[Buijs99] Jean Buijs and Michael Lew. Visual learning of simple semantics in imageScape. In *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems*, Amsterdam, The Netherlands, pages 131–138, 1999.

[Bush45] Vannevar Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[Campbell00] Iain Campbell. *The ostensive model of developing information needs*. PhD thesis, University of Glasgow, Scotland, 2000.

[Carneiro05] Gustavo Carneiro and Nuno Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, San Diego, California, USA, pages 163–168, 2005.

[Carson02] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using Expectation-Maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[Chang03] Edward Chang, Kingshy Goh, Gerard Sychay, and Gang Wu. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38, Jan 2003.

[Chang05] Edward Y. Chang. EXTENT: fusing context, content, and semantic ontology for photo annotation. In *CVDB '05: Proceedings of the 2nd International Workshop on Computer Vision Meets Databases*, Baltimore, Maryland, USA, pages 5–11, 2005.

[Chang98] Shih-Fu Chang, William Chen, and Hari Sundaram. Semantic visual templates: linking visual features to semantics. In *Proceedings of the International Conference on Image Processing*, Chicago, Illinois, USA, volume 3, pages 531–535, Oct 1998.

[Chapelle99] Olivier Chapelle, Patrick Haffner, and Vladimir Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, Sep 1999.

[Chen01] Yunqiang Chen, Xiang Sean Zhou, and Thomas Huang. One-class SVM for learning in image retrieval. In *Proceedings of the International Conference on Image Processing*, Thessaloniki, Greece, volume 1, pages 34–37, 2001.

[Cheng01] H. Cheng, X. Jiang, Y. Sun, and Jing Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259–2281, Elsevier, December 2001.

[Chevallet07] Jean-Pierre Chevallet, Joo-Hwee Lim, and Mun-Kew Leong. Object identification and retrieval from efficient image matching. Snap2Tell with the STOIC dataset. *International Journal on Information Processing and Management*, 43(2):515–530, Pergamon Press, Inc., 2007.

[Cho07] Sung-Jung Cho, Roderick Murray-Smith, and Yeun-Bae Kim. Multi-context photo browsing on mobile devices based on tilt dynamics. In *MobileHCI '07: Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services*, Singapore, pages 190–197, 2007.

[Choi02] Youngok Choi and Edie Rasmussen. Users' relevance criteria in image retrieval in American history. *Information Processing & Management*, 38(5):695 –726, Pergamon Press, Inc., 2002.

[Choi03] Youngok Choi and Edie Rasmussen. Searching for images: the analysis of users' queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology*, 54(6):498–511, John Wiley & Sons, Inc., 2003.

[Choi08] Jae Young Choi, Seungji Yang, Yong Man Ro, and Konstantinos N. Plataniotis. Face annotation for personal photos using context-assisted face recognition. In *MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, Vancouver, British Columbia, Canada, pages 44–51, 2008.

[Clarkson02] Brian Clarkson. *Life Patterns: structure from wearable sensors*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, September 2002.

[Clawson08] James Clawson, Amy Voida, Nirmal Patel, and Kent Lyons. Mobiphos: a collocated-synchronous mobile photo sharing application. In *MobileHCI '08: Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, Amsterdam, The Netherlands, pages 187–195, 2008.

[Comaniciu02] Dorin Comaniciu and Peter Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[Correia05] Nuno Correia, Luís Alves, Helder Correia, Luis Romero, Carmen Morgado, Luís Soares, José Cunha, ao Teresa Rom Eduardo Dias, and Joaquim Jorge. InStory: a system for mobile information access, storytelling and gaming activities in physical spaces. In *ACE '05: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, Valencia, Spain, pages 102–109, 2005.

[Cox00] I. Cox, M. Miller, T. Minka, T. Papathomas, and P. Yianilos. The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, Jan 2000.

[Cox96] Ingemar Cox, Matt Miller, Stephen Omohundro, and Peter Yianilos. PicHunter: Bayesian relevance feedback for image retrieval. Vienna, Austria, volume 3, pages 361–369, Aug 1996.

[Cunningham04] Sally Cunningham, David Bainbridge, and Masood Masoodian. How people describe their image information needs: a grounded theory analysis of visual arts queries. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tuscon, Arizona, USA, pages 47–48, 2004.

[Czerwinski06] Mary Czerwinski, Douglas Gage, Jim Gemmell, Catherine Marshall, Manuel Pérez-Quinones, Meredith Skeels, and Tiziana Catarci. Digital memories in an era of ubiquitous computing and abundant storage. *Communications of the ACM*, 49(1):44–50, 2006.

[Datta08] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[Davis04] Marc Davis, Simon King, Nathan Good, and Risto Sarvas. From context to content: leveraging context to infer media metadata. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, New York, New York, USA, pages 188–195, 2004.

[Davis06] Marc Davis, Michael Smith, Fred Stentiford, Adetokunbo Bamidele, John Canny, Nathan Good, Simon King, and Rajkumar Janakiraman. Using context and similarity for face and location identification. In *Proceedings of the IS&T/SPIE 18th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, California, USA. 2006.

[Deerwester90] Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[Dempster77] A. Dempster, M. Laird, and B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, Blackwell Publishing for the Royal Statistical Society, 1977.

[Dias07] Ricardo Dias, Rui Jesus, Rute Frias, and Nuno Correia. Mobile interface of the memoria project. In *SIGIR '07: Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, pages 904–904, 2007.

[Dix00] Alan Dix, Tom Rodden, Nigel Davies, Jonathan Trevor, Adrian Friday, and Kevin Palfreyman. Exploiting space and location as a design framework for interactive mobile systems. *ACM Transaction on Computer-Human Interaction*, 7(3):285–321, 2000.

[Dix02] Alan Dix. The ultimate interface and the sums of life? *Interfaces*, 50:16, 2002.

[Do02] Minh Do and Martin Vetterli. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158, Feb 2002.

[Dong03] Anlei Dong and Bir Bhanu. Active concept learning for image retrieval in dynamic databases. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Nice, France, volume 1, pages 90–95, 2003.

[Dourish00] Paul Dourish, Keith Edwards, Anthony LaMarca, John Lamping, Karin Petersen, Michael Salisbury, Douglas Terry, and James Thornton. Extending document management systems with user-specific active properties. *ACM Transactions on Information Systems*, 18(2):140–170, 2000.

[Dumais03] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, Toronto, Canada, pages 72–79, 2003.

[Duygulu02] P. Duygulu, Kobus Barnard, J. de Freitas, and David Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, Copenhagen, Denmark, pages 97–112, 2002.

[Eakins04] John Eakins, Pamela Briggs, and Bryan Burford. Image retrieval interfaces: a user perspective. In *Third International Conference on Image and Video Retrieval (CIVR)*, Dublin, Ireland, volume 3115 of *Lecture Notes in Computer Science*, pages 628–637. Springer, 2004.

[Endel02] Endel Tulving. Episodic memory: from mind to brain. *Annual Review of Psychology*, 53(1):1–25, 2002.

[Engelbart68] Douglas Engelbart and William English. A research center for augmenting human intellect. In *AFIPS Conference Proceedings of the 1968 Fall Joint Computer Conference*, volume 3, pages 395–410, December 1968.

[Exif98] Exchangeable Image File Format. http://www.exif.org, 1998, *último acesso 26/01/2009*.

[EyeToy05] EyeToy. http://www.eyetoy.com, 2005, *último acesso 13/03/2009*.

[Fan05] Xin Fan, Xing Xie, Zhiwei Li, Mingjing Li, and Wei-Ying Ma. Photo-to-search: using multimodal queries to search the web from mobile devices. In *MIR '05: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Hilton, Singapore, pages 143–150, 2005.

[Fan07] Jianping Fan, Yuli Gao, and Hangzai Luo. Hierarchical classification for automatic image annotation. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, pages 111–118, 2007.

[Feng04] S. L. Feng, R. Manmatha, and V. Lavrenko. F. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, District of Columbia, USA, volume 2, pages 1002–1009, 2004.

[Ferecatu05] M. Ferecatu, M. Crucianu, and N. Boujemaa. Improving performance of interactive categorization of images using relevance feedback. In *Proceedings of the IEEE International Conference on Image Processing*, Genoa, Italy, volume 1, pages I–1197–1200, Sept. 2005.

[Fitzgibbon03] Andrew Fitzgibbon and Ehud Reiter. Memories for Life - Managing information over a human lifetime. *Included in Grand Challenges for Computing Research, Sponsored by the UK Computing Research Committee, with support from EPSRC and NeSC*, 2003.

[Flickner95] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by Image and Video Content: The QBIC System. *Computer*, 28(9):23–32, IEEE Computer Society Press, 1995.

[Flickr04] Flickr homepage. http://www.flickr.com, 2004, *último acesso 26/01/2009*.

[Fockler05] Paul Föckler, Thomas Zeidler, Benjamin Brombach, Erich Bruns, and Oliver Bimber. PhoneGuide: museum guidance supported by on-device object recognition on mobile phones. In *MUM '05: Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia*, Christchurch, New Zealand, pages 3–10, 2005.

[Fototagger06] Fototagger. http://www.fototagger.com, 2006, *último acesso 26/01/2009*.

[Freeman96] Eric Freeman and David Gelernter. Lifestreams: a storage model for personal data. *SIGMOD Record*, 25(1):80–86, ACM, 1996.

[Frohlich02] David Frohlich, Allan Kuchinsky, Celine Pering, Abbe Don, and Steven Ariss. Requirements for photoware. In *CSCW '02: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, New Orleans, Louisiana, USA, pages 166–175, 2002.

[Frohlich08] David Frohlich and Matt Jones. Audiophoto narratives for semi-literate communities. *Interactions*, 15(6):61–64, ACM, 2008.

[Gemmell02] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. MyLifeBits: fulfilling the Memex vision. In *MULTIMEDIA '02: Proceedings of the tenth ACM International Conference on Multimedia*, Juan-les-Pins, France, pages 235–238, 2002.

[Gemmell04] Jim Gemmell, Lyndsay Williams, Ken Wood, Roger Lueder, and Gordon Bell. Passive capture and ensuing issues for a personal lifetime store. In *CARPE'04: Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, New York, New York, USA, pages 48–55, 2004.

[Gemmell06] Jim Gemmell, Gordon Bell, and Roger Lueder. MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.

[Gever00] Theo Gevers and Arnold Smeulders. PicToSeek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9(1):102 – 119, 2000.

[Giacinto07] Giorgio Giacinto. A nearest-neighbor approach to relevance feedback in content based image retrieval. In *CIVR '07: Proceedings of the 6th ACM International Conference on Image and video retrieval*, Amsterdam, The Netherlands, pages 456–463, 2007.

[Goncalves08] Duarte Gonçalves, Rui Jesus, and Nuno Correia. A gesture based game for image tagging. In *CHI '08: CHI '08 Extended Abstracts on Human Factors in Computing Systems*, Florence, Italy, pages 2685–2690, 2008.

[Goncalves08a] Duarte Gonçalves, Rui Jesus, Filipe Grangeiro Teresa Romão, and Nuno Correia. Tag Around: a 3D gesture game for image annotation. In *ACE '08: Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, Yokohama, Japan, pages 259–262, 2008.

[Goncalves08b] Duarte Gonçalves, Rui Jesus, Filipe Grangeiro, and Nuno Correia. Tag Around - interface gestual para anotação de imagens. In *3rd Portuguese Conference on Human Computer Interaction*, Évora, Portugal, 2008.

[Goncalves08c] Duarte Gonçalves. Gesture based interface for image annotation. Master's thesis, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 2008.

[Gong94] Yihong Gong, Hongjiang Zhang, C. Chuan, and M. Sakauchi. An image database system with content capturing and fast image indexing abilities. Boston, Massachusetts, USA, pages 121–130, May 1994.

[Graham02] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd. Time as essence for photo browsing through personal digital libraries. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, Oregon, USA, pages 326–335, 2002.

[Grangeiro08] Filipe Grangeiro, Rui Jesus, and Nuno Correia. Detecção e reconhecimento de faces para aplicações multimédia. In *4ª Jornadas de Engenharia de Electrónica e Telecomunicações e de Computadores*, Lisboa, Portugal, 2008.

[Grangeiro08a] Filipe Grangeiro. Detecção e reconhecimento de faces em aplicações multimédia. Master's thesis, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 2008.

[Grangeiro09] Filipe Grangeiro, Rui Jesus, and Nuno Correia. Face recognition and gender classification in personal memories. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009*, Taipei, Taiwan, pages 1945–1948, April 2009.

[Gray03] Jim Gray. What next?: A dozen information-technology research goals. *Journal of the ACM*, 50(1):41–57, 2003.

[Gupta97] Amarnath Gupta and Ramesh Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, 1997.

[Gurrin05] Cathal Gurrin, Gareth Jones, Hyowon Lee, Neil O'Hare, Alan Smeaton, and Noel Murphy. Mobile access to personal digital photograph archives. In *MobileHCI '05: Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services*, Salzburg, Austria, pages 311–314, 2005.

[Haase04] Ken Haase and David Tamés. BabelVision: better image searching through shared annotations. *Interactions*, 11(2):18–26, ACM, 2004.

[Hagita03] Norihiro Hagita, Kiyoshi Kogure, Kenji Mase, and Yasuyuki Sumi. Collaborative capturing of experiences with ubiquitous sensors and communication robots. In *Proceedings of International Conference on Robotics and Automation '03*, Taipei, Taiwan, volume 3, pages 4166–4171, 2003.

[Halaschek05] Christian Halaschek-Wiener, Jennifer Golbeck, Andrew Schain, Michael Grove, Bijan Parsia, and Jim Hendler. Photostuff - an image annotation tool for the semantic web. In *4th International Semantic Web Conference*, Galway, Ireland, 2005.

[Harada04] Susumu Harada, Mor Naaman, Yee Jiun Song, QianYing Wang, and Andreas Paepcke. Lost in memories: interacting with photo collections on PDAs. In *Proceedings of the Joint ACM/IEEE Conference on Digital Libraries*, Tucson, Arizona, USA, pages 325–333, June 2004.

[Haralick73] Robert Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, Nov. 1973.

[Hare05] Neil O Hare, Cathal Gurrin, Gareth Jones, and Alan Smeaton. Combination of content analysis and context features for digital photograph retrieval. In *Proceedings of the IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, London, England, pages 323–328, 2005.

[Healey98] Jennifer Healey and Rosalind Picard. StartleCam: A cybernetic wearable camera. In *ISWC '98: Proceedings of the 2nd IEEE International Symposium on Wearable Computers*, Pittsburgh, Pennsylvania, USA, page 42, 1998.

[Heesch04] Daniel Heesch and Stefan Rüger. Three interfaces for content-based access to image collections. In *Third International Conference on Image and Video Retrieval (CIVR)*, Dublin, Ireland, volume 3115 of *Lecture Notes in Computer Science*, pages 491–499. Springer, 2004.

[Hirata92] Kyoji Hirata and Toshikazu Kato. Query by visual example - Content Bbased Image Retrieval. In *EDBT '92: Proceedings of the 3rd International Conference on Extending Database Technology*, Vienna, Austria, pages 56–71, 1992.

[Hoi04] Chu-Hong Hoi and Michael Lyu. A novel log-based relevance feedback technique in content-based image retrieval. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM International Conference on Multimedia*, New York, New York, USA, pages 24–31, 2004.

[Hori03] Tetsuro Hori and Kiyoharu Aizawa. Context-based video retrieval system for the life-log applications. In *MIR '03: Proceedings of the 5th ACM SIGMM International workshop on Multimedia Information Retrieval*, Berkeley, California, USA, pages 31–38, 2003.

[Huang07] Gary Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labelled Faces in the Wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst, Technical Report*, 57(2):07–49, 2007.

[Huang97] Jing Huang, Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pages 762–768, Jun 1997.

[Huynh02] David Huynh, David R. Karger, and Dennis Quan. Haystack: A platform for creating, organizing and visualizing information using RDF. In *Semantic Web Workshop*, Honolulu, Hawaii, USA, 2002.

[Hwang07] Amy Hwang, Shane Ahern, Simon King, Mor Naaman, Rahul Nair, and Jeannie Yang. Zurfer: mobile multimedia access in spatial, social and topical context. In *MULTIMEDIA '07: Proceedings of the 15th International Conference on Multimedia*, Augsburg, Germany, pages 557–560, 2007.

[Ishikawa98] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. MindReader: Querying databases through multiple examples. In *VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, New York, New York, USA, pages 218–227, 1998.

[Jansen00] Bernard Jansen, Abby Goodrum, and Amanda Spink. Searching for multimedia: analysis of audio, video and image Web queries. *World Wide Web Journal*, 3(4):249–254, Kluwer Academic Publishers, 2000.

[Jeon03] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, Toronto, Canada, pages 119–126, 2003.

[Jesus06] Rui Jesus, Arnaldo Abrantes, and Nuno Correia. Photo retrieval from personal memories using generic concepts. In *Advances in Multimedia Information Processing - PCM 2006, 7th Pacific Rim Conference on Multimedia*, Hangzhou, China, volume 4261 of *Lecture Notes in Computer Science*, pages 633–640. Springer, 2006.

[Jesus06a] Rui Jesus, Tiago Martins, Rute Frias, Arnaldo Abrantes, and Nuno Correia. PhotoNav: a system to capture, share and access personal memories. In *Memories for Life Colloquium*, London, England, 2006.

[Jesus07] Rui Jesus, Ricardo Dias, Rute Frias, Arnaldo Abrantes, and Nuno Correia. Sharing personal experiences while navigating in physical spaces. In *ACM SIGIR Conference on Research and Development in Information Retrieval, Multimedia Information Retrieval Workshop*, Amsterdam, The Netherlands, July 2007.

[Jesus07a] Rui Jesus, Ricardo Dias, Rute Frias, and Nuno Correia. Geographic image retrieval in mobile guides. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical Information Retrieval*, Lisbon, Portugal, pages 37–38, 2007.

[Jesus07b] Rui Jesus, Edgar Santos, Rute Frias, and Nuno Correia. An interface to explore personal memories. In *15th Portuguese Computer Graphics Group Conference*, Porto Salvo, Portugal, 2007.

[Jesus08] Rui Jesus, Duarte Goncalves, Arnaldo Abrantes, and Nuno Correia. Playing games as a way to improve automatic image annotation. In *CVPRW 08: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Workshop on Semantic Learning Applications in Multimedia (SLAM08) ,2008*, Anchorage, Alaska, USA, pages 1–8, June 2008.

[Jesus08a] Rui Jesus, Ricardo Dias, Rute Frias, Arnaldo Abrantes, and Nuno Correia. Memoria mobile: sharing pictures of a point of interest. In *AVI '08: Proceedings of the Working Conference on Advanced Visual Interfaces*, Napoli, Italy, pages 412–415, 2008.

[Jiayi03] C. Jiayi, T. Tan, and P. Mulhem. Using speech annotation for home digital image indexing and retrieval. In *Content Based Multimedia Indexing Conference (CBMI)*, Rennes, France, pages 195–200, 2003.

[Jing03] F. Jing, M. Li, L. Zhang, H. Zhang, and B. Zhang. Learning in region-based image retrieval. In *International Conference on Image and Video Retrieval*, Urbana-Champaign, Illinois, USA, volume 2728 of *Lecture Notes in Computer Science*, pages 198–207. Springer, 2003.

[Jing05] Feng Jing, Mingjing Li, HongJiang Zhang, and Bo Zhang. A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 14(7):979–989, 2005.

[Jones03] Karen Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, MCB UP Ltd, 1972.

[Jones05] William Jones. A review of Personal Information Management. *IS-TR-2005-11-01. The Information School Technical Repository, University of Washington, Seattle*, 2005.

[Joshi06] Dhiraj Joshi, Ritendra Datta, Ziming Zhuang, W. Weiss, Marc Friedenberg, Jia Li, and James Wang. PARAgrab: a comprehensive architecture for web image management and multimodal querying. In *VLDB '06: Proceedings of the 32nd International Conference on Very Large Databases*, Seoul, Korea, pages 1163–1166. 2006.

[Joshi08] Dhiraj Joshi and Jiebo Luo. Inferring generic activities and events from image content and bags of geo-tags. In *CIVR '08: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, Niagara Falls, Canada, pages 37–46, 2008.

[Kang00] Hyunmo Kang and Ben Shneiderman. Visualization methods for personal photo collections: browsing and searching in the PhotoFinder. In *IEEE International Conference on Multimedia and Expo*, New York, New York, USA, volume 3, pages 1539–1542, 2000.

[Khella04] Amir Khella and Benjamin Bederson. Pocket PhotoMesa: a zoomable image browser for PDAs. In *MUM '04: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia*, College Park, Maryland, pages 19–24, 2004.

[Kim05] Suckchul Kim, Yoonsik Tak, Yunyoung Nam, and Eenjun Hwang. mCLOVER: mobile content-based leaf image retrieval system. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM International Conference on Multimedia*, Hilton, Singapore, pages 215–216, 2005.

[Kim06] Jeong Kim and John Zimmerman. Cherish: smart digital photo frames for sharing social narratives at home. In *CHI '06: CHI '06 Extended Abstracts on Human Factors in Computing Systems*, Montréal, Québec, Canada, pages 953–958, 2006.

[Kuchinsky99] Allan Kuchinsky, Celine Pering, Michael Creech, Dennis Freeze, Bill Serra, and Jacek Gwizdka. FotoFile: a consumer multimedia organization and retrieval system. In *CHI '99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, Pennsylvania, USA, pages 496–503, 1999.

[Kustanowitz05] Jack Kustanowitz and Ben Shneiderman. Motivating annotation for personal digital photo libraries: lowering barriers while raising incentives. Technical report, HCIL-2004-18, U. Maryland, 2005.

[Lamming92] Michael Lamming and William Newman. Activity-based information retrieval: Technology in support of personal memory. In *Proceedings of the IFIP 12th World Computer Congress on Personal Computers and Intelligent Systems - Information Processing '92 - Volume 3*, pages 68–81, 1992.

[Lansdale88] M. Lansdale. The psychology of personal information management. *Applied Ergonomics*, 19(1):55–66, Elsevier, March 1988.

[Lansdale89] Mark Lansdale, D. Young, and C. Bass. MEMOIRS: a personal multimedia information system. In *Proceedings of the fifth Conference of the British Computer Society, Human-Computer Interaction Specialist Group on People and Computers V*, Nottingham, England, pages 315–327, 1989.

[Lapedriza06] Agata Lapedriza, Manuel Maryn-Jimenez, and Jordi Vitria. Gender recognition in non controlled environments. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, Hong Kong, China, pages 834–837, 2006.

[Levasseur83] M. Levasseur and E. Veron. *Ethographie de l'exposition*. Bibliotheque publique d'Information, Centre Georges Pompidou, 1983.

[Lew06] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[Li03] Jia Li and James Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[Li03a] Beitao Li, Kingshy Goh, and Edward Chang. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *MULTIMEDIA '03: Proceedings of the eleventh ACM International Conference on Multimedia*, Berkeley, CA, USA, pages 195–206, 2003.

[Li06] Jia Li and James Wang. Real-time computerized annotation of pictures. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM International Conference on Multimedia*, Santa Barbara, California, USA, pages 911–920, 2006.

[Lieberman01] Henry Lieberman, Elizabeth Rozenweig, and Push Singh. Aria: An Agent for Annotating and Retrieving Images. *Computer*, 34(7):57–62, IEEE Computer Society, 2001.

[Likert32] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

[Lindley06] Siân Lindley and Andrew Monk. Designing appropriate affordances for electronic photo sharing media. In *CHI '06: CHI '06 Extended Abstracts on Human Factors in Computing Systems*, Montréal, Québec, Canada, pages 1031–1036, 2006.

[Lowe04] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal Computer Vision*, 60(2):91–110, Kluwer Academic Publishers, 2004.

[Lu03] Ye Lu, Hongjiang Zhang, Liu Wenyin, and Chunhui Hu. Joint semantics and feature based image retrieval using relevance feedback. *IEEE Transactions on Multimedia*, 5(3):339–347, Sept. 2003.

[Lucas81] Bruce Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, Vancouver, Canada, pages 674–679, April 1981.

[Luo06] Jiebo Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum - an overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine*, 23(2):101–114, Mar 2006.

[MLDCPort05] Microsoft Language Development Center em Portugal. http://www.microsoft.com/portugal/mldc/default.mspx, 2005, *último acesso 23/08/2009*.

[Ma97] Wei-Ying Ma and Bangalore Manjunath. NeTra: a toolbox for navigating large image databases. In *ICIP '97: Proceedings of the 1997 International Conference on Image Processing*, Washington, District of Columbia, USA, volume 1, pages 568–571, 1997.

[Magalhaes07] João Magalhães and Stefan Rüger. Information-theoretic semantic multimedia indexing. In *CIVR '07: Proceedings of the 6th ACM International Conference on Image and video retrieval*, Amsterdam, The Netherlands, pages 619–626, 2007.

[Mallows72] C. L. Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43(2):508–515, 1972.

[Malone83] Thomas Malone. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems*, 1(1):99–112, 1983.

[Manjunath96] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[Mann96] Steve Mann. "Smart Clothing": wearable multimedia and 'Personal Imaging' to restore the balance between people and their intelligent environments. In *Proceedings of the ACM Multimedia 96*, Boston, Massachusetts, USA, pages 163–174, Nov. 18-22 1996.

[Maron98] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *NIPS '97: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*, Denver, Colorado, United States, pages 570–576, 1998.

[Matkovic04] Kresimir Matkovic, Thomas Psik, Ina Wagner, and Werner Purgathofer. Tangible image query. In *Proceedings of Smart Graphics 2004*, Banff, Canada, pages 31–42. 2004.

[Mehrotra95] Rajiv Mehrotra and James E. Gary. Similar-shape retrieval In shape data management. *Computer*, 28(9):57–62, IEEE Computer Society Press, 1995.

[MemoryNet06] Global Memory Net homepage. http://www.memorynet.org, 2006, *último acesso 26/01/2009*.

[Mikolajczyk04] Krystian Mikolajczyk and Cordelia Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, Kluwer Academic Publishers, 2004.

[Moghaddam00] Baback Moghaddam and Ming-Hsuan Yang. Gender classification with Support Vector Machines. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, pages 306–611, 2000.

[Moghaddam04] Baback Moghaddam, Qi Tian, Neal Lesh, Chia Shen, and Thomas Huang. Visualization and user-modeling for browsing personal photo libraries. *International Journal of Computer Vision*, 56(1-2):109–130, Kluwer Academic Publishers, 2004.

[Monay03] Florent Monay and Daniel Gatica-Perez. On image auto-annotation with latent space models. In *MULTIMEDIA '03: Proceedings of the eleventh ACM International Conference on Multimedia*, Berkeley, CA, USA, pages 275–278, 2003.

[Mori99] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*, Orlando, Florida, USA, 1999.

[Mukherjea99] Sougata Mukherjea, Kyoji Hirata, and Yoshinori Hara. AMORE: A World Wide Web image retrieval engine. *World Wide Web*, 2(3):115–132, Kluwer Academic Publishers, 1999.

[Muller01] Klaus-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, Mar 2001.

[Mweb08] Memoria Web. http://di205.di.fct.unl.pt/instory-web, 2008, *último acesso 27/02/2009*.

[Naaman05] Mor Naaman, Ron Yeh, Hector Garcia-Molina, and Andreas Paepcke. Leveraging context to resolve identity in photo albums. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, Colorado, USA, pages 178–187, 2005.

[Nakazato01] Munehiro Nakazato and Thomas Huang. 3D MARS: immersive virtual reality for content-based image retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2001)*, Tokyo, Japan, pages 44–47, Aug. 2001.

[Naphade06] Milind Naphade, John Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[Nelson65] Ted Nelson. Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th National Conference*, Cleveland, Ohio, United States, pages 84–100, 1965.

[Nelson99] Ted Nelson. Xanalogical structure, needed now more than ever: parallel documents, deep links to content, deep versioning, and deep re-use. *ACM Computing Surveys*, 31(4es), 1999.

[Nicholas07] Nicholas Diakopoulos and Patrick Chiu. PhotoPlay: A collocated collaborative photo tagging game on a horizontal display. In *Proceedings addendum of User Interface Software Technology (UIST)*, Newport, Rhode Island, USA, pages 53–54, 2007.

[Nielsen90] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Seattle, Washington, United States, pages 249–256, 1990.

[Noda02] Makiko Noda, Hirotaka Sonobe, Saeko Takagi, and Fujiichi Yoshimoto. Cosmos: convenient image retrieval system of flowers for mobile computing situations. In *Proceedings of the IASTED International Conference on Information Systems and Databases (ISDB 2002)*, Tokyo, Japan, pages 25–30. 2002.

[Nowak06] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of the European Conference on Computer Vision*, Graz, Austria, volume 3954 of *Lecture Notes in Computer Science*, pages 490–503. Springer, 2006.

[Over06] Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan Smeaton. TRECVID 2006 - an overview. In *TRECVid 2006 - Text REtrieval Conference TRECVid Workshop*, 2006.

[Papadopoulos07] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis. Combining global and local information for knowledge-assisted image analysis and classification. *EURASIP Journal on Advances in Signal Processing*, 2007(2):18–18, Hindawi Publishing Corp., 2007.

[Patel06] Dynal Patel, Gary Marsden, Matt Jones, and Steve Jones. Improving photo searching interfaces for small-screen mobile computers. In *MobileHCI '06: Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*, Helsinki, Finland, pages 149–156, 2006.

[Pereira01] Fernando Pereira and Rob Koenen. MPEG-7: A standard for multimedia content description. *International Journal of Image and Graphics*, 1(3):527–546, World Scientific, 2001.

[Platt99] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, MIT Press, 1999.

[Poggio03] Tomaso Poggio and Steve Smale. The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society*, 50:537–544, 2003.

[Porkaew99] Kriengkrai Porkaew and Kaushik Chakrabarti. Query refinement for multimedia similarity retrieval in MARS. In *MULTIMEDIA '99: Proceedings of the seventh ACM International Conference on Multimedia (Part 1)*, Orlando, Florida, United States, pages 235–238, 1999.

[Pratt06] Wanda Pratt, Kenton Unruh, Andrea Civan, and Meredith M. Skeels. Personal health information management. *Communications of the ACM*, 49(1):51–55, 2006.

[Quack04] Till Quack, Ullrich Mönich, Lars Thiele, and B. S. Manjunath. Cortina: a system for large-scale, content-based web image retrieval. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, New York, NY, USA, pages 508–511, 2004.

[Rabiner90] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in Speech Recognition*, pages 267–296, Morgan Kaufmann Publishers Inc., 1990.

[Retrievr06] Retrievr. http://labs.systemone.at/retrievr, 2006, *último acesso 26/01/2009*.

[Riya05] Riya Visual Search. http://riya.com, 2005, *último acesso 26/01/2009*.

[Rocchio71] J. J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

[Rodden01] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? In *CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Seattle, Washington, United States, pages 190–197, 2001.

[Rodden03] Kerry Rodden and Kenneth Wood. How do people manage their digital photographs? In *CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Ft. Lauderdale, Florida, USA, pages 409–416, 2003.

[Rowe05] Lawrence Rowe and Ramesh Jain. ACM SIGMM retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(1):3–13, 2005.

[Rubner00] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, Kluwer Academic Publishers, 2000.

[Rubner97] Yossi Rubner, Leonidas Guibas, and Carlo Tomasi. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA Image Understanding Workshop*, New Orleans, Louisiana, USA, pages 661–668, 1997.

[Rui00] Yong Rui and Thomas Huang. Optimizing learning in image retrieval. In *Proceedings of the IEEE Confeference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, volume 1, pages 236–243, 2000.

[Rui97] Yong Rui, Thomas Huang, and Sharad Mehrotra. Content-based image retrieval with relevance feedback in MARS. In *Proceedings of the IEEE International Conference on Image Processing*, Washington, District of Columbia, USA, pages 815–818, 1997.

[Rui98] Yong Rui, Thomas Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[Russell08] Bryan Russell, Antonio Torralba, Kevin Murphy, and William Freeman. LabelMe: A aatabase and Web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, Kluwer Academic Publishers, 2008.

[Salton75] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[Salton86] Gerard Salton and Michael McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[Sarvas04] Risto Sarvas, Mikko Viikari, Juha Pesonen, and Hanno Nevanlinna. MobShare: controlled and immediate sharing of mobile images. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, New York, NY, USA, pages 724–731, 2004.

[Sclaroff94] S. Sclaroff, Rosalind Picard, and Alexander Pentland. Photobook: Tools for content-based manipulation of image databases. In *Proceedings of the Conference on Storage and Retrieval for Image and Video Database II, SPIE*, San Jose, California, USA, volume 2368, pages 37–50, 1994.

[Sebe02] Nicu Sebe and Michael Lew. Robust Shape Matching. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, London, England, pages 17–28, 2002.

[Sebe03] Nicu Sebe and Michael Lew. Comparing salient point detectors. *Pattern Recognition Letters*, 24(1):89–96, Elsevier Science, 2003.

[Shi00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, Aug 2000.

[Shneiderman00] Ben Shneiderman and Hyunmo Kang. direct annotation: a drag-and-drop strategy for labeling photos. In *Proceedings International Conference Information Visualisation (IV2000)*, London, England, pages 88–95. 2000.

[Sinnott84] R. Sinnott. Virtues of the Haversine. *Sky and Telescope*, 68:158, December 1984.

[Smeulders00] Arnold Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[Smith96] John Smith and Shih-Fu Chang. Automated binary texture feature sets for image retrieval. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, volume 4, pages 2239–2242, May 1996.

[Smith96a] John R. Smith and Shih-Fu Chang. VisualSEEk: a fully automated content-based image query system. In *MULTIMEDIA '96: Proceedings of the fourth ACM International Conference on Multimedia*, Boston, Massachusetts, USA, pages 87–98, 1996.

[Sonobe04] Hirotaka Sonobe, Saeko Takagi, and Fujiichi Yoshimoto. Mobile computing system for fish image retrieval. In *Proceedings of International Workshop on Advanced Image Technology (IWAIT)*, Singapore, pages 33–37, 2004.

[Srihari00] Rohini K. Srihari, Aibing Rao, Benjamin Han, Srikanth Munirathnam, and Xiaoyun Wu. A model for multimodal information retrieval. In *IEEE International Conference on Multimedia and Expo (II) 2000*, New York, New York, USA, pages 701–704. 2000.

[Stricker95] Markus Stricker and Markus Orengo. Similarity of color images. In W. Niblack and R. Jain, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 2420 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 381–392, March 1995.

[Sumi04] Yasuyuki Sumi, Sadanori Ito, Tetsuya Matsuguchi, Sidney Fels, and Kenji Mase. Collaborative capturing and interpretation of interactions. In *Proceedings of the Pervasive 2004 Workshop on Memory and Sharing of Experiences*, Vienna, Austria, pages 1–7, 2004.

[Surface07] Microsoft Surface. http://www.microsoft.com/surface, 2007, *último acesso 26/01/2009*.

[Swain91] Michael Swain and Dana Ballard. Color indexing. *International Journal Computer Vision*, 7(1):11–32, Kluwer Academic Publishers, 1991.

[Tamura78] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, June 1978.

[Tancharoen05] Datchakorn Tancharoen, Toshihiko Yamasaki, and Kiyoharu Aizawa. Practical experience recording and indexing of Life Log video. In *CARPE '05: Proceedings of the 2nd ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, Hilton, Singapore, pages 61–66, 2005.

[Tieu04] Kinh Tieu and Paul Viola. Boosting Image Retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, Kluwer Academic Publishers, 2004.

[Tokela08] Tero Jokela, Jaakko T. Lehikoinen, and Hannu Korhonen. Mobile multimedia presentation editor: enabling creation of audio-visual stories on mobile devices. In *CHI '08: Proceeding of the twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy, pages 63–72, 2008.

[Tong01] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *MULTIMEDIA '01: Proceedings of the ninth ACM International Conference on Multimedia*, Ottawa, Canada, pages 107–118. 2001.

[Toyama03] Kentaro Toyama, Ron Logan, and Asta Roseway. Geographic location tags on digital images. In *MULTIMEDIA '03: Proceedings of the eleventh ACM International Conference on Multimedia*, Berkeley, California, USA, pages 156–166, 2003.

[Trecvid05] TREC Video Retrieval Evaluation (TRECVID). http://www-nlpir.nist.gov/projects/tv2005/tv2005.html.

[Turk91] Matthew Turk and Alex Pentland. Face recognition using Eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, USA, pages 586–591, 1991.

[Tuulos07] Ville Tuulos, Jürgen Scheible, and Heli Nyholm. Combining Web, mobile phones and public displays in large-scale: Manhattan story Mashup. In *Proceedings of 5th International Conference on Pervasive Computing*, Toronto, Ontario, Canada, pages 37–54. 2007.

[Urban03] Jana Urban, Joemon M. Jose, and C. J. Van Rijsbergen. An adaptive approach towards content-based image retrieval. In *Proceedings International Workshop on Content-Based Multimedia Indexing (CBMI03)*, Rennes, France, pages 119–126, 2003.

[Vailaya01] Aditya Vailaya, Mário Figueiredo, Anil Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, Jan 2001.

[Vasconcelos00] Nuno Vasconcelos and Andrew Lippman. Learning from user feedback in image retrieval Systems. In T.K. Leen S.A. Solla and K.-R. Müller, editors, *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, Denver, Colorado, USA. 2000.

[Veltkamp00] Remco Veltkamp and Mirela Tanase. Content-based image retrieval systems: A survey. Technical report, UU-CS-2000-34, 2000.

[Vezhnevets03] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva. A survey on pixel-based skin color detection techniques. In *Proceedings of the GRAPHICON*, Moscow, Russia, pages 85–92, 2003.

[Viola04] Paul Viola and Michael Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, Kluwer Academic Publishers, 2004.

[VonAhn04] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria, pages 319–326, 2004.

[VonAhn06] Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, pages 55–64, 2006.

[VonAhn06a] Luis von Ahn. Games with a purpose. *IEEE Computer Magazine*, 39(6):92–94, 2006.

[Wahba99] Grace Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. *Advances in Kernel Methods: Support Vector Learning*, pages 69–88, MIT Press, 1999.

[Walter07] Andreas Walter and Gábor Nagypál. Imagenotion - methodology, tool support and evaluation. In *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS*, volume 4803 of *LNCS*, pages 1007–1024. Springer, 2007.

[Wang01] James Wang, Jia Li, and Gio Wiederhold. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, September 2001.

[Wang07] Shuo Wang, Feng Jing, Jibo He, Qixing Du, and Lei Zhang. IGroup: presenting web image search results in semantic clusters. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, USA, pages 587–596, 2007.

[Wang97] James Wang, Gio Wiederhold, Oscar Firschein, and Sha Wei. Content-based image indexing and searching using Daubechies' Wavelets. *International Journal on Digital Libraries*, 1(4):311–328, Springer Berlin / Heidelberg, 1997.

[Weber00] Markus Weber, Max Welling, and Pietro Perona. Unsupervised learning of models for recognition. In *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part I*, Dublin, Ireland, pages 18–32, 2000.

[Wenyin00] Liu Wenyin, Yanfeng Sun, and Hongjiang Zhang. MiAlbum - a system for home photo managemet using the semi-automatic image annotation approach. In *MULTIMEDIA '00: Proceedings of the eighth ACM International Conference on Multimedia*, Marina del Rey, California, United States, pages 479–480, 2000.

[Wenyin01] Liu Wenyin, Susan Dumais, Yanfeng Sun, Hongjiang Zhang, Mary Czerwinski, and Brent Field. Semi-automatic image annotation. In *INTERACT2001, 8th IFIP TC.13 Conference on Human-Computer Interaction*, Tokyo, Japan, pages 326–333. 2001.

[Whittaker06] Steve Whittaker, Victoria Bellotti, and Jacek Gwizdka. Email in personal information management. *Communications of the ACM*, 49(1):68–73, 2006.

[Wood98] M. Wood, B. Thomas, and N. Campbell. Iterative refinement by relevance feedback in content-based digital image retrieval. In *MULTIMEDIA '98: Proceedings of the sixth ACM International Conference on Multimedia*, Bristol, United Kingdom, pages 13–20, 1998.

[Wu01] P. Wu and B. Manjunath. Adaptive nearest neighbor search for relevance feedback in large image databases. In *MULTIMEDIA '01: Proceedings of the ninth ACM International Conference on Multimedia*, Ottawa, Canada, pages 89–97, 2001.

[Yan07] Rong Yan, Apostol Natsev, and Murray Campbell. An efficient manual image annotation approach based on tagging and browsing. In *MS '07: Workshop on Multimedia Information Retrieval on The Many Faces of Multimedia Semantics*, Augsburg, Bavaria, Germany, pages 13–20, 2007.

[Yang05] Changbo Yang, Ming Dong, and Farshad Fotouhi. Semantic feedback for interactive image retrieval. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM International Conference on Multimedia*, Singapore, pages 415–418, 2005.

[Yang06] Changbo Yang, Ming Dong, and Jing Hua. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, New York, USA, pages 2057–2063, 2006.

[Yates99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. Addison Wesley, May 1999.

[Yavlinsky05] Alexei Yavlinsky, Edward Schofield, and Stefan Rüger. Automated Image annotation using global features and robust nonparametric density estimation. In D. Polani, B. Browning, A. Bonarini, and K. Yoshida, editors, *Proceedings of the 4th International Conference on Image and Video Retrieval*, Singapore, Singapore, volume 3568 of *Lecture Notes in Computer Science (LNCS)*, pages 507–517, Springer 2005.

[Yeh05] Tom Yeh, , Kristen Grauman, Konrad Tollmar, and Trevor Darrell. A picture is worth a thousand keywords: image-based object search on a mobile platform. In *CHI '05: CHI '05 Extended Abstracts on Human Factors in Computing Systems*, Portland, Oregon, USA, pages 2025–2028, 2005.

[Youtube05] YouTube. http://www.youtube.com, 2005, *último acesso 27/02/2009*.

[Yu04] Yuh-Shyan Chen Gwo-Jong Yu and Kuei-Ping Shih. A content-based image Retrieval system for outdoor ecology learning: a firefly watching system. In *AINA '04: Proceedings of the 18th International Conference on Advanced Information Networking and Applications*, page 112, 2004.

[Yu07] Ning Yu, Khanh Vu, and Kien A. Hua. An in-memory relevance feedback technique for high-performance image retrieval Systems. In *CIVR '07: Proceedings of the 6th ACM International Conference on Image and video retrieval*, Amsterdam, The Netherlands, pages 9–16, 2007.

[Zhang00] Hongiang Zhang, Liu Wenyin, and Chunhul Hu. iFind - a system for semantics and feature based image retrieval over Internet. In *MULTIMEDIA '00: Proceedings of the eighth ACM International Conference on Multimedia*, Marina del Rey, California, USA, pages 477–478, 2000.

[Zhang01] Lei Zhang, Fuzong Lin, and Bo Zhang. Support vector machine learning for image retrieval. In *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, volume 2, pages 721–724 vol.2, Oct 2001.

[Zhou02] Xiang Zhou and Thomas Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 9(2):23–33, Apr-Jun 2002.

[Zhou03] Xiang Zhou and Thomas Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, Springer Berlin Heidelberg, April 2003.