Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Contribuições para a Localização e Mapeamento em Robótica através da Identificação Visual de Lugares

Francisco Mateus Marnoto de Oliveira Campos

Doutoramento em Informática, especialidade de Engenharia Informática.

2015

Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Contribuições para a Localização e Mapeamento em Robótica através da Identificação Visual de Lugares

Francisco Mateus Marnoto de Oliveira Campos

Tese orientada pelo Prof. Doutor Luís Miguel Parreira e Correia e co-orientada pelo Prof. Doutor João Manuel Ferreira Calado, especialmente elaborada para a obtenção do grau de doutor em Informática, especialidade de Engenharia Informática.

2015

Resumo

Em robótica móvel, os métodos baseados na aparência visual constituem uma abordagem atractiva para o tratamento dos problemas da localização e mapeamento. Contudo, para o seu sucesso é fundamental o uso de características visuais suficientemente discriminativas. Esta é uma condição necessária para assegurar o reconhecimento de lugares na presença de factores inibidores, tais como a semelhança entre lugares ou as variações de luminosidade.

Esta tese debruça-se sobre os problemas de localização e mapeamento, tendo como objectivo transversal a obtenção de representações mais discriminativas ou com menores custos computacionais. Em termos gerais, dois tipos de características visuais são usadas, as características locais e globais.

A aplicação de características locais na descrição da aparência tem sido dominada pelo modelo BoW (*Bag-of-Words*), segundo o qual os descritores são quantizados e substituídos por palavras visuais. Nesta tese questiona-se esta opção através do estudo da abordagem alternativa, a representação não-quantizada (NQ). Em resultado deste estudo, contribui-se com um novo método para a localização global de robôs móveis, o classificador NQ. Este, para além de apresentar maior precisão do que o modelo BoW, admite simplificações importantes que o tornam competitivo, também em termos de eficiência, com a representação quantizada.

Nesta tese é também estudado o problema anterior à localização, o da extracção de um mapa do ambiente, sendo focada, em particular, a detecção da revisitação de lugares. Para o tratamento deste problema é proposta uma nova característica global, designada LBP-Gist, que combina a análise de texturas pelo método LBP com a codificação da estrutura global da imagem, inerente à característica Gist. A avaliação deste método em vários *datasets* demonstra a viabilidade do detector proposto, o qual apresenta precisão e eficiência superiores ao *state-of–the-art* em ambientes de exterior.

Palavras chave: localização por visão, métodos baseados na aparência, características visuais locais, características visuais globais, detecção de revisitação.

Abstract

In the mobile robotics field, appearance-based methods are at the core of several attractive systems for localization and mapping. To be successful, however, these methods require features having good descriptive power. This is a necessary condition to ensure place recognition in the presence of disturbing factors, such as high similarity between places or lighting variations.

This thesis addresses the localization and mapping problems, globally seeking representations which are more discriminative or more efficient. To this end, two broad types of visual features are used, local and global features.

Appearance representations based on local features have been dominated by the BoW (Bag of Words) model, which prescribes the quantization of descriptors and their labelling with visual words. In this thesis, this method is challenged through the study of the alternative approach, the non-quantized representation (NQ). As an outcome of this study, we contribute with a novel global localization method, the NQ classifier. Besides offering higher precision than the BoW model, this classifier is susceptible of significant simplifications, through which it is made competitive to the quantized representation in terms of efficiency.

This thesis also addresses the problem posed prior to localization, the mapping of the environment, focusing specifically on the loop closure detection task. To support loop closing, a new global feature, LBP-Gist, is proposed. As the name suggests, this feature combines texture analysis, provided by the LBP method, with the encoding of global image structure, underlying the Gist feature. Evaluation on several datasets demonstrates the validity of the proposed detector. Concretely, precision and efficiency of the method are shown to be superior to the state-of-the-art in outdoor environments.

Keywords: visual localization, appearance-based methods, local image features, global image features, loop closure detection.

Agradecimentos

Um trabalho que, como este, decorre ao longo de vários anos é, não só resultado de um esforço pessoal, mas também da ajuda de várias pessoas que, directa ou indirectamente, contribuiram para ele.

Em primeiro lugar, agradeço aos meus orientadores, Prof. Luís Correia e Prof. João Calado, pelas sugestões importantes, apoio e encorajamento dados durante todo o trabalho que conduziu a esta tese.

Agradeço também ao Prof. Carlos Cardeira, pela disponibilização do robô móvel que foi usado no ISEL. O Pedro Santana, colega do curso de Doutoramento, teve um papel essencial, pelas interessantes discussões que tivemos e que se reflectiram no meu trabalho. Existe ainda uma dívida permanente para com os meus colegas de trabalho, Fernando Carreira, Mário Mendes e Pedro Silva, pela cooperação sem limites dentro da Secção de Controlo de Sistemas.

Índice

Lista de Acrónimos

AGV	*Automated Guided Vehicle*
BoW	*Bag of Words*
BRIEF	*Binary Robust Independent Elementary Features*
BRISK	*Binary Robust Invariant Scalable Keypoints*
CIE	*Commission Internationale de l'Eclairage*
DoG	*Difference of Gaussians*
DoH	*Determinant of Hessian*
E2LSH	*Exact Euclidean Locality Sensitive Hashing*
FAB-Map	*Fast Appearance Based Mapping*
FAST	*Features from Accelerated Segment Test*
FD	*Feature Deletion*
FM	*Feature Mearging*
FREAK	*Fast Retina Keypoint*
GPS	*Global Positioning System*
GS	*Gist Selection*
HMM	*Hidden Markov Model*
HSL	*Hue-Saturation-Lightness*
HSV	*Hue-Saturation-Value*
ICRA	*International Conference on Robotics and Automation*
idf	*Inverse Document Frequency*
IDOL	*Image Database for rObot Localization*
IE	*Inverse Entropy*
IEAT	*Inverse Entropy with Average Threshold*
LBP	*Local Binary Patterns*

LDB	*Local Difference Binary*
LoG	*Laplacian of Gaussian*
LSH	*Locality Sensitive Hashing*
MPEG	*Moving Picture Experts Group*
MSER	*Maximally Stable Extremal Region*
NBNN	*Naive Bayes Nearest Neighbour*
NE	*Negative Entropy*
NEAT	*Negative Entropy with Average Threshold*
ORB	*Oriented Fast and Rotated BRIEF*
PCA	*Principal Component Analysis*
PS	*Progressive Selection*
RBF	*Radial Basis Function*
RGB	*Red-Green-Blue*
RST	*Relative Similarity Threshold*
SIFT	*Scale Invariant Feature Detector*
SURF	*Speed-Up Robust Feature*
SVM	*Support Vector Machine*
tf-idf	*Term Frequency - Inverse Document Frequency*
WTA	*Winner Take All*
YCbCr	*Y(Luminance)-Chroma Blue-Chroma Red*

Lista de Figuras

Lista de Tabelas

1. Introdução

Esta tese lida com problemas da robótica móvel, a disciplina em que confluem os conhecimentos de áreas como a electrónica, mecânica, computação, inteligência artificial e controlo, com o objectivo de desenvolver plataformas móveis autónomas ou semi-autónomas. Desde os primeiros trabalhos sobre o robô Shakey (Nilsson, 1984), o primeiro a recorrer a métodos de inteligência artificial, em 1966, foram realizados progressos assinaláveis, contribuindo para um grau de maturidade que torna hoje possível a utilização de robôs móveis na Indústria, Agricultura e Serviços, bem como a sua aplicação em missões espaciais, de salvamento em zonas de desastre, de vigilância e de desminagem, entre outras.

Apesar destes sucessos, a actividade de investigação dedicada à robótica móvel continua ainda hoje com grande vitalidade. Tal deve-se, em parte, às inovações introduzidas em áreas complementares, absorvidas pela robótica móvel com vista a gerar novas funcionalidades, maior eficiência e autonomia, ou menor custo. Um exemplo recente dessa tendência é dado pela introdução de câmaras de profundidade, que podem vir a substituir os *scanners* laser, mais dispendiosos. A segunda motivação que continua a impulsionar os esforços de investigação é, naturalmente, a existência de ambições por cumprir, associadas por vezes a novas aplicações que não eram perspectivadas há algumas décadas atrás. Hoje, por exemplo, começa a admitir-se a viabilidade de veículos autónomos circularem nas estradas, algo que só se tornou possível com o grau de sofisticação que a robótica móvel atingiu. Entre as ambições mais antigas, a ideia de autonomia, embora genérica, define os objectivos que têm guiado uma grande parte dos trabalhos de investigação. Essencialmente, entende-se por autonomia a redução ao mínimo da intervenção humana que é necessária para garantir a operacionalidade do robô. Enquanto algumas aplicações tradicionais de robôs móveis exibiam uma autonomia reduzida – os AGVs (*Autonomous Guided Vehicle*), por exemplo, têm os seus movimentos pré-definidos pelos percursos fixos marcados no chão – actualmente o objectivo é muitas vezes dotá-los das capacidades cognitivas necessárias para lidar com ambientes pouco estruturados e dinâmicos.

Independentemente do seu grau de sofisticação, a faculdade mínima exigida num robô móvel é a capacidade de navegação, isto é, a capacidade de se deslocar no espaço por forma a cumprir uma missão. Na classe de soluções que aspiram a uma maior

versatilidade, autonomia e eficiência, a navegação é baseada em representações do ambiente, i.e., mapas. Segundo Levitt e Lawton (1990) a navegação que segue este paradigma pode ser decomposta em 3 subproblemas:

1. Construção de um mapa do ambiente
2. Localização
3. Planeamento de acções.

Genericamente, esta tese trata o reconhecimento de lugares, o qual é fundamental nos dois primeiros subproblemas mencionados acima. No primeiro, o robô deve construir um mapa do ambiente, através de informação recolhida durante uma fase de exploração. No segundo, o robô usa informação recolhida recentemente para se localizar sobre o mapa existente. Em ambos os casos a informação relevante provém de duas origens: dos sinais recolhidos pelos sensores e dos dados sobre os movimentos realizados pelo próprio robô. Se o caminho para a autonomia completa passa pela restrição a estas fontes de informação, este não é um objectivo trivial, dada a incerteza que lhes está associada. Em princípio, é possível inferir a posição de um robô através dos seus movimentos elementares, no entanto, este processo gera desvios progressivamente maiores relativamente à posição correcta, pois os erros individuais sobre cada movimento são acumulados ao longo do tempo. Estes erros podem ser reduzidos recorrendo à informação de sensores, mas também aqui existem dois factores de ambiguidade, que dificultam a tarefa. Em primeiro lugar temos a dinâmica do ambiente, devido à qual um lugar pode apresentar-se ao robô sob diferentes aspectos, em diferentes momentos. Em segundo lugar temos a ambiguidade que está inerente a ambientes com alguma homogeneidade. No limite imagine-se, por exemplo, a condução sobre uma autoestrada no deserto: aqui, diferentes lugares têm provavelmente a mesma aparência, tornando a localização baseada apenas nas leituras por sensores extremamente difícil.

Actualmente existe um tipo de sensor que, por fornecer a posição métrica absoluta, não está sujeito aos inconvenientes das abordagens anteriores: o GPS (*Global Positioning System*). Este sistema apresenta vários atractivos, como o seu custo reduzido e ausência de erros cumulativos, tendo, por isso, sido usado em alguns estudos como auxiliar à navegação de robôs (Agrawal e Konolige, 2006; Schleicher et al., 2009). No entanto, algumas características deste sensor impedem que ele substitua completamente as abordagens tradicionais. Em ambientes de interior o sinal GPS não

está disponível, inviabilizando a sua utilização neste cenário. Além disso, mesmo em ambientes de exterior o sinal pode ser interrompido, devido à presença de obstáculos como edifícios, árvores ou outras estruturas elevadas (Cummins e Newman, 2008; Badino, Huber e Kanade, 2012).

Na impossibilidade de se recorrer ao sinal GPS em todas as circunstâncias, o uso de sensores internos torna-se essencial. Neste sentido, as câmaras digitais são uma alternativa atractiva, dado o seu baixo custo e reduzida dimensão, e por darem acesso a um sinal rico em informação. Nesta tese é abordado o reconhecimento de lugares por visão, mais especificamente, sob o ponto de vista da aparência dos lugares. Os métodos baseados na aparência são específicos, no contexto da localização por visão, por não exigirem a detecção de marcos de navegação pré-definidos, nem recorrerem à representação, sobre um mapa, dos objectos do ambiente. Em vez disso, procura-se tirar partido da riqueza do sinal visual, no sentido de construir representações que sejam suficientemente discriminativas para suportar o reconhecimento de lugares. Nesta linha de pensamento, o reconhecimento de lugares passa, muitas vezes, pela simples comparação de imagens, uma operação que é comum a outros problemas da visão computacional, como o da pesquisa de imagens por conteúdo (Datta et al., 2008). Não é de admirar, por isso, que muitas das técnicas usadas nos métodos baseados na aparência sejam partilhadas nesse domínio, o qual é ocasionalmente referido nesta tese.

Este capítulo apresenta material introdutório que enquadra os desenvolvimentos posteriores, começando por descrever os principais paradigmas de modelação do ambiente e por contextualizar os mapas de aparência, na secção 1.1. Posteriormente são apresentadas as características visuais através das quais se obtêm representações compactas das imagens, na secção 1.2. A secção 1.3 descreve as formas como essas características visuais têm sido usadas na representação da aparência em robótica móvel; a secção 1.4 aborda os problemas da localização e mapeamento, introduzindo também a abordagem probabilística que permite lidar com as incertezas em robótica. Em 1.5 e 1.6 apresentam-se os aspectos motivadores e as contribuições desta tese, e, por fim, a secção 1.7 resume a estrutura da tese.

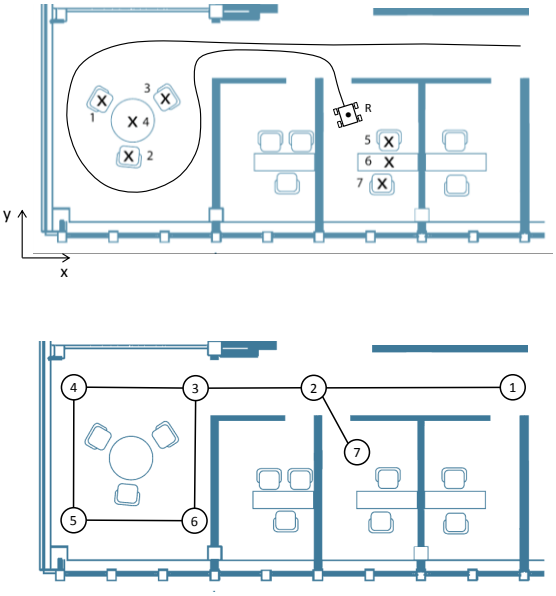


Figura 1.1. Principais paradigmas de representação do ambiente. Em cima: mapa métrico. Note-se que os objectos identificados são apenas ilustrativos, pois, tipicamente, os mapas representam estruturas de mais baixo nível, como arestas e cantos. Em baixo: mapa topológico. Aqui, a posição dos nós sobre a planta é também ilustrativa, pois estes mapas podem não incluir informação métrica.

1.1 Paradigmas de representação do ambiente

Tradicionalmente, as representações do ambiente têm sido distinguidas entre as que usam mapas métricos e as que recorrem a mapas topológicos. No primeiro caso as posições do robô e dos objectos do ambiente são representadas num referencial cartesiano, comum e fixo (Figura 1.1). Contrariamente a esta abordagem, que é puramente geométrica, os mapas topológicos incluem informação qualitativa, expressa na forma de um grafo (Figura 1.1). Neste grafo, os nós correspondem a lugares no ambiente e as ligações representam a acessibilidade entre eles.

Cada uma destas representações apresenta vantagens e inconvenientes relativamente à outra. Os mapas métricos são construídos através da fusão de informação interna (odometria) com informação sobre o ambiente, obtida por sensores, a qual é transferida para o referencial cartesiano através de modelos dos sensores. Nestas representações é dada uma forte ênfase à precisão das estimativas, a qual é desejável para a posterior utilização do mapa, mas, sobretudo, é essencial para manter a consistência geométrica durante a sua construção. Desta condição decorrem as principais dificuldades na extracção de mapas métricos, já que as estimativas baseadas na odometria estão sujeitas a erros cumulativos, e os modelos dos sensores não garantem precisão suficiente. Nas formas de mapeamento mais comuns a atenuação

dos erros acumulados ocorre naturalmente, quando os objectos do ambiente são observados em momentos distintos. No entanto, esse processo envolve custos computacionais elevados, além de depender de forma crucial da identificação correcta dos objectos que são re-observados.

Os mapas topológicos aliviam as dificuldades anteriores por substituírem a necessidade de consistência geométrica pela consistência topológica. Aqui, um mapa é considerado correcto desde que a sua topologia corresponda à da estrutura do ambiente. Além disso, a discretização do espaço num número finito de lugares, associada ao facto de estes mapas dispensarem a representação dos objectos do ambiente, torna-os mais aptos a tratar ambientes de elevada dimensão. Esta discretização tem, no entanto, justificado o argumento de que a representação topológica é menos precisa, pois, ao contrário da representação métrica, não realiza estimativas num espaço contínuo. Por forma a superar esta limitação, muitos sistemas têm complementado a informação topológica com dados métricos, resultando em mapas que se convencionou designar por híbridos. Dentro desta categoria encontra-se uma gama de soluções que podem simplesmente registar a distância e orientação das ligações entre nós (Simmons e Koenig, 1995; Kunz, Willeke e Nourbakhsh, 1997; Shatkay e Kaelbling, 2002) ou recorrer a estruturas mais complexas, como descrito por Tomatis, Nourbakhsh e Siegwart (2003). Neste caso, a informação métrica é colocada sobre um conjunto de referenciais cartesianos independentes, dispostos sobre o ambiente para auxiliar a navegação, de forma precisa, em cada uma das regiões de influência.

Alguns trabalhos têm salientado que a maior simplicidade dos mapas topológicos decorre de usarem representações centradas no robô, em lugar de centradas nos objectos do ambiente. Esta ideia está na base de vários estudos que defendem ser possível fazer o reconhecimento de lugares sem representar explicitamente os objectos, registando, em vez disso, a experiência sensorial do robô nesses lugares. Nos mapas resultantes, cada lugar tem associada uma assinatura sensorial, que resume aquela experiência e que deve ser tão discriminativa quanto possível. Em (Tapus, Tomatis e Siegwart, 2006), por exemplo, os autores sumarizam a informação extraída de um *scanner* laser e de imagens omni-direccionais numa sequência de caracteres a que chamaram *fingerprint*. Quando a principal modalidade sensorial é a visão, os mapas resultantes são normalmente designados por *mapas de aparências*, pois as

assinaturas sensoriais são um resumo da aparência das imagens (Jones, Andresen e Crowley, 1997; Ulrich e Nourbakhsh, 2000; Booij et al., 2007).

Os métodos baseados na aparência tiveram um primeiro fôlego durante a década de 90, por influência de duas áreas de investigação contemporâneas: o reconhecimento visual de objectos e a navegação de robôs baseada em memórias. No primeiro caso recorreu-se à comparação de medidas globais da imagem para a identificação de classes de objectos. Alguns dos trabalhos precursores nesta área dedicaram-se ao reconhecimento facial (Turk e Pentland, 1991) e ao reconhecimento de objectos sujeitos a variações de iluminação e postura (Murase e Nayar, 1995). A técnica comum a estes trabalhos, e que foi também adoptada na localização baseada na aparência, consiste na aplicação de PCA (*Principal Component Analysis*) na caracterização do espaço das imagens.

Os sistemas de navegação baseados em memórias assentam no princípio simples de que comportamentos úteis à navegação podem ser obtidos definindo as respostas a gerar para cada padrão sensorial, em lugar de serem definidas leis abstractas e mais genéricas para essas respostas. Estas soluções encontram um paralelo na área da robótica baseada em comportamentos, onde o controlo se baseia em acções reactivas. A navegação baseada em memórias parte do mesmo tipo de tratamento de baixo nível das percepções, mas acentua o facto de os sinais sensoriais serem comparados com os que se encontram guardados num banco de memórias, construído previamente e onde se encontra também a resposta associada. Naturalmente, quanto mais complexo for o tipo de comportamento procurado, ou a variedade de ambientes e de reacções a tratar, maior será a dimensão desse banco de memórias. No entanto, segundo alguns autores, uma das inspirações deste tipo de soluções é o exemplo dos insectos que, apesar de possuírem um sistema neuronal mínimo, apresentam uma capacidade de se auto-localizarem muito robusta (Jogan e Leonardis, 2003).

Já nos anos 2000, os métodos baseados na aparência tiveram um novo ímpeto, sobretudo devido à introdução das técnicas de análise de imagens baseadas em características locais. Os métodos desenvolvidos neste período serão abordados em maior detalhe nas secções 1.2 e 1.3.

1.2 Características visuais

As características visuais podem ser amplamente divididas entre as características locais e globais, cada uma delas reflectindo preocupações distintas na análise de imagens. Com as primeiras pretende-se obter uma descrição da imagem que seja robusta às transformações mais comuns, de perspectiva, iluminação, oclusão, etc. Para fazer face a essas variações, a abordagem por características locais centra-se na descrição de regiões de interesse da imagem, as quais são seleccionadas independentemente e em função do seu padrão visual. Em si, cada uma das características extraídas oferece alguma invariância a rotações, mudança de escala, etc., promovendo assim a robustez da representação que se obtém do *conjunto* de características.

As características globais têm como principal atractivo a sua simplicidade e eficiência, já que dispensam a segmentação da imagem em regiões de interesse. Apesar dessa simplicidade, o tratamento uniforme do espaço da imagem tem produzido características com bom poder descritivo que, aliado à sua eficiência, justificou o uso em problemas que envolvem bases de imagens de elevada dimensão (Torralba, Fergus e Weiss, 2008; Deselaers, Keysers e Ney, 2008; Douze et al., 2009).

De seguida são introduzidos os conceitos e soluções que foram desenvolvidos, no seio da comunidade de Visão Computacional, para tratar cada uma das categorias de características. No contexto das características locais será ainda dado destaque à chamada representação BoW (*Bag of Words*), pelo papel essencial que tem na construção de representações mais compactas das imagens. Mais à frente, a secção 1.3 referir-se-á à integração destes conceitos em representações da aparência que são úteis no domínio da robótica móvel.

1.2.1 Características locais

A primeira etapa na extracção de características locais é a da detecção de regiões de interesse da imagem. A literatura relacionada é muito rica em algoritmos que cumprem este objectivo, variando nas ferramentas algébricas usadas e nos critérios que orientam a selecção das regiões. De um modo geral, todos estes detectores têm por objectivo produzir regiões ricas em conteúdo, por um lado, e além disso

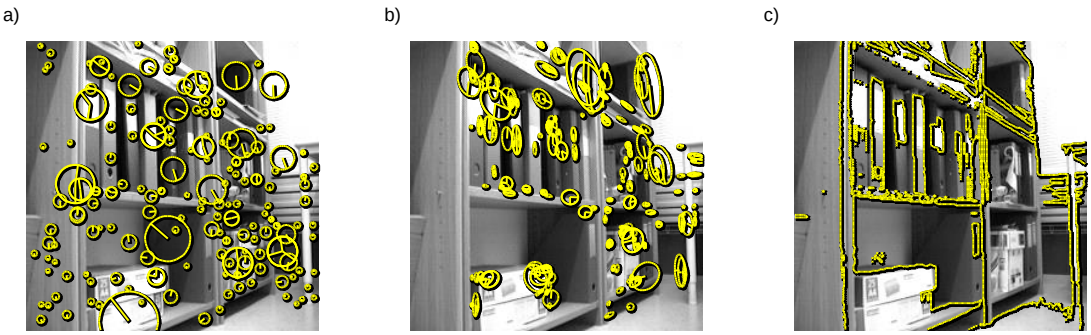


Figura 1.2. Regiões de interesse identificadas por diferentes detectores. A) DoG, b) Hessian-Affine e c) MSER.

bem definidas e estáveis perante transformações geométricas e de iluminação. O objectivo subjacente às últimas condições é o de que, perante duas imagens do mesmo objecto ou cena, sejam detectadas as mesmas regiões de interesse.

Ao longo das últimas décadas, a evolução registada nestes algoritmos tem sido motivada por duas preocupações principais: a de desenvolver detectores mais rápidos e a de acrescentar dimensões de invariância. Uma classe importante de detectores oferece invariância à escala, obtida pela aplicação de filtros gaussianos com múltiplas aberturas – à representação resultante dá-se a designação de *espaço de escalas*. Com base nesta representação foram desenvolvidos os detectores DoG, LoG e DoH que recorrem aos conceitos de Diferença de Gaussianas (Lowe, 1999), Laplaciano de Gaussianas e Determinante da Hessiana (Lindeberg, 1998), respectivamente. Todos estes detectores incorporam, para além da invariância à escala, invariância à translação, pois envolvem uma pesquisa completa no plano da imagem, e invariância à rotação, dado que os filtros aplicados são simétricos. Na Figura 1.2.a mostram-se exemplos de regiões identificadas pelo detector DoG.

Por forma a melhorar a robustez perante transformações de perspectiva, Mikolajczyk e Schmid (2002) propuseram dois detectores que incluem também invariância a transformações afins, o Harris-Affine e o Hessian-Affine, ilustrado na Figura 1.2.b. Com o mesmo objectivo, Matas et al. (2004) desenvolveram o detector MSER (*Maximally Stable Extremal Regions*), que se destaca dos anteriores por não tratar a imagem na representação espaço-escala. Em vez disso, as imagens são analisadas no domínio dos *thresholds* de binarização, sendo escolhidas as regiões que apresentam maior estabilidade à variação daquele parâmetro (ver Figura 1.2.c).

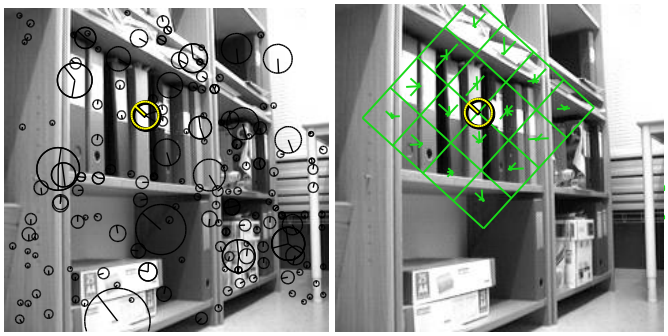


Figura 1.3. Descritor SIFT. À direita: visualização do descritor calculado sobre a região de interesse realçada na figura à esquerda.

A segunda etapa na extracção de características locais é a da codificação das regiões seleccionadas em descritores que sejam suficientemente representativos e, ao mesmo tempo, compactos. Um dos primeiros métodos a ser proposto para este efeito, e o que se mantém mais popular, foi desenvolvido por Lowe (1999), que o denominou SIFT (*Scale Invariante Feature Transform*). Este descritor, tal como outros que se seguiram, capta as texturas na região de suporte, recorrendo, para isso, aos gradientes medidos nessa região. Por forma a descrever a distribuição espacial dos gradientes, a região de suporte é dividida numa grelha de 4×4 e para cada divisão é calculado o histograma dos gradientes medidos em 8 orientações (ver Figura 1.3). Na construção destes histogramas, os gradientes são ainda modulados por uma função gaussiana, colocada no centro da região e com desvio-padrão igual a metade da sua extensão. Através deste procedimento, dá-se maior peso aos píxeis centrais, os quais são menos sensíveis a deformações da imagem. O descritor SIFT é um vector de dimensão 128 que resulta da concatenação dos 4×4 histogramas de comprimento 8. Por fim, este vector é sujeito a uma operação de normalização relativa à luminosidade. A característica SIFT está associada ao detector DoG, o qual selecciona as regiões de interesse, mas não fornece a orientação do padrão a descrever. A invariância à rotação envolve uma etapa anterior ao cálculo do descritor, na qual é determinada a orientação dominante na região de suporte. Este ângulo é usado para orientar a grelha aplicada à região e, por isso, o descritor extraído é independente da orientação particular do padrão visual.

Posteriormente à introdução da característica SIFT, foram investigadas outras técnicas, tendo em vista descritores mais compactos. Contudo, num amplo estudo em que Mikolajczyk e Schmid (2005) comparam um conjunto de descritores sob

condições de transformações geométricas e fotométricas, e em diversos *datasets*, o descritor SIFT apresentou melhor desempenho global. Mais tarde, Bay et al. (2008) desenvolveram a característica SURF, que tem recebido especial atenção por ser de extracção mais rápida do que os anteriores. Tal como o SIFT, a característica SURF inclui um detector associado, que consiste numa simplificação do detector DoH. A rapidez de execução resulta da utilização de imagens integrais, quer no cálculo daquela aproximação, quer na extracção do descritor, baseada em filtros Haar (Viola e Jones, 2001). A comparação dos descritores SIFT e SURF foi realizada em vários estudos que não produziram conclusões unânimes, o que pode ser explicado pela utilização de diferentes versões dos descritores e pela avaliação em diferentes problemas (Bauer, Sunderhauf e Protzel, 2007; Valgren e Lilienthal, 2010). De um modo geral, tem sido apontado que eles são semelhantes em termos de precisão, apresentando cada um deles maior robustez relativamente a perturbações específicas da imagem (Khan, McCane e Wyvill, 2011).

Devido à importância que as características locais têm ganho na Visão Computacional e à necessidade de reduzir os custos computacionais, a investigação sobre novas características continua muito activa. Os exemplos mais notáveis de esforços recentes encontram-se nas características BRIEF (Calonder et al., 2012), ORB (Rublee et al., 2011) e BRISK (Leutenegger, Chli e Siegwart, 2011), tendo em comum a utilização de descritores binários que, para além de serem de extracção rápida, facilitam a comparação através da distância Hamming. A fase de detecção foi também abordada nos casos do ORB e BRISK, os quais foram associados ao detector FAST (Rosten e Drummond, 2006), com modificações específicas que visaram produzir invariância à rotação e escala, respectivamente. Um estudo comparativo de características recentes demonstrou que estes descritores binários apresentam precisão semelhante às do SIFT e do SURF, ao mesmo tempo que oferecem maior rapidez de execução (Miksik e Mikolajczyk, 2012).

Modelo BoW

Os sistemas que usam características locais encontram um obstáculo na pesquisa em bases de dados de grandes dimensões, devido à necessidade de comparar todos os pares constituídos pelas características da imagem de teste e das imagens da base de dados. Sivic e Zisserman (2003) debruçaram-se sobre este problema e propuseram um modelo, *Bag of Words* (BoW), através do qual um conjunto de características pode

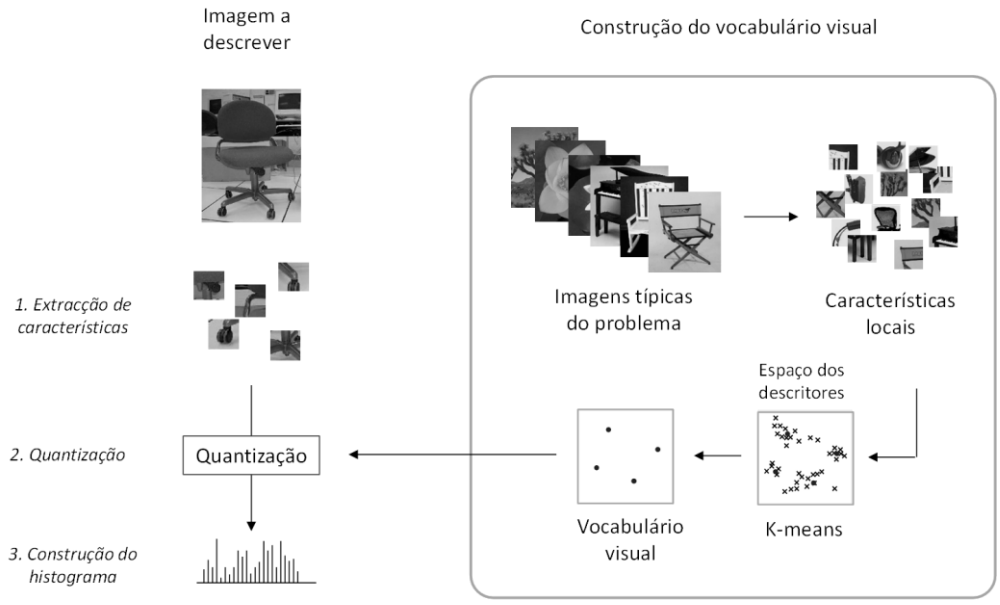


Figura 1.4. Resumo das etapas envolvidas na representação de uma imagem sob o modelo BoW.

ser transformado numa representação compacta da imagem. A designação *Bag of Words* tem origem no domínio da pesquisa de textos, referindo-se a um modelo em que estes são encarados como conjuntos não ordenados de palavras e representados pelos histogramas da frequência de palavras. Na construção destes vectores, cada texto é sujeito a um pré-processamento, em que palavras pouco discriminativas como 'e' 'o' e 'de' são eliminadas, e palavras derivadas são substituídas pela palavra primitiva. Por exemplo, palavras como 'utilização' e 'utilizador' são substituídas pelo seu tronco comum, 'utilizar'. Já nesta forma, é calculada a frequência de cada palavra e estes valores são reunidos num vector que constitui um descritor do documento.

Na adaptação destas ideias à visão computacional, Sivic e Zisserman (2003) fizeram equivaler as características locais às palavras de um texto, e encararam uma imagem como um conjunto não ordenado destas características. Uma dificuldade encontrada na adaptação do modelo BoW reside na definição do conceito de vocabulário de palavras visuais, já que não existe uma correspondência imediata com o de palavras numa linguagem natural. Por exemplo, no domínio das características visuais não existem regras para identificar aquelas que não são discriminativas, nem para associar características semelhantes a uma raiz comum. A solução proposta por Sivic e Zisserman passa por construir um vocabulário visual a partir das características extraídas de um conjunto representativo de imagens (ver Figura 1.4). Nesse processo, os descritores encontrados são sujeitos a aglomeração, pelo algoritmo *k-means*, que

devolve um número pré-definido de centróides. Estes centróides constituem o vocabulário visual e têm o papel de ícones, representativos das características visuais que lhes são próximas. Dispondo destes vectores, qualquer imagem pode ser representada no mesmo vocabulário, através dos seguintes passos (Figura 1.4): i) extracção das características locais, ii) quantização das características para as palavras visuais mais próximas iii) construção de um histograma da ocorrência das palavras visuais.

1.2.2 Características globais

Desde cedo que a abordagem por características globais explorou as diferentes dimensões de conteúdo visual: cor e textura. O trabalho de Swain e Ballard (1991) constituiu um marco na utilização da cor, ao mostrar que é possível fazer o reconhecimento de objectos através de características globais de cor, mesmo na presença de mudanças de perspectiva e de oclusão parcial. Essencialmente, naquele artigo é proposta i) a utilização de histogramas tridimensionais de cor e ii) a sua comparação através de uma medida designada por intersecção de histogramas. Na construção dos histogramas, píxeis no espaço de cores RGB foram em primeiro lugar transformados para um espaço de cores oponentes, com eixos branco-preto, vermelho-verde e azul-amarelo discretizados em 8, 16 e 16 intervalos, respectivamente. Definida esta grelha tridimensional, o cálculo do descritor passou simplesmente pela contagem do número de píxeis com coordenadas de cor dentro de cada volume elementar.

O método proposto por Swain e Ballard inspirou inúmeros trabalhos posteriores onde esteve, muitas vezes, subjacente a procura de espaços de cores mais robustos. Em alternativa ao espaço RGB, encontram-se na literatura referências aos espaços HSV, HSL, CIELAB e CIELUV, entre outros, caracterizando-se por apresentarem maior *uniformidade perceptiva*. Esta propriedade, aplicada a um espaço de cor, indica que variações na mesma quantidade sobre as suas coordenadas são percepcionadas de forma semelhante pelo sistema visual humano. Como notaram Smeulders et al. (2000), o interesse sobre estes espaços de cor é especialmente relevante em sistemas de pesquisa de imagens por conteúdo, em que a avaliação de resultados é feita por humanos. A comparação entre espaços de cor, realizada em dois estudos, revelou a superioridade da representação HSV relativamente aos espaços RGB, YUV, e

Munsell (Wan e Kuo, 1996) e RGB, CIELAB, CIELUV, e YCbCr (Pickering e Rüger, 2003).

Outros trabalhos na linha de (Swain e Ballard, 1991) dedicaram-se ainda às formas de discretização do espaço de cores (Xia e Kuo, 1998; Mojsilovic e Soljanin, 2001) ou à avaliação de diferentes medidas de comparação de histogramas (Rui et al., 2008; Puzicha et al., 1999). A evolução verificada no uso da cor foi também acompanhada pela introdução de novos descritores, como os correlogramas de cor (Jing et al., 1997), os vectores de coerência de cor (Pass, Zabih e Miller, 1996) e as características definidas na norma MPEG-7 (Manjunath, Salembier e Sikora, 2002).

A textura é uma propriedade fundamental das imagens, evidenciada quer no papel que assume na percepção visual humana quer na eficácia demonstrada em sistemas automáticos – de pesquisa de imagens, segmentação, classificação de materiais, etc. Na construção de descritores baseados na textura têm sido sobretudo usados métodos i) estatísticos, ii) de processamento de sinal, iii) baseados em experiências psico--físicas e iv) geométricos. De seguida sumarizam-se os desenvolvimentos mais importantes dentro de cada categoria.

i) Dentro da primeira categoria encontram-se as medidas retiradas da matriz de co-ocorrência, proposta por Haralick (1979). Nesta matriz, cada entrada de coordenadas (i,j) contém a frequência de ocorrência de dois píxeis respectivamente com níveis de cinzento i e j. Cada matriz de co-ocorrência está associada a um vector d, que define o deslocamento entre os dois píxeis, e que é usado no cálculo daquelas estatísticas. Com base nesta matriz, diversas características de textura foram inicialmente propostas, tais como a entropia, contraste, homogeneidade, energia e correlação (Gotlieb e Kreyszig, 1990). Mais recentemente, foram propostos desenvolvimentos que estendem o uso da matriz de co-ocorrência à análise multi-resolução (Roberti de Siqueira, Robson Schwartz e Pedrini, 2013) e à combinação de textura e cor (Palm, 2004).

ii) A análise de experiências psicométricas realizadas com pessoas permitiu a Tamura, Mori e Yamawaki (1978) desenvolverem modelos matemáticos de 6 medidas, conhecidas como características de Tamura, que estão correlacionadas com a experiência visual humana. Entre estas, as três características de *granularidade*, *contraste* e *direccionalidade* foram reconhecidas como sendo mais úteis, por serem independentes e mais relevantes na percepção visual. Recentemente, modelos mais

precisos para o cálculo da granularidade e da direccionalidade foram avançados respectivamente por Chamorro-Martínez et al. (2007) e por Islam, Dengsheng e Guojun (2008).

iii) A análise de texturas tem sido muitas vezes realizada através de métodos oriundos da área de processamento de sinais. As primeiras aplicações desenvolvidas neste domínio usaram a transformada de Fourier (Bajcsy, 1973; Matsuyama, Miura e Nagao, 1983) mas posteriormente esta abordagem foi sendo substituída pela decomposição por *wavelets* (Mallat, 1989) e por filtros de Gabor (Daugman, 1985), que permitem a análise multiresolução. A análise por *wavelets*, a um dado nível de resolução, consiste na decomposição da imagem em 4 sub-bandas de frequência, a primeira contendo as baixas frequências e as restantes retendo as variações de alta frequência nas direcções horizontal, vertical e diagonal. Dentro deste esquema, uma representação multiresolução é obtida de forma iterativa, com a imagem de baixas frequências sendo fornecida à etapa seguinte, em que é feita nova decomposição em 4 sub-bandas. Esta representação é tipicamente convertida num descritor global que reúne algumas estatísticas calculadas sobre cada sub-banda, tais como a entropia, desvio padrão ou energia (Laine e Fan, 1993; Kokare, Biswas e Chatterji, 2005).

Traçando um histórico do interesse que os filtros de Gabor têm recebido ao longo das últimas décadas, na sua origem está o artigo de Daugman (1985), que revelou as propriedades óptimas destes filtros, em termos da sua resolução nos domínios do espaço e frequência, e que demonstrou serem bons modelos da actividade das células simples do córtex visual. Essencialmente, estes são filtros passa-banda locais, descritos por um *kernel* bidimensional que resulta do produto de uma função gaussiana com uma sinusóide. A primeira determina a região de suporte do filtro e é parametrizada pelo desvio-padrão e excentricidade, enquanto a segunda é uma onda no plano caracterizada pela sua frequência e orientação. Tipicamente, a aplicação dos filtros de Gabor passa por estabelecer uma função de Gabor de base que é depois rodada e ampliada, gerando um banco de filtros que captam estruturas do tipo arestas em diferentes orientações e resoluções (Bianconi e Fernández, 2007). Após a convolução destes filtros com a imagem original, pode ser obtido um descritor de texturas calculando a média e o desvio padrão das respostas, ou calculando um histograma das energias medidas sobre as imagens de saída (Deselaers, Keysers e Ney, 2008).

iv) Na categoria dos métodos geométricos encontram-se as abordagens que encaram as imagens como uma composição de padrões primitivos, retirados de um conjunto pré-definido. Nesta linha, uma técnica de assinalável sucesso é a que recorre ao conceito de *textons* (Leung e Malik, 2001), padrões de textura contidos num dicionário construído por aglomeração (ex. *k-means*). Dentro desta categoria encontra-se também o método LBP, que analisa a vizinhança de cada píxel através de uma geometria pré-definida, à qual está associado um número de padrões possíveis. Segundo Mäenpää e Pietikäinen (2005), o método LBP constitui uma unificação das abordagens geométrica e estatística já que, associada à extracção de padrões pré-definidos, está a construção de descritores da imagem como histogramas da ocorrência desses padrões. O método tem sido muitas vezes escolhido em problemas que requerem a análise de texturas (Takala, Ahonen e Pietikäinen, 2005; Liao, Law e Chung, 2009; Nanni, Lumini e Brahnam, 2010; Di et al., 2011), devido ao seu bom poder descritivo e à sua rapidez de execução.

A comparação de descritores globais foi realizada em diversos estudos ao longo da última década. Em (Howarth e Ruger, 2005) é feita a comparação entre as características baseadas na matriz de co-ocorrência, as características de Tamura e as características obtidas por filtros de Gabor, as quais se revelaram superiores às anteriores. Na comparação de vários descritores de textura, Deselaers, Keysers e Ney (2008) verificaram que as características de Gabor e histogramas de características de Tamura superavam as definidas pela norma MPEG e as extraídas da matriz de co-ocorrência. Schwartz, Roberti de Siqueira e Pedrini (2012) realizaram um estudo sobre classificação de texturas em que um conjunto abrangente de descritores foi considerado. Nesse trabalho, as características de Gabor e LBP apresentaram melhor desempenho do que os métodos baseados em *wavelets*, *Markov Random Fields* ou no espectro de Fourier. Um aspecto comum a alguns estudos é a constatação de que a combinação de diferentes descritores de texturas (Deselaers, Keysers e Ney, 2008; Schwartz, Roberti de Siqueira e Pedrini, 2012) ou a combinação de textura e cor (Howarth e Ruger, 2005) oferecem uma descrição mais robusta das imagens.

Nesta secção cabe ainda mencionar a característica Gist, pela atenção que tem recebido nos últimos anos. Ao contrário dos métodos descritos anteriormente, que tinham um âmbito de aplicação genérico, esta característica foi desenvolvida especi-

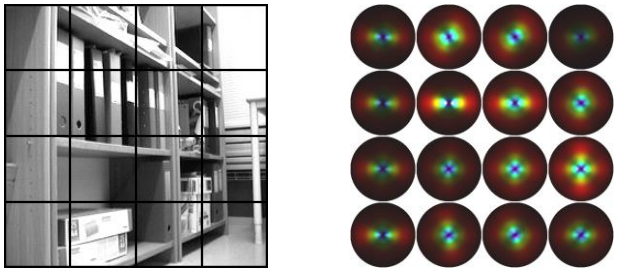


Figura 1.5. Visualização da característica Gist. À esquerda: imagem a descrever, sobreposta com a grelha de partição. À direita: ilustração da informação extraída, na forma de gráfico polares. Cada gráfico representa a informação de uma das 4×4 parcelas da imagem. Nestes gráficos, a luminosidade representa a média da resposta de um filtro, com a coordenada de ângulo correspondendo à orientação do filtro e o raio correspondendo à sua escala (raio menor – frequências baixas; raio maior – frequências altas).

ficamente para a descrição de cenas – ainda que, para isso, recorra a métodos clássicos de análise de texturas. Esta característica foi pela primeira vez descrita no influente artigo de Oliva e Torralba (2001), onde os autores estudam a percepção visual de cenas. Através das experiências reportadas, os autores concluem que a categorização semântica de cenas é baseada em cinco propriedades da imagem: naturalidade, abertura, rugosidade, expansão e robustez[1]. Uma das teses fortes daquele trabalho é a de que a categorização é feita previamente à discriminação dos objectos que constituem a cena e que aquelas propriedades são, por isso, medidas dispensando a segmentação da imagem. Na perspectiva da visão computacional, a contribuição fundamental do artigo foi a proposta da característica global Gist, da qual se podem derivar as propriedades perceptivas mencionadas atrás, ou realizar directamente a classificação semântica de cenas.

Existem diversas formas de obter a característica Gist, mantendo os princípios de concisão e a utilização de operações de baixo nível. Na abordagem mais comum, esta característica é construída a partir das respostas de filtros de Gabor com diversas orientações e escalas. Por forma a incorporar informação espacial, o domínio da imagem é dividido numa grelha de $N \times N$ e cada parcela é representada pelas médias das respostas dos filtros nessa região (ver Figura 1.5). Em resultado destas operações obtém-se um descritor de dimensão $N \times N \times K$, onde K é o número de filtros, dado por nº de orientações×nº de escalas. Com o objectivo de gerar um descritor mais conciso, e

[1] Estas são designações dadas em linguagem natural para propriedades da cena, com uma interpretação que é detalhada por Oliva e Torralba (2001). Por exemplo, o grau de naturalidade distingue as cenas de natureza das de construção humana; o grau de abertura distingue as cenas em que a linha do horizonte é maioritariamente visível daquelas em que o campo de visão é ocupado por elementos próximos.

que reflicta as variações do Gist mais relevantes para a descrição de imagens naturais, é geralmente aplicada análise PCA. Através desta obtêm-se as componentes principais do descritor, que passa a ser representado por um vector de dimensão inferior (Oliva e Torralba, 2006).

1.3 Representações da aparência

1.3.1 Representação por características locais

Os progressos realizados sobre a extracção de características locais, noutros domínios da visão computacional, têm sido genericamente integrados em sistemas de localização. Em 2001, o autor da característica SIFT apresentou a sua utilização num sistema de localização, mas a estratégia desenvolvida era do tipo métrica (Se, Lowe e Little, 2001). Posteriormente, Kosecka e Fayin (2004) realizaram localização sobre mapas topológicos, usando conjuntos de descritores SIFT como representação da aparência. Nessa abordagem, mais tarde adoptada também em (Goedemé et al., 2007; Andreasson, Duckett e Lilienthal, 2008; Valgren e Lilienthal, 2010), a comparação entre duas imagens passa pela comparação de todos os pares de descritores retirados de uma e de outra, por forma a encontrar correspondências. O número de correspondências é usado como medida de semelhança entre imagens, por vezes com modificações informadas por dados geométricos (Goedemé et al., 2007).

Com a introdução do modelo BoW, este paradigma passou a dominar as representações baseadas em características locais. O primeiro trabalho reportado que faz uso deste modelo foi apresentado por Wang, Cipolla e Zha (2005). Nesse estudo, tal como noutros que se seguiram, existe uma fase inicial em que o ambiente é modelado e é construído o vocabulário visual. Nessa fase, são recolhidas imagens do ambiente onde o robô se irá movimentar e extraídas as características locais de cada uma. Estas passam de seguida pelo processo aglomeração *k-means*, resultando num conjunto de descritores que constitui o vocabulário. Este vocabulário é usado na quantização de descritores SIFT, permitindo a representação da aparência de forma compacta, através dos histogramas da ocorrência de palavras visuais. Adicionalmente, em (Wang, Cipolla e Zha, 2005) estes vectores são modificados para a forma tf-idf (*term frequency-inverse document frequency*; Sivic e Zisserman, 2003).

Tendo em vista a aplicação do modelo BoW em ambientes de dimensão elevada, Fraundorfer, Engels e Nister (2007), exploraram uma técnica usada na pesquisa rápida de documentos de texto, baseada em ficheiros invertidos. Segundo esta técnica, para cada palavra visual é mantida uma lista das imagens onde ela ocorreu. Na fase de pesquisa, são extraídas as palavras visuais da imagem de teste e as listas são usadas para encontrar as imagens do modelo onde elas existem. Desta forma, obtém-se um conjunto reduzido de imagens que serão testadas pela comparação de histogramas ou por verificação geométrica.

A introdução de descritores alternativos ao SIFT reflectiu-se também na robótica móvel, existindo sistemas que recorreram ao descritor SURF (Valgren e Lilienthal, 2010; Cummins e Newman, 2008) e ao descritor BRIEF (Galvez-López e Tardós, 2012). Uma prática que tem vindo a ser adoptada nestas soluções é a eliminação da invariância à rotação das características. As vantagens desta abordagem, pela primeira vez identificadas por Williams e Ledwich (2004), são o menor número de características extraídas e a eliminação do ruído devido à estimação incorrecta da orientação. A etapa de detecção de regiões de interesse foi também estudada no contexto da robótica móvel, por Ramisa et al. (2009). Naquele trabalho, verificou-se que o detector Harris-Affine apresenta os melhores resultados num problema de localização e que a combinação de diferentes detectores, naquele caso Harris-Affine, Hessian Affine e MSER, é a estratégia que oferece maior robustez.

1.3.2 Representação por características globais

O trabalho de Ulrich e Nourbakhsh (2000) foi um marco na localização por características globais, na medida em que representou, pela primeira vez, a aparência na forma de uma estatística da imagem. Naquela investigação usaram-se características de cor retiradas de imagens omnidireccionais e mostrou-se que, mesmo recorrendo a características muito simples e sem aplicar uma análise geométrica, é possível usar informação visual para localizar um robô. Por forma a obter a descrição da aparência, os valores de cor foram projectados nos espaços HLS e RGB de onde se extraíram histogramas independentes para cada um dos canais. Como representação da assinatura visual foi usado o conjunto dos seis histogramas assim obtidos.

Devido à sua simplicidade e poder descritivo, a caracterização da aparência por histogramas de cor foi mais tarde aplicada em estudos de localização (Blaer e Allen,

2002) e em trabalhos que estendem o âmbito do problema à extracção de mapas topológicos (Werner, Sitte e Maire, 2007; Werner, Maire e Sitte, 2009).

A associação de imagens panorâmicas ou omnidireccionais a características globais, mencionada por Ulrich e Nourbakhsh (2000), foi frequentemente adoptada por outros autores. Esta opção – que pode ser explicada pelo facto de estas imagens serem mais informativas sobre os lugares, permitindo assim o uso de características com discriminatividade inferior – levou ao desenvolvimento de métodos específicos para visão omnidireccional. Em (Ishiguro et al., 2003), as imagens omnidireccionais são em primeiro lugar transformadas para o formato panorâmico e, posteriormente, é aplicada a transformada de Fourier unidimensional sobre as linhas destas imagens. Notando que a rotação do robô se manifesta em translações nas imagens panorâmicas, e que o espectro de Fourier não contém informação sobre a localização da ocorrência das frequências, os autores concluem que este espectro constitui uma representação invariante à rotação. Sobre esta propriedade, é proposto um descritor para estas imagens, a assinatura de Fourier, constituída pelos primeiros 15 coeficientes do espectro, que tipicamente contêm as amplitudes mais elevadas. A mesma estratégia é usada, numa versão mais simples, por Ranganathan e Dellaert (2005) e Gerstmayr-Hillen et al. (2011). Em (Gerstmayr-Hillen et al., 2011), para além da assinatura de Fourier, são usados histogramas de níveis de cinzento, os 4 primeiros momentos estatísticos da imagem (média, variância, assimetria e curtose) e uma medida do centro de massas da imagem. No sistema RatSlam, de Milford e Wyeth (2010), é igualmente usada a transformada de Fourier sobre as linhas de imagens panorâmicas, reservando-se os coeficientes da partes real e imaginária como descritor da imagem. Estes valores são usados posteriormente para comparar imagens através da sua correlação.

A característica Gist, tendo sido desenhada para modelar a percepção visual de espaços, constituiu uma escolha natural para o desenvolvimento de sistemas de localização. Nalguns trabalhos esta característica foi usada na sua forma original mas, ocasionalmente, foram feitas adaptações ao problema em estudo. Murillo e Kosecka (2009) introduziram o conceito de panoramas Gist, a adaptação daquela característica a imagens panorâmicas. Na solução proposta, o panorama é dividido em quatro imagens e o descritor standard é calculado para cada uma delas. Uma representação mais compacta da aparência é obtida pela quantização de cada um dos descritores, o

que resulta num descritor final composto por apenas quatro índices. A invariância à rotação é conseguida na fase de pesquisa, em que são testadas quatro translações dos índices por forma a identificar o melhor alinhamento.

Schubert et al. (2007) inspiraram-se também na característica Gist, mas substituíram os filtros de Gabor por filtros mais simples, de arestas e cantos. Na construção do descritor aplicaram uma divisão diferente da imagem e mediram a energia e curtose das respostas de cada filtro. Kai et al. (2008) combinaram o Gist com o conceito de epítome (Jojic, Frey e Kannan, 2003), com vista a obter uma característica com maior invariância à translação e escala, em imagens de perspectiva. A preocupação subjacente à modificação de Sunderhauf e Protzel (2011) foi a de criar uma versão do Gist de extracção mais rápida, tendo para isso substituído os filtros de Gabor pelo descritor BRIEF, que foi aplicado em cada divisão da imagem.

1.4 Localização e Mapeamento

1.4.1 Modelação probabilística

A modelação probabilística é uma forma conveniente de lidar com os problemas de estimação na robótica móvel, por representar de forma explícita os tipos de incerteza presentes nos dados. Esta abordagem assenta no compromisso epistemológico de que as crenças de um agente autónomo podem ser descritas por valores de probabilidade (Russell e Norvig, 2009). Neste contexto, uma distribuição de probabilidades sobre a variável aleatória X é entendida como as crenças depositadas pelo agente sobre as hipóteses para aquela variável.

Para além de proporcionar uma representação conveniente, a modelação probabilística fornece um aparato formal em que a combinação de informação é teoricamente justificada. A situação típica em que este procedimento é útil pode descrever-se como: um robô deve estimar uma variável X, não observável, inferindo informação sobre ela a partir de uma variável Y, que é possível medir. No centro deste processo, designado por inferência probabilística, está a regra de Bayes, expressa por:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1.1}$$

Esta regra constitui um mecanismo através do qual o conhecimento sobre a variável *X* é modificado por forma a reflectir a nova informação, dada pela leitura de *Y*. Segundo esta interpretação, *P*(*X*) sumariza o conhecimento inicial sobre *X* e é designada por distribuição à priori. A probabilidade de *X* condicionada pela leitura de *Y*, *P*(*X*|*Y*), representa a informação final sobre *X*, que resulta da integração daquela leitura com a informação inicial. Inerente à utilização da regra de Bayes está a ideia de que existe uma relação causal entre as duas variáveis, que justifica a inferência de uma a partir da outra. Nessa relação, expressa por *P*(*Y*|*X*), *Y* é entendida como um efeito, cuja causa é a variável a estimar, *X*. Finalmente, na Eq. (1.1), *P*(*Y*) é a probabilidade da observação, que tem a função de normalizar a distribuição posterior.

Os problemas da robótica móvel lidam com contextos dinâmicos – por um lado os robôs mudam de posição e, por outro, o ambiente em que se movem pode sofrer transformações. Embora o ambiente seja muitas vezes assumido como estático, considerar a dinâmica do robô é fundamental. Estudar estes problemas à luz da inferência bayesiana passa por reconhecer que, para além das observações, existe informação adicional relativa às acções do robô. A integração desta informação no contexto probabilístico conduz a regras de inferência designadas por filtros bayesianos.

No problema de localização, a variável a estimar, *X*, é a posição do robô e *y* designa as leituras feitas pelos sensores. O objectivo é estimar *X* usando toda a informação disponível, que inclui as leituras realizadas desde o instante inicial ao actual, $y_0, \ldots, y_t$, e também as acções do robô nesses momentos, $u_0, \ldots, u_t$. Os filtros bayesianos assentam na premissa markoviana de que o estado é completo, isto é, contém toda a informação que determina as observações e o estado seguinte. Este pressuposto permite que a estimação possa ser realizada de forma recursiva, integrando, em cada momento, a informação mais recente. Este processo é realizado em duas etapas descritas por:

Predição:

$$P(x_t|X_{t-1}, u_t) = \int P(x_t|x_{t-1}, u_t)\, P(x_{t-1}) dx_{t-1} \qquad (1.2)$$

Actualização:

$$P(x_t|X_{t-1}, u_t, z_t) = \eta P(z_t|x_t) P(x_t|X_{t-1}, u_t) \qquad (1.3)$$

Na Eq. (1.2), $P(x_t|x_{t-1},u_t)$ é designada por modelo das acções e descreve a forma como a distribuição estimada é modificada através da acção u_t. Na etapa de predição, a aplicação do modelo das acções gera uma primeira versão da estimativa de x_t, aquela que é prevista com base na acção realizada. Na etapa de actualização é incorporada a informação da observação, através da regra de Bayes. A distribuição resultante é normalizada pelo factor η.

Sobre estas expressões foram desenvolvidos diversos estimadores recursivos que diferem na forma como cada um dos componentes do estimador é concretizado. Entre estes, o aspecto que determina mais profundamente a natureza da solução é o tipo de representação para a distribuição X. Devido à dificuldade de tratar distribuições contínuas que possam tomar uma forma arbitrária, torna-se necessário introduzir algumas simplificações. O tipo de abordagem mais simples consiste em representar a distribuição por uma função gaussiana, o que conduz a uma descrição compacta da estimativa da posição, dada apenas pela sua média e matriz de covariância. Estas soluções, genericamente designadas por filtros de Kalman, são extremamente atractivas dada a sua simplicidade e a sua robustez, que é garantida na condição de se conhecer uma estimativa inicial suficientemente precisa (Thrun, Burgard e Fox, 2005). Contudo, devido à impossibilidade de representar adequadamente múltiplas hipóteses, esta abordagem falha em situações em que aquela condição não se verifica. Por esta razão, o filtro de Kalman torna-se inadequado ao tratamento de dois problemas paradigmáticos da localização: o da localização global e o do robô raptado (Thrun, Burgard e Fox, 2005). Por forma a lidar com estes problemas, foram desenvolvidas duas soluções alternativas ao filtro de Kalman: a localização com *Hidden Markov Models* (HMM) e a localização de Monte Carlo, que serão tratadas no ponto seguinte.

1.4.2 Localização

Além da abordagem probabilística, que será discutida mais à frente, alguns trabalhos relevantes seguiram outra via, que em lugar de encarar a localização como um problema de inferência, tratam-no na perspectiva da classificação. Um exemplo desta abordagem é dado no artigo de Pronobis et al. (2006), em que um classificador discriminativo do tipo SVM (*Support Vector Machine*) é usado para integrar informação de características globais e locais. Apesar dos bons resultados obtidos,

este tipo de classificador é limitado no sentido de não permitir a inclusão, de forma simples, de novos lugares no ambiente modelado. Outra abordagem que tem sido mais usada recorre à classificação do tipo vizinho mais próximo. Os trabalhos que seguem esta via baseiam-se tipicamente em características locais, na forma não quantizada (Kosecka e Fayin, 2004; Ramisa et al., 2009; Valgren e Lilienthal, 2010) ou sob o modelo BoW (Wang, Cipolla e Zha, 2005; Fraundorfer, Engels e Nister, 2007). No primeiro caso a medida de semelhança é calculada pelo número de correspondências encontradas entre as características da imagem de teste e da imagem de referência, usando o critério de Lowe (2004), enquanto no segundo esta medida é dada pela comparação de vectores BoW. Uma vez dispondo da distribuição de semelhanças sobre as imagens de referência, o resultado da localização é dado como o lugar cuja imagem obteve maior pontuação. Um aspecto comum a vários trabalhos nesta linha foi a constatação de que a medida de semelhança pode ser aperfeiçoada com a introdução de constrangimentos geométricos. Consequentemente, Wang, Cipolla e Zha (2005), Ramisa et al. (2009) e Valgren e Lilienthal (2010) calculam uma medida final de semelhança igual ao número de correspondências que verificam as condições da geometria epipolar.

No campo das abordagens probabilísticas, os sistemas de localização distinguem-se, em primeiro lugar, pelo tipo de representação usada na distribuição da posição. Na literatura encontram-se exemplos de soluções que recorrem à *localização com HMM* e *localização de Monte Carlo*.

A localização com HMM designa as soluções em que o ambiente é discretizado, o que permite a descrição do problema na forma de um modelo de Markov escondido. Este tratamento foi usado no âmbito dos mapas métricos, em que o ambiente é discretizado em grelhas de ocupação (Fox, Burgard e Thrun, 1999), e também na localização topológica, em que a discretização está implícita na representação por grafos. A especificação do modelo de Markov escondido envolve a definição de um modelo para as transições, o qual pode ser facilmente derivado da topologia do mapa ou dos dados de exploração. Por exemplo, Torralba et al. (2003) estimam a probabilidade de transição pelo número de transições realizadas entre dois lugares durante a recolha de dados de modelação, enquanto Li, Yang e Kosecka (2005) definem, para cada lugar, probabilidades iguais de transição a partir dos lugares adjacentes. Goedemé et al.

(2007) estimam a probabilidade de transição como uma função da distância entre os dois lugares, medida sobre o mapa topológico.

Uma vez estabelecidas as probabilidades de transição, o modelo probabilístico completa-se com a definição do modelo do sensor. No caso da localização com HMM, este é uma função que gera a probabilidade da observação num dos lugares do ambiente discretizado. Kosecka e Fayin (2004) e Goedemé et al. (2007) oferecem exemplos do cálculo desta probabilidade com características locais, o qual é baseado no número de correspondências encontradas entre imagens. Em ensaios em que a aparência foi descrita por características globais, foi sugerido o cálculo daquela probabilidade com uma função gaussiana, centrada num descritor representativo do lugar (Kosecka e Fayin, 2004), ou mistura de gaussianas (Torralba et al., 2003).

Como fizeram notar Fox, Burgard e Thrun (1999), a localização com HMM pode colocar dificuldades quando a resolução usada na discretização do ambiente é demasiado larga, o que obriga a um ajuste dos modelos do sensor e das acções àquela representação. Uma alternativa a este método, que elimina esta dificuldade, é a localização de Monte Carlo, em que a distribuição posterior é representada por um conjunto de amostras (também designadas por partículas) sobre um espaço contínuo. Esta solução é usada por Siagian e Itti (2009), que discretizam vários ambientes de larga dimensão num pequeno número de lugares. Naquele trabalho cada lugar é entendido como cobrindo um segmento do ambiente, o que permite a introdução de variáveis contínuas para descrever o espaço percorrido nesse segmento. Sob este modelo, as partículas que representam a distribuição de probabilidades são descritas num vector de estado dado por dois valores: o valor discreto do lugar e o valor contínuo, entre 0 e 1, do percurso realizado. A aplicação da localização de Monte Carlo sobre esta configuração permitiu realizar a localização com erros de posição geralmente inferiores a 3m, num ambiente de dimensão 137.16×178.31m, apesar da sua discretização num número reduzido de lugares. A localização de Monte Carlo com representações baseadas na aparência foi também reportada por Gross et al. (2003), com erros de posição baixos. Contudo, neste sistema o mapa do ambiente é métrico, o que exigiu a estimação das posições do modelo com precisão, ao contrário das abordagens topológicas.

1.4.3 Construção de mapas de aparência

Numa fase anterior à localização é necessário que o robô construa um mapa do ambiente onde vai operar. Focando esta tese os métodos baseados na aparência, os mapas de interesse são do tipo topológico, i.e., constituídos por nós, que representam lugares, e ligações, que representam a acessibilidade entre eles. Tipicamente, parte desta informação decorre directamente da forma como os dados são recolhidos. Em trabalhos como os de Galvez-López e Tardós (2012) e de Cummins e Newman (2008) as imagens são capturadas a uma frequência constante, ou a um distância métrica constante, e a cada imagem é associado um lugar (nó do mapa topológico). Além disso, a sequência de recolha estabelece ligações triviais entre os nós, já que imagens sucessivas provêm de lugares ligados entre si. Contudo, a topologia resultante deste processo é normalmente incompleta, pois não prevê a revisitação de lugares. Por forma a reflectir a estrutura do ambiente torna-se então necessário que, nos momentos em que o robô atinge lugares já visitados, sejam introduzidas as ligações respectivas no mapa. Neste cenário, o problema da construção de um mapa consiste no estabelecimento de ligações, para além das triviais, partindo do pressuposto de que qualquer nó existente pode estar ligado a quaisquer outros.

Alguns dos primeiros trabalhos sobre mapeamento topológico procuraram estimar simultaneamente todas as ligações do mapa. Um exemplo deste esforço é dado por Shatkay e Kaelbling (2002), onde o problema de mapeamento é colocado como o da estimação de um modelo de Markov escondido, o qual foi resolvido pelo método da maximização da expectativa. Observando que este método pode convergir para mínimos locais, Ranganathan e Dellaert (2004) propuseram a extracção de mapas realizando inferência probabilística sobre o espaço de todas as topologias possíveis. Sob essa perspectiva, a dificuldade do problema reside na elevada dimensionalidade do espaço de inferência, que cresce de forma hiper-exponencial com o número de observações (Ranganathan e Dellaert, 2004). Estes dois autores lidaram com esta complexidade através da estimação de Monte Carlo, avançando com um algoritmo que gera uma distribuição sobre o espaço de topologias, na forma de amostras sobre esse espaço. Assim, para além de fornecer uma topologia mais provável, os resultados deste método incorporam a incerteza sobre a estimação realizada.

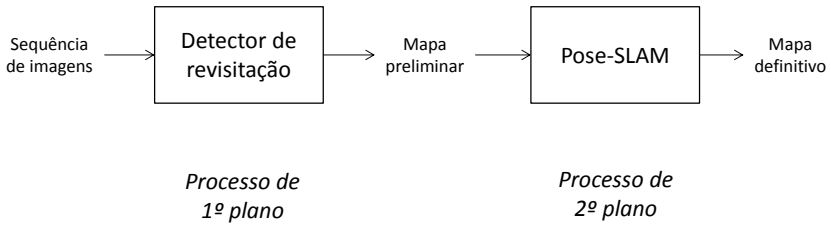


Figura 1.6. Construção de mapas topológicos em duas etapas. Processo de primeiro plano: detector de revisitação; processo de segundo plano: pose-SLAM.

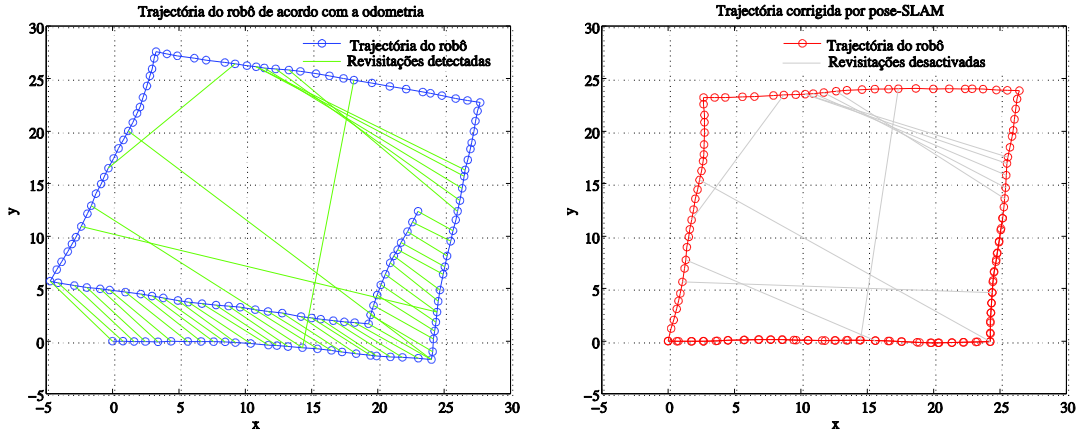


Figura 1.7. Ilustração dos resultados obtidos nas duas etapas de construção do mapa. À esquerda trajectória estimada por odometria e ligações estabelecidas por detecção de revisitação; à direita: correcção do mapa por pose-SLAM. Figura adaptada de (Sunderhauf e Protzel, 2012).

Nos últimos anos o mapeamento topológico tem sido encarado sob um prisma diferente, mais simples, por tratar independentemente cada potencial ligação entre nós. Nesta abordagem, a extracção da topologia reduz-se à detecção da revisitação de lugares, i.e., à identificação do lugar presente como um lugar já visitado, seguida da introdução da ligação respectiva. Sobre este mecanismo foi desenvolvido um paradigma completo que prevê ainda a correcção do mapa mediante a informação recolhida. Estes sistemas têm sido descritos como constituídos por dois processos, um de primeiro plano, o detector de revisitação, e um de segundo plano designado *pose-SLAM* (*pose-Simultaneous Localization And Mapping*), que realiza a correcção do mapa sob os constrangimentos métricos e topológicos disponíveis (ver Figuras 1.6 e 1.7).

A detecção de revisitação foi extensivamente estudada no passado, frequentemente associada ao modelo BoW. Um dos primeiros trabalhos que segue esta via foi desenvolvido por Fraundorfer, Engels e Nister (2007). Nele, a detecção de revisitação assenta unicamente no poder discriminativo da comparação de vectores BoW, seguida da verificação geométrica das correspondências. Mais recentemente, a mesma

estratégia foi também usada por Galvez-López e Tardós (2012), com excelentes resultados. Outros autores contribuíram com modelos probabilísticos, com o objectivo de superar a simples comparação de vectores BoW. Por exemplo, Angeli et al. (2008) propuseram um filtro de Bayes que integra informação temporal na tomada de decisão. Numa perspectiva diferente, Cummins e Newman (2008) focaram-se no desenvolvimento de um modelo probabilístico gerador da aparência. O sistema resultante, conhecido por FAB-Map (*Fast Appearance Based Mapping*), foi muitas vezes considerado o estado da arte na detecção de revisitação. A originalidade deste sistema consistiu na modelação da aparência, não apenas como função da probabilidade individual de cada palavra visual, mas também como função da probabilidade da sua detecção e da probabilidade conjunta de ocorrência, modelada por árvores Chow-Liu (Chow e Liu, 1968).

A detecção de revisitação foi também abordada na perspectiva das características globais, em vários estudos que exploram a eficiência destas características. Focando a navegação em ambientes interiores, Gerstmayr-Hillen et al. (2011) avaliaram vários descritores globais e funções de comparação para imagens panorâmicas, concluindo que é possível realizar a detecção de revisitação com vectores de dimensão entre 16 e 128. A navegação em ambientes mais extensos, de exterior, foi também estudada em vários trabalhos. Sunderhauf e Protzel (2011) sugeriram pela primeira vez a extracção da característica Gist com descritores de textura mais simples, utilizando a característica BRIEF. O sistema resultante demonstrou boa precisão, mas é atractivo sobretudo pela rapidez de execução. Liu e Zhang (2012) recorreram à característica Gist na sua definição original, integrada num filtro de Bayes, o que conduziu a um detector com excelente precisão num cenário de exterior. Contudo, esta solução apresenta tempos de computação largamente superiores aos do BRIEF-Gist, proposto por Sunderhauf e Protzel (2011).

A característica Gist foi ainda usada em ambientes exteriores descritos através de imagens panorâmicas por Murillo et al. (2013) e Arroyo et al. (2014b). Em ambos está patente a preocupação de reconhecer os lugares sob mudança de orientação e para isso os panoramas são divididos em sub-panoramas e testadas as várias possibilidades de correspondência mediante rotações do robô. No estudo de Arroyo et al. (2014b) são ainda avaliados os descritores binários ORB, BRISK, FREAK, LDB na extracção do Gist, concluindo-se que o último é o mais adequado ao problema de revisitação.

A correcção topológica dos mapas extraídos por detecção de revisitação depende unicamente da precisão do detector, pelo que os valores de *recall* conseguidos estão limitados pela necessidade de garantir elevada precisão. Os métodos de segundo plano, mencionados anteriormente, permitem realizar a correcção das coordenadas métricas dos nós e, em desenvolvimentos recentes, identificar ligações topológicas erradas. Ao contrário do SLAM métrico tradicional, em que a posição do robô e o mapa são estimados de forma incremental por filtros de Bayes, estas soluções, genericamente designadas por pose-SLAM, estimam simultaneamente todas as posições do robô introduzidas no mapa. Como dados de entrada, o pose-SLAM recebe um mapa topológico complementado com informação métrica, a das posições relativas dos nós e as incertezas que lhes estão associadas. Em casos típicos, em que existe revisitação, estes dados métricos são globalmente inconsistentes, devido aos erros de odometria. Os algoritmos de pose-SLAM procuram minimizar esta inconsistência, através de métodos de optimização que calculam novas posições para os nós, sob os constrangimentos métricos e topológicos fornecidos no mapa inicial. Esta ideia foi inicialmente introduzida em 1997 por Lu e Milios (1997), mas, devido às dificuldades no cálculo da solução pela máxima verosimilhança, em espaços de elevada dimensão, recebeu pouca atenção por parte da comunidade científica. Mais tarde, vários autores retomaram esta ideia, propondo algoritmos de optimização mais eficientes como os métodos de relaxamento (Frese, Larsson e Duckett, 2005), de gradiente descendente (Olson, Leonard e Teller, 2006), de suavização (Kaess, Ranganathan e Dellaert, 2007) e hierárquicos (Grisetti et al., 2010). Recentemente foram realizados avanços muito relevantes no pose-SLAM, com a introdução de algoritmos robustos a *outliers*. Estas soluções, inicialmente estudadas por Sunderhauf e Protzel (2012) e mais tarde por Agarwal et al. (2013), admitem a existência de *outliers* no mapa topológico, i.e., ligações erradas estabelecidas pelo detector de revisitação. Através de algoritmos de optimização robustos, Sunderhauf e Protzel (2012) demonstraram ser possível corrigir não apenas a informação métrica mas também a informação topológica do mapa inicial. Segundo os autores, este tipo de solução estreita a relação entre os métodos de primeiro e de segundo plano, pois permite aliviar as exigências sobre o detector de revisitação, cujos erros podem ser corrigidos numa fase posterior.

1.5 Motivação

Actualmente, pode dizer-se que a localização visual de robôs é um problema em larga medida resolvido, dada a existência de enquadramentos teóricos (filtros de Bayes, geometria epipolar) e técnicas de extracção de características que suportam vários sistemas bem sucedidos. Este cenário tem justificado a maior ênfase dada pela comunidade científica, nos últimos anos, ao problema de extracção de mapas, o qual é teórica e tecnicamente mais desafiador. No entanto, desde os trabalhos de Jogan et al. (2002) e de Pronobis et al. (2006), que são conhecidas as dificuldades da localização sob diferentes condições de luminosidade. Mais recentemente, Neubert, Sunderhauf e Protzel (2013) observaram que "os ambientes variáveis colocam problemas sérios aos sistemas robóticos actuais que ambicionem uma operacionalidade de longo termo. Os sistemas de reconhecimento de lugares resultam razoavelmente bem em ambientes estáticos ou de dinâmica reduzida, contudo, variações severas de aparência, que ocorram entre o dia e a noite, entre diferentes estações ou diferentes condições atmosféricas, permanecem um desafio". De facto, a formulação tradicional do problema de localização, do ponto de vista probabilístico, é inadequada nestes casos, já que a premissa de Markov pressupõe ambientes estáticos. Por outro lado, a verificação geométrica pelos constrangimentos da geometria epipolar, embora robusta, é uma medida insuficiente perante variações severas de aparência, devido à forte presença de *outliers*.

O problema descrito foi recentemente abordado por Churchill e Newman (2012) e Krajnik et al. (2014), que optaram por modelar explicitamente as variações temporais de aparência. Complementar a estes esforços está a necessidade de se desenvolverem representações da aparência mais robustas. Esta é a via seguida na presente tese, que estuda formas de representação e comparação da aparência na perspectiva do poder descritivo que proporcionam. O âmbito deste estudo é o da descrição da aparência por características locais, em particular do tipo SIFT, a qual constitui a abordagem mais popular e bem documentada na literatura relacionada. Subjacente a este trabalho está a ideia de que a representação que tem vindo a ser genericamente aceite para estas características, dentro do modelo BoW, limita o seu poder descritivo. Esta observação foi pela primeira vez feita por Jurie e Triggs (2005) no contexto de características densamente extraídas na imagem. Calculando a distribuição de características numa partição regular do espaço de descritores, os autores verificaram que as características

encontradas em imagens naturais seguem aproximadamente uma distribuição da lei de potência, à semelhança da lei de Zipf para as palavras em documentos de texto. Com base neste dado foi concluído que os centróides resultantes da aglomeração por *k-means* se concentram nas áreas mais densas do espaço de descritores. Mais tarde, Boiman, Shechtman e Irani (2008) aprofundaram a análise das propriedades do modelo BoW, verificando que os erros introduzidos na etapa de quantização de descritores aumentam com a discriminatividade das características. Em suma, as características de maior discriminatividade perdem maior poder descritivo, pelo facto de aquelas regiões do espaço de descritores estarem pouco representadas na distribuição de centróides. O artigo de Boiman, Shechtman e Irani (2008) teve repercussões importantes no seio da comunidade científica, tendo sido desenvolvidas diversas extensões ao classificador *Naive Bayes Nearest Neighbour*, então proposto, devidas a Lowe (2012), Timofte, Tuytelaars e Van Gool (2013) e Rematas, Fritz e Tuytelaars (2013).

Esta tese explora as conclusões de Jurie e Triggs (2005) e de Boiman, Shechtman e Irani (2008), apresentando um método de localização visual baseada na representação não-quantizada (NQ) das características. A ideia fundamental na base desta proposta é a de que, através da eliminação dos erros de quantização, é possível obter-se representações de aparência mais discriminativas e, por isso, mais robustas perante variações ambientais severas. A apresentação do método original é acompanhada da comparação das representações quantizada (Q) e NQ no que diz respeito à sua precisão, tempo de comutação e memória utilizada.

Em termos da avaliação de métodos candidatos, o trabalho desenvolvido segue a linha de (Pronobis et al., 2006; Valgren e Lilienthal, 2010; Jianxin e Rehg, 2011) em que a localização é entendida como um problema de classificação – para cada imagem de teste, o localizador deve devolver um lugar do ambiente, apenas com base nessa imagem. Este é o contexto da localização global, em que o robô deve situar-se no ambiente, partindo de informação nula sobre o seu estado.

Nesta tese é também estudado o problema anterior à localização, o da extracção de um mapa do ambiente, em particular na perspectiva da detecção de revisitação. Tal como no tratamento da localização, o trabalho realizado sobre este tema foca-se no desenvolvimento de representações de aparência com precisão e requisitos computacionais mais adequados ao problema. A abordagem adoptada apresenta

pontos em comum com os trabalhos de Sunderhauf e Protzel (2011) e de (Liu e Zhang, 2012), onde a detecção de revisitação é baseada em características globais da imagem. À semelhança daqueles trabalhos, neste ponto da tese a análise por características locais será substituída por características globais. Esta mostrar-se-á especialmente adequada para uma categoria de ambientes e é justificada pelo facto de as revisitações realizadas durante o mapeamento estarem normalmente sujeitas a variações na aparência que são menos severas do que aquelas que podem ocorrer em fase de localização.

Relativamente aos dois trabalhos acima citados, esta tese inova pelo uso de uma característica global original, designada LBP-Gist, que combina a análise de texturas pelo método LBP com a codificação da estrutura global da imagem, inerente ao Gist. A característica proposta revelar-se-á mais precisa do que a característica BRIEF-Gist, avançada por Sunderhauf e Protzel (2011), e mais eficiente do que a característica Gist original, aplicada na detecção de revisitação por Liu e Zhang (2012).

No decorrer desta tese foram conduzidas experiências sobre um robô móvel baseado no sistema de motorização RD02 (Robot Electronics, 2014) e implantado no Instituto Superior de Engenharia de Lisboa. Estes ensaios foram úteis na validação de algumas das técnicas estudadas na tese, no entanto, o seu objectivo fundamental foi o de explorar a informação veiculada pelo sensor Kinect, especificamente os dados de profundidade. Verificou-se, contudo, que a dimensão de dados corrompidos que estão presentes no sinal de profundidade, resultantes do desvio de feixes infravermelhos em em superfícies brilhantes ou sob determinadas orientações, torna esta informação menos fiável do que o sinal visual, que será explorado nesta tese. Por esta razão, não foram incluídos na tese resultados respeitantes a estas experiências, optando-se por usar as bases de imagens públicas.

1.6 Contribuições

Nesta tese são apresentadas duas contribuições principais, relativas aos dois problemas estudados, o da localização global e o da detecção de revisitação. No âmbito do primeiro é proposto um método de localização baseado na representação NQ das características visuais. Este método é mais discriminativo do que a representação Q e, simultaneamente, mais discriminativo do que outras utilizações da

representação NQ. Para a detecção de revisitação foi desenvolvida uma nova característica global, LBP-Gist, que proporciona um bom equilíbrio entre precisão e eficiência. Mais detalhadamente, são feitas as contribuições adicionais descritas de seguida.

- É apresentada a análise de discriminatividade das representações Q e NQ, revelando os factores que prejudicam a representação quantizada. Esta análise aprofunda aquela que foi descrita por Boiman, Shechtman e Irani (2008), estudando a influência dos parâmetros e configurações das duas representações.
- É feita a comparação das duas representações relativamente à sua precisão, no contexto da localização global e em situações de luminosidade variável, uma avaliação que não tinha sido realizada em trabalhos anteriores.
- Por forma a explorar todo o potencial da representação NQ, ela deve ser acompanhada de operações de comparação de imagens adequadas. Este problema é estudado no âmbito da combinação de classificadores, um enquadramento ajustado à combinação de características NQ. Dentro deste enquadramento, é apresentada a avaliação comparativa de vários combinadores algébricos, a qual suportará a configuração do método de localização proposto.
- Se adoptada na sua forma mais imediata, a comparação de imagens pela representação NQ acarreta custos computacionais largamente superiores aos da representação Q. Este facto deve-se à necessidade de, no primeiro caso, se compararem todos os pares de características formados por características da imagem de teste e da imagem do modelo. Nesta tese são propostos meios de reduzir a disparidade entre as duas representações relativamente à rapidez de execução e memória ocupada.
- Relativamente à detecção de revisitação, é apresentada uma análise detalhada das capacidades do método LBP para a representação de cenas. Desta análise resulta a configuração proposta para a característica LBP-Gist, que toma a forma de um vector de dimensão 995. Por forma a reduzir o peso computacional do detector, é proposta a utilização do algoritmo de *hashing Winner Take All*. Sobre este são propostas duas modificações, que beneficiam a sua relação *fall-out* vs *recall*.

No decorrer do trabalho que conduziu a esta tese foram realizadas as publicações enunciadas de seguida.

Trabalho preliminar a esta tese foi apresentado na conferência ECAI 2010, na sessão de posters (Campos, Correia e Calado, 2010); o desenvolvimento do método de localização baseado na representação NQ foi apresentado na conferência ICAR 2011 (Campos, Correia e Calado, 2011) e posteriormente publicado em forma extendida em (Campos, Correia e Calado, 2012); a combinação de características não quantizadas foi abordada em (Campos, Correia e Calado, 2013b) e mais tarde aprofundada no artigo (Campos, Correia e Calado, 2015); os aspectos computacionais do detector de revisitação foram descritos em (Campos, Correia e Calado, 2013a) e o desenvolvimento do detector e da característica LBP-Gist foram apresentados no artigo (Campos, Correia e Calado, 2013c), o qual recebeu o *Best Student Paper Award* na conferência EPIA 2013.

1.7 Estrutura da tese

A presente tese está estruturada da forma descrita de seguida.

O **capítulo 2** introduz o método de localização global proposto nesta tese, que recorre à representação não-quantizada das características visuais. Nesse capítulo são apresentados os *datasets* que serão usados em todas as experiências sobre localização e é comparado o classificador NQ com dois classificadores baseados na representação Q: o Naive Bayes e o SVM. As propriedades das duas representações são analisadas à luz da discriminatividade que proporcionam para a descrição de lugares.

O **capítulo 3** foca-se apenas na representação NQ, no sentido de encontrar a melhor estratégia para a combinação de características. Para este estudo, o problema é enquadrado no contexto da combinação de múltiplos classificadores e a fusão é realizada através de combinadores algébricos. Também aqui as propriedades das soluções candidatas são analisadas através da discriminatividade que oferecem.

No **capítulo 4** retoma-se a comparação das representações NQ e Q, agora em termos dos seus requisitos computacionais. Nesse capítulo são propostas medidas para reduzir o tempo de computação e a memória usados pela representação NQ. Através da aplicação desses métodos definem-se diversas versões do classificador NQ, as

quais são confrontadas com uma gama de configurações do classificador Q, obtida pela variação da dimensão do vocabulário visual.

O **capítulo 5** aborda a construção de mapas topológicos através da detecção de revisitação de lugares. Nesse capítulo, a descrição da aparência é feita através de características globais, sendo proposta, para esse efeito, uma característica original, designada LBP-Gist. No desenvolvimento do detector de revisitação são detalhados os métodos LBP, para a extracção de texturas, e o método de *hashing Winner Take All*, necessário para a comparação rápida de descritores. O detector obtido é avaliado em vários *datasets* para os quais existem resultados publicados baseados em características locais, a abordagem que mais frequentemente tem sido usada neste problema.

O **capítulo 6** sumariza os principais resultados e conclusões retirados dos capítulos precedentes e aponta direcções de trabalho futuro levantadas pela presente tese.

2. Localização global com a representação não--quantizada de características visuais

2.1 Introdução

Neste capítulo introduz-se o método de localização global proposto nesta tese, que recorre à representação não-quantizada (NQ) das características visuais. A motivação para o desenvolvimento deste método tem origem na observação, feita por vários autores, de que a representação alternativa, fazendo uso de características quantizadas (Q), está sujeita a erros que prejudicam a sua precisão.

Na essência deste problema encontra-se a operação de quantização que, ao aproximar cada característica a uma palavra visual de uma lista finita, introduz um erro, dado pela distância entre elas. Para além da evidência de que estes erros estão presentes, há a acrescentar que eles se distribuem de forma desigual pelas características, já que dependem do posicionamento das palavras visuais no espaço de características. Esta questão está relacionada com a construção automática do vocabulário visual, a qual foi analisada em (Jurie e Triggs, 2005). Naquele trabalho, os autores fazem notar que as características encontradas em imagens naturais se distribuem de forma não uniforme no espaço de características, i.e., que alguns tipos de padrões visuais se repetem mais frequentemente do que outros. Em resultado disso, os centróides (palavras visuais) devolvidas pelo método *k-means* concentram-se nas áreas mais densas do espaço de características, enquanto as regiões pouco ou mediamente populadas são menos representadas. Uma vez que as características menos frequentes, que incidem nestas regiões, são potencialmente mais discriminativas, este tipo de vocabulário é sub-óptimo para classificação.

A identificação destes problemas inspirou uma série de trabalhos que procuram atenuar os erros de quantização, contudo, a abordagem que garante a sua eliminação é aquela que parte de um princípio diferente, o da não quantização das características. Esta abordagem foi proposta no influente artigo de Boiman, Shechtman e Irani (2008), onde são realçados os benefícios dos classificadores do tipo vizinho mais próximo e da representação NQ. A partir destas observações, os autores desenvolveram um classificador baseado nas características não-quantizadas, designado por *Naive Bayes Nearest Neighbour* (NBNN). À semelhança deste, o

método de localização proposto nesta tese também assenta na representação NQ mas, enquanto o NBNN combina as características num classificador do tipo Naive Bayes, no método proposto o problema é enquadrado na combinação de múltiplos classificadores e a regra da soma é usada. No capítulo 3 desta tese será apresentada a avaliação comparativa com outros métodos de combinação.

As principais contribuições deste capítulo são, em primeiro lugar, o método de localização baseado na representação NQ; em segundo lugar, a comparação do desempenho das representações Q e NQ no problema da localização global, a qual estava ausente da literatura anterior a este trabalho. Como terceira contribuição, apresenta-se a análise da discriminatividade das características nas representações Q e NQ. Esta análise estende a que foi apresentada por Boiman, Shechtman e Irani (2008), focando o problema de localização e considerando a influência dos parâmetros e configurações das duas representações.

O capítulo está organizado da seguinte forma: na secção 2.2 introduz-se o método de localização proposto e descrevem-se os classificadores baseados na representação Q que serão usados na avaliação; a secção 2.3 apresenta a análise de discriminatividade levada a cabo; na secção 2.4 são apresentados os *datasets* que serão usados neste capítulo e nos capítulos seguintes; a secção 2.5 analisa o desempenho dos classificadores baseados nas duas representações e compara-os em termos da sua precisão; finalmente, a secção 2.6 discute alguns dos resultados apresentados ao longo do capítulo.

2.2 Localização global usando as representações NQ e Q

A localização global pode ser entendida como um problema de classificação, em que se pretende estabelecer a correspondência entre os dados sensoriais recebidos pelo robô e um lugar existente num mapa do ambiente. Neste contexto, cada lugar é representado por uma classe, cujo modelo é construído a partir de imagens recolhidas numa fase inicial de exploração do ambiente. O resultado da classificação baseia-se na distribuição de probabilidades dos lugares dada a imagem actual, ou, mais genericamente, numa distribuição de pontuações atribuídas aos lugares. De seguida apresenta-se o método proposto para a estimação daquela distribuição de probabilidades utilizando a representação NQ e na secção 2.2.2 apresentam-se os métodos alternativos, baseados na representação Q.

2.2.1 Método proposto – representação NQ

Defina-se a aparência I da imagem de teste como um conjunto de nf descritores, $\{d_1,d_2,\ldots, d_{nf}\}$. Designando por L a variável aleatória que representa a localização do robô, esta é definida num mapa contendo np lugares conhecidos e descrito pelo conjunto de lugares $\{l_1,l_2,...,l_{np}\}$. O método proposto para a estimação das probabilidades dos lugares é baseado na inferência de Bayes, segundo a qual a distribuição posterior dada a característica d_i é expressa por

$$P(l_j|d_i) = \frac{P(d_i|l_j)P(l_j)}{P(d_i)} \tag{2.1}$$

onde

- $P(l_j|d_i)$ é a probabilidade do lugar j dada a observação da característica i;
- $P(d_i|l_j)$ é o modelo probabilístico das observações, designado por *função de verosimilhança;*
- $P(l_j)$ é a probabilidade à priori do lugar j;
- $P(d_i)$ é a probabilidade de observação da característica i.

O termo $P(d_i)$ é calculado por forma a que a probabilidade acumulada dos lugares seja igual a 1, através de

$$P(d_i) = \sum_j P(d_i|l_j)P(l_j). \tag{2.2}$$

O termo $P(l_j)$ reflecte o conhecimento prévio sobre a localização do robô e pode tomar o valor $1/np$, no caso de todos os lugares serem igualmente prováveis, ou ser retirado de uma distribuição não-uniforme, no caso de existir informação anterior à observação presente. No âmbito do estudo presente, consideraremos apenas a primeira situação, pois o foco é o uso da aparência visual e não a fusão com informação adicional.

Enquanto os termos $P(d_i)$ e $P(l_j)$ podem ser definidos da forma genérica apresentada acima, a função de verosimilhança é desenhada tendo em conta os tipos de características e modelos das classes usados pelo classificador. No caso presente, pretende-se estimar a probabilidade de ocorrência da característica d_i condicionada pelo lugar l_j, o qual é descrito pelo conjunto de nl_j características observadas durante a

fase de exploração do ambiente, $\{d_1^j, d_2^j, \ldots, d_{nl_j}^j\}$. No método proposto, esta estimação é feita através de um estimador de densidade de *Kernel*, segundo

$$P(d_i|l_j) = \frac{1}{nl_j} \sum_{m=1}^{nl_j} K(d_i, d_m^j) \tag{2.3}$$

onde $K(d_i, d_m^j)$ é uma função de *Kernel* que compara os descritores d_i e d_m^j. Aquela expressão pode ser sujeita a uma simplificação, sugerida no trabalho de Boiman, Shechtman e Irani (2008), onde se mostra empiricamente que uma boa aproximação daquela função pode ser obtida usando apenas o descritor d_m^j mais próximo de d_i. Esta aproximação, em que a verosimilhança de uma característica é calculada por

$$P(d_i|l_j) = \max_m K(d_i, d_m^j) \tag{2.4}$$

será também avaliada, com vista à simplificação do algoritmo de localização. Para além destas opções, o algoritmo é determinado pela escolha da função $K(\cdot)$ que, na estimação de densidade de *Kernel*, pode ser uniforme, gaussiana, triangular, entre outras (Wand e Jones, 1994). Duas funções pertinentes para o tratamento de características locais, a gaussiana e a de Weibull, serão avaliadas na secção 2.5.2.

Uma vez definidos todos os termos da Eq. (2.1), a apresentação do método proposto completa-se com a fusão dos resultados obtidos para todas as características da imagem de teste. Para esta operação, considerou-se que cada uma das distribuições à posteriori dadas pela Eq. (2.1) corresponde à saída de um classificador individual, colocando o problema de fusão no âmbito da combinação de múltiplos classificadores. Neste contexto, adoptou-se a regra da soma para a fusão de resultados, dada a sua reconhecida robustez, o que conduz à seguinte expressão:

$$P(l_j|I) = \frac{1}{nf} \sum_{i=1}^{nf} P(l_j|d_i). \tag{2.5}$$

Por fim, o método proposto adopta uma decisão do tipo Máximo à Posteriori, devolvendo o lugar que obteve maior probabilidade.

2.2.2 Métodos de localização baseados na representação Q

Nesta secção apresentam-se dois métodos, o classificador Naive Bayes e o classificador *Support Vector Machine* (SVM), que serão aplicados à representação Q e que

serão a base da comparação desta com a representação NQ. Cada um destes classificadores admite uma definição genérica, onde o caso do tratamento de características quantizadas é facilmente integrado. No estudo realizado, estes classificadores terão, portanto, em comum o facto de receberem como entrada informação que resulta da quantização das características visuais. Este é um mecanismo em que:

- Se usa um vocabulário de palavras visuais, construído através da aplicação do algoritmo *k-means* a um conjunto representativo de características;
- Todas as características usadas quer na modelação quer na localização são sujeitas à aproximação a uma palavra visual, a mais próxima existente no vocabulário. Assim, após a quantização, a representação de uma característica d_i passa a ser simplesmente o índice, w_i, que a palavra visual escolhida ocupa no vocabulário.

No caso do classificador SVM, as entradas são vectores que resumem o conjunto das características encontradas numa imagem. Uma vez que se trata aqui de características quantizadas, estes descritores tomam a forma de um histograma da ocorrência de palavras visuais na imagem.

2.2.2.a Naive Bayes

O classificador Naive Bayes é um método probabilístico onde o cálculo da probabilidade conjunta das características é simplificado, sob pressuposto de que as características são independentes dada a classe. Nestas condições, a verosimilhança do conjunto de características dada a classe *j* é obtida pelo produto das probabilidades individuais de cada característica (Csurka et al., 2004). Relembrando que, na representação Q, uma característica é substituída pela correspondente palavra visual, aquela distribuição pode ser expressa por

$$P(d_1, \dots, d_{nf} | l_j) = P(w_1, \dots, w_{nf} | l_j) = \prod_{i=1}^{nf} P(w_i | l_j). \qquad (2.6)$$

Apesar do forte pressuposto em que se baseia, é reconhecida a boa performance do classificador Naive Bayes, que tem sido observada mesmo em condições em que a independência das características não se verifica (Domingos e Pazzani, 1996). Por esta razão, este classificador é considerado competitivo em relação a métodos mais

complexos e muitas vezes usado como termo de comparação na avaliação de outros classificadores (Rennie et al., 2003).

Um trabalho de referência sobre a utilização do classificador Naive Bayes em conjunção com a representação Q é apresentado por Csurka et al. (2004), que se debruçam sobre vários problemas de classificação visual. Notando que a verosimilhança de uma característica pertence a uma distribuição categórica, naquele trabalho ela é estimada usando a técnica de suavização aditiva (Chen e Goodman, 1996), que será igualmente usada nesta tese e é calculada por:

$$P(w_i|l_j) = \frac{\alpha + n_i}{\alpha \cdot n_c + nl_j}. \qquad (2.7)$$

Nesta expressão, n_i é o número de ocorrências da palavra w_i nas imagens do modelo do lugar j, n_c é a dimensão do vocabulário e, como anteriormente, nl_j é o número de características extraídas no lugar j. A suavização da estimação através deste método, também designado por suavização de Laplace, é determinada pelo parâmetro α e tem por objectivo evitar probalidades iguais a zero, ao mesmo tempo que atenua erros resultantes da quantização.

2.2.2.b SVM

Os classificadores do tipo SVM são actualmente considerados dos mais bem sucedidos de entre o estado da arte no reconhecimento de padrões, o que se deve, entre outros aspectos, i) à sua capacidade de lidar com espaços de características de elevada dimensão, ii) à teoria bem fundamentada que os suporta e iii) à sua elevada precisão. Para além de exibirem estas propriedades, estes classificadores têm sido extensivamente usados na classificação visual com características quantizadas (Chapelle, Haffner e Vapnik, 1999; Jianguo et al., 2006; Jiang, Ngo e Yang, 2007), o que justifica a escolha deste modelo para a avaliação comparativa apresentada neste capítulo.

Segundo a formulação original do método SVM, este trata problemas de classificação binária, associando ao vector de teste *x'* a classe resultante de

$$f(x') = sign\left(\sum_{i=1}^{l}(\alpha_i y_i K(x_i, x') + b)\right). \qquad (2.8)$$

Nesta expressão x_i são os vectores de treino identificados como vectores de suporte, y_i designa a classe destes vectores, b é o offset da superfície de decisão relativamente à origem e α_i são os multiplicadores de Lagrange, obtidos por optimização de uma função de custo. No contexto da classificação não linear com o método SVM, $K(\cdot,\cdot)$ é uma função de *Kernel*, que avalia a similaridade entre dois vectores e respeita a condição de Mercer (Scholkopf e Smola, 2001).

A aplicação deste método à localização baseada na representação Q passa por i) definir os vectores x_i e x', como sendo histogramas de palavras visuais, correspondentes respectivamente a imagens de treino e de teste e ii) à extensão do método ao problema multi-classe, já que o número de lugares é geralmente superior a dois. Neste caso, é também possível recorrer a classificadores SVM binários, existindo para isso métodos baseados no treino de múltiplos classificadores binários (Chih-Wei e Chih-Jen, 2002). Entre as alternativas existentes, tem-se verificado que os resultados são semelhantes (Scholkopf e Smola, 2001; Jianguo et al., 2006). Neste trabalho usou-se o método que, num problema envolvendo n_p classes, treina n_p classificadores binários, cada um deles com o objectivo de distinguir uma classe de todas as restantes. Nesta abordagem, a saída dos classificadores é agora um valor de probabilidade, calculado pelo método proposto por Wu, Lin e Weng (2004) e o resultado da classificação consiste na classe cujo classificador binário produziu o valor mais alto de probabilidade.

O tipo de *Kernel* escolhido para a comparação de descritores é um dos factores que tem maior peso no desempenho dos classificadores SVM. Dada a sua importância, a comparação entre diferentes opções foi levada a cabo nos estudos de Jianguo et al. (2006) e Jiang, Ngo e Yang (2007), focando o caso particular da classificação visual baseada em características locais. Considerando os resultados destes estudos, onde o *Kernel* χ^2 apresenta o melhor desempenho (Jianguo et al., 2006) ou um desempenho entre os melhores (Jiang, Ngo e Yang, 2007), neste trabalho utilizou-se aquela função de *Kernel*. Este tipo de *Kernel* é uma instância das *Radial Basis Functions* (RBF) generalizadas, que se definem como (Chapelle, Haffner e Vapnik, 1999):

$$K_{RBF}(x, x') = e^{-\rho \cdot d(x,x')}. \qquad (2.9)$$

Na expressão anterior ρ, é um parâmetro de escala e $d(x,x')$ é uma medida de distância que, no caso do *Kernel* χ^2 é calculada por:

$$d(x, x') = \sum_i \frac{(x_i - \mathrm{x}')^2}{x_i + \mathrm{x}'}. \qquad (2.10)$$

2.3 Análise de discriminatividade

Nesta secção analisam-se as representações Q e NQ relativamente à discriminatividade patente nas características visuais, quando estas são descritas em cada uma daquelas representações. No que diz respeito à representação Q, a análise será feita no contexto do classificador Naive Bayes, por duas razões. Por um lado, tratando-se este de um classificador que não está sujeito a uma fase de aprendizagem, e que por isso não depende de pesos calculados automaticamente, permite uma análise simples sobre a contribuição de cada característica. Por outro lado, como se verá na secção 2.5, este classificador apresenta genericamente melhor desempenho que o classificador SVM, nas nossas experiências.

A discriminatividade é entendida como uma medida que quantifica a contribuição de uma característica para um resultado correcto na classificação. Dada esta definição genérica, esta medida pode ser calculada pela comparação da contribuição da característica para a probabilidade à posteriori do lugar correcto com a contribuição relativa ao lugar incorrecto mais bem classificado. Embora esta medida avalie a contribuição individual de cada característica, é importante que ela contemple a forma como as diversas contribuições são combinadas, por forma a relectir o desempenho global de um conjunto de características. Este aspecto exige formas de cálculo diferentes para diferentes classificadores, conduzindo, no presente trabalho, às seguintes expressões:

Representação NQ – Tendo em conta que a contribuição de múltiplas características é feita, segundo a Eq. (2.5), pela soma das probabilidades à posteriori associadas a cada característica, pode definir-se a medida de discriminatividade na representação NQ como:

$$discNQ_i = P(l_t|d_i) - P(l_{nt}|d_i) \qquad (2.11)$$

onde l_t designa o lugar correcto (*target*) e l_{nt} designa o candidato incorrecto (*non-target*) que obteve melhor classificação. Daquela expressão se conclui que esta medida de discriminatividade toma valores no intervalo [-1, 1].

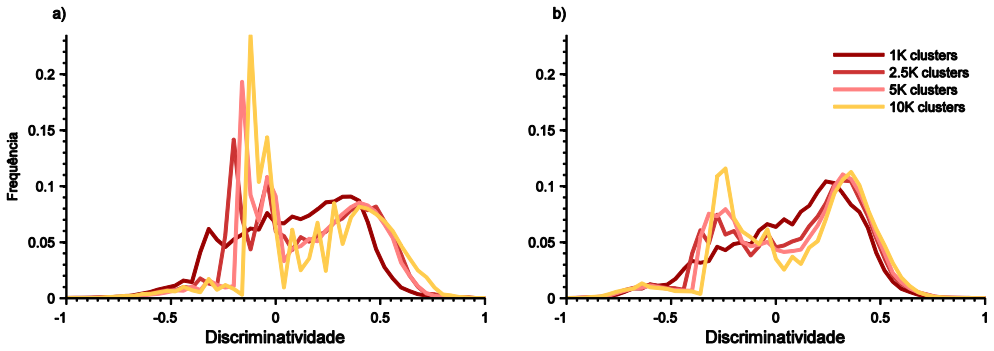


Figura 2.1. Perfis de discriminatividade na representação Q, para diferentes dimensões do vocabulário. À esquerda, α=1, à direita, α=0.05.

Representação Q – No caso do classificador Naive Bayes, a probabilidade de um lugar é dada pelo produto das probabilidades à posteriori associadas a cada característica. Uma vez que aqui as contribuições das diversas características são combinadas pelo produto, a discriminatividade deve ser calculada pela razão entre $P(l_t|d_i)$ e $P(l_{nt}|d_i)$. Por forma a facilitar a comparação das representações Q e NQ, aplicaremos o logaritmo àquele quociente e normalizaremos o resultado para o intervalo [-1,1]. Assim, o cálculo da discriminatividade é feito por

$$discQ_i = \frac{1}{sc}\Big(log\big(P(l_t|d_i)\big) - log\big(P(l_{nt}|d_i)\big)\Big) \qquad (2.12)$$

onde *sc* é um factor de normalização. A análise conduzida nesta secção baseia-se na distribuição da discriminatividade estimada sobre dados obtidos com o *dataset* IDOL, detalhado na secção 2.4. Para este efeito, foram seleccionados 4 vídeos de treino e 4 vídeos de teste, configurando assim 16 situações distintas de localização. Em cada uma destas situações, as imagens de teste foram processadas no sentido de calcular a discriminatividade das suas características na representação Q e NQ. Globalmente, um total de 801985 características foram consideradas. A estimação da distribuição de discriminatividade foi feita na forma de histogramas, que contabilizam a frequência de ocorrência de valores de discriminatividade em intervalos de 0.04 no domínio [-1,1].

2.3.1 Discriminatividade na representação Q

Nesta secção analisa-se o perfil de discriminatividade da representação Q, com o objectivo de identificar as limitações desta representação e o efeito que estas produzem para diferentes dimensões do vocabulário visual e valores do parâmetro α.

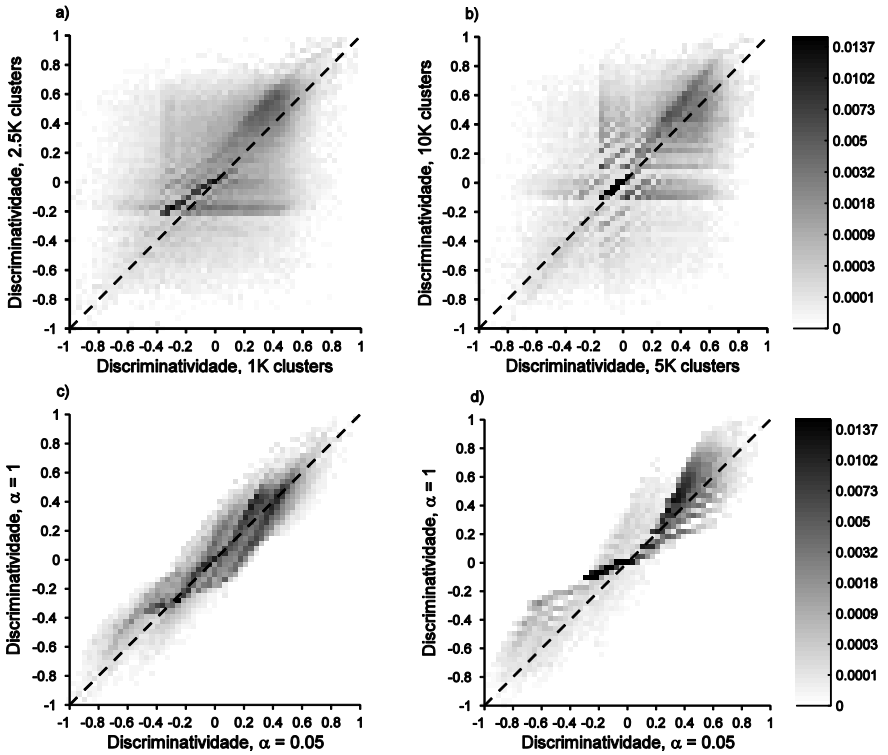


Figura 2.2. Distribuição conjunta da discriminatividade para diferentes combinações de parâmetros. Na primeira linha: α=1, a) comparação de n_c=2.5K vs n_c=1K, b) comparação de n_c=10K vs n_c=5K. Na segunda linha: c) comparação de α=1 vs α=0.05, com n_c=1K, d) comparação de α=1 vs α=0.05, com n_c=10K.

Com este objectivo, estimaram-se os perfis de discriminatividade com vocabulários de dimensão 1K, 2.5K, 5K e 10K e valores de α=1 e α=0.05. A informação sobre a discriminatividade de uma representação será ilustrada por dois tipos de gráficos: o primeiro descreve a distribuição de discriminatividade, estimada na forma de um histograma (Figura 2.1); o segundo é usado na comparação de duas configurações (diferentes dimensões do vocabulário ou diferentes valores de α) e descreve, num plano, a distribuição conjunta da discriminatividade das duas configurações, estimada por um histograma bidimensional (Figura 2.2). Este tipo de gráfico ilustra mais detalhadamente a forma como a discriminatividade é modificada quando se passa de uma configuração para outra. Por exemplo, a densidade de características acima da diagonal identifica aquelas cuja discriminatividade aumenta da primeira configuração (eixo x) para a segunda (eixo y), enquanto a densidade abaixo da diagonal corresponde a características cuja discriminatividade diminui.

É genericamente aceite que os vocabulários de dimensão reduzida limitam a precisão dos classificadores baseados na representação Q. Esta ideia assenta no facto de vocabulários mais extensos permitirem a emergência de palavras visuais mais

específicas, e por isso mais discriminativas, o que conduz às tendências verificadas nas Figuras 2.1.a, 2.2.a e 2.2.b. De facto, é possível observar, na Figura 2.1.a, que o aumento de n_c resulta em perfis de discriminatividade apresentando densidade significativa para valores mais próximos de +1. As Figuras 2.2.a e 2.2.b, onde no 1º quadrante há uma acumulação de densidade acima da diagonal, mostram que aquele efeito resulta de um ganho de discriminatividade sobre as características de discriminatividade positiva, obtido pelo incremento de n_c.

Para além do comportamento das características de discriminatividade alta, outro aspecto determinante no desempenho da representação Q prende-se com as características de discriminatividade próxima de zero, as quais têm forte relevância para n_c elevados (ver Figura 2.1.a). A ocorrência destas características é explicada pelo mecanismo descrito seguidamente. Com o aumento de n_c, é maior a probabilidade de qualquer palavra visual não constar no modelo de um lugar, já que as células de voronoi associadas aos *clusters* do vocabulário apresentam menor suporte. Consequentemente, é também maior a probabilidade de uma palavra visual que é observada na imagem de teste não ter sido observada nos lugares *t* nem *nt*. Quando isso acontece, o numerador da fracção do lado direito na Eq. (2.7) toma o mesmo valor, α, quer no cálculo de $P(l_t|d_i)$ como de $P(l_{nt}|d_i)$, e o valor de discriminatividade seria zero, caso o denominador fosse também igual para os dois lugares. No entanto, devido aos diferentes números de características nos modelos dos lugares, nl_t e nl_{nt}, aquelas probabilidades não são em geral iguais e os dados experimentais representados na Figura 2.1.a demonstram que a discriminatividade, em lugar de ser zero, é nestes casos geralmente negativa. A identificação deste mecanismo sugere um beneficio adicional no incremento de n_c, já que, para n_c superiores, a dependência do denominador da Eq. (2.7) relativamente a nl_j é atenuada, aproximando assim os valores de $P(l_t|d_i)$ e $P(l_{nt}|d_i)$. Este benefício é observado na Figura 2.1.a, onde se verifica que para os valores mais altos de n_c os picos de densidade na vizinhança negativa de zero se aproximam de zero.

As Figuras 2.1.b, 2.2.c e 2.2.d, onde se apresentam dados relativos à aplicação de α=0.05 revelam que as conclusões anteriores não se estendem para toda a gama de α. De facto, a Figura 2.1.b mostra que o aumento de n_c pode não ser benéfico, conduzindo, neste caso à acumulação de densidade em valores negativos e afastados da origem. Dado o exposto anteriormente, este efeito é explicado pelo facto de o

número de características não observadas nos lugares *t* e *nt* aumentar com n_c. Neste caso a discriminatividade destas características não se concentra próximo de zero pois, para valores reduzidos de α, o peso do termo $\alpha.n_c$ é menor no denominador de (2.7) e por isso atenua menos eficazmente as diferenças entre nl_t e nl_{nt}.

A comparação entre a discriminatividade obtida para α=0.05 e α=1 é fornecida nas Figuras 2.2.c e 2.2.d que representam a distribuição conjunta respectivamente para os vocabulários n_c =1K e n_c =10K. Lembrando que reduzir o parâmetro α é equivalente a atenuar o efeito da suavização aditiva, a Figura 2.2.d ilustra o efeito benéfico da suavização, evidenciando que para α reduzido as características de discriminatividade negativa tomam valores mais próximos de -1. É assim evidente que para vocabulários de dimensão alta a aplicação da suavização é fundamental, para evitar probabilidades próximas de zero e, genericamente, porque a estimação de probabilidades nestes casos é menos precisa. Contrastando com este cenário, os vocabulários de dimensão reduzida, como é o caso de n_c = 1K, não estão sujeitos à mesma imprecisão e o efeito da suavização pode implicar alguma perda de informação. A Figura 2.2.c sugere, de facto, que pode haver algum beneficio na redução da suavização nestes casos, dada a acumulação de densidade abaixo da diagonal.

2.3.2 Discriminatividade na representação NQ

Neste ponto examinam-se os perfis de discriminatividade da representação NQ, colocando a ênfase na comparação destes perfis com os da representação Q e na avaliação da aproximação da estimação de densidade através da Eq. (2.4).

No que diz respeito à simplificação da estimação de densidade por *Kernel*, a Figura 2.3 sugere que as propriedades da representação NQ são semelhantes, com e sem simplificação. Uma perspectiva mais detalhada da relação entre as duas técnicas é fornecida na Figura 2.4.a. Esta figura mostra que, embora se verifique alguma divergência entre a discriminatividade das duas representações, a densidade desta distribuição é aproximadamente simétrica relativamente à diagonal, o que justifica a semelhança entre os dois perfis de discriminatividade na Figura 2.3 e sugere que o seu desempenho será idêntico.

Na comparação dos perfis de discriminatividade da representação Q e NQ, os dados das Figuras 2.3 e 2.4.b evidenciam que estas representações diferem significativa-

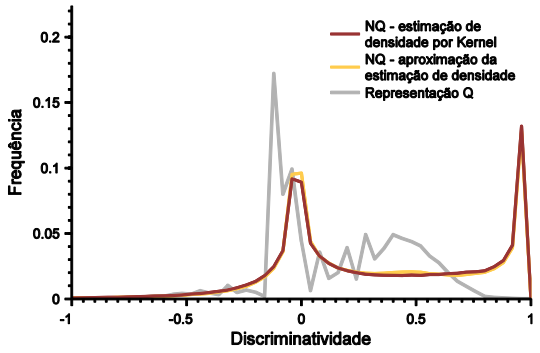


Figura 2.3. Perfis de discriminatividade na representação NQ e comparação com a representação Q, com α=1 e n_c =10K.

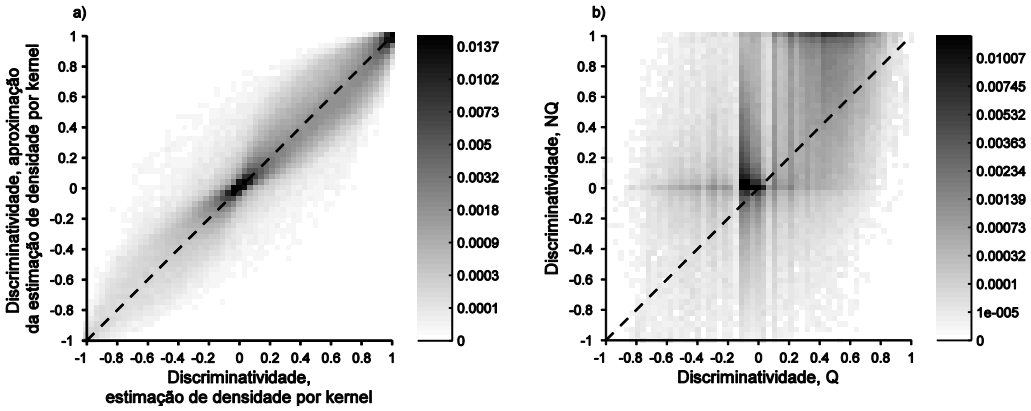


Figura 2.4. Distribuição conjunta da discriminatividade. A) representação NQ, comparação da estimação de densidade com e sem aproximação, b) comparação das representações Q e NQ.

mente nas suas propriedades. Em particular, a grande dispersão de densidade no plano de distribuição conjunta (Figura 2.4.b) sugere que não há uma relação simples entre as duas representações. Da análise das figuras, a representação NQ destaca-se por:

- Permitir a representação de um elevado número de características com discriminatividade máxima, ou próxima de 1;
- Representar um menor número de características de discriminatividade próxima de zero. Como foi mencionado no ponto anterior, a representação Q produz um elevado número de características de discriminatividade próxima de zero, que ocorrem quando uma palavra visual não foi observada nos lugares *t* nem *nt*. A Figura 2.4.b mostra que, na representação NQ, a discriminatividade destas características se distribui por toda a gama [-1,1], no entanto a sua densidade é superior no intervalo positivo.

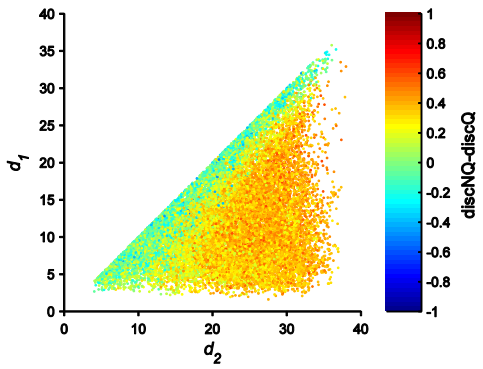


Figura 2.5. Gráfico de dispersão das características no plano d_1 vs d_2.

Apesar de, como se disse, não existir uma relação simples entre a discriminatividade das duas representações, é possível identificar as condições em que a representação NQ não produz benefícios relativamente à representação Q. Com esse objectivo, defina-se d_1 e d_2 como as menores distâncias encontradas entre a característica de teste e as características dos lugares l_t e l_{nt}, ou seja, as distâncias que determinam o valor da discriminatividade na representação NQ. Na Figura 2.5 apresenta-se o gráfico de dispersão das características no plano d_1 vs d_2, para as características no primeiro quadrante da Figura 2.4.b. O valor da diferença de discriminatividade entre as duas representações é ilustrado pela cor atribuída aos pontos no gráfico. Da observação da figura conclui-se que as características para as quais a diferença de discriminatividade é reduzida ou negativa apresentam valores idênticos de d_1 e d_2 (posições próximas da diagonal na Figura 2.5). Este dado indica que, quando as características dos lugares são quase equidistantes da característica de teste, o estabelecimento de correspondências com valores binários (representação Q) pode ser mais distintiva do que com valores contínuos (representação NQ), pois neste caso os dois lugares recebem valores de correspondência semelhantes.

2.4 Datasets

2.4.1 *Dataset* IDOL

O *dataset* KTH-IDOL2 (Luo et al., 2006), doravante designado simplesmente por IDOL (*Image Database for rObot Localization*), foi criado pelo grupo Computational Vision and Active Perception Laboratory do KTH Royal Institute of Technology, com o objectivo de avaliar a robustez e a adaptabilidade de algoritmos de reconhecimento visual, quando aplicados a ambientes dinâmicos.

Tabela 2.1. Características principais dos *datasets* usados na avaliação.

Dataset	Dimensões do ambiente [m]	Amostragem	Sub-amostragem	Nº médio de imagens por sequência	Formato de imagem
IDOL	10.5×19.6	5 fps	-	953	240×320 RGB
FDF Park	137.2×178.3	30 fps	7.5 fps	9052	240×320 RGB

O *dataset* contém 24 sequências de imagens, captadas por dois robôs móveis, cada uma correspondendo a uma passagem pelo ambiente, e todas descrevendo aproximadamente a mesma trajectória (na Tabela 2.1 compilam-se alguns dos dados relevantes sobre os *datasets* usados). As imagens descrevem um ambiente de interior, especificamente o laboratório daquele grupo de investigação, composto por um corredor, dois gabinetes, uma cozinha e uma zona de impressoras (na Figura 2.6 apresentam-se imagens típicas deste *dataset*). As variações de aparência que se verificam entre as diferentes sequências de imagens têm diversas origens: variações de perspectiva, resultantes de desvios de posição na captura de imagem, mudança de posição de objectos, presença/ausência de pessoas e, sobretudo, variações de luminosidade. De maneira a registar as variações de luminosidade que ocorrem em ambientes reais, as sequências de imagens foram recolhidas em diferentes alturas do dia e espaçadas no tempo por alguns meses. Referindo-se às diferentes condições de luminosidade, os autores do *dataset* reuniram as sequências de imagens em conjuntos, que identificaram com os termos '*cloudy*' '*sunny*' e '*night*'. Para cada uma destas condições estão disponibilizadas 4 sequências, totalizando 12 sequências para cada uma das plataformas robóticas. Dada a semelhança entre os dados capturados por cada uma das plataformas, neste trabalho usou-se apenas o *dataset* referente à plataforma *PowerBot Dumbo*.

A cada uma das imagens neste *dataset* está associada uma posição do robô no espaço métrico, estimada com o auxílio de um dispositivo *laser rangefinder*. A existência deste tipo de informação permitiu que, neste *dataset*, quer a modelação do ambiente quer o cálculo da precisão se baseassem em informação métrica. No respeitante à modelação do ambiente, o objectivo é dividir uma sequência de imagens por forma a que cada subconjunto corresponda a um lugar distinto. Designando por x_j e θ_j respectivamente a posição e o ângulo representativos do lugar j, e x_i, θ_i as mesmas variáveis para uma imagem i, o mecanismo usado na modelação do ambiente descreve-se como:

- O modelo é inicializado com um lugar apenas, com x_j, θ_j tomando os valores da primeira imagem da sequência,
- Para cada nova imagem i é verificada a condição de esta pertencer ao último lugar j que foi inicializado. Se a condição for verdadeira x_j, θ_j são actualizados, como a média dos valores de todas as imagens que foram associadas a esse lugar, incluindo a presente; se a condição não se verificar, é inicializado um novo lugar, com os valores da imagem i.
- Por forma a obter um mapa topológico conciso, foi definida empiricamente a condição seguinte como limite para a inclusão de uma imagem: uma imagem com coordenadas métricas x_i, θ_i pertence ao último lugar inicializado se

$$\begin{cases} \|x_i - x_j\| < 3m \\ |\theta_i - \theta_j| < 55° \end{cases}. \quad (2.13)$$

Como mencionado acima, a informação métrica foi usada também para o cálculo da precisão neste *dataset*. A este respeito, uma classificação será considerada correcta se verificar o seguinte teste: para a classe devolvida pelo classificador, é determinada a imagem do modelo desta classe que é mais próxima da imagem de teste, e o resultado será positivo se as distâncias linear e angular entre as duas imagens verificarem os mesmos limites da Eq. (2.13).

2.4.2 *Dataset* FDF Park

O *dataset* FDF Park foi usado no passado nos estudos de localização de robôs de Siagian e Itti (Siagian e Itti, 2007; Siagian e Itti, 2009) e, contrastando com o *dataset* IDOL, retrata um ambiente de exterior, especificamente o campus da University of Southern California. Este ambiente é representado por um conjunto de 9 lugares distintos, sendo disponibilizadas várias sequências de imagens para cada um dos lugares, captadas por uma câmara transportada manualmente. Pelo facto de frequência de captura ser alta para a velocidade de deslocamento em causa, os vídeos foram sub-amostrados de acordo com os dados da Tabela 2.1. Para o efeito de avaliação de sistemas de localização, estas sequências foram agrupadas, pelos autores do *dataset*, em 8 conjuntos para modelação do ambiente e 7 conjuntos de teste. A principal fonte de variabilidade registada neste *dataset* é a luminosidade, já que este foi recolhido em diferentes dias e horas do dia (ver Figura 2.7 com exemplos de imagens típicas deste

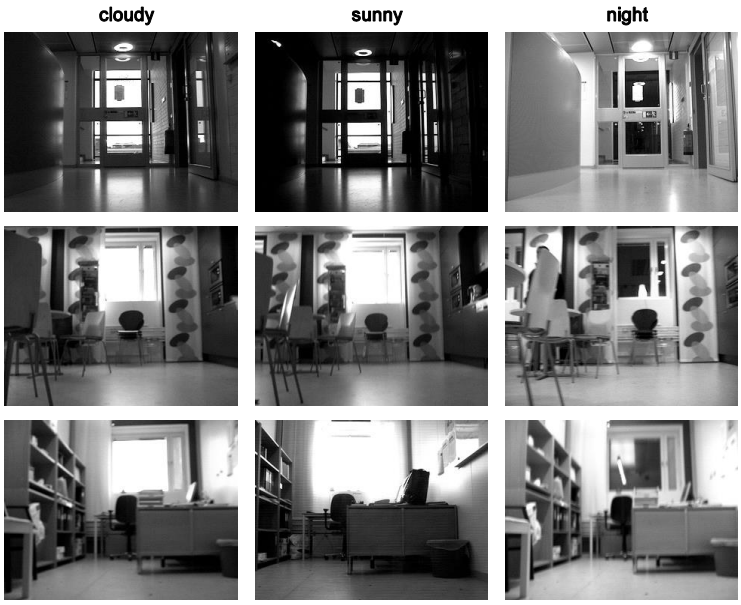


Figura 2.6. Exemplos de imagens do *dataset* IDOL. Da esquerda para a direita as colunas dizem respeito às condições cloudy, sunny e night. Cada uma das linhas apresenta imagens do mesmo lugar.

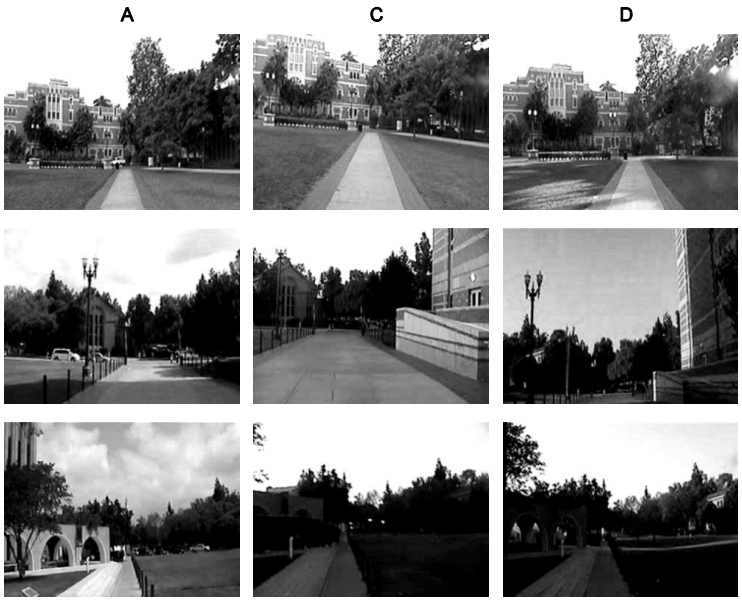


Figura 2.7. Exemplos de imagens do *dataset* FDF Park. Da esquerda para a direita as colunas dizem respeito às sequências de modelação, A, B e C. Cada uma das linhas apresenta imagens do mesmo lugar.

dataset). Tendo em conta os conjuntos de modelação e teste disponibilizados, várias condições de avaliação podem ser criadas. As nossas experiências preliminares revelaram que os conjuntos de modelação A, C e D são os que colocam condições de localização mais adversas e por isso foram seleccionados para a avaliação das abordagens em estudo neste trabalho.

Contrariamente ao *dataset* IDOL, este *dataset* não dispõe de informação métrica sobre as posições de captura das imagens. Por esta razão, na modelação do ambiente e na avaliação dos classificadores fez-se uso da partição original do ambiente em 9 lugares. Em concreto, foram usados modelos do ambiente com um número superior de lugares, obtidos pela divisão das sequências de imagens de cada um dos lugares originais em várias sub-sequências. Neste processo, em que se impôs que cada subsequência contivesse 30 imagens, pretendeu-se evitar modelos de lugares de elevada dimensão. A dimensão dos modelos dos lugares como factor no desempenho dos classificadores é um aspecto que será discutido no capítulo 3 desta tese.

No que diz respeito à avaliação de resultados, a inexistência de informação métrica impede o estabelecimento de correspondências entre os sub-lugares criados em diferentes sequências de imagens, pelo que aqui se recorreu à partição original. Assim, considerou-se que um resultado é correcto se o sub-lugar devolvido pelo classificador pertence ao lugar original onde foi recolhida a imagem de teste.

2.5 Precisão das representações Q e NQ

Nesta secção avaliam-se os classificadores baseados nas representações Q e NQ, em termos da sua precisão no problema de localização, definida como a percentagem de imagens correctamente classificadas num conjunto de imagens de teste. Em cada avaliação é considerada uma sequência de imagens de exploração do ambiente e uma sequência de imagens de uma travessia de teste. A partir da primeira é construído o modelo do ambiente, no caso dos classificadores probabilísticos ou, no caso do classificador SVM, são ajustados por treino os parâmetros do classificador. A precisão é medida aplicando estes classificadores a cada uma das imagens da sequência de teste e calculando a percentagem de imagens correctamente classificadas. Com o objectivo de apresentar os resultados de forma mais concisa, nalguns casos será apresentada a média das precisões obtidas para várias condições de modelação/teste.

Na secção 2.5.1 é avaliado o desempenho dos classificadores baseados na representação Q e a sua dependência relativamente à dimensão do vocabulário visual. A secção 2.5.2 foca a configuração do classificador NQ, abordando a selecção da função de *Kernel*, a sua parameterização e a simplificação da estimação de densidade por *Kernel*. Nesta secção compara-se ainda a precisão das duas representações.

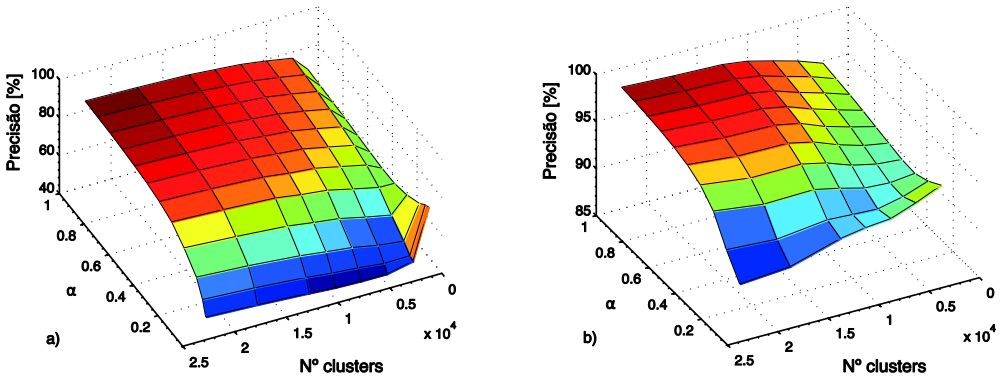


Figura 2.8. Precisão média obtida com o classificador Naive Bayes em função dos parâmetros α e n_c. À esquerda, resultados do *dataset* IDOL; à direita resultados do *dataset* FDF Park.

2.5.1 Representação Q

2.5.1.a Impacto dos parâmetros do classificador Naive Bayes

A Figura 2.8, que representa a precisão média obtida pelo classificador Naive Bayes nos dois *datasets*, em função do parâmetro α e da dimensão do vocabulário, mostra que o melhor desempenho é obtido para α=1 e para os vocabulários de dimensão mais alta. A tendência de melhoria da precisão com o aumento de n_c é genérica, com as excepções localizando-se apenas nos valores mais baixos de α. Com efeito, para estes valores verifica-se que a precisão é maior com o vocabulário mais reduzido. Este facto é consistente com as conclusões da secção 2.3.1, segundo as quais, para valores de α reduzidos, a dependência da estimação pela Eq. (2.7) relativamente ao número de características nos modelos prejudica o desempenho do classificador. Verifica-se também que, para estes valores de α, o pico de desempenho se encontra no valor mínimo de n_c nos dois *datasets*, no entanto estes apresentam tendências ligeiramente diferentes quando n_c aumenta. Esta diferença dever-se-á ao facto de os modelos do *dataset* IDOL serem compostos de lugares com grande disparidade no número de características, o que origina uma queda muito acentuada de desempenho imediatamente após o pico referido.

2.5.1.b Classificador Naive Bayes vs SVM

As Figuras 2.9 e 2.10 comparam a precisão obtida pelos classificadores Naive Bayes e SVM em função da dimensão do vocabulário visual, respectivamente nos *datasets* IDOL e FDF Park. Para essa avaliação foram considerados vocabulários com n_c entre 2.5K e 22.5K. Estas figuras mostram que o classificador Naive Bayes tem gene-

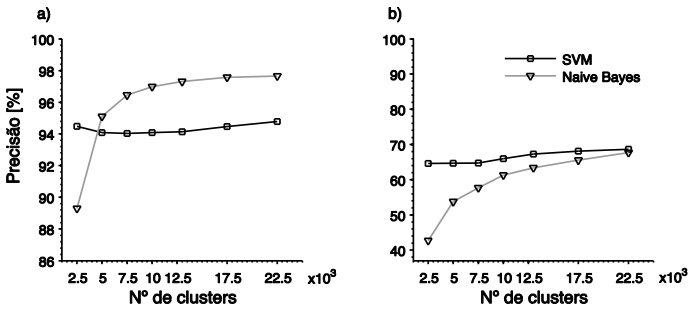


Figura 2.9. Precisão dos classificadores Naive Bayes e SVM no *dataset* IDOL. À esquerda, resultados obtidos nas combinações de luminosidade mais próximas, à direita resultados para as combinações de luminosidade mais adversas.

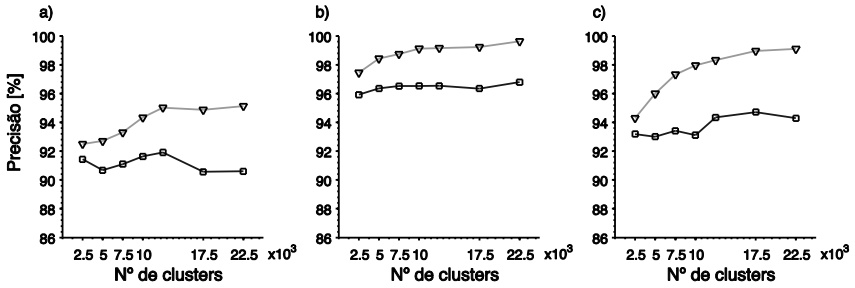


Figura 2.10. Precisão dos classificadores Naive Bayes e SVM no *dataset* FDF Park. A) sequência de treino A, b) sequência de treino C e c) sequência de treino D.

ricamente um desempenho superior, com excepção dos casos, no *dataset* IDOL, em que há variações de luminosidade severas. O bom desempenho do classificador Naive Bayes, notável dada a sua simplicidade, encontra uma explicação plausível nos argumentos avançados por (Boiman, Shechtman e Irani, 2008) na sua defesa dos classificadores do tipo 'vizinho mais próximo'. Segundo estes autores, a modelação das classes por imagens é prejudicial, por não facilitar a generalização na classificação. A forma de modelação das classes é precisamente uma das distinções fundamentais entre os dois classificadores: enquanto o classificador SVM recebe como exemplos de treino histogramas que caracterizam cada imagem individualmente, no classificador Naive Bayes o modelo de classes que está inerente à Eq. (2.7) é um conjunto de características indiscriminadas relativamente às imagens em que foram captadas. Como apontado por Rematas, Fritz e Tuytelaars (2013), o último tipo de modelo permite que, na comparação de uma imagem de teste com uma classe, a imagem de teste possa ser vista como uma composição de características observadas em diferentes imagens dessa classe. Esta propriedade, apesar de assumir a independência das características e daí incorrer nalguma perda de informação,

nalguns casos beneficia o processo de classificação, por promover a generalização (Rematas, Fritz e Tuytelaars, 2013).

Outro aspecto que diferencia as duas abordagens, e que poderá contribuir para o bom desempenho relativo do classificador Naive Bayes, prende-se com a forma como cada uma das abordagens trata as características encontradas no modelo da classe mas não observadas na imagem de teste. Enquanto o classificador Naive Bayes apenas inclui, no calculo da probabilidade à posteriori (Eq. (2.6)), as características observadas na imagem de teste, o classificador SVM compara os histogramas de palavras visuais em toda a sua dimensão pela distância χ^2 (Eq. (2.10)), a qual é maior se as características na imagem de treino não estão presentes na imagem de teste. Este dado pode ser a razão por que não se verifica um incremento consistente de desempenho no classificador SVM com o aumento da dimensão do vocabulário, já que, com este aumento, diminui a probabilidade de as características de teste e de treino serem atribuídas às mesmas palavras visuais.

2.5.2 Representação NQ

2.5.2.a Selecção de parâmetros da estimação de densidade por *Kernel*

Na estimação da verosimilhança, através da estimação de densidade por *Kernel* (Eq. (2.3)) ou da sua simplificação (Eq. (2.4)), foram consideradas duas funções de *Kernel* candidatas: a distribuição gaussiana e a distribuição de Weibull. A primeira foi adoptada por ser a de uso mais comum neste método de estimação, enquanto a segunda foi escolhida por ter sido teoricamente demonstrado por Burghouts, Smeulders e Geusebroek (2007) que a função de Weibull descreve a distribuição de distâncias entre características visuais quando medidas por normas-L_p. A selecção de parâmetros para estas funções foi realizada empiricamente, pela avaliação da precisão em gamas adequadas de valores de cada um deles. No caso da função gaussiana, o valor da função de *Kernel* entre os descritores d_1 e d_2 é calculado por

$$K(d_1.d_2) = exp\left(-\frac{\|d_1 - d_2\|^2}{2\sigma^2}\right) \qquad (2.14)$$

onde σ é a largura de banda do *Kernel*; no caso da função de Weibull aquele valor é dado pela expressão:

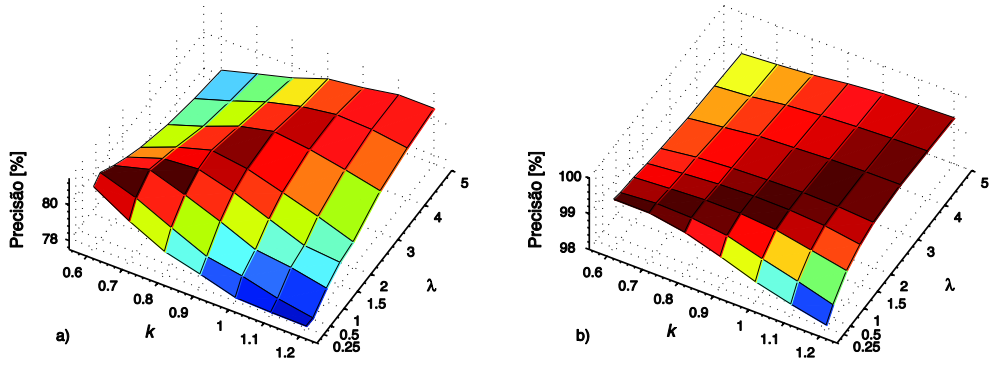


Figura 2.11. Precisão média como função dos parâmetros do *Kernel* de Weibull. À esquerda, resultados do *dataset* IDOL; à direita, do *dataset* FDF Park.

$$K(d_1.d_2) = \left(\frac{\|d_1 - d_2\|}{\lambda}\right)^{k-1} exp - \left(\frac{\|d_1 - d_2\|}{\lambda}\right)^{k}. \qquad (2.15)$$

Nesta expressão k é o parâmetro de forma da função e λ o parâmetro de escala.

As Figuras 2.11.a e 2.11.b representam a precisão média obtida com o *Kernel* de Weibull, em função dos parâmetros k e λ. Como se pode observar, a precisão máxima obtida para k=1 é aproximadamente igual à máxima encontrada em toda a gama de k considerada, a qual ocorre para valores de k abaixo de 1. Como no caso de k=1 a função de Weibull se reduz à função exponencial, de cálculo mais simples, doravante considerar-se-á apenas o caso de k=1.

Com vista à análise das diferentes configurações da estimação de densidade por *Kernel*, na Figura 2.12 apresenta-se a precisão média obtida nos dois *datasets* com os *Kernel*s gaussiano e exponencial. Na figura incluem-se também os resultados obtidos quando se aplica a estimação com e sem aproximação. Nesta figura é em primeiro lugar relevante a superioridade da estimação de probabilidades com aproximação. Os benefícios deste método encontram-se sobretudo na sua maior precisão, a que se acrescenta uma menor sensibilidade à selecção de parâmetros do que na estimação sem aproximação. Estes dados, além de validarem o método com aproximação, sugerem algumas propriedades da localização baseada na representação NQ. Na estimação de probabilidades sem aproximação, o valor de probabilidade inclui, para além da contribuição da característica do modelo de um lugar mais próxima da característica de teste, contribuições significativas das características abrangidas pela

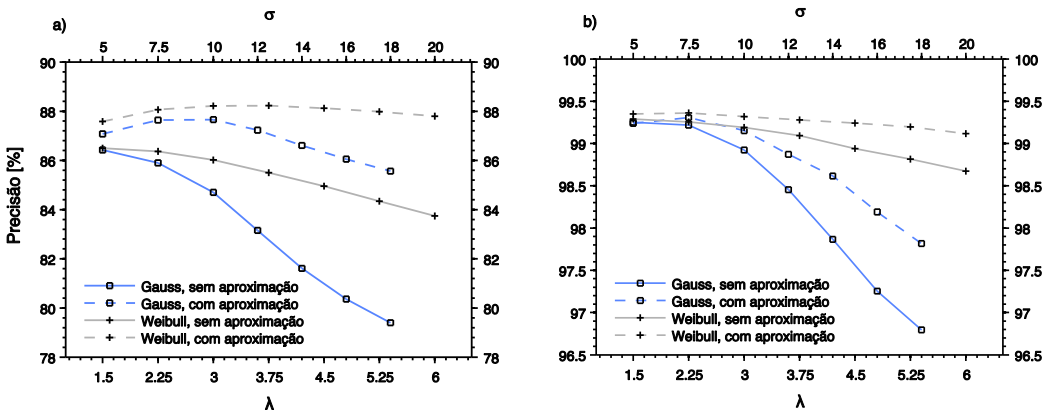


Figura 2.12. Precisão média obtida com as duas funções de *Kernel* como função dos seus parâmetros. À esquerda, resultados do *dataset* IDOL; à direita, do *dataset* FDF Park.

largura de banda do *Kernel*. Pode dizer-se que, neste caso, contabiliza-se a multiplicidade de características do modelo próximas da característica de teste, enquanto na estimação com aproximação considera-se apenas a semelhança visual da característica mais próxima. Dada a superioridade do último método, os resultados sugerem que a multiplicidade das características é uma propriedade pouco discriminativa, quando comparada com a simples semelhança entre descritores. Deve notar-se também que no método de localização usado, a multiplicidade de características pode não resultar da sua ocorrência na mesma imagem, o que contribuiria para uma caracterização do lugar, mas da sua ocorrência em múltiplas imagens, sendo resultado da forma como um lugar é amostrado. De facto, lembrando que o modelo proposto para os lugares na representação NQ não associa as características a uma imagem particular, a multiplicidade de características pode dever-se à sua detecção em imagens sucessivas, dependendo então do posicionamento do robô entre diferentes capturas de imagem, factor que não é descritivo do lugar mas da forma como é explorado.

Os resultados da Figura 2.12 mostram ainda que o *Kernel* exponencial oferece maior precisão do que o *Kernel* gaussiano, reforçando a ideia de que a família de funções de Weibull se adequa melhor à descrição da distribuição de distâncias entre descritores. Os valores máximos de desempenho encontrados nos dois *datasets* verificam-se para valores ligeiramente diferentes dos parâmetros de *Kernel*. Por forma a garantir um desempenho próximo do máximo para ambos, seleccionou-se como valor de compromisso λ =3, que será usado daqui em diante.

2.5.2.b Comparação da precisão das duas representações

Com vista à comparação da precisão das representações NQ e Q, nas Tabelas 2.2 a 2.4 apresenta-se a precisão medida na representação NQ, com *Kernel* exponencial, λ=3, e aproximação da estimação de densidade, e na representação Q, aplicando os classificadores Naive Bayes e SVM. Para estes casos utilizou-se a dimensão mais elevada de vocabulários considerada, com n_c=22.5K, e, no caso particular do classificador Naive Bayes, α=1.

A Tabela 2.2 colige os resultados relativos ao *dataset* IDOL. Uma vez que neste *dataset* as sequências de imagens estão agrupadas por tipo de luminosidade, naquela tabela apresenta-se a precisão média obtida sobre todas as combinações de sequências que correspondem a uma dada configuração de luminosidades na modelação e no teste. O *dataset* FDF Park, por outro lado, não agrupa as sequências de imagens por condições de captura, pelo que se apresentam os resultados individuais para cada par de sequências modelo/teste deste *dataset* (Tabelas 2.3 a 2.5).

Nos resultados apresentados é notória a superioridade da representação NQ, que obteve maior precisão em todos os casos, excepto um. O *dataset* IDOL é aquele onde se encontram as maiores diferenças de desempenho, que ocorrem nas situações de localização mais difíceis, *i.e.*, em que as condições de luminosidade na modelação e no teste são mais díspares. Nestes casos o desempenho de todos os classificadores cai, contudo, o classificador baseado na representação NQ supera o classificador Naive Bayes em cerca de 8 pontos percentuais no par *cloudy*/*night* e de 12 pontos percentuais no par *night/sunny*.

Os resultados dos vários classificadores no *dataset* FDF Park evidenciam que este coloca situações de localização menos exigentes do que o *dataset* IDOL. Também aqui as maiores diferenças de desempenho coincidem com os casos em que a precisão global de todos os classificadores diminui. Estas situações verificam-se na sequência de modelação A, onde a representação NQ supera o classificador Naive Bayes num máximo de cerca de 5 pontos percentuais.

Nas Figuras 2.13 e 2.14 complementam-se os resultados já apresentados com dados sobre a distribuição da precisão pelos lugares de um modelo. Para esse fim, considerou-se cada um dos lugares de um modelo do ambiente e determinou-se a

Tabela 2.2. Precisão [%] dos diferentes classificadores no *dataset* IDOL.

Condições de modelação	Condições de teste	SVM	Naive Bayes	Representação NQ
Cloudy	*Cloudy*	94.93±2.86	98.16±1.34	**98.29±1.49**
	Sunny	93.32±5.18	97.25±2.92	**97.64±2.63**
	Night	77.55±7.57	77.84±7.75	**85.46±7.54**
Sunny	*Cloudy*	94.92±3.19	97.37±2.27	**97.71±2.54**
	Sunny	96.47±2.35	**98.49±1.28**	98.3±1.50
	Night	62.53±7.13	59.74±9.75	**67.07±6.95**
Night	*Cloudy*	75.6±9.50	78.02±5.97	**87.19±5.83**
	Sunny	58.95±10.78	60.00±11.71	**71.54±8.44**
	Night	94.29±4.4	97.83±1.74	**98.3±1.13**

Tabela 2.3. Precisão [%] dos diferentes classificadores no *dataset* FDF Park, sequência de treino A.

Conjunto de teste	SVM	Naive Bayes	Representação NQ
A	94.64	98.08	**100.00**
B	83.91	91.22	**96.71**
C	95.99	97.96	**99.45**
D	95.73	97.27	**98.80**
H	83.56	92.05	**96.93**
I	85.26	92.01	**94.80**
J	93.85	94.64	**97.79**
K	91.92	97.76	**100.00**

Tabela 2.4. Precisão [%] dos diferentes classificadores no *dataset* FDF Park, sequência de treino C.

Conjunto de teste	SVM	Naive Bayes	Representação NQ
A	93.03	99.62	**99.85**
B	93.12	99.41	**100.00**
C	100.00	100.00	**100.00**
D	100.00	100.00	**100.00**
H	93.08	98.03	**100.00**
I	99.77	100.00	**100.00**
J	100.00	100.00	**100.00**
K	95.36	99.91	**100.00**

Tabela 2.5. Precisão [%] dos diferentes classificadores no *dataset* FDF Park, sequência de treino D.

Conjunto de teste	SVM	Naive Bayes	Representação NQ
A	80.92	96.25	**99.77**
B	94.51	99.20	**100.00**
C	97.80	99.61	**99.84**
D	98.72	100.00	**100.00**
H	93.78	98.82	**100.00**
I	100.00	100.00	**100.00**
J	99.76	99.76	**99.68**
K	88.74	99.14	**100.00**

respectiva precisão média, calculada sobre os resultados de todas as sequências de teste. Nas Figuras 2.13.a a 2.13.c apresentam-se os resultados para 3 sequências de modelação do *dataset* IDOL, respectivamente nas condições *cloudy*, *sunny* e *night*; a Figura 2.14 mostra os resultados para a sequência de modelação A do *dataset* FDF Park.

Estas figuras mostram, antes de mais, que a dificuldade na localização distribui-se de forma pouco uniforme pelos diferentes lugares, em virtude de as variações de luminosidade afectarem diferentemente cada um deles. No caso do *dataset* de interior, há a acrescentar que a disposição das estruturas arquitectónicas e objectos em torno do robô determina a forma como as mudanças de perspectiva influenciam a aparência observada. Há por isso a considerar que os lugares são também diferentemente afectados, neste *dataset*, pelos desvios de posição do robô entre as fases de modelação e de teste.

À semelhança do que foi constatado relativamente à precisão global, as diferenças na precisão por lugar sugerem a superioridade da representação NQ. Contudo, os ganhos obtidos por esta representação são significativamente variáveis sendo que num número considerável de casos a representação Q apresenta maior precisão. Este dado sugere que as especificidades dos lugares, tais como o número de imagens de modelação, presença de ruído, influência de *aliasing*, podem ter respostas diferentes por cada uma das representações.

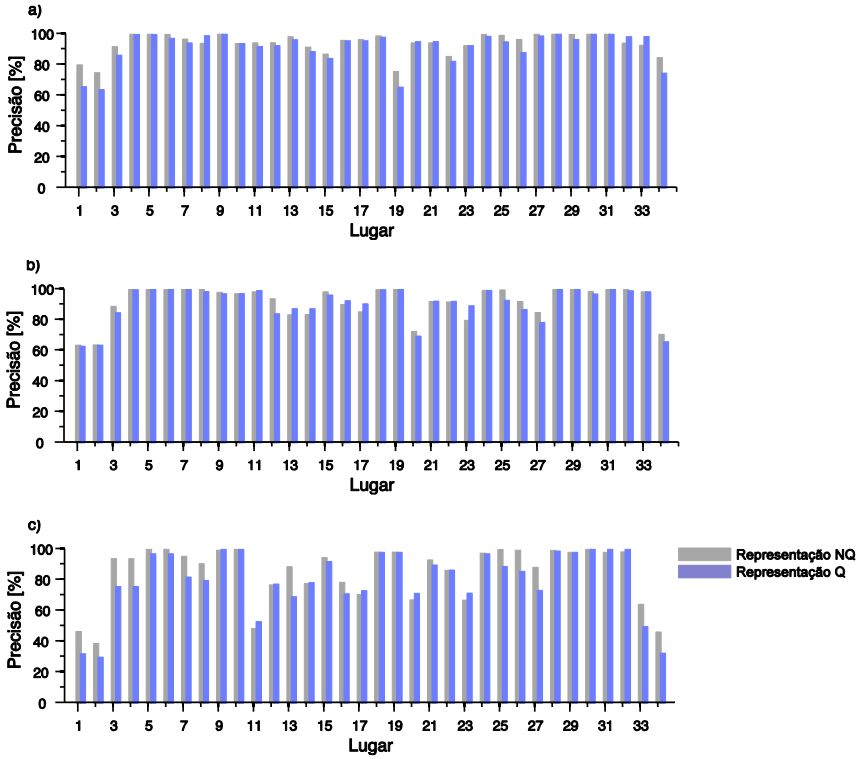


Figura 2.13. Precisão média por lugar no *dataset* IDOL. Resultados sobre 3 modelos, representativos das condições a) cloudy, b) sunny e c) night, respectivamente.

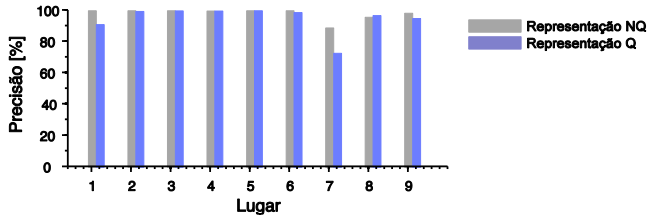


Figura 2.14. Precisão média por lugar, medida para a sequência de treino A do *dataset* FDF Park.

2.6 Discussão

A investigação apresentada neste capítulo encontra um paralelo, na área da classificação visual genérica, no trabalho de Boiman, Shechtman e Irani (2008). Nesse trabalho, os autores desenvolvem um classificador que é igualmente baseado em características não quantizadas, mas em que estas são combinadas num classificador do tipo Naive Bayes. Apesar das diferenças entre aquele classificador e o proposto nesta tese, algumas observações são suscitadas pela comparação entre os resultados de Boiman, Shechtman e Irani (2008) e os do presente trabalho.

Em primeiro lugar, a análise da discriminatividade em (Boiman, Shechtman e Irani, 2008) sugere que a aproximação da estimação de densidade por *Kernel* introduz uma

perda de desempenho na classificação. Contrastando com aquela conclusão, a análise da discriminatividade apresentada na secção 2.3.2 é inconclusiva a este respeito, e os resultados de localização da secção 2.5.2 favorecem o método de estimação com aproximação. Por outro lado, na comparação entre as representações Q e NQ, aqueles autores mediram uma diferença de precisão de 20 pontos percentuais. Embora a comparação realizada neste capítulo revele uma superioridade genérica da representação NQ, o ganho de desempenho obtido por esta representação não atinge um nível tão elevado.

Dado que a principal diferença entre os dois trabalhos reside no problema que cada um deles aborda, cremos que a classificação de lugares num ambiente conhecido coloca condições distintas da classificação genérica de objectos, e que estas estão na origem das diferenças encontradas. De facto, na classificação genérica de objectos, pretende-se distinguir objectos entre várias categorias, podendo, cada uma delas, incluir instâncias com aparências muito diversas, ou estarem colocadas em contextos muito diferentes. De forma diferente, pode dizer-se que na localização pretende-se distinguir entre um conjunto específico de objectos (lugares) previamente visualizados e que estão sujeitos à variabilidade decorrente da captura de imagem em momentos diferentes. Comparativamente, o problema da classificação genérica de objectos inclui maior variabilidade dentro de cada classe, podendo configurar um problema de maior dificuldade na classificação. Neste caso, os benefícios do uso da representação NQ terão maior impacto, explicando o maior ganho observado por Boiman, Shechtman e Irani (2008).

3. Fusão de características visuais por combinação de múltiplos classificadores

3.1 Introdução

No capítulo 2 desta tese foi apresentado um método de localização baseado na informação extraída de características locais das imagens. Uma questão central na utilização de características locais para este fim é a da fusão da informação proveniente de múltiplas características. A este respeito, e a par da abordagem usada no capítulo 2, encontram-se na literatura métodos de votação que usam a comparação binária de características. Este tipo de solução foi adoptada, por exemplo, nos trabalhos de Li, Yang e Kosecka (2005), Goedemé et al. (2007) e Ramisa et al. (2009). Apesar da viabilidade destas abordagens e da proposta no capítulo 2, segundo o nosso melhor conhecimento não foram anteriormente feitos estudos comparativos sobre a fusão de características com vista à localização de robôs. Neste capítulo realiza-se este tipo de avaliação, focando a fusão de informação no contexto da combinação de múltiplos classificadores. A opção por este enquadramento ocorre naturalmente, uma vez consideradas as semelhanças entre a classificação baseada em características locais e os sistemas de múltiplos classificadores. Por exemplo, em ambos os casos pretende-se obter um sistema de reconhecimento robusto que recorra a informação de múltiplas fontes: na localização visual esta informação provém de múltiplas características, enquanto nos sistemas de múltiplos classificadores ela provém das saídas de vários classificadores de base. Por outro lado, ambas as abordagens visam obter um classificador robusto construído sobre múltiplos classificadores simples, evitando assim a necessidade de se desenvolver um classificador individual, mas mais complexo, para o mesmo problema.

Apesar destes pontos em comum, as especificidades da classificação com características locais determinam o tipo de ferramentas, do âmbito da combinação de classificadores, que são aplicáveis a este problema. Estas especificidades resultam, em particular, do facto de as características serem extraídas de forma dinâmica, em regiões da imagem selecionadas por um detector de pontos de interesse. Devido a este mecanismo, não é possível considerar quer o número quer as regiões da imagem em que as características são extraídas como sendo fixos. A primeira consequência disso é a de não ser possível treinar classificadores de base para tratar características

específicas, recorrendo-se então a classificadores probabilísticos genéricos. A segunda consequência é a de também não ser possível considerar combinadores com aprendizagem, o que elimina da gama de combinadores aplicáveis técnicas como o *bagging*, o *Adaboost* ou a mistura de peritos. Dentro dos constrangimentos apresentados, neste capítulo são avaliados os dois classificadores sem aprendizagem mais populares: as regras do produto e da soma, bem como duas extensões a estas regras. Enquanto as anteriores operam sobre classificadores de saídas contínuas, um terceiro método, de votação, será igualmente considerado, por ter sido usado em trabalhos anteriores de localização.

Importa nesta introdução referir as dificuldades colocadas pelo problema de localização que poderão ser atenuadas na fase de fusão de informação. Em primeiro lugar, devido ao número tipicamente elevado de características extraídas de imagens naturais, um nível significativo de ruído é introduzido quando as condições de visualização variam entre a fase de modelação e de teste. Neste caso, ilustrado na Figura 3.1, duas imagens do mesmo lugar podem partilhar apenas um pequeno número de características, enquanto as restantes contribuem com ruído. Como se verá, as modificações às regras algébricas podem ser usadas para aliviar este efeito. Por outro lado, quando os modelos dos lugares são construídos sobre dezenas de imagens, o número de características que contêm é também muito elevado, da ordem dos milhares. Nestes casos, os modelos podem ser demasiado abrangentes, em termos da diversidade de características incluídas, aumentando por isso a probabilidade de um lugar produzir probabilidade significativa para características extraídas de outros lugares. Este aspecto, que pode igualmente ser aliviado na fase de fusão, será estudado através da análise da granularidade do ambiente e do seu impacto sobre os resultados de classificação.

Este capítulo apresenta a seguinte estrutura: a secção 3.2 oferece uma panorâmica dos trabalhos mais relevantes que estão relacionados com o presente estudo; na secção 3.3 definem-se os métodos de combinação que serão avaliados, as suas extensões e os classificadores de base que usam; a secção 3.4 apresenta a análise de discriminatividade, à semelhança do capítulo 2, mas aqui aplicada com o intuito de

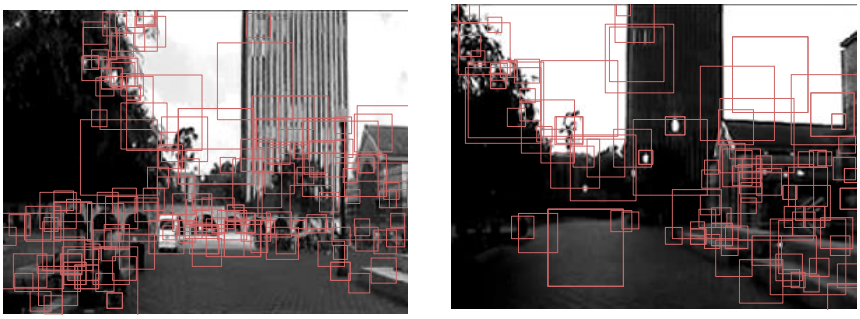


Figura 3.1. Duas imagens do mesmo lugar em que, devido a variações de perspectiva e luminosidade, apenas um pequeno número de características é comum às duas.

revelar as propriedades dos métodos de combinação; a secção 3.5 colige os resultados de localização, focando em particular o impacto da granularidade do ambiente e o desempenho dos métodos em avaliação; a secção 3.6 conclui o capítulo com a discussão dos resultados mais importantes.

3.2 Trabalhos relacionados

3.2.1 Combinadores de saídas contínuas

Na literatura relacionada, existe uma vasta gama de métodos estudados para a fusão de classificadores com saídas contínuas, incluindo regras algébricas, estatísticas e combinadores com aprendizagem (Kuncheva, 2004). De entre estes métodos, as regras da soma e do produto têm recebido especial atenção, devido à sua simplicidade e ao bom desempenho que oferecem. Na prática, a regra da soma tem sido muitas vezes preferida devido à sua robustez, no entanto, em termos teóricos não encontra uma fundamentação na fusão Bayesiana (Tax et al., 2000). Contrariamente, a regra do produto decorre directamente da regra de Bayes, sob o pressuposto da independência dos classificadores de base. Apesar desta sólida fundamentação teórica, tem sido observado que a regra do produto é mais sensível ao ruído e a erros de estimação nas saídas dos classificadores (Tax et al., 2000). Este problema foi estudado por Alkoot e Kittler (2002), num trabalho onde desenvolvem um método de truncagem através do qual melhoram significativamente os resultados daquela regra. Uma estratégia diferente que pode ser usada para elevar o desempenho dos combinadores algébricos consiste em ponderar a contribuição dos classificadores de base. No âmbito desta estratégia, Fumera e Roli (2005) debruçaram-se sobre a família de combinadores lineares, concluindo que a combinação pela média é óptima quando os classificadores

de base exibem taxas de erro idênticas bem como correlações idênticas entre elas. Ainda segundo estes autores, a média ponderada pode atingir melhores resultados quando aquelas condições não são válidas, desde que seja possível calcular valores de pesos adequados.

Nos últimos anos, diversos estudos têm confirmado a superioridade da média ponderada em várias aplicações práticas. Entre eles, os trabalhos sobre reconhecimento de discurso áudio (Misra, Bourlard e Tyagi) e sobre autenticação de pessoas (Sanderson e Paliwal, 2003) demonstraram que o uso de pesos adaptados dinamicamente é muito eficaz no tratamento de ruído variável no tempo. O tema da ponderação dinâmica foi ainda tratado noutros estudos que salientam a relação entre os pesos e algumas medidas de fiabilidade. Por exemplo, em aplicações sobre sinal áudio, foram desenhadas medidas de fiabilidade baseadas no nível de ruído do sinal de entrada (Sanderson e Paliwal, 2003; Heckmann, Berthommier e Kroschel, 2002). Uma abordagem mais genérica, que não se limita a sinais áudio, passa pela análise dos resultados dos classificadores de base, com vista ao cálculo da sua fiabilidade.

Dentro desta estratégia, foram desenvolvidos métodos que usam a diferença entre as duas melhores classificações (Wark e Sridharan, 2001), a dispersão das probabilidades posteriores (Wark e Sridharan, 2001), ou os máximos destas distribuições (Seymour, Stewart e Ming). Ainda na mesma linha, uma abordagem que tem recebido especial atenção é aquela que se centra na entropia das distribuições posteriores para o cálculo da fiabilidade (Misra, Bourlard e Tyagi, 2003). Contrastando com outros métodos que usam apenas o máximo da distribuição ou os dois melhpres resultados, este método, segundo Misra, Bourlard e Tyagi (2003), apresenta maior potencial, pois faz uso de toda a informação disponível na distribuição. Por outro lado, num estudo de reconhecimento de discurso áudio, a medida de entropia suplantou outra medida que usa a distribuição completa, designadamente a dispersão da distribuição (Heckmann, Berthommier e Kroschel, 2002).

Uma questão levantada na utilização da entropia relaciona-se com o mapeamento dos seus valores para pesos a aplicar aos classificadores. A este respeito, a abordagem mais comum tem sido a de considerar os pesos inversamente proporcionais à entropia, reflectindo a intuição de que distribuições com entropia reduzida são geradas por classificadores fiáveis, enquanto classificadores que devolvem distribuições próximas da uniforme são menos informativos. Apesar da sua concordância com esta intuição, a

proporcionalidade inversa pode não oferecer um mapeamento óptimo, pelo que alguns estudos foram dedicados ao desenvolvimento e avaliação de funções alternativas (Misra, Bourlard e Tyagi, 2003; Gurban e Thiran, 2008). Em (Misra, Bourlard e Tyagi, 2003), os autores propõem duas variantes do inverso da entropia, desenhadas por forma a atenuar significativamente a contribuição dos classificadores com entropia elevada. Entre estas, a mais bem sucedida é a que calcula a média da entropia e atribui pesos negligenciáveis aos classificadores com entropia acima desse valor. Num estudo similar, Gurban e Thiran (2008) introduzem um mapeamento alternativo, em que os pesos são definidos como proporcionais ao simétrico da entropia, e mostram a sua superioridade relativamente ao mapeamento por proporcionalidade inversa, numa aplicação de reconhecimento de linguagem.

3.2.2 Combinadores de saídas binárias

Dentro desta categoria existem alguns sistemas desenvolvidos para a localização de robôs, recorrendo à combinação de características pelo método da votação. Todos estes métodos são centrados na comparação de características SIFT através da técnica proposta por Lowe (2004). Segundo este método, cada imagem dos modelos é considerada individualmente e, para cada uma das características de teste, é atribuído um valor binário, indicando a sua presença (1) ou ausência (0) nessa imagem. Desta forma, uma característica de teste pode colocar votos em diferentes lugares, se forem encontradas correspondências nas imagens dos seus modelos. Em (Li, Yang e Kosecka, 2005) a decisão de localização é tomada escolhendo o lugar com o maior número de votos acumulados na mesma imagem, ou convertendo o número de votos num valor de probabilidade a usar com o método Hidden Markov Model. Uma abordagem semelhante é descrita em (Goedemé et al., 2007), onde o resultado final é suportado por uma medida de distância que agrega informação de : i) o número de correspondências encontradas, ii) o número total de características nas imagens de teste e do modelo e iii) uma distância de alinhamento geométrico. Para além do trabalho de Goedemé et al. (2007), a ideia de incluir informação sobre a distribuição espacial das características foi aprofundada numa série de trabalhos, que recorrem aos constrangimentos da geometria epipolar para validar as correspondências encontradas (Wang, Cipolla e Zha; Fraundorfer, Engels e Nister, 2007; Ramisa et al., 2009). Nestes trabalhos, a etapa de verificação passa pela estimação da transformação geométrica entre as duas imagens, e.g., na forma da matriz fundamental (Wang,

Cipolla e Zha) ou da matriz de homografia (Fraundorfer, Engels e Nister, 2007), e pela selecção das características que satisfazem esta transformação. Dado o elevado número de correspondências incorrectas (*outliers*) normalmente presentes, a estimação da transformação é auxiliada por métodos de estimação robusta do tipo RANSAC (Fischler e Bolles, 1981).

3.3 Fusão de características através da combinação de classificadores

3.3.1 Classificadores de base

3.3.1.a Classificadores de base com saídas contínuas

No presente enquadramento, que encara a fusão de características na perspectiva da combinação de classificadores, cada uma das características da imagem de teste é associada a um classificador de base. Para este efeito, considera-se um classificador probabilístico genérico, o qual, mediante uma característica de teste e o modelo do ambiente, produz uma distribuição de probabilidades sobre os lugares do ambiente.

No capítulo 2 desta tese definiu-se já o classificador que será usado neste capítulo como classificador de base e que, resumidamente, assenta na inferência de Bayes, dada por

$$P(l_j|d_i) = \frac{P(d_i|l_j)P(l_j)}{P(d_i)} \tag{3.1}$$

e no cálculo da verosimilhança através de

$$P(d_i|l_j) = \max_m K(d_i, d_m^j). \tag{3.2}$$

Neste capítulo, a expressão anterior será ainda modificada por forma a incluir informação espacial sobre as características. O procedimento usado para introduzir esta informação assenta nas etapas: i) determinação da característica do modelo, designada por *z*, cuja aparência é mais próxima da característica de teste e ii) cálculo da verosimilhança, que combina a aparência e a distância espacial. Esta última é incluída através de um *Kernel* gaussiano, *K'*, que compara a posição p_i da característica de teste e a posição p_z da característica *z*. Globalmente, o cálculo da verosimilhança é descrito por:

$$\begin{cases} z = \text{ind}\max_m K(d_i, d_m^j) \\ P(d_i|l_j) = K(d_i, d_z^j) \cdot K'(p_i, p_z^j) \end{cases}. \tag{3.3}$$

Para a caracterização do *Kernel* geométrico definiremos p_i e p_z como os vectores que reúnem as coordenadas x e y das características, na forma

$$p_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad p_z^j = \begin{bmatrix} x_z^j \\ x_z^j \end{bmatrix}. \tag{3.4}$$

No caso mais geral este *Kernel* incluirá a informação sobre as duas coordenadas, sendo descrito por

$$K'_{xy}(x_i, x_z^j) = exp\left(-\frac{1}{2}(p_i - p_z^j)^{\mathrm{T}}\Sigma^{-1}(p_i - p_z^j)\right), \tag{3.5}$$

onde Σ é a matriz de covariância, que por simplicidade será considerada diagonal e composta por $\Sigma = diag(\sigma_x^2, \sigma_y^2)$. Os parâmetros σ_x e σ_y designam as larguras de banda do *Kernel* nas direcções x e y. Em algumas experiências, como as da selecção destes parâmetros, serão aplicados separadamente *Kernels* unidimensionais, sobre cada uma das direcções. Nestes casos a Eq. (3.5) reduz-se a

$$\begin{aligned} K'_x(x_i, x_z^j) &= exp\left(-\frac{(x_i - x_z^j)^2}{2\sigma_x^2}\right) \\ K'_y(y_i, y_z^j) &= exp\left(-\frac{(y_i - y_z^j)^2}{2\sigma_y^2}\right), \end{aligned} \tag{3.6}$$

respectivamente para as coordenadas x e y.

3.3.1.b Classificadores de base com saídas binárias

Estes classificadores distinguem-se dos anteriores não apenas pelo tipo de resultado que produzem mas também por avaliarem individualmente cada uma das imagens dos lugares. Esta propriedade decorre do uso do critério de Lowe (2004), através do qual se estabelecem correspondências estritamente entre as características de duas imagens. Considerando o modelo do lugar l_j como sendo definido por um conjunto de nI_j imagens, $\left\{I_1^j, I_2^j, \ldots, I_{nI_j}^j\right\}$, cada uma destas imagens recebe uma avaliação pelo classificador associado à característica d_i. Esta avaliação indica a presença ou não da característica na imagem do lugar e é calculada por:

$$m(d_i, I_k^j) = \begin{cases} 1 & \text{se } \|d_i - d_1\| \leq 0.8 \cdot \|d_i - d_2\| \\ 0 & \text{caso contrário} \end{cases}. \quad (3.7)$$

Nesta expressão, d_1 e d_2 designam os descritores da imagem I_k^j respectivamente mais próximo e segundo mais próximo de d_i e 0.8 é a razão máxima de distâncias, tal como proposto por Lowe (2004).

3.3.2 Métodos de combinação de classificadores

A ideia subjacente à combinação de saídas contínuas é a de determinar uma melhor estimativa das probabilidades das classes, conseguida pela combinação adequada das estimativas dos classificadores de base. No contexto da localização global, este objectivo traduz-se na estimação de $P(l_j|d_1, \ldots, d_{nf})$ a partir das saídas dos classificadores de base, $P(l_j|d_i)$. De seguida definem-se os métodos de combinação avaliados neste capítulo, designadamente o método de votação, as duas regras algébricas e as extensões destas.

Regra da soma – Através da regra da soma, a probabilidade associada ao lugar l_j é calculada pela média das saídas dos classificadores de base:

$$P(l_j|d_1, \ldots, d_{nf}) = \frac{1}{nf} \sum_{i=1}^{nf} P(l_j|d_i). \quad (3.8)$$

Regra do produto – Segundo esta regra, a probabilidade associada ao lugar l_j é determinada pelo produto das saídas dos classificadores de base:

$$P(l_j|d_1, \ldots, d_{nf}) = \frac{1}{Z} \prod_{i=1}^{nf} P(l_j|d_i). \quad (3.9)$$

Nesta expressão Z é uma constante calculada por forma a que a probabilidade acumulada totalize 1, através de

$$Z = \sum_{j=1}^{np} \left(\prod_{i=1}^{nf} P(l_j|d_i) \right). \quad (3.10)$$

Método da votação – Este método é aplicado sobre os classificadores de saídas binárias, produzindo, para cada imagem dos modelos, uma pontuação que consiste simplesmente na soma dessas saídas:

$$S(I_k^j) = \sum_{i=1}^{nf} m(d_i, I_k^j). \tag{3.11}$$

Uma vez determinadas as pontuações por imagens, a pontuação atribuída a cada lugar é a da imagem que lhe está associada e que obteve melhor classificação:

$$S(l_j) = \max_k S(I_k^j). \tag{3.12}$$

Sobre este método é ainda possível aplicar os constrangimentos da geometria epipolar. Para esse efeito recorreu-se ao método usado por Fraundorfer, Engels e Nister (2007), em que se aplica o conceito de homografia na verificação do conjunto de correspondências, relativamente à visualização da mesma cena sob diferentes perspectivas. Após a estimação da homografia, são modificadas as correspondências que não foram validadas, atribuindo o valor zero à saída do respectivo classificador. Assim, nesta versão do método de votação, a pontuação atribuída a uma imagem corresponde ao número de características validadas na verificação geométrica.

Regras ponderadas – Tanto a regra da soma como a do produto podem ser modificadas no sentido de se ponderar a contribuição dos classificadores de base. Este tipo de modificação dá origem a dois métodos de combinação que designaremos respectivamente por *regra da soma ponderada* e *regra do produto ponderado*. A primeira é definida como a média ponderada das saídas dos classificadores de base:

$$P(l_j|d_1, \dots, d_{nf}) = \frac{1}{Z}\sum_{i=1}^{nf} w_i P(l_j|d_i), \tag{3.13}$$

onde w_i é o peso atribuído ao classificador *i*.

Na definição da regra do produto ponderado faz-se notar que o logaritmo de $P(l_j|d_1, \dots, d_{nf})$, dado pela regra do produto, se relaciona com a média dos logaritmos de $P(l_j|d_i)$. A regra ponderada é definida neste domínio como uma combinação linear de $\log P(l_j|d_i)$, o que se traduz, no domínio original, na expressão:

$$P(l_j|d_1, \dots, d_{nf}) = \frac{1}{Z}\prod_{i=1}^{nf} P(l_j|d_i)^{w_i}. \tag{3.14}$$

Nas Eqs. (3.13) e (3.14), à semelhança da Eq. (3.9), *Z* é uma constante de normalização.

Modificação por *Threshold* – Esta modificação visa atenuar o impacto negativo que os erros de estimação produzem sobre a regra do produto. Especificamente, a introdução deste método, por Alkoot e Kittler (2002), foi motivada pela observação de que as probabilidades próximas de zero têm um efeito dominante sobre o resultado do produto, e de que estes valores podem ocorrer por via de uma subestimação das probabilidades. Por forma a atenuar este efeito, Alkoot e Kittler (2002) propõem a modificação das saídas dos classificadores de base que tomam valores abaixo de um limite (*threshold*), segundo

$$\begin{cases} P(l_j|d_i) = Th, & P(l_j|d_i)_o < Th \\ P(l_j|d_i) = P(l_j|d_i)_o, & P(l_j|d_i)_o \geq Th \end{cases}, \tag{3.15}$$

onde *o* se refere às estimativas iniciais e *Th* designa o parâmetro de *threshold*. Esta regra estabelece que aos valores abaixo de *Th* é atribuído o valor de *Th*, previamente à aplicação do combinador.

3.3.3 Estimação da confiança nos classificadores de base

Na secção anterior foram introduzidas as regras de combinação ponderadas, as quais recorrem a factores de peso para modular a contribuição dos classificadores de base. Estes pesos, que devem reflectir a confiança nos classificadores, serão calculados a partir da entropia da distribuição produzida por cada um deles, dada por

$$H_i = -\sum_{j=1}^{np} P(l_j|d_i) \log P(l_j|d_i). \tag{3.16}$$

Por forma a usar esta medida numa regra ponderada é necessário que a entropia seja mapeada para factores de peso, através de uma transformação adequada. Nesta tese consideram-se quatro métodos de mapeamento: o inverso da entropia (IE) (Misra, Bourlard e Tyagi, 2003; Gurban e Thiran, 2008), o negativo da entropia (NE) (Gurban e Thiran, 2008), o inverso da entropia com limite pela média (IEAT- *Inverse Entropy with Average Threshold*) (Misra, Bourlard e Tyagi, 2003) e o negativo da entropia com limite pela média (NEAT-*Negative Entropy with Average Threshold*) proposto no âmbito desta tese em (Campos, Correia e Calado, 2015).

IE – Neste mapeamento o peso atribuído a cada classificador é inversamente proporcional à sua entropia:

$$w_i = H_i^{-1}. \tag{3.17}$$

NE – Através deste mapeamento os pesos relacionam-se proporcionalmente com o simétrico da entropia, por

$$w_i = H_{max} - H_i. \tag{3.18}$$

Nesta expressão H_{max} é o valor máximo que a entropia pode tomar, o qual, num problema com *np* lugares candidatos, é calculado como

$$H_{max} = \log np. \tag{3.19}$$

IEAT – Neste mapeamento calculam-se valores de pesos preliminares, w'_i, segundo o método IE, que são posteriormente modificados por

$$\begin{cases} w_i = 10^{-4}, & w'_i < \overline{w'} \\ w_i = w'_i, & w'_i \geq \overline{w'} \end{cases}, \tag{3.20}$$

onde $\overline{w'}$ designa a média de w'_i.

NEAT – Este método resulta do cálculo de pesos preliminares, de acordo com a Eq. (3.18), que são de seguida modificados pela mesma regra de IEAT.

3.4 Análise de discriminatividade

À semelhança do estudo realizado no capítulo 2, nesta secção analisa-se a discriminatividade das características visuais, neste caso com o intuito de comparar as propriedades dos dois combinadores algébricos e das suas extensões.

3.4.1 Cálculo da discriminatividade

Evocando a definição de discriminatividade introduzida no capítulo 2, e tendo em conta que a combinação pela regra da soma coincide com o classificador NQ então avaliado, a discriminatividade desta regra é calculada pela Eq. (2.11). Relativamente à regra do produto, faz-se notar que esta tem por base o classificador Naive Bayes, cuja discriminatividade foi definida no capítulo 2, com vista à avaliação da representação Q. Assim, e recuperando as expressões introduzidas na secção 2.2, a discriminatividade das regras da soma e do produto são calculadas respectivamente por

$$discS_i = P(l_t|d_i) - P(l_{nt}|d_i) \tag{3.21}$$

$$discP_i = \frac{1}{sc}\Big(log\big(P(l_t|d_i)\big) - log\big(P(l_{nt}|d_i)\big)\Big). \tag{3.22}$$

No que diz respeito às modificações daquelas regras, as expressões anteriores são directamente aplicáveis à avaliação da modificação por *Threshold*, já que esta actua sobre as probabilidades e não sobre a forma de combinação. Por outro lado, nas regras ponderadas a combinação dos classificadores de base é feita modulando a sua contribuição por factores de peso. Nestes casos, o cálculo da discriminatividade é igualmente modulado por esses pesos, resultando nas seguintes expressões para a regra da soma e do produto:

$$discS_i = \frac{w_i}{sc}\left(P(l_t|d_i) - P(l_{nt}|d_i)\right) \tag{3.23}$$

$$discP_i = \frac{w_i}{sc}\Big(log\big(P(l_t|d_i)\big) - log\big(P(l_{nt}|d_i)\big)\Big). \tag{3.24}$$

Como se pode concluir pela Eq. (3.23), nesta versão da regra da soma há também a necessidade de aplicar um factor de escala *sc* de maneira a normalizar a distribuição de discriminatividade, pois os pesos não foram restringidos a valores menores que 1.

Tal como na secção 2.3, a análise que se segue será baseada nos perfis de discriminatividade (Figura 3.2) e nos histogramas de discriminatividade conjunta de dois combinadores (Figura 3.3).

3.4.2 Comparação das duas regras algébricas e suas extensões

A comparação dos perfis das regras simples da soma (Figura 3.2.a) e do produto (Figura 3.2.b) torna evidente que estes combinadores apresentam características muito distintas. Por simplicidade, de seguida designaremos as características de discriminatividade próxima de zero como características *não informativas* e as de discriminatividade positiva como características *positivas*.

Fundamentalmente, os perfis dos dois combinadores distinguem-se pelo facto de na regra da soma as características não informativas estarem bem separadas das características positivas, que se distribuem de forma aproximadamente uniforme sobre o intervalo positivo e apresentam um pico na discriminatividade máxima. Tal não se verifica na regra do produto, em que, por um lado as características não informativas se distribuem de forma mais dispersa e, mais relevante, as características positivas decaem com o aumento de discriminatividade. Ambas as tendências são explicadas pela sensibilidade do produto aos erros de estimação, quando a saída dos classi-

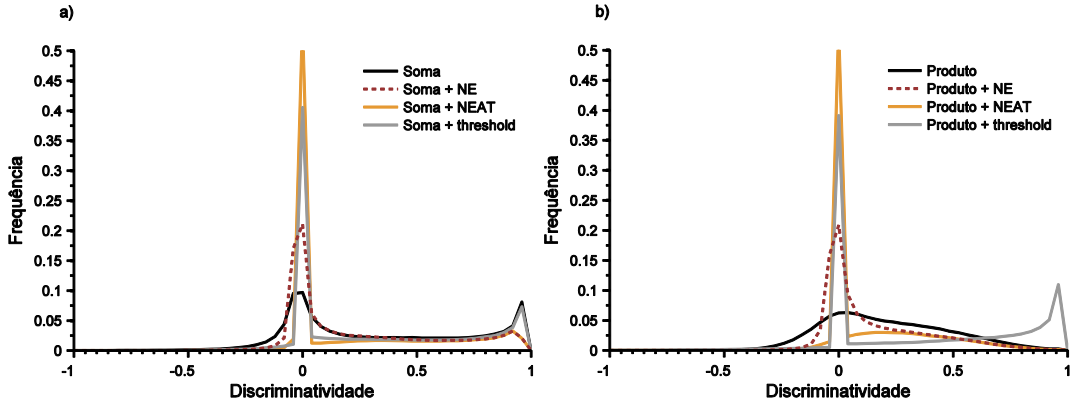


Figura 3.2. Perfis de discriminatividade das duas regras algébricas e das suas extensões. À esquerda, regra da soma; à direita, regra do produto.

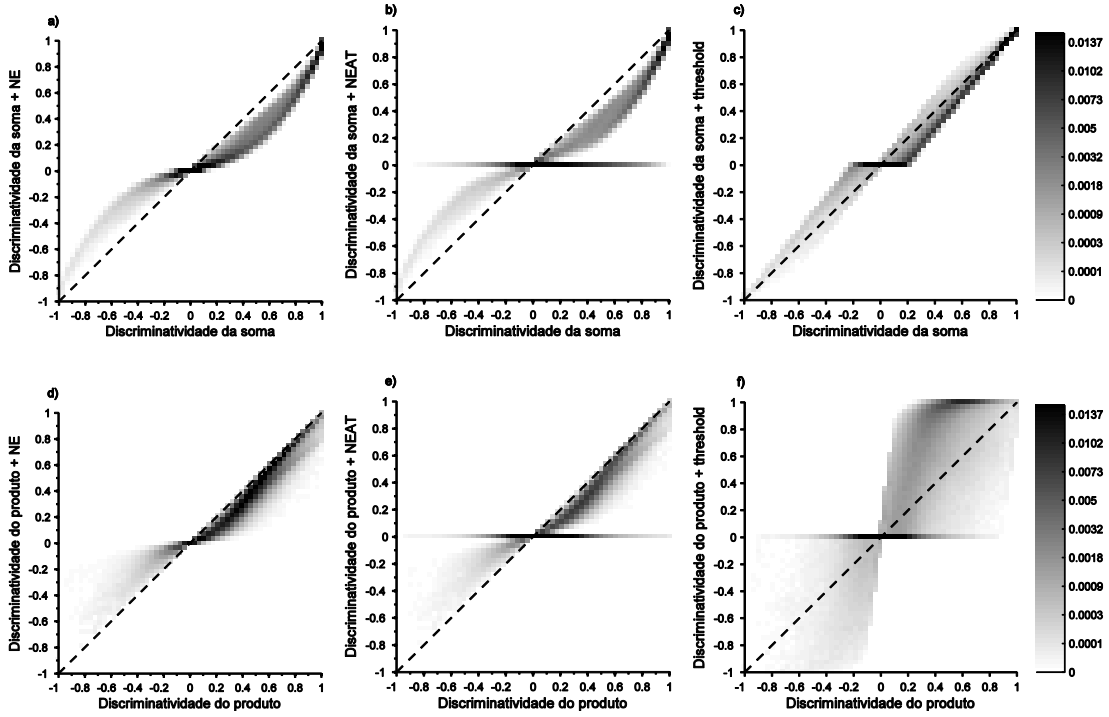


Figura 3.3. Distribuição conjunta da discriminatividade entre as regras algébricas e as suas extensões. Em cima, distribuições relativas à regra da soma; em baixo, distribuições relativas à regra do produto.

ficadores é próxima de zero. De facto, notando que, para esta regra, as derivadas parciais da discriminatividade têm a forma

$$\frac{d}{dP(l_t|d_i)}discP_i = \frac{1}{sc}\cdot\frac{1}{P(l_t|d_i)}$$
$$\frac{d}{dP(l_{nt}|d_i)}discP_i = -\frac{1}{sc}\cdot\frac{1}{P(l_{nt}|d_i)}\ , \qquad (3.25)$$

conclui-se que para valores de $P(l_t|d_i)$ ou $P(l_{nt}|d_i)$ próximos de zero, variações, ainda que reduzidas, destes valores são amplificadas nos termos $P(l_t|d_i)^{-1}$ e $P(l_{nt}|d_i)^{-1}$. Este efeito influencia a discriminatividade das características não informativas já que, tipicamente, estas não foram observadas nos modelos e como tal recebem probabilidades baixas. Relativamente às características positivas, estas ocorrem quando foram detectadas no lugar correcto, resultando em $P(l_t|d_i)$ próximo de 1 e não estão presentes no modelo do lugar incorrecto, o que significa que $P(l_{nt}|d_i)$ é próximo de zero. De forma indesejada, também neste caso a discriminatividade é afectada pelos erros de estimação de $P(l_{nt}|d_i)$ e, como consequência disso, apenas algumas características adquirem valores de *discP* próximos da discriminatividade máxima.

Os restantes perfis da Figura 3.2, juntamente com os gráficos da Figura 3.3 ilustram a influência que cada uma das modificações promove sobre as regras originais. Da análise destes resultados, as seguintes observações podem ser extraídas:

i) Todas as extensões às regras aproximam a discriminatividade das características não informativas de zero. Este efeito pode ser considerado positivo, pois estas características introduzem ruído na classificação.

ii) Aludindo às regras ponderadas, o efeito descrito em i) é explicado pelo facto de a entropia daquelas características ser alta e o mapeamento para pesos atribuir-lhes pesos reduzidos. No caso da modificação com limite pela entropia média (NEAT), aquele efeito é largamente acentuado, pela atribuição de um peso negligenciável às características de entropia inferior à média.

iii) A operação de *Threshold* produz o efeito i) através de um mecanismo diferente. Recordando que as características não informativas ocorrem normalmente quando $P(l_t|d_i)$ e $P(l_{nt}|d_i)$ são baixas, a aplicação do *Threshold* diminui a diferença entre elas e, consequentemente, aproxima a discriminatividade de zero. Concretamente, para cerca de 50% das características a discriminatividade passa a ser zero, pelo facto de $P(l_t|d_i)$ e $P(l_{nt}|d_i)$ assumirem o mesmo valor, o de *Threshold*, após esta operação.

iv) Além disso, a operação de *Threshold* é especialmente benéfica na regra do produto, por modificar também a distribuição das características positivas. Como referido anteriormente, a discriminatividade destas características é prejudicada pelos erros de estimação de $P(l_{nt}|d_i)$. A operação de *Threshold* tem aqui um efeito

positivo, já que impõe um limite mínimo a $P(l_{nt}|d_i)$ e evita assim valores próximos de zero, para os quais os erros de estimação são amplificados de forma acentuada. Nesta extensão da regra do produto a distribuição das características positivas aproxima-se da distribuição encontrada na regra da soma.

3.5 Resultados

3.5.1 Granularidade do modelo do ambiente

Por forma a analisar a influência da granularidade do ambiente no desempenho dos combinadores, foram criados diversos modelos do ambiente FDF Park, com granularidades sucessivamente mais finas. A escolha deste *dataset* deve-se ao facto de o *dataset* IDOL apresentar alguns lugares para os quais o número de imagens é reduzido, não permitindo por isso testar uma gama de granularidades que suportasse a análise que se segue.

Na construção dos modelos do ambiente foram usados dois processos distintos: o primeiro, que designaremos por *partição original*, parte da definição original do *dataset* em 9 lugares e divide as sequências de imagens de cada lugar no mesmo número de sub-sequências. Aplicando partições em 3, 5, 10 e 20 sub-sequências, obtiveram-se respectivamente modelos com 27, 45, 90 e 180 lugares. No segundo processo, designado por *partição fixa*, adopta-se uma abordagem diferente, em que cada sequência original é dividida em sub-sequências com um número fixo de imagens. Aplicando valores para este parâmetro de 50, 30, 15 e 7 imagens, obtiveram-se modelos com os números de lugares indicados na Tabela 3.1.

A Figura 3.4 apresenta a precisão obtida através dos dois processos de divisão, para cada um dos conjuntos de modelação. Para além dos resultados gerados pelas regras da soma e produto, apresentam-se os resultados obtidos por uma das modificações que potencia o desempenho destas, designadamente a operação de *Threshold*. Entre os aspectos salientes da Figura 3.4 destaca-se o hiato de desempenho entre os modelos obtidos pela partição original e pela partição fixa. Esta diferença deve-se ao facto de, na partição original, o número de imagens que constituem os modelos apresentar grandes variações de lugar para lugar. Como mostra a Tabela 3.2, as razões entre o número de imagens da sequência mais extensa e o da mais curta é superior a 2 nos três conjuntos de modelação, e irá manter-se nas subdivisões pelo método da partição

Tabela 3.1. Número de lugares obtidos pela partição fixa do *dataset* FDF Park.

Nº de imagens	Conjunto de modelação		
	A	C	D
50	26	17	21
30	49	33	35
15	99	63	70
7	209	136	145

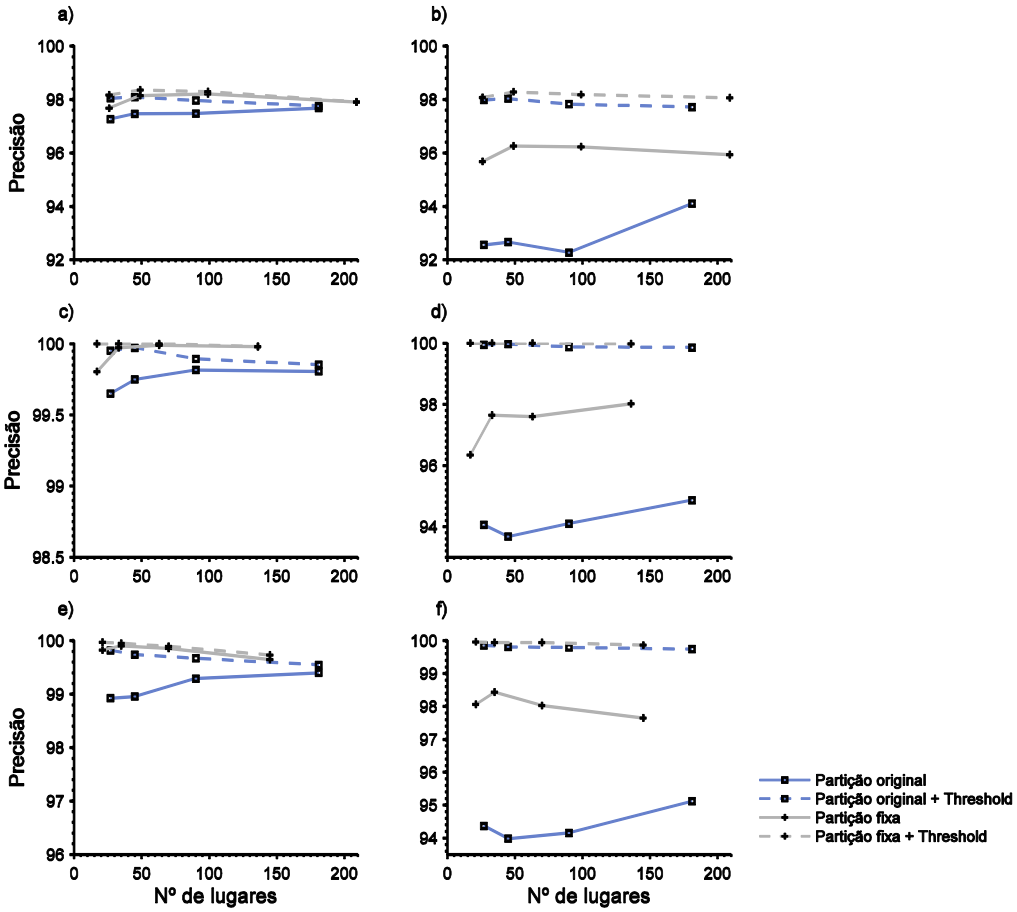


Figura 3.4. Precisão como função da granularidade do ambiente. A coluna da esquerda apresenta os resultados da regra da soma, enquanto a coluna da direita os resultados respeitantes à regra do produto. As linhas referem-se respectivamente aos dados de modelação A, C e D.

original. A discrepância no número de imagens tem como consequência que o número de características que constituem os modelos de uns lugares é significativamente maior do que noutros e, por isso, a probabilidade de se encontrar uma característica próxima da característica de teste é também maior. Como tal, os classificadores em estudo apresentam, na presença de modelos não-balanceados, uma preferência pelos lugares com modelos mais extensos, o que prejudica o desempenho e justifica a diferença de precisão encontrada.

Tabela 3.2. Números máximo e mínimo de imagens dos lugares definidos na partição original do *dataset* FDF Park.

Nº de imagens	Conjunto de modelação		
	A	C	D
máximo	284	202	214
mínimo	131	81	93

Os resultados apresentados demonstram também que a precisão depende do grau de granularidade dos modelos. A este respeito, verifica-se que, no caso dos modelos obtidos pela partição original, o desempenho aumenta com a granularidade, o que se explicará pelo menor impacto do efeito descrito acima, quando todos os modelos têm um menor número de imagens, ainda que se mantenham não-balanceados.

Quando se consideram os resultados da partição fixa verifica-se que a precisão aumenta quando se passa do primeiro nível de granularidade para o segundo e posteriormente reduz-se, em níveis de granularidade superior. Esta tendência sugere que, nas melhores condições, i.e., com partição fixa, o melhor desempenho é atingido como um compromisso entre os modelos conterem suficientes imagens para oferecer uma descrição rica do lugar e, simultaneamente, não serem demasiado abrangentes. A última condição relaciona-se com o facto, já observado na partição original, de que os modelos mais extensos promovem a existência de probabilidades significativas para características extraídas em lugares não correspondentes.

Doravante, a avaliação de classificadores sobre o *dataset* FDF Park será realizada sobre a partição que genericamente conduz ao melhor desempenho, i.e., a partição fixa com o segundo nível de granularidade (30 imagens por lugar).

3.5.2 Selecção de parâmetros

Nesta secção descreve-se a análise que conduziu à determinação dos parâmetros usados nos combinadores algébricos. Especificamente, serão abordados os parâmetros *Th*, relativo à modificação por *Threshold*, e os parâmetros de largura de banda do *Kernel* geométrico. Ao longo da secção constatar-se-á que os *dataset*s FDF Park e IDOL apresentam especificidades próprias as quais determinam sensibilidades diferentes relativamente àqueles parâmetros. Tendo em conta essas especificidades, e

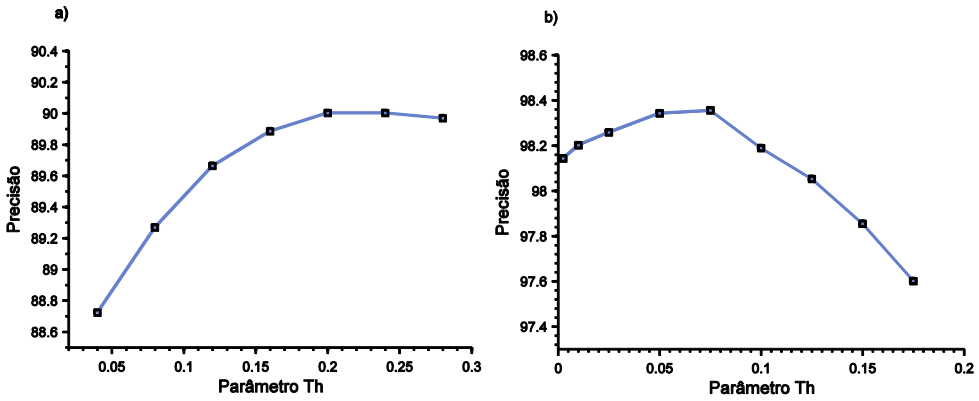


Figura 3.5. Precisão global como função do parâmetro *Th*. A) *dataset* IDOL, b) *dataset* FDF Park.

com o objectivo de avaliar o melhor desempenho proporcionado pelos classificadores, serão escolhidos parâmetros individualizados para cada um dos *dataset*s.

3.5.2.a Parâmetro *Th*

A Figura 3.5, que apresenta a evolução da precisão como função do parâmetro *Th*, mostra que o benefício da operação de *Threshold* apresenta um pico, a partir do qual o aumento de *Th* acarreta perda de informação e, consequentemente, degradação do desempenho. No trabalho de Alkoot e Kittler (2002), em que a operação de *Threshold* foi introduzida, a selecção de *Th* foi feita de forma empírica, uma opção que reflecte a dificuldade em definir-se uma regra genérica para a sua selecção. No presente problema essa dificuldade é acrescida, pelo facto de os valores de *Th* óptimos dependerem, como se pode ver na Figura 3.5, do *dataset* em causa. Esta variabilidade pode, no entanto, ser relacionada com os valores de probabilidades típicos que se esperam num dado *dataset*. De facto, constatando que $P(l_j|d_i)$ é determinado, não só pela verosimilhança da característica *i* no lugar *j*, mas também pela sua probabilidade nos restantes lugares, através do termo de normalização $P(d)$ (ver Eq. (3.1)), conclui-se, por exemplo, que em ambientes com maior número de lugares, os valores de $P(l_j|d_i)$ são tipicamente mais baixos. De seguida mostrar-se-á que é possível prever, com alguma precisão, o valor de *Th* óptimo, através da modelação do termo variável de $P(d)$. Lembrando que $P(l_j|d_i)$ é dado pela Eq. (3.1) e que a distribuição à priori dos lugares é considerada uniforme, $P(l_j|d_i)$ pode ser escrita como:

$$P(l_j|d_i) = \frac{P(d_i|l_j)}{\sum_j P(d_i|l_j)}. \qquad (3.26)$$

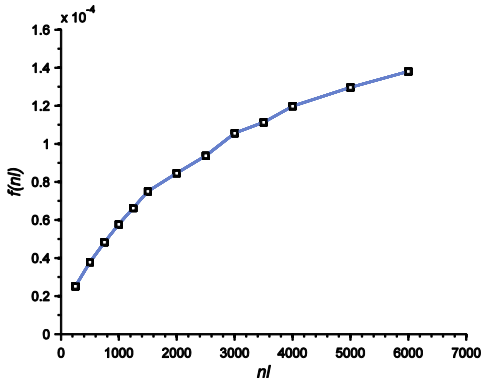


Figura 3.6. Função *f*(*nl*) que descreve a evolução de E[$P(d_i \mid l_j)$] com o número de características no modelo de um lugar.

Com vista a prever o valor médio que o denominador nesta expressão irá tomar, estimar-se-á o seu valor esperado, que designaremos por λ, pela expressão

$$\lambda = E\left[\sum_j P(d_i|l_j).\right] = \sum_j E\left[P(d_i|l_j)\right]. \qquad (3.27)$$

Na aplicação da expressão anterior há a necessidade de modelar E[$P(d_i \mid l_j)$], a qual depende do número de características, nl_j, presentes no modelo do lugar l_j. A estimação desta função foi feita através da média empírica dos valores medidos num número elevado de experiências. Nestas experiências foram construídos modelos de lugares com um número variável de características, escolhidas aleatoriamente no *dataset* FDF Park. De seguida calculou-se a probabilidade de 50×10^3 características de teste, escolhidas também aleatoriamente, relativamente àqueles modelos. Pela média dos valores medidos para cada valor de *nl*, estimou-se a função *f*(*nl*) da Figura 3.6 que representa a evolução de E[$P(d_i \mid l_j)$], com *nl*. A forma crescente da curva traduz a intuição de que em lugares com maior número de características é maior a probabilidade de se encontrar uma que seja próxima de uma característica de teste escolhida aleatoriamente.

Uma vez conhecida a função *f*(*nl*) é possível calcular λ para um dado modelo do ambiente, caracterizado por um número *np* de lugares, cada um deles contendo nl_j características:

$$\lambda = E\left[\sum_j P(d_i|l_j)\right] = \sum_{j=1}^{np} f(nl_j). \qquad (3.28)$$

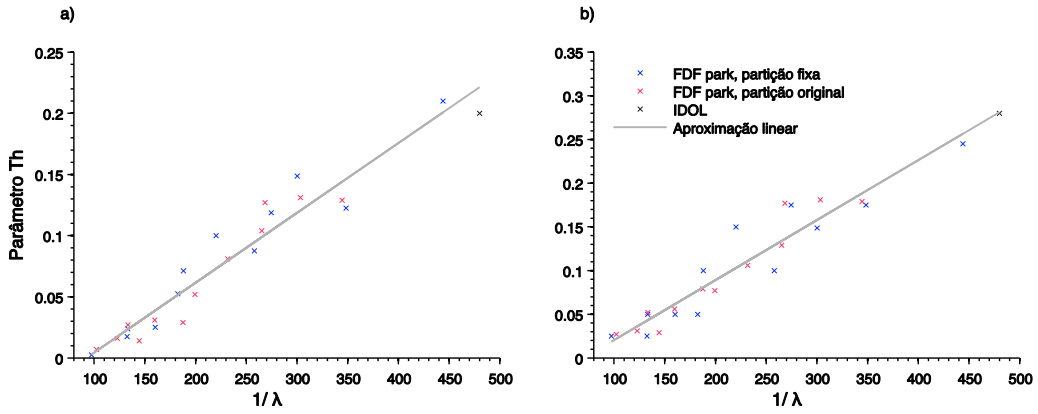


Figura 3.7. Valores óptimos de *Th* medidos no *dataset* IDOL e em diferentes representações do *dataset* FDF Park. A) valores para a regra da soma, b) para a regra do produto.

Na Figura 3.7 apresentam-se os valores óptimos de *Th* encontrados para os *datasets* FDF Park e IDOL, como função de 1/λ, já que as probabilidades à posteriori dependem inversamente desta variável. Como se pode observar, os valores de *Th* óptimos evoluem de forma aproximadamente linear com 1/λ, o que motivou o ajuste de uma função linear a estes dados, apresentada também na figura. Esta função, cujos parâmetros foram determinados por regressão linear, permite prever os valores óptimos de *Th* para um dado modelo do ambiente, apresentando um coeficiente de determinação de R^2=0.94 para a regra da soma e de R^2=0.93 para a regra do produto. Nas experiências seguintes serão usados os valores de *Th* dados por esta aproximação.

3.5.2.b Largura de banda do *Kernel* geométrico

Com vista à selecção dos parâmetros do *Kernel* geométrico estudou-se, de forma independente, a influência dos parâmetros σ_x e σ_y. Para tal aplicou-se, na expressão (3.3), *Kernels* unidimensionais sobre cada uma das coordenadas da imagem, do que resultaram os dados de precisão apresentados na Figura 3.8. Os ganhos expectáveis através da aplicação destes *Kernels* prendem-se, por um lado, com a quantidade de informação que os constrangimentos geométricos introduzem e, por outro, com a repetibilidade das posições das características entre diferentes passagens no mesmo lugar.

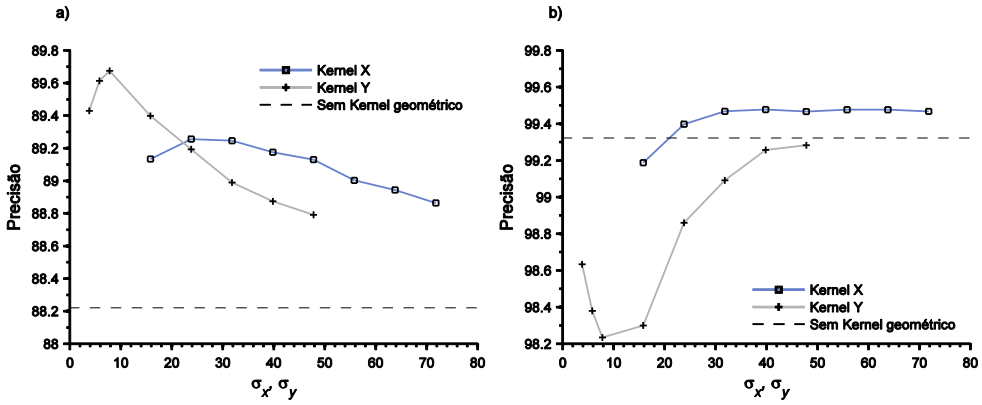


Figura 3.8. Precisão global medida nos *datasets* a) IDOL e b) FDF Park com a integração da informação espacial em x e y.

Os resultados da Figura 3.8.a mostram que a introdução de informação espacial, em qualquer uma das dimensões, produz um ganho de desempenho no *dataset* IDOL, todavia este é mais acentuado na coordenada vertical. Para compreender esta diferença, é necessário recordar que este *dataset* é captado por um robô que se desloca sobre o plano do chão num ambiente de interior. Por esta razão, imagens do mesmo lugar podem ser captadas com alguns desvios de posição e orientação no plano, mas sempre aproximadamente à mesma altura. Este constrangimento leva a que as medidas da coordenada y apresentem maior consistência do que as medidas em x, explicando assim o maior benefício da introdução da informação em y.

Os dados da Figura 3.8.b conduzem a diferentes conclusões quando o *dataset* FDF Park é considerado. Neste caso, a aplicação do *Kernel* em x produz um ligeiro ganho na precisão, enquanto o *Kernel* em y introduz uma perda de desempenho. Este efeito negativo é explicado, em primeiro lugar, pelo facto de a informação sobre a posição vertical ser aqui menos informativa, já que as imagens deste *dataset* são em grande parte ocupadas superiormente por céu e inferiormente pelo chão. Daí resulta que as características mais informativas se concentrem em torno da linha do horizonte e estejam, em termos da coordenada vertical, próximas entre si. Por outro lado, as imagens deste *dataset* são recolhidas manualmente, o que introduz ruído na captura e reduz a repetibilidade na detecção das posições das características.

Mediante os resultados observados na Figura 3.8, nas experiências posteriores em que se recorre à informação geométrica serão usados σ_x=24 píxeis e σ_y=8 píxeis para o *dataset* IDOL e, no respeita ao *dataset* FDF Park, apenas será aplicado o *Kernel* em x, com σ_x=40 píxeis.

3.5.3 Desempenho dos métodos de combinação

Nas Tabelas 3.3 a 3.5 apresentam-se os resultados de classificação obtidos para o *dataset* IDOL e no Anexo A os resultados relativos ao *dataset* FDF Park. A par dos resultados obtidos exclusivamente pela comparação de descritores (Tabelas 3.3, A.1, A.3 e A.5), apresenta-se a precisão medida após a integração de informação geométrica (Tabelas 3.4, A.2, A.4 e A.6). Da análise destes resultados é possível extrair os seguintes dados mais salientes:

- O método de votação tem um desempenho inferior aos combinadores algébricos, com uma única excepção (*dataset* FDF Park, conjunto C, em que é superior ao produto). Nas condições de localização mais difíceis, a diferença de desempenho é muito acentuada, atingindo cerca de 30 pontos percentuais relativamente à soma, no par *night*/*sunny* do *dataset* IDOL.
- A regra da soma suplanta a do produto em todas as comparações das precisões médias. Em termos individuais, a diferença atinge máximos de aproximadamente 8 pontos percentuais no *dataset* IDOL (no par *night*/*sunny*) e de 6 pontos percentuais no *dataset* FDF Park (par A/I).
- Tanto o método NEAT como o de *Threshold* beneficiam, em geral, as regras algébricas. Os ganhos obtidos são mais relevantes no caso do produto, o qual passa a apresentar desempenhos próximos da regra da soma, após a aplicação daquelas modificações.

Considerando de seguida os resultados que integram a informação geométrica, observa-se que:

- A introdução desta informação produz ganhos de desempenho e atenua as diferenças de precisão encontradas anteriormente, entre o método de votação e as regras algébricas e entre a regra da soma e a do produto.
- Os benefícios da aplicação dos constrangimentos geométricos são mais significativos no método da votação. Todavia, nestas condições encontram-se ainda hiatos de desempenho acentuados, em particular no *dataset* IDOL, em que a diferença de precisão média relativamente à regra da soma é de 10.6 pontos percentuais.

Tabela 3.3. Precisão [%] dos métodos de combinação sobre o *dataset* IDOL.

Condições de modelação	Condições de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
	Cloudy	94.68	98.29	96.49	98.65	98.46	98.55	98.54
Cloudy	*Sunny*	92.45	97.64	95.96	97.89	97.6	97.86	97.81
	Night	68.34	85.46	80.32	90.09	89.13	88.73	87.96
	Cloudy	88.52	97.71	95.8	98.22	97.99	98.19	98.19
Sunny	*Sunny*	94.37	98.30	96.54	99.03	98.47	99.11	98.91
	Night	38.99	67.07	62.26	71.46	69.54	70.5	70.11
	Cloudy	69.23	87.19	80.85	90.31	89.23	89.94	89.37
Night	*Sunny*	41.99	71.54	63.41	74.51	73.54	75.08	74.63
	Night	94.54	98.30	96.2	98.57	98.25	98.62	98.58
Média		**75.90**	**89.06**	**85.31**	**90.97**	**90.25**	**90.73**	**90.46**

Tabela 3.4. Precisão [%] dos métodos de combinação sobre o *dataset* IDOL, integrando os constrangimentos geométricos.

Condições de modelação	Condições de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
	Cloudy	95.59	98.84	98.02	99.06	98.77	99.06	98.98
Cloudy	*Sunny*	93.81	98.58	97.61	98.38	98.29	98.52	98.47
	Night	74.15	88.7	85.76	90.5	89.84	89.55	89.1
	Cloudy	94.76	98.1	97.12	98.39	98.15	98.29	98.17
Sunny	*Sunny*	96.53	98.41	97.55	98.69	98.26	98.52	98.45
	Night	48.35	73.05	69.51	75.46	73.2	74.2	73.25
	Cloudy	76.49	89.9	85.22	92.32	91.98	91.07	90.7
Night	*Sunny*	48.21	74.78	68.76	78.37	77.48	76.51	75.77
	Night	95.87	98.68	97.5	98.81	98.41	98.8	98.73
Média		**80.42**	**91.00**	**88.56**	**92.22**	**91.60**	**91.61**	**91.29**

Tabela 3.5. Precisão [%] obtida pelas funções de mapeamento da entropia para os pesos das regras ponderadas. Resultados calculados sobre o *dataset* IDOL.

Condições de modelação	Condições de teste	IE	NE	IEAT	NEAT	IE	NE	IEAT	NEAT
	Cloudy	98.51	98.67	98.52	98.65	97.41	98.31	97.49	98.46
Cloudy	*Sunny*	97.40	97.87	97.41	97.89	97.00	97.45	97.09	97.60
	Night	89.17	89.83	89.66	90.09	87.08	88.30	88.31	89.13
	Cloudy	97.97	98.23	97.98	98.22	97.10	97.73	97.28	97.99
Sunny	*Sunny*	98.92	99.01	98.95	99.03	97.35	98.24	97.53	98.47
	Night	71.07	71.27	71.62	71.46	68.60	68.62	69.81	69.54
	Cloudy	89.50	90.21	89.78	90.31	87.38	88.56	88.71	89.23
Night	*Sunny*	73.25	74.76	73.38	74.51	71.34	72.68	72.78	73.54
	Night	97.85	98.6	97.86	98.57	96.36	98.20	96.42	98.25
média		**90.40**	**90.94**	**90.57**	**90.97**	**88.85**	**89.79**	**89.49**	**90.25**

Na Tabela 3.5 estão coligidos os resultados calculados através dos diversos mapeamentos da entropia. Esta análise incide sobre o *dataset* IDOL em virtude de o desempenho das regras modificadas sobre o *dataset* FDF Park ser elevado e, por isso, menos sensível ao mapeamento escolhido.

Nestes resultados são identificáveis duas tendências, observadas na comparação dos métodos que usam proporcionalidade directa (NE, NEAT) com aqueles que usam a proporcionalidade inversa (IE, IEAT) e na comparação dos que aplicam a atenuação das características com entropia inferior à média (IEAT, NEAT) com os que não aplicam esta operação (IE, NE). As diferenças de desempenho, mais evidentes na regra do produto, indicam que o mapeamento linear é preferencial à proporcionalidade inversa. Além disso, foi verificado em 3.4.2 que os mapeamentos com a operação *Average Threshold* incidem sobretudo sobre as características não informativas, reduzindo significativamente o seu contributo para o resultado final. Os resultados presentes na Tabela 3.5 indicam que esta operação é benéfica, dada a superioridade de IEAT e NEAT, quando comparados, respectivamente, com IE e NE.

3.6 Discussão

Com base nos resultados apresentados na secção 3.5.3 é possível retirar algumas conclusões sobre o desempenho relativo dos combinadores em estudo. Apesar do relativo sucesso do método de votação, usado em diversos trabalhos (Li, Yang e Kosecka, 2005; Goedemé et al., 2007; Ramisa et al., 2009), este apresenta um desempenho inferior aos combinadores alternativos. Existem duas razões que podem concorrer para este resultado: por um lado, o tipo de saídas, binário, utilizado pelos classificadores de base pode ser menos informativo do que as saídas contínuas. A representação por saídas contínuas é interessante, por exemplo, por permitir que o resultado relativo a um lugar seja função também da ocorrência da característica noutros lugares. Efectivamente, a operação de normalização das probabilidades (Eq. (3.1)) modifica os valores individuais das saídas, em função da frequência da característica em todo o ambiente, o que pode ser importante por, implicitamente, atribuir menor peso às características mais comuns. A segunda razão que identificamos para a precisão limitada do método de votação prende-se com a comparação por imagens – em lugar da comparação por modelos – que é usada. Este tipo de comparação pode ser entendido como um caso extremo da comparação por

modelos, em que, neste caso, o modelo de um lugar inclui apenas uma imagem. À luz da análise sobre a granularidade do ambiente, descrita em 3.5.1, conclui-se que um modelo composto por uma imagem pode carecer de robustez, prejudicando por isso o desempenho deste método.

A aplicação dos constrangimentos geométricos contribui para uma melhoria acentuada do método de votação. Contudo, também neste caso o seu desempenho é inferior ao das regras algébricas. Este dado é especialmente significativo sobre a superioridade dos combinadores algébricos, já que a verificação geométrica baseada na homografia é mais adequada ao problema de localização do que a verificação associada às regras algébricas. De facto, enquanto aquela contempla variações de perspectiva, o *Kernel* geométrico impõe posições fixas e aplica uma comparação suave como forma de tratar os desvios existentes. A opção pelo uso do *Kernel* geométrico em detrimento dos métodos da geometria epipolar deve-se ao facto de estes não serem directamente aplicáveis aos classificadores por saídas contínuas, por duas razões: i) as saídas contínuas contrastam com as saídas binárias, que implicitamente circunscrevem o conjunto das correspondências a ser verificado através do constrangimento geométrico; no caso das saídas contínuas, não havendo uma separação binária, não é possível identificar esse conjunto; ii) os classificadores com saídas contínuas comparam a imagem de teste com o modelo de um lugar, que agrega informação de várias imagens; nessa comparação são encontradas as características do modelo que são mais próximas da imagem de teste, as quais pertencem, normalmente, a diversas imagens; consequentemente, não são aplicáveis, neste caso, as condições da geometria epipolar, já que estas estabelecem relações entre *duas* imagens.

Os resultados de 3.5.3 demonstram também a superioridade da regra da soma relativamente à regra do produto, quando consideradas nas suas versões originais. Este resultado era esperado, uma vez que a regra do produto é mais sensível a dados com ruído, o qual tem um impacto considerável nas situações de localização mais adversas. Como demonstrado na secção 3.4, ambas as modificações a estas regras atenuam a contribuição das características não informativas, mecanismo através do qual se obtém ganhos na localização, observados em 3.5.3. No que diz respeito à modificação por ponderação, verificou-se que os mapeamentos da entropia mais eficazes são os IEAT e NEAT, que, à semelhança da operação de *Threshold*, levam

aquela atenuação a um extremo, em que uma grande percentagem das características passa a ter discriminatividade zero. Comparativamente, a operação de *Threshold* é em geral mais eficaz do que as operações de ponderação, quando aplicadas à regra do produto. Como demonstrado pelas relações de discriminatividade desta regra, a operação de *Threshold* modifica o perfil das características positivas, fazendo emergir um número maior de características próximas da discriminatividade 1. Este efeito, que não é conseguido pelos métodos de ponderação, explicará o maior benefício da modificação por *Threshold* na combinação pelo produto. É interessante constatar que, através desta modificação, os perfis da regra do produto assemelham-se aos da soma, e que os seus resultados de localização aproximam-se igualmente dos resultados obtidos pela soma. Estes factos sugerem que, de entre os métodos estudados, a operação soma se afigura como o padrão de melhor desempenho.

A análise relativa à granularidade do ambiente, apresentada em 3.5.1, conduziu a duas conclusões fundamentais. A primeira diz respeito ao uso de modelos não balanceados, que se verificou ser prejudicial ao desempenho dos combinadores algébricos. A segunda conclusão refere-se ao efeito da granularidade em modelos balanceados. Neste caso verificou-se que há em geral perda de desempenho quando se avança para os extremos de granularidades baixas ou de granularidades altas. Uma constatação importante sobre os resultados de 3.5.1 é a de que a dependência daqueles dois factores – balanceamento dos modelos e granularidade – é, em parte, atenuada pela aplicação das modificações às regras algébricas. Este dado realça a importância destas extensões, por exemplo, em ambientes que pela natureza da sua arquitectura são intrinsecamente não balanceados, como é o caso do *dataset* IDOL.

Por fim, um resultado relevante deste capítulo prende-se com a utilização da modificação por *Threshold* na operação soma. Esta modificação foi desenvolvida com vista a melhorar a combinação pelo produto, e não havia sido anteriormente aplicada à soma. O presente estudo mostrou que este método é em geral tão eficaz como a ponderação, colocando-o entre os métodos a considerar no desenvolvimento de combinadores baseados na regra da soma.

4. Métodos de redução do peso computacional

4.1 Introdução

Neste capítulo desenvolvem-se os dois temas que constituem os principais obstáculos à utilização da representação NQ: os requisitos de memória e o tempo de computação do algoritmo. Numa discussão sobre estes temas deve, em primeiro lugar, fazer-se notar que eles estão relacionados. Por um lado, os requisitos de memória são impostos pelo número de descritores usados no modelo do ambiente e, por outro, o tempo de computação está dependente do número de comparações realizadas com esses descritores. Daqui se conclui que uma redução conseguida no tamanho de memória é acompanhada de uma redução nos custos de computação.

O segundo aspecto a ter em conta nesta discussão é o de que a limitação imposta por estes factores depende da dimensão do ambiente e se, para ambientes de larga escala, eles podem constituir um obstáculo, em ambientes de reduzida ou média dimensão, os valores envolvidos serão, provavelmente, aceitáveis. Tal como nos anteriores capítulos, também neste a avaliação das técnicas introduzidas é feita sobre os *datasets* IDOL e FDF Park. Em si, estes *datasets* constituem uma amostra limitada dos ambientes em que um robô móvel pode ser implantado, no entanto, é possível a partir deles obter uma avaliação comparativa entre a representação NQ e aquela que é genericamente considerada a mais eficiente, a representação Q.

Em ambientes modelados na forma de mapas topológicos o número de lugares é geralmente reduzido (da ordem das dezenas, nas nossas experiências), levando a que os requisitos de memória da representação Q não sejam dominados pela assinatura dos lugares mas antes pela dimensão do vocabulário visual. O termo dominante na memória usada tem por isso a dimensão $n_c \times s_d$, em que n_c é a dimensão do vocabulário e s_d o número de bytes ocupados por um descritor (128 bytes nos descritores SIFT). Também devido ao número reduzido de lugares, o tempo de computação desta representação não é dominado pela comparação de assinaturas visuais dos lugares, mas pela quantização dos descritores da imagem de teste, que envolve a sua comparação com os descritores do vocabulário. O custo desta operação tem ordem $O(nf \cdot n_c)$, onde nf designa o número de características da imagem de teste. As Figuras 4.1 e 4.2 ilustram os mecanismos inerentes às representações Q e NQ, respecti-

vamente, e evidenciam, para cada uma, os dados que é necessário manter em memória e as operações realizadas em fase de localização.

Na localização baseada na representação NQ, quando encarada na sua forma mais simples − apresentada no capítulo 2 − , os custos envolvidos podem ser radicalmente superiores aos da representação Q. Esta relação deve-se ao facto de os descritores guardados serem os de *todas* as características extraídas das imagens de modelação, correspondendo a um espaço de memória de dimensão $nl \times np \times s_d$, onde *nl* designa o número médio de descritores que representam um lugar e *np* o número de lugares. O custo computacional, que se deve neste caso à comparação um-a-um dos descritores da imagem de teste com os dos lugares, apresenta ordem O($nf \cdot nl \cdot np$). Destas expressões é fácil concluir que, para aproximar a complexidade da representação NQ da verificada na representação Q, é desejável reduzir os termos *nl* e *np*. Neste capítulo são introduzidos métodos que visam actuar sobre cada um destes termos.

A primeira categoria de métodos tem por objectivo reduzir *nl*, explorando, para isso, as especificidades das imagens de modelação de lugares. A motivação para o desenvolvimento destes métodos encontra-se, de facto, no contraste existente entre os dados típicos de uma tarefa de classificação visual genérica e os dados de modelação de um ambiente. Enquanto no primeiro caso os dados de modelação de uma categoria de objectos são imagens de *diferentes* instâncias dessa categoria, no segundo, os dados são imagens recolhidas em sequência do mesmo objecto, i.e., do mesmo lugar. Esta especificidade permite prever que nas imagens de modelação de lugares existirá redundância entre características que são visualizadas em diferentes imagens, o que sugere uma estratégia na redução dos modelos, pela fusão de características repetidas. Por outro lado, é possível dizer-se que as características que não se repetem entre imagens consecutivas são pouco robustas e contribuirão menos significativamente para a descrição do lugar. Sob estes dois pressupostos, serão desenhadas duas técnicas que visam reduzir *nl* e manter a precisão do localizador próxima da obtida com os modelos originais.

A segunda categoria de métodos compreende aqueles que reduzem o valor de *np*. Na apresentação destes métodos é necessário referir, em primeiro lugar, que estes se dirigem à redução da computação em fase de localização, i.e., a redução obtida em *np* é apenas aplicável à complexidade do algoritmo. No que diz respeito ao custo de

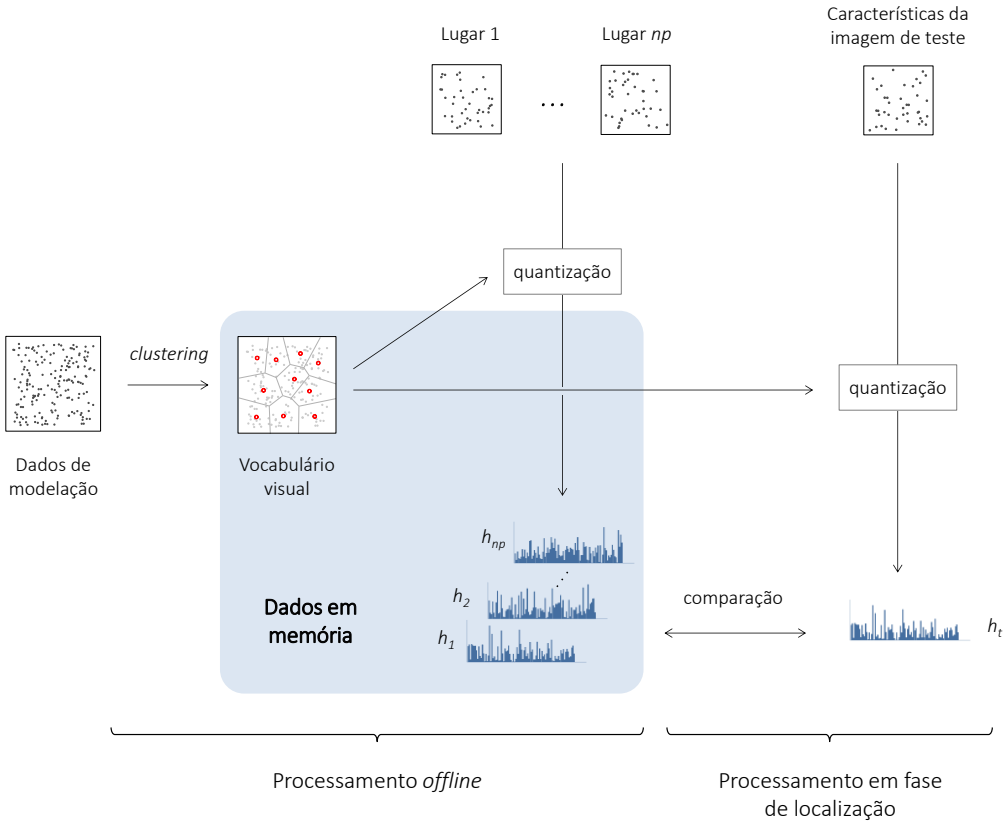


Figura 4.1. Resumo dos dados e operações envolvidos na representação Q.

representação, todos os lugares têm que ter representação em memória e o valor de *np* no cálculo deste custo é o número total de lugares.

Essencialmente, os métodos em causa produzem uma simplificação da computação através da escolha criteriosa dos lugares que serão processados pelo algoritmo de localização. Este objectivo é atingido aplicando uma estratégia hierárquica em que os lugares são seleccionados, numa primeira fase, recorrendo à aparência global da imagem, codificada na característica Gist. Numa fase posterior essa selecção é ainda depurada, à medida que as características locais da imagem de teste são processadas.

Este capítulo está organizado da seguinte forma: a secção 4.2 cobre os trabalhos relacionados mais relevantes; na secção 4.3 apresentam-se os métodos de redução dos modelos do ambiente; a secção 4.4 apresenta os métodos de selecção de lugares e a análise dos critérios de decisão associados; a secção 4.5 avalia o impacto dos métodos propostos sobre o desempenho na localização e compara as representações Q e NQ em termos de precisão e custos computacionais; por fim, a secção 4.6 sumariza os resultados do capítulo.

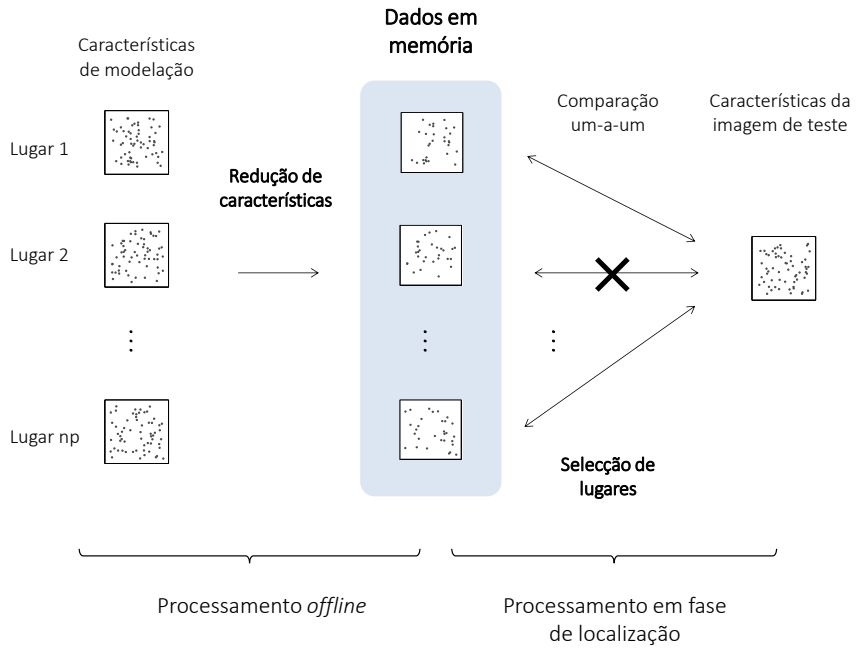


Figura 4.2. Resumo dos dados e operações envolvidos na representação NQ.

4.2 Trabalhos relacionados

Apesar da importância do custo computacional, alguns dos primeiros trabalhos que usaram características locais (Se, Lowe e Little, 2002; Li, Yang e Kosecka, 2005) não contemplam a redução do modelo do ambiente, podendo ser entendidos como provas-de-conceito onde a preocupação de escalabilidade não está presente. Aquele que é provavelmente o primeiro trabalho a revelar esta preocupação foi apresentado por Williams e Ledwich (2004). Na sua investigação os autores propõem a eliminação da invariância à rotação das características SIFT, evitando assim a multiplicidade de características que frequentemente são extraídas sobre a mesma posição na imagem. Esta é uma estratégia que foi também adoptada nesta tese. Outra abordagem, em que se realiza igualmente a selecção de características, foi proposta por Zhang (2011). No seu estudo, Zhang baseia esta selecção em dois factores, a escala da característica e a sua repetibilidade, o que se demonstrou eficaz numa tarefa de detecção da revisitação de lugares, mas não está demonstrado que resulte na tarefa de localização, em que há maiores variações de aparência.

Modelos reduzidos do ambiente também podem ser obtidos pela utilização de descritores mais compactos. Para este objectivo, existem duas linhas de investigação, activas nos últimos anos, que podem ser úteis. A primeira é aquela que se tem dedicado à transformação de descritores existentes, como o SIFT, em representações

mais compactas e é exemplificada nos trabalhos de Stommel e Herzog (2009), Ventura e Hollerer (2011), Diephuis et al. (2011) e Peker (2011). Outros investigadores têm adoptado a segunda abordagem, que consiste em desenhar de raiz características intrinsecamente mais compactas. Esta abordagem tem produzido alguns resultados muito relevantes, de que se destacam as características BRIEF (Calonder et al., 2010), ORB (Rublee et al., 2011) e BRISK (Leutenegger, Chli e Siegwart, 2011), codificadas em descritores binários que, para além da reduzida dimensão, permitem a comparação rápida pela distância Hamming. Apesar das boas propriedades destes descritores, os trabalhos que utilizam a representação Q têm preferencialmente usado o descritor SIFT, para o qual o procedimento de *clustering* com a distância euclideana está bem estudado. Uma vez que um dos objectivos desta tese é a comparação daquela representação com a representação NQ, optou-se por usar o descritor SIFT nas duas.

O tipo de simplificação dos modelos desenvolvida neste capítulo, por via da fusão de características, encontra, paradoxalmente, um paralelo com o trabalho de Filliat (2007) no âmbito do modelo BoW. Naquele estudo, o vocabulário visual é construído de forma incremental durante a exploração do ambiente: à medida que as características são recolhidas, elas são comparadas com as palavras visuais existentes e, se a menor distância exceder um limite, é inicializada uma nova palavra. Apesar das semelhanças existentes, naquele trabalho o modelo BoW não é abandonado e, ao contrário da nossa abordagem, as palavras visuais não ficam associadas a um lugar específico, sendo usadas, quando possível, na quantização de características provenientes de qualquer lugar visitado após a sua inicialização.

Na segunda categoria de métodos para a redução da computação faz-se a selecção dos lugares que são processados em fase de localização. O tipo de estratégia, hierárquica, que estrutura essa selecção tem sido aplicada sob diversas formas noutros trabalhos. Uma abordagem popular é a que adia a verificação pela geometria epipolar para uma fase tardia do algoritmo, sendo inicialmente aplicado um teste de semelhança, usando o modelo BoW, que permite eliminar a maioria dos lugares candidatos (Wang, Cipolla e Zha, 2005). Noutros trabalhos, a divisão do algoritmo em diversas etapas está relacionada com a utilização de diferentes descritores visuais em cada uma delas. Por exemplo, em (Murillo et al., 2007) a selecção inicial é baseada em descritores globais de cor e a fase seguinte na comparação de características locais pelo método

pyramidal matching. A utilização de diferentes descritores é também a ideia que está na base dos métodos propostos neste capítulo, com a particularidade de *i*) na fase inicial ser usado o descritor Gist (Oliva e Torralba, 2001), de extracção rápida e boa discriminatividade e *ii*) na fase posterior a selecção ser continuamente refinada, à medida que os descritores locais são processados.

Associada à selecção de lugares está a aplicação de um critério que, mediante um descritor da imagem de teste e um conjunto de descritores de diferentes lugares, decida sobre aqueles a eliminar. Este tópico relaciona-se com o problema da pesquisa de imagens por exemplos, extensivamente estudado na área da visão computacional. Este problema parte da apresentação de uma imagem de exemplo *q* (*query*), fornecida por um utilizador, em resultado da qual deve ser apresentada uma lista de imagens, i_1, ..., i_k, existentes numa base de dados. Tipicamente, a determinação desta lista é baseada na pontuação atribuída a cada imagem *i*, a qual resulta da aplicação de uma função de distância ou de semelhança entre imagens. Os estudos sobre estes sistemas têm-se focado essencialmente sobre a qualidade da ordenação das imagens, cabendo ao utilizador a escolha do número de imagens a inspeccionar. Este cenário é distinto do presente problema, já que a selecção dos lugares é o principal objectivo e esta deve ser feita de forma automática. Apesar de esta questão ter sido frequentemente ignorada na área da visão computacional, alguns estudos no campo da pesquisa de textos têm-se focado sobre o número *k* de documentos a devolver ao utilizador (Zhang e Callan, 2001; Arampatzis, Kamps e Robertson, 2009). Nesses trabalhos, quer a ordenação dos documentos, quer a sua selecção é baseada na noção de relevância, a qual reflete o interesse que um documento apresenta para o utilizador. Segundo alguns autores, esta definição de relevância envolve alguma subjectividade, já que resulta da avaliação pessoal que o utilizador faz de um resultado da pesquisa (Crestani et al., 1998). Por esta razão, todos os sistemas que não se baseiam na supervisão pelo utilizador (e que têm interesse no sentido de se obter sistemas completamente autónomos) adoptam algum tipo de pressuposto sobre o conceito de relevância. Para além deste grau de subjectividade, existe incerteza associada à pesquisa de informação, em virtude de, tipicamente, o utilizador procurar documentos relacionados com a informação introduzida, em lugar de documentos *exactamente* iguais. Por forma a lidar com essa incerteza, foram desenvolvidos diversos modelos probabilísticos, em que a relevância é entendida como uma variável aleatória. Dentro

deste paradigma, uma das abordagens mais populares assenta no princípio de que a informação patente na distribuição de pontuações é suficiente para se concluir sobre a relevância dos documentos (Arampatzis e Robertson, 2011).

A ideia fundamental que é explorada naquela abordagem, proposta inicialmente por Swets (1963), é a de que a distribuição de pontuações pode ser entendida como uma mistura de duas distribuições, correspondentes aos documentos relevantes e não relevantes. Na prática esta ideia torna-se útil quando se admite que as duas distribuições têm uma forma paramétrica e que esta pode ser estimada com base na amostra resultante de uma pesquisa. Por exemplo, Arampatzis, Kamps e Robertson (2009) apresentaram um método para a estimação de uma mistura em que a distribuição de documentos relevantes é modelada por uma função normal e a de documentos não relevantes tem a forma exponencial. No mesmo trabalho os autores vão mais além, mostrando que o conhecimento daquelas distribuições permite seleccionar o valor de *k* óptimo, mediante critérios baseados nos valores de *recall*, precisão ou outras medidas de desempenho. Um aspecto interessante desta estratégia é o da selecção dinâmica de *k*, que é calculado para cada evento de pesquisa, em função da informação de entrada e da distribuição de pontuações que ela induz. Num trabalho de pesquisa multimodal texto/imagem (Arampatzis, Zagoris e Chatzichristofis, 2011), mostrou-se que este aspecto confere maior eficácia e robustez ao sistema de pesquisa, quando comparado com a utilização de *k* fixo. Apesar do sucesso desta abordagem, a sua aplicabilidade depende da possibilidade de se estimarem os parâmetros das distribuições (Arampatzis, Zagoris e Chatzichristofis, 2011), o que pode inviabilizá-lo no que diz respeito ao problema de localização. De facto, lembrando que o número de lugares que compõem um ambiente é normalmente muito menor que o número de documentos numa base de dados de imagens, conclui-se que, na selecção de lugares, a estimação das distribuições seria necessariamente imprecisa, devido à pequena dimensão da amostra.

Na literatura encontram-se, no entanto, abordagens alternativas, que levam também à selecção de dinâmica de documentos e que, por serem mais simples, contornam as dificuldades da anterior. Uma perspectiva útil é dada por Wong e Yao (1995) que relacionam a pesquisa de informação à inferência probabilística sobre documentos. Nesse trabalho introduzem-se duas medidas de relevância (*R*) que, quando usadas como critério de selecção, refletem diferentes preocupações na pesquisa de infor-

mação. A primeira medida faz equivaler a relevância à probabilidade posterior de um documento:

$$R(i,q) = P(i|q) \tag{4.1}$$

e é entendida pelos autores como uma medida de relevância orientada para o valor de *recall*. Esta distribuição é também adoptada por Vasconcelos (2004), estando no centro do seu sistema *Minimum Probability of Error Image Retrieval*. A outra medida, descrita por Wong e Yao (1995) como orientada para a precisão faz equivaler a relevância à verosimilhança de q dado i:

$$R(i,q) = P(q|i). \tag{4.2}$$

Neste capítulo serão considerados 4 critérios candidatos para a selecção de lugares:

i) O primeiro aplica um limite de decisão à função de pontuação. Este método pode ser interpretado como uma aplicação da Eq. (4.2), se se considerar que a conversão de pontuação para verosimilhança é uma função monótona;

ii) O segundo método não avalia o número de lugares a escolher como função dos valores de pontuação, devolvendo sempre um número fixo k dos lugares mais pontuados;

iii) O terceiro método é a aplicação de Eq. (4.1), e

iv) O quarto método é uma regra empírica, desenhada para a selecção através da característica Gist, e será descrita em pormenor na secção 4.4.1.

4.3 Compactação dos modelos dos lugares

4.3.1 Fusão de características

Através deste método, designado por FM (*Feature Merging*), obtém-se a redução de *nl* por meio da fusão de características semelhantes no mesmo descritor. Embora a determinação de características semelhantes possa ser realizada com o auxílio da geometria epipolar, neste trabalho propôs-se um método mais simples e que se revelou adequado, uma vez que proporciona uma forte compactação dos modelos sem perda de desempenho significativa.

Neste método, descrito na Listagem 1, o conjunto C_j das características do lugar j é construído de forma progressiva, à medida que as imagens do lugar são processadas

Listagem 1. Construção do modelo do ambiente pelo método FM

```
Para cada lugar l_j
    Inicializar C_j com as características da primeira imagem do lugar;
    Inicializar os contadores c_1, …, c_nlj destas características com o valor 1;
    Para cada imagem i seguinte
        Para cada característica d da imagem
            Encontrar a característica d_k em C_j mais próxima;
            Se dist(d_k,d)<d_FM
                d_k= d_k×c_k/(c_k+1) + d/(c_k+1);
                c_k=c_k+1;
            caso contrário,
                adicionar d ao conjunto C_j;
            Fim
        Fim
    Fim
Fim
```

na mesma sequência em que foram recolhidas. Após a inicialização deste conjunto com as características da primeira imagem, cada iteração consiste na comparação das características da nova imagem com as de C_j e, para aquelas em que a distância é inferior a um valor limite pré-especificado, d_{FM}, realiza-se a fusão com as características correspondentes, através da média. As características que não encontram correspondência pelo critério anterior são adicionadas ao conjunto C_j.

Uma questão central na concretização deste método é a da selecção da distância limite. Este valor deve ser tal que uma grande percentagem de correspondências correctas seja abrangida, ao mesmo tempo que a percentagem de correspondências incorrectas é mantida baixa. A escolha deste parâmetro pode ser fundamentada na análise das distribuições das distâncias observadas em correspondências correctas e incorrectas, em condições em que a *ground truth* é conhecida. Com vista a proceder a esta análise, determinaram-se as correspondências entre imagens consecutivas da sequência *dum_sunny1* do *dataset* IDOL, aplicando a verificação geométrica baseada na homografia, já usada no capítulo 3 desta tese (ver secção 3.3.2). Para além das distâncias entre as características seleccionadas por este processo, calculou-se a distância entre as características mais próximas que não verificam a condição geométrica. A Figura 4.3.a apresenta as distribuições da distância entre características correspondentes e não-correspondentes, as quais foram estimadas na forma de

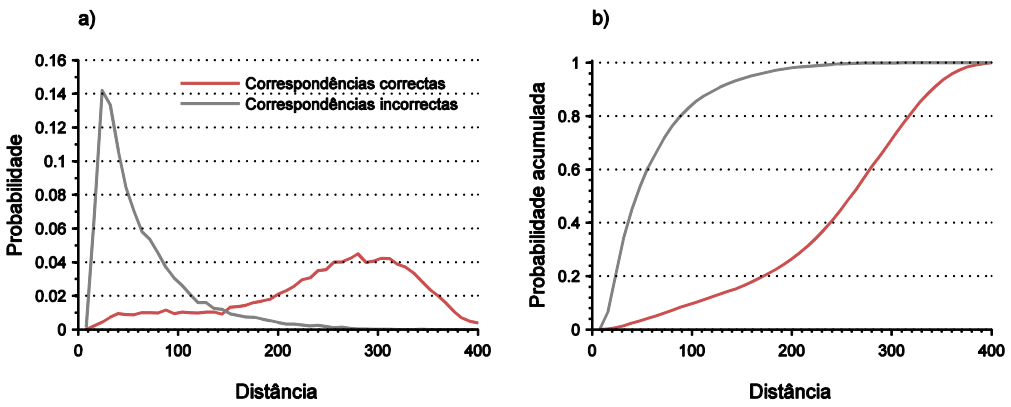


Figura 4.3. Distribuição das distâncias entre características correspondentes e não-correspondentes (à esquerda). À direita: distribuições acumuladas.

histogramas com intervalos de 10. Adicionalmente, a Figura 4.3.b mostra a distribuição acumulada em cada um dos casos.

Dada a forma daquelas distribuições, seleccionou-se d_{FM}=100 como um valor de compromisso, que garante a detecção de cerca de 85% das correspondências correctas e para o qual a detecção de correspondências incorrectas é de 10%.

As Figuras 4.4 e 4.5 ilustram os efeitos da aplicação deste método, apresentando os valores totais das características contidas nos modelos do ambiente, antes e depois da sua aplicação. Para cada uma das sequências de modelação são apresentados estes valores bem como o factor de redução correspondente. Os dados apresentados mostram que a fusão de características é muito eficaz na redução de características em ambos os *datasets*, resultando em factores de redução entre 0.4 e 0.5 no *dataset* IDOL e entre 0.45 e 0.55 no *dataset* FDF Park.

Uma outra perspectiva deste método é dada na Figura 4.6, onde se apresentam os histogramas dos factores de redução medidos em todos os lugares dos *datasets* IDOL e FDF Park. Assim, enquanto os dados das Figuras 4.4 e 4.5 apresentam factores de redução globais, esta figura é indicativa da forma como os factores de redução se distribuem pelos lugares de um ambiente. Os dados da Figura 4.6 mostram que no *dataset* FDF Park a distribuição subjacente é aproximadamente unimodal e centrada em 0.5, no entanto, a mesma distribuição apresenta maior dispersão no caso do *dataset* IDOL, a qual é claramente multimodal. Neste caso, muitos lugares apresentam factores de redução vantajosos, próximos de 0.4, mas verifica-se também que cerca de metade dos lugares têm factores de redução menos significativos, entre 0.6 e 1.

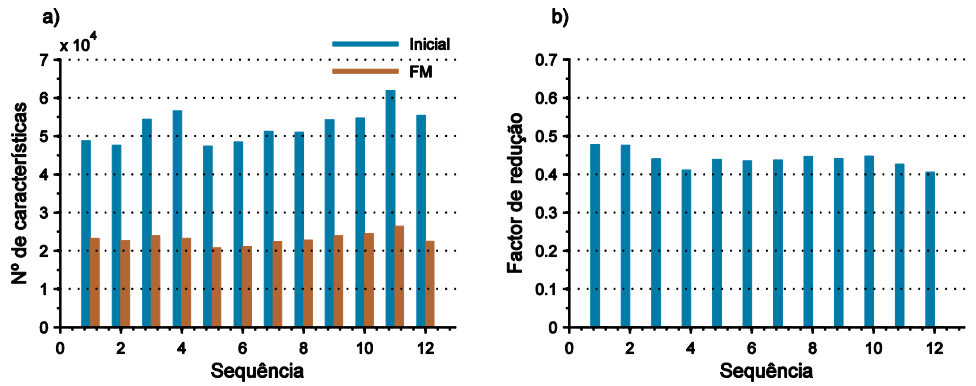


Figura 4.4. Redução do nº de características pelo método FM, no *dataset* IDOL. À esquerda: número de características contidas nos modelos do ambiente construídos sobre cada uma das sequências do *dataset*; à direita: factor de redução obtido pela aplicação do método FM.

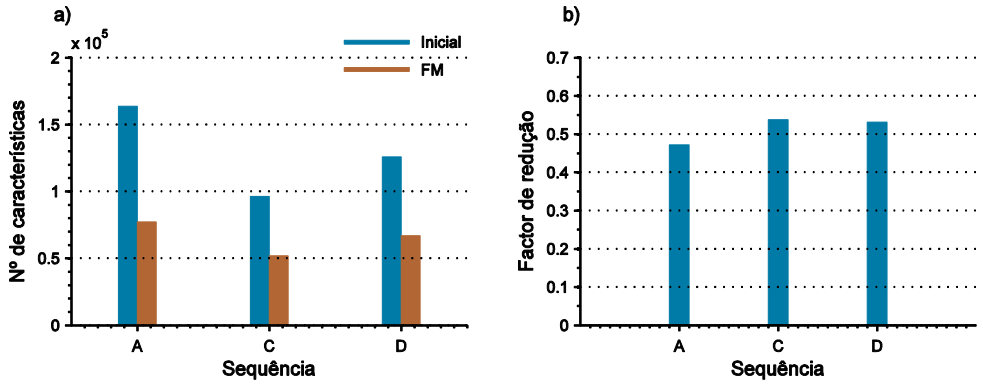


Figura 4.5. Redução do nº de características pelo método FM, no *dataset* FDF Park. À esquerda: número de características contidas nos modelos do ambiente construídos sobre cada uma das sequências do *dataset*; à direita: factor de redução obtido pela aplicação do método FM.

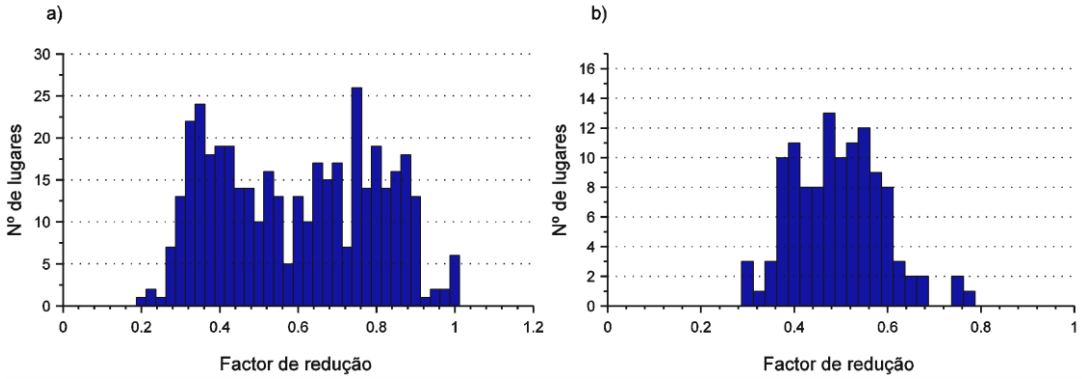


Figura 4.6. Ocorrência dos factores de redução (método FM) sobre os lugares dos *datasets* IDOL (à esquerda) e FDF Park (à direita).

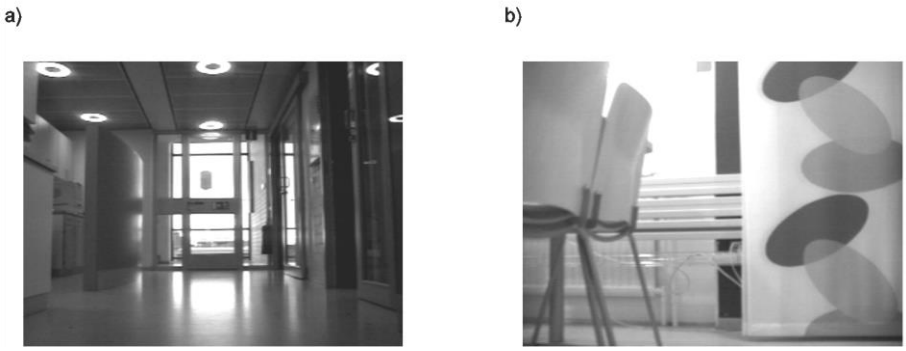


Figura 4.7. Exemplos de lugares com factores de redução díspares, respectivamente de a) 0.31 e b) 0.84.

Presume-se que estas diferenças estão relacionadas com o tipo de lugar e com a sequência de imagens que o caracteriza. Exemplos típicos de lugares com factores de redução díspares são dados na Figura 4.7, onde se apresentam imagens de lugares com valores de 0.31 e 0.84. Pela inspecção destas imagens verifica-se que a fusão de características é mais significativa nas sequências em que o robô tem movimento de avanço, normalmente através de um corredor (Figura 4.7.a). Nestes casos, a aparência das imagens muda suavemente entre *frames*, promovendo a repetição de características ao longo da sequência. Em contraste, a Figura 4.7.b, mostra um espaço confinado, em que o movimento do robô necessariamente inclui mudança de orientação. Neste caso, a mudança de aparência entre imagens é mais abrupta e consequentemente o número de características repetidas entre imagens é menor. No *dataset* FDF Park o movimento é apenas de avanço em espaço aberto, pelo que os factores de redução são mais consistentes do que no *dataset* IDOL. Observou-se também que os lugares do *dataset* IDOL em que o factor de redução é menos significativo incluem, tipicamente, poucas imagens, pelo que o seu impacto na compactação global é reduzido.

4.3.2 Eliminação de características

Este método parte do pressuposto de que as características que não foram sujeitas a fusão pelo método anterior apresentam reduzida repetibilidade, i.e., são identificadas de forma pouco robusta pelo detector de pontos de interesse. Numa primeira versão deste método, apresentada em (Campos, Correia e Calado, 2011) estas características eram simplesmente eliminadas dos modelos. A forte compactação dos modelos obtida por esta via é no entanto acompanhada de alguma perda de desempenho na classi-

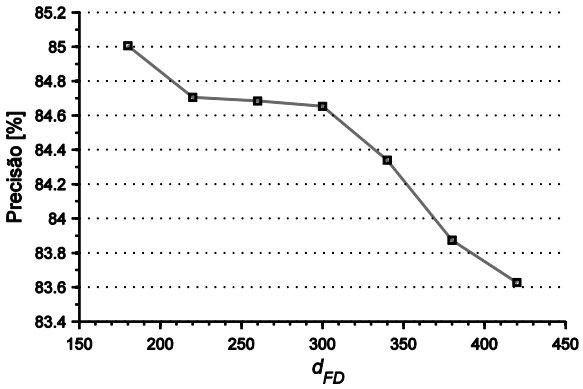


Figura 4.8. Precisão na localização vs limite de fusão d_{FD} usado no método FD.

ficação, pelo que foi desenvolvida uma nova versão deste método, com vista a obter-se um equilíbrio entre os dois efeitos.

Tal como na versão original, o método proposto incide sobre as características que não foram sujeitas a fusão pelo método FM mas, em lugar de as eliminar, recorre à sua aglomeração por forma a reduzir *nl*. O método de aglomeração aplicado é idêntico ao usado pelo método FM, com a diferença de a distância limite ser agora maior, permitindo a fusão de características com menor semelhança. Embora nesta versão não ocorra efectivamente eliminação de características, usaremos a designação FD (*Feature Deletion*) para este método, tal como em (Campos, Correia e Calado, 2011). De facto, na presente versão aquela designação também é aplicável, já que a fusão de características muito díspares altera significativamente o seu conteúdo.

A Figura 4.8 mostra a precisão média obtida sobre a sequência *dum_sunny1* do *dataset* IDOL, em função da distância limite para aglomeração, d_{FD}. Para valores elevados de d_{FD} a fusão resulta na redução drástica do número de características, aproximando este mecanismo da eliminação completa de características. Como é visível, para estes valores há alguma perda de desempenho relativamente a uma fusão criteriosa, determinada por valores de d_{FD} mais baixos. Dada a tendência observada naquela curva, seleccionou-se d_{FD}=300, como um valor de equilíbrio entre perda de desempenho e compactação dos modelos.

Os dados apresentados nas Figuras 4.9 e 4.10 caracterizam o efeito da aplicação do método FD sobre o número total de características. Estes resultados mostram que, após a aplicação do método FM, há uma grande percentagem de características que não foram sujeitas a fusão e que são alvo do método FD. Pela aplicação deste método

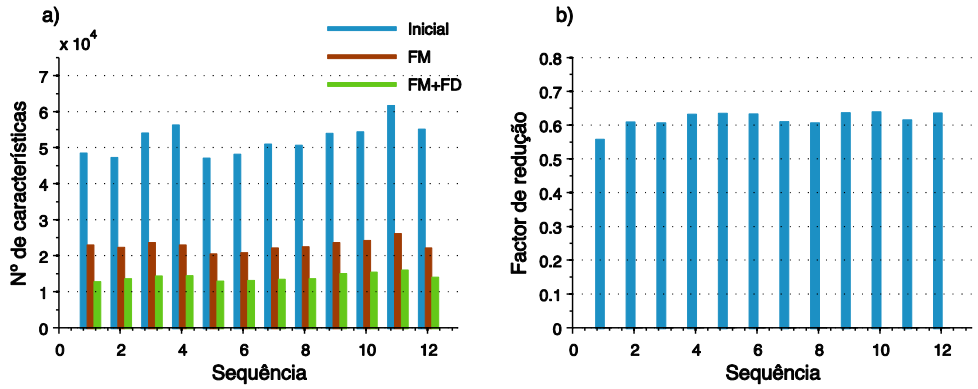


Figura 4.9. Redução do nº de características pelo método FD, no *dataset* IDOL.A) número total de características nas diversas versões dos modelos; b) factor de redução obtido pelo do método FD.

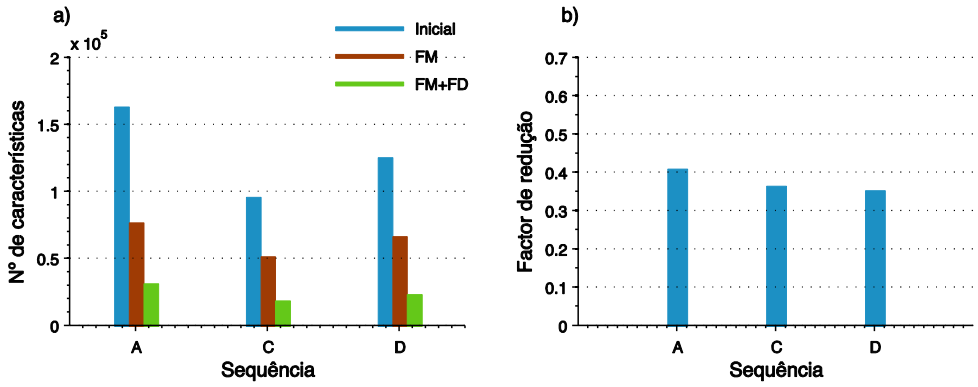


Figura 4.10. Redução do nº de características pelo método FD, no *dataset* FDF Park. A) número total de características nas diversas versões dos modelos;b) factor de redução obtido pelo método FD.

o número total de características é significativamente reduzido, por factores entre 0.5 e 0.6 no *dataset* IDOL e 0.35 e 0.45 no *dataset* FDF Park. O impacto que esta redução pode induzir no desempenho de localização será analisado na secção 4.5.

Relativamente à forma como o factor de redução se distribui pelos lugares, os histogramas da Figura 4.11 mostram que, dentro do mesmo *dataset*, este método é mais consistente do que FM. Este aspecto dever-se-á à aplicação de d_{FD} consideravelmente superiores aos usados em FM, aliviando o método FD da dependência de existirem características semelhantes entre imagens.

4.4 Selecção de lugares

Esta secção refere-se à redução do peso computacional por via da redução de *np*, o número de lugares usados pelo algoritmo. Como mencionado na introdução do capítulo, a selecção de lugares é fundamentada, numa primeira fase, na informação

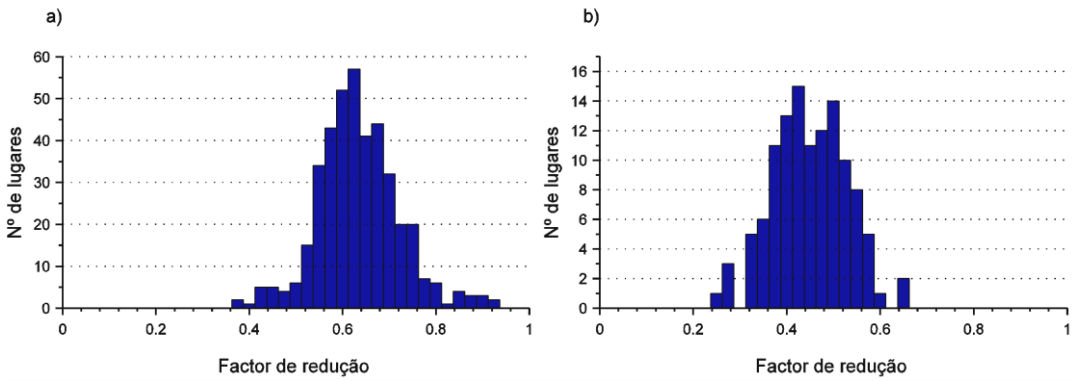


Figura 4.11. Ocorrência dos factores de redução (método FD) sobre os lugares do *dataset* IDOL (à esquerda) e FDF Park (à direita).

proveniente da característica Gist (secção 4.4.1) e, posteriormente, na informação das características locais (secção 4.4.2). Em cada uma destas fases é aplicado um critério de decisão, que determina os lugares a ignorar em função da informação disponível. Nesta secção são avaliados diversos critérios, usando a sequência *dum_sunny1* como modelo do ambiente, o que conduzirá à escolha de um critério para cada uma das fases e à sua parametrização. Posteriormente, na secção 4.5 estes métodos são validados sobre todas as sequências dos *datasets*.

4.4.1 Selecção pela característica Gist

A característica Gist visa condensar o conteúdo global de uma imagem num descritor que, apesar de não oferecer a robustez da análise por características locais, possibilita a análise rápida da aparência visual. O método GS (*Gist Selection*) explora esta propriedade, com o objectivo de avaliar de forma expedita a semelhança entre imagens e restringir a avaliação subsequente apenas aos lugares mais relevantes.

Sendo originalmente associada à análise de texturas com filtros de Gabor (Torralba, 2003), nesta tese a característica Gist é extraída através do método LBP (Ojala, Pietikäinen e Harwood, 1996), o qual é mais interessante pela sua rapidez de execução. Segundo este método, a textura associada a um píxel é identificada pela análise de uma vizinhança de raio R e considerando n píxeis dessa vizinhança. Mais detalhes sobre este método serão apresentados no capítulo 5, em que ele é aplicado ao problema da detecção da revisitação de lugares. Para efeitos de selecção de lugares, o operador LBP foi parametrizado com n=8, R=5 e uniformidade 2, e a construção do descritor Gist envolveu as seguintes etapas: i) determinação do código LBP para cada píxel da imagem, ii) cálculo do histograma de códigos LBP em duas sub-janelas

horizontais e sem sobreposição e iii) concatenação dos dois histogramas, resultando num vector *g* de dimensão 118.

A selecção de lugares é feita comparando o descritor Gist da imagem de teste, *g*, com os descritores representativos dos lugares, através da distância Chi-quadrado, e aplicando um critério adequado sobre os valores de distância obtidos. Por forma a simplificar a comparação, cada lugar l_j é representado por um único descritor, g_j, calculado como a média dos descritores Gist das imagens usadas na modelação desse lugar.

Um aspecto determinante para a eficácia deste método é o critério usado na escolha dos lugares. Neste trabalho foram estudados 4 métodos candidatos, referidos de seguida como M1, … , M4. Qualquer um dos métodos visa a selecção de um conjunto *CL* de lugares, tendo por base a distribuição de distâncias, e é parametrizado por um valor *th* que estabelece o limite de aceitação ou rejeição de um lugar. Idealmente, o método adequado deve oferecer boa selectividade, isto é, escolher um pequeno número de lugares, entre os quais se encontra o lugar correcto.

O primeiro método considerado selecciona os lugares simplesmente comparando as distâncias com um valor fixo. Neste método o conjunto CL é definido por

$$\text{M1: } CL = \{l_j, dg_j < th_1\}.$$

Na expressão anterior e seguintes, dg_j designa a distância entre os descritores Gist da imagem de teste e do lugar l_j.

O segundo método devolve um número fixo de lugares, seleccionando os th_2 lugares com menor valor de distância:

$$\text{M2: } CL = \{l_1, l_2, \dots, l_{th_2}\}, \;\; dg_j < dg_{j+1}.$$

Enquanto os métodos anteriores são baseados nos valores de distância, M3 e M4 tomam a decisão com base em valores de probabilidade. Para este efeito, a probabilidade de ocorrência de um valor de distância é modelada como uma distribuição normal centrada em zero:

$$P(g|l_j) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{dg_j{}^2}{2\sigma^2}\right). \qquad (4.3)$$

Nesta expressão, usou-se o valor de desvio padrão σ=1700, estimado empiricamente por forma a obter-se um bom desempenho global dos métodos seguintes.

No método M3 usa-se uma regra empírica que pretende compensar a falta de robustez da característica Gist mediante variações de aparência. De facto, verificou-se que, sob condições de luminosidade diferentes, as distâncias relativas aos lugares relevantes crescem significativamente, o que justifica a aplicação de um limite que seja adaptável à imagem de teste apresentada. Neste método estima-se a probabilidade de observação das distâncias aos lugares relevantes pela média dos dois valores mais elevados de probabilidade, sendo o limite de selecção definido de forma relativa a essa estimativa. Assim, o conjunto CL é dado por:

$$\text{M3: } CL = \{l_j, P(g|l_j) > P_r - th_3\}.$$

Nesta expressão P_r designa a estimativa da probabilidade sobre os lugares mais relevantes, calculada por

$$P_r = 0.5\big(P(g|l_1) + P(g|l_2)\big), \tag{4.4}$$

com l_1 e l_2 designando os lugares com os valores mais elevados de $P(g|l_j)$.

O método M4 faz igualmente depender a selecção de um lugar da probabilidade sobre os restantes lugares, mas neste caso utilizando um *threshold* fixo e recorrendo às probabilidades posteriores, obtidas pela regra de Bayes:

$$P(l_j|g) = \frac{P(g|l_j)P(l_j)}{\sum_i P(g|l_j)P(l_j)}, \tag{4.5}$$

em que se utilizou uma distribuição à priori $P(l_j)$ uniforme. A selecção do conjunto CL é dada por:

$$\text{M4: } CL = \{l_j, P(l_j|g) > th_4\}.$$

Na avaliação dos quatro métodos recorreu-se às medidas de *recall* e *fall-out*, que quantificam respectivamente a capacidade de abranger os lugares relevantes e a capacidade de excluir os lugares não-relevantes.

Designando por N_{rs} e N_r respectivamente o número de lugares relevantes seleccionados e o número total de lugares relevantes, a medida de *recall* é definida por:

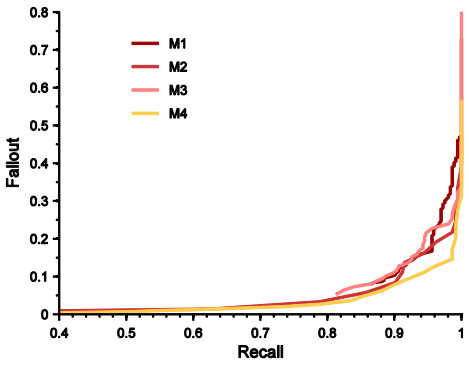


Figura 4.12. Curvas *fall-out* vs *recall* obtidas com cada um dos critérios de selecção.

$$recall = \frac{N_{rs}}{N_r}. \tag{4.6}$$

Esta medida, quando aplicada a um dado critério, pode ser entendida como a probabilidade que este apresenta de seleccionar lugares relevantes.

De forma semelhante, designam-se por N_{nrs} e N_{nr} respectivamente o número de lugares não relevantes selecionados e o número total de lugares não relevantes, conduzindo à definição de *fall-out* como:

$$fall - out = \frac{N_{nrs}}{N_{nr}}. \tag{4.7}$$

A interpretação desta medida é a da probabilidade de se selecionar um lugar não relevante e como tal deve ser tão baixa quanto possível.

A Figura 4.12 oferece uma perspectiva do desempenho de cada um dos métodos, através da curva *fall-out* vs *recall* medida com a sequência de modelação *dum_sunny1* e de teste *dum_night1*. A região deste gráfico que contém informação importante para efeitos de selecção de lugares é a vizinhança de *recall*=1, que corresponde às condições em que cerca de 100% dos lugares correctos estão incluídos no conjunto CL. Naturalmente, dentro desta vizinhança um método é considerado superior a outro se produzir menor valor de *fall-out*. Embora esta figura seja ilustrativa do comportamento dos 4 métodos, não oferece informação sobre a sua generalização a outras sequências de teste e modelação. Em particular, é possível que um valor de *threshold* que garante *recall* igual a 1 numa sequência de teste não o garanta noutra.

Com vista a analisar a robustez dos métodos, realizaram-se testes sobre 11 sequências de teste do *dataset* IDOL. Os resultados destes testes são sintetizados na Figura 4.13

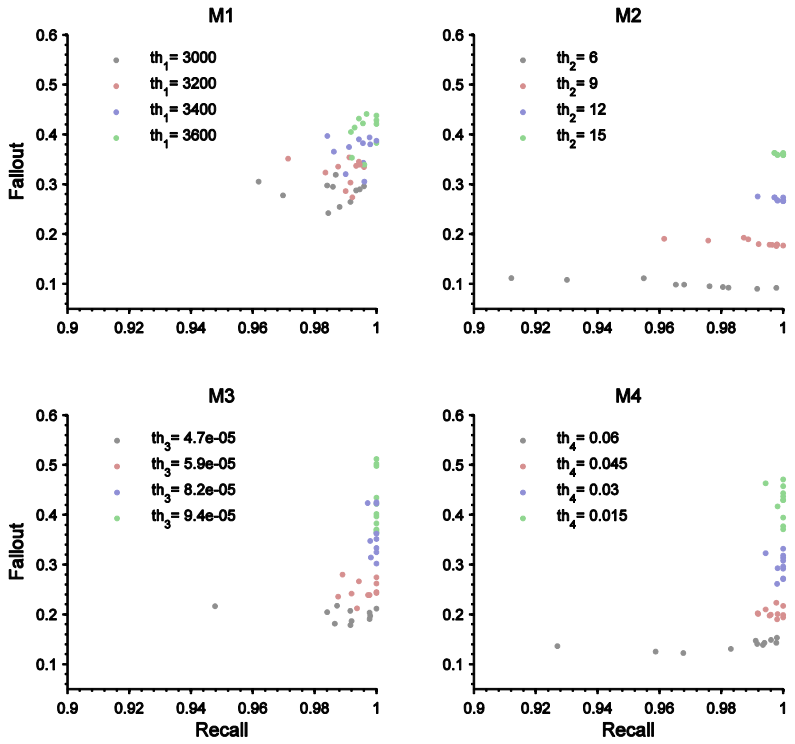


Figura 4.13. *Fall-out* vs *recall* medidos na selecção pela característica Gist, com diversos valores de *th* e sobre 11 sequências de teste.

onde se apresentam, no plano *fall-out* vs *recall*, as combinações destas variáveis que foram medidas para quatro valores de *th*.

Entre os métodos testados, M1 é o que apresenta menor robustez, evidenciando que a aplicação de um *threshold* fixo à distância entre descritores Gist é pouco flexível, dada a sensibilidade desta característica à variação global de aparência. O método M2 apresenta melhores propriedades do que o anterior, proporcionando valores de *recall* próximos de 1 para th_2=15. Contudo, devido ao número fixo de lugares seleccionados por este critério, não há uma adaptação do *threshold* às condições do problema. Este é um inconveniente deste método, já que a escolha de um *threshold* conservativo impede que, em situações de localização mais simples, seja rejeitado um número maior de lugares. A propriedade de adaptabilidade é proporcionada pelos métodos M3 e M4, que apresentam comportamentos idênticos neste sentido. Contudo, enquanto no método M4 não foi encontrado um valor de th_4 que garanta *recall*=1, em M3, com $th_3=9.4\times10^{-3}$, esta condição foi atingida nestes dados de teste. Por esta razão escolheu-se M3, e aquele valor de th_3, para os ensaios posteriores em que o método GS é aplicado à selecção de lugares.

4.4.2 Selecção progressiva

A ideia central do algoritmo PS (*Progressive Selection*) é de, após a selecção de lugares através do Gist, realizar um refinamento dessa selecção, à medida que as características locais são avaliadas. Para tal, o conjunto das características da imagem de teste é subdividido, e os conjuntos resultantes são tratados sequencialmente. Em cada etapa desse processo, é calculada a distribuição de probabilidades sobre os lugares, em função da qual se decide sobre a eliminação dos lugares candidatos menos plausíveis. Nesta secção foram considerados 3 métodos de selecção, idênticos aos métodos M2 a M4 do ponto anterior. O método M1 do ponto anterior não foi aqui considerado, visto que a selecção será baseada numa distribuição de probabilidades posteriores, enquanto aquele método assentava numa distribuição de distâncias.

A aplicação da selecção progressiva passa pelo ordenamento aleatório das características e pelo seu agrupamento em subconjuntos de *w* características. Em cada etapa *f* do algoritmo é calculada, à semelhança da Eq. (2.5), a distribuição posterior por

$$P_f\left(l_j\middle|d_1,\dots,d_{f\times w}\right)=\frac{1}{f\times w}\sum_{i=1}^{f\times w}P\left(l_j\middle|d_i\right). \tag{4.8}$$

Designando por *t* o conjunto de características $d_1, \dots, d_{f\times w}$, os métodos de selecção usados de seguida são definidos por:

M1- selecciona uma percentagem fixa do número de lugares avaliados na última etapa:

$$CL_f=\{l_1,l_2,\dots,l_k\},\ \ P_f\left(l_j\middle|t\right)>P_f\left(l_{j+1}\middle|t\right).$$

Definindo o número de lugares selecionados na etapa *f* por np_f, o valor de *k* é calculado por $k=th_1\times np_{f-1}$, onde $th_1 \in [0,1]$ é o parâmetro ajustável deste critério.

M2- dado o sucesso do método M3 usado no ponto anterior, testou-se aqui uma estratégia semelhante, que opera, neste caso, sobre probabilidades à posteriori:

$$CL_f=\left\{l_j,P_f\left(l_j\middle|t\right)>P_r-th_2\right\},$$

$$P_r=0.5\left(P_f(l_1|t)+P_f(l_2|t)\right).$$

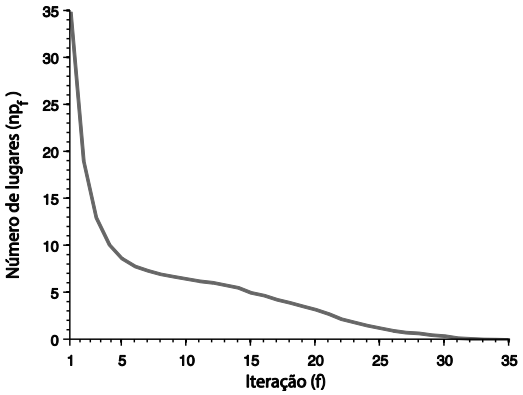


Figura 4.14. Evolução do número de lugares durante o processo de selecção progressiva.

M3- Finalmente, o método M3 selecciona lugares pela aplicação de um *threshold* à distribuição posterior. Na definição deste limite há a ter em conta que os níveis globais de $P_f(l_j|t)$ estão relacionados com o número de lugares envolvidos na etapa f (np_f). Por forma a contemplar essa relação, o limite foi definido como proporcional à probabilidade numa distribuição uniforme com np_f hipóteses:

$$CL_f = \left\{ l_j, P_f(l_j|t) > \frac{1}{np_f} th_3 \right\}.$$

A título de exemplo, na Figura 4.14 mostra-se a evolução de np_f obtida com a aplicação do método M3. Esta curva demonstra o efeito produzido pela selecção progressiva, colocando em evidência a redução faseada do número de lugares. A rapidez na redução de np_f é, em cada um dos critérios, determinada pelo parâmetro de *threshold*, o qual deve ser escolhido por forma a obter-se um compromisso entre rapidez e precisão na selecção dos lugares. Tal como no ponto anterior, a avaliação dos três critérios foi baseada nas medidas de *recall* e *fall-out*. Contudo, o facto de a selecção progressiva não se restringir a uma única operação mas a uma sequência de selecções, torna necessário redefinir estas medidas para este caso.

No que diz respeito à probabilidade de um lugar relevante ser escolhido (*recall*), a selecção a ter em conta deve ser o conjunto obtido na etapa final, já que é entre estes lugares que será escolhido um pelo classificador. Assim, na Eq. (4.6) a determinação do número de lugares correctos que foram seleccionados é feita sobre o conjunto obtido no final da selecção progressiva.

De modo diferente, o cálculo da medida *fall-out* deve reflectir a capacidade de serem excluídos lugares não relevantes em todas as etapas, já que o objectivo é simplificar a

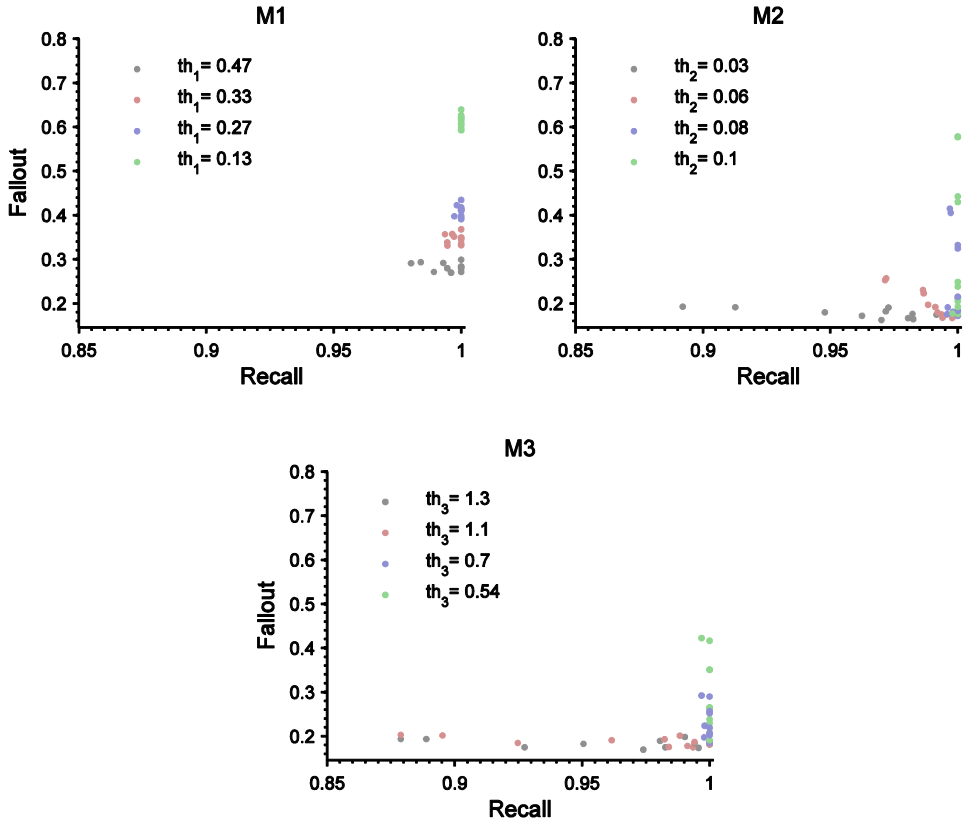


Figura 4.15. *Fall*-out vs *recall* medidos na selecção progressiva, com diversos valores de *th* e sobre 11 sequências de teste.

computação envolvida em todo o processo. Por esta razão, na Eq. (4.7) o valor de N_{nrs} usado é a média dos lugares não relevantes seleccionados em todas as etapas.

Tal como em 4.4.1 os critérios de selecção são analisados recorrendo às combinações (*recall, fall-out*) medidas num conjunto de testes em que se aplicaram diversos valores de *th*. Na Figura 4.15., onde se apresentam esses resultados, é visível que o método M1, como esperado, não realiza a selecção em função da dificuldade do problema, originando valores de *fall-out* semelhantes para o mesmo valor de *th*. A razão por que os valores de *fall-out* variam ligeiramente resulta de esta medida depender do número de características da imagem de teste, o qual é variável. Com M1, apenas se garante valores de *recall* próximos de 1 quando o *fall-out* é superior a 0.4. Os métodos M2 e M3 oferecem mais flexibilidade, garantindo, para valores de *th* adequados, *recall* próximo de 1 e valores de *fall-out* que, nas melhores condições, se aproximam de 0.2. Estes dois métodos apresentam comportamentos semelhantes,

tendo sido escolhido para os estudos seguintes o método M3, com th_3=0.7, por produzir valores de *fall-out* ligeiramente inferiores aos de M2.

4.5 Custos computacionais e precisão

Nesta secção são avaliados os métodos introduzidos neste capítulo, em termos dos seus custos computacionais e da sua precisão no reconhecimento de lugares. Estes resultados são comparados com os da representação Q, tendo sido todos os algoritmos programados no ambiente Matlab e executados num computador com as seguintes características:

- processador: Intel Core i5-4440
- velocidade de processamento: 3.10 GHz
- RAM: 16.0 GB.

As Figuras 4.16 e 4.17 apresentam os dados relativos aos tempos de computação e factores de redução dos quatro métodos. A aplicação dos quatro métodos foi feita de forma cumulativa, significando que, se tomarmos como exemplo o caso de PS, e sendo este o último método usado, os dados dizem respeito à aplicação em simultâneo dos quatro métodos. Os dados relativos ao *dataset* IDOL são divididos em dois grupos, por forma avaliar separadamente os desempenhos nas situações de localização mais simples e nas mais desafiadoras. No primeiro caso, incluem-se as condições em que a sequência de teste e modelação foram obtidas nas mesmas condições de luminosidade ou envolvendo as condições *sunny* e *cloudy*, que são semelhantes. No segundo caso incluem-se as combinações que envolvem a configuração *night* e qualquer uma das outras.

Da análise destas figuras, e considerando os métodos FM e FD, verifica-se que a redução nos tempos de computação está em linha com a redução no número de características dos modelos, obtida em 4.3. Globalmente, a aplicação dos dois métodos é bastante eficaz, resultando num factor de redução conjunto, dado pelo produto dos factores de redução individuais, de cerca de 0.21 nos modelos do *dataset* FDF Park e de 0.32 nos modelos do *dataset* IDOL.

Enquanto os métodos anteriores dependem apenas do modelo do ambiente, a eficácia dos métodos GS e PS está relacionada com os resultados da comparação da imagem de teste com o modelo. Por essa razão, a eficácia destes métodos apresenta uma varia-

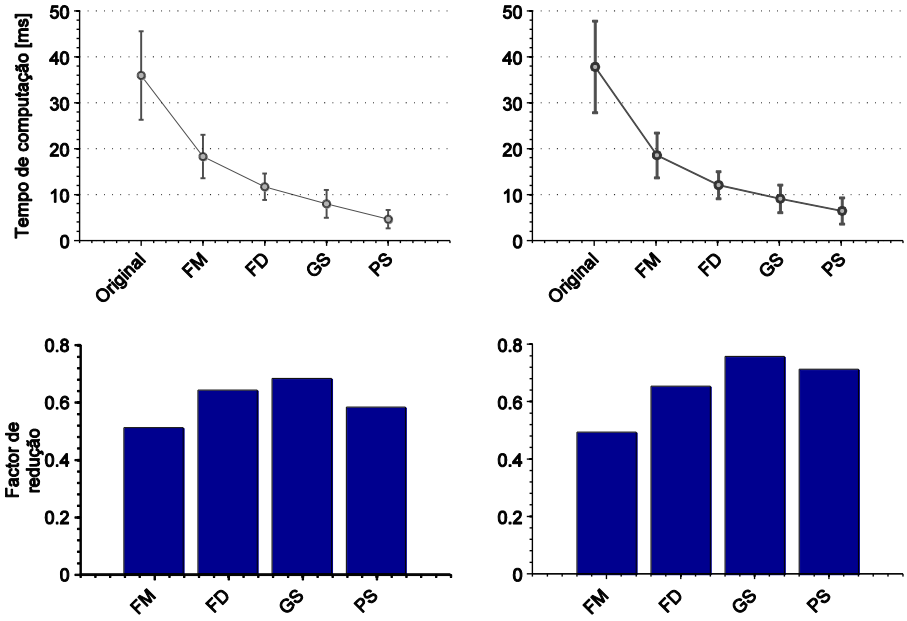


Figura 4.16. Tempos de computação e factores de redução obtidos no *dataset* IDOL. Os gráficos da primeira coluna referem-se a condições de luminosidade semelhantes; os da segunda coluna a condições de luminosidade díspares.

bilidade que está relacionada com a dificuldade do problema. No que diz respeito ao método GS, verifica-se, por exemplo, que nas sequências C e D do *dataset* FDF Park o método é menos eficaz, porque nestas sequências as variações de aparência entre as imagens do modelo e de teste são pouco toleradas pela característica Gist. A variabilidade do método PS é particularmente evidente no *dataset* IDOL, onde se verifica um factor de redução de 0.58 nas condições de luminosidade semelhantes e de 0.71 nas condições mais adversas. Tanto o método GS como PS envolvem alguma computação extra, que corresponde respectivamente à extracção e comparação da característica Gist e à actualização progressiva das distribuições de probabilidade. Os custos destas operações estão incluídos nos valores das Figuras 4.16 e 4.17 e são apresentados separadamente na Tabela 4.1.

As Figuras 4.18 e 4.19 apresentam os resultados de precisão associados a cada um dos métodos. Para além dos resultados obtidos com o classificador NQ simples (Eq. 2.5), apresenta-se, para o *dataset* IDOL, a precisão da modificação por *Threshold* (Eq. 3.15) que, como demonstrado no capítulo 3, beneficia significativamente o classificador. Com vista à comparação das representações Q e NQ, estas figuras incluem também os valores de precisão do classificador Naive Bayes (representação Q), obtidos com o mesmo conjunto de vocabulários usados em 2.5.1 (ver, por exemplo, as

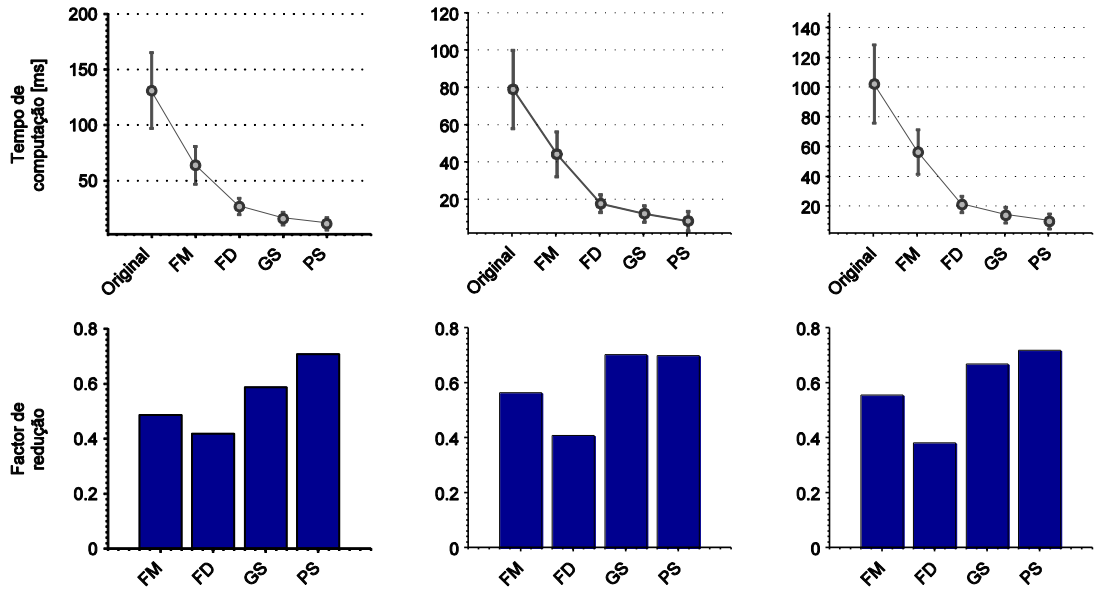


Figura 4.17. Tempos de computação e factores de redução obtidos no *dataset* FDF Park. As colunas 1 a 3 referem-se às sequências de modelação A, C e D respectivamente.

Tabela 4.1. Tempos de computação [ms] dos cálculos adicionais envolvidos nos métodos GS e PS.

Extracção e comparação da característica Gist - método GS	Actualização faseada das distribuições de probabilidade - método PS	
	IDOL	FDF Park
1.43±0.18	0.67±0.18	0.90± 0.26

Figuras 2.9 e 2.10). Nesta comparação, a precisão é representada como função do tempo de computação (primeira linha de gráficos das figuras) e como função dos requisitos de memória (segunda linha de gráficos).

Da observação destas figuras verifica-se que, à excepção de um caso (Figura 4.19.a), o impacto de FM sobre a precisão é negligenciável. A aplicação dos métodos seguintes acarreta, em geral, alguma perda de precisão. Contudo, a redução de precisão na representação NQ é normalmente menos acentuada do que a redução na representação Q, para os mesmos níveis de tempo de computação. Daqui resulta que o hiato de desempenho entre as duas representações se mantém ou aumenta, com a aplicação dos métodos de redução do peso computacional. Nas Tabelas 4.2 e 4.3 apresenta-se a diferença de precisão e a memória ocupada para cada uma das representações e considerando os tempos de computação mais baixos atingidos na representação NQ. Relativamente aos valores da representação Q, estes são obtidos por interpolação dos dados medidos com as dimensões de vocabulários testadas e

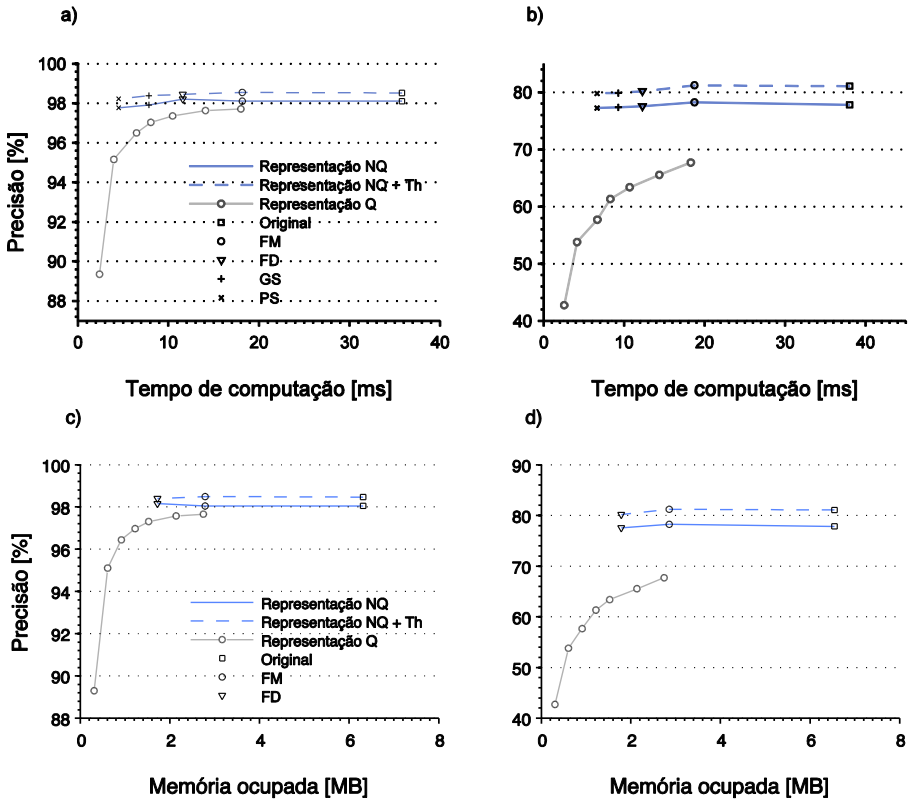


Figura 4.18. Precisão como função dos requisitos computacionais no *dataset* IDOL. Em cima: precisão vs tempo de computação; em baixo: precisão vs dimensão de memória; primeira coluna: condições de luminosidade semelhantes; segunda coluna: condições de luminosidade díspares.

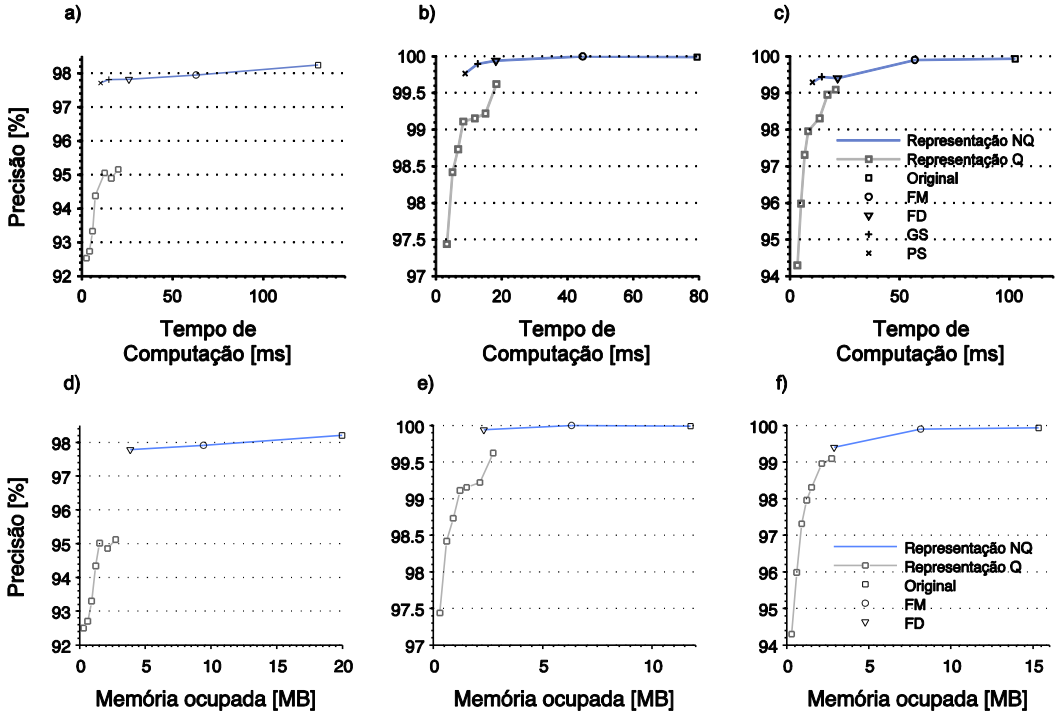


Figura 4.19. Precisão como função dos requisitos computacionais no *dataset* FDF Park. Em cima: precisão vs tempo de computação; em baixo: precisão vs dimensão de memória. As colunas 1 a 3 referem-se às sequências de modelação A, C e D respectivamente.

Tabela 4.2. Perda de precisão e memória ocupada para as mesmas condições de tempo de computação nas duas representações - *dataset* IDOL.

Condições de iluminação	Diferença de precisão [pontos percentuais]		Memória ocupada [MB]		Tempo de computação médio [ms]
	Q	NQ	Q	NQ	
Semelhantes	-1.13	-0.33	0.75	1.72	4.63
Díspares	-6.38	-0.56	1.05	1.79	6.62

Tabela 4.3. Perda de precisão e memória ocupada para as mesmas condições de tempo de computação nas duas representações - *dataset* FDF Park.

Sequência de modelação	Diferença de precisão [pontos percentuais]		Memória ocupada [MB]		Tempo de computação médio [ms]
	Q	NQ	Q	NQ	
A	-0.40	-0.53	1.39	3.87	11.08
B	-0.50	-0.23	1.26	2.31	8.93
C	-1.00	-0.64	1.34	2.89	10.37

usando os tempos de computação da representação NQ como referência. A diferença de precisão refere-se à diferença entre os valores interpolados e os valores obtidos com os vocabulários mais extensos, que oferecem a melhor precisão.

4.6 Sumário

Neste capítulo foram introduzidas duas categorias de métodos que têm o objectivo de reduzir os custos computacionais da representação NQ. Na primeira categoria, os métodos FM e FD produzem a compactação dos modelos do ambiente, em resultado da qual há uma redução da memória ocupada bem como do tempo de execução. Com os métodos GS e PS, na segunda categoria, pretende-se acelerar o algoritmo de localização, através da selecção dos lugares que devem ser considerados.

O estudo levado a cabo sobre os custos da representação NQ foi acompanhado da comparação com a representação Q, usando dois *datasets* como casos de estudo. No caso da representação Q, foi observado no capítulo 2 que a sua precisão está relacionada com a dimensão do vocabulário visual, n_c. Na secção 4.5 deste capítulo caracterizou-se a relação entre a precisão desta representação e o seus custos (ver Figuras 4.18 e 4.19), que aumentam também com n_c. Em suma, a representação Q oferece uma gama de configurações em que é possível, através do parâmetro n_c, selecionar um compromisso entre precisão e custos computacionais.

Neste capítulo demonstrou-se que, também na representação NQ, é possível obter uma gama de configurações, através da aplicação dos métodos propostos, que oferecem diferentes compromissos entre precisão e custos computacionais. Na comparação desta gama de configurações com as da representação Q verificou-se que, na maior parte dos casos, a última acarreta representações mais compactas. No que diz respeito ao tempo de execução, a representação NQ beneficia adicionalmente da aceleração devida aos métodos GS e PS, que aproximam o seu custo do da representação Q. No caso do *dataset* IDOL, os tempos de execução atingidos são próximos dos observados na representação Q com vocabulários de dimensão de 5000 a 7500, enquanto no *dataset* FDF Park os valores são próximos para vocabulários de 10000 a 12500 palavras visuais. Os tempos de computação médios medidos naquelas condições são inferiores a 12ms, o que torna os métodos propostos viáveis para execução em tempo real. Em todas as configurações testadas da representação NQ a sua superioridade em termos de precisão mantém-se. Além disso, os testes realizados sugerem que a degradação de desempenho nesta representação é mais suave, pelo que a diferença de desempenho entre as duas representações tende a acentuar-se nas configurações de menor custo computacional.

5. Detecção da revisitação de lugares com a característica LBP-Gist

5.1 Introdução

Este capítulo introduz, na presente tese, o tratamento do problema da revisitação de lugares. Este problema ocorre aquando da exploração de um novo ambiente, em que o estabelecimento de correspondências entre múltiplas passagens pelo mesmo lugar é vital para a construção de um mapa. O tipo de sistema desejado pertence à categoria dos detectores, tendo por objectivo assinalar o evento da revisitação de um lugar já conhecido. A capacidade de um robô detectar estes eventos é essencial, quer para a geração de mapas topológicos correctos, quer para a correcção de informação métrica sujeita a erros de odometria. Por outro lado, a indicação destes eventos quando eles não ocorrem tem efeitos muito adversos sobre os mapas resultantes, pelo que a existência de falsos positivos é altamente indesejável. Este requisito é traduzido, em termos do desempenho procurado, na necessidade de garantir precisão muito elevada, ao mesmo tempo que a taxa de *recall* deve ser tão alta quanto possível.

Este capítulo marca também uma diferente direcção no tipo de técnicas empregues, com a abordagem por características locais a ser substituída pelo uso de características globais. Estas, que tinham sido introduzidas no capítulo 4 no sentido de acelerar um sistema de localização, estarão, aqui, na base do detector. Esta mudança de abordagem é justificada, em primeiro lugar, pelo facto de os problemas de localização e de revisitação de lugares serem colocados em condições distintas. No primeiro caso, o intervalo de tempo que dista entre o momento da exploração do ambiente do momento de localização pode ser de várias horas ou dias. Nos capítulos anteriores, estas condições estiveram patentes nos *datasets* usados e reflectiam-se em mudanças de iluminação com grande impacto sobre a aparência das imagens. Neste caso, a abordagem por características locais é a mais adequada, dada a sua robustez perante estas transformações. De forma distinta, a detecção de revisitação é aplicada apenas durante a primeira exploração do ambiente, em que o intervalo entre a primeira passagem num lugar e as seguintes é muito menor – no *dataset* mais extenso usado neste capítulo, a duração da experiência é de 44min. Nestas condições não são esperadas variações de luminosidade significativas, tornando viável a aplicação de

características que, embora de menor robustez, oferecem poder descritivo suficiente para suportar a detecção de revisitação.

A segunda razão que motiva a substituição de características locais por globais diz respeito à complexidade e custo computacional de cada abordagem. As soluções mais comuns baseadas em características locais assentam no modelo BoW, resultando em sistemas com alguma complexidade, por envolverem as etapas de i) detecção de pontos de interesse, ii) extracção de características, iii) quantização e iv) pesquisa de imagens por ficheiros invertidos. Muitas vezes, as imagens candidatas passam ainda por uma etapa de verificação de coerência geométrica. A abordagem por características globais, por seu lado, origina sistemas mais simples e de menor custo computacional, essencialmente por dispensar as etapas de detecção de pontos de interesse e de verificação geométrica. Em vez destas, a cada imagem é aplicada uma partição fixa, de onde são extraídas estatísticas de cada bloco que depois são comparadas com as dos blocos homólogos de outras imagens.

Com o objectivo de realizar a detecção de revisitação com características globais, neste capítulo é desenvolvida uma característica original, designada LBP-Gist. Como o nome sugere, esta é inspirada no conceito de representação holística de imagens, que está na base do Gist, e no método LBP, que introduz a análise de texturas com grande eficiência e bom poder descritivo. Neste capítulo é descrito o desenvolvimento desta característica e é feita a sua validação, em quatro *datasets* para os quais existem resultados publicados baseados no modelo BoW.

Neste capítulo apresenta-se ainda um método de pesquisa rápida de imagens adequado à pesquisa por descritores LBP-Gist. Através deste método, baseado nos princípios de *Locality Sensitive Hashing* (LSH), é possível seleccionar as imagens mais relevantes que posteriormente são analisadas por uma medida de semelhança mais precisa. Este módulo de pesquisa revelar-se-á essencial na obtenção de um detector eficiente, dado que a elevada dimensionalidade do descritor implicaria custos computacionais elevados, se todas as imagens da base de dados tivessem se ser consideradas. De entre os métodos de LSH existentes usar-se-á o algoritmo *Winner Take All hashing* (WTA) que, para além de fazer a indexação de imagens, oferece uma estimativa de semelhança entre imagens. Esta propriedade será explorada em duas modificações propostas, em que i) essa estimativa de semelhança é melhorada

modulando as contribuições pelo termo *inverse-document-frequency* e ii) a selecção das imagens é determinada por uma condição definida sobre a razão de semelhanças.

O resto deste capítulo desenvolve-se pela seguinte ordem: em 5.2 perspectiva-se o presente trabalho face à literatura relacionada; na secção 5.3 introduzem-se os conceitos essenciais na análise de texturas pelo método LBP; em 5.4 apresenta-se o desenvolvimento da característica LBP-Gist, focando as opções de configuração tomadas com base em resultados sobre dois *datasets*; em 5.5 descreve-se o algoritmo de pesquisa WTA e as modificações introduzidas à versão original; a secção 5.6 caracteriza o detector em termos de custo computacional e de desempenho na detecção de revisitação e compara-o com o modelo BoW; por fim, na secção 5.7, são analisados os resultados mais importantes do capítulo.

5.2 Trabalhos relacionados

A par de inúmeros estudos sobre o modelo BoW (Angeli et al., 2008; Cummins e Newman, 2008; Galvez-López e Tardós, 2012) encontram-se na literatura alguns exemplos da aplicação de características holísticas no reconhecimento de lugares. Os sistemas mais recentes que seguem esta abordagem são inspirados na característica Gist, introduzida por Oliva e Torralba (2001) para a modelação do conteúdo semântico de cenas. Tal como definida nesse trabalho, a extracção do Gist envolve a aplicação de filtros direccionais, com diversas escalas e orientações, e a divisão da imagem por uma grelha de 4 por 4. De seguida é calculado o valor médio da resposta de cada filtro, em cada sub-bloco, e esses valores são reunidos no descritor Gist.

A ideia, patente na definição anterior, de se aplicar uma geometria fixa e calcular estatísticas sobre essa partição tem sido adoptada e desenvolvida em diversos trabalhos sobre reconhecimento de lugares. No trabalho de Siagian e Itti (2009), a localização de um robô é efectuada por um sistema de visão inspirado na biologia, o qual é integrado no algoritmo de localização de Monte Carlo. O bom desempenho obtido por esta solução deve-se à combinação, biologicamente plausível, da característica Gist com características locais baseadas na saliência visual.

A característica Gist foi também adaptada a imagens omni-direccionais, por Murillo e Kosecka (2009), que introduziram uma representação específica para este tipo de câmaras, designada por panorama Gist. A utilização da característica Gist naquele estudo é inovadora também pela quantização do descritor, por forma a reduzir o custo

computacional nas comparações, e pela aplicação de regras de alinhamento dos panoramas que oferecem alguma invariância ao descritor.

Essencialmente, o método proposto neste capítulo contrasta com os trabalhos mencionados anteriormente pela utilização de características LBP, que, comparativamente, são de extracção muito rápida. Este cuidado colocado na escolha de características mais simples verifica-se também em estudos mais recentes, que demonstram ser possível obter resultados satisfatórios com ganhos significativos nos custo computacional. Neste contexto, uma abordagem semelhante à que propomos é a introduzida por Jianxin e Rehg (2011), que recorrem à transformada *census* na extracção de texturas da imagem. Embora aquela transformada seja semelhante ao operador LBP, não oferece a mesma flexibilidade, devido ao suporte espacial de 3x3 píxeis. Além disso, naquele trabalho apenas o problema de localização global foi tratado, tendo sido usado um classificador treinado que não é aplicável ao problema da revisitação de lugares. Outro exemplo de utilização de características de extracção rápida é dado por Sunderhauf e Protzel (2011), que desenvolveram a característica BRIEF-Gist. A abordagem proposta por estes autores concilia a ideia de se registar a estrutura espacial da imagem, tal como na característica Gist, com a característica BRIEF, desenvolvida por Calonder et al. (2012) como descritor local. A rapidez deste sistema é garantida pelas boas propriedades da característica BRIEF, a qual é baseada na simples comparação de píxeis e é representada por um descritor binário, permitindo a comparação rápida de descritores pela distância Hamming. Apesar da eficiência deste método, os seus autores reconhecem que ele pode ser sensível a desvios de rotação e translação, os quais contribuem para diminuir a sua precisão.

Nos últimos anos as técnicas de LSH têm suscitado grande interesse por parte da comunidade de Visão Computacional, motivado pela crescente dimensão das bases de imagens públicas e da capacidade destes algoritmos pesquisarem de forma rápida e precisa. Este interesse resultou no desenvolvimento de diversos algoritmos que têm sido aplicados com sucesso na comparação de características globais (Weiss, Fergus e Torralba, 2012) e de características locais (Paulevé, Jégou e Amsaleg, 2010). Entre os algoritmos existentes, a técnica WTA (Yagnik et al., 2011), adoptada neste trabalho, destaca-se pela sua simplicidade e por apresentar o melhor desempenho na comparação de descritores Gist (Yagnik et al., 2011). Além disso, esta técnica enquadra-se na categoria de métodos que devolvem uma estimativa de semelhança,

permitindo seleccionar as imagens a devolver pela comparação com um valor de semelhança mínima. Esta propriedade é explorada neste trabalho através de duas modificações ao algoritmo WTA original que se revelam úteis na selecção dos lugares mais relevantes. A primeira é inspirada na modulação pelo factor *tf-idf*, desenvolvida inicialmente para a pesquisa de textos (Jones, 1972) e mais tarde aplicada à pesquisa de imagens com o modelo BoW (Sivic e Zisserman, 2003). Com a segunda modificação pretende-se reduzir o número de imagens devolvidas nas situações em que existem inúmeros lugares com aparências semelhantes e em que a aplicação de um limite fixo à semelhança resultaria numa lista extensa. A modificação proposta sugere a aplicação de um limite à razão de semelhanças, o que resulta efectivamente na aplicação de um limite adaptativo, mais adequado àqueles casos.

5.3 Análise de texturas através do método LBP

A primeira etapa do método LBP consiste na geração de uma imagem de códigos, em que cada píxel da imagem original é convertido num código que sumariza o padrão de textura em torno desse píxel (ver Figura 5.1). A eficiência desta operação resulta do cálculo dos códigos a partir de simples comparações binárias entre o valor de intensidade dos píxeis na imagem original. Na formulação inicial do método LBP (Ojala, Pietikäinen e Harwood, 1996), ilustrada na Figura 5.2, são consideradas apenas vizinhanças de dimensão 3×3 e a codificação do píxel central resulta da comparação do seu nível de cinzento, g_c, com os dos oito píxeis na sua vizinhança g_0, g_1,..., g_7, por uma função de *threshold* definida como

$$Th(g_c, g_i) = \begin{cases} 1, g_i \geq g_c \\ 0, g_i < g_c \end{cases}. \qquad (5.1)$$

Os oito bits assim obtidos formam um byte que identifica o padrão de textura actual como um de 2^8 padrões possíveis.

Num desenvolvimento posterior, a limitação no suporte do operador foi ultrapassada com a introdução do operador multi-escala (Ojala, Pietikainen e Maenpaa, 2002), o qual generaliza o operador original e admite quaisquer valores na dimensão da vizinhança e número de pontos na amostragem dessa região (ver Figura 5.2). Nesta versão, a vizinhança é definida como um conjunto de *P* pontos distribuídos uniformemente sobre uma circunferência centrada no píxel a codificar. A análise

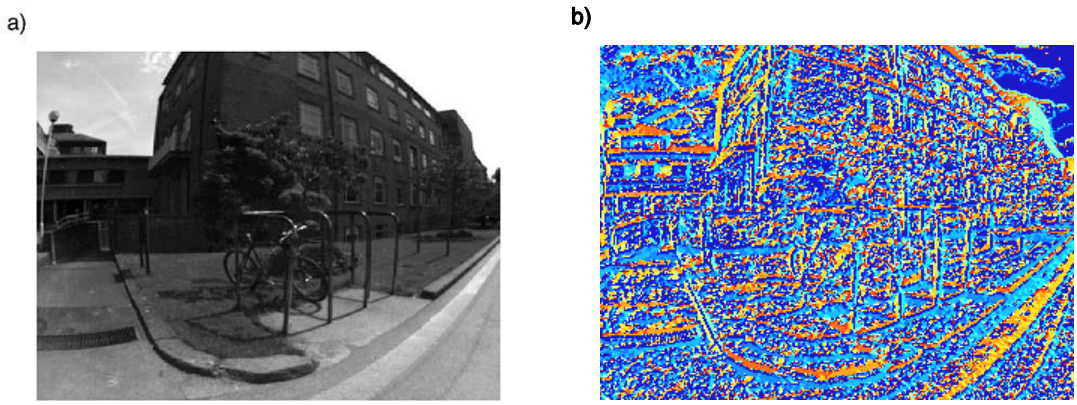


Figura 5.1. Primeira etapa do método LBP: conversão da imagem em níveis de cinzento para uma imagem de códigos de textura. A) imagem original, b) imagem de códigos.

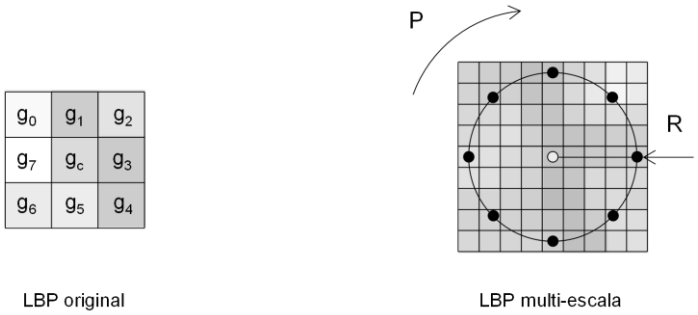


Figura 5.2. Pontos de amostragem da vizinhança no operador LBP original e multi-escala, com *P*=8 e *R*=4.

multi-escala é possível neste método pela variação do raio *R* desta circunferência. Além disso, a escolha de *P* determina a resolução dos padrões que podem ser representados. Por forma a manter o número de códigos dentro de limites aceitáveis, e tendo em conta que este número aumenta exponencialmente com *P*, este valor é normalmente mantido baixo, mesmo quando *R* aumenta. Nesta situação, é sugerido pelos autores Mäenpää e Pietikäinen (2003), que o nível de cinzento de um ponto de amostragem represente uma área mais vasta do que o próprio pixel. Para esse efeito, propuseram a suavização da imagem por um filtro gaussiano, previamente à aplicação do operador LBP.

Em várias aplicações, os códigos LBP obtidos na comparação de píxeis têm sido mapeados para conjuntos mais restritos, como medida de redução dos padrões possíveis. Este mapeamento é fundamentado no conceito de Uniformidade, desenvolvido por Topi et al. (2000) para analisar as estatísticas dos diversos códigos.

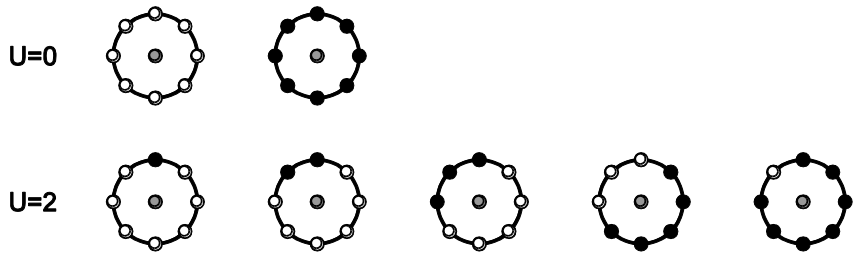


Figura 5.3. Em cima, os dois únicos padrões com *U*=0, em baixo, exemplos de padrões com *U*=2.

Tal como definida naquele trabalho, a medida de uniformidade, *U*, é o número de transições de 0 para 1, ou vice-versa, que se encontram na cadeia de bits, quando esta é vista de forma circular. Segundo esta definição, e de forma algo contra-intuitiva, os códigos com menor valor de *U* são mais uniformes, o que levou à designação de *uniformes* todos os padrões com $U \leq 2$ (na Figura 5.3 apresentam-se exemplos deste caso). Notando que os padrões uniformes representam uma grande percentagem dos códigos extraídos de imagens naturais, em (Topi et al., 2000) define-se o chamado mapeamento uniforme, o qual retém todos os padrões uniformes e atribui um código único aos padrões com $U > 2$. Na prática, este mapeamento trará uma simplificação dos descritores LBP, à custa de alguma perda de informação contida nos padrões que são classificados como não-uniformes. O operador LBP multi-escala que o usa mapeamento uniforme é genericamente designado por $LBP_{R,P}^{u2}$.

O último passo no método LBP refere-se à construção do descritor da imagem. Este é definido como o histograma dos padrões LBP presentes na imagem de códigos, podendo ser calculado sobre a imagem completa ou sobre sub-blocos, caso em que o descritor é obtido pela concatenação dos histogramas referentes a cada bloco.

5.4 Característica LBP-Gist

Essencialmente, a característica LBP-Gist combina o operador LBP com o conceito de representação holística que está na origem do Gist, o que resulta num descritor de computação rápida e que oferece bom poder descritivo de cenas. Apesar de a integração entre os dois conceitos ser simples, dado que o método LBP facilmente admite o cálculo de histogramas sobre partições da imagem, verifica-se que, para obter precisão ao nível do estado da arte, é necessário ajustar o operador ao problema da revisitação de lugares. Esta adaptação envolve modificações ao operador LBP e decisões no desenho da característica LBP-Gist que são resumidas de seguida:

Função de *threshold* - A função de *threshold* proposta originalmente é extremamente sensível ao ruído da imagem em regiões uniformes. Nestas regiões, o nível de cinzento deve ser aproximadamente igual, mas, devido à existência de ruído, estão normalmente presentes variações não negligenciáveis. Uma vez que a função de *threshold* realiza uma simples comparação de níveis de cinzento, estas variações afectam o resultado da operação, produzindo códigos diferentes daquele que é esperado numa região perfeitamente uniforme. Na secção 5.4.1 define-se uma extensão desta função que supera esta limitação.

Parâmetros *R* e *P* - No problema de classificação de materiais pela sua textura, para o qual o operador LBP foi inicialmente desenvolvido, foram testadas combinações de R e P com valores em $R \in \{1, 2, 3\}$ e $P \in \{8, 16, 24\}$ (Ojala, Pietikainen e Maenpaa, 2002). Embora estes valores se mostrem adequados àquele problema, os padrões mais discriminativos na descrição de lugares poderão ocorrer em escalas diferentes daquelas, pelo que estes parâmetros devem ser analisados à luz do presente problema. Na escolha destes parâmetros foram colocadas algumas restrições, com o intuito de manter a dimensão do descritor reduzida e diminuir os custos computacionais associados. A primeira diz respeito ao valor de P, que foi limitado a 8. Embora o uso de mais pontos de amostragem pudesse trazer alguma discriminatividade adicional, fazemos notar que o incremento de P se reflecte exponencialmente no número de códigos, o que rapidamente levaria o descritor a tomar dimensões não tratáveis em tempo real. Para P=8, a dimensão de um histograma é de 255, sem a aplicação de mapeamentos simplificativos. A segunda restrição refere-se à análise multi-escala das imagens. Nalguns trabalhos, o operador LBP é aplicado em múltiplas escalas, com vista a obter-se maior poder descritivo, contudo, para cada escala adicional é adicionado 2^P à dimensão do descritor. Também para manter a dimensão do descritor baixa, optou-se por extrair códigos LBP a uma única escala. Dadas estas restrições, o único parâmetro a configurar é R, o qual deverá tomar o valor de uma escala adequada, e suficientemente genérica, para a descrição de lugares. Este parâmetro será analisado na secção 5.4.2.

Função de mapeamento - Em várias aplicações do método LBP a redução dos descritores tem sido obtida através do mapeamento dos códigos originais para um subconjunto mais pequeno, recorrendo-se para isso ao mapeamento uniforme e, por vezes também, invariante à rotação (Ojala, Pietikainen e Maenpaa, 2002). A rotação

no plano da imagem, embora presente nas imagens de lugares devido a variações de perspectiva, não tem o efeito preponderante encontrado na classificação de materiais. Além disso, a orientação em que os padrões são encontrados fornece informação discriminativa, pelo que a invariância à rotação não foi considerada nos mapeamentos testados nesta tese. A uniformidade do mapeamento é também uma propriedade que deve ser reconsiderada, quando se trata da descrição de lugares. De facto, a justificação para este tipo de mapeamento assenta na raridade dos códigos de uniformidade superior a 2 e, consequentemente, na dificuldade em estimar a sua distribuição (Ojala, Pietikainen e Maenpaa, 2002). Estes argumentos, especialmente válidos quando a escala do operador é reduzida, não têm o mesmo peso quando R aumenta. Vários estudos (Ojala, Pietikainen e Maenpaa, 2002; Shu et al., 2006) mostram que, com o aumento de R, os padrões não uniformes ganham representatividade e, por isso, a sua estimação torna-se viável. Este dado, associado ao facto de que estes códigos podem conter poder descritivo importante, justificou o estudo de um mapeamento alternativo que será apresentado na secção 5.4.2.

Seccionamento da imagem - A definição da característica Gist dada por Oliva e Torralba (2001) prescreve a divisão da imagem por uma grelha de dimensão 4×4, uma prática que foi também seguida em aplicações posteriores (Douze et al., 2009; Siagian e Itti, 2009). Embora esta divisão da imagem se tenha revelado adequada para a classificação semântica de imagens, a detecção de lugares distingue-se daqueles estudos por colocar condicionalismos diferentes. Enquanto no primeiro problema se pretende que o descritor ofereça boa generalização dentro de categorias de imagens, no segundo procura-se um descritor que apresente boa especificidade na caracterização de lugares, mesmo na presença de alguma variação de perspectiva. Neste caso, o excessivo particionamento da imagem pode diminuir a tolerância do descritor às variações de perspectiva e, por isso, reduzir o desempenho do detector. Numa discussão sobre este tema, a desenvolver na secção 5.4.3, deve ser considerada também a influência sobre a dimensão do descritor, o qual cresce proporcionalmente ao número de divisões da imagem.

Função de comparação - A decisão de assinalar um lugar como já tendo sido visitado é tomada a partir de uma medida de comparação entre descritores LBP-Gist. Normalmente, a comparação entre descritores LBP é feita através da distância chi-quadrado, aplicada aos vectores que resultam da concatenação dos histogramas LBP

(Ahonen, Hadid e Pietikäinen, 2004). Designando por $D_i = \{h_i^1, \dots, h_i^{nB}\}$ o descritor da imagem i, que reúne os histogramas de nB blocos da imagem, a medida de distância tradicional é dada por:

$$d_{LBP}(D_i, D_j) = CHI2\left([h_i^1, \dots, h_i^{nB}], \left[h_j^1, \dots, h_j^{nB}\right]\right). \quad (5.2)$$

Nos pontos seguintes, e antes de ser introduzida uma forma de comparação mais complexa, esta será a expressão usada para avaliar a característica LBP-Gist. No ponto 5.4.3, será introduzida uma medida de semelhança que oferece melhores propriedades na detecção de lugares.

O desenvolvimento da característica LBP-Gist, apresentado nas secções seguintes, será acompanhado de resultados obtidos sobre dois *datasets*, designados por *Malaga* e *City Centre*. Estes *datasets* foram escolhidos por serem representativos de duas condições distintas de visualização do ambiente, no primeiro caso através de uma câmara frontal e, no segundo, com duas câmaras laterais. Mais à frente, na secção 5.6 estes e outros *datasets* serão descritos em detalhe, e esse conjunto mais alargado de dados suportará a validação do sistema.

Nas secções seguintes, a avaliação de diferentes opções na configuração da característica LBP-Gist é feita recorrendo a uma medida de desempenho ajustada à revisitação de lugares. Em problemas de detecção ou de pesquisa de imagens, é comum a avaliação ser baseada na curva de precisão vs *recall*, nomeadamente através do parâmetro de precisão média, que mede a área sob aquela curva. No cálculo deste parâmetro são contemplados todos os pontos da curva, mesmo aqueles com precisão baixa e que representam parametrizações menos interessantes do sistema. No caso da revisitação de lugares estes pontos são particularmente indesejáveis pois a existência de falsos positivos é muito prejudicial na construção de mapas topológicos. Por esta razão, a avaliação deve ser feita sobre as zonas da curva em que realisticamente o detector pode ser útil, isto é, nos pontos próximos da precisão=1. Face a este requisito definiu-se uma medida que calcula a área abaixo da curva precisão vs *recall* e acima de um valor elevado de precisão, neste trabalho de 0.96. Esta medida, ilustrada na Figura 5.4, pode ser entendida como o *recall* acumulado no intervalo de precisão [0.96, 1] e será designada por Rac_{96}. Segundo esta medida, o desempenho de um detector é tanto melhor quanto mais próximo Rac_{96} se aproxima da área que seria coberta por um detector ideal (representada na figura) e que tem o valor de 0.04.

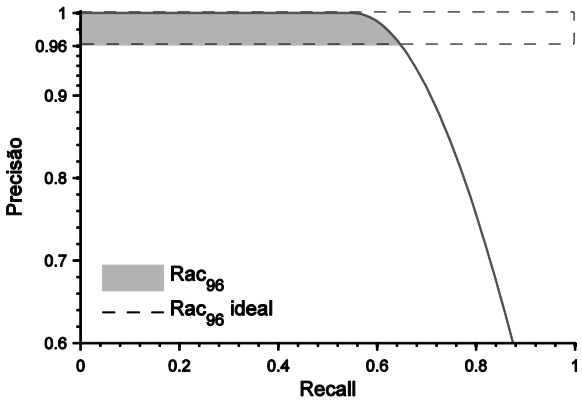


Figura 5.4. Exemplo de medição do parâmetro Rac_{96}. A cinzento, área correspondente a Rac_{96} da curva de exemplo; a tracejado, Rac_{96} de um detector ideal.

5.4.1 Função de *threshold*

Em estudos anteriores, tem sido reconhecido que o operador LBP é pouco robusto na descrição de regiões uniformes, tipicamente encontradas em zonas da imagem ocupadas por céu ou paredes lisas, e onde pequenas variações de intensidade resultam em códigos LBP diferentes. Uma vez que estes padrões estão frequentemente presentes em imagens de lugares, devem ser adoptadas estratégias para minimizar aquele efeito e produzir um descritor mais robusto.

No artigo dos autores Heikkila e Pietikainen (2006) a técnica usada consiste em estabelecer um nível fixo e diferente de zero a partir do qual a função de *threshold* muda de valor, levando assim a que pequenas variações no nível de cinzento originem consistentemente uma comparação de valor binário 0. Posteriormente, Shengcai et al. (2010) fizeram notar que o limite constante não é invariante a transformações lineares dos níveis de cinzento e propuseram um limite variável, proporcional ao valor do píxel central. Contudo, esta estratégia origina limites pequenos em regiões escuras, pelo que o operador pode ainda ser sensível a ruído nestas regiões.

Por forma a dotar o operador de maior robustez, propusemos a utilização de um limite que combina um termo variável com um termo fixo. A função de *threshold* que resulta desta opção é definida por

$$Th(g_c, g_i) = \begin{cases} 1, g_i \geq (1+\tau_1)g_c + \tau_0 \\ 0, g_i < (1+\tau_1)g_c + \tau_0 \end{cases}, \qquad (5.3)$$

onde τ_0 designa o limite constante e $1+\tau_1$ é o factor de proporcionalidade que determina o limite variável. O resultado da aplicação desta função é ilustrado na

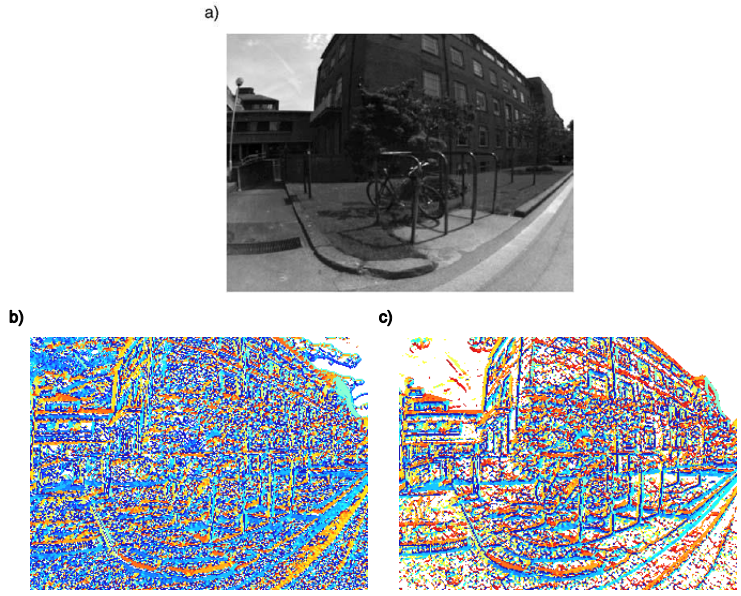


Figura 5.5. Atenuação de ruído com a função de *threshold* modificada. A) imagem em níveis de cinzento, b) códigos obtidos com a função de *threshold* original e c) com a função de *threshold* modificada.

Figura 5.5.c. Nas nossas experiências não foram encontrados valores destes parâmetros que fossem óptimos em todos os *datasets*, no entanto, para τ_0=5 e τ_1=0.03, obteve-se resultados satisfatórios em todos eles.

5.4.2 Raio do operador e mapeamento de códigos

Neste ponto analisa-se a influência do parâmetro R sobre o poder descritivo do operador LBP, bem como a importância do mapeamento de códigos usado. Dado que o mapeamento uniforme acarreta alguma perda de informação, foi considerado o mapeamento que representa, para além dos códigos com U=2, os códigos com uniformidade imediatamente acima, isto é, U=4. Estes códigos, que correspondem a cadeias de bits onde o número de transições 0→1 e 1→0 totalizam 4, são ilustrados através de alguns exemplos na Figura 5.6. O operador que aplica este tipo de mapeamento será designado por $LBP_{R,P}^{u4}$. A Tabela 5.1 apresenta o número de códigos, com P=8, que são abrangidos por cada um dos valores de uniformidade. Lembrando que, em qualquer dos mapeamentos, os padrões não representados são mapeados para um único código, conclui-se que o uso de $LBP_{R,P}^{u4}$ acarreta, relativamente a $LBP_{R,P}^{u2}$, um aumento da dimensão dos histogramas de 59 para 199.

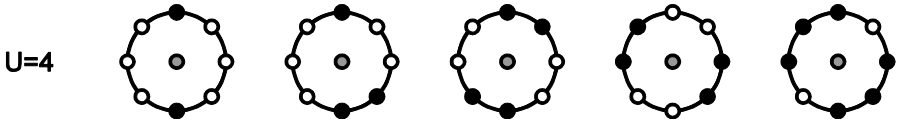


Figura 5.6. Exemplos de padrões com *U*=4.

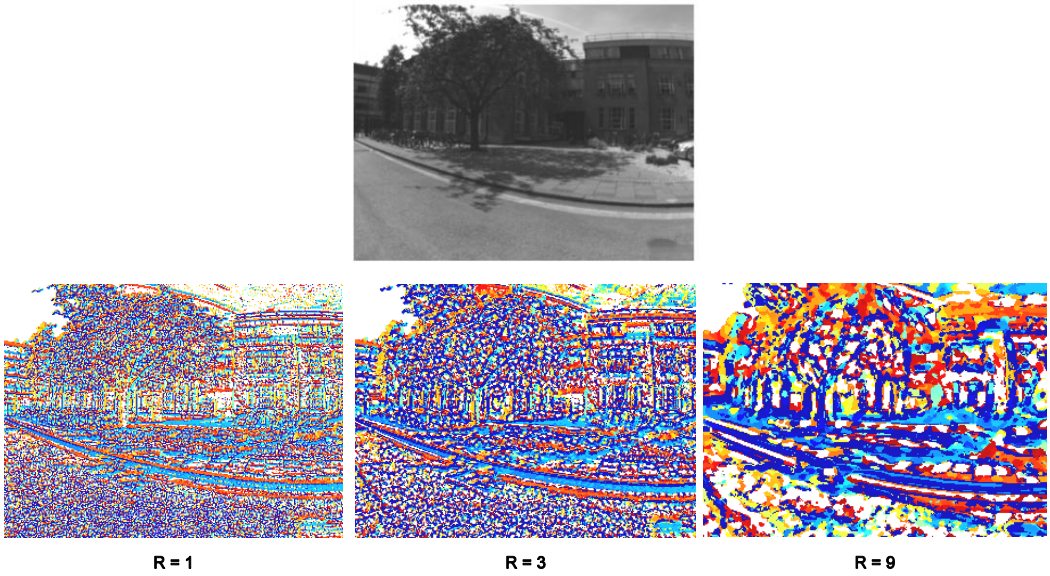


Figura 5.7. Imagens de códigos LBP obtidas com o operador multi-escala e diferentes raios da vizinhança. Em cima: imagem original, em baixo: imagens de códigos com *R*=1,3 e 9.

Tabela 5.1. Número de códigos existentes em cada categoria de Uniformidade, com *P*=8.

Uniformidade	0	2	4	>4
Nº de códigos	2	56	140	58

Para além do mapeamento a usar, o raio do operador é um parâmetro determinante na extracção da informação mais relevante, pois define a escala de análise da imagem. Na Figura 5.7 ilustra-se o impacto deste parâmetro sobre a distribuição de códigos no plano da imagem e, na Figura 5.8, apresenta-se a evolução de Rac_{96} em função do raio do operador, incluindo as curvas obtidas com os mapeamentos de $LBP_{R,8}^{u2}$ e $LBP_{R,8}^{u4}$. Nestes e posteriores resultados foram processadas imagens com dimensão 240×320. De acordo com Ojala, Pietikainen e Maenpaa (2002) e Shu et al. (2006), a ocorrência de padrões com *U*>2 aumenta com o raio do operador, o que sugere que o benefício de incluir estes códigos acompanha também a evolução de *R*. No caso do *dataset*

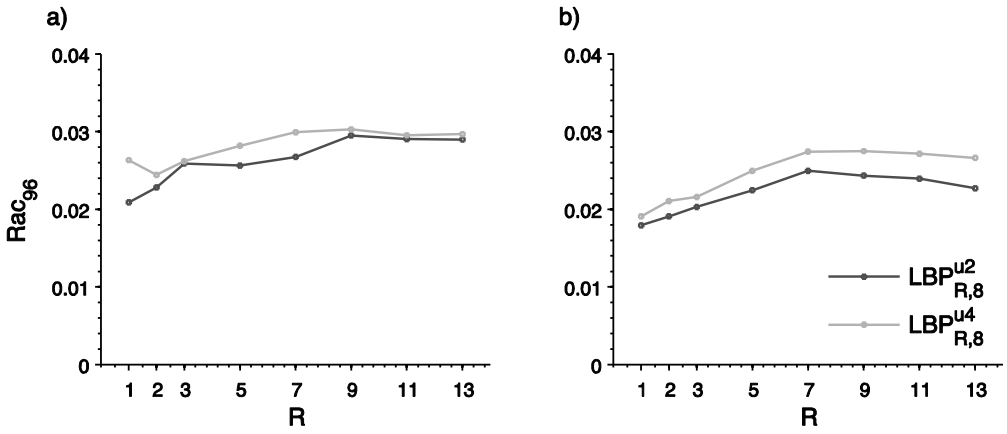


Figura 5.8. Evolução de R_{ac96} com o raio do operador LBP no a) *dataset* Malaga e b) *dataset* City Centre.

City Centre verifica-se esta tendência, com o desempenho de $LBP_{R,8}^{u4}$ distanciando-se de $LBP_{R,8}^{u2}$ à medida que o raio aumenta. No entanto, esta tendência não está patente no *dataset* Malaga, o qual apresenta evoluções menos consistentes em qualquer dos mapeamentos. Apesar desta diferença de comportamentos, o mapeamento $LBP_{R,8}^{u4}$ é sempre superior e será por isso adoptado na definição da característica LBP-Gist. Os resultados da Figura 5.8 são também coerentes na posição do pico de desempenho daquele mapeamento, que ocorre para R=9. Por esta razão será usado o operador $LBP_{9,8}^{u4}$ na construção do descritor LBP-Gist.

5.4.3 Seccionamento da imagem e operação de comparação

Tradicionalmente, o descritor Gist reúne informação extraída de sub-blocos que resultam de um seccionamento em grelha da imagem. Neste seccionamento, a geometria da grelha é definida com um número igual de divisões horizontais e verticais. Embora esta metodologia tenha sido bem sucedida na análise semântica de cenas, o objectivo da característica LBP-Gist é o da identificação de lugares, o qual coloca diferentes constrangimentos e pode, talvez, ser atingido mais eficazmente com outro tipo de geometria. A pertinência desta questão é particularmente evidente quando se nota que o movimento do robô está constrangido ao plano horizontal e, que por isso, a maior variabilidade nas imagens consiste em translações horizontais. Uma vez que os particionamentos da imagem em linhas ou colunas são afectados diferentemente por aquele efeito, foi estudado o impacto destes parâmetros no descritor. Para além disso, foram consideradas diferentes estratégias na divisão da imagem, definidas de seguida e ilustradas na Figura 5.9.

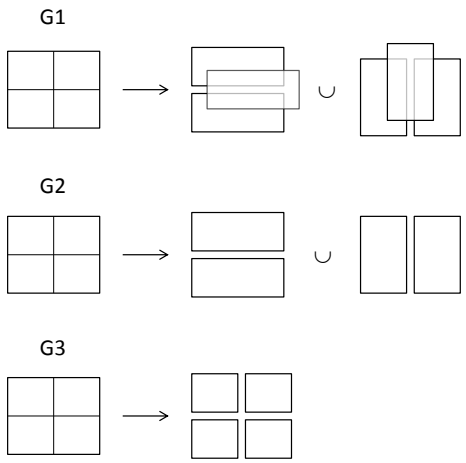


Figura 5.9. Com *nh*=2 e *nv*=2, a figura ilustra as três estratégias de seccionamento da imagem estudas em 5.4.3.

Designando por *nh* e *nv* respectivamente o número de linhas e colunas da grelha, os métodos de particionamento G1, G2 e G3 são descritos por:

G1: Nesta forma de seccionamento, as linhas horizontais e verticais da grelha não são intersectadas, logo, os sub-blocos são *nh* faixas que se estendem em todo o comprimento da imagem e *nv* faixas que se estendem por toda a altura. Adicionalmente, são extraídos blocos verticais e horizontais intermédios que cobrem metade de cada um dos adjacentes. Estes blocos visam adicionar alguma robustez relativamente aos movimentos que deslocam os padrões visuais entre blocos adjacentes.

Neste método, o número total de blocos é dado por $nB = nh \times 2 - 1 + 2 \times nv - 1$. Nos casos em que $nh=1$ (ou $nv=1$) o bloco horizontal (vertical) corresponde à imagem completa. Nestes casos, ignorou-se esse bloco já que o histograma calculado sobre a imagem completa é pouco discriminativo, e então $nB = 2 \times nh - 1$ ($nB = 2 \times nv - 1$).

G2: Esta forma de seccionamento é idêntica a G1, diferindo apenas na não utilização dos blocos intermédios. O número total de blocos é de $nB = nh + nv$. Também aqui se ignoram os blocos que correspondam à imagem completa, quando *nh* ou *nv* são iguais a 1.

G3: Esta é a forma de seccionamento que tem sido usada na característica Gist. Aqui, os blocos extraídos resultam da intersecção das divisões horizontais e verticais, resultando em $nB = nh \times nv$.

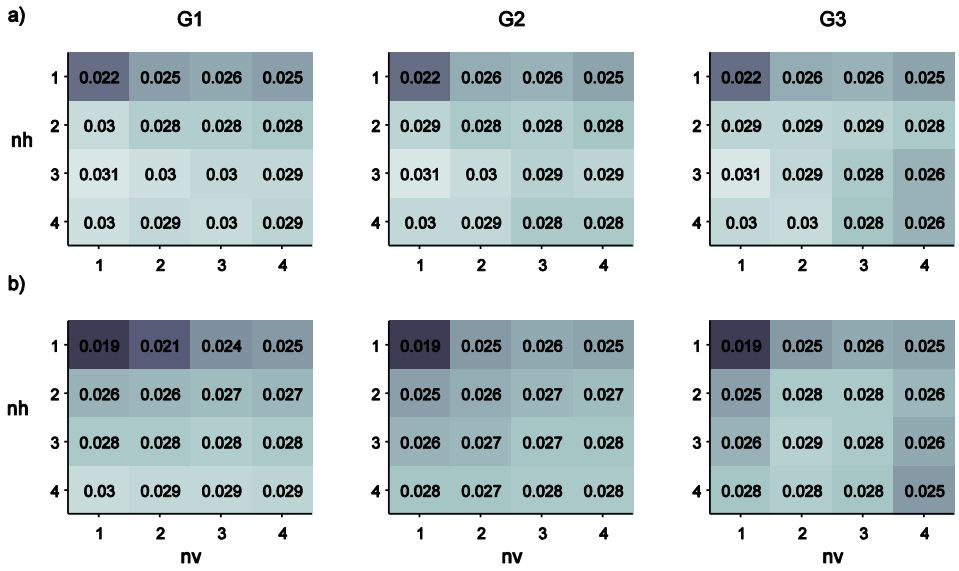


Figura 5.10. Desempenho de cada uma das geometrias (valores de Rac_{96}) em função do número de divisões horizontais e verticais. A) dataset Malaga, b) dataset City Centre. O sombreado identifica, a claro, os valores mais elevados de Rac_{96}.

A Figura 5.10 ilustra os valores de Rac_{96} medidos com cada uma das geometrias e valores de *nh* e *nv* entre 1 e 4. Um dado relevante destes resultados é o de que, com o seccionamento mais comum, G3, o desempenho se degrada quando *nh* e *nv* são simultaneamente elevados. Este facto mostra que a divisão da imagem em blocos excessivamente pequenos é contraproducente, por retirar tolerância ao descritor relativamente às translações no plano da imagem. Este resultado mostra também que a prática comum de dividir a imagem numa grelha de 4×4 não é a mais adequada ao presente problema.

Estas figuras permitem identificar também diferenças no comportamento do descritor relativamente à sucessiva divisão da imagem horizontalmente (quando *nh* aumenta) ou verticalmente (quando *nv* aumenta). Estas diferenças são mais evidentes no *dataset* Malaga, onde os valores abaixo da diagonal são tipicamente mais altos do que os correspondentes acima da diagonal. Por exemplo, o caso *nh*=2, *nv*=1 é claramente superior, em todas as geometrias, ao caso *nh*=1, *nv*=2. No *dataset* City Centre estas diferenças são menos notórias, no entanto, quando existem, seguem a mesma tendência.

Previamente à selecção de uma geometria definitiva a ser usada no descritor LBP-Gist, importa neste ponto introduzir a comparação entre descritores que será usada daqui em diante. Com a técnica que será desenvolvida, pretende-se superar a

abordagem anterior, em que os histogramas dos diversos blocos são agrupados e a distância entre os vectores assim obtidos é calculada pela Eq. (5.2). A limitação encontrada nesta abordagem é a do tratamento de todos os blocos da imagem de igual forma, o que não permite atribuir a uns blocos maior relevância do que a outros. Esta questão foi abordada, por exemplo, por Ahonen, Hadid e Pietikäinen (2004), onde foi desenvolvido um sistema de reconhecimento facial em que se agregam as distâncias calculadas sobre os blocos individuais através de uma soma ponderada. Naquele trabalho os pesos atribuídos aos blocos eram estáticos e calculados *offline*, o que contrasta com os requisitos da revisitação de lugares, em que o detector deve ajustar-se a um ambiente desconhecido. Neste caso, não é possível saber previamente qual a relevância dos diferentes blocos da imagem e, além disso, essa relevância pode mudar com a região do ambiente em que o robô se move. Por estas razões, foi proposta uma medida de semelhança dependente do contexto, a qual atribui relevância a um bloco com base na ocorrência do mesmo padrão em imagens semelhantes. O procedimento de cálculo passa por, em primeiro lugar, dispor as imagens da base de dados em ordem ascendente das suas distâncias à imagem de teste, medidas pela Eq. (5.2). Numa base de dados contendo *np* imagens, este procedimento gera o conjunto ordenado $\{D_1, ..., D_{np}\}$. No passo seguinte calcula-se uma medida de semelhança para cada um dos blocos da imagem. Designando por $d_{LBP}(h^i, h_j^i)$ a distância chi-quadrado entre os blocos homólogos *i* da imagem de teste e a imagem *j*, esta distância é convertida para um valor de semelhança, s_j^i através da função exponencial:

$$s_j^i = exp\left(-\frac{d_{LBP}(h^i, h_j^i)}{\alpha}\right), \tag{5.4}$$

onde α é um factor de escala, ajustado para α=200. A relevância atribuída a um bloco *i*, representada pelo peso w^i, decorre dos valores de semelhança que este bloco recebeu num subconjunto de imagens:

$$w^i = \left(\sum_{j=k}^{k+9} s_j^i\right)^{-1}. \tag{5.5}$$

Segundo esta expressão, a relevância atribuída ao bloco *i* é maior se o padrão visual que contém for raro (semelhanças s^i_j baixas) e menor quando esse padrão é mais frequente (s^i_j elevados). Os valores s^i_j usados são os de um subconjunto de 10 imagens

escolhidas entre as imagens mais pertinentes para o teste actual, isto é, entre as mais bem posicionadas no ordenamento feito atrás. Na prática, não se usaram as primeiras imagens, já que na situação de estarmos perante uma verdadeira revisitação de um lugar, s^i_j será particularmente elevado nas imagens correspondentes, o que iria reduzir w^i de um modo indesejado. Assim, optou-se por excluir daquela selecção as imagens que, de acordo com a distância $d_{LBP}(D, D_j)$ são possíveis correspondências com a imagem actual. O critério usado foi o de definir *k* como a primeira imagem do ordenamento para a qual a distância $d_{LBP}(D, D_j)$ é superior a 4500. Este valor foi encontrado empiricamente, correspondendo à distância máxima entre imagens do mesmo lugar captadas em momentos diferentes da exploração do ambiente.

A medida de semelhança que integra a informação de todos os blocos é finalmente calculada por:

$$s_j = \sum_i w^i s_j^i. \qquad (5.6)$$

A Figura 5.11 apresenta os valores de Rac_{96} obtidos através desta medida de semelhança e nas restantes condições da Figura 5.10. Também incluídos na Figura 5.11, terceira linha, estão os números de blocos resultantes de cada geometria e combinação de *nh* e *nv*. Confrontando as Figuras 5.10 e 5.11, torna-se evidente que a comparação de descritores pela Eq. (5.6) é favorável relativamente à medida $d_{LBP}(D, D_j)$. Verifica-se também aqui, e de forma mais acentuada, que a divisão da imagem horizontalmente é mais benéfica do que o seccionamento vertical. Entre as geometrias consideradas, G1 é aquela que mais consistentemente oferece melhor desempenho, apesar de ser também aquela que conduz a descritores mais extensos (maior número de blocos da imagem), apenas superada, neste aspecto, por G3 quando *nh*=4 e *nv*=4. Contudo, se considerarmos o caso em que *nv*=1, o número de blocos por G1 é relativamente pequeno e o desempenho obtido é muito bom, em particular quando $nh \geq 3$. Focando o caso de *nh*=3 e *nv*=1, teremos com G1 um número de blocos igual a 5 e um desempenho elevado, superior a todas as outras condições com número de blocos igual ou inferior. Por esta razão, foi seleccionada para a definição do descritor LBP-Gist a geometria G1 com *nh*=3 e *nv*=1. Em resumo, a aplicação da característica LBP-Gist na detecção da revisitação de lugares segue o seguinte procedimento:

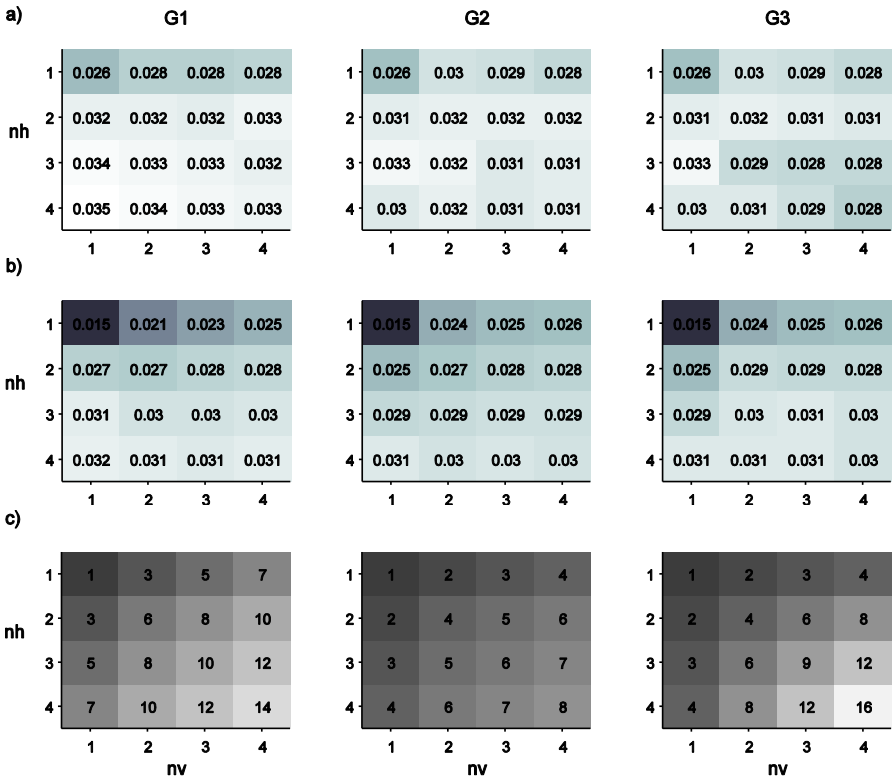


Figura 5.11. Linhas a) e b) valores de R_{ac96} para diferentes geometrias de seccionamento e função de semelhança dada pela Eq. (5.6). A) *dataset* Malaga, b) *dataset* City Centre. Linha c) número de blocos correspondente a cada seccionamento. O sombreado identifica, a claro, os valores mais elevados de R_{ac96} (linhas a e b) ou do nº de blocos da imagem (linha c).

1) Para todas as imagens é extraído o descritor LBP-Gist, envolvendo
    i) Conversão da imagem original para níveis de cinzento e resolução 240×320;
    ii) Aplicação do operador $LBP_{9,8}^{u4}$ com a função de *threshold* dada pela Eq. (5.6) e τ_0=5 e τ_1=0.03;
    iii) Construção do descritor LBP-Gist como o conjunto dos histogramas calculados sobre a geometria G1 com *nh*=3 e *nv*=1.
2) A decisão sobre a revisitação de um lugar é positiva quando uma imagem *j* do lugar verifica s_j>0.7. Este valor foi ajustado por forma a garantir precisão de 100% em todos os *datasets*.

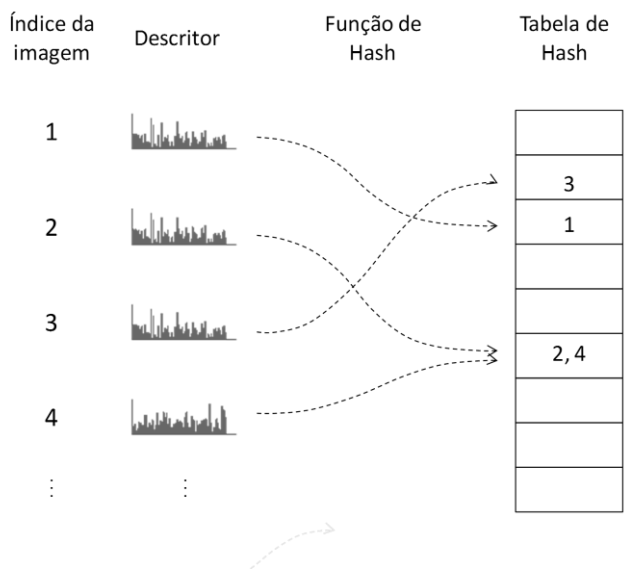


Figura 5.12. Conceitos fundamentais envolvidos nos algoritmos de *Locality Lensitive Hashing.*

5.5 Pesquisa de imagens por Winner Take All hashing

Os métodos de *Locality Sensitive Hashing* (LSH) facilitam a pesquisa rápida de imagens pela aplicação de um mapeamento em que vectores no espaço de representação escolhido são indexados numa *tabela de Hash*, através das chamadas *funções de dispersão*. A aplicação destes métodos pressupõe que, durante a construção da base de dados, todos os descritores são sujeitos a este procedimento, resultando no preenchimento da tabela com os índices das imagens (ver Figura 5.12). Na fase de pesquisa, a mesma função de dispersão é aplicada ao descritor de teste, apontando para uma entrada da tabela de *Hash*. Os índices das imagens contidos nessa entrada são encarados como os das imagens relevantes. Associado à pesquisa por LSH, está o conceito de colisão, que corresponde à situação em que duas imagens são mapeadas para a mesma entrada da tabela. Face a esta definição, pode dizer-se que as imagens devolvidas pelo algoritmo são aquelas que colidem com a imagem de teste.

No cerne da construção de um algoritmo de LSH está o desenvolvimento de funções de dispersão que promovem a colisão entre imagens semelhantes e evitam-na entre imagens distintas. Este objectivo é normalmente dado através de condições probabilísticas garantidas pelo algoritmo:

$$\begin{aligned} &P\big(f(D_1) = f(D_2)\big) > P_1 \\ &P\big(f(D_1) = f(D_3)\big) < P_2, \qquad P_1 > P_2 \end{aligned} \tag{5.7}$$

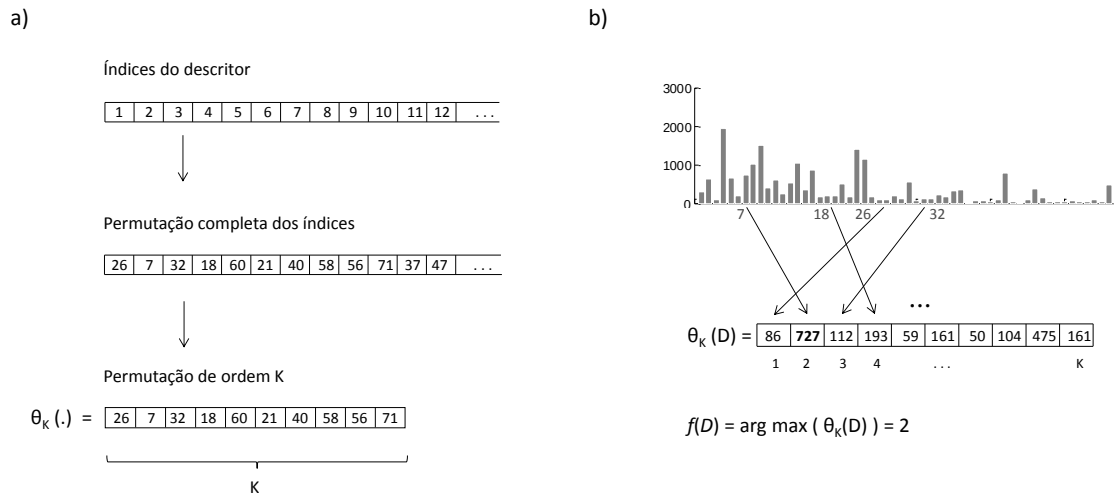


Figura 5.13. *Winner Take All hashing*: a) construção da função de permutação, b) aplicação da permutação e resultado da função de dispersão.

onde f é uma função de dispersão, D_1 e D_2 são descritores de imagens semelhantes e D_3 é distinto daquelas.

Estes são também os princípios na base do método WTA (*Winner Take All*), utilizado neste capítulo. Nesta secção descreve-se esse método e apresentam-se duas modificações ao algoritmo original, com vista a melhorar o seu desempenho.

5.5.1 *Winner Take All hashing*

O algoritmo WTA desenvolvido por Yagnik et al. (2011) assenta na utilização de funções de dispersão f_i com as características descritas de seguida. Em primeiro lugar, é criada uma função de permutação dos elementos do descritor e os primeiros K componentes dessa permutação são reservados, resultando numa permutação parcial θ_{Ki}. O valor de saída da função f_i, calculada sobre o descritor D, é obtido aplicando a permutação parcial e extraindo a posição do componente de maior valor do vector resultante, como ilustrado na Figura 5.13. Segundo Yagnik et al. (2011), a probabilidade de colisão entre duas imagens, através destas funções, é uma estimativa de uma medida de semelhança na categoria das semelhanças de comparação de ordem. Nesta classe de medidas, a semelhança entre dois descritores não depende dos valores absolutos dos seus componentes mas antes das relações *'maior que'* que se estabelecem entre estes e que são comuns aos dois descritores.

Tal como noutros métodos de LSH, a eficiência do algoritmo WTA pode ser melhorada com a formação de funções compostas designadas por *sketches*. Nesta

abordagem, um *sketch* sk_j de comprimento T é definido como um n-tuplo $sk_j = (f_{j,1}, f_{j,2},..., f_{j,T})$ em que $f_{j,i}$ são funções de dispersão geradas independentemente. Assumindo que, através de uma função $f_{j,i}$, a probabilidade de colisão entre duas imagens distintas é menor que P_2, este valor passa a P_2^T, quando se aplica um *sketch* de dimensão T. Consequentemente, esta abordagem é útil na redução do número de imagens irrelevantes que serão devolvidas. Por outro lado, ainda que a probabilidade de colisão entre imagens semelhantes seja maior do que entre imagens distintas, este valor será também reduzido ao ser elevado a T. Por esta razão, e para garantir uma probabilidade suficiente de se encontrar as imagens relevantes, é comum aplicar-se um número L de *sketches* em paralelo e devolver as imagens em que ocorrem colisões em pelo menos um deles.

Em muitas aplicações de LSH, a pesquisa termina após serem verificadas as colisões, e a lista devolvida é a das imagens em que ocorre pelo menos uma colisão. Este é o caso, por exemplo, do algoritmo E2LSH (Datar et al., 2004) que resolve o problema de encontrar os R-vizinhos próximos e que, como tal, foi desenvolvido para garantir que os vizinhos a uma distância menor que R do descritor de teste são devolvidos com probabilidade elevada. Contudo, outros algoritmos, tais como o WTA e o *min-Hash* (Broder, 1997), foram desenvolvidos tendo em vista garantir que o número de colisões aproxima uma medida de semelhança, permitindo que a informação contida nesse número possa ser usada com vantagens. Neste trabalho é adoptada essa perspectiva e são introduzidas duas modificações que beneficiam o algoritmo de pesquisa.

Seja $c_j(D_1,D_2)$, $c_j \in \{0,1\}$ a função indicadora da colisão entre os descritores D_1 e D_2 através de sk_j, com c_j=1quando ocorre colisão e c_j=0 no caso contrário. A estimativa de semelhança entre os descritores pode então ser definida pela acumulação de colisões encontradas em todos os sketches:

$$S(D_1, D_2) = \sum_{j=1}^{L} c_j(D_1, D_2). \qquad (5.8)$$

De seguida são introduzidos os dois métodos, modulação pelo termo *idf* e *threshold* sobre a relação de semelhanças (RST - *Relative Similarity Threshold*), que visam respectivamente melhorar a medida de semelhança e reduzir o número de imagens irrelevantes devolvidas pelo algoritmo.

Método *idf* - A técnica *tf-idf* (*term frequency - inverse document frequency*), tal como é definida no âmbito dos sistemas baseados em vocabulários, modula a contribuição de cada palavra com um factor dado pelo produto de dois termos, *tf* e *idf*. O primeiro indica a importância de uma palavra no documento de teste e é dado pela sua frequência nesse documento. Através do termo *idf* pretende-se diminuir a contribuição de palavras que são globalmente muito frequentes, fazendo depender esse termo da frequência da palavra no *corpus* de documentos. A modulação *tf-idf* foi anteriormente integrada num sistema de *hashing*, num trabalho em que as imagens eram representadas no modelo BoW (Chum, Philbin e Zisserman, 2008). Ao contrário desse estudo, no presente trabalho a representação da imagem é feita através de histogramas LBP, onde o conceito de palavra visual não é aplicável. No entanto, é também aqui possível aplicar a modulação *tf-idf* fazendo equivaler as entradas das tabelas de *Hash* a palavras visuais. Neste caso, o termo *tf* será igual a 1, já que cada imagem tem apenas uma entrada na tabela de dispersão e o termo *idf* é definido como

$$idf_j = log\left(\frac{N}{n_j}\right), \tag{5.9}$$

onde N é o número de imagens indexadas na tabela e n_j é o número de imagens que colidem na entrada j. A medida de semelhança modificada através deste esquema acumula o número de colisões, tal como na Eq. (5.8), que são agora pesadas pelo termo *idf*:

$$S(D_1, D_2) = \sum_{j=1}^{L} idf_j \cdot c_j(D_1, D_2). \tag{5.10}$$

Método RTS - Este é um método adaptativo que visa reduzir o número de imagens irrelevantes que são devolvidas, recorrendo para isso à medida de semelhança desenvolvida acima. A ideia essencial é a de substituir a aplicação de um *threshold* fixo aos valores de semelhança por um *threshold* variável, o que na prática se traduzirá numa condição expressa sobre a *razão* de semelhanças. Um critério deste tipo pode ser vantajoso em ambientes com muitos lugares idênticos, que produzem valores de semelhança altos. Nestes casos, um *threshold* fixo resultaria numa extensa lista de imagens, enquanto um *threshold* aplicado à razão de semelhanças pode ser útil na rejeição de lugares com menor probabilidade de corresponderem ao lugar actual.

Desenvolvendo esta ideia, definiu-se o método RST como um critério de selecção em que as imagens escolhidas verificam a condição:

$$\frac{S(D_1, D_2)}{S_{max}} > \tau, \tag{5.11}$$

onde S_{max} é o valor máximo de semelhança obtido na pesquisa e τ é uma constante inferior a 1. Neste trabalho τ foi ajustado empiricamente para o valor 0.8, produzindo resultados satisfatórios.

5.6 Resultados

5.6.1 *Datasets*

As técnicas propostas neste capítulo foram testadas em quatro *datasets*, correspondendo a quatro ambientes distintos e resumidos abaixo. Na Figura 5.14 apresentam-se imagens típicas de cada um deles e a Tabela 5.1 compila os dados mais importantes sobre eles. Nos casos em que o *dataset* é extraído de uma colecção mais vasta foram escolhidas as sequências para as quais existem resultados publicados.

Malaga - Este *dataset* corresponde à sessão campus-6L da colecção mais vasta conhecida por Malaga datasets 2009 (Blanco, Moreno e Gonzalez, 2009). O *dataset* retrata uma zona de parqueamento automóvel no campus da Universidade de Malaga e é caracterizado por descrever um espaço aberto, em resultado da pouca ocupação do ambiente e do posicionamento da câmara a uma elevação maior do que é habitual nestes sistemas. Em consequência disso, áreas significativas das imagens são ocupadas por céu.

City Centre - Este *dataset*, apresentado por Cummins e Newman (2008), descreve parte da zona urbana de Oxford e retrata um ambiente com muitos elementos dinâmicos devido ao tráfego e passagem de transeuntes. Relativamente aos restantes, este *dataset* tem a particularidade de as imagens serem captadas em sentido perpendicular ao movimento do robô, com uma câmara direccionada para a esquerda e outra para a direita.

New College - Recolhido nas imediações do edifício do New College em Oxford (Smith et al., 2009), este *dataset* descreve um campus universitário em grande parte ajardinado. Pelo facto de o robô se mover frequentemente em espaço aberto, e devido

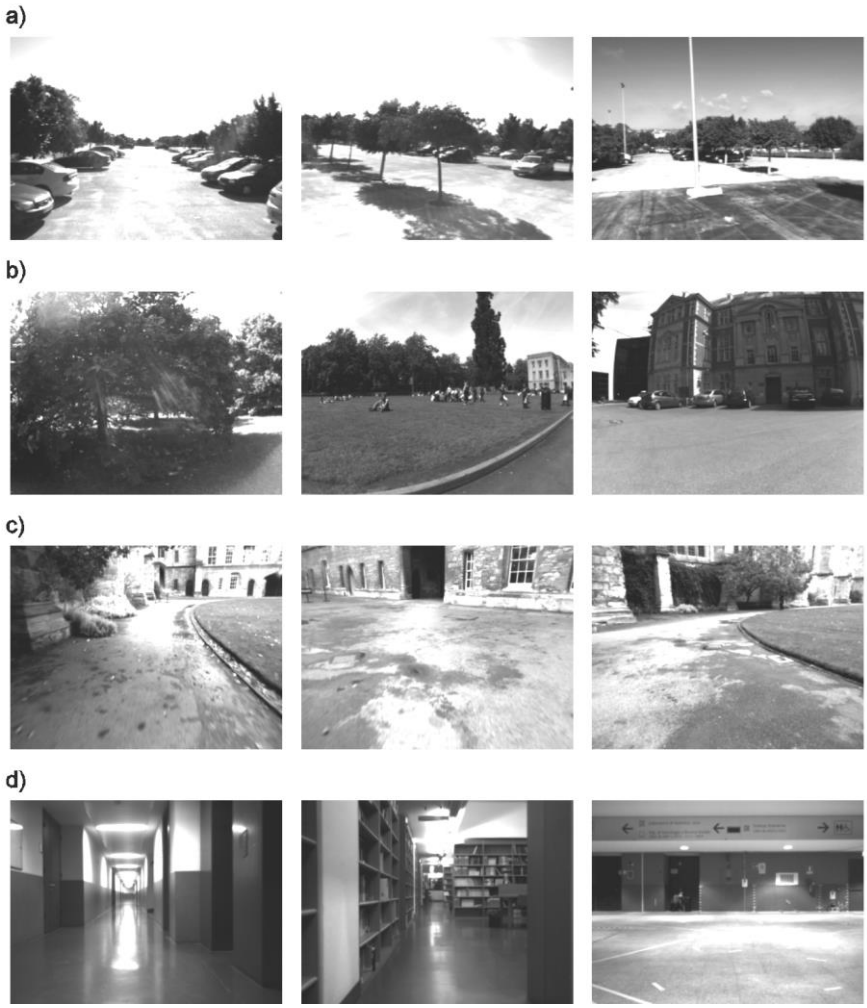


Figura 5.14. Imagens típicas dos *datasets* a) Malaga, b) City Centre, c) New College e d) Bicocca.

Tabela 5.2. Características principais dos *datasets* usados na avaliação.

Dataset	Distância percorrida [m]	Distância revisitada [m]	Amostragem	Sub-amostragem	Formato de imagem	Nº de imagens avaliadas
Malaga	1192	162	7.5fps	3.25 fps	768×1024 RGB	1737
City Centre	2025	801	1.5m	-	480×640 RGB	2474
New College	2260	1570	20 fps	2 fps	384×512 níveis de cinzento	5248
Bicocca	760	113	15 fps	1.875 fps	480×640 níveis de cinzento	5268

à orientação da câmara, os padrões captados são predominantemente do terreno do ambiente e, por isso, pouco informativos.

Bicocca - O *dataset* que designamos por Bicocca corresponde à sessão 2009-02-25b de captura de dados do projecto Rawseeds (Fontana, Matteucci e Sorrenti, 2014). O ambiente retratado é o de um piso de dois edifícios da Universidade de Milão-Bicocca, ligados entre si, e tem como elementos principais uma grande extensão de corredores uniformes, vários átrios e uma biblioteca.

À excepção do *dataset* City Centre, os restantes integram informação visual de duas câmaras formando um par stereo. Nestes casos, foram usadas apenas as imagens captadas pela câmara esquerda, já que a informação da segunda câmara é largamente redundante, quando não se extraem dados de profundidade. Relativamente ao *dataset* City Centre, as câmaras presentes estão orientadas em direcções opostas, o que justifica o uso das imagens de ambas pois fornecem informação distinta.

Os *datasets* de exterior são acompanhados de dados GPS do robô, permitindo validar a detecção de revisitação a partir dessa informação métrica. Assim, foi determinada a *groundtruth* associada a cada percurso considerando que existe revisitação quando a posição do robô dista menos de 6.5m de um lugar previamente ocupado. O *dataset* de interior, Bicocca, inclui também informação métrica, resultante da fusão de localizadores baseados em visão e *laser scanner*, a partir da qual se obteve a *groundtruth*, estabelecendo como raio de decisão 3m.

5.6.2 *Hashing* com WTA e tempos de execução

Na avaliação de desempenho do algoritmo WTA e das suas modificações foram realizados vários testes de pesquisa de imagens sobre os *datasets* Malaga e City Centre. Nas nossas experiências verificou-se que, para $T>3$, o algoritmo não é significativamente sensível à variação dos parâmetros T e K, tendo sido escolhidos para todos os testes os valores de $K=10$, $T=4$. O número de sketches (L) é importante no algoritmo, por representar o tamanho da amostra usada na estimação da semelhança entre descritores. Assim, o aumento de L é positivo para a precisão do algoritmo, embora acarrete maior peso computacional. Como valor de compromisso entre eficiência e boa precisão foi seleccionado $L=100$ para os testes seguintes. À semelhança do procedimento seguido em 4.4 a avaliação de desempenho é feita recorrendo às medidas de *recall* e *fall-out*, definidos naquela secção.

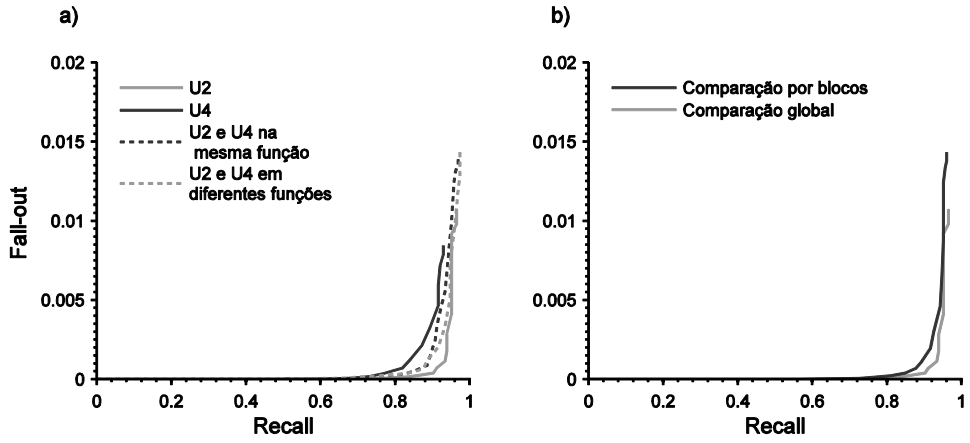


Figura 5.15. *Fall-out* vs *recall* medidos no *dataset* Malaga com diferentes configurações das funções de dispersão.

Na aplicação do algoritmo WTA sobre descritores que, tal como o LBP-Gist, têm os seus componentes agrupados por categorias, neste caso de Uniformidade e de blocos da imagem, deve ser considerada a forma como os diversos subgrupos são tratados nas funções de dispersão. Neste contexto, pode ser analisada a importância que cada subgrupo tem na precisão do algoritmo e também duas opções de configuração, em que as funções de dispersão são aplicadas independentemente para cada subgrupo, ou aplicadas sobre o descritor completo.

A Figura 5.15.a ilustra esta ideia, representando as curvas de *fall-out* vs *recall* obtidas quando se usam apenas os componentes de U=4 ou U=2 e também quando se usa a combinação dos dois. O último caso apresenta duas possibilidades de implementação, correspondentes à separação dos componentes U=4 e U=2 em diferentes funções de dispersão ou à utilização de funções que combinam os dois tipos de padrões.

Os resultados mostram que, comparativamente, os padrões com U=4 oferecem reduzido desempenho pois os valores de *fall-out* desta curva são significativamente maiores, para os mesmos valores de *recall*. Este facto dever-se-á à menor frequência de ocorrência destes padrões, o que torna a estimação da relação entre eles menos robusta. Este dado, aliado à maior diversidade destes padrões, a qual requer um maior número de tabelas de *Hash* para uma melhor estimação da semelhança, explica o pior desempenho destes padrões relativamente àqueles com U=2. As combinações dos dois tipos de padrões não representam um acréscimo de precisão relativamente ao caso U=2, justificando que a aplicação do algoritmo seja doravante restringida a este subgrupo.

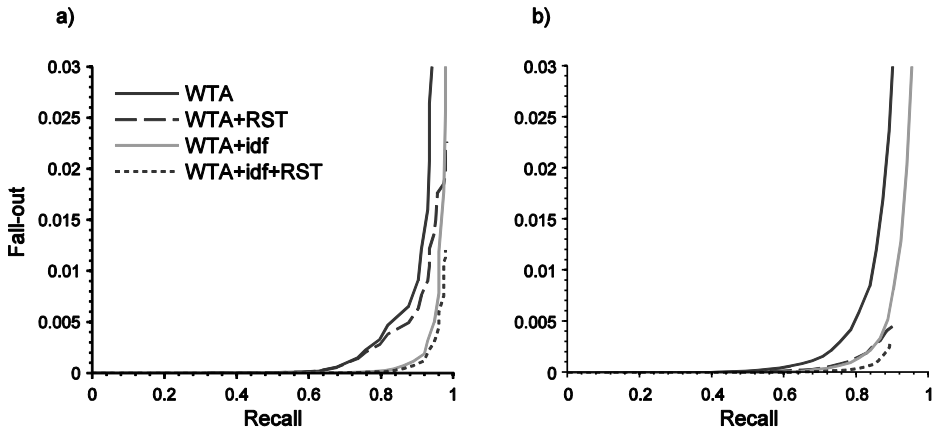


Figura 5.16. *Fall-out* vs *recall* para as diferentes versões do algoritmo WTA. A) *dataset* Malaga, b) *dataset* City Centre.

Os resultados anteriores foram obtidos ignorando a distribuição dos componentes do descritor pelos blocos da imagem. Na Figura 5.15.b a configuração escolhida atrás é comparada com aquela em que as funções de dispersão são aplicadas separadamente para cada bloco. Este caso corresponde a calcular a semelhança para cada bloco independentemente e, posteriormente, somar os resultados individuais. De forma diferente, no primeiro caso cada função de dispersão contribui para a estimação da semelhança global entre imagens. Os resultados da figura mostram que a comparação global é superior, o que é explicado pelo facto de as funções de dispersão globais introduzirem relações adicionais, entre blocos, e por isso usarem mais informação do que no caso em que os blocos são tratados separadamente.

A melhor configuração encontrada na Figura 5.15 (subgrupo U=2 e comparação global) é usada na Figura 5.16 para demonstrar os benefícios das modificações propostas ao algoritmo WTA. Nesta figura apresentam-se as curvas *fall-out* vs *recall* do algoritmo original e das suas modificações, aplicados aos *datasets* Malaga e City Centre. Os resultados da figura mostram que o método *idf* é muito eficaz na redução do número de imagens irrelevantes que são devolvidas, enquanto os benefícios do método RST não são tão consistentes. Apesar da redução ténue de *fall-out* no *dataset* Malaga, verifica-se uma redução importante no *dataset* City Centre o que justificou a utilização deste método nos ensaios seguintes.

Os tempos de computação associados ao algoritmo WTA são apresentados na Tabela 5.3. Estes valores e todas as restantes medições do peso computacional foram

Tabela 5.3. Tempos de computação [ms] envolvidos na selecção de imagens por WTA.

Dataset	Inserção na tabela de *Hash*		Pesquisa na tabela de *Hash*	
	média	máx.	média	máx.
Malaga	0.3	0.5	0.5	0.9
City Centre	0.3	0.6	0.5	0.9
New College	0.3	0.5	0.5	0.9
Bicocca	0.4	0.6	0.5	1.0
New College 10Hz	0.3	0.5	1.6	2.6

Tabela 5.4. Tempos de computação [ms] do sistema BoW no *dataset* New College 10Hz.

Extracção de características	Pesquisa e comparação no modelo BoW	Verificação geométrica	Sistema Completo
14.4	6.9	1.6	21.6

realizadas num computador com as características descritas no capítulo 4, secção 4.5. À semelhança do procedimento seguido por Galvez-López e Tardós (2012) considerou-se, para além dos *datasets* descritos em 5.5.1, uma versão do *dataset* New College em que a frequência das imagens é de 10 Hz. Este *dataset*, designado por *New College 10Hz*, será usado para avaliar a eficiência do algoritmo em *datasets* com dimensões elevadas, aqui de 26240 imagens. Para efeitos de comparação, na Tabela 5.4 apresentam-se os tempos de computação de um sistema BoW, apresentado por Galvez-López e Tardós (2012) como sendo mais rápido do que outros sistemas assentes no mesmo modelo.

O tempo de cálculo dos *sketches* e inserção na tabela de *Hash* é praticamente constante, de valor médio 0.3ms, enquanto a pesquisa na tabela de *Hash* já depende da dimensão do *dataset*. No caso mais exigente em termos computacionais, New College 10Hz, o tempo total das operações de *hashing* é em média inferior a 2ms, mostrando que esta forma de indexação é muito competitiva relativamente ao modelo BoW onde as operações de pesquisa ocuparam 6.9ms.

A Tabela 5.5 reúne os tempos de computação relativos à extracção da característica LBP-Gist e à comparação de descritores pela Eq. (5.6). Os tempos de extracção apresentam alguma variabilidade entre *datasets*, que está relacionada com a conversão das imagens originais para o formato de 240×320 píxeis em níveis de cinzento. No caso do *dataset* Malaga, em que esta conversão é mais exigente, a extracção tem custo

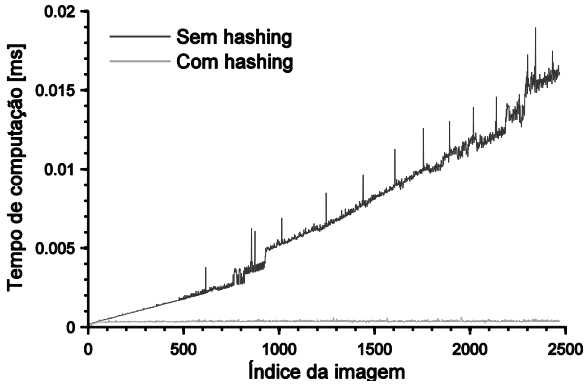


Figura 5.17. Tempos de comparação de descritores LBP-Gist no *dataset* City Centre.

Tabela 5.5. Tempos de computação [ms] na extracção e comparação de descritores LBP-Gist.

Dataset	Extracção do LBP-Gist		Cálculo de semelhança c/ *Hashing*		Cálculo de semelhança s/ *Hashing*	
	média	máx.	média	máx.	média	máx.
Malaga	4.1	4.8	0.3	0.6	4.0	10.3
City Centre	3.1	4.0	0.3	0.6	6.8	18.9
New College	2.5	3.2	0.3	0.6	22.7	51.7
Bicocca	2.5	3.8	0.3	0.6	22.7	50.4
New College 10Hz	2.5	3.4	0.3	1.0	132.1	283.9

médio de 4.1ms, significativamente menor que o valor de 14.4ms medido no trabalho de Galvez-López e Tardós (2012). É de salientar que naquele estudo as características usadas são consideradas muito eficientes, dentro do modelo BoW, por usarem o detector de pontos de interesse FAST e as características binárias BRIEF.

A Figura 5.17, que mostra os tempos de comparação de descritores com e sem pesquisa por *hashing* tornam evidente a importância de se realizar a selecção prévia de imagens. No segundo caso, o número de imagens a comparar cresce linearmente com o número de imagens capturadas e o tempo de comparação segue também essa tendência. Através da técnica de *hashing*, o número de imagens envolvidas na comparação é drasticamente menor, tendo valor médio de 49 no *dataset* New College 10Hz. Em consequência disso, o tempo de comparação é agora, em média, de 0.3ms, enquanto sem a aplicação de *hashing* era de 132.1ms. É de salientar também o facto de o tempo médio de computação com *hashing* não ser visivelmente afectado pela dimensão do *dataset*, o que sugere boas propriedades de escalabilidade do detector. Considerando os tempos de computação do sistema completo, apresentados na

Tabela 5.6. Tempos totais de computação [ms] do detector LBP-Gist.

Dataset	Tempo total de computação	
	média	máx.
Malaga	5.3	6.5
City Centre	4.3	5.6
New College	3.8	5.1
Bicocca	3.7	5.6
New College 10Hz	4.7	7.1

Tabela 5.6, o valor médio mais elevado é de 4.7ms, medido no *dataset* New College 10Hz, o qual representa uma redução relativamente ao sistema BoW (Tabela 5.4) por um factor de 0.22.

5.6.3 Precisão na detecção de lugares revisitados

A Figura 5.18 caracteriza o desempenho do detector LBP-Gist através das curvas de precisão e de *recall* em função do *threshold* aplicado à medida de semelhança dada pela Eq. (5.6). Assinalado nestes gráficos está também o limite mínimo de semelhança que foi usado para assinalar a revisitação com precisão de 100%. Os valores de *recall* correspondentes a esta condição estão reunidos na Tabela 5.7, onde se compara o desempenho do detector LBP-Gist com o detector BoW. Uma outra perspectiva do desempenho do detector proposto é dada na Figura 5.19, onde se apresentam as trajectórias descritas juntamente com as posições em que a detecção foi assinalada. Nesta figura excluiu-se o *dataset* New College em virtude de as falhas de GPS presentes nos dados não permitirem gerar uma trajectória inteligível.

Os dados da Tabela 5.7 mostram que o detector LBP-Gist é superior em dois casos, dos *datasets* Malaga e City Centre. No segundo, a diferença de desempenho é muito significativa, com valor de 38 pontos percentuais. No *dataset* New College, o detector LBP-Gist apresenta resultados idênticos aos do modelo BoW.

O *dataset* Bicocca destaca-se pelas dificuldades colocadas ao detector LBP-Gist. Neste caso o detector proposto apresenta um *recall* de 48.3%, enquanto o *recall* obtido pelo detector BoW é de 81.2%. Esta diferença de desempenho mostra que o detector LBP-Gist pode ser limitado no tratamento de ambientes interiores, que colocam condições distintas dos de exterior. A primeira dificuldade encontrada relaciona-se com a presença de *aliasing*, isto é, o facto de diferentes lugares terem aparências muito semelhantes. Este efeito é exemplificado na Figura 5.20, em que se

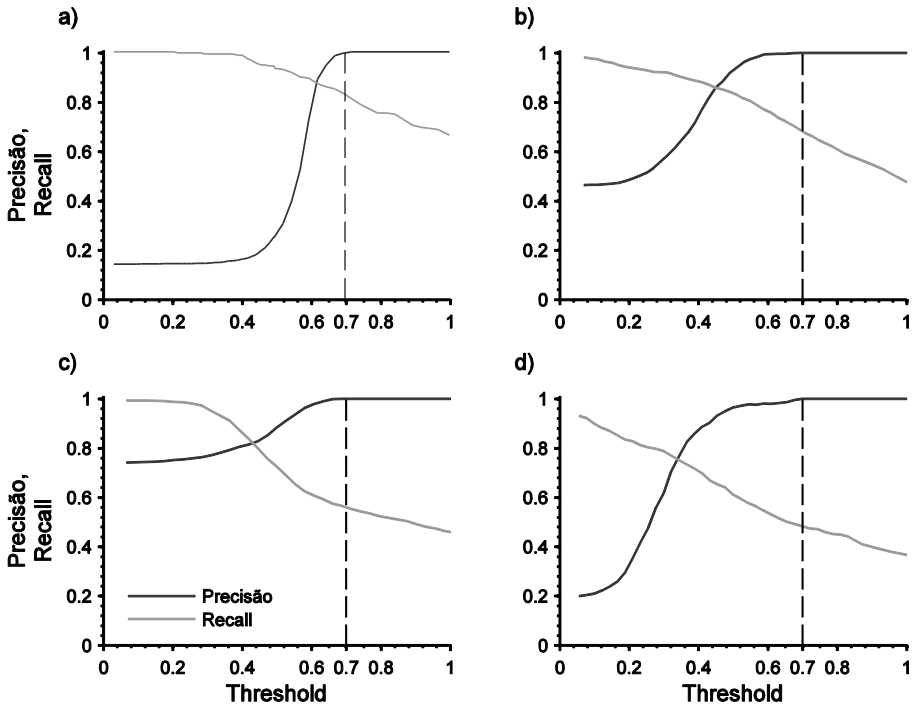


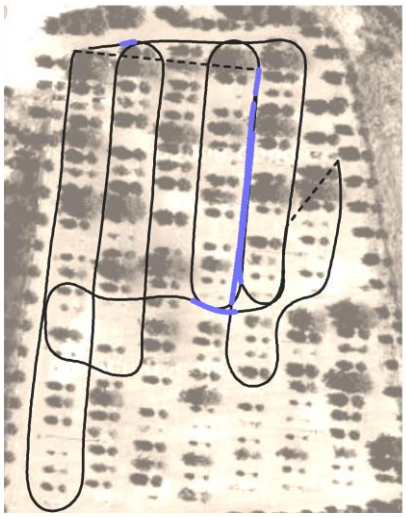
Figura 5.18. Precisão e *recall* como função do *threshold* aplicado à medida de semelhança. Os gráficos a) a d) dizem respeito aos *datasets* Malaga, City Centre, New College e Bicocca, respectivamente.

Tabela 5.7. Valores de *recall* [%] com precisão =100%.

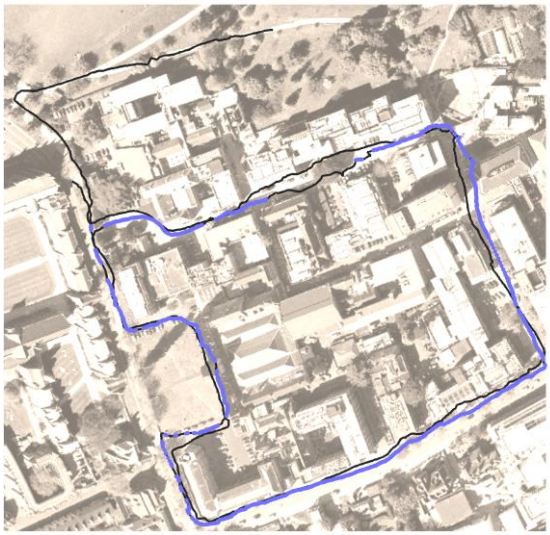
Dataset	LBP-Gist	BoW
Malaga	82.1	74.8
City Centre	68.7	30.6
New College	56.0	55.9
Bicocca	48.3	81.2

mostram imagens de dois lugares distintos, mas visualmente idênticos. Os ambientes de interior são mais susceptíveis de ocorrência destas condições, que, associadas ao facto de serem também mais pobres em texturas, colocam problemas a descritores globais como o LBP-Gist. A segunda dificuldade deve-se à forma abrupta como a aparência muda em resultado de pequenos desvios de posição. Este aspecto é ilustrado na Figura 5.21 onde se mostram imagens captadas em duas posições que distam 1.8m. A forte variação de aparência global levou a que não fosse detectada revisitação pelo detector LBP-Gist. Neste caso, um tratamento por características locais oferece maior probabilidade de detecção, dada a robustez dessa abordagem relativamente a variações de perspectiva.

a)



b)



c)

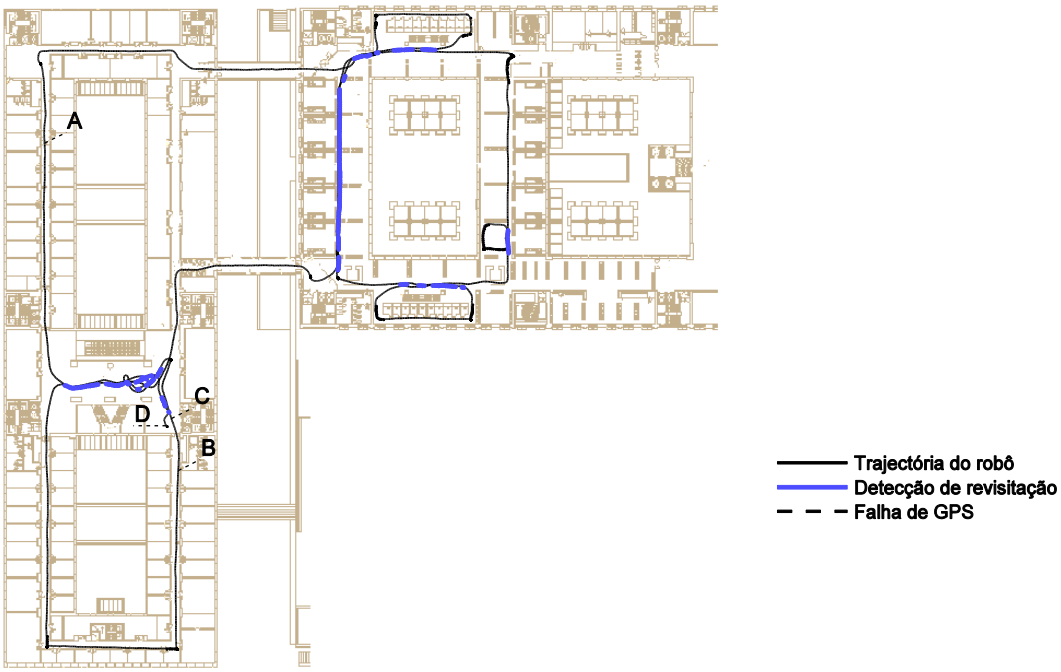


Figura 5.19. Percursos do robô e posições de revisitação que foram assinaladas nos *datasets* a) Malaga, b) City Centre e c) Bicocca.

Figura 5.20. Imagens de dois lugares distintos mas visualmente semelhantes (*dataset* Bicocca). As posições de captura estão assinaladas na Figura 5.19 com as letras A e B.

Figura 5.21. Imagens de um lugar em que pequenos desvios de posição produzem variações significativas na aparência global. As posições de captura estão assinaladas na Figura 5.19 com as letras C e D.

5.7 Discussão

Neste capítulo introduziu-se a característica LBP-Gist, a qual foi aplicada à detecção da revisitação de lugares. No desenvolvimento desta característica foram colocadas algumas restrições, com vista a limitar a dimensão do descritor. Com as restrições da análise de texturas a uma única escala e de P=8, a dimensão do descritor ficou dependente apenas do número de divisões da imagem. Por fim, a escolha de 3+2 divisões da imagem resultou em descritores de dimensão 995. Sob esta configuração, a característica LBP-Gist revelou bom desempenho nos *datasets* de exterior; num dos casos com uma diferença relativamente ao modelo BoW muito significativa.

É, no entanto, plausível que o desempenho da característica possa ser ainda incrementado, se a restrição ao tamanho do descritor for aliviada. Neste caso, a análise da imagem em múltiplas escalas pode ser particularmente vantajosa, dado o sucesso demonstrado, quer noutras implementações do Gist, quer na abordagem por características locais.

Relativamente ao *dataset* de interior, as experiências realizadas mostraram um desempenho claramente inferior do LBP-Gist relativamente à abordagem BoW. Este resultado, cremos, não se deve tanto a uma limitação particular da característica LBP-Gist, ou da configuração usada, mas a uma limitação transversal às características globais. De facto, as experiências realizadas sugerem-nos que os ambientes de interior colocam condições em que o modelo BoW é naturalmente superior à abordagem por características globais, devido i) à ocorrência de imagens pobremente texturadas e ii) ao forte impacto que a mudança de posição/orientação exerce sobre a aparência visual nestes ambientes.

Importa ainda nesta secção posicionar a característica LBP-Gist relativamente a outras características globais. Apesar de os trabalhos publicados sobre esta abordagem serem escassos, os autores Sunderhauf e Protzel (2011) bem como Liu e Zhang (2012) apresentam resultados de dois tipos de descritores aplicados à detecção de revisitação. No trabalho de Sunderhauf e Protzel (2011) é desenvolvido o descritor BRIEF-Gist, que, apesar de não ter os seus tempos de computação reportados, é mais simples do que o LBP-Gist. No entanto, o seu desempenho no *dataset* City Centre é inferior, com uma diferença de *recall* de 36.7 pontos percentuais. No caso dos autores Liu e Zhang (2012) foi usado o descritor Gist original, o que resultou num melhor desempenho no mesmo *dataset*, com uma diferença de *recall* de 18.3 pontos percentuais. Contudo a extracção desta característica tem custo computacional de 160ms, superior ao do LBP-Gist em 40 vezes, considerando o valor mais elevado da Tabela 5.5. Dados estes resultados, pode dizer-se que a característica LBP-Gist oferece uma solução intermédia, de equilíbrio entre rapidez de execução e boa precisão.

A eficiência do detector proposto é decorrente, em grande medida, da selecção de imagens pelo algoritmo WTA. Este algoritmo é, em si, de execução muito rápida e além disso revelou-se muito adequado à comparação de descritores LBP-Gist, o que é patente na relação *fall-out* vs *recall* observada nas experiências. Relativamente ao método mais popular, E2LSH (Datar et al., 2004), que compara descritores pela distância euclidiana, a validade do método WTA pode ser confirmada na Figura 5.22. Nesta figura, em que se compara a distância euclidiana com a distância calculada por WTA, verifica-se que a primeira proporciona uma pior relação *fallout* vs *recall*.

Entre as modificações propostas ao algoritmo WTA, o método *idf* destaca-se pela sua consistência na melhoria da medida de semelhança estimada. Embora as experiências

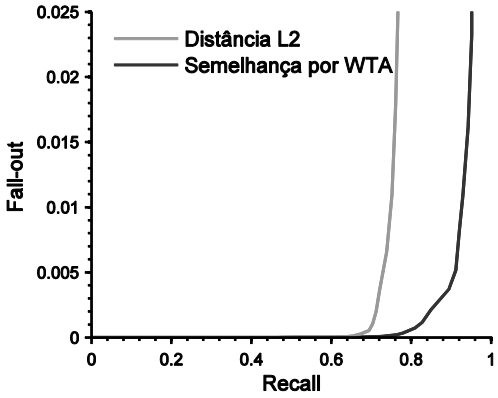


Figura 5.22. Comparação da distância euclideana com a distância estimada por WTA na pesquisa de imagens pelo descritor LBP-Gist. Estes resultados foram calculados sobre o *dataset* Malaga.

realizadas se restrinjam ao algoritmo WTA com descritores LBP-Gist, os benefícios encontrados poderão ser extensíveis a outros algoritmos de LSH e outros descritores. Além disso, este método constitui uma forma simples de adaptação do algoritmo de *hashing* aos dados de pesquisa, dispensando o treino de funções de dispersão, a solução encontrada noutros vários estudos para aumentar a precisão na pesquisa (Shakhnarovich, Viola e Darrell, 2003; Ruei-Sung, Ross e Yagnik, 2010; Strecha et al., 2012).

6. Conclusões

A investigação realizada nesta tese debruçou-se sobre a utilização da aparência visual em dois problemas fundamentais da robótica móvel: a localização global e a detecção da revisitação de lugares. Relativamente ao primeiro, foi proposto o classificador NQ, um método que faz uso da representação não quantizada de características locais de forma inovadora relativamente a trabalhos anteriores. No que concerne ao segundo problema, foi introduzida uma nova característica global para a descrição da aparência, o LBP-Gist, o qual se revelou adequado no tratamento da detecção de revisitação, especialmente em ambientes de exterior. Neste capítulo resumem-se as principais conclusões e contribuições desta tese e apontam-se algumas direcções para trabalho futuro.

6.1 Resumo e Contribuições

6.1.2 Localização

Nos últimos anos, a representação quantizada (Q) tem-se consolidado como o formato preferido para a utilização das características locais, devido à descrição concisa das imagens que dela decorre, bem como à rapidez de execução que lhe está associada. Na primeira parte desta tese, que abrange os capítulos 2 a 4, questiona-se esta opção através do estudo da abordagem alternativa, a representação não-quantizada (NQ). A motivação para este estudo tem origem na constatação, feita por vários autores, de que a operação de quantização introduz aproximações na descrição das características, que, por sua vez, limitam a precisão dos classificadores. Apesar do desempenho reconhecidamente sub-óptimo da representação Q, esta tem sido maioritariamente preferida, porque o grau de simplificação do classificador e a redução dos custos computacionais são, presumivelmente, mais importantes do que a perda de precisão. Esta premissa pode no entanto ser questionada, por duas razões: por um lado o problema da localização permanece desafiador quando há variações ambientais severas, e, nestes casos, os limites de precisão impostos pela representação Q têm maior relevância; por outro lado, a representação NQ não foi, no passado, amplamente estudada no sentido da simplificação dos seus algoritmos.

Nesta tese contribuiu-se para colmatar esta lacuna, recorrendo-se, para isso, ao classificador NQ proposto. Este, para além de apresentar maior precisão do que outros

classificadores baseados na representação NQ, admite simplificações importantes ao seu algoritmo que o tornam competitivo com a representação Q.

No capítulo 2 realizou-se uma avaliação inicial das duas representações, tendo sido seleccionados, para a localização com a representação Q, os classificadores SVM e Naive Bayes. Entre estes, o classificador Naive Bayes revelou maior precisão na maior variedade de condições de localização. Dado o seu bom desempenho, simplicidade e rapidez de execução, este classificador foi escolhido para as comparações posteriores com a representação NQ. Para além destas propriedades favoráveis, o classificador Naive Bayes admite uma análise simples da discriminatividade associada à representação Q, o que permitiu revelar propriedades importantes desta representação, e compará-la nesse aspecto com a representação NQ.

No capítulo 3 o classificador NQ foi estudado em maior detalhe, com a avaliação de diversas alternativas à combinação de características pela regra da soma. Algumas das extensões a esta regra produzem incrementos de precisão significativos relativamente aos resultados apresentados no capítulo 2, acentuando as diferenças de desempenho entre as representações Q e NQ.

No capítulo 4 desta tese estudaram-se os factores relativamente aos quais a representação NQ é, em princípio, menos atractiva do que a representação Q: os custos de memória e de tempo de execução. O peso destes factores é de facto importante pois o tempo de execução do classificador NQ, quando considerado na sua forma original pode atingir os 135ms, cerca de 5 vezes superior aos tempos mais elevados medidos com a representação Q. O objectivo no capítulo 4 foi o de demonstrar que a representação NQ admite simplificações significativas, através das quais é possível torná-la mais competitiva com a representação Q. Para este efeito foram propostas duas categorias de métodos que actuam respectivamente sobre os dois parâmetros determinantes para o custo da representação NQ: o número médio de características nos modelos dos lugares e o número de lugares processados em fase de localização.

Através dos métodos *Feature Merging* e *Feature Deletion*, da primeira categoria, é conseguida a compactação dos modelos dos lugares, no primeiro caso pela substituição de descritores semelhantes por um único descritor e, no segundo caso, pela aglomeração de características pouco frequentes. A aplicação destes métodos

produziu factores de redução no tempo de computação de 0.32 e 0.21 e na memória ocupada de 0.27 e 0.19, respectivamente para os *datasets* IDOL e FDF Park.

As técnicas na segunda categoria visam reduzir, de forma dinâmica, o número de lugares que são considerados em fase de localização. Esta redução acarreta a aceleração do algoritmo de localização, não tendo consequências relativamente à memória ocupada, já que os modelos de todos os lugares têm de ser mantidos em memória. A execução destes métodos é feita de forma faseada, com a técnica *Gist Selection* sendo aplicada na fase inicial e a técnica *Progressive Selection* realizada sequencialmente durante o processamento das características locais. Para o desenvolvimento da técnica *Gist Selection* recorreu-se a uma característica adicional, o Gist, que, apesar de ser menos robusto do que a análise por características locais, fornece informação suficiente para que se possa descartar lugares menos plausíveis. A pré-selecção obtida pelo *Gist Selection* é posteriormente refinada pela técnica *Progressive Selection*. Aqui, a evolução das distribuições à posteriori é analisada, à medida que as características locais são processadas, também no sentido de se eliminar os lugares menos plausíveis. Através das experiências realizadas, concluiu-se que a eficácia destes métodos varia com a dificuldade do problema, todavia, obteve-se sempre factores de redução importantes, com valores de 0.54 no pior caso e de 0.40 no caso mais favorável.

A comparação entre as duas representações completou-se, no capítulo 4, com a caracterização da precisão de cada uma delas como função dos custos associados. Esta análise é motivada, em primeiro lugar, pelo facto de a representação Q oferecer uma precisão que depende da dimensão do vocabulário visual, que, por sua vez, determina os seus custos computacionais. Por outro lado, também na representação NQ é possível obter uma gama de configurações, através dos métodos propostos, que oferecem diferentes compromissos entre precisão e custos computacionais.

A comparação levada a cabo (ilustrada nas Figuras 4.18 e 4.19) mostrou que, para os mesmos níveis de tempo de execução, a representação NQ mantém a superioridade em termos de precisão. Relativamente à memória ocupada, não foi sempre possível atingir os níveis obtidos com a representação Q e, quando isso acontece, apenas os valores mais elevados são atingidos. De facto, a redução conseguida na memória ocupada é menos acentuada do que no tempo de execução pois apenas dois dos métodos propostos afectam esta variável. Apesar disso, mesmo nos casos mais

exigentes o espaço de memória observado foi aceitável, com um valor máximo de 3.87MB no dataset FDF Park, que cobre uma área de 137.2×178.3m.

Neste ponto das conclusões justifica-se fazer uma súmula dos capítulos 2 a 4 com o objectivo de delinear uma resposta à questão sobre a pertinência e viabilidade da utilização da representação NQ. No capítulo 2 fez-se uma comparação inicial desta representação com dois classificadores baseados na representação Q, tendo sido verificada a maior precisão da primeira; no capítulo 3 justificou-se a opção pela regra da soma no classificador NQ e mostrou-se que o seu desempenho pode ser ainda melhorado, com as modificações àquela regra; o capítulo 4 comparou as duas representações em termos dos seus custos computacionais numa gama de configurações que, na representação NQ, foi possível obter através de um conjunto de métodos simplificativos. Sob essas simplificações mantém-se a superioridade da representação NQ, verificando-se ainda que a diferença de precisão é tipicamente mais pronunciada nas configurações com menor tempo de execução.

Em face do exposto, é possível afirmar que a representação NQ constitui uma alternativa competitiva, dado que a sua maior precisão, para tempos de execução semelhantes, justifica o seu uso em detrimento da representação Q. É necessário, todavia, salientar que os ganhos conseguidos são variáveis, dependendo da dificuldade do problema de localização. Com efeito, se os acréscimos de precisão são consideráveis no ambiente de interior, IDOL, no *dataset* FDF park a precisão de ambas as representações é superior a 90% e a diferença entre elas é menos significativa. Este dado confirma o bom desempenho da representação Q verificado em vários trabalhos anteriores e sugere que a escolha pela representação NQ é sobretudo adequada para ambientes em que as variações de aparência esperadas são mais severas.

Dadas as afinidades entre o classificador NQ e o NBNN, avançado por Boiman, Shechtman e Irani (2008) no artigo em que defendem o uso de características não quantizadas, importa nestas conclusões colocar os dois em perspectiva. Aquando da sua introdução o classificador NBNN foi inovador sobretudo por conjugar a representação NQ com a ideia de comparar imagens com classes, em lugar de imagens com imagens. Esta propriedade significa que as características da imagem de teste não são comparadas com as características de cada imagem de modelação, separadamente, mas antes com o conjunto de todas as características extraídas das

imagens de modelação de uma classe. Como observaram Rematas, Fritz e Tuytelaars (2013) este mecanismo permite que a imagem de teste possa ser vista como uma composição de características observadas em diferentes imagens dessa classe, o que aumenta a capacidade de generalização do classificador. Perante a simplicidade do classificador NBNN, o facto de o seu desempenho competir com o de classificadores treinados justificou a sua boa recepção pela comunidade científica. Mais tarde, Behmo et al. (2010) apontaram algumas fragilidades a este classificador que, segundo os autores, estarão na origem de alguma inconstância do seu desempenho em diferentes *datasets*. Behmo e os seus co-autores explicam este comportamento pela aproximação realizada na estimação de densidade por *kernel*, i.e., pela substituição da expressão 2.3 por 2.4, onde o factor de normalização é eliminado. Embora essa opção seja justificada, já que o somatório de probabilidades é substituído por uma única probabilidade, Behmo et al. (2010) alegam que o número de características no modelo das classes influencia aquele valor, levando a que o classificador NBNN seja enviesado, pois favorece as classes com maior número de características. De modo a evitar esta tendência, os autores propuseram um estimador que envolve dois parâmetros por classe, os quais são ajustados por um método de optimização. Embora este procedimento melhore significativamente o desempenho do classificador, ele implica uma fase de treino, o que retira um dos atractivos do classificador original, o de não necessitar de pré-processamento.

O classificador NQ parte também dos princípios fundamentais do classificador NBNN (não-quantização e comparação de 'imagem com classe') e introduz contribuições que são atingidas por uma via diferente da seguida por Behmo et al. (2010). O desenvolvimento deste classificador teve por base o estudo sobre a combinação de características no âmbito da combinação de múltiplos classificadores, realizado no capítulo 3. Neste estudo foram comparados os combinadores mais importantes que se adequam ao tratamento de características locais, nomeadamente as regras da soma e do produto – e as suas extensões –, e o método de votação de Li, Yang e Kosecka (2005). Entre os métodos testados, a regra do produto é aquela que corresponde ao classificador NBNN, pois a combinação de múltiplos classificadores por esta regra é equivalente a um classificador Naive Bayes (Kittler et al., 1998). A comparação entre os combinadores revelou, tal como em (Behmo et al., 2010), limites no desempenho

do classificador NBNN e permitiu identificar as melhores configurações para o classificador NQ.

Fundamentalmente, a constatação da superioridade da regra da soma permite concluir que os limites do classificador NBNN são inerentes à regra do produto, a qual é especialmente sensível a dados com ruído. Em problemas de localização visual este é um aspecto essencial a ter em conta, pois as variações de aparência fazem emergir características que não foram visualizadas durante a modelação do ambiente. A análise de discriminatividade realizada no capítulo 3 revelou o impacto destas características: na regra do produto as características de discriminatividade baixa apresentam maior dispersão em torno de 0 e, igualmente importante, o número de características de discriminatividade alta é reduzido. Devido a estes factores, a regra do produto foi consistentemente inferior à regra da soma, observando-se diferenças de precisão que atingem um máximo de cerca de 8.1 pontos percentuais (Tabela 3.3, condições *night-sunny*).

No capítulo 3 identificaram-se medidas para aumentar a precisão proporcionada pelo classificador NBNN, que consistem em modificações à regra do produto. Entre as modificações testadas, a mais eficaz é a que aplica a limitação por *Threshold* aos valores de probabilidade (secção 3.3.2). Através desta operação impede-se que as probabilidades tomem valores próximos de zero, que teriam os impactos negativos descritos na secção 3.4.2. Na versão modificada, o desempenho da regra do produto é substancialmente melhor, com acréscimos de precisão que atingem os 11.2 pontos percentuais (Tabela 3.3, condições *night-sunny*). Para além destes dados, dois resultados interessantes puderam ser observados na aplicação da operação de *Threshold*. O primeiro prende-se com a regra do produto que, após modificação por *Threshold*, apresenta um perfil de discriminatividade semelhante ao da soma. Notando que esta é a versão de melhor desempenho do produto, pode concluir-se que a regra da soma constitui um padrão de desempenho a atingir. O segundo resultado relaciona-se com a aplicação do *Threshold* à regra da soma. Embora esta modificação tenha sido originalmente pensada para o produto (Alkoot e Kittler, 2002) verifica-se que ela é útil também na regra da soma.

Tendo em conta a superioridade da regra da soma relativamente ao produto, que se estende às modificações por *Threshold*, seleccionou-se esta regra para a combinação de características no classificador NQ. Com a opção por esta regra, complementada

pela operação de *Threshold*, ultrapassam-se os limites do classificador NBNN, tal como em (Behmo et al., 2010). Contudo no classificador proposto, estes benefícios são atingidos apenas com um único parâmetro, *Th*, e sem a necessidade de treino, pois, como foi demonstrado na secção 3.5.2.a é possível prever este parâmetro com base nas propriedades do modelo do ambiente.

6.1.2 Detecção da revisitação de lugares

O capítulo 5 estendeu o âmbito desta tese à construção de mapas topológicos, a etapa anterior à localização. Como foi mencionado na Introdução, actualmente este problema é resolvido pela combinação de dois processos, um de primeiro plano, que consiste na detecção de revisitação, e outro, de segundo plano, que corrige o mapa existente sob os constrangimentos métricos e topológicos recolhidos pelo processo anterior. Os métodos de correcção desenvolvidos mais recentemente são dotados de robustez à presença de *outliers*, i.e. admitem a existência de ligações incorrectas no mapa preliminar, aliviando, por isso, a exigência colocada sobre o detector de revisitação. No entanto, o peso computacional colocado sobre estes algoritmos está relacionado com a qualidade do mapa preliminar (Agarwal et al., 2013), o que justifica, ainda, a investigação sobre detectores de revisitação que superem o estado da arte.

Concretamente, os métodos que têm sido considerados no estado da arte são aqueles que recorrem ao modelo BoW, de que o sistema Fab-Map (Cummins e Newman, 2008) é um exemplo paradigmático. O trabalho desenvolvido nesta tese segue uma via diferente, inserindo-se na categoria de soluções que, em lugar de usarem características locais, recorrem a uma característica global. À luz dos capítulos 2 a 4, esta abordagem não se apresenta de forma evidente como a conclusão lógica dos resultados então obtidos, já que, como se viu, uma forma de superar o modelo BoW seria usar ainda características locais, mas na forma não quantizada. A opção por características globais é justificada pelas diferentes condições em que os problemas de localização e detecção de revisitação são colocados. Embora se tenha verificado que a análise por características locais pode ser aperfeiçoada pelo uso da representação NQ, os benefícios desta verificam-se sobretudo quando há variações de aparência muito significativas, que se devem, normalmente, ao facto de o mapeamento e a localização serem muito espaçados no tempo. No caso da revisitação, esta ocorre após intervalos

de tempo que, comparativamente, podem ser mais curtos. Por esta razão, as exigências colocadas sobre o detector de revisitação pesam menos sobre a robustez perante variações de aparência e mais sobre a capacidade de distinguir diferentes lugares. Nos trabalhos sobre características globais pretende-se demonstrar que estas características podem oferecer estas propriedades, com vantagens sobre o modelo BoW, por exigirem menores recursos computacionais e, eventualmente, por proporcionarem melhor relação precisão vs *recall*. Sunderhauf e Protzel (2011) e Arroyo et al. (2014a) propuseram características que de facto envolvem tempos de execução mais baixos. No primeiro caso, a característica BRIEF-Gist apresentou valores de *recall* ligeiramente abaixo dos do modelo BoW, no *dataset* City Centre, enquanto no segundo a característica LDB apresentou um desempenho superior, medido no *dataset* KITTI Odometry (Geiger, Lenz e Urtasun, 2012). Contudo, em ambos os casos a aplicação do modelo BoW não envolveu o teste de consistência geométrica das características locais que, como demonstrado por Galvez-López e Tardós (2012), é determinante para o sucesso deste modelo. Nos testes descritos no capítulo 5 a característica LBP-Gist é avaliada nos *datasets* usados por Galvez-López e Tardós (2012), permitindo, por isso, compara-la com o modelo BoW na sua versão mais completa. Os resultados obtidos demonstraram que o LBP-Gist é mais eficiente do que o modelo BoW, sobretudo devido ao menor tempo de extracção da característica. Em termos de *recall*, o LBP-Gist compara-se favoravelmente com o modelo BoW nos *datasets* de exterior, apresentando num dos casos uma diferença muito significativa. A mesma relação não se verifica no *dataset* de interior, em que o LBP-Gist é nitidamente inferior ao modelo BoW. Este facto dever-se-á às especificidades dos ambientes de interior, nomeadamente à ocorrência de superfícies pouco texturadas e à variação de aparência mais acentuada para pequenos desvios de posição/orientação do robô. Note-se que nos trabalhos que avaliaram outras características globais estas foram testadas apenas em *datasets* de exterior, pelo que não é possível afirmar que as limitações encontradas sejam especificas do LBP-Gist. De facto, a robustez da análise por características locais perante variações de perspectiva, aliada ao facto de as superfícies não texturadas serem ignoradas nesta abordagem, são factores que explicam o seu melhor desempenho relativamente às características globais.

6.2 Trabalho futuro

Actualmente, as características binárias têm ganho popularidade relativamente às mais tradicionais SIFT e SURF, devido à sua forma compacta e tempos de extracção e de comparação muito baixos. Estas propriedades tornam, naturalmente, as características binárias muito adequadas à utilização na forma não-quantizada, pois aliviam os custos computacionais desta representação. Contudo, num trabalho em que estas características foram usadas no reconhecimento de lugares (Galvez-López e Tardós, 2012), elas foram ainda enquadradas no modelo BoW, tendo sido desenvolvido, naquele trabalho, um método de extracção de palavras visuais específico para descritores binários. Uma questão que permanece em aberto, e que é levantada pelos resultados desta tese, refere-se ao potencial de desempenho destas características, quando usadas na forma não-quantizada. Assim, o estudo sobre as duas representações aplicadas a características binárias constitui um desenvolvimento natural do presente trabalho e é particularmente pertinente pois os efeitos da quantização sobre estes descritores não foram ainda analisados em profundidade. Para um estudo deste tipo existem vários descritores actualmente dísponíveis que constituem candidatos interessantes tais como o BRIEF, o ORB e o BRISK.

As questões levantadas na comparação entre o LBP-Gist e o modelo BoW sugerem algumas perspectivas de trabalhos futuros, com vista a melhorar o desempenho da característica proposta em ambientes de interior. Uma abordagem simples que pode ser seguida para este efeito passa por combinar a informação de textura com outras dimensões de informação, como por exemplo a disparidade obtida por câmaras stereo, que já foi usada com sucesso na detecção de revisitação (Arroyo et al., 2014a), ou a informação de cor, que não foi ainda suficientemente explorada neste problema. Outra abordagem, mais desafiadora, envolveria a combinação de texturas com mapas de saliência (Siagian e Itti, 2009) calculados sobre a imagem. Uma vez disponível, esta informação poderia ser usada de duas formas: a primeira consiste na modulação da contribuição de cada píxel para os histogramas LBP. Em suma, o cálculo dos histogramas seria feito por uma soma ponderada, em que os píxeis mais salientes receberiam maior peso. Desta forma evitar-se-ia que zonas pouco texturadas tivessem uma contribuição preponderante no descritor final e, desejavelmente, obter-se-iam descritores mais discriminativos. A segunda estratégia possível na utilização de mapas de saliência passaria pela utilização destes mapas para obter uma segmentação da

imagem. Esta segmentação constituiria uma partição alternativa à divisão fixa que foi usada nesta tese, com a vantagem de ser mais robusta relativamente a mudanças de perspectiva pois, em lugar de ser pré-definida, seria baseada no conteúdo da imagem. O desafio colocado nesta abordagem é o de desenvolver modelos de saliência que sejam suficientemente simples para a aplicação em causa. Actualmente, os detectores de saliência existentes apresentam tempos de computação que podem ser uma ordem de grandeza mais elevados do que os do detector LBP-Gist e a sua aplicação colocaria em causa os benefícios deste detector. Daí que, na investigação sobre este tipo de extensão do detector LBP-Gist, está inerente o desenvolvimento de modelos de saliência mais rápidos, mas suficientemente precisos para fornecer informação útil ao detector.

6.3 Nota final

As técnicas baseadas na aparência têm-se revelado ferramentas fundamentais na resolução dos problemas de localização e mapeamento em robótica móvel. Actualmente estes métodos são objecto de intensa investigação, com vista à sua utilização em aplicações reais. Na concretização desta ambição será necessário desenvolver soluções robustas, por estarem sujeitas à condições menos controladas do mundo real, e eficientes, já que têm de partilhar recursos computacionais com módulos de controlo, percepção, planeamento, comunicação, etc. Esperamos que o trabalho desenvolvido nesta tese contribua para o avanço destas soluções, por um lado proporcionando um conhecimento mais profundo sobre as características visuais locais, amplamente utilizadas na localização e, por outro, consolidando a abordagem por características globais como uma alternativa viável aos métodos tradicionais na detecção de revisitação.

Anexo A

Este anexo apresenta resultados relativos ao *dataset* FDF Park através dos quais se compara o desempenho de várias regras de combinação. Estes dados complementam o o capítulo 3, distribuindo-se da seguinte forma: as tabelas A.1, A.3 e A.5 apresentam a precisão do conjunto de regras nas sequências de modelação A, C e D, respectivamente, e as tabelas A.2, A.4 e A.6 apresentam, para as mesmas sequências, os valores de precisão após a inclusão do *kernel* geométrico.

Tabela A.1. Precisão [%] dos métodos de combinação sobre o *dataset* FDF Park, dados de modelação A.

Conjunto de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
A	94.87	100.00	99.23	100.00	100.00	100.00	100.00
B	82.30	96.71	92.54	97.07	95.83	97.81	97.00
C	92.37	99.45	98.51	99.06	98.19	99.21	99.06
D	94.62	98.80	99.15	98.72	98.80	98.72	98.38
H	84.74	96.93	92.29	97.80	97.01	97.95	97.88
I	83.63	94.80	88.98	96.20	93.64	96.35	96.59
J	91.72	97.79	97.24	97.48	97.24	97.56	97.40
K	92.61	100.00	99.23	99.66	99.48	99.57	99.83
Média	**89.61**	**98.06**	**95.90**	**98.25**	**97.52**	**98.40**	**98.27**

Tabela A.2. Precisão [%] dos métodos de combinação sobre o *dataset* FDF Park, dados de modelação A, integrando os constrangimentos geométricos.

Conjunto de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
A	97.93	100.00	100.00	100.00	100.00	100.00	100.00
B	87.71	97.37	95.46	97.22	97.15	97.15	96.85
C	94.73	99.45	99.92	99.45	99.61	99.21	99.37
D	96.16	99.57	99.91	99.23	99.83	99.23	99.15
H	91.11	97.88	95.44	97.95	97.56	98.11	97.64
I	86.50	96.43	92.09	96.66	94.57	97.05	96.90
J	93.14	98.11	98.58	98.03	97.63	97.79	97.71
K	95.79	100.00	99.91	99.83	99.74	99.91	100.00
Média	**92.88**	**98.60**	**97.66**	**98.55**	**98.26**	**98.56**	**98.45**

Tabela A.3. Precisão [%] dos métodos de combinação sobre o *dataset* FDF Park, dados de modelação C.

Conjunto de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
A	97.32	99.85	93.1	100.00	100.00	100.00	99.92
B	94.81	100.00	98.32	100.00	99.93	100.00	100.00
C	100.00	100.00	100.00	100.00	100.00	100.00	100.00
D	100.00	100.00	100.00	100.00	100.00	100.00	100.00
H	95.28	100.00	97.48	100.00	100.00	100.00	100.00
I	99.15	100.00	98.06	100.00	100.00	100.00	100.00
J	99.92	100.00	99.92	100.00	100.00	100.00	100.00
K	97.25	100.00	94.50	100.00	99.83	100.00	100.00
Média	**97.97**	**99.98**	**97.67**	**100.00**	**99.97**	**100.00**	**99.99**

Tabela A.4. Precisão [%] dos métodos de combinação sobre o *dataset* FDF Park, dados de modelação C, integrando os constrangimentos geométricos.

Conjunto de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
A	99.08	100.00	94.48	100.00	99.92	100.00	100.00
B	98.54	100.00	99.27	100.00	100.00	100.00	100.00
C	100.00	100.00	100.00	100.00	100.00	100.00	100.00
D	100.00	100.00	100.00	100.00	100.00	100.00	100.00
H	98.74	100.00	98.11	100.00	100.00	100.00	100.00
I	99.69	100.00	98.60	100.00	100.00	100.00	100.00
J	99.92	100.00	100.00	100.00	100.00	100.00	100.00
K	98.45	100.00	94.93	100.00	100.00	100.00	100.00
Média	**99.30**	**100.00**	**98.17**	**100.00**	**99.99**	**100.00**	**100.00**

Tabela A.5. Precisão [%] dos métodos de combinação sobre o *dataset* FDF Park, dados de modelação D.

Conjunto de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
A	80.23	99.77	92.72	99.62	99.62	99.69	99.77
B	95.17	100.00	100.00	100.00	100.00	100.00	100.00
C	96.93	99.84	99.21	99.92	100.00	99.92	99.84
D	99.15	100.00	100.00	100.00	100.00	100.00	100.00
H	95.59	100.00	100.00	100.00	100.00	100.00	100.00
I	100.00	100.00	100.00	100.00	100.00	100.00	100.00
J	96.37	99.68	96.92	99.45	99.45	99.76	99.76
K	91.57	100.00	99.05	99.91	100.00	99.91	100.00
Média	**94.38**	**99.91**	**98.49**	**99.86**	**99.88**	**99.91**	**99.92**

Tabela A.6. Precisão [%] dos métodos de combinação sobre o *dataset* FDF Park, dados de modelação D, integrando os constrangimentos geométricos.

Conjunto de teste	Votação	Soma	Produto	Soma + NEAT	Produto + NEAT	Soma + *Threshold*	Produto + *Threshold*
A	92.87	100.00	96.78	99.85	100.00	100.00	100.00
B	99.41	100.00	100.00	100.00	100.00	100.00	100.00
C	97.8	99.92	99.69	100.00	100.00	100.00	100.00
D	99.4	100.00	100.00	100.00	100.00	100.00	100.00
H	99.53	100.00	100.00	100.00	100.00	100.00	100.00
I	100	100.00	100.00	100.00	100.00	100.00	100.00
J	97.08	99.76	97.4	99.61	99.61	100.00	99.92
K	97.59	100.00	99.4	100.00	100.00	100.00	100.00
Média	**97.96**	**99.96**	**99.16**	**99.93**	**99.95**	**100.00**	**99.99**

Bibliografia

Agarwal, P., Tipaldi, G. D., Spinello, L., Stachniss, C. e Burgard, W. 2013. Robust map optimization using dynamic covariance scaling. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 6-10 May 2013.

Agrawal, M. e Konolige, K. 2006. Real-time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS. In Proc. of the 18th International Conference on Pattern Recognition (ICPR), 20-24 Aug. 2006.

Ahonen, T., Hadid, A. e Pietikäinen, M. 2004. "Face Recognition with Local Binary Patterns." In *Computer Vision - ECCV 2004*, edited by Tomás Pajdla e Jiří Matas, 469-481. Springer Berlin Heidelberg.

Alkoot, F. M. e Kittler, J. 2002. "Modified product fusion." *Pattern Recognition Letters*, no. 23 (8):957-965.

Andreasson, H., Duckett, T. e Lilienthal, A. J. 2008. "A minimalistic approach to appearance-based visual SLAM." *IEEE Transactions on Robotics*, no. 24 (5):991-1001.

Angeli, A., Filliat, D., Doncieux, S. e Meyer, J. 2008. "A Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words." *IEEE Transactions on Robotics, special issue on Visual Slam*, no. 24 (5):1027-1037.

Arampatzis, A., Kamps, J. e Robertson, S. 2009. Where to stop reading a ranked list?: threshold optimization using truncated score distributions. In *32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Boston, MA, USA: ACM.

Arampatzis, A. e Robertson, S. 2011. "Modeling score distributions in information retrieval." *Information Retrieval*, no. 14 (1):26-46.

Arampatzis, A., Zagoris, K. e Chatzichristofis, S. 2011. "Dynamic Two-Stage Image Retrieval from Large Multimodal Databases." In *Advances in Information Retrieval*, edited by Paul Clough, Colum Foley, Cathal Gurrin, et al., 326-337. Springer Berlin Heidelberg.

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Yebes, J. J. e Bronte, S. 2014a. Fast and effective visual place recognition using binary codes and disparity information. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 14-18 Sept. 2014.

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Yebes, J. J. e Gamez, S. 2014b. Bidirectional loop closure detection on panoramas for visual navigation. In Proc. of the IEEE Intelligent Vehicles Symposium, 8-11 June 2014.

Badino, H., Huber, D. e Kanade, T. 2012. Real-time topometric localization. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 14-18 May 2012.

Bajcsy, R. 1973. Computer description of textured surfaces. In *3rd International joint Conference on Artificial Intelligence*. Stanford, USA: Morgan Kaufmann Publishers Inc.

Bauer, J., Sunderhauf, N. e Protzel, P. 2007. Comparing several implementations of two recently published feature detectors. In Proc. of the International Conference on Intelligent and Autonomous Systems, 3-5 Sept. 2007.

Bay, H., Ess, A., Tuytelaars, T. e Van Gool, L. 2008. "Speeded-Up Robust Features (SURF)." *Computer Vision and Image Understanding*, no. 110 (3):346-359.

Behmo, R., Marcombes, P., Dalalyan, A. e Prinet, V. 2010. Towards optimal naive bayes nearest neighbor. In Proc. of the 11th European Conference on Computer Vision, 5-10 Sept. 2010, at Heraklion, Crete, Greece.

Bianconi, F. e Fernández, A. 2007. "Evaluation of the effects of Gabor filter parameters on texture classification." *Pattern Recognition*, no. 40 (12):3325-3335.

Blaer, P. e Allen, P. 2002. Topological mobile robot localization using fast vision techniques. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 2002.

Blanco, J.-L., Moreno, F.-A. e Gonzalez, J. 2009. "A collection of outdoor robotic datasets with centimeter-accuracy ground truth." *Autonomous Robots*, no. 27 (4):327-351.

Boiman, O., Shechtman, E. e Irani, M. 2008. In defense of Nearest-Neighbor based image classification. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

Booij, O., Terwijn, B., Zivkovic, Z. e Krose, B. 2007. Navigation using an appearance based topological map. In Proc. of the IEEE International Conference on Robotics and Automation, 10-14 April 2007.

Broder, A. Z. 1997. On the resemblance and containment of documents. In Proc. of the Compression and Complexity of Sequences, 11-13 Jun 1997.

Burghouts, G. J., Smeulders, A. W. M. e Geusebroek, J. M. 2007. The Distribution Family of Similarity Distances. In Proc. of the Advances in Neural Information Processing Systems, 3-6 Dec. 2007.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C. e Fua, P. 2012. "BRIEF: Computing a Local Binary Descriptor Very Fast." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 34 (7):1281-1298.

Calonder, M., Lepetit, V., Strecha, C. e Fua, P. 2010. "BRIEF: Binary Robust Independent Elementary Features." In *Computer Vision – ECCV*, edited by Kostas Daniilidis, Petros Maragos e Nikos Paragios, 778-792. Springer Berlin Heidelberg.

Campos, F., Correia, L. e Calado, J. M. F. 2015. "Robot Visual Localization Through Local Feature Fusion: An Evaluation of Multiple Classifiers Combination Approaches." *Journal of Intelligent & Robotic Systems*, no. 77 (2):377-390.

Campos, F. M., Correia, L. e Calado, J. 2010. A Probabilistic Approach to Appearance-Based Localization and Mapping. In Proc. of the European Conference on Artificial Intelligence (ECAI), 16-20 Aug. 2010.

Campos, F. M., Correia, L. e Calado, J. M. F. 2011. Mobile robot global localization with non-quantized SIFT features. In Proc. of the International Conference on Advanced Robotics (ICAR), 20-23 June 2011.

Campos, F. M., Correia, L. e Calado, J. M. F. 2012. "Global localization with non-quantized local image features." *Robotics and Autonomous Systems*, no. 60 (8):1011-1020.

Campos, F. M., Correia, L. e Calado, J. M. F. 2013a. Development of a Visual Loop Closure Detector in Matlab. In Proc. of the 1st International Conference on Algebraic and Symbolic Computation, 9-10 Sept. 2013, at Lisbon.

Campos, F. M., Correia, L. e Calado, J. M. F. 2013b. An evaluation of local feature combiners for robot visual localization. In Proc. of the 13th International Conference on Autonomous Robot Systems (Robotica), 24-25 April 2013.

Campos, F. M., Correia, L. e Calado, J. M. F. 2013c. "Loop Closure Detection with a Holistic Image Feature." In *Progress in Artificial Intelligence. 16th Portuguese Conference on Artificial Intelligence, EPIA 2013*, edited by Luís Correia, Luís Paulo Reis e José Cascalho, 247-258. Springer.

Chamorro-Martínez, J., Galán-Perales, E., Prados-Suárez, B. e Soto-Hidalgo, J. M. 2007. "Perceptually-Based Functions for Coarseness Textural Feature Representation." In *Pattern Recognition and Image Analysis*, edited by Joan Martí, José Miguel Benedí, Ana Maria Mendonça, et al., 579-586. Springer Berlin Heidelberg.

Chapelle, O., Haffner, P. e Vapnik, V. N. 1999. "Support vector machines for histogram-based image classification." *IEEE Transactions on Neural Networks*, no. 10 (5):1055-1064.

Chen, S. F. e Goodman, J. 1996. An empirical study of smoothing techniques for language modeling. In *34th annual meeting on Association for Computational Linguistics*. Santa Cruz, California: Association for Computational Linguistics.

Chih-Wei, H. e Chih-Jen, L. 2002. "A comparison of methods for multiclass support vector machines." *IEEE Transactions on Neural Networks*, no. 13 (2):415-425.

Chow, C. e Liu, C. 1968. "Approximating discrete probability distributions with dependence trees." *IEEE Transactions on Information Theory*, no. 14 (3):462-467.

Chum, O., Philbin, J. e Zisserman, A. 2008. Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In Proc. of the British Machine Vision Conference, 1-4 Sept. 2008.

Churchill, W. e Newman, P. 2012. Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 14-18 May 2012.

Crestani, F., Lalmas, M., Rijsbergen, C. J. V. e Campbell, I. 1998. "'Is this document relevant? ... Probably': a survey of probabilistic models in information retrieval." *ACM Computing Surveys*, no. 30 (4):528-552.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C. e Maupertuis, D. 2004. Visual Categorization with Bags of Keypoints. In Proc. of the ECCV Workshop on Statistical Learning in Computer Vision, 15 May 2004.

Cummins, M. e Newman, P. 2008. "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance." *The International Journal of Robotics Research*, no. 27 (6):647-665.

Datar, M., Immorlica, N., Indyk, P. e Mirrokni, V. S. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *20th Annual Symposium on Computational Geometry*. Brooklyn, New York, USA: ACM.

Datta, R., Joshi, D., Li, J. e Wang, J. Z. 2008. "Image retrieval: Ideas, influences, and trends of the new age." *ACM Computing Surveys*, no. 40 (2):1-60.

Daugman, J. G. 1985. "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters." *Journal of the Optical Society of America*, no. 2 (7):1160-1169.

Deselaers, T., Keysers, D. e Ney, H. 2008. "Features for image retrieval: an experimental comparison." *Information Retrieval*, no. 11 (2):77-107.

Di, H., Caifeng, S., Ardabilian, M., Yunhong, W. e Liming, C. 2011. "Local Binary Patterns and Its Application to Facial Image Analysis: A Survey." *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, no. 41 (6):765-781.

Diephuis, M., Voloshynovskiy, S., Koval, O. e Beekhof, F. 2011. Statistical analysis of binarized SIFT descriptors. In Proc. of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA), 4-6 Sept. 2011.

Domingos, P. e Pazzani, M. 1996. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In Proc. of the 13th International Conference on Machine Learning, 13-6 July 996.

Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L. e Schmid, C. 2009. Evaluation of GIST descriptors for web-scale image search. In Proc. of the ACM International Conference on Image and Video Retrieval, 2009, at Santorini, Fira, Greece.

Filliat, D. 2007. A visual bag of words method for interactive qualitative localization and mapping. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 10-14 April 2007.

Fischler, M. A. e Bolles, R. C. 1981. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." *Communications of the ACM*, no. 24 (6):381-395.

Fontana, G., Matteucci, M. e Sorrenti, D. 2014. "Rawseeds: Building a Benchmarking Toolkit for Autonomous Robotics." In *Methods and Experimental Techniques in Computer Engineering*, edited by Francesco Amigoni e Viola Schiaffonati, 55-68. Springer International Publishing.

Fox, D., Burgard, W. e Thrun, S. 1999. "Markov Localization for Mobile Robots in Dynamic Environments." *Journal of Artificial Intelligence Research*, no. 11:391-427.

Fraundorfer, F., Engels, C. e Nister, D. 2007. Topological mapping, localization and navigation using image collections. In Proc. of the EEE/RSJ International Conference on Intelligent Robots and Systems, 2007, at San Diego, CA.

Frese, U., Larsson, P. e Duckett, T. 2005. "A multilevel relaxation algorithm for simultaneous localization and mapping." *IEEE Transactions on Robotics*, no. 21 (2):196-207.

Fumera, G. e Roli, F. 2005. "A theoretical and experimental analysis of linear combiners for multiple classifier systems." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 27 (6):942-956.

Galvez-López, D. e Tardós, J. D. 2012. "Bags of Binary Words for Fast Place Recognition in Image Sequences." *IEEE Transactions on Robotics*, no. 28 (5):1188-1197.

Geiger, A., Lenz, P. e Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 16-21 June 2012.

Gerstmayr-Hillen, L., Schluter, O., Krzykawski, M. e Moller, R. 2011. Parsimonious loop-closure detection based on global image-descriptors of panoramic images. In Proc. of the 15th International Conference on Advanced Robotics (ICAR), 20-23 June 2011.

Goedemé, T., Nuttin, M., Tuytelaars, T. e Van Gool, L. 2007. "Omnidirectional Vision Based Topological Navigation." *International Journal of Computer Vision*, no. 74 (3):219-236.

Gotlieb, C. C. e Kreyszig, H. E. 1990. "Texture descriptors based on co-occurrence matrices." *Computer Vision, Graphics, and Image Processing*, no. 51 (1):70-86.

Grisetti, G., Kummerle, R., Stachniss, C., Frese, U. e Hertzberg, C. 2010. Hierarchical optimization on manifolds for online 2D and 3D mapping. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 3-7 May 2010.

Gross, H., Koenig, A., Schroeter, C. e Boehme, H. J. 2003. Omnivision-based probabilistic self-localization for a mobile shopping assistant continued. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 27-31 Oct. 2003.

Gurban, M. e Thiran, J. P. 2008. Dynamic modality weighting for multi-stream hmms in audio-visual speech recognition. In Proc. of the 10th conference on Multimodal Interfaces, 20-22 Oct. 2008.

Haralick, R. M. 1979. "Statistical and structural approaches to texture." *Proceedings of the IEEE*, no. 67 (5):786-804.

Heckmann, M., Berthommier, F. e Kroschel, K. 2002. "Noise adaptive stream weighting in audio-visual speech recognition." *EURASIP Journal on Applied Signal Processing* (11):1260-1273.

Heikkila, M. e Pietikainen, M. 2006. "A texture-based method for modeling the background and detecting moving objects." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* , no. 28 (4):657-662.

Howarth, P. e Ruger, S. 2005. "Robust texture features for still-image retrieval." *IEE Proceedings -Vision, Image and Signal Processing*, no. 152 (6):868-874.

Ishiguro, H., Ng, K. C., Capella, R. e Trivedi, M. M. 2003. "Omnidirectional image-based modeling: three approaches to approximated plenoptic representations." *Machine Vision and Applications*, no. 14 (2):94-102.

Islam, M. M., Dengsheng, Z. e Guojun, L. 2008. A geometric method to compute directionality features for texture images. In Proc. of the IEEE International Conference on Multimedia and Expo, June 23-April 26 2008.

Jiang, Y.-G., Ngo, C.-W. e Yang, J. 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. In Proc. of the 6th ACM International Conference on Image and Video Retrieval, 9-11 July 2007, at Amsterdam, The Netherlands.

Jianguo, Z., Marszalek, M., Lazebnik, S. e Schmid, C. 2006. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. In Proc. of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), 17-22 June 2006.

Jianxin, W. e Rehg, J. M. 2011. "CENTRIST: A Visual Descriptor for Scene Categorization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 33 (8):1489-1501.

Jing, H., Kumar, S. R., Mitra, M., Wei-Jing, Z. e Zabih, R. 1997. Image indexing using color correlograms. In Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 17-19 Jun 1997.

Jogan, M. e Leonardis, A. 2003. "Robust localization using an omnidirectional appearance-based subspace model of environment." *Robotics and Autonomous Systems*, no. 45 (1):51-72.

Jogan, M., Leonardis, A., Wildenauer, H. e Bischof, H. 2002. Mobile robot localization under varying illumination. In Proc. of the 16th International Conference on Pattern Recognition, 11-15 Aug. 2002

Jojic, N., Frey, B. J. e Kannan, A. 2003. Epitomic analysis of appearance and shape. In Proc. of the IEEE International Conference on Computer Vision, 13-16 Oct. 2003.

Jones, K. S. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation*, no. 28 (1):11-21.

Jones, S. D., Andresen, C. e Crowley, J. L. 1997. Appearance based process for visual navigation. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 7-11 Sep 1997.

Jurie, F. e Triggs, B. 2005. Creating efficient codebooks for visual recognition. In Proc. of the International Conference on Computer Vision (ICCV), 17-21 Oct. 2005

Kaess, M., Ranganathan, A. e Dellaert, F. 2007. iSAM: Fast Incremental Smoothing and Mapping with Efficient Data Association. In Proc. of the IEEE International Conference on Robotics and Automation, 10-14 April 2007.

Kai, N., Kannan, A., Criminisi, A. e Winn, J. 2008. Epitomic location recognition. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 23-28 June 2008.

Khan, N. Y., McCane, B. e Wyvill, G. 2011. SIFT and SURF Performance Evaluation against Various Image Deformations on Benchmark Dataset. In Proc. of the International Conference on Digital Image Computing Techniques and Applications (DICTA), 6-8 Dec. 2011.

Kittler, J., Hatef, M., Duin, R. P. W. e Matas, J. 1998. "On combining classifiers." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 20 (3):226-239.

Kokare, M., Biswas, P. K. e Chatterji, B. N. 2005. "Texture image retrieval using new rotated complex wavelet filters." *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* no. 35 (6):1168-1178.

Kosecka, J. e Fayin, L. 2004. Vision based topological Markov localization. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), April 26-May 1, 2004.

Krajnik, T., Fentanes, J. P., Mozos, O. M., Duckett, T., Ekekrantz, J. e Hanheide, M. 2014. Long-term topological localisation for service robots in dynamic environments using spectral maps. In Proc. of the International Conference on Intelligent Robots and Systems (IROS), 14-18 Sept. 2014.

Kuncheva, L. 2004. *Combining Pattern Classifiers: Methods and Algorithms*: JOHN WILEY & SONS.

Kunz, C., Willeke, T. e Nourbakhsh, I. 1997. Automatic mapping of dynamic office environments. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 20-25 Apr 1997.

Laine, A. e Fan, J. 1993. "Texture classification by wavelet packet signatures." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 15 (11):1186-1191.

Leung, T. e Malik, J. 2001. "Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons." *International Journal of Computer Vision*, no. 43 (1):29-44.

Leutenegger, S., Chli, M. e Siegwart, R. Y. 2011. BRISK: Binary Robust invariant scalable keypoints. In Proc. of the IEEE International Conference on Computer Vision (ICCV), 6-13 Nov. 2011.

Levitt, T. S. e Lawton, D. T. 1990. "Qualitative navigation for mobile robots." *Artificial Intelligence*, no. 44 (3):305-360.

Li, F., Yang, X. e Kosecka, J. 2005. "Global Localization and Relative Positioning Based on Scale-Invariant Keypoints." *Robotics and Autonomous Systems*, no. 52 (1):27-38.

Liao, S., Law, M. W. K. e Chung, A. C. S. 2009. "Dominant Local Binary Patterns for Texture Classification." *IEEE Transactions on Image Processing*, no. 18 (5):1107-1118.

Lindeberg, T. 1998. "Feature Detection with Automatic Scale Selection." *International Journal of Computer Vision*, no. 30 (2):79-116.

Liu, Y. e Zhang, H. 2012. Visual loop closure detection with a compact image descriptor. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 7-12 Oct. 2012.

Lowe, D. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision*, no. 60 (2):91-110.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In Proc. of the IEEE International Conference on Computer Vision, 1999.

Lowe, D. G. 2012. Local Naive Bayes Nearest Neighbor for image classification. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 16-21 June 2012.

Lu, F. e Milios, E. 1997. "Globally Consistent Range Scan Alignment for Environment Mapping." *Autonomous Robots*, no. 4 (4):333-349.

Luo, J., Pronobis, A., Caputo, B. e Jensfelt, P. 2006. The KTH-IDOL2 Database. Technical Report CVAP304. Stockholm, Sweden: Kungliga Tekniska Hoegskolan, CVAP/CAS.

Mäenpää, T. e Pietikäinen, M. 2003. "Multi-scale Binary Patterns for Texture Analysis." In *Image Analysis*, edited by Josef Bigun e Tomas Gustavsson, 885-892. Springer Berlin Heidelberg.

Mäenpää, T. e Pietikäinen, M. 2005. "Texture analysis with local binary patterns." *Handbook of Pattern Recognition and Computer Vision*, no. 3:197-216.

Mallat, S. G. 1989. "A theory for multiresolution signal decomposition: the wavelet representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 11 (7):674-693.

Manjunath, B. S., Salembier, P. e Sikora, T. 2002. *Introduction to MPEG-7: Multimedia Content Description Interface*: John Wiley & Sons, Inc.

Matas, J., Chum, O., Urban, M. e Pajdla, T. 2004. "Robust wide-baseline stereo from maximally stable extremal regions." *Image and Vision Computing*, no. 22 (10):761-767.

Matsuyama, T., Miura, S.-I. e Nagao, M. 1983. "Structural analysis of natural textures by Fourier transformation." *Computer Vision, Graphics, and Image Processing*, no. 24 (3):347-362.

Mikolajczyk, K. e Schmid, C. 2002. "An Affine Invariant Interest Point Detector." In *Computer Vision — ECCV 2002*, edited by Anders Heyden, Gunnar Sparr, Mads Nielsen, et al., 128-142. Springer Berlin Heidelberg.

Mikolajczyk, K. e Schmid, C. 2005. "A performance evaluation of local descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 27 (10):1615-1630.

Miksik, O. e Mikolajczyk, K. 2012. Evaluation of local detectors and descriptors for fast feature matching. In Proc. of the International Conference on Pattern Recognition (ICPR), 11-15 Nov. 2012.

Milford, M. e Wyeth, G. 2010. "Persistent navigation and mapping using a biologically inspired SLAM system." *The International Journal of Robotics Research*, no. 29 (9):1131-1153.

Misra, H., Bourlard, H. e Tyagi, V. 2003. New entropy based combination rules in HMM/ANN multi-stream ASR. In Proc. of the Acoustics, Speech, and Signal Processing, 6-10 April 2003.

Mojsilovic, A. e Soljanin, E. 2001. "Color quantization and processing by Fibonacci lattices." *IEEE Transactions on Image Processing*, no. 10 (11):1712-1725.

Murase, H. e Nayar, S. 1995. "Visual learning and recognition of 3-d objects from appearance." *International Journal of Computer Vision*, no. 14 (1):5-24.

Murillo, A. C. e Kosecka, J. 2009. Experiments in place recognition using gist panoramas. In Proc. of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 27 Sept. - 4 Oct. 2009.

Murillo, A. C., Sagues, C., Guerrero, J. J., Goedemé, T., Tuytelaars, T. e Van Gool, L. 2007. "From omnidirectional images to hierarchical localization." *Robotics and Autonomous Systems*, no. 55 (5):372-382.

Murillo, A. C., Singh, G., Kosecka, J. e Guerrero, J. J. 2013. "Localization in Urban Environments Using a Panoramic Gist Descriptor." *IEEE Transactions on Robotics*, no. 29 (1):146-160.

Nanni, L., Lumini, A. e Brahnam, S. 2010. "Local binary patterns variants as texture descriptors for medical image analysis." *Artificial Intelligence in Medicine*, no. 49 (2):117-125.

Neubert, P., Sunderhauf, N. e Protzel, P. 2013. Appearance change prediction for long-term navigation across seasons. In Proc. of the European Conference on Mobile Robots (ECMR), 25-27 Sept. 2013.

Nilsson, N. J. 1984. Shakey the robot. SRI AI Center, Menlo Park, California.

Ojala, T., Pietikäinen, M. e Harwood, D. 1996. "A comparative study of texture measures with classification based on featured distributions." *Pattern Recognition*, no. 29 (1):51-59.

Ojala, T., Pietikainen, M. e Maenpaa, T. 2002. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 24 (7):971-987.

Oliva, A. e Torralba, A. 2001. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope." *International Journal of Computer Vision*, no. 42 (3):145-175.

Oliva, A. e Torralba, A. 2006. "Building the gist of a scene: the role of global image features in recognition." In *Progress in Brain Research*, edited by S. L. Macknik L. M. Martinez J. M. Alonso S. Martinez-Conde e P. U. Tse, 23-36. Elsevier.

Olson, E., Leonard, J. e Teller, S. 2006. Fast iterative alignment of pose graphs with poor initial estimates. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 15-19 May 2006.

Palm, C. 2004. "Color texture classification by integrative Co-occurrence matrices." *Pattern Recognition*, no. 37 (5):965-976.

Pass, G., Zabih, R. e Miller, J. 1996. Comparing images using color coherence vectors. In Proc. of the 4th ACM International Conference on Multimedia, 18-22 Nov. 1996, at Boston, Massachusetts, USA.

Paulevé, L., Jégou, H. e Amsaleg, L. 2010. "Locality sensitive hashing: A comparison of hash function types and querying mechanisms." *Pattern Recognition Letters*, no. 31 (11):1348-1358.

Peker, K. A. 2011. Binary SIFT: Fast image retrieval using binary quantized SIFT features. In Proc. of the 9th International Workshop on Content-Based Multimedia Indexing (CBMI), 13-15 June 2011.

Pickering, M. J. e Rüger, S. 2003. "Evaluation of key frame-based retrieval techniques for video." *Computer Vision and Image Understanding*, no. 92 (2–3):217-235.

Pronobis, A., Caputo, B., Jensfelt, P. e Christensen, H. I. 2006. A Discriminative Approach to Robust Visual Place Recognition. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 9-15 Oct. 2006.

Puzicha, J., Buhmann, J. M., Rubner, Y. e Tomasi, C. 1999. Empirical evaluation of dissimilarity measures for color and texture. In Proc. of the IEEE International Conference on Computer Vision, 1999.

Ramisa, A., Tapus, A., Aldavert, D. e Toledo, R. 2009. "Robust vision-based robot localization using combinations of local feature region detectors." *Autonomous Robots*, no. 27 (4):373-385.

Ranganathan, A. e Dellaert, F. 2004. Inference in the space of topological maps: an MCMC-based approach. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), 28 Sept.-2 Oct. 2004.

Ranganathan, A. e Dellaert, F. 2005. Data driven MCMC for appearance-based topological mapping. In Proc. of the Robotics: Science and Systems Conference I (RSS), 8-11 June 2005.

Rematas, K., Fritz, M. e Tuytelaars, T. 2013. The pooled NBNN kernel: beyond image-to-class and image-to-image. In Proc. of the 11th Asian conference on Computer Vision, at Daejeon, Korea.

Rennie, J. D., Shih, L., Teevan, J. e Karger, D. R. 2003. Tackling the poor assumptions of naive bayes text classifiers. In Proc. of the 20th International Conference on Machine Learning, 2003.

Roberti de Siqueira, F., Robson Schwartz, W. e Pedrini, H. 2013. "Multi-scale gray level co-occurrence matrices for texture description." *Neurocomputing*, no. 120 (0):336-345.

Robot Electronics. 2014. *Drive Systems*, (acedido a 21 Jan 2015). Disponível em http://www.robot-electronics.co.uk/acatalog/Drive_Systems.html.

Rosten, E. e Drummond, T. 2006. "Machine Learning for High-Speed Corner Detection." In *Computer Vision – ECCV 2006*, edited by Aleš Leonardis, Horst Bischof e Axel Pinz, 430-443. Springer Berlin Heidelberg.

Rublee, E., Rabaud, V., Konolige, K. e Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In Proc. of the IEEE International Conference on Computer Vision (ICCV), 6-13 Nov. 2011.

Ruei-Sung, L., Ross, D. A. e Yagnik, J. 2010. SPEC hashing: Similarity preserving algorithm for entropy-based coding. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 13-18 June 2010.

Rui, H., Ruger, S., Dawei, S., Haiming, L. e Zi, H. 2008. Dissimilarity measures for content-based image retrieval. In Proc. of the IEEE International Conference on Multimedia and Expo, June 23-April 26 2008.

Russell, S. e Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*: Prentice Hall Press.

Sanderson, C. e Paliwal, K. K. 2003. "Noise compensation in a person verification system using face and multiple speech features." *Pattern Recognition*, no. 36:293-302.

Schleicher, D., Bergasa, L. M., Ocana, M., Barea, R. e Lopez, M. E. 2009. "Real-Time Hierarchical Outdoor SLAM Based on Stereovision and GPS Fusion." *IEEE Transactions on Intelligent Transportation Systems*, no. 10 (3):440-452.

Scholkopf, B. e Smola, A. J. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*: MIT Press.

Schubert, F., Spexard, T. P., Hanheide, M. e Wachsmuth, S. 2007. Active vision-based localization for robots in a home-tour scenario. In Proc. of the International Conference on Computer Vision Systems 21-24 March at Bielefeld, Germany

Schwartz, W. R., Roberti de Siqueira, F. e Pedrini, H. 2012. "Evaluation of feature descriptors for texture classification." *Journal of Electronic Imaging*, no. 21 (2):023016-1-023016-17.

Se, S., Lowe, D. e Little, J. 2001. Vision-based mobile robot localization and mapping using scale-invariant features. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 2001.

Se, S., Lowe, D. e Little, J. 2002. Global localization using distinctive visual features. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2002.

Seymour, R., Stewart, D. e Ming, J. 2007. Audio-visual integration for robust speech recognition using maximum weighted stream posteriors. In Proc. of the 8th Annual Conference of the International Speech (INTERSPEECH), 2007.

Shakhnarovich, G., Viola, P. e Darrell, T. 2003. Fast pose estimation with parameter-sensitive hashing. In Proc. of the 9th IEEE International Conference on Computer Vision, 13-16 Oct. 2003.

Shatkay, H. e Kaelbling, L. P. 2002. "Learning geometrically-constrained hidden Markov models for robot navigation: bridging the topological-geometrical gap." *Journal of Artificial Intelligence Research*, no. 16 (1):167-207.

Shengcai, L., Guoying, Z., Kellokumpu, V., Pietikainen, M. e Li, S. Z. 2010. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 13-18 June 2010.

Shu, L., Wei, F., Chung, A. S. e Dit-Yan, Y. 2006. Facial Expression Recognition using Advanced Local Binary Patterns, Tsallis Entropies and Global Appearance Features. In Proc. of the IEEE International Conference on Image Processing, 8-11 Oct. 2006.

Siagian, C. e Itti, L. 2007. "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 29 (2):300-312.

Siagian, C. e Itti, L. 2009. "Biologically Inspired Mobile Robot Vision Localization." *IEEE Transactions on Robotics*, no. 25 (4):861-873.

Simmons, R. e Koenig, S. 1995. Probabilistic robot navigation in partially observable environments. In Proc. of the 14th international joint conference on Artificial intelligence, 1995, at Montreal, Quebec, Canada.

Sivic, J. e Zisserman, A. 2003. Video Google: a text retrieval approach to object matching in videos. In Proc. of the 9th IEEE International Conference on Computer Vision, 2003.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. e Jain, R. 2000. "Content-based image retrieval at the end of the early years." *IEEE Transactions on Pattern Analysis and Machine Intelligence, ,* no. 22 (12):1349-1380.

Smith, M., Baldwin, I., Churchill, W., Paul, R. e Newman, P. 2009. "The new college vision and laser data set." *The International Journal of Robotics Research*, no. 28 (5):595-599.

Stommel, M. e Herzog, O. 2009. "Binarising SIFT-Descriptors to Reduce the Curse of Dimensionality in Histogram-Based Object Recognition." In *Signal Processing, Image Processing and Pattern Recognition*, edited by Dominik Ślęzak, SankarK Pal, Byeong-Ho Kang, et al., 320-327. Springer Berlin Heidelberg.

Strecha, C., Bronstein, A. M., Bronstein, M. M. e Fua, P. 2012. "LDAHash: Improved Matching with Smaller Descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 34 (1):66-78.

Sunderhauf, N. e Protzel, P. 2011. BRIEF-Gist - closing the loop by simple means. In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 25-30 Sept. 2011.

Sunderhauf, N. e Protzel, P. 2012. Towards a robust back-end for pose graph SLAM. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 14-18 May 2012.

Swain, M. e Ballard, D. 1991. "Color indexing." *International Journal of Computer Vision*, no. 7 (1):11-32.

Swets, J. A. 1963. "Information Retrieval Systems." *Science*, no. 141 (3577):245-250.

Takala, V., Ahonen, T. e Pietikäinen, M. 2005. "Block-Based Methods for Image Retrieval Using Local Binary Patterns." In *Image Analysis*, edited by Heikki

Kalviainen, Jussi Parkkinen e Arto Kaarna, 882-891. Springer Berlin Heidelberg.

Tamura, H., Mori, S. e Yamawaki, T. 1978. "Textural Features Corresponding to Visual Perception." *IEEE Transactions on Systems, Man and Cybernetics*, no. 8 (6):460-473.

Tapus, A., Tomatis, N. e Siegwart, R. 2006. "Topological Global Localization and Mapping with Fingerprints and Uncertainty." In *Experimental Robotics IX*, edited by Marcelo Ang, Jr. e Oussama Khatib, 99-111. Springer Berlin Heidelberg.

Tax, D. M. J., Breukelen, M. V., Duin, R. P. W. e Kittler, J. 2000. "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, no. 33:1475-1485.

Thrun, S., Burgard, W. e Fox, D. 2005. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*: The MIT Press.

Timofte, R., Tuytelaars, T. e Van Gool, L. 2013. "Naive Bayes Image Classification: Beyond Nearest Neighbors." In *Computer Vision – ACCV 2012*, edited by Kyoung Lee, Yasuyuki Matsushita, James Rehg, et al., 689-703. Springer Berlin Heidelberg.

Tomatis, N., Nourbakhsh, I. e Siegwart, R. 2003. "Hybrid simultaneous localization and map building: a natural integration of topological and metric." *Robotics and Autonomous Systems*, no. 44 (1):3-14.

Topi, M., Timo, O., Matti, P. e Maricor, S. 2000. Robust texture classification by subsets of local binary patterns. In Proc. of the 15th International Conference on Pattern Recognition, 2000.

Torralba, A. 2003. "Contextual Priming for Object Detection." *International Journal of Computer Vision*, no. 53 (2):169-191.

Torralba, A., Fergus, R. e Weiss, Y. 2008. Small codes and large image databases for recognition. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 23-28 June 2008.

Torralba, A., Murphy, K. P., Freeman, W. T. e Rubin, M. A. 2003. Context-based vision system for place and object recognition. In Proc. of the 9h IEEE International Conference on Computer Vision, 13-16 Oct. 2003.

Turk, M. e Pentland, A. 1991. "Eigenfaces for recognition." *Journal of Cognitive Neuroscience*, no. 3 (1):71-86.

Ulrich, I. e Nourbakhsh, I. 2000. Appearance-based place recognition for topological localization. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 2000.

Valgren, C. e Lilienthal, A. J. 2010. "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments." *Robotics and Autonomous Systems*, no. 58 (2):149-156.

Vasconcelos, N. 2004. "Minimum probability of error image retrieval." *IEEE Transactions on Signal Processing*, no. 52 (8):2322-2336.

Ventura, J. e Hollerer, T. 2011. Fast and scalable keypoint recognition and image retrieval using binary codes. In Proc. of the IEEE Workshop on Applications of Computer Vision (WACV), 5-7 Jan. 2011.

Viola, P. e Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In Proc. of the Computer Vision and Pattern Recognition., 2001.

Wan, X. e Kuo, C. C. J. 1996. Color distribution analysis and quantization for image retrieval. In Proc. of the Storage and Retrieval for Still Image and Video Databases IV, 1996.

Wand, P. e Jones, C. 1994. *Kernel Smoothing*: Taylor & Francis.

Wang, J., Cipolla, R. e Zha, H. 2005. Vision-based Global Localization Using a Visual Vocabulary. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 2005, at Barcelona, Spain.

Wark, T. e Sridharan, S. 2001. "Adaptive fusion of speech and lip information for robust speaker identification." *Digital Signal Processing*, no. 11 (3):169-186.

Weiss, Y., Fergus, R. e Torralba, A. 2012. Multidimensional spectral hashing. In Proc. of the 12th European conference on Computer Vision - Volume Part V, 2012, at Florence, Italy.

Werner, F., Maire, F. e Sitte, J. 2009. "Topological SLAM Using Fast Vision Techniques." In *Advances in Robotics*, edited by Jong-Hwan Kim, ShuzhiSam Ge, Prahlad Vadakkepat, et al., 187-196. Springer Berlin Heidelberg.

Werner, F., Sitte, J. e Maire, F. 2007. On the Induction of Topological Maps from Sequences of Colour Histograms. In Proc. of the Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications, 3-5 Dec. 2007.

Williams, L. e Ledwich, S. 2004. Reduced SIFT Features For Image Retrieval and Indoor Localisation. In Proc. of the Australian Conference on Robotics and Automation, 2004, at Canberra, Australia.

Wong, S. K. M. e Yao, Y. Y. 1995. "On modeling information retrieval with probabilistic inference." *ACM Transactions on Information Systems*, no. 13 (1):38-68.

Wu, T. F., Lin, C. J. e Weng, R. C. 2004. "Probability estimates for multi-class classification by pairwise coupling." *The Journal of Machine Learning Research*, no. 5:975-1005.

Xia, W. e Kuo, C. C. J. 1998. "A new approach to image retrieval with hierarchical color clustering." *IEEE Transactions on Circuits and Systems for Video Technology*, no. 8 (5):628-643.

Yagnik, J., Strelow, D., Ross, D. A. e Lin, R.-s. 2011. The power of comparative reasoning. In *International Conference on Computer Vision*: IEEE Computer Society.

Zhang, H. 2011. BoRF: Loop-closure detection with scale invariant visual features. In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), 9-13 May 2011.

Zhang, Y. e Callan, J. 2001. Maximum likelihood estimation for filtering thresholds. In Proc. of the ACM SIGIR conference on Research and Development in Information Retrieval, 2001, at New Orleans, Louisiana, USA.